31 Inverse problems

COLIN FOX, HEIKKI HAARIO
AND J. ANDRÉS CHRISTEN

31.1 Introduction

The aim of collecting data from a physical system is to gain meaningful information about the system or phenomenon of interest. However, in many situations the quantities that we wish to determine are different from the ones which we are able to measure, or have measured. Starting with the data that we have measured, the problem of trying to reconstruct the quantities that we really want is called an *inverse problem.* Loosely speaking, we say an inverse problem is where we measure an *effect* and want to determine the *cause*.

Most science and statistics is data-driven in this way, though not always called an 'inverse problem'. Here we want to discuss the features that are characteristic for the problems most typically treated under the umbrella of inverse problems. The quintessential setting is where the measurement process is a complex physical relationship, and inversion presents analytic difficulties.

In a mathematical setting, we represent the measurement process by a family of models parameterized by x, where all necessary physical parameters are contained in x, including nuisance parameters. In the language of inverse problems, simulation of the model for given x defines the *forward map* $A : x \mapsto d$ giving data d in the absence of errors. Determining and simulating the map $A : x \mapsto d$ is the *forward problem*, whereas inferring x from d is the *inverse problem*.

A mathematical model of the forward map A is usually based on some physical theory. For many physical models the mathematical analysis of the forward map is well developed; indeed, many areas of mathematics have been developed precisely to understand the structure of these mappings. Computer evaluation of $A(x)$ is typically the subject of computational science, and again, much of numerical computation has been developed to simulate these problems. For example, solving large-scale partial differential equations arising as models of physical systems drives a great deal of computational science and engineering. Thus, distinctive features of inverse problems are that the forward map is based on physics, mathematical analysis of the forward map is well developed, and evaluation of the forward map uses advanced numerical computation.

Bayesian methods are well suited to incorporating these mathematical and computational models, and for accounting for errors or uncertainties in each of these steps. In this chapter we present methodology and algorithms that are currently used for the Bayesian analysis of inverse problems.

A diverse range of researchers and practitioners work on inverse problems. There are probably as many notions of what it means to *solve* an inverse problem as there are communities of people working on inverse problems. Our notion of an inverse problem and the methods we use to solve them has been influenced by the problems in front of us, and the shared experience of trying to achieve solutions with quantified accuracy in industrial and scientific contexts. That has led

us to reformulate the inverse problem in the Bayesian (probabilistic) framework, and to employ sample-based inference to evaluate summary posterior statistics. In doing so we are outliers in the wider inverse-problems community in which deterministic 'regularization' methods (discussed in Section 31.2) are overwhelmingly the most popular. Bayesian methods have the reputation of providing the 'gold standard' amongst solutions, but also of being computationally impractical. Perhaps for those reasons, and also because regularization has a Bayesian *interpretation*, it is common to see analyses of inverse problems under the title of 'Bayesian' that amount to nothing more than regularization. While regularized solutions can be very useful, actually regularization is not a Bayesian method and our view is that scientific accuracy is served by making a linguistic distinction.

A recent development is the focus on *uncertainty quantification* (UQ) within computational models, particularly in the computational science and engineering community. We see this development as very heartening, as we are already seeing a renewed vigour in research into methods for tackling the sizable computational tasks involved in Bayesian analysis of inverse problems.

Inverse problems are often high dimensional in the sense of many unknowns and many data. When using low-level representations it is common to work with 10^3 or 10^4 unknown parameters, which we call *high* dimensional. For example, in impedance tomography about 10^3 elements are needed in an unstructured finite element mesh to ensure that the computed forward map accurately simulates the physics. A global climate model contains upwards of 10^7 unknowns, which we call *very* high dimensional. Mid-level representations, such as representations of surfaces, can effectively *reduce* the number of unknowns. In inverse problems, this reduction often leads to a *more difficult* sampling problem, that we attribute to the geometry of state space becoming more complex. Of order 10 unknowns is *low* dimensional for inverse problems, and usually arises when using parametric representations. Such problems can be very difficult when the system response is chaotic, as occurs in weather and chemical systems.

The remainder of this chapter is organized as follows. This introductory section continues with a list of representative examples of inverse problems followed by a discussion of the the key mathematical property of ill-posedness. We further discuss deterministic and regularization methods in Section 31.2. Some history of Bayesian analysis, as viewed from physics, is presented in Section 31.3. We present the framework for current methodology in Section 31.4, in the context of case studies. We also present some of the recent advances in MCMC algorithms in Section 31.4. We conclude with a glimpse of future directions in Section 31.5.

31.1.1 Examples of inverse problems

- **Compton scattering** The inelastic scattering of photons in matter can be used to probe the wave function of electrons in matter. The forward problem is to predict the angle and energy of scattered photons given the electron structure; the inverse problem is to determine electron structure from measurements of the scattering.
- **Computer axial tomography** X-rays are partially transmitted through the body, with various internal structures having different opacity to X-rays. CAT scans display a picture of that variation *in vivo*. Non-invasive measurements are made of the *total* absorption along lines through the body. Given measurement of such line integrals, how do we reconstruct the absorption as a function of position in the body?
- **Model fitting** A common task in science and engineering is to 'fit' parameters θ of a model

$$d = f(x, \theta) + \epsilon$$

for a given set of measured points $\{x_i, d_i\}_{i=1}^n$. The unknown vector θ may be low dimensional, and the fit routinely done by suitable optimization routines. But even here, with a nonlinear

model and possibly non-ideal data, only the rather recent advent of efficient Bayesian sampling algorithms has enabled us to properly analyse the reliability of parameter values and model predictions.

- **Radio-astronomical imaging** When using a multi-element interferometer as a radio telescope, the measured data is not the distribution of radio sources in the sky (called the 'sky brightness' function) but is approximately the Fourier transform of the sky brightness. It is not possible to measure the entire Fourier transform, but only to sample this transform on a collection of irregular curves in Fourier space. From such data, how is it possible to reconstruct the desired distribution of sky brightness?
- **Measuring bulk flow** Many industrial processes transport mixed phase fluids in closed pipes. Control of the process is often improved by real-time measurement of total flow of one or more of the phases. Soft-field imaging, that uses diffusive or highly scattering fields, provides a suitable non-invasive measurement that is sensitive to bulk properties. The image recovery problem is ill-posed, while the determination of bulk flow corresponds to image analysis or segmentation.
- **Geophysics** Inverse problems have always played an important role in geophysics as the interior of the Earth is not directly observable yet the surface manifestation of waves that propagate through its interior is measurable. Like many classes of inverse problems, 'inverse eigenvalue problems' were first investigated in geophysics when, in 1959, the normal modes of vibration of the Earth were first recorded and the modal frequencies and shapes were used to learn about the structure of the Earth in the large.

From this short and incomplete list, it is apparent that inverse problems occur in a myriad of settings.

31.1.2 Ill-posed and ill-conditioned

The problem of solving

$$A(x) = d \tag{31.1}$$

for x given d is called *well-posed* (in the sense of Hadamard) [42] if:

1. a solution *exists* for any data d,
2. the solution is *unique*, and
3. the inverse mapping $d \mapsto x$ is *continuous*.

Conditions 1 and 2 are equivalent to saying that the operator A is onto and one-to-one. Condition 3 is a necessary but not sufficient condition for stability of the solution.

A problem that is not well-posed is said to be *ill-posed*. So an ill-posed problem is one where an inverse does not exist because the data is outside the range of A, or the inverse is not unique because more than one value of x is mapped to the same data d, or because an arbitrarily small change in the data can cause an arbitrarily large change in the solution. Most correctly stated inverse problems turn out to be ill-posed, including all of the examples listed above.

For a well-posed problem, relative error propagation from the data to the solution is controlled by the *condition number* of A, denoted cond (A). If Δd is a variation of d and Δx the corresponding variation of x, then

$$\frac{||\Delta x||}{||x||} \leq \text{cond}\,(A)\,\frac{||\Delta d||}{||d||} \tag{31.2}$$

where (for linear forward problems) $\text{cond}\,(A) = ||A||\,\left|\left|A^{-1}\right|\right|$. When the 2-norm is used, cond (A) is just the ratio of largest to smallest singular values of A. It is possible to find a variation in data Δd for which eqn (31.2) is arbitrarily close to equality, so we usually think of eqn (31.2) with equality since the worst case behaviour will dominate the inverse.

Smaller values of cond(A) give more stable problems. If cond(A) is not too large, the problem in eqn (31.1) is said to be *well-conditioned*, otherwise the problem is said to be *ill-conditioned*. The separation between well-conditioned and ill-conditioned problems is not very sharp and depends on the computational environment. Strictly speaking, a problem that is ill-posed because it fails condition 3 must be infinite dimensional—otherwise the ratio $||\Delta x||/||\Delta d||$ is bounded. However, for ill-conditioned problems the ratio can become very large and we refer to such problems as (discrete) ill-posed problems [22].

The classical example of an ill-posed problem is a Fredholm integral equation of the first kind

$$\int_a^b k\,(t,s)\,x\,(s)\;\mathrm{d}s = d\,(t)\,, \qquad a \leq t \leq b \tag{31.3}$$

with a square integrable, or Hilbert–Schmidt, kernel k. If the solution x is perturbed by $\Delta x\,(s) = \epsilon\,\sin(2\pi ps)$, ϵ a constant, and $\Delta d\,(t)$ is the corresponding perturbation of $d\,(t)$, it follows from the Riemann–Lebesgue lemma that $\Delta d \to 0$ as $p \to \infty$. Hence, the ratio $||\Delta x||/||\Delta d||$ can become arbitrarily large by choosing the frequency p large enough, showing that eqn (31.3) is an ill-posed problem because it fails condition 3. In particular, this calculation shows that inverses of Hilbert–Schmidt integral equations are extremely sensitive to high-frequency perturbations.

