The Bayesian framework for inverse problems in heat transfer

Jari P Kaipio[*][†] Colin Fox[‡]

March 3, 2010

Abstract

The aim of this paper is to provide researchers dealing with inverse heat transfer problems a review of the Bayesian approach to inverse problems, the related modelling issues, and the methods that are used to carry out inference. In Bayesian inversion, the aim is not only to obtain a single point estimate for the unknown, but rather to characterize uncertainties in estimates, or predictions. Before any measurements are available, we have some uncertainty in the unknown. After carrying out measurements, the uncertainty has been reduced, and the task is to quantify this uncertainty, and in addition to give plausible suggestions for answers to questions of interest. The focus of this review is on the modelling-related topics in inverse problems in general, and the methods that are used to compute answers to questions. In particular, we build a scene of how to handle and model the unavoidable uncertainties that arise with real physical measurements. In addition to giving a brief review of existing Bayesian treatments of inverse heat transfer problems, we also describe approaches that might be successful with inverse heat transfer problems.

1 Introduction

1.1 Inverse problems

The classical definition of a well-posed problem, due to Hadamard, is that the solution exists, is unique and depends continuously on data [1]. If any of these conditions is not fulfilled, the problem is called *ill-posed*.

[*] Department of Mathematics, University of Auckland, Auckland, New Zealand

[†] Department of Physics and Mathematics, University of Eastern Finland, Kuopio, Finland

[‡] Department of Physics, University of Otago, Dunedin, New Zealand

Correspondence: jari@math.auckland.ac.nz

With well-posed problems, the existence problem is circumvented by looking for generalized solutions such as least squares solutions, while the uniqueness problem is approached by considering minimum-norm solutions [2]. In practice, the (lack of) continuity means that with an ill-posed problem, small errors in the data or the associated models and mappings may cause very large errors in the solution. Since measurements are always noisy, and observation models are always only *models*, ill-posed problems could be characterized as being problems that cannot be solved using straightforward least squares or minimum norm approaches. Well-posed problems by definition do not possess these problems and are stable in this sense. In this paper, we adopt the common practice of using interchangeably the notions of an ill-posed problem and an inverse problem.

The counterparts of inverse problems are the associated *forward problems.* Loosely speaking, if we knew the answer to an inverse problem, the solution of the forward problem could be interpreted as computing (predicting) the associated noiseless measurements. For example, the solution of the one-dimensional heat equation with known initial and boundary conditions would be tagged as a forward problem since it is stable. On the other hand, solving for the initial conditions when the boundary conditions are known and the temperature distribution is given at a time $T > 0$, is a highly unstable problem and is thus an ill-posed inverse problem [3, 4]. The more stable the forward problem is, the less stable is the corresponding inverse problem.

1.2 Deterministic framework for inverse problems

In the deterministic framework for inverse problems, the solution is interpreted as an unknown parameter, vector or a function. If the unknown takes, for example, the form of a projection[1], the projection coordinates are assumed to be completely unknown in the sense that any projection is equally plausible and acceptable, save for possible constraints such as positivity of the solution.

Naturally, all parameter estimation problems do not exhibit ill-posed nature [5]. Furthermore, it is sometimes difficult to make the distinction between a stable parameter estimation problem and an ill-posed problem when dealing with practical computational problems. For example, in the strict mathematical sense, an invertible finite dimensional linear problem is never discontinuous. However, these problems may well exhibit all practical problems associated with an ill-posed problem. On the other hand, using the projection approach with a low-dimensional subspace, an extremely ill-posed inverse problem is seemingly turned into a stable one. Referring to the above problem related to finding the initial condition, finding the best *constant* initial temperature to match the data at time T is a stable one

[1] With unknown functions, one usually approximates the function x as a projection Px onto a subspace spanned by a set of functions, say φ_k, so that we write $Px = \sum_k x_k \varphi_k$, where x_k are the projection coordinates. The task is then to solve for these projection coordinates and thus for Px.

and can be solved as a least squares problem. But this is not the same problem as finding the spatially inhomogeneous initial condition. Mathematically, one has then solved the (orthogonal) projection of the initial condition to the (one-dimensional) subspace spanned by the constant function.

The history of parameter estimation problems dates back at least to Gauss in the framework of general parameter estimation [6], and Laplace who developed inverse (or Bayesian) probability methods [7]. Historically, most of the parameter estimation problems studied have been stable ones. Problems with ill-posed nature were sometimes considered as unsolvable and left as such. Until the 1960's, several disciplines in physics, engineering and other fields developed proprietary methods to deal with important practical problems without a unifying general mathematical and statistical theory.

The general theory of ill-posed problems has been developed since the 60's, most notably by Tikhonov and others [8, 9, 10, 11] and produced a number of different approaches, such as truncated singular value decomposition, Tikhonov regularization, stopped iterative methods, to accompany the obvious projection approaches, see for example [12, 13, 14, 3, 15]. These methods are referred to as *regularization methods*.

Research in regularization methods has traditionally focused on two types of issues: the uniqueness and stability of the analytical errorless problems, and the convergence of the solution to the minimum norm solution when the norm of the noise tends to zero. The structure of the noise is almost always neglected and typically the norm of the errors is expected to be known. In most real world problems, the measurement errors do not vanish and, furthermore, their level (norm) is not known accurately. Also, model errors have very seldom been considered [16] for reasons that become apparent later in this paper. With real world problems, the (effects of) modelling errors tend to dominate the measurement errors, for examples related to electrical impedance tomography, see [17, 18, 19].

Regularization methods are not based on explicit models for the unknowns, except in the case of some projection methods [20, 21] in which the subspaces are constructed from explicit models for the unknowns. However, these methods can be argued to employ implicit models, which is most manifest in the case of truncated singular value decomposition [17]. Also, the deterministic methods usually seek only to find a single solution for the problem, possibly with some error estimates based, for example, on sensitivity analysis. These error estimates do not, however, generally bear any solid (statistical) interpretation. Most importantly, regularization methods are implicitly based on a number of assumptions which may not be valid [17]. For example, using 2-norms (cf. least squares) usually refers to an additive noise model with independent identically distributed Gaussian errors.

1.3 Bayesian framework for inverse problems

While the statistical framework has been systematically employed when solving stable parameter estimation problems throughout the 20th century [5], it seems to have been a notable framework with inverse problems only in geophysics and astronomy before the 90's [22, 23, 24, 25].

In contrast to the deterministic approach and the related regularization methods, the Bayesian approach for solving inverse problems is not a method. Rather, it can be described as a framework for the modelling of the entire problem in terms of probability in order to allow for inference, that is, the posing of questions in terms of statistics, and giving answers to these questions[2].

While in the deterministic paradigm one attempts to obtain a single solution for the interesting unknown, in the Bayesian framework the essence is to *explore* the posterior distribution to determine the uncertainty in the unknowns given the measurements and (prior) uncertainty inherent in all models. The exploration calls for computing different *point estimates* and *spread estimates*, as well as *marginal distributions* of individual unknowns or sets of unknowns. Also, probabilities of *events* are eventually often of interest and can be computed within this framework. In most cases, this calls for *sampling* of the posterior distribution [26].

The most important sampling algorithms are called Markov chain Monte Carlo methods (MCMC). If we had a large number of samples from a probability distribution, all statistical question that are related to the associated random variables could be answered with sample averages. Unfortunately, the implementation of sampling methods for inverse problems can turn out to be a tricky business due to the typically high number of unknowns and the narrowness of the distributions.

Bayesian inversion is a hierarchical process which first calls for the modelling of the measurement process and the unknown, with special reference to the actual uncertainties of the models. These models together with the measurements fix the uncertainty of the unknowns *given* the measurements. Formally, this uncertainty is given in the form of the *posterior distribution* which is then subject to exploration, for example, using MCMC sampling. It is of central importance that the modelling of the measurement process and the modelling of the unknowns are carried out *completely separately*. This is not usually the case with regularization methods in which a change in measurement setting may change the implicit model for unknowns.

Regularization methods and Bayesian inversion results are difficult to compare to

[2] The terms *Bayesian inverse problems* and *statistical inverse problems* have been used interchangeably in recent times. The aspect of regularization theory which takes into account the measurement error distribution is also referred to as statistical inversion. The related theory is, however, based on the same interpretation and problem formulation of inverse problems as the deterministic regularization theory. In particular, no explicit modelling of the unknowns is usually carried out.

each other since the Bayesian modelling is almost invariably more extensive and thus contains information that is not used with regularization methods. Furthermore, while regularization methods focus on providing a single (numerically) stable answer to the problem, the aim in Bayesian inversion is to provide point estimates *together with* systematic assessment of reliability and posterior uncertainty. The Bayesian approach invariably calls for significantly more elaborate modelling of the overall problem and prior uncertainties than the deterministic approaches. Furthermore, in the deterministic (regularization) framework it is completely impossible to answer questions such as *"What is the probability of thermal conductivity at the center of the slab being smaller than $\kappa_{\max}$?"* or, ultimately, *"What is the probability that the temperature gradients are so large that a composite will develop cracks?"*

In general, the Bayesian framework allows for combining data from different modalities in a single formulation in a straightforward, albeit sometimes tedious, manner. For example, let the current problem be to estimate the spatially inhomogeneous thermal conductivity based on a measurement setting. Let separate data be earlier acquired which was used to estimate the specific heat. The conventional approach would be to estimate a point estimate for the specific heat first, and use this estimate in the model used for the estimation of the thermal conductivity. In the Bayesian approach, all these data and the related uncertainties can be embedded in a single formulation. The resulting estimates for the thermal conductivity can be significantly better than the traditional approach which usually proposes unrealistically small posterior uncertainty.

Naturally, the posterior uncertainty can be, and often is, still very large. But this (large uncertainty) is an important piece of information in itself and indicates that the current data does not allow us to draw affirmative conclusions. Again, regularization methods do not provide us with statistically meaningful ways to assess the reliability of the solutions.

1.4 Modelling and methods

The notion of a "method" is used rather loosely in the literature in general. In this paper we require that a (numerical) "method" is something that should not affect the final outcome. It is not uncommon in the scientific literature to see studies in which two minimization methods are compared for the minimization of a fixed functional so that the different characteristics of the minimizers are explained to be related to the minimization algorithm. Of course, if the minimizers differ from each other, at least one of the two minimization algorithms has not converged to the (global) minimum. The functional here represents "the model" and the minimization algorithms are "methods".

We find it of central importance to distinguish models from methods. Fortunately, in the Bayesian framework this is usually straightforward. For example, the forward problems are often induced by partial differential equations and the related initial-boundary value problems. There are different approaches to approximate the mapping from the unknowns to the measurements, such as finite difference and finite element methods. When these are constructed properly, the predictions should be essentially the same. On the contrary, how the boundary conditions are modelled, can have a profound effect on the predictions.

To take the notion further, when implementing a Metropolis-Hastings sampling algorithm, one has a choice over the proposal distributions. The choice of proposal distribution *should bear no effect* to the estimates and answers to specified questions. The choice may, however, have a significant effect on the convergence rate of the algorithm.

If any of the models is infeasible, the posterior model may also be such. If the posterior model is infeasible, it is of no avail to embark on the inference and vice versa, if the models are feasible but the samplers or other computational machinery is inefficient, no inference can be made. Although we might refer, for example, to a *likelihood distribution* in the sequel, all distributions *absolutely have to be understood as models only.* Also, we use the notions of *density* and *probability* interchangeably, although the former is the correct one.

1.5 Inverse problems in the Bayesian context

What makes inverse problems a special class of problems in Bayesian inference? There are a few related issues.

In many cases, the dimension of a feasible representation of the unknowns is significantly larger than the number of measurements. Thus, for example, a maximum likelihood estimate is impossible to compute. Even in cases in which the number of unknowns would be significantly smaller than the number of measurements, the structure of the forward problem is such that maximum likelihood estimates would still be unstable. Any approach using a stabilized or regularized likelihood method inherits the problems that are related to deterministic regularization methods.

In addition to the instability, the variances of the observation (likelihood) model are almost invariably much smaller than the variances of the models for the unknowns (priors). The posterior density is often extremely narrow and, in addition, may be a nonlinear manifold. Constructing samplers for such distributions is significantly more tricky than for more regular distributions.

When dealing with physical data, it is clear that the computational forward models, which form the central part of the likelihood models, are approximate at best. This fact together with the typical narrowness of the likelihood density can lead to *infeasible* posterior models, that is, the actual unknown may have essentially zero probability with respect to the posterior distribution. With inverse problems, this can easily happen when model uncertainties are underestimated, or their structure is poorly modelled. Thus, with inverse problems, the realistic analysis of the actual measurement system is a centrally important task. This task is, however, often neglected. Very often, the standard *ad hoc* choice of an independent identically distributed Gaussian noise model is adopted, even though it may bear little resemblance to the actual noise process.

When taken in the strict sense, an extensive Bayesian analysis of an inverse problem has only seldom been carried out. There is, however, an extensive literature on statistical inverse problems from the Bayesian perspective, with the state of the art being distributed over a number of fields.

1.6 Organization of the paper

The rest of the paper is organized as follows. In Section 2, we discuss the classification of inverse heat conduction problems from the viewpoint of Bayesian inference. In Section 3, we discuss some simple examples of inverse heat transfer problems with the intent to motivate the readers to venture beyond least squares estimation and standard regularization approaches. In Section 4, we give a brief review of the philosophy and basic notions of Bayesian probability, model building, and posterior inference. In Section 5, we discuss MCMC methods and inference in general. In Sections 6 and 7, we treat the likelihood and prior models, respectively, that are most common with inverse problems. In Section 8, we discuss typical sources and types of uncertainties and approximations and how to treat these. In Section 9, we discuss nonstationary inverse problems. These are problems in which the unknowns are time-varying and that can be modelled with stochastic evolution models. In Section 10, we discuss computational issues such as model reduction, inverse crimes and computational models for the forward problems solvers. In Section 11, we discuss briefly miscellaneous topics such as model selection and problems with variable dimensions. In Section 12, we give a brief review of existing statistical inversion analyses of inverse heat transfer problems. In Section 13, we draw conclusions.

2 Classification of inverse heat transfer problems

Inverse heat transfer problems are considered in many excellent texts [27, 28, 29, 30, 4]. We follow the classification used by Özişik and Orlande [4] who consider problems in inverse heat *conduction*, *convection*, and *radiation*, according to which mode is dominant in heat transfer. The dynamics of heat within a region of space Ω is described in terms of the temperature $T(t, \vec{r})$ for $\vec{r} \in \Omega$ and t in some time interval S. We will consider cases in which Ω is a fixed region in 1-, 2-, or 3-dimensions[3],

[3] We note, however, that it is straightforward to extend the formalism we give to treat regions that vary with time.

and the time interval $S = \{t : 0 < t < t_{\mathrm{F}}\}$ represents times after some initial state at $t = 0$ up to time t_{F} that we may take to be infinite.

We first discuss the forward and inverse problems in inverse heat conduction in some detail, and then briefly discuss inverse heat convection and radiation as extensions.

2.1 Forward problem for heat conduction

For problems in heat conduction, heat flow within the medium is characterized by the *thermal conductivity* $\kappa(\vec{r})$ and the *heat capacity* $c(\vec{r}) = \rho c_p$ where ρ is the density of the medium and c_p is its specific heat.

In a fairly general formulation of heat conduction, the temperature $T(t, \vec{r})$ is governed by the initial boundary value problem (IBVP)

$$c\frac{\partial T}{\partial t} - \nabla \cdot (\kappa \nabla T) = q(t, \vec{r}) \quad \vec{r} \in \Omega, t > 0, \tag{1a}$$

$$k\frac{\partial T}{\partial n} = q_{\mathrm{N}}(t, \vec{r}) \quad \vec{r} \in \partial\Omega_{\mathrm{N}}, t > 0, \tag{1b}$$

$$T = T_{\mathrm{D}}(t, \vec{r}) \quad \vec{r} \in \partial\Omega_{\mathrm{D}}, t > 0, \tag{1c}$$

$$T = T_0(\vec{r}) \quad \vec{r} \in \Omega, t = 0. \tag{1d}$$

These equations model the situation where a medium occupying region Ω is initially at temperature $T_0(\vec{r})$, and is subsequently subject to heat sources in the interior with heat rate density $q(t, \vec{r})$, sources of heat flux $q_{\mathrm{N}}(t, \vec{r})$ on part of the surface denoted $\partial\Omega_{\mathrm{N}}$, while the remainder of the boundary denoted $\partial\Omega_{\mathrm{D}}$ is held at temperature $T_{\mathrm{D}}(t, \vec{r})$. The Neumann condition (1b) is often written as a convective condition, as discussed in Section 2.4.

We write the heat conduction operator as

$$L = c\frac{\partial}{\partial t} - \nabla \cdot (\kappa \nabla)$$

so that the spatial part is formally self-adjoint and positive definite, since $\kappa(\vec{r}) > 0$ [31]. When κ does not depend on T, or can be assumed to be a constant over the temperature range, the system in (1) is a *linear* boundary value problem and its solution is essentially a problem in linear analytical or numerical methods. When κ does depend on T, the system in (1) is *nonlinear*, requiring implicit solution methods.

When the thermophysical coefficients c and κ are known, as are all source terms q, q_{N}, T_{D}, and T_0, (and the geometry), the system in (1) defines a classical initial-boundary value problem that uniquely determines the temperature $T(t, \vec{r})$ at all points $\vec{r}$ in Ω, and for all times $t > 0$.

In cases where the finite speed of heat propagation is important, the hyperbolic heat conduction equation

$$c\left(\frac{\partial T}{\partial t}+\tau_{\mathrm{r}}\frac{\partial^2 T}{\partial t^2}\right)-\nabla\cdot(k\nabla T)=q\left(t,\vec{r}\right)+\tau_{\mathrm{r}}\frac{\partial q\left(t,\vec{r}\right)}{\partial t} \tag{2}$$

is used in place of (1a) [32, 33]. Here τ_{r} is a relaxation time (or phase lag in heat flux), giving a thermal wave speed of $\sqrt{\kappa/c\tau_{\mathrm{r}}}$.

Green's function solution. When the coefficients c and κ in system (1) do not depend on the temperature T, a formal solution may be written in terms of the Green's function, allowing a clear understanding of the nature of the forward and inverse problems. For a general inhomogeneous medium, the Green's function is not available in closed form, and must be computed numerically. We note, however, in a problem with the same structure as a stationary inverse heat conduction problem, an efficient computational scheme for Bayesian inversion has been developed that actually maintains numerically evaluated Green's functions [34], while an eigenfunction expansion was employed in [35].

The Green's function for the boundary-value problem in system (1) is the unique solution to the auxiliary equation $Lg(t,\vec{r}|\tau,\vec{\xi}\,)=\delta(\vec{r}-\vec{\xi}\,)\delta\left(t-\tau\right)$, for $(t,\vec{r})$ and $(\vec{\xi},\tau\,)\in\Omega\times S$, that satisfies the homogeneous[4] form of the boundary conditions, that is, with $q_{\mathrm{N}}=0$, $T_{\mathrm{D}}=0$, and $T_0=0$, for all t. This function gives the temperature at the field point $(t,\vec{r})$ when there is a unit heat source localized in space and time at the source point $(\tau,\vec{\xi}\,)$, with no other sources of heat. It follows that this function is causal, i.e. $Lg(t,\vec{r}|\tau,\vec{\xi}\,)=0$ for $t<\tau$ [31]. Solutions to the system (1) can then be written as

$$\begin{aligned}T\left(t,\vec{r}\right)=&\int_{\Omega\times S}q\left(\tau,\vec{\xi}\right)g\left(t,\vec{r}|\tau,\vec{\xi}\right)\mathrm{d}\tau\mathrm{d}\vec{\xi}\\&+\int_{\partial\Omega_{\mathrm{N}}\times S}q_{\mathrm{N}}\left(\tau,\vec{\xi}\right)g\left(t,\vec{r}|\tau,\vec{\xi}\right)ds\left(\xi\right)d\tau\\&-\int_{\partial\Omega_{\mathrm{D}}\times S}T_{\mathrm{D}}\left(\tau,\vec{\xi}\right)\kappa\left(\xi\right)\frac{\partial g\left(t,\vec{r}|\tau,\vec{\xi}\right)}{\partial n\left(\xi\right)}ds\left(\xi\right)d\tau\\&+\int_{\Omega}c\left(\xi\right)T_0\left(\xi\right)g\left(t,\vec{r}|\xi,0\right)d\xi\end{aligned}$$

for $(t,\vec{r})$ in the region $\Omega\times S$.

Since the Green's function depends on the thermophysical properties c and κ (and the geometry of the medium), but not on the heat sources in (1), we see

[4] This use of *homogeneous* refers to the boundary conditions being zero. It is not to be confused with a *homogeneous material* which is one that has properties that do not vary with location.

that the temperature T, and hence the forward map, is a *linear* function of the heat sources. However, the Green's function depends nonlinearly on c and κ, and hence the temperature T is, in general, a *nonlinear* function of those thermophysical properties.

When κ and c are constant the causal free-space fundamental solution in n space dimensions that satisfies $Lg(t,\vec{r}|\tau,\vec{\xi}) = \delta(\vec{r}-\vec{\xi})\delta\left(t-\tau\right)$ with homogeneous initial conditions, has the simple form [31, vol. 2, p. 60]

$$g\left(t,\vec{r}|\tau,\vec{\xi}\right) = \frac{1}{c\left(4\pi\alpha\left(t-\tau\right)\right)^{n/2}}\exp\left\{-\frac{\|r-\xi\|^2}{4\alpha\left(t-\tau\right)}\right\} \tag{3}$$

for $t > \tau$, where $\alpha = \kappa/c$ is the thermal diffusivity. We will use this fundamental solution later in Section 3.

2.2 Inverse problems in heat conduction

Inverse heat conduction problems occur when the temperature T is measured at one or several locations, and the aim is to estimate one, or more, of the arguments in the forward map. The purpose of some inverse heat conduction problems is to determine these thermophysical properties, while in others these properties are assumed known. We will classify the inverse problem according to which quantities are unknown. These are usually referred to as the *primary unknowns* in the engineering literature, or *parameters* in statistics.

We denote by d_k the measured temperature at the location and time $(\vec{r}_k, t_k)$ for $k = 1, 2, \ldots, M$ when there are M measurements in all. Typical experimental setups are when the temperature is measured at M_r locations $\vec{r}_1, \vec{r}_2, \ldots, \vec{r}_{M_r}$ at each of the M_t times $t_1, t_2, \ldots, t_{M_t}$ in which case a simple enumeration of measurements is $k = n + M_r\left(m-1\right)+1$. Thus d_k, for $k = 1, 2, \ldots, M = M_rM_t$, is the temperature measured at location and time $(t_m, \vec{r}_n)$. Hence we have made the usual assumption that measurements are made instantaneously a single known point. Practical sensors are of finite size and have a finite time response, leading to uncertainty in the exact location of measurements. Methods for dealing with these additional uncertainties are discussed in Section 8, and in [36, 37].

For computational purposes, the forward map is the composed operator consisting of the solution of the IBVP (1) and the observation process that can be as simple as a projection of the temperature field $T\left(t,\vec{r}\right) \mapsto \left\{T\left(t_1,\vec{r}_1\right), T\left(t_2,\vec{r}_2\right), \ldots, T\left(t_M,\vec{r}_M\right)\right\}$ onto the finite set of measurements.

