Supplementary Materials for 'Physics-constrained Bayesian Inference of Magnetic Resonance Velocimetry data'

Matthew Juniper,[1] Alexandros Kontogiannis,[1]

[1]Department of Engineering, University of Cambridge, United Kingdom, CB2 1PZ; email: mpj1001@cam.ac.uk

Xxxx. Xxx. Xxx. Xxx. YYYY. AA:1–8

https://doi.org/10.1146/((please add article doi))

Copyright © YYYY by the author(s).
All rights reserved

7. BAYESIAN INFERENCE

7.1. A brief history

Inference is the act of reasoning with incomplete knowledge. We start from prior knowledge, often accumulated from years of experience, observe some new data, and refine our knowledge with the information in the data. At a subconscious level, humans are extraordinarily good at inference, particularly when data is visual or aural. We are so good that we are surprised when computers have trouble with tasks such as driving a car or spotting letters hidden behind a pattern. At a conscious level, however, our skills of inference can be poor. Fortunately, there is a solid mathematical framework that can make us as good at conscious inference as we already are at unconscious inference, and can be extended to situations that are too complicated for our raw intuition to handle. This is Bayesian inference, which neatly corresponds to and improves upon our own common sense.

Bayesian inference usually finds its most receptive audience in the physical sciences. Physical scientists instinctively propose hypotheses that they think might explain some data and, consciously or not, assign each a prior probability. They then compare the hypotheses' predictions against the observed data and, consciously or not, re-estimate the probability of each hypothesis, instinctively favouring simpler hypotheses. Indeed, the framework of Bayesian Inference maps directly onto the process of *normal science* and *paradigm shifts* described in 1962 by the philosopher of science Thomas Kuhn (Kuhn 2009). During normal science, the probability of the most likely hypothesis tends inexorably towards 1 as evidence piles up in its favour. A paradigm shift occurs when a hypothesis with prior probability tending towards 1 is overthrown by a different hypothesis that had either not previously been considered, or had prior probability close to zero. An example of this is the caloric theory of Lavoisier.

Jaynes (1984) maps out the history of formal inference from Herodotus in 500 BC to Laplace (1820), via Bernoulli (1713) and a paper on chance by Bayes (1763), which was published posthumously and can be found at Bayes et al. (1963). According to Jaynes, Laplace used Bayesian inference to identify promising astronomical problems in which the discrepancy between model prediction and experimental observation was large enough to give a high probability that there was something new to be found – in other words more likely to lead to paradigm shifts rather than normal science. Because he did not waste time on unpromising research, he made more discoveries in celestial mechanics than any of his contemporaries.

Jaynes (1984, 2003) credits Sir Harold Jeffreys (Jeffreys 1983) with the renaissance of Bayesian inference in the early twentieth century. Like Laplace, Jefferys applied Bayesian inference to several fields. However, he faced fierce resistance from frequentist statisticians, who strongly objected to the concept that probability should depend on a humans' prior knowledge. In Jaynes words: *"for thirty more years Jefferys' work was under attack from the very persons who had the most to gain by understanding it, some of whom were living and eating with him daily in St John's College, Cambridge"* (see box). Today, the Bayesian approach is well-established, particularly in the physical sciences. An entertaining and useful textbook is MacKay (2003), which is written from the physicists perspective, and in which powerful ideas are illustrated with interesting problems rather than mathematical formality. Other useful textbooks are Tarantola (2005); Sivia & Skilling (2006) and Gelman (2014). The review papers by Stuart (2010) and Ghattas & Willcox (2021), and the lecture notes by Dashti et al. (2013) provide the Bayesian framework appropriate for continuum

mechanics problems, in particular fluid mechanics, and are essential reading for a thorough understanding of the subject.

PROBABILITY AND COLLEGIALITY

I (Matthew Juniper) was an undergraduate at St John's College Cambridge and return for lunch occasionally. Soon after starting this paper, I found myself sitting opposite an old friend who had been an undergraduate with me 30 years ago and who is now a distinguished Professor of Statistical Science. He asked me what I was working on these days. "Bayesian inference into physics-based models," I replied eagerly, thinking we would now have much to talk about. "Ugh!" he replied, evidently disgusted, his hands wandering menacingly towards the heavy candlestick between us, "I suppose you have swallowed the Bayesian pill." We don't talk about work any more, although we are probably (in a Bayesian sense) still friends.