Hilbert–Schmidt operators are examples of *compact* operators [45] that commonly arise in inverse problems. Since the inverse of a compact operator cannot be continuous (in standard topologies), all such inverse problems are ill-posed. Many forward problems, especially those that probe an object by the propagation of energy, are also *smoothing* operators. That is, the energy fields throughout the domain have a higher order of differentiability than the imposed excitation. It follows that the singular values of the forward map are summable to some power [3], again ensuring that the inverse is unbounded and the inverse problem is ill-posed. These considerations also explain why best-fit and maximum likelihood estimates are unreliable.

The properties of compact and smoothing both imply that the forward map is arbitrarily well approximated by a *finite-dimensional* operator, even though the spaces for parameters and data could be arbitrarily high dimensional. This means that, in the presence of uncertainty, the physical measurement process conveys only a finite amount of information about the unknowns, even when many more data are measured. Commonly the effective (local) rank can be of the order of 10 to 100. Then the physically possible data lies on a manifold of much lower dimension than data space. This explains the extreme sensitivity that inverse problems display to measurement error or model error, since measurement error will easily put data out of the range of the forward map, while modelling error will mean that the range of the model does not coincide with the physical process.

31.2 Deterministic approaches

The deterministic inverse problem is to invert the function A to obtain unknowns x as a function of data d. Mathematical studies in inverse problems typically focus on the idealized inverse problem in which *all* data is measured, and are concerned with invertability of the forward map and to what degree the inverse problem is ill-posed.

In the absence of an inverse, a solution that achieves a *best fit* to data can be computed as

$$\hat{x}_0 = \arg\min_x C(x), \qquad \text{where} \qquad C(x) = ||d - Ax||^2$$

is the *data misfit* functional, in this case the square of the norm of the residual. When A is invertible the minimum misfit is $C\left(\hat{x}_0\right) = 0$ for $\hat{x}_0 = A^{-1}d$.

However, choosing $\hat{x}$ that minimizes $C(x)$ almost always gives a poor solution. In the presence of noise, finding the (possibly non-unique) minimum of C leads to amplification of the noise because of the ill-posedness. Instead, deterministic studies often regard the data as defining a *feasible set* of solutions for which $C(x) \leq C_{\mathrm{m}}$ where C_{m} depends on the 'level' of the noise.

The primary difficulty in deterministic solutions to ill-posed inverse problems is due to small singular values of the linearized forward map. Actually, the situation is a little worse in practice since the forward map A never models the measurement process precisely. If we consider measurement error e and model error ΔA and the simple observation model $d = (A + \Delta A)x + e$ then direct inversion may be written symbolically as

$$\hat{x} = \frac{d}{A} = x + \frac{\Delta A x + e}{A}$$

Using the bases of singular vectors makes this formula precise, and shows that the direct inverse will be dominated by model error and measurement noise in the directions of singular vectors of A corresponding to small singular values.

31.2.1 Regularization methods

The most common resolution in the deterministic setting is to formulate and apply a *regular* operator that approximates the singular inverse operator A^{-1}. That is most commonly performed using the *method of regularization* introduced by Tikhonov [42], via the variational statement

$$\hat{x}_\lambda = \arg\min_x \left\{C(x) + \lambda^2 R(x)\right\} \tag{31.4}$$

Here $R(\cdot)$ is a *regularizing functional* that represents our aversion to a particular solution, with larger values being larger aversion, and λ is the *regularizing parameter*.

There are many ways of arriving at this variational form. One way is to think of minimizing the regularizing functional $R(x)$ over the set of solutions satisfying $C(x) = C_{\mathrm{m}}$, for some C_{m}. Introducing the Lagrange multiplier $1/\lambda^2$ gives the form in eqn (31.4).

The most common regularizing functional is *Tikhonov regularization*

$$R(x) = ||x||_2^2$$

Sometimes, there is a preference for solutions which are close to some *default solution* x_∞ which can be accommodated by choosing

$$R(x) = ||x - x_\infty||^2 \tag{31.5}$$

More generally, it may not be the norm of $x - x_\infty$ which needs to be small, but some linear operator acting on this difference. Introducing the operator L for this purpose, we can set

$$R(x) = ||L(x - x_\infty)||^2 = (x - x_\infty)^{\mathrm{T}} L^{\mathrm{T}} L (x - x_\infty) \tag{31.6}$$

In discrete problems the matrix L is of size $p \times n$ where $p \leq n$. Typically, L is a banded matrix approximation to the $(n-p)$th derivative. For example, when data and unknowns are one-dimensional functions discretized with interval h, approximations to the first and second derivatives are given by the matrices

$$L_1 = \frac{1}{h}\begin{pmatrix} -1 & 1 & & \\ & -1 & 1 & \\ & & \ddots & \ddots \\ & & & -1 \quad 1 \end{pmatrix} \quad \text{and} \quad L_2 = \frac{1}{h^2}\begin{pmatrix} 1 & -2 & 1 & & \\ & 1 & -2 & 1 & \\ & & \ddots & \ddots & \ddots \\ & & & 1 \quad -2 & 1 \end{pmatrix}$$

Use of the second derivative, also called *Laplacian regularization*, penalizes curvature in the solution and is commonly used when making contour maps.

In other cases, it may be appropriate to minimize some combination of the derivatives such as

$$R(x) = \alpha_0 ||x - x_\infty||^2 + \sum_{k=1}^{q} \alpha_k \left|\left|L_k (x - x_\infty)\right|\right|^2$$

where L_k is a matrix which approximates the kth derivative, and α_k are non-negative constants. Such a quantity is the square of a *Sobolev norm* that may also be written in the form of eqn (31.6).

Equation (31.4) provides a family of solutions parameterized by the regularization parameter λ. If λ is very large, the data misfit term $C(x)$ is negligible compared to $R(x)$ with $\lim_{\lambda\to\infty} \hat{x}_\lambda = x_\infty$. We effectively ignore the data (and any noise on the data) and minimize the solution seminorm by choosing the default solution. On the other hand, if λ is small, the weighting placed on the solution seminorm is small and the data misfit at the solution becomes more important. If λ is reduced to zero, the solution reduces to the least squares case.

When A is linear and the regularizing functional has the quadratic form in eqn (31.6), a solution to eqn (31.4) may readily be found by solving

$$\left(A^{\mathrm{T}}A + \lambda^2 L^{\mathrm{T}}L\right)\hat{x}_\lambda = \lambda^2 L^{\mathrm{T}}Lx_\infty + A^{\mathrm{T}}d \tag{31.7}$$

Computing the regularized solution is thus reduced to solving a (large) system of simultaneous equations with a symmetric positive definite coefficient matrix, for which there are many efficient algorithms. In stationary time-series problems sequential solutions may sometimes be implemented by repeated action of a linear operator, or *filter*. Examples are the Wiener filter and the Kalman filter.

The regularization functionals we have discussed are norms or *seminorms* on the space of solutions, as is typically the data misfit functional. There are many other regularizing functionals in common use, many designed to overcome the observation that regularization can over smooth solutions, especially at transitions in images. For example, *total variation* regularization is often used to encourage 'blocky' images [22]. Other norms are also used such as the o-norm that penalizes the number of non-zero components and hence prefers sparse solutions.

31.2.1.1 Truncated singular value decomposition

A linear operator A with rank r has the singular value decomposition (SVD)

$$A = \sum_{l=1}^{r} \sigma_l u_l^{\mathrm{t}} v_l \tag{31.8}$$

for some bases of left and right singular vectors $\{u_l\}$ and $\{v_l\}$, respectively, and singular values σ_l. The truncated SVD method is based on the observation that the components of the solution for singular vectors associated with the larger singular values of A are well determined by the data, whereas the components corresponding to smaller singular values are not. When the singular values up to $k \leq n$ are deemed to be significant the *truncated SVD* solution is

$$x'_k = \sum_{l=1}^{k} \left(\frac{u_l^{\mathrm{T}} d}{\sigma_l} \right) v_l \tag{31.9}$$

The integer k takes the role of regularizing parameter.

31.2.1.2 *Filter factors*

For the case of linear forward maps, the filter factor representation displays the solutions to the regularization problem for all values of λ in a convenient form. Here we analyse Tikhonov regularization since the SVD in eqn (31.8) suffices. Equivalent results for more general L are available using the generalized SVD.

Writing $d_l = u_l^{\mathrm{T}} d$, $\hat{x}_l = v_l^{\mathrm{T}} \hat{x}$, and $x_{\infty l} = v_l^{\mathrm{T}} x_\infty$, i.e. resolving each vector into the bases of singular vectors, the regularized solution can be written

$$\hat{x}_l = \begin{cases} \dfrac{\sigma_l^2}{\lambda^2 + \sigma_l^2} \left(\dfrac{d_l}{\sigma_l} \right) + \dfrac{\lambda^2}{\lambda^2 + \sigma_l^2} x_{\infty l} & \text{for } l = 1, 2, \ldots, r, \\ x_{\infty l} & \text{for } l = r+1, \ldots, n. \end{cases} \tag{31.10}$$

The terms d_l/σ_l and $x_{\infty l}$ give the solution coefficient in the extreme cases of no regularization ($\lambda = 0$) and no data ($\lambda = \infty$), respectively. The coefficients of these terms are the *filter factors*.