Thermal conductivity. In the inverse problem for *thermal conductivity*, the function k is unknown, while all other quantities appearing as coefficients in IBVP (1) are known. A typical experimental setup for noninvasive measurement is to apply known heat fluxes to the entire surface of the medium, i.e. q_{N} is prescribed and

$\partial\Omega_{\mathrm{N}} = \partial\Omega$, with known initial temperature, have no internal heat sources so $q = 0$, and measure the resulting temperature at several locations on the surface.

In this paper, we mainly consider the case in which k does not depend on temperature. There are, however, many materials and temperature regimes where the system (1) holds where k depends appreciably on the dependent variable T, see for example [38].

Thermal capacity. In systems where the relaxation time τ_{r} is appreciable, such as the rapid heating due to laser pulses [39], heat flow is modelled by (2).

Initial temperature distribution. Estimation of the *initial condition* is typically made from measurements of the temperature at the surface at some time $t_0 > 0$, when there are no other heat sources. In the system (1), T_0 is the primary unknown while q, q_{N}, and T_{D} are all fixed at zero. This is sometimes called "thermal archaeology", for example in [3], and is an example of an extremely ill-posed problem with eigenvalues of the forward operator decaying as $\exp(-k^2)$.

Heat sources. The inverse problem for *source strength* requires, in general, interpreting measurements of temperature in terms of internal heat sources $q\left(t,\vec{r}\right)$ that vary in space and time. A typical problem setup is where the medium has a known initial temperature and the internal heat source is to be determined for time $t > 0$ using measurements of temperature at locations on the boundary of the medium.

Conductance at interfaces. Determining the thermal conductance at contact interfaces nearly always requires solving an inverse problem as the value of conductance depends on local properties such as surface roughness and contact pressures [40] and is seldom known a priori. Contact conductance inverse problems can arise in situations that have fixed boundaries, time-varying fixed boundaries such as occurs when parts in a machine make contact and separate [4, Examples 3-7],[41], or with moving boundaries such as the interface between different phases in medium. In the latter case the boundary is defined implicitly by the relationship between temperature and phase, and therefore requires an implicit computational method for the solving the system in (1).

Truncation boundaries. A further estimation problem arises in situations where numerical solution of the system in (1) necessitates truncating the computational domain, thereby creating boundaries with an unknown relationship between temperature and heat flux. We discuss unknown boundary conditions in Section 7.3.

2.3 Forward and inverse problems in heat convection

Heat convection occurs when motion of the medium itself leads to transport of heat. Then the material derivative is required, and the equation

$$c\left(\frac{\partial T}{\partial t} + u\left(\vec{r}\right)\cdot\nabla T\right) - \nabla\cdot\left(k\nabla T\right) = q\left(t,\vec{r}\right) \quad \vec{r}\in\Omega, t>0 \tag{4}$$

holds in place of (1a). Here, the velocity of the medium is denoted $u(\vec{r})$. In *free convection* problems, the motion is caused partially by thermal effects and so $u(\vec{r})$ in unknown a priori, leading to a system of coupled PDE's, in which the other PDE describes the flow and has a driving term that depends on temperature. In *forced convection* problems, the flow is externally generated and the velocity field $u(\vec{r})$ may be known a priori, or can be determined independently of thermal effects. Then the inverse heat transfer problems have the same basic structure as those for inverse heat conduction. That is, the inverse source problems lead to linear inverse problems while inverse problems for thermophysical properties or boundary location are nonlinear.

2.4 Forward and inverse problems in heat radiation

For opaque materials, heat radiation is a surface phenomenon and may be treated as a boundary condition, with heat transport within the medium governed by the heat conduction or convection equations in Sections 2.1 and 2.3. Radiative heat transfer at a surface is nonlinear. A common linearized surface radiation condition is given by replacing the flux boundary condition (1b) with the Newton condition

$$k\frac{\partial T}{\partial n} = \beta(T - T_{\mathrm{Amb}}) \quad \vec{r} \in \partial\Omega_{\mathrm{N}}, t > 0. \tag{5}$$

that also represents a convective boundary condition.

Radiation in transparent materials is a bulk scattering phenomenon, that is outside the scope of this paper. See [4] for further details.

2.5 Numerical methods

While there are a few idealized problems in inverse heat transfer for which analytic solution of the forward map is possible (we give an example in Section 3), in most practical cases, accurate simulation of data requires computer evaluation of a discretized version of equations (1).

Time-independent problems. Discretization of the space derivatives most commonly uses the finite element method (FEM), the finite difference method (FDM) [42], or the boundary element method (BEM).

The BEM uses a discrete form of the boundary integral equations [43, 44, 45] that express fields within regions of constant thermal properties in terms of values at the boundary of the region. Hence BEM is applicable to problems where the thermophysical properties are piecewise constant in sub-domains, and has been used extensively in IHT [43, 46]. The BEM discretization results in a system with the form

$$\left(K + \frac{1}{2}I\right)T = Hq \tag{6}$$

where T is a vector of Dirichlet values, q is the vector of Neumann values, and the matrices K and H are dense, non-symmetric, and of size $N_\mathrm{b} \times N_\mathrm{b}$ when the internal and external boundaries are partitioned into N_b boundary elements in total.

In FEM discretization of equations (1), the region Ω is usually discretized as the union of triangular elements, each with constant thermophysical properties, with the temperature interpolated between nodes by piecewise linear functions [47, 48, 36], giving the expansion

$$T(\vec{r}, t) = \sum_{\ell=1}^{N} T_\ell(t) \varphi_\ell(\vec{r}). \tag{7}$$

The FEM discretization of stationary problems results in a linear system to be solved of the form

$$KT = f \tag{8}$$

where T is a vector of nodal values, over the whole mesh, and f is a forcing vector and K is the stiffness matrix modified for the Dirichlet conditions corresponding to non-zero heat sources. Notably, the matrix K is symmetric, sparse, and of size $N_\mathrm{e} \times N_\mathrm{e}$ when there are N_e nodes in the mesh.

Time dependent problems. Time-dependent problems are often solved using semi-discretization, or the 'method of lines', in which FDM, FEM, or BEM is used to discretize the space part of the IBVP (1), so as to obtain a coupled system of ordinary differential equations (ODEs) which is then solved using an ODE solver, typically a high order implicit Runge-Kutta (RK) method or similar method [48]. Evolution equations are given in Section 9.1. The classical Crank-Nicolson method for time-dependent heat equations is an example, using a suitable choices of FDM or FEM followed by Heun's implicit RK method [47, 49]. FEM methods can also be used for both space and time parts [47, 48]. More efficient numerical methods use fully adaptive methods that coarsen in both space and time directions, as solutions become smooth [50]. Fast multipole methods (FMM) implementing convolution kernels give some of the fastest BEM based methods [51].

For the theory and numerical approaches of finite element methods and methods for ordinary differential equations, see [52, 48, 53, 49].

3 Motivating examples

In this section, we give two examples that demonstrate some of the problems that can occur with least squares or regularized least squares approaches. The examples are chosen to allow simple analytic solutions so that the results are not obfuscated by numerical issues. We also briefly compare with results that would be given by Bayesian methods. Our hope is that the fundamental problems with the least squares approach in these simple examples will motivate readers to explore Bayesian methods.

For non-linear inverse problems, least squares estimates suffer from *stochastic bias*, which is a systematic offset in estimated values resulting from errors on measured temperatures, or other measurement uncertainties. Stochastic bias may be displayed in a simple form as follows: Let x be a (scalar) random variable from any distribution with mean $\mathbb{E}\left(x\right)=\mu$ and variance σ^2. Then, $\mathbb{E}(y)=\mathbb{E}\left(x^2\right)=\mu^2+\sigma^2$. That is, in the presence of noise on variable x, the mean of the square equals the square of the mean *plus the variance of the noise.* In a situation where we have measured x subject to noise, and the problem is to estimate the random variable $y=x^2$, simple estimation of y by squaring an estimate for x will lead to a systematic offset in the estimate of y. The bigger the noise, the bigger the offset. Stochastic bias is a very real effect that, for example, is one of the mechanisms that allows profit to be made in volatile markets, whether increasing or decreasing [54]. The first example demonstrates stochastic bias in least squares estimation.

The ability to calculate data-dependent or posterior variance is a distinct advantage of a Bayesian approach to inverse problems. In contrast, methods such as least squares are justified on the basis of the variance of the *estimator*, defined as the average variance over all possible measurements. Yet, in many practical inverse problems the variance of the estimator has little to do with the uncertainty in parameters estimated from the actual data set. The second example demonstrates the issue in a very simple setting, where the forward map is the identity function with uniform measurement errors.

3.1 Estimating heat capacity from impulsive heating

Consider the problem of determining the heat capacity c in a 3-dimensional homogeneous medium of large extent for which the thermal conductivity κ is known, perhaps from measurements of stationary heat flow. Since heat capacity affects the time derivative in the heat conduction equation, it is best measured using transient heating. We consider the idealized problem where an infinite medium is subject to unit impulsive heating at $r=0$ and time $t=0$, and the temperature at the point of heating is subsequently measured. For this problem the forward map is available analytically since the causal fundamental solution in equation (3) is the Green's function for this problem. In particular the noise-free temperature at time t is

$$T\left(t\right)=\frac{c^{1/2}}{\left(4\pi kt\right)^{3/2}}. \tag{9}$$

Consider the case in which the measurements d_i are made at times $t=it_{\mathrm{s}}$, $i=1,2,\ldots M$, and are subject to additive noise with zero mean and variance σ_{e}^2. Then, the least squares estimate of c is

$$\hat{c} = \arg\min_{c} \sum_{i=1}^{M} \left(d_i - b\frac{\sqrt{c}}{i^{3/2}} \right)^2$$

where $b = \left(4\pi\kappa t_{\mathrm{s}}\right)^{-3/2}$. The normal equations are solvable in this case, and give

$$\hat{c} = \left[\frac{\sum_{i=1}^{M} d_i / i^{3/2}}{b\sum_{i=1}^{M} 1/i^3} \right]^2 = w^2.$$

The term in the square brackets, denoted by w, is a weighted sum of random variables and hence w is a random variable with mean $\sqrt{c}$ and variance $\sigma_{\mathrm{e}}^2 / \left(b^2 \sum_{i=1}^{M} 1/i^3 \right) \to \sigma_{\mathrm{n}}^2 / \left(b^2 \zeta\left(3\right) \right)$ as $M \to \infty$. Here $\zeta\left(3\right) \approx 1.2021$ is the Riemann zeta function evaluated at 3 [55, Chapter 23]. Using the result quoted above, for a large (infinite) number of measurements, the least squares estimate of c has expected value $\hat{c} = c + \sigma_{\mathrm{e}}^2 / \left(1.2021 \times b^2\right)$. That is, in the presence of measurement error the least squares estimate is systematically biased for all M. Note that reducing the measurement error results in smaller bias, whereas increasing the number of measurements by increasing M actually increases the bias.

In this case the bias may be easily removed, by subtracting $\sigma_{\mathrm{e}}^2 / \left(1.2021 \times b^2\right)$. The result is a better estimator, having the same variance but lower bias, and is explicitly not the least squares estimator. Unfortunately, the obvious conclusion, that best fit to data is not the same as best fit to parameters, is not commonly observed in the inverse problems literature.

In most inverse problems, we are not able to determine the bias analytically, making the least squares estimate both biased and difficult to fix. The application of regularization actually compounds this problem. For example, for this example the Tikhonov regularized estimate may be calculated analytically, and has bias that is dependent on the *unknown* value of c, leaving an implicit problem to remove bias.

It is instructive to note that the quantity w is an unbiased estimator for $\sqrt{c}$, since the data is a linear function of $\sqrt{c}$. Hence the least squares estimate of $\sqrt{c}$ makes a good estimate, but its square is not a good estimate of c. This apparently paradoxical behavior is an example of how the algebra of random variables with uncertainty is quite different to the algebra of deterministic variables, see [56]. For this reason it is necessary to track the *distribution* of possible values that a variable can take, not just the single 'best' estimates. Maintaining and summarizing distributions over variables is a central component of the Bayesian approach.

Anticipating the framework in Section 4, we briefly describe how a Bayesian approach could solve this example. The likelihood function combines the forward map in (9) and the distribution over measurement error, or noise. As is typical in inverse problems, the range of the forward map is a small fraction of data space, and so the noise statistics may be determined from the measurements. In this case, at large times when $T(t) \approx 0$ the data solely consists of measurements of the noise. Hence sample statistics may be determined and the noise distribution modelled. For this simple single parameter estimation problem, an 'objective Bayesian' [57] analysis is feasible, by choosing the Jeffreys type [22, 58], or 'reference' [59], prior distribution that is invariant to choice of units for heat capacity, giving $\pi_{\mathrm{p}}(c) \propto c^{-1/2}$. This is an 'improper' prior distribution and would require modification if only a few measurements were available, although in such a case the data would be too poor to allow for reasonable estimation of c with any approach. When the data is adequate, the posterior mean gives a suitable point estimate for c. In the ideal case where the noise is *independent identically distributed* (iid) zero-mean Gaussian, and measurements are very accurate, there is little to choose between the least squares estimate and the posterior mean, except in computational cost. In most other circumstances the posterior mean does a much better job of estimating the unknown true heat capacity.

3.2 Uncertainty in estimates with uniform noise

Consider the simple case where a scalar quantity μ is measured directly, subject to uniform noise, with mean zero and width 2. Then the i^{th} measurement is

$$d_i = \mu + e_i$$

where each $e_i \sim \mathrm{Uniform}(-1,1)$ is uniformly distributed over the interval $[-1,1]$. Since $\mu - 1 \leq d_i \leq \mu + 1$ for all i, it follows that $\max\{d_i\} - 1 \leq \mu \leq \min\{d_i\} + 1$. In fact, these bounds are exactly what the measurements tell us about μ. We note that the likelihood distribution precisely expresses these bounds, and so they are automatically included in a Bayesian analysis.

The least squares estimate of μ from M measurements is easily seen to be

$$\hat{\mu}_{\mathrm{ls}} = \frac{1}{M}\sum_{i=1}^{M} d_i.$$

It is instructive to note that this estimate can lie *outside* the interval $[\max\{d_i\} - 1, \min\{d_i\} + 1]$, in which case it is not even consistent with the measured data and information on the distribution of the errors. For $K = 10$, this happens a little more that 30% of the time. Thus, in one out of every three experiments the least squares estimate is not even a possible value. More troublesome, in practice, is that the error often quoted for the least squares estimate has little to do with the actual uncertainty in the value of μ as determined by the data. From the considerations above, we see that $\mu \in \frac{1}{2}(\max\{d_i\} + \min\{d_i\}) \pm \frac{1}{2}(2 + \min\{d_i\} - \max\{d_i\})$ so the uncertainty is (certainly) $1 + \frac{1}{2}(\min\{d_i\} - \max\{d_i\})$. The mean-square-error for the least squares estimator of $1/\sqrt{3M}$ is often quoted as the error in the least squares estimate. Note that this is independent of the data. Three simulations for the case $\mu = 0$ and $M = 2$, returned the values $(d_1, d_2) = (-0.7477, 0.6688)$, $(-0.6112, -0.6136)$, and $(0.6278, -0.0376)$ giving estimates with posterior error ± 0.2918, ± 0.9988, and ± 0.6673. This is sometimes larger, and sometimes smaller than the least squares error of ± 0.4082.

Posterior error estimates are particularly informative when recovering spatially-varying parameters such as the thermal conductivity of an inhomogeneous material. It is clear on physical grounds that the spatial dependence of uncertainty in a reconstruction *must depend on the measurements*. For example, when a region of low thermal conductivity is surrounded by a region of high thermal conductivity, the outer region shields the inner region from external heat sources, since heat will preferentially flow through the outer region and not penetrate the inner region. Hence, the conductivity of the inner region cannot be accurately determined from measurements based on external heat sources. However, if the whole medium has similar thermal conductivity the heat can flow through all regions, resulting in more accurate estimation. Since the data reflects the distribution of thermal conductivity, the spatially varying error must be dependent on the data. An example of shielding by a region of high electrical conductivity, that shows up in posterior variance, is given in [60]. As mentioned above, the mean square error of the least squares estimate, or of any other fixed estimator, gives no clue to this effect.

4 Bayesian inference

In this section, we give a brief introduction to topics in Bayesian statistics. These topics are elaborated in later sections. We will first consider the case in which the primary unknown is the only unknown. In Section 4.5, we consider the case in which there are auxiliary unknowns.

For treatises of Bayesian statistics in general, we refer to [26, 61, 62, 63] and in connection with inverse problems [17, 64, 65].

4.1 Bayesian probability

We denote the unknown random variables with x and the measurements (data) with d. Complete statistical information of all the random variables is given by the joint distribution $\pi(x, d)$. This distribution expresses all the uncertainty of the random variables. Once the measurements d have been obtained, the uncertainty of the unknowns x is (usually) reduced. The measurements are now reduced from random variables to numbers and the uncertainty of x is expressed as the conditional distribution $\pi(x\,|\,d)$, which is also referred to as the *posterior distribution*. This distribution contains all information on the uncertainty of the unknowns x when the information on measurements d is utilized[5].

A schematic example in the case $x \in \mathbb{R},\ d \in \mathbb{R}$ is given in Fig. 1. The marginal distribution $\pi(x)$ is called the *prior distribution* and it represents the uncertainty of the unknown prior to obtaining the measurement. Two different conditional densities $\pi(x\,|\,d)$ with different measurements d are also shown. After the measurement the uncertainty regarding the unknown is significantly reduced.

The conditional distribution of the measurements given the unknown is called the *likelihood distribution* and is denoted by $\pi(d\,|\,x)$. The marginal distribution of the unknown is called the prior (distribution) and is denoted by $\pi(x)$. By the definition of conditional probability we have

$$\pi(x, d) = \pi(d\,|\,x)\pi(x) = \pi(x\,|\,d)\pi(d)\ . \tag{10}$$

Furthermore, the marginal distributions can be obtained by marginalizing (integrating) over the remaining variables, that is, $\pi(x) = \int \pi(x, d)\,\mathrm{d}d$ and $\pi(d) = \int \pi(x, d)\,\mathrm{d}x$. Note that after the measurement is obtained, $\pi(d)$ is a positive number. The following rearrangement is called Bayes' theorem

$$\pi(x\,|\,d) = \pi(d)^{-1}\pi(d\,|\,x)\pi(x)\ . \tag{11}$$

If we had access to the joint distribution, we could simply use the above definitions to compute the posterior distribution. Unfortunately, only in rare cases, the joint distribution is available in the first place. However, it turns out that in many cases the derivation of the likelihood density is a straightforward – if not always trivial – task. Also, a feasible probabilistic model for the unknown can often be obtained. Then one can use Bayes' theorem to obtain the posterior distribution. The key point here is that the posterior is obtained by using a (prior) model for the distribution of the unknown rather than the marginal density, which cannot be computed since the joint distribution is not available in the first place[6].

4.2 Point and spread estimates

Point estimates are the Bayesian counterpart of the deterministic "solutions". The most common point estimates are the *maximum a posteriori* estimate (MAP) and the

[5] To avoid confusion, all densities should have the associated subscript, here for example, $\pi_{x\,|\,d}(x\,|\,d)$. We will omit subscripts when the arguments specify the density unambiguously.

[6] Roughly speaking, the frequentist and Bayesian frameworks are separated by the following issue: If the prior is modelled separately (not by marginalization from the joint distribution), we are in the Bayesian framework.

conditional mean estimate (CM, also known as the minimum mean square estimate, MMSE). In the sequel, we consider the unknowns and measurements as random vectors (of finite dimension): $x \in \mathbb{R}^N$, $d \in \mathbb{R}^M$.

The computation of the MAP estimate is an optimization problem while the computation of the CM estimate is an integration problem:

$$\begin{aligned} x_{\mathrm{MAP}} &= \arg \max_x \pi(x\,|\,d) && (12) \\ x_{\mathrm{CM}} &= \mathbb{E}(x\,|\,d) = \int x\,\pi(x\,|\,d)\,\mathrm{d}x && (13) \end{aligned}$$

where arg reads as "argument of" the maximization problem, $\mathbb{E}(\cdot)$ denotes expectation, and the integral in (13) is an N-tuple integral.

The most common estimate of spread is the *conditional covariance*

$$\Gamma_{x|d} = \int (x - \mathbb{E}(x\,|\,d))(x - \mathbb{E}(x\,|\,d))^{\mathrm{T}}\,\pi(x\,|\,d)\,\mathrm{d}x \qquad (14)$$

Here, $\Gamma_{x\,|\,d}$ is an $N \times N$ matrix and the integral (14) refers to a matrix of associated integrals.

Often, the marginal distributions of single variables are also of interest. These are formally obtained by integrating over all other variables

$$\pi(x_\ell\,|\,d) = \int\limits_{x_{-\ell}} \pi(x\,|\,d)\,\mathrm{d}x_{-\ell} \qquad (15)$$

where the notation $(\cdot)_{-\ell}$ refers to all components *excluding* the ℓ^{th} component. Thus, (15) is an $(N-1)$-tuple integral. Furthermore, $\pi(x_\ell|d)$ is a function of a single variable, and can be visualized by plotting. The *credibility intervals* are the Bayesian counterpart to the frequentist confidence intervals, but the interpretation is different. Technically, a p%-credible interval is a subset which contains p% of the probability mass of the *posterior distribution.*

4.3 Linear models and Gaussian densities

The linear Gaussian, or normal, problems are an important class of inverse problems. For these problems, the answers to the most common questions often have analytical expressions. For this reason, Gaussian approximations for prior, likelihood and posterior models are often sought.

As an important example, we consider the additive error model in which the error e and x are jointly independent and the measurement model is linear:

$$d = Ax + e\ , \quad \pi(x,e) = \pi(x)\pi(e)$$

and $\pi(x) = \mathcal{N}(x_*, \Gamma_x)$ and $\pi(e) = \mathcal{N}(e_*, \Gamma_e)$. Then, the joint distribution of (y, x) is Gaussian and is thus completely specified by its mean and covariance only. Direct computation gives

$$\mathbb{E}\begin{pmatrix} d \\ x \end{pmatrix} = \begin{pmatrix} Ax_* + e_* \\ x_* \end{pmatrix} \tag{16}$$

$$\operatorname{cov}\begin{pmatrix} d \\ x \end{pmatrix} = \begin{pmatrix} A\Gamma_x A^{\mathrm{T}} + \Gamma_e & A\Gamma_x \\ \Gamma_x A^{\mathrm{T}} & \Gamma_x \end{pmatrix} \tag{17}$$

In the linear Gaussian case, all conditional distributions are Gaussian. It suffices therefore to compute the (conditional) means and covariances only. Furthermore, in the case of Gaussian distributions, the MAP and CM estimates are equal. From (16-17), for example, using the Schur complements, the posterior mean and covariance can be obtained

$$\begin{aligned} x_{\mathrm{CM}} &= x_* + \Gamma_x A^{\mathrm{T}}\big(A\Gamma_x A^{\mathrm{T}} + \Gamma_e\big)^{-1} \cdot \\ &\quad \big(d - Ax_* - e_*\big) \end{aligned} \tag{18}$$

$$\Gamma_{x|d} = \Gamma_x - \Gamma_x A^{\mathrm{T}}\big(A\Gamma_x A^{\mathrm{T}} + \Gamma_e\big)^{-1} A\Gamma_x \tag{19}$$

Another form for the posterior mean and covariance can be obtained by employing the matrix inversion lemma [66][7]

$$\Gamma_{x|d} = \big(A^{\mathrm{T}}\Gamma_e^{-1}A + \Gamma_x^{-1}\big)^{-1} \tag{20}$$

$$x_{\mathrm{CM}} = \Gamma_{x|d}\big(A^{\mathrm{T}}\Gamma_e(d - e_*) + \Gamma_x^{-1}x_*\big) \tag{21}$$

All marginal distributions of single variables or sets of variables can be obtained analytically by the Schur complements, see for example [17].