7.2. Bayesian and frequentist definitions of probability

To Bayesians, probability is an abstract concept that is used to represent a state of knowledge, or that is calculated from previously-assigned probabilities and experiments using the rules of probability theory. To frequentists, probability is a factual property of the real world that is measured from repeated experiments. Jaynes (1984, p12) ascribes the rivalry between Bayesian and frequentist statisticians to mis-communication. He illustrates this with a question posed by a frequentist, George Barnard, aimed at pointing out the absurdity of the Bayesian approach: *"How could the distribution of a parameter possibly become known from data which were taken with only one value of the parameter actually present?"* As Jaynes points out, the mis-communication arises by confusing "distribution of the parameter" with "distribution of the probability of the parameter". In the Bayesian approach, we calculate the posterior probability distribution of the parameter. In other words, by combining prior knowledge and data, we are describing our posterior knowledge about the unknown constant value that the parameter had when the data were taken. In the frequentist approach, we are deducing, from prior knowledge of the frequency distribution of the parameter over some large class of repetitions of the whole experiment, the frequency distribution that it has in the subclass of cases that yield the same data. The two approaches share the same words and often the same solution procedures but solve fundamentally different problems.

7.3. Bayesian inverse problems in infinite dimensions

In Jaynes (2003), Jeffreys (1983), and MacKay (2003), Bayesian inference is expressed through the probabilities, $p(A)$, of propositions such as '$A \equiv$ the coin landed on heads'. Jaynes (2003, §4.7) introduces compound hypotheses, which combine governing equations with model parameters in which the probability of these model parameters is expressed through the probability, $p(A)$, of propositions such as '$A \equiv$ the parameter x has the particular value x'' (Jaynes 2003, §B.1). MacKay (2003) follows the same approach. This language is descriptive and useful, but becomes problematic if the parameter can take an infinite number of values because then $p(A) = 0$. Although Jaynes (2003, p xxi and §B.6) advises against using infinite sets, more recent developments in Bayesian inference for PDEs

use the language of measure theory, which enables the rigorous formulation in this paper. This section summarizes the pertinent points of Section 2 and Section 6 of Stuart (2010).

Stuart (2010) formulates Bayesian inference for PDEs within a general abstract continuous *infinite-dimensional* function space setting, which reveals the required properties of observable models, $\mathcal{Z}(\boldsymbol{x})$, and prior pdfs, $\pi^\circ(\boldsymbol{x})$, for Bayesian inverse problems to be well-posed. These properties carry over to the finite-dimensional formulation, thus placing the formulation in this paper on firm theoretical grounds.

A *measure space* is a triplet $(\boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{F}}, \mu)$ containing the *sample space*, $\boldsymbol{\mathcal{X}}$, a *σ-algebra* of events in this space, $\boldsymbol{\mathcal{F}}$, and the *measure*, μ, of these events. We consider a subset of events $\boldsymbol{\mathcal{A}} \in \boldsymbol{\mathcal{F}}$. A discrete example helps to illustrate this: if $\boldsymbol{\mathcal{X}}$ represents a fair die roll then one subset $\boldsymbol{\mathcal{A}}_1$ is $\{1,2,3,4,5,6\}$ with $\mu = \{\frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}\}$ and another subset $\boldsymbol{\mathcal{A}}_2$ is $\{\text{even}, \text{odd}\}$ with $\mu = \{\frac{1}{2}, \frac{1}{2}\}$. If $\mu(\boldsymbol{\mathcal{X}}) = 1$ then the measure space has *full measure* and is a *probability space*.