Notice how the filter factors sum to one, and the first filter factor smoothly decreases to zero as the singular values gets smaller, or as λ increases. The value of λ sets the boundary between 'small' and 'large' singular values. In contrast, the filter factors for the truncated SVD method are equal to unity for those singular values which are deemed to be non-negligible ($l \leq k$) and to zero for those singular values which are negligible ($l > k$). That sharp cutoff typically leads to ringing[36] in solutions. Thus, Tikhonov regularization may be viewed as a type of *windowing* as employed in signal processing.

31.2.1.3 *Choosing the regularization parameter*

We have seen that λ sets the balance between minimizing the residual norm $||d - Ax||$ and minimizing the solution seminorm $||L(x - x_\infty)||$. There is no single rule for selecting λ that works in all cases. Perhaps the most convenient graphical tool is the *L-curve* [22], that is a parametric plot of log of the solution seminorm versus log of the data misfit. One of the simplest methods is the *Morozov discrepancy principle* that sets λ so that the data misfit equals the measurement error 'level'. Another method is *generalized cross validation* (GCV) for selecting the parameter in ridge regression [15], which is equivalent to regularized inversion.

[36] More formally known as Gibbs' phenomenon.

31.3 A subjective history of subjective probability

For the many physicists and astronomers who were applying Bayesian analysis to inverse problems in the 1980s, the history of Bayesian methods is synonymous with the development of probabilistic methods in the physical sciences. This viewpoint is supported by many key components in Bayesian methodology being developed in response to problems arising in physics, including the Metropolis algorithm. This section presents a history of Bayesian methods as imbibed by one of us (CF) while studying inverse problems amongst *Bayesian physicists*.[37,38]

The name of Bayes was attached to Bayes' theorem by Poincaré around 1886, in his own work on probability. Bayes never wrote Bayes' theorem in the modern form. He did, however, give a method for finding *inverse probability* while solving an unfinished problem stated by Bernoulli. That method was reasoned by lengthy arguments and appeared in a paper published in 1763, after Bayes' death in 1760.

The first clear statement and use of Bayes' theorem was given by Laplace in almost his first published work in 1774. Laplace rediscovered Bayes' principle in greater clarity and generality, and then for the next 40 years applied it to scientific and civic problems. Laplace published in 1812 his two-volume treatise *Théorie Analytique des Probabilités* in which the analytical techniques for Bayesian calculations were developed. The second volume contains Laplace's definition of probability, Bayes' theorem, remarks on moral and mathematical hope (or expectation), a discussion of the method of least squares, Buffon's needle problem, and inverse probability. Later editions also contain supplements which consider applications in physics and astronomy. Laplace was mainly concerned with overdetermined problems (many observations and few unknowns) and solely used the *principle of insufficient reason*[39] to determine prior probabilities.

Laplace's calculus of probability was soon applied to explaining physical phenomena. The physicist James Clerk Maxwell said in 1850 [32],

> the true logic for this world is the calculus of Probabilities, which takes account of the magnitude of the probability which is, or ought to be, in a reasonable man's mind.

Even though Maxwell was only 19 years old at the time, he was already a formidable scientist and these principles remained in Maxwell's later work. In his kinetic theory of gases Maxwell determined the distribution over molecular velocities, effectively determining a prior probability distribution by 'pure thought' [27]. Experimental verification promoted the Maxwell distribution to the status of physical *law*, founding the subject of statistical physics.

However, among those looking to develop a theory of uncertain events the concept of probability as representing *a state of knowledge* was rejected, from about 1850, and replaced by the notion that probability must refer to *frequency in a random experiment*. Largely that rejection took place when it was realized that the notion of equiprobable, encapsulated in Laplace's principle of insufficient reason, gave results that depended on the parameterization chosen, and since Laplace had based his notion of *probable* on the more fundamental notion of *equiprobable* the whole theory was rejected. Bertrand constructed his paradox in 1889, as a transformation of Buffon's needle problem, to demonstrate the difficulties.

By the beginning of the twentieth century, application of Bayes' theorem was severely criticized with a growing tendency to avoid its application [9]. The new statistics[40] was connected with

[37] This term was apparently coined by Brian Ripley as a pejorative.

[38] A more extensive early history can be found in the first section of [28].

[39] Renamed the *principle of indifference* by Keynes [32].

[40] Interestingly, Cramér referred to Bayesian methods as 'classical' in 1945.

the theory of *fiducial probabilities* due to R. A. Fisher and the theory of *confidence intervals* due to J. Neyman. These methods became so dominant that for half a century from 1930 a student of statistics could easily not know that any other conception had existed. In that period, von Mises said that Bertrand's paradox did not even belong to the field of probability, apparently unaware of the Boltzmann (including Maxwell) distributions in physics that resolve problems of the same type.

In the 1930s, Harold Jeffreys found himself unconvinced by Fisher's arguments and rediscovered Laplace's rationale while working on 'extracting signals from noise' in geophysics. In 1939 he published his *Theory of Probability* in which he extended Bayesian inference, explaining the theory much more clearly than did Laplace. In the 1948 edition Jeffreys gave a much more general *invariance theory* for determining ignorance priors, which remains of importance today in the form of *reference priors*.

For many physicists the question of whether one can or cannot use Bayes' theorem to quantify uncertainty was answered by the physicist Richard T. Cox in 1946 and 1961 [7]. Instead of asking whether or not Laplace gave us the right 'calculus of inductive reasoning', he raised the question of what such a calculus must look like. Supposing that degrees of plausibility are to be represented by real numbers, he found the functional conditions that such a calculus be consistent and showed that the general solution uniquely determines the product and sum rules for probability to within a change of variables. An immediate consequence is Bayes' theorem. This does not answer the question of how to assign probabilities, but it does determine how they must be manipulated once assigned.

The reappearance of Bayesian methods in the physical sciences from about 1970 can in many cases be traced to the physicist Edwin T. Jaynes who, from the 1960s to 1980s, championed Bayesian methods as an inductive extension of deductive logic. While looking to unify statistical physics and Shannon's new theory of communication, he observed that methods that were experimentally verified in statistical physics appeared to be derided in statistics, and set about formalizing the basis of those methods. This led Jaynes to formulate the *maximum entropy* principle for prior distributions [28], as an extension of Jeffreys' uninformative prior. Jaynes also adapted the group invariance methods, that are standard in physics for deriving the mathematical form of physical laws, to the method of *transformation groups* for determining prior probabilities. Notably, this resolved Bertrand's 'paradox', showing that it is actually well posed [27]. Jaynes had rephrased Laplace's *indifference between events* to an *indifference between problems*. An anonymous poet celebrated this contribution in the lines:

> So, are you faced with problems you can barely understand?
> Do you have to make decisions, though the facts are not in hand?
> Perhaps you'd like to win a game you don't know how to play.
> Just apply your lack of knowledge in a systematic way.

By the 1980s, a number of groups in physics and astronomy saw Bayesian analysis as the correct route to resolving inverse problems in the presence of 'incomplete and noisy data' [18]. The advanced state of computational optimization allowed Bayesian MAP estimates to be calculated in large-scale problems, with some notoriety being achieved by the *maximum entropy method* (MEM). The practical properties and limitations of MEM were pointed out by a number of statisticians, most influentially in [12]. In the same period, inverse problems became a 'topic' in statistics, though analysis was limited to regularization *estimators* [38], or Bayesian analyses that used an artificial likelihood conditioned on a regularized solution.

The renewed appreciation of MCMC following the publication of Gelfand and Smith in 1990 influenced those applying Bayesian methods to inverse problems, with the first substantive analyses of inverse problems using MCMC appearing in 1997 [14, 37]. The analysis in [34] of a realistic problem in geophysics also appeared in that year using a Metropolis algorithm, apparently (though

somewhat implausibly) unaware of Gelfand and Smith, or Hastings' improvement. That work followed the direction set by Albert Tarantola in formulating inverse problems in a Bayesian framework [41]. The title *Inverse Problems = Quest for Information*, alone, of the 1982 paper by Tarantola and Valette had motivated many in the inverse problems community to explore Bayesian methods.

For Bayesian statisticians the early impact of MCMC was summed up by Peter Clifford when he wrote in 1993,

> from now on we can compare our data with the model we actually want to use rather than with a model which has some mathematical convenient form.

The situation for Bayesian physicists was somewhat different since they were already using physically realistic models (at least for the forward map) but lacked the computational tools for unhindered exploration of the posterior distribution. MCMC provided that tool, though the computational challenges were formidable. Exposure to spatial statistics brought the mid-level and high-level representations [26], that don't fit into a regularization framework, with a clear route for inference having been charted by Grenander and Miller [17].

31.4 Current Bayesian methodology

The methods of regularization and truncation in Section 31.2 provide valid algorithms to tame ill-posed computational problems. They also come close to the Bayesian approach in the sense that a regularization can be interpreted as equivalent to setting prior knowledge—or guess—to some characteristics of the solution. The estimate then will be a compromise produced by the regularization and the measurement data. But a crucial component of the Bayesian approach is still missing: how to produce a proper analysis of the certainty, or rather uncertainty, of the estimates? How much, indeed, can we trust the predictions given by our models, often simulating complex physical systems? Here, we believe, is the main contribution that present day Bayesian Monte Carlo algorithms are able to provide.