The simplicity of using Gaussian models applies only to the computation of the posterior statistics *given that the models are available.* Consider the case where the parametrization of the primary unknown is $x \in \mathbb{R}^{1000}$. The covariance Γ_x is a symmetric positive semi-definite matrix which means that Γ_x has about $500,000$ real numbers to be specified. The question is, how does one fix these numbers so that Γ_x is a feasible and sustainable model for the prior uncertainty? Some nontrivial constructions of Gaussian prior models are mentioned in Section 7.

There is a widespread confusion about the relationship between Bayesian inversion and Tikhonov regularization. Assume that the additive noise is zero mean Gaussian iid noise, that is, $e \sim \mathcal{N}(0, \sigma^2 I)$. Assume further that we can model the unknown as a (spatial) zero mean white noise process with covariance $\Gamma_x = \beta^{-2} I$,

[7] We have equivalence if all covariances and their inverses exist. If not, one of these forms may be feasible. In the general case, reparametrizations of the random variables may be necessary.

and that we can assume that the additive noise and the unknown are mutually independent. The conditional mean estimate is

$$x_{\mathrm{CM}} = \arg\min_x \left\{ \sigma^{-2} \|d - Ax\|^2 + \beta^2 \|x\|^2 \right\}. \tag{22}$$

Then, the CM (and MAP) estimates coincide exactly with the Tikhonov regularized solution

$$x_{\mathrm{Tik}} = \arg\min_x \left\{ \|d - Ax\|^2 + \alpha^2 \|x\|^2 \right\} \tag{23}$$

only if we have set $\alpha = \sigma\beta$. But in the Tikhonov regularization approach, we would use, for example, the Morozov discrepancy principle to adjust α so that the Tikhonov solution $x_{\mathrm{Tik}}(\alpha)$ satisfies

$$\|d - Ax_{\mathrm{Tik}}(\alpha)\| = \sqrt{N}\sigma \tag{24}$$

where it is assumed that σ is known exactly. Here, the notation $x_{\mathrm{Tik}}(\alpha)$ indicates that the estimate has been computed using a fixed α in (23).

Thus, it seems that we have done something that resembles Bayesian inversion. The particular problems here are the following. If the models for the additive noise and the unknown were feasible and could be supported, there is no reason to tamper with these indirectly by adjusting α. Moreover, how does one compute the posterior covariance, if the task has to be to minimize (23) subject to (24) in the first place?

4.4 Exploration by sampling

In the general case, point and spread estimates, as well as answers to other questions cannot be computed analytically. In such cases there are roughly two possibilities: either one tries to approximate the posterior model, for example with a Gaussian model or, one has to resort to the sampling methods.

The motivation behind sampling is straightforward. Assume that one has an ensemble of independent samples from the posterior distribution $\pi(x\,|\,d)$. We write $\{x^{(k)}, k = 1, \ldots, N_{\mathrm{s}}\} \sim \pi(x\,|\,d)$ where N_{s} is the number of samples (draws) from the posterior distribution. Then, the mean of *any function $g(x)$ of x* can be approximated by the sample average

$$\mathbb{E}(g(x)\,|\,d) = \int g(x)\, \pi(x\,|\,d)\, \mathrm{d}x \approx \frac{1}{N_{\mathrm{s}}} \sum_{k=1}^{N_{\mathrm{s}}} g\left(x^{(k)}\right). \tag{25}$$

Furthermore, the law of large numbers guarantees that (under quite mild assumptions) the variance of the sample average behaves like $\propto N_{\mathrm{s}}^{-1}$.

For example, for the posterior mean of x we would set $g(x) = x$ and for the posterior covariance $g(x) = (x - \mathbb{E}(x\,|\,d))(x - \mathbb{E}(x\,|\,d))^{\mathrm{T}}$. Furthermore, the answer to a question such as "What is the posterior probability that the k^{th} component of x is between 1 and 5", is obtained as

$$P(x_k \in [1,5]) \approx \frac{\text{number of samples with } x_k \in [1,5]}{N_{\mathrm{s}}}.$$

Furthermore, marginal densities of single random variables x_k are simply obtained by considering the k^{th} components of the samples only.

The key problem is, of course, how to obtain an ensemble from the posterior distribution in the first place. This turns out not to be a trivial problem. The most important class of methods for the generation of samples from an arbitrary probability distribution are the *Markov chain Monte Carlo* methods, which are discussed in Section 5.

4.5 Auxiliary variables and uncertainties

We very seldom face a problem in which the *primary* unknowns are the only unknowns. Most often, we have a number of *auxiliary* unknowns; We denote these variables with χ in a specific *parametrization* of the uncertainties.

An example of an almost ubiquitous auxiliary variable is the noise on measurements. In the sequel, we shall distinguish between the measurement errors e and other auxiliary unknowns, which we denote with μ, so that we have $\chi = (\mu, e)$. For example, consider a linear deconvolution problem

$$d = A_\mu x + e$$

where A_μ is the convolution operator with a convolution kernel that can be completely specified with a single real number $\mu > 0$. We assume that μ is not known exactly but can take any value in the interval $[\mu_1, \mu_2]$. Furthermore, we know that all values on this interval are not equally probable, and we would then aim to construct a model $\pi(\mu)$ accordingly. Furthermore, the additive noise (vector) e is known to obey the Gaussian probability distribution with known mean and covariance $e \sim \pi(e) = \mathcal{N}(e_*, \Gamma_e)$. Then, given the measurements d, we have the unknowns (x, μ, e) and thus $\chi = (\mu, e)$ when we are mainly interested in the reconstruction of x only.

In some fortunate cases, such as additive noise that is independent of other unknowns, we can marginalize over some variables. Specifically, if e and (x, μ) are mutually independent, we have

$$d - A_\mu x \sim \pi_e = \mathcal{N}(e_*, \Gamma_e)$$

and we have

$$\pi(d|x,\mu) = \int_e \pi(d, e|x, \mu)\,\mathrm{d}e = \pi_e(d - A_\mu x).$$

In the special case of Gaussian distribution with zero mean and iid components, we have $\mathbb{E}(e) = e_* = 0$ and $\Gamma_e = \sigma^2 I$. This gives us the likelihood distribution in which the additive errors have been marginalized

$$\pi(d|x, \mu) \propto \exp\left\{-\frac{1}{2\sigma^2}\|d - A_\mu x\|^2\right\}.$$

Unfortunately, the variable μ cannot usually be marginalized so that we would have an analytical form for $\pi(d|x)$. In such cases, we have to treat both x and μ as unknowns that have to be estimated simultaneously. But this will almost invariably necessitate sampling, for example, to compute $\mathbb{E}(x|d) = \int \pi(x, \mu|d)\, \mathrm{d}\mu$.

The key is to realize that the uncertainty of the primary unknown x is given by the posterior density

$$\pi(x|d) = \int \pi(x, \mu|d)\, \mathrm{d}\mu$$

and not by

$$\pi(x|d) \neq \pi(x|d, \mu_*)$$

generally with *any choice* for a fixed μ_*. In particular, the conditional mean

$$\mathbb{E}(x|d) \neq \mathbb{E}(x|d, \mu_*)$$

generally, and these two estimates may differ significantly.

5 Exploration of the posterior distribution

The posterior distribution for practical problems usually does not allow analytic evaluation of posterior statistics. However, the posterior distribution may be evaluated using Bayes' theorem (11) and posterior statistics evaluated using computational methods. In a Bayesian analysis, model space will consist of primary unknowns as well as parameterizations of other uncertainties, and may have many parameters. When model space is high dimensional, expectations need to be evaluated using Monte Carlo integration as in (25).

These computational methods rely on drawing samples from the posterior distribution, which is achieved by Markov chain Monte Carlo (MCMC) algorithms. A general introduction to MCMC can be found in [67]. These algorithms generate a sequence of states in model space $\{x^{(n)}, n = 0, 1, 2, \ldots,\}$ that converge in distribution to the desired distribution as $n \to \infty$. The resulting sequence of states forms a random walk through feasible solutions, with the proportion of time spent at any state being equal to the relative probability of the state.

The chain is constructed as a Markov chain, that is, $P(x^{(n+1)}|x^{(n)}, x^{(n-1)}, \ldots, x^{(0)}) = P(x^{(n+1)}|x^{(n)})$, so the transition from state $x^{(n)}$ to $x^{(n+1)}$ depends on $x^{(n)}$, but not explicitly on previous states.

Below, we only describe the discrete state Markov chain. A homogeneous (discrete state) Markov chain is defined by the transition matrix

$$K_{ij} = P(x^{(n+1)} = j | x^{(n)} = i) \tag{26}$$

giving the conditional probability to enter state number j on the next step, given that the current state is numbered i. The chain is initialized by drawing $x^{(0)}$ according to some distribution $\pi^{(0)}$ and then iterating the random update. Let $\pi^{(n)}$ be the n-step distribution, which is the row vector with j^{th} component $\pi_j^{(n)} = P(x^{(n)} = j)$. Then $\pi^{(n)} = \pi^{(0)} K^n$. When $\pi^{(n)} \to \pi$ for some fixed distribution π, independent of $\pi^{(0)}$, the Markov chain is said to be *ergodic* [61]. Then we are able to replace integrals over π by averages over the chain as required in (25), with convergence guaranteed by appropriate central limit theorems [68, 69, 70].

Ergodicity is guaranteed when K has a single eigenvalue of 1 in which case π is the corresponding left eigenvector[8]. This determines the *equilibrium* distribution π from the transition matrix K.

MCMC algorithms require the converse, that is, determining a transition matrix K that has the desired equilibrium distribution π, in our case the posterior distribution for the inverse problem being considered. Even for univariate models, model space can be huge and so it is not feasible to actually assemble the transition matrix. Instead, MCMC algorithms *simulate* operation by a suitable transition matrix at each step.

5.1 Metropolis-Hastings algorithm

Almost all implementations of MCMC sampling employ the Metropolis-Hastings (MH) algorithm, or some variant of it. This algorithm was originally developed for applications in statistical physics [71], and was later generalized to allow general proposal distributions [72], and then allowing transitions in state space with differing dimension [73], allowing insertion and deletion of parameters [74, 75]. Even though we do not always use variable-dimension models, we prefer this ‘reversible jump’ formulation of MH as it greatly simplifies calculation of acceptance probabilities for the subspace moves that are frequently employed in inverse problems.

The MH algorithm generates a Markov chain with the desired equilibrium distribution by simulating a suitable transition kernel at each step. The *reversible jump* MH algorithm [73] can be described as follows. When at state x, we generate, say, r random numbers γ from a known density $q(\gamma)$ and then form a proposed new state x' as some suitable deterministic function of x and γ. This gives a random proposal depending on the current state. The proposal is accepted or rejected according to a rule that ensures the desired equilibrium distribution.

[8] All eigenvalues of K have magnitude less than or equal to 1.

The reversible jump formalism considers the composite parameter (x, γ), and (x', γ') which is the composite parameter for the reverse proposal. One step of the MCMC sampling algorithm with MH dynamics can be written as:
Let the chain be in state $x_n = x$, then x_{n+1} is determined in the following way:

1. *Propose a new candidate state x' from x depending on random numbers γ with density $q(\gamma)$*
2. *Calculate the MH acceptance ratio*
$$\alpha(x, x') = \min\left(1, \frac{\pi(x'|d)q(\gamma')}{\pi(x|d)q(\gamma)}\left|\frac{\partial(x', \gamma')}{\partial(x, \gamma)}\right|\right) \tag{27}$$
3. *Set $x_{n+1} = x'$ with probability $\alpha(x, x')$ (accept the proposed state), otherwise set $x_{n+1} = x$ (reject).*

The last factor in equation (27) denotes the magnitude of the Jacobian determinant of the transformation from (x, γ) to (x', γ').

The only choice one has in the MH algorithm, is *how* to propose a new state x' when at state x. The popular choice of Gibbs sampling is the special case where x' is drawn from a (block) conditional distribution, giving $\alpha(x, x') = 1$. The choice of the proposal density is largely arbitrary, with convergence guaranteed when the resulting chain is irreducible and aperiodic [61, 76]. However, the choice of proposal distribution critically affects efficiency of the resulting sampler. We discuss efficiency further in Section 10.5.

The most common MH variants employ *random walk* proposals that set $x' = x+\gamma$ where γ is a random variable with density $q(\cdot)$, usually centered about zero.

To demonstrate the algorithm we give an example of sampling from the univariate posterior distribution in Section 3.1 when there are $M = 100$ measurements. The posterior probability density over heat capacity c for one measurement set is shown in Fig. 2.

We use the simple random-walk proposal $c' = c+w\gamma$ where $\gamma \sim \text{Uniform}(-1/2, 1/2)$ so that w sets the width of a uniform *window* about the current state c.

The reverse proposal is $c = c' + w\gamma'$ giving $\gamma' = -\gamma$. Hence the Jacobian is

$$\frac{\partial\,(c', \gamma')}{\partial\,(c, \gamma)} = \begin{pmatrix} \frac{\partial c'}{\partial c} & \frac{\partial \gamma'}{\partial c} \\ \frac{\partial c'}{\partial \gamma} & \frac{\partial \gamma'}{\partial \gamma} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ w & -1 \end{pmatrix}$$

which has determinant with unit magnitude. Since $q(\gamma) = q(\gamma') = 1$, the acceptance ratio in 27 simplifies to

$$\alpha(x, x') = \min\left(1, \frac{\pi(c'|d)}{\pi(c|d)}\right)$$

which depends only on the ratio of the posterior densities at the current and proposed states.

Fig. 3 shows three traces of a sequence of 2000 states generated by the MH algorithm for three choices of proposal window, $w = 0.1$, $w = 1$, and $w = 10$.

Each of the resulting chains is guaranteed to converge to the posterior distribution, but, as can be seen the nature of the three traces is quite different. The choice $w = 1$ gives a proposal window that is similar in scale to the width of the posterior distribution, has about 50% of proposals accepted, and the chain efficiently produces samples from the posterior distribution. When $w = 0.1$, proposals are only a small change on the current state, and are accepted about 90% of the time. However, because the proposal window is small, adjacent states are strongly correlated and the chain moves slowly through state space, taking a long time to generate independent samples. When the proposal window is large, $w = 10$, a very small proportion of proposals are accepted which shows up by the chain remaining constant for many steps, and again the chain is slow to produce independent samples. We look at the efficiency of MCMC in more detail in Section 10.5.

For this univariate problem, expectations over the posterior are easily calculated using numerical quadrature and MCMC sampling is simple though not particularly efficient. However, for models with more than about 5 components, numerical quadrature is computationally infeasible while Monte Carlo integration remains feasible.

6 Likelihood models

The likelihood distribution $\pi(d|x)$ introduced in Section 4.1 is a probabilistic model for the distribution over measurements d given that the unknowns have value x. Design of the likelihood, in the case of inverse problems, often involves modelling the forward problem along with a separate model for measurement errors. In that case, the forward map is the model for errorless observations. This is a feasible interpretation in most cases, such as in the case of additive noise models. However, in many important measurement processes, for example with counting observations of radioactive decay, there is no such thing as an errorless observation.

While the construction of the likelihood can generally be considered as a straightforward problem, this task is not always trivial, especially when real data is to be used and the actual uncertainties that are related to the measurement setting are to be considered. Improvement of likelihood models is often best achieved by examining residuals to ensure they conform to the assumed error distribution, particularly when tracking down physical processes that are not part of the intended measurement procedure. A practical limit to likelihood modelling is often set by complexity or computational limitations. We discuss some frameworks for dealing with the resulting approximate likelihoods in Sections 8.5 and 10.6.

Partial differential equations and initial-boundary value problems arguably form the most common class of forward problems that are relevant in the case of inverse problems. As discussed in Section 2.5, implementation of the model usually requires numerical solution. Before one embarks on the computational implementation, the physical setting should be thoroughly investigated and the uncertainties identified and modelled. A suitable computational framework should then be chosen that can accommodate the models and parameterizations for all unknowns.

It is often the case, for example, that measurement locations are not exactly known, or the geometry of the medium required to specify the outer boundary of the computational domain is not known exactly, and it may be necessary to include these uncertainties in the likelihood model. Computational limitations often require truncating computational domains, for example when a large body is investigated and the measurements are carried out locally, in which case the boundary conditions holding at this truncation boundary are unknown. These topics are discussed later in Section 8.

6.1 Forward models

Forward models define (conditionally) the deterministic part of the likelihood distribution. As in Section 4.5, let the primary and secondary unknowns be x and μ, respectively, and e denote the (classical) noise variable. In the case of additive noise, we have

$$d = A_\mu(x) + e$$

where $A_\mu(x)$ is the mapping $(x, \mu) \mapsto d$ which predicts the *noiseless* measurement when both x and μ are known. We also presume that the distribution $\pi(e)$ is known, or if applicable, the joint distribution $\pi(x, \mu, e)$ is known. Thus, if both (x, μ) were known, we could compute the errorless observations and evaluate the likelihood. As it turns out, when MCMC methods are to be used, this is basically the only requirement.

In the case of non-additive noise models, we need the more general data simulation $d = A_\mu(x, e)$ when a *realization* of the random variables (x, μ, e) is available. In the most general form, we have to be able to evaluate the likelihood distribution, see Section 6.3.

In inverse heat transfer problems, the mapping A_μ is commonly related to partial differential equations and the related initial-boundary value problems, see Section 2. Then the numerical schemes for the IBVP discussed in Section 2.5 provide a simple and tractable form for the model $A_\mu(x)$. For example, let μ be the unknown Dirichlet data on part of the boundary of the computational domain, and let us use FEM for the numerical approximation of an elliptic problem. Then, if we write μ in the same basis as the FEM basis on the boundary, we can write

$$A_\mu(x) = Bx + C\mu$$

where B and C are matrices provided by the FEM formulation, see Section 9 for an example. Modelling of the unknown μ could be achieved by using a mid- or high level representation that is then mapped onto the FEM basis, see Section 7.

The computational implementation of the likelihood and forward models are generally approximative, due to the following issues:

- The mapping that is used in the computations is always a computational and numerical approximation for the PDE and IBVP.
- The PDE itself may be an approximation, for example, that neglects radiation effects.
- Initial conditions may be approximate.
- All boundary data and/or parameters of the boundary models are seldom known even in controlled laboratory experiments.
- The measurement sensor locations might not be exactly known and an idealized pointwise measurement model might not be an adequate model for the actual sensors.

Possibilities for how to handle these model errors, will be discussed in Section 8.

The construction of the likelihood density can thus roughly be divided into two tasks: the construction of the deterministic forward model that assumes a value for all unknowns, and the extension to a full statistical model that gives the likelihood distribution over measurements.

6.2 Additive noise models

The construction of the likelihood model in the case of additive noise is the easiest to implement. The additive noise model was briefly visited in Section 4.3 in the case of Gaussian errors that are mutually independent with other unknowns.

The additive noise model is of the form

$$d = A_\mu(x) + e$$

with a marginal distribution model for $e \sim \pi_e(e)$. Generally, however, the variables (x, μ, e) are not mutually independent. When (x, μ, e) are given, we have formally $\pi(d|x, \mu, e) = \delta(d - A_\mu(x) - e)$ which gives us

$$\pi(d|x, \mu) = \pi_{e\,|\,x,\mu}(d - A_\mu(x)\,|\,x, \mu)$$

where we write $\pi(x, \mu, e) = \pi(e|x, \mu)\pi(x, \mu)$, see, for example, [17]. The special case in which e and (x, μ) are mutually independent, gives the simple form

$$\pi(d|x, \mu) = \pi_e(d - A_\mu(x)),$$

and hence the likelihood modelling task consists only of the separate tasks of modelling $\pi_e(e)$ and $A_\mu(x)$. If e and (x, μ) are not mutually independent, a feasible starting point is usually to consider modelling $\pi(e|x, \mu)$ first.

The most commonly used model for the distribution over e is the Gaussian distribution. Moreover, almost always the special choice $e \sim \mathcal{N}(0, \sigma^2 I)$ is employed, that is, the zero mean iid error model[9]. While there are situations in which this is a feasible model, in the majority of problems there are more realistic models. Note that the Gaussian distribution decays very fast and thus if the mean and covariance are badly modelled, the actual x might correspond to an essentially vanishing likelihood and thus also to a vanishing posterior. In other words, according to our posterior model, the true x would be essentially impossible.

Even when the Gaussian additive noise model is sustainable, the assumptions of zero mean, diagonal covariance structure and identical variances (diagonal elements of Γ_e) can usually be challenged in real physical situations. For example, errors due to model reduction and approximate physical models are very unlikely to have zero mean. Also the Gaussianity and the iid covariance structure are often not feasible assumptions.

As an extreme example, consider the following industrial measurement setting. There are 10 temperature sensors attached to the target, the measurements are carried out simultaneously and there are high power electric motors nearby. For example, in a typical thermomechanical pulp plant there can be twenty 2 MW motors within 20 m radius. The electric and magnetic fields are very strong with most energy around 25 Hz, and induce large currents into the measurement leads. The resulting noise is typically several orders of magnitude larger than any measurement instrument noise. While the instrument noise could be modelled as iid errors, the measurement errors due to the external fields are very highly correlated. In this case, let σ_1^2 and σ_2^2 be the variances of a single measurement that are due to the instrument noise and the external fields, respectively. Then, a model of the form

$$e \sim \mathcal{N}(0, \sigma_1^2 I + \sigma_2^2 \mathbf{1}\mathbf{1}^{\mathrm{T}})$$

where $\mathbf{1} = (1, \ldots, 1)^{\mathrm{T}}$ and $\mathbf{1}\mathbf{1}^{\mathrm{T}}$ is thus a matrix of all ones, could be more sustainable. The use of an independent (uncorrelated) error model in such a situation is bound to lead to significantly misleading estimates, and, in particular, significantly overoptimistic error estimates.

[9] In the Gaussian case, uncorrelatedness implies mutual independence.

Fortunately, the measurement errors and their distribution can almost always be determined experimentally. Also, the designers of measurement systems should be able to provide operational specifications of the system, in particular, the joint statistics of internal errors.

There are several common sources of non-Gaussian additive errors. In many measurement situations there are infrequent large errors that are often referred to as anomalies or outliers. In one-off investigations the most obvious outliers could be removed after visual examination of the data, but this is seldom a feasible or optimal way of treating the data. Heavy tailed distribution models such as the Cauchy distribution and the L_1-norm induced distribution $e \sim \exp(-c\|e\|_1)$ can mitigate the effects of outliers [17], while a mixture model in a hierarchical Bayesian formulation can be used to explicitly model and detect outliers [59].