If $\mu^\bullet$ and μ° are two measures on the same measure space then $\mu^\bullet$ is *absolutely continuous* with respect to μ° if $\mu^\circ(\boldsymbol{\mathcal{A}}) = 0$ implies $\mu^\bullet(\boldsymbol{\mathcal{A}}) = 0$. This is written $\mu^\bullet \ll \mu^\circ$ and means that a non-negative function $f(x)$ can be defined such that $\mu^\bullet(\boldsymbol{\mathcal{A}}) = \int_{\boldsymbol{\mathcal{A}}} f(x)\,\mathrm{d}\mu^\circ(x)$. This is known as the *Radon-Nikodym derivative* of $\mu^\bullet$ with respect to μ° and is written:

$$\frac{\mathrm{d}\mu^\bullet}{\mathrm{d}\mu^\circ}(x) = f(x) \tag{13.}$$

The function $f(x)$ is often called the *density* of $\mu^\bullet$ with respect to μ° and, because it is non-negative, can be written as an exponential, $f = \exp(-\Phi)$, of a potential Φ. If μ° is the Lebesgue measure in finite dimensions then $f(x)$ is the *probability density function* $\pi(x)$. In infinite dimensions, the Lebesgue measure does not exist, so probability density functions (pdfs) are not defined.

We now associate $\mu^\bullet$ with posterior knowledge, μ° with prior knowledge, and f with the discrepancy between data and model predictions, meaning that Equation 13 represents how our knowledge changes when the data arrives. Bayes rule (Equation 5a) can be re-arranged to:

$$\frac{\pi^\bullet}{\pi^\circ} \propto \pi(\boldsymbol{u}^\star|\boldsymbol{x}, \mathcal{Z}) \propto \exp(-\Phi) \quad \text{where} \quad \Phi \equiv \frac{1}{2}\|\boldsymbol{u}_1^\star - \mathcal{S}\boldsymbol{u}\|^2_{\mathcal{C}_{\boldsymbol{u}^\star}} \quad \text{and} \quad \mathcal{S}\boldsymbol{u} = \mathcal{Z}(\boldsymbol{x}) \tag{14.}$$

In the infinite dimensional setting, the equivalent expression is that the posterior measure, $\mu^\bullet$, and the prior measure, μ°, are related through the Radon-Nikodym derivative (Equation 13), in which $f(x)$ is the data likelihood:

$$\frac{\pi^\bullet}{\pi^\circ} \to \frac{\mathrm{d}\mu^\bullet}{\mathrm{d}\mu^\circ}(x) \propto \exp\big(-\Phi(\boldsymbol{x}; \boldsymbol{u}^\star)\big) \tag{15.}$$

This form generalizes to situations in which $\boldsymbol{\mathcal{X}}$ and possibly $\boldsymbol{D}$ are infinite-dimensional. Our goal now is to obtain, in the continuous infinite dimensional setting, a well-defined posterior measure $\mu^\bullet$, given a prior measure μ°, a model $\mathcal{Z}(\boldsymbol{x})$, and data, $\boldsymbol{u}^\star$. The well-posedness of this Bayesian inverse problem depends on (i) the properties of the forward problem expressed in terms of $\Phi(\boldsymbol{x}; \boldsymbol{u}^\star)$ or $\mathcal{Z}(\boldsymbol{x})$, and (ii) the properties of the prior measure μ°.

The requirements on $\Phi(\boldsymbol{x}, \boldsymbol{u}^\star)$ are that Φ has an upper bound, a lower bound, is Lipschitz continuous in $\boldsymbol{u}^\star$, and is Lipschitz continuous in $\boldsymbol{x}$ (Stuart 2010, Assumption 2.6). When the data space, $\boldsymbol{D}$, is finite dimensional and the measurement noise is Gaussian with covariance operator $\mathcal{C}_{\boldsymbol{u}^\star}$ then $\Phi(\boldsymbol{x}; \boldsymbol{u}^\star) = \frac{1}{2}\|\boldsymbol{u}^\star - \mathcal{Z}\boldsymbol{x}\|^2_{\mathcal{C}_{\boldsymbol{u}^\star}}$. Then the requirements on $\mathcal{Z}$ are that

measure space: a generalization of the concept of *volume*.

sample space, $\boldsymbol{\mathcal{X}}$: a set containing all possible values of $\boldsymbol{x}$.

σ-algebra, $\boldsymbol{\mathcal{F}}$: the set of all subsets of $\boldsymbol{\mathcal{X}}$.

measure, μ: a value assigned to each event in $\boldsymbol{\mathcal{F}}$.

probability measure: a measure for which $\mu(\boldsymbol{\mathcal{X}}) = 1$.

probability density function: $\pi(x)$ defined such that $\mu(\boldsymbol{\mathcal{A}}) = \int_{\boldsymbol{\mathcal{A}}} \pi(x)\mathrm{d}x$ in finite dimensions

Lebesgue measure: the volume of a space, e.g. length in 1D, area in 2D.