In this section we discuss our computational approaches to the statistical aspects of inverse problems, as well as the spirit in which we see ourselves as 'Bayesians'. The discussion is largely influenced by the applied projects from our own experience, and so inevitably is subjective again.

All available data contains measurement errors, so the estimated unknowns are more or less uncertain. A natural question then arises: if measurement noise corrupting the data follows some statistics, what is the *distribution* of the possible solutions after the estimation procedure? Bayesian thinking explicitly allows for the unknown vector x to be interpreted as a *random variable* with a distribution of its own. In addition, the approach typically emphasizes the use of *prior knowledge* in the estimation process, even subjective. As we all know, and we alluded to in our 'history', these questions have been the focus of a longstanding dispute between the two opposing views:

- Frequentists argue that analysis should be driven by the data as much as possible, and that attaching a distribution to a parameter based on one's subjective belief should not be a part of valid statistical analysis. Moreover, parameters indeed are constants without distributions of their own.
- Bayesians argue that treating solutions as random variables is actually more realistic and, by considering different choices for distributions, Bayesian analysis is perfectly valid. Moreover, scientific research most often contains strong *hidden* prior information, such as the choice of model used to explain the phenomena under study.

A practically oriented researcher might find the dispute somewhat academic. In a real modelling project, are we really so concerned about the 'true' interpretation of parameters? In any case we all certainly should be interested in the *reliability* of model predictions. Naturally, the estimates for unknowns should be physically plausible. We have experience in geophysics applications where it is necessary that estimates show a sub-surface structure that is believable to a geologist, before the predictions will be trusted.

But as the solution is estimated from noisy data, some uncertainty always remains, whether we interpret the 'truth' as fixed or random. So, it is essential, in any case, to realize that estimation problems do not have a unique solution. A numerical optimizer may find a true global minimum for a given least squares function with fixed data values. However, a multitude of different solutions may fit the data 'equally well', when we take into account the noise in the measurements. The practical essence of the Bayesian approach, in our experience, is to to find *all* those possible solutions, as well as the respective model predictions, as probability distributions. An added value is also the interpretation of those probabilities in a clear formal perspective (see e.g. [25]), that permit not only useful engineering solutions but valid 'scientific' answers as well.

For many of us, the Bayesian approach is almost synonymous with the use of MCMC methods. The advantages of using MCMC for solving inverse problems are various: full characterization of (non-Gaussian) posterior distributions is possible. We have full freedom in implementing prior information. Even modelling errors can be taken into account in a flexible way. Moreover, we are less likely to get trapped in local minimums than when employing optimization methods to get MAP estimates.

31.4.1 Mathematical formulation

The framework for Bayesian analysis of inverse problems is straightforward in concept; one formulates the likelihood function by modelling the measurement process and errors, specifies a prior distribution over unknowns, and then performs posterior inference (commonly by MCMC).

A general stochastic model for the measurement process is

$$d = G(x, v) \tag{31.11}$$

where x represents the deterministic unknowns, typically physical constants, and v is a random variable accounting for variability between 'identical' experiments. In practice the separation between 'deterministic' and 'random' is a modelling choice, since all effects may be modelled as random. We find that better results are given by modelling as many deterministic processes as possible. However, modelling practicalities often demand that some residual deterministic processes are treated as random.

In the *state space approach*, eqn (31.11) is the *observation equation* in a problem that does not vary with time. The time-varying problem [29, 44] is commonly treated as inference for a hidden Markov model [5], for which sequential Monte Carlo methods are applicable.

In the simplest formulation the stochastic part is attributed to measurement error that is additive and independent of x so that

$$G(x, v) = A(x) + v$$

where $v \sim \pi_{\mathrm{n}}(\cdot)$ comes from the *noise* distribution. Then, when the forward map is treated as certain, the distribution over data conditioned on x is is given by

$$\pi(d|x) = \pi_{\mathrm{n}}(d - A(x)) \tag{31.12}$$

The likelihood function for given data d is the same function considered as a function of the unknown variables x. Hence, formulating the likelihood function requires modelling the forward map as well as the distribution over measurement errors. Evaluation of the likelihood function requires simulation of the forward map, and hence is typically computationally expensive.

Given measurements d, the focus for inference, at least in parameter estimation, is the posterior distribution given by Bayes' theorem

$$\pi(x|d) = \frac{\pi(d|x)\pi(x)}{\pi(d)} \propto \pi(d|x)\pi(x) \tag{31.13}$$

where $\pi(x)$ is the prior distribution and $\pi(d)$ is often called the evidence. Note that we take the usual (and sometimes dangerous) liberty with notation where each density function is implicitly distinguished by the type of argument it takes.

31.4.2 Models for model error

All models are wrong, and particularly so in inverse problems. We are not aware of any inverse problem where the measurement error is greater than the model uncertainty. Perhaps this is because measurements may be made more accurately whereas more accurate physical modelling requires conceptual advances.

It is useful to distinguish between the physical process, the mathematical model and the computational model, that we denote A_{p}, A_{m} and A_{c}, respectively. Kennedy and O'Hagan [31] introduced a model for inadequacy in computer models, writing

$$A_{\mathrm{p}}(x) = A_{\mathrm{c}}(x) + D(x)$$

where the *model inadequacy* $D(x)$ was modelled nonparametrically as a Gaussian process (GP), as was A_{c}.[41] This approach would be familiar in machine learning. While a nonparametric model for model inadequacy seems very sensible, the use of Gaussian process models is somewhat unsatisfactory for inverse problems. For example, formulating a GP is prohibitive in high dimensions. Instead, modelling D by a Gaussian distribution is feasible, as we will see. Also, building a GP surrogate to the forward map is problematic since the complex input/output structure is effectively only captured in the mean process, but that amounts to tabulating input/output pairs, which is prohibitive. More successful is using a reduced order (computational) model (ROM) A_{c}^* of the computational forward map A_{c} that approximately captures that structure with a cheap computation.

The use of ROMs is almost mandatory in large-scale inverse problems, to reduce computational cost of the forward map. There are many schemes for building ROMs, such as local linearization, coarse numerical discretization, or low-order expansions. A systematic approach can be found in [1].

A ROM necessarily introduces a model error that we can analyse as

$$A_{\mathrm{c}}(x) = A_{\mathrm{c}}^*(x) + B(x) \tag{31.14}$$

We call $B(x) = A_{\mathrm{c}}(x) - A_{\mathrm{c}}^*(x)$ the *model reduction error*.

The approximation error model (AEM) of Kaipio and Somersalo [30] has proved effective in mitigating the effects of model reduction error. They modelled the model reduction error as being

[41] A taxonomy for the *arguments* of these functions, that fits well in the inverse problem context, was given by Campbell and McKay in the discussion of [31].

independent of the model parameters and normally distributed. Then the observation process is reduced to

$$d = A_c^*(x) + B + v \tag{31.15}$$

where $B \sim N(\mu_B, \Sigma_B)$, when we assume that the accurate computational model is correct, i.e. $A_p = A_c$, as in [30]. However, it is interesting to note that if the model inadequacy D is also taken to be Gaussian, then eqn (31.15) still holds without the assumption $A_p = A_c$, with the distribution over B also accounting for bias and uncertainty in the mathematical model.

31.4.3 Prior information

Most often, we do not really want to specify a non-trivial prior distribution for the solution. We may just know that the solution components must have some bounded and positive values, leading to *uninformative* or *flat* priors. Naturally, one must remember that 'flatness' depends on the parameterization of the model. For instance, a flat prior for the conductivity σ is non-flat for the resistivity $1/\sigma$, while a model could be equally written in terms of either parameterization.[42]

A practical guide for parameterization is simply to try to write a model that is easily identified by available data. For a given parameterization then, we may just set a box of 'simple bounds', just lower and upper bounds, to constrain the solutions. The analysis is now fully driven by data, supposing that the posterior distribution of the parameters is well inside the given bounds.

If, on the other hand, the posterior does not stay inside any reasonable bounds, we must observe that the available data is *not sufficient to identify the parameters*. This is an important conclusion, and not too unusual! We can then consider a few options:

- *Design of Experiments*. If non-identifiability of parameters is due to lack of data, an obvious remedy is to design new experiments to gain more informative measurements. Several classical linearization-based methods exist. Bayesian analysis and MCMC sampling provide a comprehensive way to design simulation based experiments for, e.g. situations where the classical criteria can not be implemented due to a singular information matrix [35].
- *Model reductions*. Often, however, the non-identifiability is an inherent feature of the forward map and no practically measurable data (or just reparameterization) is able to correct the situation. This occurs when parts of a physics-based model are unobservable due to, e.g. different scales in time or space: fast equilibria of certain parts of chemical kinetics, or negligible diffusion due to small catalyst particles are typical examples. An alternative option for fixing priors for unidentified parameters is then to simplify the model, and thus reduce the list of parameters to be identified. Again, MCMC sampling gives an algorithmic tool here. Instead of 'political decisions' on how to reduce the model, we may create parameter posterior distributions by MCMC to see which model parameters remain unidentified, and reduce the model accordingly. The reduction process itself may require special tools, such as the singular perturbation methods (see [19] for an example).

Typically, Bayes' theorem is seen as the way of putting together a *fixed* prior, data, and model. We may observe that the approaches suggested above rather employ Bayesian sampling techniques as flexible, algorithmic tools for *model development*, that may guide all the relevant steps of modelling: not only the analysis of model parameter identifiability, but also the design of measurements as well as testing different versions of the model for the phenomenon under study. Only if such measures

[42] The Jeffreys' prior for a *scale parameter* works here, that is uniform in $\log(\sigma)$.

are not available, one should carefully seek true prior information to be included as the prior distribution in the estimation process.