6.3 Other noise and likelihood models

We consider two other common likelihood types, the multiplicative noise case and the Poisson distributed observations [17]. For uniformly distributed additive noise, see also Section 3.

Multiplicative noise. It is quite common that errors have a multiplicative component. Modulation type observations are a typical example. As a very simple example, consider an observation model that is linear with respect to the primary unknown, with a noisy amplifier in the measurement chain, so the likelihood model has the form

$$d = A(h_* + \nu)x + e$$

where the nominal amplification, say, $h_* = 1$ is corrupted by noise $\nu \sim \pi(\nu)$. Such a noise model is called (partially) *multiplicative* due to the term $A\nu x$.

Counting distributions. Another example of a non-additive noise likelihood model is that of Poisson distributed measurements. Such a measurement model is typical for situation in which the measurements are counts, the most common example being radioactive decay of low activity samples. Although such measurements are not that common with inverse heat transfer problems, Poisson distributed observations form a simple example in understanding likelihood models beyond standard additive errors models.

In the simplest case, we can write for the measurements d_k

$$d_k \sim \text{Poisson}\left(A_k(x)\right)$$

where $d_k \in \mathbb{N}$. If the random variables d_k are mutually independent, we have

$$\begin{aligned}\pi(d|x) &= \prod_{k=1}^{M} \frac{A_k(x)^{d^k}}{d_k!}\exp(-A_k(x))\\ &\propto \exp\left(d^{\mathrm{T}}\log(A(x)) - \|A(x)\|_1\right).\end{aligned}$$

There is no additive error form for $\pi(d|x)$. Most importantly, the notion of "error" or "noise" is irrelevant, since the measurement d can only be considered as a draw from the conditional distribution $\pi(d|x)$.

In most practical cases, relevant error models are more complex than these simple models. In addition to the combination of additive and multiplicative errors, combination of Poisson distributed variables and additive errors may be a relevant model [77, 78]. In Section 8.5 we consider likelihood models that arise when approximate marginalization is used to remove auxiliary unknowns, and when model reduction is employed.

6.4 Comments

One of the most important principles in likelihood modelling for inverse problems, is to *not underestimate* any of the errors and uncertainties. While overestimating errors can be interpreted as squandering the accuracy of the measurement information and thus reducing the accuracy of estimates, this is not a capital crime since the true unknowns would still be supported by the posterior model. In contrast, underestimating the errors will typically provide a posterior model with respect to which the actual unknown is impossible, and thus cannot be recovered.

7 Prior modelling

One of the practical advantages of the Bayesian formulation for inverse problems is that it provides the framework whereby uncertainties can be included in a way that is informed by stochastic modelling. Prior modelling is perhaps the most distinctive of these, since separate modelling of the measurement process and the unknowns is a distinguishing feature of Bayesian methods.

While the construction of the likelihood models can be described as a straightforward albeit sometimes tedious task, prior modelling typically is considered at least partially as an "art". This is due to the fact that there are always several plausible models for the unknowns. These models may, however, differ significantly as to how easily they can implemented, and especially how efficient the related computational machinery can be implemented.

7.1 General considerations

Prior modelling has the same flavor as mathematical modelling in general. That is, one describes the key physical features of unknowns, checks plausibility of the model and robustness of solutions to particular assumptions, and refines the model as needed. In particular, there is no single 'correct' prior model, since the purpose of a model depends on the purpose of performing the inverse problem. A prior model will, in general, be informative with respect to some questions but non-informative with respect to others, hence different prior models are required to give quantitatively accurate answers to different questions, see for example [79]. Competing models may be tested using Bayesian *model comparison* [80, 61], as discussed in Section 11.5.

The central components of prior models are the *representation* of primary unknowns, that is, the parameters or coordinates used, and a normalizable *prior density* over allowable values. The choice of representation is largely determined by the modelling assumptions one wants to enforce and the properties that are being sought. Stronger modelling assumptions leading to reduced models can reduce ill-posedness of the inverse problem, however extreme models that excessively restrict model space can lead to large approximation errors and vastly over-confident estimates of accuracy [17, 81]. Since a Bayesian analysis primarily quantifies uncertainty in models, there is no need to restrict models to give a unique solution.

As discussed in Section 2, primary unknowns in inverse heat transfer problems can be the thermophysical properties, heat sources, boundary conditions, or some combination of these. Since these unknowns are spatially-varying functions, models developed for *spatial statistics* [82, 83] are directly applicable.

In that field, representations and priors are classified as low-level, mid-level, and high-level [84, 85]. Low-level representations are local and generic, and usually very high-dimensional, such as gray-scale pixel images, or the vector of element coefficients in a FEM discretization. These representations are typical in regularized inversion and can be used for any image, but are inconvenient for stating or calculating anything other than local structural information. Mid-level models are also generic, but provide convenient ways of expressing quantities of interest such as geometric features of objects, or between objects. An example is the representation of boundaries using implicit functions or patches [85, 86, 87]. High-level models capture important, possibly complex, features of the images and are useful for answering global questions, such as counting the number of heat sources [88].

The type of model that is applicable to a given problem depends on what is known about the primary unknowns, and the purpose of the analysis. We now look at modelling issues in the context of heat transfer.

7.2 Modelling thermophysical properties

If the thermophysical properties are unknown a priori, the estimation problem requires determining all of the thermal conductivity κ, the heat capacity c and the relaxation time τ_{r}. In the following, we discuss models for the conductivity κ, though the models are equally applicable to all other thermophysical properties.

Modelling of the coefficient κ depends on modelling of the medium. When the medium is isotropic, κ is a scalar since heat flows in the direction of the gradient of temperature. For laminated materials that are orthotropic, κ can be written as a diagonal matrix in suitable orthonormal coordinates, while for general composite materials k is a tensor of a more general structure. Combinations of these properties are possible, for example in a medium that is homogeneous and orthotropic, κ can be modelled as a constant diagonal matrix [4, 35].

When the material is homogeneous but the constant value of k is unknown, a univariate prior distribution is required over the unknown value of κ. Positivity of conductance can be asserted by requiring the prior density to be zero for $\kappa \leq 0$. As in the first example of Section 3, a reference prior [59] based on transformation groups [89] ensures that inference does not depend on choice of units, see also [57]. These prior distributions are typically *improper*, i.e. do not have finite integral, and hence can only be used in an MCMC when there is sufficient data to ensure that the likelihood effectively constrains κ to a finite set. Prior knowledge of a range for κ, such as the possible values for a given material, may be asserted by using a prior distribution with finite support.

More generally, κ will depend on position and we write $\kappa(\vec{r})$. In layered three-dimensional materials, or rod-like geometries, κ can be represented by a one-dimensional function, while more generally k may be an unknown function of two or three dimensions. In exploratory analyses, or when little is known about the medium, it is typical to use a low-level representation such as a pixel image, and a *non-parametric* prior distribution such as a Markov prior model with the Gibbs form

$$\pi(\kappa) \propto \exp\left\{-\sum_{C \in \text{cliques}} \Psi_C(\kappa)\right\} \tag{28}$$

where Ψ is a potential function and the sum is over *cliques*, i.e. sets of pixels that are mutually neighbors. When pixels have no neighbors, so the cliques are individual pixels, and $\Psi_i(\kappa) = \beta|\kappa_i|^2$, the prior density is the Gaussian density of Section 4.1 with $\Gamma_\kappa = \beta^{-1}I$. More typically smoothness of the unknown function k is asserted by considering nearest-neighbor interaction giving

$$\pi(\kappa) \propto \exp\left\{-\sum_{i=1}^{M}\sum_{j\sim i} \Psi(\kappa_i, \kappa_j)\right\} \tag{29}$$

where the first sum is over the M pixels and the second is over pixels j that neighbor pixel i, denoted $j \sim i$. An example is the Gaussian Markov random field (GMRF) when $\Psi(\kappa_i, \kappa_j) = \beta_{ij}\left(\kappa_i - \kappa_j\right)^2$, though other potentials and neighborhood structures are possible. Small neighborhood structures lead to sparse precision matrices (inverse of the covariance matrix) which is useful for computational purposes [90]. An alternative is to specify the covariance matrix directly, often by writing $\pi(\kappa)$ as a Gaussian process specified by the covariance matrix $\Gamma_{ij} = \phi\left(\|\vec{r}_i - \vec{r}_j\|\right)$ where $\phi(\cdot)$

is a suitable covariance function [83]. Commonly used is the exponential covariance function

$$\phi(r) = c_0 \exp\{-r/r_0\} \tag{30}$$

where c_0 sets the variance of each location, and r_0 sets a length scale, or 'correlation length'. This function is often generalized to allow different length scales in different directions, when, for example, modelling time and space dependence, or when length scales differ along differing coordinates as occurs in orthotropic materials.

The parameters c_0 and r_0 appearing in (30) are examples of *hyperparameters* that appear in the distributions over primary unknowns. In some cases these hyperparameters may be given fixed values, but more typically are modelled as uncertain and ascribed distributions, often called *hyperpriors*. This multi-level, or 'hierarchical', approach to prior modelling is handled easily in the Bayesian framework, and is discussed extensively in [36] in the context of inverse heat conduction problems. Further applications of hyperpriors are discussed in Section 8.4.

The inhomogeneous nature of a material may be modelled more explicitly, such as when the medium is known to be a composite of several homogeneous materials. In that case κ can be represented as being piecewise constant [46], or more generally as piecewise smooth. When the medium is known to consist of two types of material, each with known κ but with unknown distribution, the prior distribution may be specified as a discrete Gibbs distribution over type defined on the pixel lattice. When there are two material types, a suitable potential is

$$\Psi\left(\kappa_i, \kappa_j\right) = \begin{cases} \beta & \kappa_i = \kappa_j \\ 0 & \kappa_i \neq \kappa_j \end{cases}$$

in (29), giving the familiar Ising model [91]. The Potts model generalizes this setting to more than two material types [85]. If the conductivity κ of each type is also uncertain than a two-level prior distribution can be used such as a segmented Gaussian field [60] in which an Ising or Potts distribution models the material type, with smoothness asserted within regions of a given type, but not across type boundaries. In regularization methods, the total-variation (TV) semi-norm is often used as an attempt to produce solutions of this type [92]. The TV seminorm also lends itself to an improper exponential prior distribution, however, see the related comment in Section 7.5.

Type-field models are examples of mid-level priors since a classification of material type is included in the representation of unknowns. The Ising and Potts models give a type field defined over a pixel lattice, which can be high dimensional and lead to slow computing. An alternative is to use continuum processes, such as the change-point models in 1-dimension [73], and colored continuum triangulations in two dimensions [93]. For example, a change-point model would be useful when modelling a layered material of thickness R in which layers occur at depths

$0 = r_0 < r_1 < r_2 < \ldots, < r_n = R$ with conductivity κ_i in the interval (r_{i-1}, r_i), $i = 1, 2, \ldots, n$ in which the $\{r_i\}$, $\{\kappa_i\}$ and n are unknown. This is a variable dimension model since the *number* and *location* of material boundaries is unknown. Performing inference on this model is straightforward using reversible jump MCMC described in Section 5.1, see also [73]. In contrast, this would be a near impossible task using standard regularization methods, see for example [46]. A further possibility is to define auxiliary unknowns, in this case a (spatially smooth) random process that defines the (spatially varying) smoothness of the primary unknown, see [94].

7.3 Modelling unknown boundary conditions

In a traditional deterministic approach it would be necessary to either estimate or assume the effective boundary conditions that hold at computational boundaries. Straightforward simultaneous estimation of material parameters and the boundary conditions may lead to an identifiability problem as will be explained in Section 8.2. In the general case, a joint distribution model for the primary unknowns and the secondary unknowns is needed – in which these unknowns are not independent.

However, the Bayesian approach allows truncation boundaries to be included in the prior model for thermophysical properties. The key realization here is that the boundary conditions on the truncation boundaries depend on the distribution of the material parameters *outside* the computational domain. If we can model the material parameters, for example, as GMRF's, this model links (statistically) the material parameters inside the computational domain and the boundary conditions implicitly. This leads to a joint distribution model for the primary unknowns and the boundary data, here playing the role of the secondary unknowns. See [95] for an example for how the approximation error approach can be used to handle the truncation boundaries.

7.4 Modelling heat sources

In many cases there is strong prior knowledge about aspects of the source such as it strength, or location in time and/or space. For example, a heat source that is known to be localized at the point $\vec{r}_0$ may be written as $q(\vec{r}, t) = q(t)\,\delta(\vec{r} - \vec{r}_0)$ and the estimation problem reduces to finding the unknown function of time $q(t)$. Alternatively, a heat source may be known to be essentially instantaneous in which case we may write $q(\vec{r}, t) = q(\vec{r})\,\delta(t - t_0)$ and the estimation problem is to find the spatial dependence of the heat source $q(\vec{r})$ and also the time at which it acts, t_0, if that is not known a priori.

Inverse problems particular to forced convection include estimation of unknown temperature of the flow at an inlet. If the inlet temperature is stationary in time, then the inverse problem reduces to finding the spatial variation of temperature over the inlet, only. Alternatively it may be known that the inlet temperature is constant in space, but varies in time. In that case the inverse problem has the unknown inlet temperature as a function of time as the primary unknown.

7.5 Comments

There is an important recent observation regarding the discretization and representation of the unknown, especially in relation to non-Gaussian prior models. The total variation prior is commonly used to represent "blocky" objects, that is, unknowns that are assumed to have the tendency to be spatially piecewise constant. The total variation is qualitatively the 1-norm of first order smoothness, and the corresponding prior model is an exponential distribution with this norm as the potential. In [96], it was shown that when the discretization is made spatially denser, the non-Gaussian TV distribution actually tends to a Gaussian distribution, thereby losing the desired prior characteristics.

The Gaussian smoothness type priors are often considered trivial and restrictive. This is definitely true with priors of type

$$\pi(x) \propto \left(-\alpha \|D_p x\|^2\right)$$

where L_p is a standard discretization of a p^{th} order differential operator, with a null space of dimension p. But the smoothness can be made inhomogeneous and, in particular, anisotropic [97, 17]. In addition, the smoothness prior can be made proper by a conditioning process. The draws from the resulting prior models are often difficult to perceive to be from a Gaussian distribution. Such inhomogeneity and anisotropicity information is often available from complementary information, such as other measurement modalities. See [98] for an example of in-painting, that is, how to fill in missing parts of an image.

8 Model uncertainty

With model errors and uncertainties, we refer to all unknowns (and their parameterizations) other than the primary unknown. It is to be noted, of course, that with different formulations of the problems, some unknowns, such as the initial temperature distribution, might be either primary unknowns or uncertainties.

8.1 Rough categorization of errors and uncertainties

We classify the model errors and uncertainties roughly into the following categories:

1. *Model reduction errors.* These errors are due to using a reduced order model, usually for computational efficiency. In other words, given more time and greater computational arsenal, these errors could be avoided in principle .
2. *Uncertainties related to the physical models.* These errors include missing boundary and initial data, and approximative (simplified) physical models, for example, neglecting radiation effects. Also, uncertainty of geometry belongs to this category.
3. *Uncertainties related to the behavior of the measurement system.* These include unknown measurement noise variance, covariance structure, or statistics in general, as well as measurement system specific issues such as cross-talk in multichannel measurement systems.
4. *Prior and other uncertainties.* Generally, uncertainties in the prior model: (covariance) structure and the distribution itself.

Some of the uncertainties could be avoided in a straightforward manner, such as the missing boundary data on the truncation boundaries of the computational domain: simply enlarge the computational domain so that the uncertain boundary conditions have no effect to prediction of the measurements. Similarly, with pure discretization errors of the forward solver, we could simply use more accurate approximations. Many others cannot be avoided in a straightforward way, and feasible prior models have to be constructed for these uncertainties. For example, parametrization of uncertain geometry in 3-dimensional situations, and simultaneous estimation with primary unknowns can turn out to be a prohibitively complex undertaking.

Thus, having parameterized the overall problem, we have two extreme possibilities: straightforward but often prohibitively computationally complex "estimate all unknowns (including uncertainties) simultaneously" or to "approximate all auxiliary unknowns with ad hoc guesses", and anything in between.

8.2 Identifiability of problems

It must be noted that if all uncertainties are parameterized, the overall structure of the problem might not be *identifiable.* If we are honest with the likelihood and prior modelling, it may turn out that the posterior model is *improper*, that is, it is technically not integrable[10]. Thus a MAP estimate would not be a point estimate but rather a manifold (or a linear or affine subspace), and could not be computed. Also, computation of MCMC is typically not possible.

[10] The problem of improper posterior distributions tends not to happen when reference priors are used [99].

As a simple example, consider the linear additive Gaussian noise case with Gaussian prior model with $\pi(e) = \mathcal{N}(0, \Gamma_e)$, $\pi(x) = \mathcal{N}(x_*, \Gamma_x)$. The posterior distribution can be written in the form

$$\pi(x\,|\,d) \propto \exp\left\{ -\frac{1}{2} \left\| \tilde{A}x - \tilde{d} \right\|^2 \right\} \tag{31}$$

where $\Gamma_e^{-1} = L_e{}^{\mathrm{T}} L_e$ and $\Gamma_x^{-1} = L_x{}^{\mathrm{T}} L_x$ and

$$\tilde{A} = \begin{pmatrix} L_e A \\ L_x \end{pmatrix} \quad \text{and} \quad \tilde{d} = \begin{pmatrix} L_e d \\ L_x x_* \end{pmatrix}.$$

If the models for A, Γ_e and Γ_x are such that $\tilde{A}$ has a nontrivial null space,[11] the conditional mean does not exist and the MAP estimates are not points but linear manifolds with the same dimension as the null space of $\tilde{A}$, see for example [17].

Of course, in such a case, regularization methods could be applied to obtain a *numerical* solution for MAP or CM estimates. But this approach would inherit the problems that are associated with regularization methods, most importantly: the estimates would not have any statistical interpretation

The correct conclusion here is that the *information contained in the models and measurements* is not adequate to allow for estimation of the unknowns. Thus, more information has to be acquired, either in terms of measurements or more informative models. This might not necessarily mean that *more* measurements have to made. Changing the way measurements are carried out changes the model A and has an effect (in the above case) on $\tilde{A}$ and its properties, and thus the characteristics of the posterior distribution. Optimization of measurements can thus be carried out in relation to the prior model. For an example in electrical impedance tomography see [100].

8.3 Strategy

The feasible choice for the strategy depends on realizability and other implementation issues, specifications for accuracy, and small versus large dimensional parameterizations for uncertainties.

Below, we discuss two classes of approaches with which to handle uncertainties. The use of *hyperprior* models is especially useful for uncertainties that have a low-dimensional parametrization. The other class is the *approximation error approach*, in which stochastic simulation over the uncertainties is performed and approximate marginalization is then carried out.

[11] We would have nonzero vectors z so that $\tilde{A}z = 0$.

8.4 Hyperpriors

Prior and likelihood modelling is typically a hierarchical process, with a density over primary unknowns depending on parameters which themselves may be uncertain, and hence are modelled as random variables with some distribution. These secondary parameters are called *hyperparameters* while the associated distributions are called *hyperpriors*. The term “hyperprior” is slightly misleading since the approach is equally well suited to modelling parameters in any distribution, not only the prior.

Modelling uncertainty in hyperparameters appearing in the prior distribution was discussed in Section 7.2. Extensive discussions can be found in [36, 37]. Unknown parameters in the likelihood can also modelled using hyperpriors. A natural example of such a hyperparameter is the variance, or noise level, σ^2 appearing in a Gaussian model for measurement error, as in Sections 4.3 and 6.2, that can be modelled using a hyperprior to good advantage. As noted in Section 4.3, it is a common mistake to view this parameter as fixed and hence equate MAP estimation to Tikhonov regularized inversion. However, it is an important difference that this variable has the distribution of the hyperprior and is not a fixed value. In an “empirical Bayes” analysis, a ‘best’ estimate of this variable is made, often by maximum likelihood, and subsequent analysis conditioned on this value [59, 63], hence mimicking a regularization approach. However, a notable feature is that the value is determined [101] *based on data*. The empirical Bayes approach is recognized as an approximation, perhaps necessitated by computational considerations, and in many cases gives much worse results compared to the correct action of marginalizing over this nuisance parameter.

An informative discussion of the importance of modelling auxiliary variables with uncertainty, in the context of heat transfer, is in [37]. See also [17] for an example of simultaneous deconvolution and estimation of the width of the convolution kernel.

8.5 Second order additive error model for approximation and modelling errors

The approximation error approach was introduced in [17, 81] originally to cope with model reduction related errors in the likelihood, where model reduction can be carried out both for the computational forward problem and the representation of the unknown. In this context, the approximation error approach can be described as follows. Let the accurate physical model be

$$d = \bar{A}_\mu(\bar{x}) + e$$

where $\bar{x}$ is typically an infinite dimensional distributed parameter, such as thermal conductivity, and μ represents other uncertainties, which in this example are related to the forward map only.

Let us fix the approximate computational model A_{μ_*} where we have fixed the unknown $\mu = \mu_*$, and a finite dimensional representation x, typically $x = P\bar{x} = \sum_j \tilde{x}_j \varphi_j$, where P is a projection onto the subset $\{\varphi_j\}$ and $\tilde{x}_j$ are the projection coefficients. In the following, we identify x and the set of coefficients $\{\tilde{x}_j\}$. We can then write

$$\begin{aligned} d &= A_{\mu_*}x + (\bar{A}_\mu \bar{x} - A_{\mu_*}x) + e \\ &= A_{\mu_*}x + \varepsilon(\bar{x}, \mu) + e \end{aligned}$$

where the random variable $\varepsilon(\bar{x}, \mu)$ is called the *approximation error.*

In the approximation error approach, the aim is to carry out approximate marginalization over $\varepsilon + e$ *prior to inference* on the interesting variable x.

This is carried out by approximating the joint distribution $\pi(x, \varepsilon, e)$ and the likelihood distribution with Gaussian models with possibly a nonlinear forward model. First, as in Section 6.2, note that

$$\begin{aligned} \pi(d|\bar{x}, e, \mu) &= \delta(d - A_\mu(\bar{x}) - e) \\ &= \delta(d - A_{\mu_*}(x) - e - \varepsilon(\mu, \bar{x})) \end{aligned}$$

where μ_* is fixed, for example, $\mu_* = \mathbb{E}(\mu)$, and $\varepsilon(\mu, \bar{x}) = A_\mu(P\bar{x}) - A_{\mu_*}(\bar{x})$. Hence, we write for the likelihood

$$\begin{aligned} \pi(d|x) &= \int \pi(d, e, \mu|x)\, de\, d\mu = \int \pi(d, e, \varepsilon|x)\, de\, d\varepsilon \\ &= \pi_{e+\varepsilon|x}(d - A_{\mu_*}(x)|x). \end{aligned}$$

The core of the (Gaussian) approximation error approach is, then, to approximate the distribution $\pi_{e+\varepsilon|x}$ with a Gaussian distribution $\pi_{e+\varepsilon|x} \approx \mathcal{N}(e_{*|x}+\varepsilon_{*|x}, \Gamma_{e+\varepsilon|x})$. For details, see [17, 81, 102] for formulation for model reduction only, and [103] for a more general formulation with uncertainties.