Unfortunately, sometimes one is forced to consider prior information in detail. In problems with more than a few unknowns, such as inverse problems, simply setting 'flat' priors over each coordinate direction can result in the prior being highly 'informative' for posterior statistics of interest. We first saw this effect pointed out in the context of estimating occupancy time, or *span*, of archaeological sites from radiocarbon dating of artifacts [36], where a uniform prior over the date of each artifact leads to a strong bias towards larger estimates of span, to the point where a short span is effectively ruled out. We have had to correct for this effect when using electrical capacitance tomography (ECT) to make quantitatively accurate measures of the cross-sectional area of water inclusions in oil pipe lines [44]. Interestingly, in the presence of uncertainties, correcting the prior to give quantitatively accurate estimates of area produces bias in estimates of the boundary length, and reminds us that information is a *relative* concept; uninformative with respect to one question is typically informative with respect to another.

31.4.4 Exploration by sampling

The Metropolis–Hastings (MH) algorithm is the basis of nearly all sampling algorithms that we currently use. This algorithm was originally developed for applications in statistical physics, and was later generalized to allow general proposal distributions [23], and then allowing transitions in state space with differing dimension [16]. Even though we do not always use variable-dimension models, we prefer this Metropolis–Hastings–Green (MHG) 'reversible jump' formulation of MH as it greatly simplifies calculation of acceptance probabilities for the subspace moves that are frequently employed in inverse problems. One step of MHG dynamics can be written as:

Algorithm 1 (MHG)

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. With probability

$$\alpha(x, x') = \min\left(1, \frac{\pi(x'|d)q(\gamma')}{\pi(x|d)q(\gamma)} \left|\frac{\partial(x', \gamma')}{\partial(x, \gamma)}\right|\right) \tag{31.16}$$

accept the proposed state by setting $x_{n+1} = x'$. Otherwise reject by setting $x_{n+1} = x$.

The last factor in eqn (31.16) denotes the magnitude of the Jacobian determinant of the transformation from (x, γ) to (x', γ'), as implemented in computer code for the proposal. A few details remain to be specified such as the choice of starting state, and the details of the proposal step.

The only choice one has within the MHG 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. However, the choice of proposal distribution critically affects efficiency of the resulting sampler. 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 centred about zero. In high-dimensional problems, global proposals that attempt to change all components of the state usually have vanishingly small acceptance probability, so are

not used. Since ill-posedness results in extremely high correlations, single-component proposals result in slow mixing. Hence, a multi-component update is usually required, that is problem specific.

In problems where the posterior distribution is essentially unimodal, computational cost can be minimized by starting at the MAP estimate computed by computational optimization. Indeed, the optimization step can provide useful input to the MCMC, such as a low rank approximation to the Hessian of the log of the target density when using BFGS optimization. This has been used to seed the proposal covariance in the adaptive Metropolis (AM) algorithm. For multi-modal target distributions, or when debugging code, it is often necessary to start from a randomized starting state drawn from an 'over-dispersed' distribution, though this can be very computationally expensive as the MCMC may require many iterations to find the support of the posterior distribution.

31.4.4.1 Algorithm performance

Since many steps of the MHG algorithm are typically required for convergence, and each step requires a computationally expensive evaluation of the forward map, it is important to evaluate and tune on the computational efficiency of a sampling algorithm.

A common measure of *statistical efficiency* is the integrated auto-correlation time (IACT) that measures the number of samples from the chain that have the same variance reducing power as one independent sample. It is desirable to have a small number of steps per IACT, so that estimates evaluated over the chain converge more quickly for a given number of steps.

However, statistical efficiency is not a sufficient measure of algorithmic performance in inverse problems where the CPU time taken per step can vary, such as when using a ROM. For example, in the delayed acceptance algorithms we consider later, the computational cost of a rejection step is much smaller than the computational cost of an acceptance. Hence, it is also necessary to measure the average CPU time per step.

We measure *computational efficiency* as the product of these two terms, to give the *CPU time per IACT*. This then measures the CPU time (sometimes called the wall-clock time) required to reduce the variance in estimates by the same amount that one independent sample would achieve. Clearly, small CPU time per IACT is desirable.

Unfortunately some papers showing new sampling algorithms only report statistical efficiency. We know of several such papers where an 'improved' algorithm is correctly reported as increasing statistical efficiency, but actually decreases computational efficiency, and hence would take longer to produce estimates with a given accuracy compared with the unimproved algorithm.

31.4.5 Atmospheric remote sensing and adaptive Metropolis

As an example of an ill-posed inverse problem, we discuss in some detail the recovery of ozone profiles by satellite measurements. This case study also provides an example on how practical challenges from real-life projects may give us impetus to develop new computational methods.

Remote sensing techniques are today routinely used for atmospheric research. The data processing of these instruments typically involve solving nonlinear inverse problems. GOMOS (Global Ozone Monitoring by Occultation on Stars) is one of the 10 instruments on board the European Space Agency's Envisat satellite which is targeted on studying the Earth's environment. The Envisat satellite was launched on the 1st of March in 2002 to a polar, sun-synchronous orbit at about 800 km above the Earth. It is still fully operational now in 2012. The main objective of GOMOS is to measure the atmospheric composition and especially the ozone concentration in the stratosphere and mesosphere with high vertical resolution. The GOMOS instrument was the first operational instrument that uses the stellar occultation technique to study the Earth's atmosphere. The measurement principle, demonstrated in Figure 31.1, is elegant: the stellar spectrum seen through the atmosphere is compared with the reference spectrum measured above the atmosphere. Due to

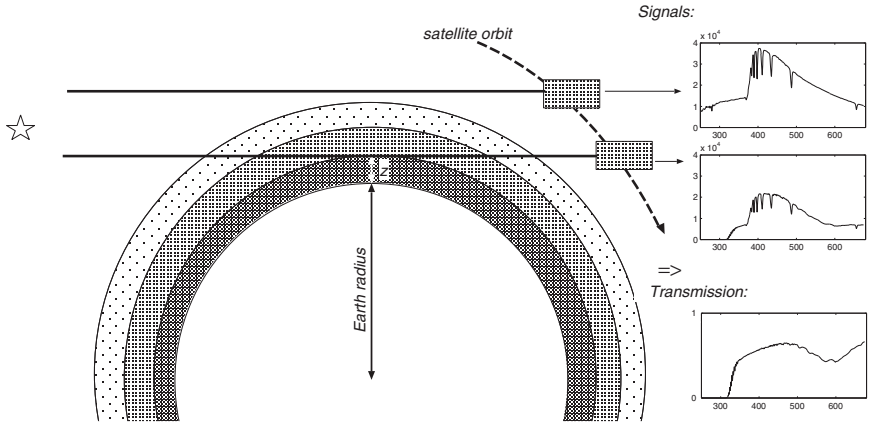

Figure 31.1 GOMOS measurement principle. The horizontal transmission of the atmosphere at tangent altitude z is obtained by dividing the attenuated stellar spectrum with the reference spectrum measured above the atmosphere.

the absorption and scattering in the atmosphere the light measured through the atmosphere is attenuated and the attenuation is proportional to the amount of constituents in the atmosphere. The measurements are repeated at different tangential altitudes to obtain vertical profiles of the concentrations of different atmospheric constituents. The advantages of the GOMOS instrument compared to other instruments measuring ozone are the fairly good global coverage, with 300–400 occultations daily around the Earth combined with the excellent vertical resolution (sampling resolution 0.3–1.7 km). The altitude range which can be covered by GOMOS is large: 15–100 km and the brightest stars can be followed even down to 5 km. Each occultation consists of about 70–100 spectra measured at different tangential altitudes and each UV-vis spectra includes measurements at 1416 different wavelengths. Because of the multitude of stars it is important that the optimal set of stars is selected for each orbit. This optimization was included in the GOMOS mission planning.

In the GOMOS data processing constituent densities are retrieved from stellar spectra attenuated in the atmosphere. The GOMOS inverse problem can be considered as an exterior problem in tomography, but in practice it is solved locally considering only data collected from one occultation at a time. This inverse problem is as follows. By dividing the stellar spectrum measured through the atmosphere with the reference spectrum measured above the atmosphere we obtain a so-called transmission spectrum. The transmission at wavelength λ, measured along the ray path ℓ, includes a term $T^{\text{abs}}_{\lambda,\ell}$ due to absorption and scattering by atmospheric constituents and a term $T^{\text{ref}}_{\lambda,\ell}$ due to refractive attenuation and scintillations, that is, $T_{\lambda,\ell} = T^{\text{abs}}_{\lambda,\ell} T^{\text{ref}}_{\lambda,\ell}$. The dependence of the transmission on the constituent densities along the line of sight ℓ is given by Beer's law:

$$T^{\text{abs}}_{\lambda,\ell} = e^{\left[-\int_\ell \sum_{\text{gas}} \alpha^{\text{gas}}_\lambda (z(s)) \rho^{\text{gas}}(z(s)) ds\right]}$$

where $\rho^{gas}(z)$ gives the constituent density at altitude z and α denotes the cross-sections. Each atmospheric constituent has typical wavelength ranges where the constituent is active either by absorbing, scattering or emitting light. The cross-sections reflect this behaviour and their values are considered to be known from laboratory measurements. In the equation above the sum is over

different gases and the integral is taken over the ray path. The problem is ill-posed in the sense that continuous profile is retrieved from a discrete set of measurements. Therefore some additional regularization or prior information is required to make the problem well-posed and solvable. In practice this is done by discretizing the atmosphere into layers and assuming some smoothness prior, or even just constant or linearly varying density inside layers.