Naturally, ε and x are *not* independent, but in the *enhanced error model* we make the additional approximation $\pi(e, \varepsilon|x) \approx \pi(e, \varepsilon)$, which has been proven to be a feasible model with many applications, [17].

The approximation error approach has proven to be a feasible and computationally attractive alternative to simultaneous estimation of x and μ and to using computationally accurate forward (PDE) solvers. In addition to coping with model reduction (such as finite elements discretization) related errors, also errors due to using approximative physical models have turned out to be negotiable. Model reduction and unknown anisotropy structures in optical diffusion tomography were treated in [104, 105, 106]. Missing boundary data in the case of image processing and geophysical ERT/EIT were considered in [107] and [95], respectively. Furthermore, overcoming errors in domain geometry was treated in [18, 108]. Also, in [18, 108] the problem of recovery from simultaneous geometry errors *and* drastic model reduction was found to be possible. In [109], an approximative physical model (diffusion model instead of the radiative transfer model) was used for the forward problem. In [103], an unknown distributed parameter (scattering coefficient) was treated with the approximation error approach.

Furthermore, the error estimates given by the approximation error approach are feasible and often only slightly larger than when accurate computational models are employed [110].

8.6 Comments

The central issue here, again, is not to underestimate the errors and uncertainties. The effects of either underestimating or overestimating the errors are the same as with those with general likelihood modelling, see Section 6.4.

The approach to dealing with uncertainties is mostly related to availability of computational resources as well as the available computational time. In process industry both may be constrained and, for example, heavy model reduction may be needed.

9 Nonstationary problems

This far, we have mainly addressed problems in which the measurements may have been time-varying but the unknowns have been constants with respect to time. Inverse problems in which the unknowns are time-varying, are referred to as *nonstationary inverse problems* [17, 111]. In addition to describing the state estimation approach for the estimation of nonstationary quantities, the purposed here is to give an example of how to approach the modelling of uncertainties.

Several inverse problems are nonstationary in the sense that the unknown is naturally a time-varying entity. These problems are also naturally cast in the Bayesian framework. Nonstationary inverse problems are usually written as evolution-observation models in which the evolution of the unknown is typically modelled as a stochastic process. The related algorithms are sequential and in the most general form are of the Monte Carlo type [112]. However, the most commonly used algorithms are based on the Kalman recursions [113, 114, 17]. A review that covers most of the topics in this section, is given in [115].

We note that if we are studying a transient problem and carrying out temperature measurements at a number of locations, we may, of course, wish to compute the temperatures also at locations where measurements were not made. Then, the overall temperature evolution is also unknown and time-varying. Depending on the characteristics of the measurement setting and uncertainties, the estimation of the overall temperature evolution may or may not be an ill-posed problem. When the material parameters and the initial and boundary data were known, the estimation of the overall temperature evolution would reduce to a stable model-based interpolation problem. Kalman filtering was suggested as suitable for estimation of inverse heat transfer problems in [30].

Problems in which the primary unknowns are *explicitly* time-varying and can be modelled as stochastic processes, are called *state estimation problems*. Examples of explicitly time-varying variables include the temperature evolution itself, and the boundary heat flux, which may be poorly known and not completely specified by the temperature distribution. Naturally, unknown heat sources are usually to be considered as (explicitly) time-varying. On the other hand, the thermal diffusivity κ, for example, usually depends on time through the temperature evolution $T(t)$ *only*.

What is the difference then to standard least squares fitting of, say, a time-varying boundary flux to transient temperature measurement data? The answer is the same as with time-invariant Bayesian problems, which is that the state estimation formalism takes systematically into account all uncertainties and errors, and also yields systematic error estimates for the unknowns. In many seemingly simple problems, the propagation of errors through the model can turn out to be highly non-trivial.

The state estimation formalism is also relevant in the estimation of time-invariant parameters, when it is called *state space identification*, see for example [116, 117]. Also here, the motivation to use the state space formalism instead of least squares estimation is the same as above.

We note that there is a major difference between state estimation approach, and in fitting type algorithms in which the heat equation is (at least implicitly) taken as accurate and is used as a *constraint*. The constraint type approach should be avoided unless the accuracy of the deterministic model is confirmed. The authors can think of very few inverse problems in which the constraint approach would be justified.

9.1 Stochastic heat equation

To set the scene, we consider the following state space identification problem. The temperature evolution is governed by the basic heat equation in IVBP (1), with the insulating boundary condition $q_{\mathrm{N}} \equiv 0$, and where the Dirichlet condition is $T_{\mathrm{D}} = f(t)$ where f is a time-varying boundary temperature representing heating by, for example, a gas torch.

Let us assume that the insulating boundary condition is not exactly fulfilled, and that the model for the boundary temperature f is based on a single measurement only. Furthermore, assume that the overall problem is to estimate the specific heat capacity c_p and the thermal diffusivity coefficient κ when the density ρ is *treated as fixed*, and transient temperature data is collected at specified locations on the insulating boundary Ω_{N} while the boundary control $f(t)$ is time-varying.

We have at least the following sources of error and uncertainty:

- The insulating boundary condition is not exactly fulfilled but since the ambient temperature is lower than in Ω, we know, from an improved condition such as (5), that the heat flux is negative.
- The temperature on Ω_{D} is not constant and thus cannot be completely specified by the single temperature measurement.
- We only know the mean density of the object and use a spatially homogeneous model for the density. In reality, we know that the density is inhomogeneous and that it could be modelled as a Markov random field with some correlation length.
- We use a basic projection approach to reduce the degree of ill-posedness and use a sparse piecewise constant parametrization for κ and c, while the actual scale of inhomogeneities is possibly smaller.
- We might be forced to employ an approximate reduced order computational model that would give biased predictions for the measurements even without any other uncertainties.

It should be clear that if these uncertainties and errors are neglected in the modelling, the accuracy and reliability of the estimates for κ and c can be highly questionable. In particular, how reliable are the respective error estimates, such as posterior covariance, for these parameters? Furthermore, it is evident that the statistics of the related errors will be time-varying.

To proceed, assume that we are to employ FEM semi-discretization (7), giving the time-dependent formulation

$$\bar{G}_\delta(c;\rho)\frac{\partial T}{\partial t} = \bar{K}_\delta(\kappa)T + \bar{A}_\delta(\kappa)T + \bar{B}_\delta(\kappa)f \tag{32}$$

in which the abstract parameter δ refers to the errors due to discretization and other uncertainties. The matrices $\bar{A}_\delta$ and $\bar{B}_\delta$ are related to the boundary conditions, and $\bar{G}_\delta$ and $\bar{K}_\delta$ are the mass and stiffness matrices, respectively. Due to the uncertainties and errors, all these matrices are in fact stochastic in the sense that they depend on the realization of the experiment, that is, the *actual unknown coefficients*[12]. This means that the model (32) is to be interpreted as a stochastic system of differential equations which in turn means that the solution of (32) is to be interpreted as a

[12] Note that the term *stochastic matrix* is also used to refer to matrices whose row and/or column sums are unity. This is not the case here.

probability distribution $\pi(T, t)$ at each time t. The evolution of probability distributions of drift-diffusion type model is governed by the Fokker-Planck equation [118].

We interpret the model (32) as follows. Assume that the initial temperature distribution is known. Let $T_*(\vec{r}, t)$ be the true temperature evolution of the real physical experiment at times $t = t_k$ and locations $\vec{r} = \vec{r}_\ell$. Then, the predictions of the model (32) with the stochastic models for all related matrices should be consistent with the realization T_*. As in the previous sections, this means that the probability $\pi(T_*, t)$ is relatively high for all t when compared to $\max_T \pi(T, t)$.

Let us fix the density and consider only the explicit dependence of the FEM matrices on the primary unknowns, and subsequently integrate (32) using, for example, a single step implicit Runge-Kutta scheme [49] over the time intervals (t_k, t_{k+1}) to give

$$\begin{aligned} T(k+1) &= F(\kappa, c)T(k) + B(\kappa, c)f(k) \\ &\quad + C(\kappa, c) + W(k) \end{aligned} \tag{33}$$

where W_k is a stochastic process which means that given a fixed $T(k)$, $T(k+1)$ is a probability distribution. The model (33) is referred to as the *(state) evolution model* with respect to the variable $T(k)$. If we start at a spatial temperature distribution $T(t)$ at time t_k, the distribution of $T(k+1)$ should be consistent with the actual state $T_*(t_{k+1})$ as defined above. With the model (33), this means that we have to be able to model the process $W(k)$ accordingly.

9.2 State space representation

A suitable statistical framework for dealing with unknowns that are modelled with stochastic processes and which are observed either directly or indirectly, is the *state estimation framework*. In this formalism, the unknown is referred to as *the state variable*, or simply *the state*. For treatises on state estimation and Kalman filtering theory in general, see for example [114, 117]. For the general nonlinear non-Gaussian treatment, see [112], and state estimation with inverse problems, see [17].

In the following, we consider only the state estimation problems with the most common assumptions. It must be noted that these assumptions are not necessary for the general state estimation problems, but the associated estimation procedures may become much more involved and that the exact interpretation of the results may change.

The standard discrete time *state space representation* of a dynamical system is of the form

$$x_{k+1} = F_k(x_k, w_k) \tag{34}$$

$$d_k = G_k(x_k, v_k) \tag{35}$$

where w_k is the *state noise process* and v_k is the *observation noise process*, while (34) and (35) are the *evolution model* and *observation model*, respectively. We do not state the exact assumptions here, since the assumptions may vary somewhat resulting in different variations of Kalman recursions, see for example [114, 119]. It suffices here to state that all sequences of matrices are assumed to be known and that the state and observation noise processes are temporally uncorrelated and that their (second order, possibly time-varying) statistics are known. Under these assumptions, the state process is a first order Markov process. The first order Markov property facilitates recursive algorithms for the state estimation problem. The Kalman recursions were first derived in [113].

9.3 Prediction, filtering and smoothing

Formally, the state estimation problem is to compute the distribution of a state variable $x_k \in \mathbb{R}^N$ given a set of observations $d_j \in \mathbb{R}^M, \ j \in \mathcal{I}$ where $\mathcal{I}$ is a set of time indices. In particular, the aim is to compute the related conditional means and covariances. Usually, $\mathcal{I}$ is a contiguous set of indices and we denote $D_\ell = (d_1, \ldots, d_\ell)$.

We can then state the following common state estimation problems:

- *Prediction.* Compute the conditional distribution of x_k given D_ℓ, $k > \ell$.
- *Filtering.* Compute the conditional distribution of x_k given D_ℓ, $k = \ell$.
- *Smoothing.* Compute the conditional distribution of x_k given D_ℓ, $k < \ell$.

The solution of the state estimation problems in linear Gaussian cases is usually carried out by employing the Kalman filtering or smoothing algorithms that are based on Kalman filtering. These are recursive algorithms and may be either real-time, on-line or batch type algorithms. In more general cases, one has the choice between the MCMC type particle filtering algorithms and the approximate extended Kalman filtering variants.

9.4 The linear Gaussian case

For linear Gaussian state estimation problems, all posterior densities are Gaussian and one only needs to compute the conditional means and covariances. We write the state space representation in the form

$$x_{k+1} = F_k x_k + B_k u_k + s_k + w_k \tag{36}$$

$$d_k = G_k x_k + v_k \tag{37}$$

where u_k is the control input, B_k the related control response model (if applicable), and s_k is a deterministic term that can be due to, for example, nonzero state noise mean. The term s_k practically never appears in the literature although it is almost always nonzero in real problems.

For these problems, the densities and, in particular, the conditional mean (minimum mean square) estimates, can be computed analytically. Furthermore, these can be computed recursively with the Kalman filtering algorithm. Let us denote $\mathbb{E}(x_k \,|\, D_\ell) = x_{k|\ell}$ and $\mathrm{cov}\,(x_k \,|\, D_\ell) = \Gamma_{k|\ell}$. The standard *innovation form* Kalman filter and the (one step) predictor recursions take the form

$$\begin{aligned}
x_{k|k-1} &= F_{k-1}x_{k-1|k-1} + s_{k-1} + B_{k-1}u_{k-1} && (38)\\
\Gamma_{k|k-1} &= F_{k-1}\Gamma_{k-1|k-1}{F_{k-1}}^{\mathrm{T}} + \Gamma_{w_{k-1}} && (39)\\
K_k &= \Gamma_{k|k-1}{G_k}^{\mathrm{T}}\big(G_k\Gamma_{k|k-1}{G_k}^{\mathrm{T}} + \Gamma_{v_k}\big)^{-1} && (40)\\
\Gamma_{k|k} &= \big(I - K_kG_k\big)\Gamma_{k|k-1} && (41)\\
x_{k|k} &= x_{k|k-1} + K_k\big(y_k - G_kx_{k|k-1}\big) && (42)
\end{aligned}$$

where $x_{k|k}$ and $x_{k|k-1}$ are the conditional means of the filtering and prediction densities, respectively, and K_k is the so-called Kalman gain. Equations (38-39) are often called the *time update* while equations (40-42) are called the *measurement update.*

As noted above, on-line imaging, perhaps for quality assurance purposes, but *without* automatic control is feasible even when the state estimates are not obtained immediately after the observation y_t. Furthermore, in transient type situations the estimates can possibly be computed completely off-line. The two relevant schemes for these two cases are the *fixed-lag smoother* $x_{k-h\,|\,k}$, where $h > 0$ is the lag in the estimation and the *fixed-interval smoother* $x_{k\,|\,t_{\mathrm{F}}}$, where t_{F} is the final time of the observations.

How much better the smoothed estimates are when compared to the real-time Kalman filtered estimates, depends on the overall state space model in a complex way. In some cases the filtered estimate errors possess a delay type structure which is largely absent in the smoothed estimates. This is typical behavior especially in cases in which the observation model is not exceptionally informative[13] and the state evolution model is not very accurate. With this we refer to the situation in which the uncertainties are significant and thus $\mathrm{var}\,\|x_k\| \gg \mathrm{var}\,\|w_k\|$ does not necessarily hold.

9.5 Nonlinear non-Gaussian cases

The *extended Kalman filter* algorithms (EKF) form a family of estimators that do not possess any optimality properties. For many problems, however, the EKF algorithms provide feasible state estimates. For EKF algorithms, see for example [114, 17].

[13] In these cases the maximum likelihood problem $\max \pi(y_t|x_t)$ is not stable

The idea in extended Kalman filters is straightforward: the nonlinear mappings are approximated with the affine mappings given by the first two terms of the Taylor expansion. We note that the notion of extended Kalman filter is not completely fixed and several levels of refinement can be referred to with this term. The state in which the linearization is computed, will be denoted with x_t^*. We describe briefly the three most common variants of the extended Kalman filters.

In the *global linearization* approach, the mappings F_t and G_t are linearized at some fixed (time-invariant) state $x_t^* \equiv x^*$ so that we have $F_t(x_t) \approx F_t(x^*) + J_{F_t}|_{x^*}(x_t - x_*) = b_t + J_{F_t}|_{x^*} x_t$, and similarly for G_t. Here J_{F_t} is the Jacobian mapping of F_t. The rationale behind global linearization is that the Jacobian does not have to be recomputed during the iteration. It is clear that a good guess of the "mean state" x^* is a prerequisite for this approximation to be successful.

The version of extended Kalman filter that is most commonly used, is the *local linearization version*, in which version the mappings are linearized at the best currently available state estimates, either the predicted or the filtered state. This necessitates the recomputation of the Jacobians at each time instant. Note that the linearization is *not necessarily* needed in the time update when the predictor $x_{k|k-1}$ is computed. This applies also for the measurement update. Furthermore, one does not have to approximate the control term when the state estimates are computed. The recursions take the form

$$x_{k|k-1} = F_{k-1}(x_{k-1|k-1}) + s_{k-1} + B_{k-1}(u_{k-1}) \tag{43}$$

$$\Gamma_{k|k-1} = J_{F_{k-1}} \Gamma_{k-1|k-1} {J_{F_{k-1}}}^{\mathrm{T}} + \Gamma_{w_{k-1}} \tag{44}$$

$$K_k = \Gamma_{k|k-1} {J_{G_k}}^{\mathrm{T}} \big(J_{G_k} \Gamma_{k|k-1} {J_{G_k}}^{\mathrm{T}} + \Gamma_{v_k} \big)^{-1} \tag{45}$$

$$\Gamma_{k|k} = \big(I - K_k J_{G_k} \big) \Gamma_{k|k-1} \tag{46}$$

$$x_{k|k} = x_{k|k-1} + K_k \big(y_k - G_k(x_{k|k-1}) \big) \tag{47}$$

The linearizations are needed only in the computation of the covariances and the Kalman gain. However, if the computation of the Jacobians is faster than for example $G_k(x_{k|k-1})$, then the affine approximations could be used in (43) and (47).

The third version is the *iterated extended Kalman filter*. Assume here that the state evolution equation is linear and thus there are no problems in computing the prediction covariances, and let the observation model be nonlinear. The idea is easiest explained based on the Bayesian interpretation of the Kalman filter. To keep the treatment brief, we note that the measurement update is equivalent to the computation of the conditional mean estimate for the state x_k given the measurement history $(d_1, \ldots, d_k)$, which is $E(x_k \,|\, d_1, \ldots, d_k)$. In Section 4.3 we saw that in the Gaussian case the computation of the mean of the posterior density $\pi(x_k \,|\, d_1, \ldots, d_k)$ coincides with the computation of the maximum of the density. It can be shown that we have

$$\pi(x_k | d_1, \ldots, d_k) \propto \pi(d_k | x_k) \pi(x_k | d_1, \ldots, d_{k-1}) \tag{48}$$

where the first density on the right hand side is the likelihood density and the latter is called the prediction density.

Assume that the predictor covariance $\Gamma_{k|k-1}$ and Γ_{v_k} are positive definite for all k so that the Cholesky factorizations $\Gamma_{k|k-1}^{-1} = L_2{}^{\mathrm{T}} L_2$ and $\Gamma_{v_k}^{-1} = L_1{}^{\mathrm{T}} L_1$ exist. In our case the maximization of the posterior density is equivalent with the minimization of the following quadratic functional so that we can write

$$\begin{aligned} x_{k|k} &= \arg\min_x \big\{ \|L_1(y_k - G_k(x))\|_2^2 \\ &\qquad + L_2(x - x_{k|k-1})\|_2^2 \big\} \end{aligned} \tag{49}$$

which can be solved for example with the Gauss-Newton algorithm. Thus, in the iterated extended Kalman filter, we would compute (43-44) as before, but (45-47) would be replaced by first computing the estimate $x_{k|k}$ by minimizing (49) and then computing $\Gamma_{k|k}$ from (45-46) so that the Jacobians J_{G_k} are recomputed at $x_{k|k}$. The case of nonlinear state evolution equation is an isolated problem of solving a nonlinear differential (difference) equation. The evaluation of the predictor covariance, however, may call for a Taylor series approximation. This is turn may be a complex undertaking, see [120] for an example that is related to hydrological flows.

9.6 State space identification

In practical problems it is usually the case that one or more of the state space terms is at least partially unknown. While careful analysis of the measurement system may determine the observation model relatively accurately, the state evolution model is always inaccurate to some extent.

In the batch type optimization approaches the task is to estimate the *time-invariant* parameters for example by the maximum likelihood method. However, if the unknown parameters are assumed to be *time-varying*, it is possible to augment the state variable to include the unknown parameters.

Denote the unknown physical parameters that are to be estimated by μ. Several parameters of the state space representation might depend on the parameters μ. As a typical example, we might be able to sustain the model $\Gamma_{w_t} \equiv \sigma_w^2 I$ but cannot specify a fixed σ_w^2, and thus have to come up with a model $\pi(\mu)$ with $\mu = \sigma_w^2$.

The idea is to compute the likelihood of the observations $D_{t_{\mathrm{F}}} = (d_1, \ldots, d_{t_{\mathrm{F}}})$ given the parameters μ and maximize the likelihood with respect to μ, that is,

$$\max_{\mu} \pi(D_{t_{\mathrm{F}}} \,|\, \mu). \tag{50}$$

The likelihood depends on the unknown parameters. For example, consider the case in which the initial state distribution is known. Then, in the case of a linear

Gaussian problem, the likelihood $\pi(D_{t_\mathrm{F}}\,|\,\mu)$ can be written in the form [117]

$$\pi(D_{t_\mathrm{F}}\,|\,\mu) = C - \frac{1}{2}\sum_{t=1}^{t_\mathrm{F}} \big(\log|\Gamma_{t|t-1}| + {e_t}^\mathrm{T}\Gamma_{t|t-1}e_t \mid \mu \big) \tag{51}$$

where C is a constant, $|\cdot|$ denotes determinant of a matrix, $e_t = d_t - G_t x_{t|t-1}$ is the prediction error and the notation $(\cdot\,|\,\mu)$ refers to all variables and their time evolutions being calculated with parameter μ.

There are certain other analytical forms for the likelihood with various types of unknown parameters. However, most often it is necessary to use Newton or quasi-Newton type methods with numerically approximated gradients to compute the maximum of (51). A particularly suitable algorithm is the BFGS quasi-Newton algorithm. Some numerical considerations are given in the general state space references and more details can be found for example in [121]. See also [122] for the application of the expectation-maximization (EM) algorithm in this problem.

9.7 Approximation errors and model reduction in nonstationary problems

The treatment of the systematic approximation error approach for nonstationary inverse problems is outside the scope of this paper. As with stationary inverse problems, the approach is usually relatively straightforward but may be tedious in the sense that it again calls for systematic modelling of the uncertainties.

The nonstationary approximation error approach was developed in [123, 110, 124], in which linear state estimation and nonlinear state space identification problems that were related to heat transfer, are discussed. This approach is applicable to handling model reduction as well as long time stepping. Both the evolution and observation models are modified when approximation errors are present. The approach was further developed in [125, 120] to employ an importance sampling type modification which involves iterative recomputation of statistics of the approximation error while the data is accumulated.

We stress again the importance of the feasibility and consistency of the models with reality. For an example, we refer to [110] in which the estimation of diffusivity and specific heat capacity is considered. It is shown that if approximation errors are not taken into account systematically, then error estimates are significantly over optimistic, and it is essentially impossible to recover the true values from the resulting model.

10 Computational aspects

As with all iterative methods, computational efficiency of sample-based inference is critical if estimates are to be evaluated in any reasonable time. In this section we discuss model reduction and other ways of speeding up evaluation of the forward map, and considerations for accelerating MCMC sampling.

10.1 Model reduction

In Section 9, we referred to model reduction related issues, and discussed one route for coping with model reduction in Section 8.5. The use of reduced models within MCMC is discussed in Section 10.6. We now discuss briefly *how* model reduction is carried out. A systematic approach to model reduction, especially in the context of dynamical systems, can be found in [126].

With forward problems that are induced by PDE's and the related IBVP's, the most obvious approach is to use coarse meshes. As in Section 2.5, consider the finite element approximation (7). How accurate the finite element solution with this approximation is, depends on the particular mesh used (how well its density is adapted to the local temperature gradients), the dimension N and the integration method used to solve the system of ODEs.