The measurements are modelled by

$$y_{\lambda,\ell} = T^{\text{abs}}_{\lambda,\ell} T^{\text{ref}}_{\lambda,\ell} + \epsilon_{\lambda,\ell},$$

$\epsilon_{\lambda,\ell,} \sim N(0, \sigma^2_{\lambda,\ell}), \lambda = \lambda_1, \ldots, \lambda_\Lambda, \ell = \ell_1, \ldots, \ell_M$. The likelihood function for the constituent profiles then reads as

$$P(y|\rho(z)) \propto e^{-\frac{1}{2}(T-y)C^{-1}(T-y)}$$

with $C = \text{diag}(\sigma^2_{\lambda,\ell})$ and $y = (y_{\lambda,\ell})$, $T = (T_{\lambda,\ell})$. The true statistics is Poisson, but can be safely treated as Gaussian. The inverse problem is to estimate the constituent profiles $\rho(z) = (\rho^{\text{gas}}(z))$, gas $= 1, \ldots, n_{\text{gas}}$.

In the operational data processing of GOMOS the problem is divided into two parts. The separation is possible if the measurement noise is independent between successive altitudes and the temperature-dependent cross-sections can be sufficiently well approximated with 'representative' cross-sections (e.g. cross-sections at the temperature of the tangent point of the ray path). In the operational algorithm these simplifications are assumed and the problem is solved in two steps. The *spectral inversion* is given by

$$T^{\text{abs}}_{\lambda,\ell} = \exp\left[-\sum_{\text{gas}} \alpha^{\text{gas}}_{\lambda,\ell} N^{\text{gas}}_{\ell}\right], \ \lambda = \lambda_1, \ldots, \lambda_\Lambda,$$

which is solved for the horizontally integrated line-of-sight densities N^{gas}_{ℓ}. The *vertical inversion*

$$N^{\text{gas}}_{\ell} = \int_{\ell} \rho^{\text{gas}}(z(s))ds, \ \ell = \ell_1, \ldots, \ell_M$$

is solved for local constituent densities ρ^{gas} using the line-of-sight densities from the previous step as the data. Naturally, it is also possible to solve the problem directly in one step by inverting the local densities from the transmission data. This approach is here referred to as the one-step inversion.

The first step of the operational GOMOS data processing, the spectral inversion problem, is nonlinear, with all the usual advantages available if solved using the MCMC technique. At each line-of-sight, the dimension of the problem is small, only some five parameters (horizontally integrated line-of-sight densities of different constituents) to be retrieved. However, the estimation is done repeatedly at each altitude, about 70–100 times for each occultation. The natural way of implementing the MCMC technique is to use random walk MH algorithm. But here we meet the difficulty of tuning the proposal distribution to obtain efficient sampling. The special feature in the GOMOS data processing is that the posterior distributions of the spectral inversion vary strongly. They depend on the tangential altitude and also on the star used for the occultation. The line-of-sight densities vary typically several decades between 15 to 100 km for ozone vertical profile measured by GOMOS. When the star is dim (and hence the signal-to-noise ratio is low) the

posterior distributions become many times wider compared with the ones obtained for a bright star. In such a setup it is impossible to find any fixed proposal distribution that would work at all altitudes and for all stars. Therefore, the proposal distributions need to be optimized for each altitude and for each occultation separately. However, any offline manual tuning of the proposal distributions is also impossible to realize because of the huge number of datasets. Automatic algorithms for tuning the proposal distribution were therefore needed.

To overcome these problems of GOMOS spectral inversion problems the adaptive MCMC algorithms were originally developed, AM for the two-step algorithm and adaptive MwG (SCAM) for the one-step inversion. The advantage of these algorithms is that they make the implementation of the MCMC easy; the adaptation can be used in a fully automatic way without increasing the computational time dramatically.

The adaptation only requires a small change in the MHG algorithm. The basic adaptive Metropolis (AM) [21] version uses a Gaussian (and thus symmetric) proposal q, whose covariance is updated by the empirical covariance of the chain:

Algorithm 2 (AM)

At step n, with state $x_n = x$ and covariance C_n, determine x_{n+1} and C_{n+1} in the following way:

1. Generate a proposal $x' \sim \mathrm{N}(x, C_n)$.
2. Accept with probability

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

   setting $x_{n+1} = x'$. Otherwise reject by setting $x_{n+1} = x$.
3. Update the proposal covariance by $C_{n+1} = s_d \,\mathrm{Cov}(x_1, x_2, ..., x_{n+1})$.

The covariance here is scaled down with the parameter s_d with $1/d$ dependence on the dimension d. The adaptation naturally can be started only when there are enough different accepted samples in the chain to compute the covariance. This may be a drawback if the initial proposal is too large; see below the discussion on the DRAM version for a remedy. Also, it may be better, especially in higher-dimensional problems, to keep adapting at some fixed intervals rather than at every step. Note also that the ergodicity does not require of use the *whole* chain but an *increasing* part of it, e.g. the last half.

31.4.6 Cheap MCMC tricks

We now present several other advances to the MHG algorithm that we have developed in response to particular inverse problems. These represent the state-of-the-art for sampling in inverse problems.

31.4.6.1 Delayed rejection AM

The delayed rejection (DR) method [16] uses several proposals: when a proposed candidate point in a Metropolis–Hastings chain is rejected, a second stage move is proposed from another proposal distribution. For example, one can use downscaled versions of a 'basic' proposal, with the motive to get acceptance after rejection. Delayed rejection can be combined with AM, as done in [20]. This method (DRAM) has been shown to be efficient in many applications, see e.g. [43]. It is helpful to get the sampler moving, especially in the beginning of the MCMC run, since AM can easily

correct a proposal that is too small, but needs accepted points for the adaptation to take place. The DR step can provide such points. In a computationally demanding situation, such as the parameter tuning of a climate model, no standard ways (i.e. preliminary parameter fitting together with the Jacobian-based approximation for the covariance) of getting an initial proposal may be available. In addition, only short chains may be simulated. In such a case, DRAM typically turned out to be a reliable approach to get a reasonably well mixed chain created.

The adaptation in DRAM could be performed in various ways. We have found it enough to keep it simple: only have two proposals, compute the empirical covariance from the chains just as in AM, and keep an identical but down-scaled version of it for the second stage proposal.

31.4.6.2 Parallel adaptive chains

Parallelizing the adaptive MCMC algorithms has been studied relatively little. In [4] a parallel MCMC implementation in the context of regeneration was studied. Combining parallel computing and MCMC is inherently difficult, since MCMC is serial by nature. Running many parallel chains independent of each other may not be satisfactory, since it takes time for each single chain to find the mode(s) of the target and for the proposal to adapt. The question whether it is better to run multiple (non-adaptive) short chains or a single long chain has been considered in many studies. In the present case with extremely time-consuming calculations, this question is not relevant, since running a single long chain is simply not possible. Instead, several short chains can be run, and parallel communicating adaptive chains can speed up the mixing of the MCMC chains considerably. For this purpose, we employ a parallel chain version of the AM algorithm. To parallelize AM, we use a simple mechanism called inter-chain adaptation, recently introduced in [8]. In inter-chain adaptation one uses the samples generated by all parallel chains to perform proposal adaptation and the resulting proposal is used for all the chains. This naturally means that one has more points for adaptation and the convergence of every individual MCMC chain is expected to speed up.

The parallel chain approach is rather straightforward to implement. The only difference to running independent parallel AM samplers is that each sampler uses and updates the same joint proposal covariance. Covariance updating can be performed at any given update interval, for instance using the rank-1 covariance update formulas, see [21]. Note that also more advanced adaptation schemes, such as the DRAM and SCAM methods discussed above, can easily be combined with the inter-chain adaptation.

31.4.6.3 Early rejection

CPU can also be saved at no cost just by looking closer at the steps of calculations. Suppose the current state in the MH algorithm is x_i. Recall that MH proceeds by proposing a candidate value x' and accepting the proposed value with probability $\alpha = \min(1, \pi(x'|d)/\pi(x)|d)$. In practice, one first evaluates $\pi(x'|d)$, then simulates a uniform random number $u \sim U(0,1)$ and accepts x' if $u < \alpha$. Thus, a point will be rejected if $u > \pi(x'|d)/\pi(x|d)$.

In numerous applications the likelihood can be divided into n independent parts $\pi(d_i|x)$, $i = 1, 2, \ldots, n$. Moreover, the partial unnormalized posterior densities $\tilde{\pi}_k(x|d) = \pi(x) \prod_{i=1}^{k} \pi(d_i|x)$ may be monotonically decreasing with respect to the index k, $k = 1, 2, \ldots, n$. This is the situation, for example, if the likelihood has an exponential form $\pi(d|x) \propto \exp(-l(d|x))$, with $l(d_i|x) \geq 0$, as in the Gaussian case. In these situations, we can reject as soon as $\tilde{\pi}_k(x'|d)/\pi(x|d) < u$ for some value of k. Thus, we can speed up the sampling simply by switching the order of the calculations: generate the random number u first, evaluate the likelihood part by part, and check after each evaluation, if the proposed value will end up being rejected. Naturally, before evaluating any likelihood terms, we can check if the proposed point will be rejected based on the prior only.