Here, the dimension N is the obvious target for model reduction, as is the length of the time steps $\delta t = t_{k+1} - t_k$. The theory of finite element methods and numerical methods for ODE's give error estimates to the solutions but these are only up to (usually multiplicative) constants, see references in Section 2.5. In practice, simulations over the expected range of material and other parameters have to be carried out to find out about the actual errors of the forward solvers.

One approach to the assessment of tolerable errors would be that the maximum errors, over the expected range of uncertainties, should be smaller than the standard deviation of the (additive) errors. This will typically lead to a very dense discretization and dense time stepping. Due to the nature of inverse problems, if the model errors are larger than, say, 1-2 standard deviations and they are not taken into account in any way, there is a high risk that the estimates and error estimates are highly misleading.

The related errors can, however, be handled up to a degree as explained in Sections 8.5 and in the references given in Section 9.7, as well as the methods in Section 10.6.

Another model reduction type is to write the *unknown* $\bar{x}$ as a projection x onto a small-dimensional subspace, that is,

$$\bar{x} \approx x = \sum_{k=1}^{p} x_k \vartheta_k \tag{52}$$

Naturally, the basis $\{\vartheta_k\}$ determines the characteristics of $\bar{x}$ that can be retrieved in theory. Very often with finite elements, for example, the basis $\{\vartheta_k\}$ is adapted to the FEM basis functions φ_ℓ in the sense that ϑ_j are characteristic functions of (unions of) the elements used to construct the FEM basis. This approach makes it easy to compute the integrals in the variational form. A multilevel coarsening scheme, in which the FEM basis functions are also variationally coarsened, is applied to an inverse problem in [127].

The state basis is usually obtained by sampling the parametric input space, propagating through the forward model, and computing a basis which spans the space of the resulting states. Popular methods for sampling the inputs vary little in methodology but go by many different names. One very popular method is to select the inputs one expects for the particular problem. Then, the resulting states are converted to an efficient basis by means of a singular value decomposition (SVD). This method goes by many titles including proper orthogonal decomposition (POD) [68], principal components analysis (PCA) [60], [20], Karhunen-Loeve (K-L) expansion [46], and empirical orthogonal functions (EOF) [62], depending on the community.

In [128] a greedy sampling algorithm is employed to determine appropriate sample locations to reduce errors between full-order and reduced-order outputs, giving an algorithm that is suitable for large-scale systems. An example is given for the thermal design and analysis of heat conduction fins.

The SVD provides insight into redundant data and guides the user in deciding how many basis vectors to retain[14]. In fact, the resulting basis can be shown to be optimal in the sense that the representation of the states in the basis produces minimal error in the 2-norm.

The essence of POD is as follows. Let again $\bar{x} \in \mathbb{R}^N$ be the approximation as in (52), where N might be too large for our purposes. Assume that the prior model $\pi(x)$ for the unknown be proper with finite covariance matrix Γ_x. Note that we do not require $\pi(x)$ to be a Gaussian distribution. Let the eigenvalue decomposition of Γ_x be

$$\Gamma_x = \sum_{k=1}^{N} \lambda_k \varphi_k {\varphi_k}^{\mathrm{T}}$$

and let the ordering be such that $\lambda_k \geq \lambda_{k+1}$ for all k. It can be shown that if the basis in (52) is chosen so that $\vartheta_k = \varphi_k$, the approximation in (52) is the best in the mean square sense for all p, that is,

$$\mathbb{E}\|x - \bar{x}\|^2 = \mathbb{E}\left\| x - \sum_{k=1}^{p} \langle x, \vartheta_k \rangle \vartheta_k \right\|^2$$

[14] Note that for symmetric matrices, the singular value decomposition is equivalent with the eigenvalue decomposition.

attains its minimum if we set $\vartheta_k = \varphi_k$. This holds for all p. It is also important to note, that the approximation $\bar{x} \in \mathcal{S}_p \subset \mathbb{R}^N$, that is, $\bar{x}$ is represented in a subspace of dimension p, but is a vector in $\mathbb{R}^N$.

How to handle the related approximation error due to the truncation of the series representation? First, note that when $p = N$, we have $\bar{x} = x$ since the eigenvectors $\{\vartheta_k\}$ are orthonormal and form a complete set in $\mathbb{R}^N$. Consider again the linear additive error model $d = Ax + e$. Then, we can write

$$x = \sum_{k=1}^{p} \langle x, \vartheta_k \rangle + \sum_{k=p+1}^{N} \langle x, \vartheta_k \rangle = \bar{x} + \tilde{x}$$

and further

$$\begin{aligned} d &= Ax + e = A\bar{x} + A\tilde{x} + e \\ &= A\bar{x} + \varepsilon(x) + e \end{aligned}$$

This means that the effects of the truncated series representation for the unknown can be handled within the approximation error formalism explained in Section 8.5. The above hold for all orthonormal bases $\{\vartheta_k\}$. However, if we have set $\vartheta_k = \varphi_k$, the covariance of $\varepsilon(x)$ is minimized, that is, the least bad approximation error is induced.

10.2 Forward solvers

Because computation of the forward map happens within each step of the MCMC, the speed of the overall computation depends critically on efficient solution of the governing PDEs.

For FEM meshes or FDM grids with up to tens of thousands of nodes, or in BEM partitions with up to thousands of elements, efficient solution of the systems (6) and (8) can be performed by first factorizing matrices. For 2-dimensional problems, efficient solution for FEM or FDM is achieved by first operating by a bandwidth reducing permutation of the sparse system matrix, followed by Cholesky factoring [129] and solution. For 3-dimensional problems with fine meshes, multigrid solvers are significantly faster, and also provide access to cheap solutions at coarse scales that may be utilized within the MCMC to decrease overall compute time [127]. When time-dependence is evaluated by convolution, interpolation methods similar to fast multipole methods (FMM) [130] are efficient [51].

When changing the thermophysical properties in an iterative optimizer or MCMC, a straightforward approach is to reform the system matrices at each iteration. However, it is also feasible to directly maintain the QR factorization in BEM, and the Cholesky factorization in FEM or FDM [129], with gain in computational efficiency.

Also possible is direct updating of *solutions* using the Woodbury formula [34] though this is numerically unstable in the long term.

Both FEM and BEM formulations also allow efficient calculation of derivatives. In a FEM formulation, operation by the Jacobian may be performed using only solutions at the current state [131]. In BEM formulations, expressions for the Fréchet derivative of the forward map allow evaluation of the gradient with respect to the *boundary* by solving a non-homogenous equation with the current BEM matrices [132].

10.3 Precomputations

Computation of fixed quantities needed in each step of the MCMC may be performed once at initialization, for substantial gain in efficiency. For FEM systems with a fixed mesh, the linear map from thermophysical properties to system matrix may be precomputed, as can the bandwidth-reducing permutation. In solvers based on BEM or FMM, precomputation and tabulation of the fundamental solution and geometric terms is critical to producing efficient code.

When using fast surrogates for the forward map [133, 134] or model reduction, these can be precomputed along with the enhanced error model discussed in Section 8.5.

Computational cost of evaluating the prior distribution is often neglected, though it can be significant particularly once efficient methods for the forward solver and likelihood evaluation have been implemented. For prior models using neighborhood structure, or other spatial features, it is often important to use efficient data structures with precomputed geometric terms.

10.4 MCMC practicalities

Practical computation generates a chain of finite length, so an obvious question is: how large does N_s need to be for estimates evaluated via (25) to be sufficiently accurate? While the central limit theorem guarantees finite variance in estimates, it gives no hint to the size of the variance in practice. Similarly, the ergodic properties of a finite chain may be imperfect so that there remains some residual effect of the starting state, and the chain may not be effectively irreducible, perhaps getting stuck in a mode of the distribution and giving no clue to the presence of other modes. These practical issues need to be diagnosed from the computational implementation.

To a large degree the problem of dependence on starting state may be mitigated by discarding the 'burn-in', that is, throwing away the m samples at the beginning of the chain before the chain has become independent of the starting state [135]. Inference is then based on the N_s samples $\left\{x^{(i)}\right\}_{i=m+1}^{N_\mathrm{s}+m}$.

A strategy that can be taken, particularly when debugging code, is to run multiple chains (typically of the order ~ 10) using the same MCMC algorithm but starting from randomly chosen starting states. The problem of the chain getting stuck is usually identifiable this way. This type of an approach is also often used when diagnosing whether optimization algorithms have got stuck in a local optimum. Ideas for selecting a starting distribution are discussed in [136]. The sample mean of any property g is approximated by computing over each chain,

$$\bar{g}_{N_\mathrm{s}} = \frac{1}{N_\mathrm{s}} \sum_{i=1}^{N_\mathrm{s}} g\left(x^{(i)}\right) \tag{53}$$

and the sample variance

$$\sigma^2_{N_\mathrm{s}} = \frac{1}{(N_\mathrm{s}-1)} \sum_{i=1}^{N_\mathrm{s}} \left(g\left(x^{(i)}\right) - \bar{g}_{N_\mathrm{s}}\right) \tag{54}$$

and check that the intervals $\bar{g}_{N_\mathrm{s}} \pm 3\sigma_{N_\mathrm{s}}$ substantially overlap. We increase N_s until this check is satisfied. This gives a check against multiple possible problems with MCMC including dependence on the starting state, the possibility of the chain getting stuck in local modes, etc. This exhaustive computing may be called the "many long chains" approach [136, 69], and is a good self-consistency check of ergodic properties. Of course, these checks indicate that the chain is mixing and converging in distribution; they do not prove that the chain is converging to the *correct* distribution. To check convergence to the desired distribution one often performs transformations on the MCMC that should have no effect on the equilibrium distribution, such as varying move ratios, and verify that the output statistics are indeed unchanged.

10.5 Efficiency of the MCMC

Efficiency of an MCMC algorithm can be measured by how quickly the sample mean in (53) converges to the asymptotic value $\mathbb{E}(g(x)\,|\,d)$, as N_s increases, that we want in (25). Although we are guaranteed that $\mathbb{E}(\bar{g}_{N_\mathrm{s}}) = g$ and that $\lim_{N_\mathrm{s}\to\infty} \bar{g}_{N_\mathrm{s}} = g$ by the central limit theorem [68], that gives no idea how accurate this estimate is for finite N_s. In practice the accuracy needs to be determined from the chain, which can be done using standard results from time series.

Since the sequence of states $\left\{x^{(i)}\right\}_{i=1}^{N_\mathrm{s}}$ is a realization from a Markov chain then so is $\left\{g\left(x^{(i)}\right)\right\}_{i=1}^{N_\mathrm{s}}$. The sum defining $\bar{g}_{N_\mathrm{s}}$ may be viewed as the output at one time of a running-mean filter when the input is the (infinite) sequence $\left\{g\left(x^{(i)}\right)\right\}$. Since the filter relates the autocorrelation of input and output sequences, and the variance is just the autocovariance at zero lag, it follows that

$$\text{var}\ (\bar{g}_{N_\text{s}}) = \frac{\text{var}\,(g)}{N_\text{s}} \left[\sum_{j=-N_\text{s}+1}^{N_\text{s}-1} (1 - \frac{j}{N}) \rho_j \right] \tag{55}$$

where ρ_j is the correlation coefficient of the input with lag j. The term in the brackets is the integrated autocorrelation time (IACT), denoted by τ_g. Since var $(\bar{g}_{N_\text{s}}) =$ var $(g)/N_\text{s}$ for independent samples, as quoted in Section 4.4, the IACT gives the effective number of samples with the same variance reducing power as one independent sample. When $N_\text{s} \gg \tau_g$ we have $\tau_g = \sum_{j=-\infty}^{\infty} \rho_j$. As with estimation of power spectra, time windowing or spectral smoothing is needed for practical estimation of τ_g. A suitable practical estimator is given by using sample estimates of $\gamma_m = \rho_{2m} + \rho_{2m+1}$ and truncating the sum at $2n$ terms when $\gamma_{n+1} > \gamma_n$ or $\gamma_{n+1} < 0$, that is, the sample estimates are non-decreasing or go negative [69, 137, 138].

The IACT is a measure of relative efficiency of a MCMC scheme since small IACT means that a shorter chain, using less compute time, is required to calculate results to a given accuracy, and hence is more *statistically efficient.* In a standard MH scheme, this means tuning the proposal distribution to give a minimal value for IACT. Proposal distributions may be tuned in many ways, for example in *random walk* proposals, the width of the window can be tuned for particular problems since it greatly affects statistical efficiency and the IACT, see [67, 139]. The example in Section 5.1 shows the effect of choosing window sizes. In inverse problems, it is common to construct proposal distributions using a number of *moves*, giving a chain that has a mixture kernel [131]. Each move has an associated proposal distribution that is chosen with probability p_i, $\sum_i p_i = 1$. Tuning the p_i can also greatly affect statistical and computational efficiency, see for example [140].

10.6 Acceleration schemes for Metropolis-Hastings MCMC

The advantages of MCMC based inference come at the computational cost of solving the forward map typically hundreds of thousands times to explore the posterior distribution. Hence much research effort has gone into finding ways of accelerating the basic MH algorithm. We review several schemes that have proved useful in solving inverse problems, see also [141]. As with all variants of the MH algorithm, these schemes may be viewed as ways of improving the proposal distribution to give better mixing.

Simulated tempering. Consider the case where the MH algorithm is used to sample from posterior distribution $\pi(\cdot)$ using some proposal distribution, and it is found that the resulting chain is evolving slowly, or worse still, is getting stuck. Simulated tempering [142] (with the name and idea adapted from simulated annealing) defines a sequence of distributions $\{\pi_\ell(\cdot)\}_{k=0}^{P}$ where $\pi_0 = \pi$ and $\pi_1(\cdot), \pi_2(\cdot), \ldots, \pi_P(\cdot)$ are distributions that are increasingly easier to sample from. The distribution over the augmented space is taken as

$$\pi(x, \ell) = \lambda_\ell \pi_\ell(x) \tag{56}$$

where $\lambda_0, \lambda_1, \cdots, \lambda_N$ are pseudo prior constants with $\sum_{\ell=1}^{P} \lambda_\ell = 1$. Transitions for a fixed k are derived from the original proposal, which are interspersed with proposals that change k (perhaps by a random walk in k) with both accepted/rejected by a standard MH algorithm. Samples from the conditional density $\pi(x, \ell | \ell = 0)$, are samples from the desired distribution.

A simple tempering scheme, used to overcome difficulties with multi-modal distributions, is given by the sequence of distributions $\pi_\ell(x) = \lambda_\ell \pi^{\beta_\ell}(x)$ which are increasingly unimodal, where $1 = \beta_0 < \beta_1 < \cdots < \beta_P$ are inverse temperatures. The opposite regime, of increasing temperature, has found greater success in inverse problems where high-accuracy data leads to posterior distributions that are too narrow to easily sample [143].

Parallel tempering is similar to simulated tempering except that the P chains (one for each value of ℓ) are maintained simultaneously. An example is the Metropolis coupled MCMC in [144] that simultaneously runs chains with the spatial parameters increasingly coarsened, defining a sequence of distributions as above. Evolutionary Monte Carlo algorithms are examples of parallel tempering with moves inspired by the genetic optimization algorithms [145].

Parallel rejection. The parallel rejection algorithm utilizes m computer processors to give a straightforward parallelizing of the serial MH algorithm [146, 147]. Each processor runs an *independent* instance of the MH algorithm initialized at state $x^{(n)}$ to give the m independent Markov chains $\left\{\phi^{(r,k)}\right\}_{k=0}^{\infty}$ for $r = 1, 2, \ldots, m$ with $\phi^{(r,0)} = x^{(n)}$. Enumerate the resulting states by $s\,(r, k) = r + m\,(k-1)$ for $r = 1, 2, \ldots, m$ and $k = 1, 2, \ldots$ giving the total ordering $s = 1, 2, \ldots$. The m parallel chains are run until the first non-trivial acceptance (in the order s) occurring at $s_{\min}$, that is, the minimum s for which $\phi(s) \neq x(t)$. Then set $x^{(j)} = x^{(n)}$ for $j = n+1, n+2, \ldots, n+s_{\min}-1$ and $x\,(n+s_{\min}) = \phi^{(s_{\min})}$, and reinitialize.

For the acceptance rate α and time for transactions relative to the forward map β, the speedup factor is

$$\frac{(1-(1-\alpha)^n)}{\alpha} \frac{1}{1+n\beta}.$$

Using approximations to the forward map. The model reduction methods of section 10.1, and surrogate models mentioned in section 10.3, can achieve significant speedup. However, by approximating the forward map, these methods only give access to an approximation of the true posterior distribution, see also [148]. The approximation error approach, described in Section 8.5 mitigates the effects of this approximation.

Sampling from the exact posterior distribution while utilizing a fast approximation to the forward map can be achieved using the 'delayed acceptance' algorithm [131]. This algorithm allows a state-dependent approximation $\pi_x^*(\cdot|d)$ to the posterior distribution calculated using a cheap approximation to the forward map. Once a proposal is generated, to avoid calculating $\pi(x'|d)$ for proposals that are rejected, the algorithm first tests the proposal using the approximation $\pi_x^*(x'|d)$ to create a second proposal distribution that is then used in a standard MH algorithm. Approximations based on local linearizations [131], coarse partitioning in BEM [140], and coarsened solutions available in a multi-level (multigrid) solver [127], have been used, in the context of inverse problems.

Adaptive MCMC. Adaptive MCMC algorithms optimize performance by learning better proposal distributions from the past output of the chain [149, 150]. Recent developments in the theory of adaptive algorithms have produced simplified requirements that ensure ergodicity, allowing wide-ranging application [151], including automatic tuning of proposal windows [152].

11 Further topics

In standard inference problems, such as the conventional parameter estimation problems, the end task is to obtain, say, a point estimate or a few estimates as well as some information on the spread of the variables. However, very often the choice of unknown parameters is dictated by the structure of the forward problem and not by the end objective of the overall task. Although it might be interesting to view the time-varying temperature distribution in a target volume visually, the end task might, however, be to use these state estimates and their distribution as a control input to the process. The end objective in such a case would probably be the yield of the process, or prevention of a system failure.

In this section, we treat some additional topics and types of problems, for which the Bayesian approach provides a natural framework. In particular, *outside* the Bayesian framework, these may not be accessible at all. As an example, problems with variable (unknown) dimensions are practically impossible to formulate as a feasible optimization problem. Another example is that of model selection, often between two possible prior models.

In Bayesian statistics, these are more or less standard topics. With inverse problems, however, many of these possibilities have seldom or never been investigated, yet. Partly, this is due to the special structure of inverse problems, the often great dimensionality and/or the poor information content of the measurements.

In addition to the topics treated in this section, several important ones, such as model validation, robustness to prior design and invariance, have been omitted in this paper. As general references to this section, including the topics that are not discussed here, see [153, 26, 154, 61].

11.1 Predictive inference

Assume that instead of finding out the unknown x, the final goal is actually predicting a random variable g, possibly a future observation, with weather forecasting serving as a natural example. Assume further that data (measurements) d has been obtained to estimate x, that is, to obtain the posterior distribution $\pi(x|d)$. We can often assume that the random variables d and g are mutually independent.

The traditional (non-Bayesian) approach would be to compute a point estimate for x and then estimate g based on this point estimate. In the Bayesian framework, however, this is handled differently since there is uncertainty in x (even) given d [26]. Thus, assuming mutually independent d and g and using the Bayes theorem repeatedly, we get

$$\pi(g|d) = \int \pi(g|x)\pi(x|d)\,\mathrm{d}x$$

which gives the uncertainty in g given the only available data d and the associated likelihood and prior models. The predictive distribution $\pi(g|d)$ implicitly incorporates the uncertainty in the unknown x. As mentioned in Section 4.5, note that

$$\pi(g|d) \neq \pi(g|x_*)$$

generally for any fixed x_*, including $x_* = \mathbb{E}(x|d)$, and also that the posterior uncertainty proposed by $\pi(g|x_*)$ is usually significantly too optimistic.

11.2 Combining data from different experiments

Assume that we carry out two different experiments with the unknowns x and μ, which experiments carry complementary information on the unknowns. Let y and d be the (mutually independent) data and $\pi(d|x,\mu)$ and $\pi(y|x,\mu)$ be the associated likelihoods.

Then, it is straightforward to see that

$$\pi(x,\mu|d,y) \propto \pi(y|x,\mu)\pi(d|x,\mu)\pi(x,\mu)$$

that is, the product of the two likelihood and the prior models. This is the optimal and natural way to combine data and to take into account the information provided by the experiments, as well as the relative uncertainties.

Sometimes one experiment can be arranged so that the associated likelihood depends only on one of the unknowns, for example, $\pi(y|x,\mu) = \pi(y|\mu)$. Assume further that we are primarily interested in x, so that we write

$$\pi(x|d,y) \propto \int \pi(y|\mu)\pi(d|x,\mu)\pi(x,\mu)\,\mathrm{d}\mu$$

which may yield itself to computationally efficient implementations, possibly via the approximation approach in Section 8.5. Again, $\pi(x\,|\,d,y) \neq \pi(x\,|\,\mu_*,d,y)$ for any μ_* and the same comments as in Section 11.1 apply here.

11.3 Optimization of experiments

First, we make a reference to the philosophy of deterministic regularization approaches. In these methods, regularization is forced to make the computation of the solutions possible in the first place. Thus, the measurement model, such as $d = Ax + e$ comes first and regularization is arranged to counter the ill-posedness of the operator A.

In the Bayesian framework, we *first* have to model the uncertainty in unknowns, that is, we construct the prior model $\pi(x)$ or $\pi(x,\mu)$. Then, the measurements are carried out and the posterior uncertainty is assessed. If this uncertainty is too large or the posterior model is improper, we don't go back to the prior model and adjust it to make the posterior model "nicer".[15] As noter earlier, one of the most appealing topics in the Bayesian framework is that the modelling of the uncertainties is separate from the modelling of the measurement process

What we can adjust in Bayesian statistics, is how *the measurements are carried out*, and furthermore, how to do this so that the measurements convey maximally complementary information relative to the prior uncertainty. Note that *how the measurements are made* fixes the model A.

To illustrate this idea, consider the standard case of linear Gaussian likelihood and prior model:

$$d = Ax + e \ , \quad e \sim \mathcal{N}(0, \Gamma_e) \ , \quad x \sim \mathcal{N}(x_*, \Gamma_x)$$

with mutually independent (e, x). The posterior covariance is

$$\Gamma_{x|d} = \left(A^{\mathrm{T}}\Gamma_e^{-1}A + \Gamma_x^{-1}\right)^{-1}$$

see Section 4.3. In the worst case, assume that A and Γ_x^{-1} have the same null space. Then, the matrix $A^{\mathrm{T}}\Gamma_e^{-1}A + \Gamma_x^{-1}$ is not invertible and the posterior distribution is improper. In other words, the posterior uncertainty is infinite in the following sense: there is at least one vector x' of unit norm so that $\pi(x + cx'\,|\,d) = \pi(x\,|\,d)$ for arbitrarily large c.

The task is to make such measurements that the uncertainty in the likelihood, which is determined (in the above simple example) by the joint structure of A and Γ_e

[15] At least when Bayesian inference is carried out honestly. Some Bayesians interpret the notion of "subjective probability" in the sense that they can do this. It has also been claimed that employing "too good physical models" is no longer Bayesian because there is not enough room for subjectivity. See [61, Chapter 11] for a brief discussion on this topic.

is complementary to the prior uncertainty. We need a measure for this uncertainty, such as the product of posterior variances, that is, $\aleph = \Pi_k \Gamma_{x|d}(k,k)$. We would then try to optimize the measurements, and thus A, so that $\aleph$ is minimized, or at made least tolerably small. See [100] and [155] for an impedance tomography related example in stationary and nonstationary cases, respectively.

11.4 Models of variable dimensions

Consider the case where we know the unknown physical coefficient to be piecewise constant, but that we don't know the number of subdomains nor the boundaries. In such a case the unknown naturally assumes a representation whose dimension is variable.

As an example, let there be p models $\mathcal{M}_k,\ k = 1, \ldots, r$. Let us fix the *prior probabilities of these models* as $p_k = \mathbb{P}(\mathcal{M}_k)$ with $\sum_k p_k = 1$, and where $\mathbb{P}(\cdot)$ denotes probability. We write

$$\mathcal{M}_k : \quad d \sim \pi_k(d|x_k)\ , \quad x_k \in \mathcal{X}_k.$$

As an example, consider the reconstruction of a spatially one-dimensional variable $x(t)$ which we discretize in N points. We know that $x(t)$ is usually a spatially smooth function but that occasionally there can be a number k of significant jumps. Thus, we model $x(t)$ as the combination of a spatially smooth process $z(t)$ and a jump process. In such a case, we can set $x_k = (z_1, \ldots, z_N, t_1, \ldots, t_k, \beta_1, \ldots, \beta_k)$, where t_ℓ specify the locations of the jumps, β_ℓ the jumps and k is the number of jumps. Thus $\mathcal{X}_k \subset \mathbb{R}^{N+2k},\ k = 0, \ldots, p-1$ and the *physical variable* $x(t)$ can be written as

$$x(t) = z(t) + \int \sum_{\ell=1}^{k} \beta_\ell \delta(\tau - t_\ell)\, \mathrm{d}\tau. \tag{57}$$

Thus, the models $\mathcal{M}_k$ correspond to different numbers of jumps.

It is clear that for such a model we cannot compute a conventional optimization type solution, not to speak of posterior uncertainty of x_k. Thus, sampling using the reversible jump MCMC formulation seems to be the only option here.

It is to be noted that even when we have computed the samples, we cannot compute the posterior mean of x_k since they are of different dimensions. We can, however, easily compute the conditional mean of (57) a well as spread estimates. See [93] for an example of such a procedure.

11.5 Model selection

This topic is closely related to that of Section 11.4, but the objective is different. As an example, assume that we carry out measurements on a target and the objective is to determine whether the target is faulty, that is, the objective is quality control. Assume also that we have constructed prior models $\pi_1(x)$ and $\pi_2(x)$ for faulty and intact targets, respectively. If we carry out exploration fixing either of these two models as a prior, we will obtain estimates that conform to the employed prior. If the model $\pi_2(x)$ corresponds to a spatially smooth function and we use this prior model, we get smooth reconstructions. Correspondingly, if the prior model $\pi_1(x)$ allows for jumps, we are likely to obtain jumps whether the actual target contains jumps or not. Rather than reconstructing the variable x, the objective here is to determine which of the models $\pi_1(x)$ or $\pi_2(x)$ is better supported by the measurements, that is, to *select the (prior) model.*

More generally, let there be *p competing models* which we denote by $\mathcal{M}_k,\ k = 1, \ldots, p$. Let the prior probabilities of these models be $p_k = \mathbb{P}(\mathcal{M}_k)$. As in Section 11.4, these models may not allow for parametrizations of the same type and the parametrizations often have different dimensions. For example, let x represent the inhomogeneous thermal conductivity and let $\mathcal{M}_1$ be a model in which x is constant in two different subdomains while in $\mathcal{M}_2$ there is a third subdomain. Furthermore, let $\mathcal{M}_3$ be a model with $x(\vec{r}) = \sum_{\ell=1}^{20} x_3(\ell)\varphi_\ell(\vec{r})$ with spatially smooth $\varphi_\ell(\vec{r})$, and so on.

The task is then to select a model which is best supported by the measurements d. First, note that formally

$$\pi(x, \mathcal{M}) = \pi(x \,|\, \mathcal{M})p(\mathcal{M})$$

and also

$$\pi(x) = \sum_\ell \pi(x \,|\, \mathcal{M}_\ell)p(\mathcal{M}_\ell).$$

Let the prior model for $\mathcal{M}_k$ be π_k. The posterior $\pi(\mathcal{M} \,|\, d)$ is a discrete distribution, and we have

$$\mathbb{P}(\mathcal{M} = \mathcal{M}_k \,|\, d) = \frac{p_k \int_{\mathcal{X}_k} \pi_k(d \,|\, x_k)\pi_k(x_k)\, dx_k}{\sum_\ell p_\ell \int_{\mathcal{X}_\ell} \pi_\ell(d \,|\, x_\ell)\pi_\ell(x_\ell)\, dx_\ell}.$$

The natural choice is to select the model for which $\mathbb{P}(\mathcal{M} \,|\, d)$ is highest, that is, the discrete MAP estimate. It is clear that the conditional mean $\mathbb{E}(\mathcal{M} \,|\, d)$ does not make any sense here.

Again, the most feasible approach to carry out model selection is usually via MCMC. The posterior probabilities of the models are then obtained simply from the dwell times on the models along the chain. The construction of a feasible reversible jump transition kernel, however, may be a tedious task. For general accounts on Bayesian model selection and validation, see for example [61, 154].

12 Review of Bayesian treatises on inverse heat transfer problems

While there is a relatively small body of existing work in Bayesian analyses of problems in inverse heat transfer, the literature already contains several quite sophisticated applications and is well worth consulting [133, 156, 157, 158, 36, 159, 35, 160, 161, 162, 163, 101, 42].

In [133], examples are given of building surrogates to a complex forward map based on measured data, integrated with analytical and numerical modelling, for an application in thermal design of wearable computers, see Section 10.3. These surrogates use ‘kriging’, that is, fitting of Gaussian process models, to allow fast simulation of the forward map for subsequent inference, including parameter estimation in inverse heat transfer, see Section 7.2 for some details on Gaussian processes.

Modelling and recovery of thermal history is considered in [156]. The work considers both simulated and real measurements in developing and implementing MCMC based inference. Prior modelling, data modelling, and interpretation of estimates, are prominent in this well executed example of Bayesian methodology applied to a scientific question.

In the series of articles [157, 158, 36, 159], recent methods in computational Bayesian inference and spatial statistics are applied to a range of problems in inverse heat transfer, with emphasis on inverse heat conduction problems. These articles provide a thorough development of formulation of the likelihood, prior distribution modelling including hierarchical modelling and the use of hyperpriors, formation of the posterior distribution, and design of efficient MCMC samplers. In [157], application to the IHCP of boundary heat flux identification in one and two dimensions is given, with demonstration of the methods using simulated data. The accuracy of point estimates with quantification of uncertainties is considered in [158], along with model reduction, with application to detecting heat sources. Extensive discussion of hierarchical methods in Bayesian modelling is presented in [36], to model uncertainties in sensor location as well as the primary unknowns of boundary flux and heat sources. Notable is the use of conjugate priors over hyperparameters. Posterior mean and uncertainty estimates are presented that demonstrate the value of the Bayesian approach that is taken. In particular, the value of using ‘conjugate’ hyperpriors is established. In Sections 3.1 and 8.4 we have highlighted the use of ‘reference’ hyperpriors that can have additional desirable properties, see [57]. In these articles, the unknowns are represented using the same FEM basis used for numerical simulation with GMRF priors for unknowns, as discussed in section 7.2.

Steady-state heat conduction problems are considered in [161, 162] using a hierarchical Bayesian approach, as in [36], to explore the posterior distribution over unknown thermal conductivity and unobserved boundary temperature. Estimates with uncertainties and marginal distributions over hyperparameters are presented, clearly showing the method and results. Efficiency of the MCMC sampling is further developed in [162], through the use of two surrogate models achieving model reduction. The (reciprocal of) the IACT is plotted as a function of proposal window as a diagnostic for efficiency, see Section 10.5.

Parameter estimation of the heat conduction in orthotropic media is developed in [134], using an eigenfunction expansion for the forward map, and MCMC sampling to perform the estimation of parameters and uncertainties. Results using simulated data show that estimates of heat conduction are accurate, and provide very good estimation of true temperatures, effectively implementing model-based smoothing of data.

Transient data has been considered in [164, 165] in which thermophysical parameters and the boundary heat flux have been estimated simultaneously. Simultaneous estimation of thermal conductivity and heat capacity is presented in [160], based on transient data at boundaries and modelling the unknowns as Markov random fields. In that paper, the reconstruction is fully tomographic and no layer type structures are forced. A state estimation approach to estimate inhomogeneous thermal diffusivity was carried out in [166] related to ultrasound induced heating and magnetic resonance based thermal mapping.

An interesting study of forward modelling and hierarchical prior modelling for the estimation of specific heat is presented in [101]. Bayesian inference from measured data is effected by MCMC sampling. A detailed comparison of least squares, regularized least squares, and Bayesian inference is presented in [37] in the practical case where experimental conditions are somewhat uncertain and nuisance variables are required in the prior modelling. The value of applying Bayesian inference is clearly displayed.

13 Discussion

We have discussed the Bayesian framework for modelling inverse problems as an alternative to the common (regularized) data fitting approach. In particular, we have pointed out some of the potential pitfalls in least squares, minimum norm and the related regularization methods.

There are several inverse problems that are moderately stable in the sense that the measurements carry out adequate information on the unknowns, and with which regularization approaches work perfectly well. But this may also be due to the parametrization of projection of the unknowns, such as when an unknown function is approximated, for example, as a second order polynomial. The interpretation of the results then lies heavily on the associated assumption that then unknown really is that smooth. But even if this were true, the posterior distribution of the coefficients and/or the physical coefficient might have long tails. The answers to probabilistic questions would especially in such cases call for sampling methods.

Another central topic in this review is the importance of the feasibility of the models. This calls for, at least at some stage of the overal process, the construction of computationally accurate forward models, and the modelling of *all* uncertainties. If the models are not feasible, it is not sensible to embark on accurate inference. Of course, the same applies also to regularization approaches: if the models are not feasible and consistent with the real measurements, the computation of the minimizers of regularized functionals is in vain. On the other hand, if we have a feasible posterior model but no practically feasible and efficient methods for exploration, we have gained very little.

In the Bayesian framework, all uncertainties are modelled using distributions. However, it is important to remember that all models are just *models*, that need to be tested, validated, and improved if necessary. The cost of not including uncertainties in models can be overly optimistic estimates from simulated data, and the impossibility of recovering true parameter values from real data.

One of the most appealing properties of the Bayesian framework is that the modelling of the measurement process is completely separate from the modelling of all uncertainties. Another motivation for adopting the Bayesian framework is that probabilistic questions can be answered. On the other hand, Bayesian inference is almost invariably much more tedious than implementing regularization methods. The deciding factor is what kind of questions one has to answer, and what are the specifications for the accuracy of the answers. If one only needs to have a visual clue on what an unknown might look like, especially the coarse structure, the authors of this paper would probably use some regularization method.

Although Bayesian computation comes with a heavy computational cost, its ability to handle uncertainty in primary unknowns and other model parameters makes it the best option for quantitatively accurate solution of inverse problems. With increased computing power and improved algorithms, this cost is becoming less and less expensive. And, again, to give answers to questions that have been posed in terms of probabilities, the Bayesian framework is the only alternative.

We have also considered such topics that cannot be casted in the regularization framework, such as model selection and variable dimensional problems. It may be possible that old problems can be cast in such a way that existing measurements and measurement systems may provide information that has been hitherto considered inaccessible.

References

[1] J. Hadamard. *Lectures on Cauchy's Problem in Linear Differential Equations.* Yale University Press, New Haven, CT, 1923.

[2] S.L. Campbell and C.D. Meyer. *Generalized Inverses of Linear Transformations.* Dover, 1991.

[3] C. W. Groetsch. *Inverse Problems in the Mathematical Sciences.* Vieweg, 1993.

[4] M. N. Özişik and H. R. B. Orlande. *Inverse Heat Transfer.* Taylor & Francis, 2000.

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

[6] C.F. Gauss. *Theoria combinationis observationum erroribus minimis obnoxiae. 2 Teile und Supplement.* Dieterich, 1823-1826.

[7] P. S. Laplace. Mêmoire sur la probabilité des causes par les évènemens. *Mem. Acad. Roy. Sci.*, vol. 6, pp. 621–656, 1774.

[8] A. N. Tikhonov and V. Y. Arsenin. *Solution of ill-posed problems.* Winston & Sons, Washington, DC, 1977.

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

[10] D. L. Phillips. A technique for the numerical solution of certain integral equations of the first kind. *J Assoc Comput Mach*, vol. 9, pp. 84–97, 1962.

[11] S. Twomey. On the numerical solution of Fredholm integral equations of the first kind by the inversion of the linear system produced by quadrature. *J Assoc Comput Mach*, vol. 10, pp. 97–101, 1963.

[12] H.W. Engl, M. Hanke, and A. Neubauer. *Regularization of Inverse Problems.* Kluwer, 2000.

[13] M. Hanke. *Conjugate Gradient Type Methods for Ill-posed Problems.* Longman Scientific & Technical, 1995.

[14] M. Hanke and P. C. Hansen. Regularization methods for large-scale problems. *Surveys math Ind*, vol. 3, pp. 253–315, 1993.

[15] P. C. Hansen. *Rank-Deficient and Discrete Ill-Posed Problems. Numerical Aspects of Linear Inversion.* SIAM, 1998.

[16] K. Y. Dorofeev and A. G. Yagola. The method of extending compacts and a posteriori error estimates for nonlinear ill-posed problems. *Inverse Ill-Posed Probl*, vol. 12, pp. 627–636, 2004.

[17] J.P. Kaipio and E. Somersalo. *Statistical and Computational Inverse Problems.* Springer-Verlag, 2005.

[18] A. Nissinen, L. M. Heikkinen, and J. P. Kaipio. Approximation errors in electrical impedance tomography - an experimental study. *Meas Sci Technol*, vol. 19, pp. doi:10.1088/0957–0233/19/1/015501, 2008.

[19] A. Nissinen, L. M. Heikkinen, V. Kolehmainen, and J. P. Kaipio. Compensation of errors due to discretization, domain truncation and unknown contact impedances in electrical impedance tomography. *Meas Sci Technol*, vol. 20, pp. doi:10.1088/0957–0233/20/10/105504, 2009.

[20] M. Vauhkonen, J. P. Kaipio, E. Somersalo, and P. A. Karjalainen. Electrical impedance tomography with basis constraints. *Inverse Probl*, vol. 13, pp. 523–530, 1997.

[21] M. Vauhkonen, D. Vadász, P. A. Karjalainen, E. Somersalo, and J.P. Kaipio. Tikhonov regularization and prior information in electrical impedance tomography. *IEEE Trans Med Imaging*, vol. 17, pp. 285–293, 1998.

[22] H. Jeffreys. *Scientific Inference.* Cambridge University Press, December 1931.

[23] S. F. Gull and G. J. Daniell. Image reconstruction from incomplete and noisy data. *Nature*, vol. 272, pp. 686–690, April 1978.

[24] A. Tarantola and B. Valette. Inverse problems = quest for information. *J Geophys*, vol. , pp. 159–170, 1982.

[25] E. T. Jaynes. Prior information and ambiguity in inverse problems. In D. W. McLaughlin, editor, *Inverse Problems: SIAM-AMS Proceedings*, volume 14. American Mathematical Society, 1984.

[26] J. O. Berger. *Statistical Decision Theory and Bayesian Analysis.* Springer, 1980.

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

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

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

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

[31] I. Stakgold. *Boundary Value Problems of Mathematical Physics (Classics in Applied Mathematics, 29) 2 volume set.* Society for Industrial Mathematics, January 1987.

[32] C. Catteneo. A form of heat conduction equation which eliminates the paradox of instantaneous propagation. *Compte Rendus*, vol. 247, pp. 431–433, 1958.

[33] P. Vernotte. Some possible complications in the phenomenon of thermal conduction. *Compte Rendus*, vol. 252, pp. 2190–2191, 1961.

[34] C. Fox, G. Nicholls, and M. Palm. Efficient solution of boundary-value problems for image reconstruction via sampling. *Journal of Electronic Imaging*, vol. 9, pp. 251–259, July 2000.

[35] H. R. B. Orlande, G. S Dulikravich, and M. J. Colaço. Bayesian estimation of the thermal conducitivity components of orthotropic solids. In *Anais do V Congresso Nacional de Engenharia Mecânica*, Salvador, Bahia, Brasil, 2008.

[36] J. Wang and N. Zabaras. Hierarchical Bayesian models for inverse problems in heat conduction. *Inverse Problems*, vol. 21, pp. 183–206, 2005.

[37] A F Emery, E Valenti, and D Bardot. Using bayesian inference for parameter estimation when the system response and experimental conditions are measured with error and some variables are considered as nuisance variables. *Measurement Science and Technology*, vol. 18, pp. 19–29, 2007.

[38] C.-S. Liu. Identification of temperature-dependent thermophysical properties in a partial differential equation subject to extra final measurement data. *Numer Meth Partial Diff Eq*, vol. 23, pp. 1083–1109, 2007.

[39] R. T. Al-Khairy and Z. M. AL-Ofey. Analytical solution of the hyperbolic heat conduction equation for moving semi-infinite medium under the effect of time-dependent laser heat source. *Journal of Applied Mathematics*, vol. 2009, pp. 18, 2009.

[40] Y. A. Çengel and A. Ghajar. *Heat and mass transfer: fundamentals and applications*. McGraw-Hil, New York, 4 edition, 2010.

[41] J. R. Howard. An experimental study of heat transfer through periodically contacting surfaces. *Int. J. Heat Mass Transfer*, vol. 19, pp. 367–372, 1976.

[42] L. Yan, F. Yang, and C. Fu. A Bayesian inference approach to identify a Robin coefficient in one-dimensional parabolic problems. *Journal of Computational and Applied Mathematics*, vol. 231, pp. 840 – 850, 2009.

[43] D. Lesnic, L. Elliot, and D. B. Ingham. Application of the boundary element method to inverse heat conduction problems. *Int. J. Heat Mass Transfer*, vol. 39, pp. 1503–1517, 1996.

[44] L. C. Wrobel. *The Boundary Element Method*. Wiley, Chichester, 2002.

[45] S. Langdon. A boundary integral equation method for the heat equation. In A. Struthers and B. Bertram, editors, *Proc. 5th Int. Conf. on Integral Methods in Science and Engineering (IMSE98)*, pp. 211–216, Boca Raton FL, 2000. CRC.

[46] B. T. Johansson and D. Lesnic. A method of fundamental solutions for transient heat conduction in layered materials. *Engineering Analysis with Boundary Elements*, vol. 33, July 2009.

[47] A. J. Davies. *The finite element method: a first approach*. Clarendon Press: Oxford, New York, 1980.

[48] A. Ern and J.L. Guermond. *Theory and practice of finite elements*. Springer, 2004.

[49] J. C. Butcher. *Numerical Methods for Ordinary Differential Equations*. Wiley, 2008.

[50] U. Ascher, R. Mattheij, and R. Russell. *Numerical Solution of Boundary Value Problems for Ordinary Differential Equations*. SIAM, Philadelphia, PA, 1995.

[51] J. Tausch. A fast method for solving the heat equation by layer potentials. *J. Comput. Phys.*, vol. 224, pp. 956–969, 2007.

[52] S. C. Brenner and L. R. Scott. *The Mathematical Theory of Finite Element Methods*. Springer, 1994.

[53] P. Solin, K. Segeth, and I. Dolezel. *Higher Order Finite Element Methods*. Chapmann and Hall, Suffolk, 2004.

[54] E. Sinclair. *Volatility Trading.* Wiley, 2008.

[55] M. Abramowitz and I. A. Stegun. *Handbook of Mathematical Functions.* Dover books on mathematics. Dover Publications, 1 edition, 1965.

[56] M. D. Springer. *The Algebra of Random Variables.* John Wiley & Sons Inc, New York, April 1979.

[57] J. O. Berger, V. De Oliveira, and B. Sanso. Objective Bayesian analysis of spatially correlated data. *J. American Statistical Assoc.*, vol. 96, pp. 1361–1374, 2001.

[58] J. A. Hartigan. *Bayes theory.* Springer Series in Statistics. Springer-Verlag, New York, 1983.

[59] C. P. Robert. *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation.* Springer Verlag, New York, 2 edition, 2001.

[60] G. K. Nicholls and C. Fox. Prior modelling and posterior sampling in impedance imaging. In A. Mohammad-Djafari, editor, *Proc SPIE, vol. 3459, " Bayesian Inference for Inverse Problems"*, pp. 116–127. SPIE, P.O.Box 10, Bellinghan WA 98227-0010, USA, 1998.

[61] C.P. Robert and G. Casella. *Monte Carlo Statistical Methods.* Springer, New York, 2004.

[62] P. M. Lee. *Bayesian Statistics: An Introduction.* Wiley, Chichester, 3 edition, 2009.

[63] B. P. Carlin and T. A. Louis. *Bayesian Methods for Data Analysis.* Chapman & Hall/CRC, Boca Raton, FL, 3 edition, 2008.

[64] A. Tarantola. *Inverse Problem Theory and Methods for Model Parameter Estimation.* SIAM, Philadelphia, 2004.

[65] D. Calvetti and E. Somersalo. *An Introduction to Bayesian Scientific Computing - Ten Lectures on Subjective Computing.* Number ISBN 978-0-387-73393-7. Springer, 2007.

[66] R. A. Horn and C. R. Johnson. *Matrix Analysis.* Cambridge University Press, 1985.

[67] W. R. Gilks, S. Richardson, and D. J. Spiegelhalter. Introducing Markov chain Monte Carlo. In W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, editors, *Markov Chain Monte Carlo in Practice*, pp. 1–19. Chapman & Hall, Suffolk, 1996.

[68] C. Kipnis and S. R. S. Varadhan. Central limit theorem for additive functionals of reversible Markov processes and applications to simple exclusions. *Comm. Math. Phys.*, vol. 104, pp. 1–19, 1986.

[69] C. J. Geyer. Practical Markov chain Monte Carlo. *Statistical Science*, vol. 7, pp. 473–483, 1992.

[70] J. S. Liu. *Monte Carlo Strategies in Scientific Computing.* Number ISBN 0-387-95230-6 in Springer Series in Statistics. Springer-Verlag, New York, 2005.

[71] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, vol. 21, pp. 1087–1092, 1953.

[72] W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, vol. 57, pp. 97–109, 1970.

[73] P. J. Green. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, vol. 82, pp. 711–732, 1995.

[74] K. E. Andersen, S. P. Brooks, and M. B. Hansen. Bayesian inversion of geoelectrical resistivity data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 65, pp. 619–642, 2003.

[75] S. P. Brooks, P. Giudici, and G. O. Roberts. Efficient construction of reversible jump Markov chain Monte Carlo proposal distributions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 65, pp. 3–39, 2003.