The amount of calculation saved by ER depends on the problem (amount of data, properties of the model, shape of the posterior distribution) and on the tuning of the proposal. In cases where the topology of the posterior distribution is complicated (strongly nonlinear, thin 'bananas', or multimodal), the MH sampler, even if properly tuned, results in low acceptance rates and potentially large performance gains can be achieved through ER. The same is true if the initial proposal covariance is too large: many points are rejected and ER is beneficial again. We have found that this 'cheap trick' may save computational time between around 10% and 80%. In cases with well-posed Gaussian-type posteriors the benefit is lowest. However, these are the situations for which MCMC is not even needed in the first place, as the classical linearization-based Fisher information matrix approach already works quite well.

31.4.6.4 Delayed acceptance

The delayed acceptance Metropolis–Hastings [6] (DAMH) algorithm improves computational efficiency of MCMC sampling by taking advantage of approximations to the forward map that are available in many inverse problems. The approximation to the forward map is used to evaluate a computationally fast approximation $\pi_x^*(\cdot)$ to the desired target distribution $\pi(\cdot|d)$, that can depend on the current state x.

Given a proposal drawn from the distribution $q(x, y)$, DAMH first 'tests' the proposal with the approximation $\pi_x^*(y)$ to create a modified proposal distribution $q^*(x, y)$ that is used in a standard MH. DAMH gains computational efficiency by avoiding calculation of $\pi(y|d)$ for poor proposals that are rejected by $\pi_x^*(y)$. One iteration of DAMH is given by:

Algorithm 3 (DAMH)

At step n, with state $x_n = x$, determine x_{n+1} in the following way:

1. Generate a proposal y from $q(x, \cdot)$.
2. When $x \neq y$, with probability

$$\alpha(x, y) = \min\left[1, \frac{\pi_x^*(y)q(y, x)}{\pi_x^*(x)q(x, y)}\right]$$

   continue to step 3. Otherwise reject by setting $x_{n+1} = x$ and exit.
3. With probability

$$\beta(x, y) = \min\left[1, \frac{\pi(y|d)q^*(y, x)}{\pi(x|d)q^*(x, y)}\right]$$

   accept y setting $x_{n+1} = y$, where $q^*(x, y) = \alpha(x, y)q(x, y)$. Otherwise reject y setting $x_{n+1} = x$.

For a state-dependent approximation we can assume that the approximation is exact when evaluated at the current state, i.e., $\pi_x^*(x) = \pi(x|d)$. Then the second acceptance probability can be simplified to

$$\beta(x, y) = \min\left[1, \frac{\min\left\{\pi(y|d)q(y, x), \pi_y^*(x)q(x, y)\right\}}{\min\left\{\pi(x|d)q(x, y), \pi_x^*(y)q(y, x)\right\}}\right] \tag{31.17}$$

If the approximation does not depend on the current state, we write $\pi^*(\cdot)$ in place of $\pi_x^*(\cdot)$ and the second acceptance probability simplifies to

$$\beta(x,y) = \min\left[1, \frac{\pi(y|d)\pi^*(x)}{\pi(x|d)\pi^*(y)}\right] \tag{31.18}$$

which is exactly the *surrogate transition method* introduced by Liu [33].

DAMH necessarily reduces statistical efficiency, but a good approximation will produce $\beta(x,y) \approx 1$ ([6] Theorem 2) and can increase computational efficiency by up to the inverse of the acceptance ratio. Christen and Fox gave an example in electrical impedance tomography (EIT) using the local linear approximation

$$A_x^*(x + \Delta x) = A(x) + J\Delta x,$$

where J is the Jacobian of A evaluated at state x, that improved computational efficiency by a factor of 25.

31.4.6.5 *Adaptive approximation error*

One way to construct an approximation is to directly replace the forward model A by a reduced-order model (ROM) A^* in evaluating the likelihood function in eqn (31.12). With forward problems that are induced by PDEs, the most obvious approach is to use coarse meshes. These induce a global, or state-independent, approximation. However, as we will see, a substantial improvement in efficiency is achieved by using a local correction that leads to a state-dependent approximation.

Not accounting for model reduction error in eqn (31.15) can give poor results. For example, in an inverse problem in geothermal reservoir modelling [10], we found that simply using a coarse model for A^* in place of A achieved only 17% acceptance in step 3 of DAMH. The reduction in statistical efficiency, by about a factor of 5, nullified any potential gain in computational efficiency.

Kaipio and Somersalo [30] estimated the mean μ_B and covariance Σ_B of the AEM off-line by drawing M samples from the prior distribution over x and used the sample mean and covariance of $\{A(x_i) - A^*(x_i)\}_{i=1}^M$. This AEM will be accurate over the support of the prior distribution, but will not necessarily be accurate over the posterior distribution. Instead, Cui *et al.* [10, 11] constructed the AEM over the posterior distribution adaptively, within the DAMH algorithm. Using this adaptive AEM, and a local correction explained next, resulted in an increase of the second acceptance ratio from 17%, quoted above, to 95%; so the stochastically corrected approximation is effectively perfect.

When implementing a state-independent ROM within DAMH, we have found it is always advantageous to make the zeroth-order local correction

$$A_x^*(y) = A^*(y) + \left[A(x) - A^*(x)\right]$$

which has virtually no computational cost since both $A(x)$ and $A^*(x)$ have been computed when at state x. The resulting approximation $A_x^*(\cdot)$ now depends on the state x, so DAMH is required in eqn (31.17), rather than surrogate transition eqn (31.18). This corrected approximation has the property that AEM has mean of zero [11] and hence the adaptive AEM converges to a zero mean Gaussian. We find in practice that simply setting the mean to zero in the adaptive algorithm gives best results.

One step of the resulting adaptive delayed acceptance Metropolis–Hastings (ADAMH) algorithm is:

Algorithm 4 (ADAMH)

At step n, with state $x_n = x$, approximate target distribution $\pi^*_{x,n}(\cdot)$, and proposal distribution $q_n(x, \cdot)$, determine x_{n+1} and updated distributions in the following way:

1. Generate a proposal y from $q_n(x, \cdot)$.
2. When $x \neq y$, with probability
$$\alpha(x, y) = \min\left[1, \frac{\pi^*_{x,n}(y) q_n(y, x)}{\pi^*_{x,n}(x) q_n(x, y)}\right]$$
continue to step 3. Otherwise reject by setting $x_{n+1} = x$ and goto step 4.
3. With probability
$$\beta(x, y) = \min\left[1, \frac{\pi(y|d) q^*_n(y, x)}{\pi(x|d) q^*_n(x, y)}\right]$$
accept y setting $x_{n+1} = y$, where $q^*_n(x, y) = \alpha(x, y) q_n(x, y)$. Otherwise reject y setting $x_{n+1} = x$.
4. Update the AEM covariance by
$$\Sigma_{B,n+1} = \frac{1}{n}\left[(n-1)\,\Sigma_{B,n} + \left[A(x_{n+1}) - A^*_x(x_{n+1})\right]\left[A(x_{n+1}) - A^*_x(x_{n+1})\right]^T\right].$$
5. Update the proposal to $q_{n+1}(x_{n+1}, \cdot)$.

Using this algorithm, Cui *et al.* [11] increased computational efficiency by a factor of 8 in a large-scale nonlinear inverse problem in geothermal modelling with 10^4 continuous unknowns. This reduced computing time from 8 months to 1 month, which is significant. Actually, the performance of ADAMH in that example was remarkable, drawing each *independent* sample from the correct posterior distribution at a cost of only 25 evaluations of the accurate model.

We have not yet given the form of the proposal, yet the choice of proposal distribution is critical in achieving computational feasibility, as with any MH MCMC. While adaptation can remove the need for tuning of proposals, choosing the *structure* of the proposal to adapt to remains something of an art. In high dimensional inverse problems neither of the extremes of single-component proposals (e.g. SCAM) or global proposals (e.g. AM) is optimal; see e.g. [11] for a discussion on this point, and [24] for a demonstration of the failure of AM. Instead, proposing block updates over highly correlated sets of variables, as in [11], can be very effective, although requires some exploration to find a suitable blocking scheme.

31.5 Future directions

Alan Sokal introduced his lecture notes on Monte Carlo methods [40] with the warning,

> Monte Carlo is an extremely bad method; it should be used only when all alternative methods are worse.

We wholeheartedly agree, and add that in practice the situation can be desperate, when we have no decent proposal distribution. Adaptive MCMC methods are useful here by automatically tuning

proposals, but even they can never exceed the performance with an optimal proposal. However, sometimes one must sin[43] when there is no alternative route to solving a problem, and we do so nowadays routinely for large classes of models. This leaves a pressing need to improve MCMC sampling.

There are now many options for performing MCMC sampling such as the random-walk MH, hybrid Monte Carlo, proposals based on Langevin diffusions, and many others. A significant issue in inverse problems is not just to rely on algorithms that are provably convergent, but to make sensible algorithmic choices in terms of computational efficiency, and particularly how the algorithm cost scales with problem size.