[76] L. Teirney. Introduction to general state-space Markov chain theory. In W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, editors, *Markov chain Monte Carlo in Practice*, pp. 59–74. Chapman & Hall, Suffolk, 1996.

[77] S. Siltanen, V. Kolehmainen, S. Järvenpää, J. P. Kaipio, P. Koistinen, M. Lassas, J. Pirttilä, and E Somersalo. Statistical inversion for medical x-ray tomography with few radiographs i: General theory. *Phys Med Biol*, vol. 48, pp. 1437–1463, 2003.

[78] V. Kolehmainen, S. Siltanen, S. Järvenpää, J. P. Kaipio, P. Koistinen, M. Lassas, J. Pirttilä, and E Somersalo. Statistical inversion for medical x-ray tomography with few radiographs i: Application to dental radiology. *Phys Med Biol*, vol. 48, pp. 1465–1480, 2003.

[79] D. Watzenig and C. Fox. A review of statistical modelling and inference for electrical capacitance tomography. *Measurement Science and Technology*, vol. 20, pp. 22pp, 2009.

[80] B. P. Carlin and S. Chib. Bayesian model choice via Markov chain Monte Carlo methods, vol. 57, pp. 473–484, 1995.

[81] J. Kaipio and E. Somersalo. Statistical inverse problems: discretization, model reduction and inverse crimes. *J Comput Appl Math*, vol. 198, pp. 493–504, 2007.

[82] S. Banerjee, B. P. Carlin, and A. E. Gelfand. *Hierachical Modeling and Analysis for Spatial Data.* Chapman & Hall/CRC, New York, 2004.

[83] D. Higdon. A primer on space-time modelling from a Bayesian perspective. In B. Finkenstadt, L. Held, and V. Isham, editors, *Statistics of Spatio-Temporal Systems*, pp. 217–279, New York, 2006. Chapman & Hall/CRC.

[84] M. A. Hurn, O. Husby, and H. Rue. Advances in Bayesian image analysis. In P. J. Green, N. Hjort, and S Richardson, editors, *Highly Structured Stochastic Systems*, pp. 302–322. Oxford: Oxford University Press, 2003.

[85] P. J. Green. MCMC in image analysis. In W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, editors, *Markov chain Monte Carlo in Practice*, pp. 381–399. Chapman & Hall, Suffolk, 1996.

[86] F. Lindgren. Reconstruction of flames. In K. V. Mardia, C. A. Gill, and R. G. Aykroyd, editors, *The art and science of Bayesian image analysis*, pp. 52–59. Leeds University Press, Leeds, 1997.

[87] K.M. Hanson, G.S. Cunningham, and R.J. McKee. Uncertainty assessment for reconstructions based on deformable geometry. *Int. J. Imaging Syst. Technol.*, vol. 8, pp. 506–512, 1997.

[88] G. Stawinski, A. Doucet, and P. Duvaut. Reversible jump Markov chain Monte Carlo for Bayesian deconvolution of point sources. In A. Mohammad-Djafari, editor, *Proc SPIE, vol. 3459, " Bayesian Inference for Inverse Problems"*, pp. 179–190. SPIE, P.O.Box 10, Bellinghan WA 98227-0010, USA, 1998.

[89] E. T. Jaynes. Prior probabilities. *IEEE Transactions on Systems Science and Cybernetics*, vol. 4, pp. 227–241, 1968.

[90] H. Rue. Fast sampling of Gaussian Markov random fields. *Journal of the Royal Statistical Society: Series B*, vol. 63, pp. 325–338, 2001.

[91] S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Trans Pattern Anal Mach Intell*, vol. 6, pp. 721–741, 1984.

[92] D. Dobson and F. Santosa. Recovery of blocky images from noisy and blurred data. *SIAM J Appl Math*, vol. , pp. 1181–1198, 1996.

[93] G. K. Nicholls. Bayesian image analysis with Markov chain Monte Carlo and coloured continuum triangulation models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 60, pp. 325–338, 1998.

[94] A. Voutilainen, F. Stratmann, and J.P. Kaipio. A non-homogeneous regularization method for estimation of narrow aerosol size distributions. *J Aerosol Sci*, vol. 31, pp. 1433–1445, 2000.

[95] A. Lehikoinen, S. Finsterle, A. Voutilainen, L. M. Heikkinen, M. Vauhkonen, and J. P. Kaipio. Approximation errors and truncation of computational domains with application to geophysical tomography. *Inverse Probl Imaging*, vol. 1, pp. 371–389, 2007.

[96] M. Lassas and S. Siltanen. Can one use total variation prior for edge preserving Bayesian inversion. *Inverse Probl*, vol. 20, pp. 1537–1564, 2004.

[97] V. Kolehmainen, S. R. Arridge, W. R. B. Lionheart, M. Vauhkonen, and J. P. Kaipio. Recovery of region boundaries of piecewise constant coefficients of an elliptic PDE from boundary data. *Inverse Probl*, vol. 15, pp. 1375–1391, 1999.

[98] D. Calvetti and E. Somersalo. Image inpainting with structural bootstrap priors. *Image Vision Comput*, vol. 24, pp. 782–793, 2006.

[99] J. O. Berger. The case for objective Bayesian analysis. *Bayesian Analysis*, vol. 1, pp. 1–17, 2004.

[100] J.P. Kaipio, A. Seppänen, E. Somersalo, and H. Haario. Posterior covariance related optimal current patterns in electrical impedance tomography. *Inverse Probl*, vol. 20, pp. 919 – 936, 2004.

[101] D Bardot and A F Emery. Combining discrepancy models with hierarchical Bayesian inference for parameter estimation of very ill posed thermal problems. *Journal of Physics: Conference Series*, vol. 135, pp. 012013 (8pp), 2008.

[102] S.R. Arridge, J.P. Kaipio, V. Kolehmainen, M. Schweiger, E. Somersalo, T. Tarvainen, and M. Vauhkonen. Approximation errors and model reduction with an application in optical diffusion tomography. *Inverse Probl*, vol. 22, pp. 175–195, 2006.

[103] V. Kolehmainen, T. Tarvainen, S. R. Arridge, and J. P. Kaipio. marginalization of uninteresting distributed parameters in inverse problems – application to optical tomography. *Int J Uncertainty Quantification*, vol. , 2009. in review.

[104] S. R. Arridge, J. P. Kaipio, V. Kolehmainen, M. Schweiger, E. Somersalo, T. Tarvainen, and M. Vauhkonen. Approximation errors and model reduction with an application in optical diffusion tomography. *Inverse Probl*, vol. 22, pp. 175–195, 2006.

[105] J. Heino and E. Somersalo. A modelling error approach for the estimation of optical absorption in the presence of anisotropies. *Phys Med Biol*, vol. 49, pp. 4785–4798, 2004.

[106] J. Heino, E. Somersalo, and J. P. Kaipio. Compensation for geometric mismodelling by anisotropies in optical tomography. *Optics Express*, vol. 13, pp. 296–308, 2005.

[107] D. Calvetti, J. P. Kaipio, and E. Somersalo. Aristotelian prior boundary conditions. *International Journal of Mathematics*, vol. 1, pp. 63–81, 2006.

[108] A. Nissinen, L. M. Heikkinen, V. Kolehmainen, and J. P. Kaipio. Compensation of modelling errors due to unknown domain boundary in electrical impedance tomography. *Meas Sci Technol*, vol. , 2009. In review.

[109] T. Tarvainen, V. Kolehmainen, A. Pulkkinen, M. Vauhkonen, M. Schweiger, S. R. Arridge, and J. P. Kaipio. Approximation error approach for compensating modelling errors between the radiative transfer equation and the diffusion approximation in diffuse optical tomography. *Inverse Probl*, vol. 26, pp. doi:10.1088/0266–5611/26/1/015005, 2010.

[110] J.M.J. Huttunen and J.P. Kaipio. Approximation error analysis in nonlinear state estimation with an application to state-space identification. *Inverse Problems*, vol. 23, pp. 2141–2157, 2007.

[111] J. P. Kaipio and E. Somersalo. Nonstationary inverse problems and state estimation. *J. Inv. Ill-Posed Problems*, vol. 7, pp. 273–282, 1999.

[112] A. Doucet, N. de Freitas, and N. Gordon. *Sequential Monte Carlo Methods in Practice*. Springer, 2001.

[113] R. E. Kalman and R. S. Bucy. New results in linear filtering and prediction theory. *Trans. ASME J Basic Engineering*, vol. 83, pp. 95–108, 1961.

[114] B.D.O. Anderson and J.B. Moore. *Optimal Filtering*. Prentice Hall, 1979.

[115] J.P. Kaipio, S. Duncan, A. Seppanen, E. Somersalo, and A. Voutilainen. State estimation. In D. Scott and H. McCann, editors, *Handbook of Process Imaging for Automatic Control*, pp. 207–235. CRC Press, 2005.

[116] N. K. Sinha and B. Kuszta. *Modeling and Identification of Dynamic Systems.* Van Nostrand Reinhold, 1983.

[117] J. Durbin and J. Koopman. *Time Series Analysis by State Space Methods.* Oxford University Press, 2001.

[118] H. Risken. *The Fokker-Planck Equation.* Springer, 1989.

[119] P. J. Brockwell and R. A. Davis. *Time Series: Theory and Methods.* Springer, 1991.

[120] A. Lehikoinen, J.M.J. Huttunen, A. Voutilainen S. Finsterle, M.B. Kowalsky, and J.P. Kaipio. A new dynamic inversion approach for hydrological process monitoring. *Water Resources Res*, vol. , 2009. In press.

[121] J. F. Bonnans, J. C. Gilbert, C. Lemarechal, and C. A. Sagastizabal. *Numerical Optimization. Theoretical and Practical Aspects.* Springer, 2003.

[122] R. H. Shumway. *Applied Statistical Time Series Analysis.* Prentice-Hall, 1988.

[123] J.M.J. Huttunen and J.P. Kaipio. Approximation errors in nostationary inverse problems. *Inverse Problem and Imaging*, vol. 1, pp. 77–93, 2007.

[124] J.M.J. Huttunen and J.P. Kaipio. Model reduction in state identification problems with an application to determination of thermal parameters. *Applied Numerical Mathematics*, vol. 59, pp. 877–890, 2009.

[125] J.M.J. Huttunen, A. Lehikoinen, J. Hämäläinen, and J.P. Kaipio. Importance filtering approach for the nonstationary approximation error method. *Inverse Problems*, vol. , 2009. in review.

[126] A.C. Antoulas. *Approximation of Large-Scale Dynamical Systems.* SIAM, 2005.

[127] J. D. Moulton, C. Fox, and D. Svyatskiy. Multilevel approximations in sample-based inversion from the Dirichlet-to-Neumann map. *J. Phys.: Conf. Ser.*, vol. 124, pp. 012035+, 2008.

[128] T. Bui-Thanh, K. Willcox, and O. Ghattas. Model reduction for large-scale systems with high-dimensional parametric input space. *SIAM Journal on Scientific Computing*, vol. 30, pp. 3270–3288, 2008.

[129] G. H. Golub and C. F. van Loan. *Matrix Computations*, volume 3rd. The Johns Hopkins University Press, Baltimore, MD, 1996.

[130] R. K. Beatson and L. Greengard. A short course on fast multipole methods. In M. Ainsworth, J. Levesley, W. Light, and M. Marletta, editors, *Wavelets, Multilevel Methods and Elliptic PDEs*, pp. 1–37. Oxford University Press, 1997.

[131] J. A. Christen and C. Fox. Markov chain Monte Carlo using an approximation. *Journal of Computational and Graphical Statistics*, vol. 14, pp. 795–810, December 2005.

[132] F. Hettlich and W. Rundell. A second degree method for nonlinear inverse problems. *SIAM Journal on Numerical Analysis*, vol. 37, pp. 587–620, 1999.

[133] N. Leoni and C.H. Amon. Bayesian surrogates for integrating numerical, analytical, and experimental data: application to inverse heat transfer in wearable computers. *Components and Packaging Technologies, IEEE Transactions on*, vol. 23, pp. 23–32, March 2000.

[134] H. R. B. Orlande, M. J. Colaço, and G. S. Dulikravich. Approximation of the likelihood function in the Bayesian technique for the solution of inverse problems. *Inverse Problems in Science and Engineering*, vol. 16, pp. 677–692, 2008.

[135] C. J. Geyer. Markov chain Monte Carlo maximum likelihood calculations. In *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface*, pp. 156–163. Interface Foundation, 1991.

[136] A. Gelman and D. B. Rubin. Inference from iterative simulation using multiple sequences. *Statistical Science*, vol. 7, pp. 457–472, 1992.

[137] A. D. Sokal. Monte Carlo methods in statistical mechanics: Foundations and new algorithms. In *Lectures at the Cargése summer school on "Functional Integration: Basics and Applications"*, 1996.

[138] G. O. Roberts. Markov chain concepts related to sampling algorithms. In W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, editors, *Markov Chain Monte Carlo in Practice*, pp. 45–57. Chapman & Hall, Suffolk, 1996.

[139] G. O. Roberts and J. S. Rosenthal. Optimal scaling for various Metropolis-Hastings algorithms. *Statistical Science*, vol. 16, pp. 351–367, 2001.

[140] J. E. Lee. Sample based inference for inverse obstacle scattering. Master's thesis, Department of Mathematics, The University of Auckland, New Zealand, 2005.

[141] C. Fox. Recent advances in inferential solutions to inverse problems. *Inverse Problems Sci. Eng.*, vol. 16, pp. 797–810, 2008.

[142] E. Marinari and G. Parisi. Simulated tempering: A new Monte Carlo scheme. *Europhysics Letters*, vol. 19, pp. 451–458, 1992.

[143] M. Palm. Monte Carlo methods in electrical conductance imaging. Master's thesis, Department of Mathematics, The University of Auckland, New Zealand, 1999.

[144] D. Higdon, H. Lee, and C. Holloman. Markov chain Monte Carlo-based approaches for inference in computationally intensive inverse problems. In J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith, and M. West, editors, *Bayesian Statistics 7*. Oxford University Press, 2003.

[145] F. Liang and W. Wong. Real parameter evolutionary Monte Carlo with applications to Bayesian mixture models. *J. Amer. Stat. Assoc.*, vol. 96, pp. 653–666, 2001.

[146] T. Cui. Bayesian inference for geothermal model calibration. Master's thesis, Department of Mathematics, The University of Auckland, New Zealand, 2005.

[147] T. Cui, C. Fox, M. O'Sullivan, and G. K. Nicholls. Using parallel MCMC sampling to calibrate a computer model of a geothermal resevoir. *manuscript*, vol. , 2010.

[148] C. Fox and G. Nicholls. Sampling conductivity images via MCMC. In K. V. Mardia, C. A. Gill, and R. G. Aykroyd, editors, *" The art and science of Bayesian image analysis ". Proceedings of the Leeds annual statistics research workshop*, pp. 91–100, Leeds, UK, 1-4 July 1997. Leeds university press.

[149] H. Haario, E. Saksman, and J. Tamminen. An adaptive Metropolis algorithm. *Bernoulli*, vol. 7, pp. 223–242, 2001.

[150] C. Andrieu and J. Thoms. A tutorial on adaptive MCMC. *Statistics and Computing*, vol. 18, pp. 343–373.

[151] Y. Bai, G. O. Roberts, and J. S. Rosenthal. On the containment condition for adaptive Markov chain Monte Carlo algorithms. Technical report, University of Toronto, 2008.

[152] G. O. Roberts and J. S. Rosenthal. Examples of adaptive MCMC. *Journal of Computational and Graphical Statistics*, vol. 18, pp. 349–367, June 2009.

[153] G.E.P. Box and G.C. Tiao. *Bayesian Inference in Statistical Analysis*. Wiley, 1992 (1973).

[154] A. Gelman, J.B. Carlin, H.S. Stern, and D.B. Rubin. *Bayesian Data Analysis*. Chapman & Hall, Suffolk, 1995.

[155] J.P. Kaipio, A. Seppänen, A. Voutilainen, and H. Haario. Optimal current patterns in nonstationary electrical impedance tomography. *Inverse Probl*, vol. 23, pp. 1201–1214, 2007.

[156] C. Ferrero and K. Gallagher. Stochastic thermal history modelling. 1. constraining heat flow histories and their uncertainty. *Marine and Petroleum Geology*, vol. 19, pp. 633 – 648, 2002.

[157] J. Wang and N. Zabaras. A Bayesian inference approach to the inverse heat conduction problem. *International Journal of Heat and Mass Transfer*, vol. 47, pp. 3927 – 3941, 2004.

[158] J. Wang and N. Zabaras. Using Bayesian statistics in the estimation of heat source in radiation. *International Journal of Heat and Mass Transfer*, vol. 48, pp. 15–29, January 2005.

[159] N. Zabaras. Inverse problems in heat transfer. In W. J. Minkowycz, E. M. Sparrow, and J. Y. Murthy, editors, *Handbook of Numerical Heat Transfer*, chapter 17, pp. 525–558. Wiley, Chichester, 2 edition, March 2006.

[160] V. Kolehmainen, J. P. Kaipio, and H. R. B. Orlande. Reconstruction of thermal conductivity and heat capacity using a tomographic approach. *International Journal of Heat and Mass Transfer*, vol. 51, pp. 1866 – 1876, 2008.

[161] B. Jin and J. Zou. A Bayesian inference approach to the ill-posed Cauchy problem of steady-state heat conduction. *International Journal for Numerical Methods in Engineering*, vol. 76, pp. 521–544, 2008.

[162] B. Jin. Fast Bayesian approach for parameter estimation. *International Journal for Numerical Methods in Engineering*, vol. 76, pp. 230–252, 2008.

[163] S. Parthasarathy and C. Balaji. Estimation of parameters in multi-mode heat transfer problems using bayesian inference - effect of noise and a priori. *International Journal of Heat and Mass Transfer*, vol. 51, pp. 2313 – 2334, 2008.

[164] C. A. A. Mota, H. R. B. Orlande, M. O. M. Carvalho, V. Kolehmainen, and J. P. Kaipio. Bayesian estimation of temperature-dependent thermophysical properties and boundary heat flux. *Heat Transfer Eng.*, vol. 31, pp. 570–580, 2010.

[165] C. A. A Mota, H. R. B. Orlande, O. J. M. Wellele, V. Kolehmainen, and J. P. Kaipio. Inverse problem of simultaneous identification of thermophysical properties and boundary heat flux. *High Temperatures - High Pressure*, vol. , 2010. In press.

[166] J. Huttunen, M. Malinen, T. Huttunen, and J.P. Kaipio. Determination of heterogenous thermal parameters using ultrasound induced heating and mr thermal mapping. *Phys Med Biol*, vol. 51, pp. 1011–1032, 2006.

Figures

Figure 1: The joint distribution (density) $\pi(x,d)$, the marginal density $\pi(x)$ that describes the prior uncertainty in x, and the conditional densities corresponding to two different observations $\pi(x\,|\,d=0.2)$ and $\pi(x\,|\,d=0.8)$.

Figure 2: The posterior distribution $\pi(c|d)$ over c for one measurement set.

Figure 3: Output traces for c from 2000 steps of a random-walk MCMC using different window sizes.

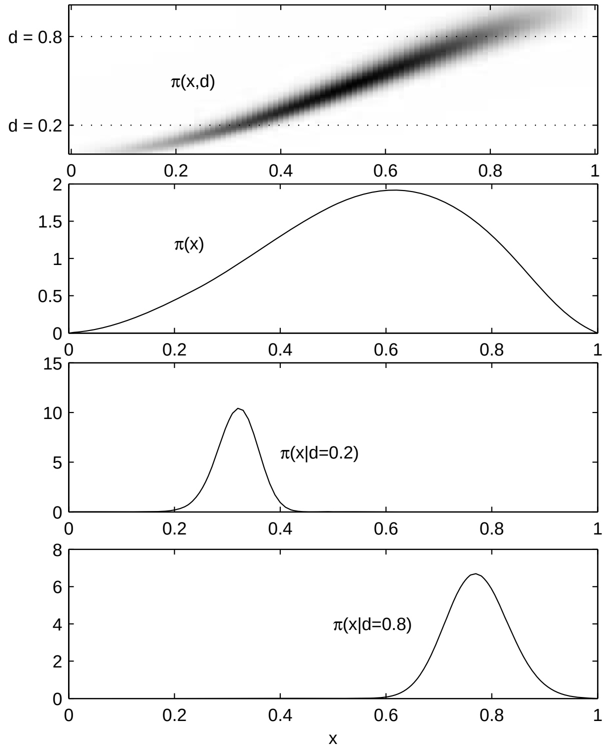

Figure 1

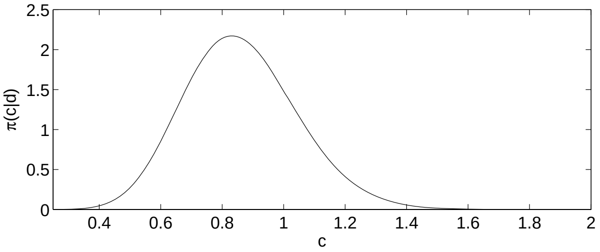

Figure 2

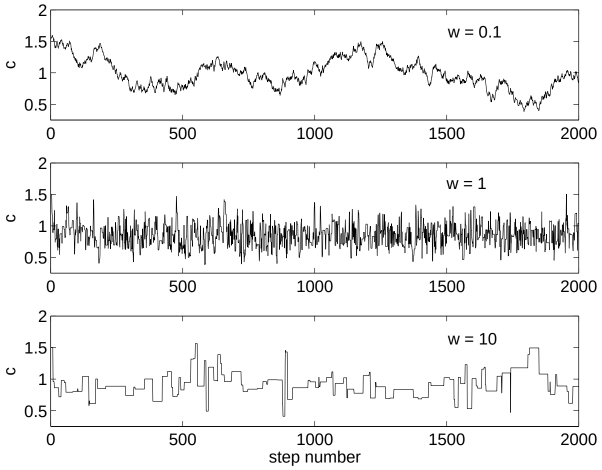

Figure 3

Biographies

Jari P. Kaipio got his MSc and PhD degrees from the University of Kuopio, Finland, where he has served as a professor of mathematical methods in physics at the Department of Physics since 1996. He has coauthored about 110 papers, mostly in inverse problems, as well as the book "Statistical and Computational Inverse Problems" with Erkki Somersalo. He is a Fellow of the Institue of Physics, UK and serves in the editorial board of Inverse Prolems, International Journal for Uncertainty Quantification, and Inverse Problems and Imaging. Dr. Kaipio is currently with the Department of Mathematics, University of Auckland but also leads the inverse problems research group in University of Eastern Finland.

Colin Fox has been Associate Professor of Physics at the University of Otago for the past two years, where he directs the program in Electronics. For seventeen years prior to that he taught Applied Mathematics at The University of Auckland, and directed the Acoustics Research Centre that undertakes research and commercial testing in building acoustics. He completed his PhD in the Radio Astronomy group at Cambridge University in 1989, on Bayesian methods for conductivity imaging. His research is in mathematical and computational methods, with long-term applications in sea ice, building acoustics, and inverse problems. His current research is on large-scale computational inference, particularly MCMC methods, with application to physical inverse problems and imaging.