We expect that lessons learned in computational optimization will be valuable for future improvements in MCMC. In that field many sophisticated algorithms have been developed such as the Krylov space methods that go by the acronyms PCG, Bi-CGSTAB, and GMRES, and the quasi-Newton methods including BFGS. These optimizers navigate high-dimensional surfaces with minimal need to evaluate a complex function, which is a requirement shared by efficient MCMC for inverse problems. There are already sampling algorithms that use these ideas. In [2] LBFGS optimization is used to construct approximate filtering in state spaces that are too high dimensional for the usual extended Kalman filtering. The same approach has been tested for ensemble filtering, and provides a way to high-dimensional MC sampling, without MCMC. The CG sampling algorithm for Gaussian distributions presented in 2001 by Schneider and Willsky, was improved in [39] and characterized for finite precision calculations. The observation that Gibbs samplers are essentially identical to stationary iterative linear solvers that are now considered very slow (see [39] for references) provides a perspective on MCMC in relation to linear solvers, and points towards fundamental improvements.

These algorithms hold the promise of drawing independent samples with the same computational cost as the optimization required for regularized solution. While that would be a dramatic improvement over the current situation, even then the reality is that sample-based inference will only become routine in engineering if the entire cost is no more than a few times the cost of optimization. That means, even with such improvements, that for the foreseeable future 'solutions' will need to be based on at most a handful of samples drawn from the posterior distribution.

If we set aside the goal of accurate estimates of errors on estimates, and set the more modest goal of improving on current practice in inverse problems, we have a chance. As argued in [13], a single sample drawn from the posterior distribution can be better than the regularized solution, in the sense of being more representative. One could then improve substantially by drawing a few samples, since that would at least give some indication of variability in solutions, while a few dozen samples would often be good enough to show the extent of posterior variability (although *which few dozen* might be difficult to determine). This is, especially, true if those few samples already are enough to verify the *negative* conclusion: that our unknown is far from being identified. We should keep in mind that in truly high-dimensional inverse problems the number of samples most likely remains far fewer than the dimension of the unknown, so any discussion on assured convergence of posterior estimates, in the usual sense, remains academic too.

References

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

[2] Auvinen, H., Bardsley, J., Haario, H. and Kauranne, T. (2010). Variational Kalman filter and an efficient implementation using limited memory BFGS. *International Journal for Numerical Methods in Fluids*, **64**(3), 314–335.

[43] John von Neumann is quoted as saying: 'anyone using Monte Carlo is in a state of sin'.

[3] Birman, M. S. and Solomyak, M. Z. (1977). Estimates of singular numbers of integral operators. *Uspekhi Mat. Nauk*, **32**(1), 17–84. Engl. transl. in: Russian Math. Surveys **32**(1977), no. 1, 15–89.

[4] Brockwell, A. (2006). Parallel Markov chain Monte Carlo simulation by pre-fetching. *Journal of Computational and Graphical Statistics*, **15**(1), 246–260.

[5] Cappé, O., Moulines, E. and Rydén, T. (2005). *Inference in Hidden Markov Models*. Springer Series in Statistics. Springer.

[6] Christen, J. A. and Fox, C. (2005). Markov chain Monte Carlo using an approximation. *Journal of Computational and Graphical Statistics*, **14**(4), 795–810.

[7] Cox, R. T. (1961). *The Algebra of Probable Inference*. Johns Hopkins.

[8] Craiu, R. V., Rosenthal, J. and Yang, C. (2009). Learn from thy neighbor: Parallel-chain and regional adaptive MCMC. *Journal of the American Statistical Association*, **104**(488), 1454–1460.

[9] Cramér, H. (1946). *Mathematical Methods of Statistics* (First US edn). Princeton University Press.

[10] Cui, T., Fox, C. and O'Sullivan, M. J. (2011). Adaptive error modelling in MCMC sampling for large scale inverse problems. Technical Report no. 687, University of Auckland, Faculty of Engineering.

[11] Cui, T., Fox, C. and O'Sullivan, M. J. (2011). Bayesian calibration of a large scale geothermal reservoir model by a new adaptive delayed acceptance Metropolis-Hastings algorithm. *Water Resources Research*, **47**. 26 pp.

[12] Donoho, D. L., Johnstone, I. M., Hoch, J. C. and Stern, A. S. (1992). Maximum entropy and the nearly black object. *Journal of the Royal Statistical Society. Series B*, **54**, 41–81.

[13] Fox, C. (2008). Recent advances in inferential solutions to inverse problems. *Inverse Problems Sci. Eng.*, **16**(6), 797–810.

[14] Fox, C. and Nicholls, G. K. (1997). Sampling conductivity images via MCMC. In *The Art and Science of Bayesian Image Analysis* (ed. K. Mardia, R. Ackroyd, and C. Gills), pp. 91–100. Leeds Annual Statistics Research Workshop: University of Leeds.

[15] Golub, G. H., Heath, M. and Wahba, G. (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, **21**(2), 215–223.

[16] Green, P. J. and Mira, A. (2001). Delayed rejection in reversible jump Metropolis-Hastings. *Biometrika*, **88**, 1035–1053.

[17] Grenander, U. and Miller, M. (1994). Representations of knowledge in complex systems. *Journal of the Royal Statistical Society. Series B*, **56**(4), 549–603.

[18] Gull, S. F. and Daniell, G. J. (1978). Image reconstruction from incomplete and noisy data. *Nature*, **272**(5655), 686–690.

[19] Haario, H., Kalachev, L. and Laine, M. (2009). Reduced models for algae growth. *Bulletin of Math. Biology*, **71**(7), 1626–1648.

[20] Haario, H., Laine, M., Mira, A. and Saksman, E. (2006). DRAM: Efficient adaptive MCMC. *Statistics and Computing*, **16**(3), 339–354.

[21] Haario, H., Saksman, E. and Tamminen, J. (2001). An adaptive Metropolis algorithm. *Bernoulli*, **7**(2), 223–242.

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

[23] Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, **57**(1), 97–109.

[24] Higdon, D., Reese, C. S., Moulton, J. D., Vrugt, J. A. and Fox, C. (2011). Posterior exploration for computationally intensive forward models. In *Handbook of Markov Chain Monte Carlo* (ed. S. Brooks, A. Gelman, G. Jones, and X.-L. Meng), pp. 401–418. Chapman & Hall/CRC.

[25] Howson, C. and Urbach, P. (2005). *Scientific Reasoning: The Bayesian Approach* (3 edn). Open Court.
[26] Hurn, M. A., Husby, O. and Rue, H. (2003). Advances in Bayesian image analysis. In *Highly Structured Stochastic Systems* (ed. P. J. Green, N. Hjort and S. Richardson), pp. 302–322. Oxford: Oxford University Press.
[27] Jaynes, E. T. (1973). The well-posed problem. *Foundations of Physics*, **3**(4), 477–492.
[28] Jaynes, E. T. (1978). Where do we stand on maximun entropy? In *Maximum Entropy Formalism* (ed. R. D. Levine and M. Tribus), p. 16. MIT Press.
[29] Kaipio, J. and Fox, C. (2011). The Bayesian framework for inverse problems in heat transfer. *Heat Transfer Engineering*, **32**(9), 718–753.
[30] Kaipio, J. and Somersalo, E. (2007). Statistical inverse problems: discretization, model reduction and inverse crimes. *J Comput Appl Math*, **198**, 493–504.
[31] Kennedy, M. C. and O'Hagan, A. (2001). Bayesian calibration of computer models (with discussion). *Journal of the Royal Statistical Society: Series B*, **63**, 425–464.
[32] Keynes, J. M. (1921). *A Treatise on Probability*. Macmillan and Co.
[33] Liu, J. S. (2005). *Monte Carlo Strategies in Scientific Computing*. Springer.
[34] Mosegaard, K., Singh, S. C., Snyder, D. and Wagner, H. (1997). Monte Carlo analysis of seismic reflections from Moho and the W-reflector. *Journal of Geophysical Research B*, **102**, 2969–2981.
[35] Müller, P. (1999). Simulation-based optimal design. In *Bayesian Statistics 6* (ed. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith), pp. 459–474. Oxford University Press. **6**, 459–474.
[36] Nicholls, G. and Jones, M. (2001). Radiocarbon dating with temporal order constraints. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, **50**, 503–521.
[37] Oliver, D. S., Cunha, L. B. and Reynolds, A. C. (1997). Markov chain Monte Carlo methods for conditioning a permeability field to pressure data. *Mathematical Geology*, **29**(1), 61–91.
[38] O'Sullivan, F. (1986). A statistical perspective on ill-posed inverse problems. *Statistical Science*, **1**(4), 502–527.
[39] Parker, A. and Fox, C. (2011). Sampling Gaussian distributions in Krylov spaces with conjugate gradients. *SIAM Journal on Scientific Computing*. In the press.
[40] Sokal, A. D. (1996). Monte Carlo methods in statistical mechanics: Foundations and new algorithms. In *Lectures at the Cargése summer school on 'Functional Integration: Basics and Applications'*.
[41] Tarantola, A. (1987). *Inverse Problem Theory: Methods for Data Fitting and Model Parameter Estimation*. Elsevier.
[42] Tikhonov, A. N. and Arsenin, V. Y. (1977). *Solutions of Ill-posed Problems*. Scripta series in mathematics. Winston.
[43] Villagran, A., Huerta, G., Jackson, C. S. and Sen, M. K. (2008). Computational methods for parameter estimation in climate models. *Bayesian Analysis*, **3**(3), 1–27.
[44] Watzenig, D. and Fox, C. (2009). A review of statistical modelling and inference for electrical capacitance tomography. *Measurement Science and Technology*, **20**(5), 22 pp.
[45] Young, N. (1998). *An Introduction to Hilbert Space*. Cambridge University Press.