Lipschitz continuity: The first derivative is bounded but not necessarily continuous. Lipschitz continuous functions are equivalent to Sobolev space $\boldsymbol{W}^1_\infty$

$\mathcal{Z}(\boldsymbol{x})$ is bounded and is Lipschitz continuous in $\boldsymbol{x}$ (Stuart 2010, Assumption 2.7). These requirements carry over to any discretization of the forward problem.

The role of the prior probability measure, μ°, is to restrict the continuous space from which $\boldsymbol{x}$ is sampled to that on which the forward problem is well-posed. In this paper, as in Stuart (2010), we always assign prior measures to be Gaussian because their regularity properties can be established easily from their covariance operator. Fortunately, in physical problems, the most suitable prior is usually Gaussian. This is because our prior knowledge of a parameter is usually limited to an expected value and a rough idea of its uncertainty. A Gaussian is the least informative (highest entropy) prior that contains this information.

8. Small scale Problems

8.1. Viscous channel flow with unknown boundary positions

The code `ToyOne.m` is a single stand-alone file that creates figure 4. The projection operator $\mathcal{S}$ is $\mathcal{I}$. The models being considered are for flow with uniform viscosity μ in a channel with half-width D, pressure gradient p', and centreline position y_c. Four models can be compared:

$$\mathcal{Z}(\boldsymbol{x}) = u = \frac{1}{2}x_1\left(x_2^2 - (y - x_3)^2\right) \qquad \texttt{3param_D}$$

$$\mathcal{Z}(\boldsymbol{x}) = u = \frac{1}{2}x_1\left(x_2^2 - y^2\right) \qquad \texttt{2param_D}$$

$$\mathcal{Z}(\boldsymbol{x}) = u = \frac{1}{2}x_1\left(\mathrm{e}^{2x_2} - (y - x_3)^2\right) \qquad \texttt{3param_logD}$$

$$\mathcal{Z}(\boldsymbol{x}) = u = \frac{1}{2}x_1\left(\mathrm{e}^{2x_2} - y^2\right) \qquad \texttt{3param_logD}$$

where

$$x_1 \equiv \mu^{-1}(p') \qquad \text{in all models}$$

$$x_2 \equiv \begin{cases} D & \text{in } \texttt{3param_D} \text{ and } \texttt{2param_D} \\ \log(D) & \text{in } \texttt{3param_logD} \text{ and } \texttt{2param_logD} \end{cases}$$

$$x_3 \equiv \begin{cases} y_c & \text{in } \texttt{3param_D} \text{ and } \texttt{3param_logD} \\ 0 & \text{in } \texttt{2param_D} \text{ and } \texttt{2param_logD} \end{cases}$$

Comments within the code indicate which parts of Section 3 are being executed.

8.2. Viscous channel flow with three viscosity models

The code `ToyTwo.m` creates figure 5. It calls `fun_MAP.m`, which calls a set of low level functions. This code structure is more versatile than that in Section 8.1 and new analytic or numerical models can be added easily.

The models $\mathcal{Q}(\boldsymbol{x})$, their derivatives $\mathcal{G}_k$, and their projection operators, $\mathcal{S}$, are held in low-level functions labelled `fun_QdQ_`$\star$`.m`, where $\star$ is `uniform` for Newtonian viscosity, `PowerLaw` for power law viscosity, and `Carreau` for Carreau viscosity. Within `fun_QdQ_uniform.m` and `fun_QdQ_PowerLaw.m`, $\mathcal{Q}(\boldsymbol{x})$ can be calculated analytically or with a finite element method. Within `fun_QdQ_Carreau.m`, only a finite element method is possible. These functions are called by `fun_ZdZ.m`, which maps $\mathcal{Q}(\boldsymbol{x})$ from the state space, $\boldsymbol{U}$, to the data space, $\boldsymbol{D}$. This function is called by `fun_JdJ.m`, which calculates the cost function $\mathscr{J}$ (Equations 4

and 5) and its first variation, $\mathcal{D}_z \mathscr{J}$ (Equation 6). The derivatives of all the above low level functions can be tested with the Taylor Tests `TT_Q.m`, `TT_Z.m`, and `TT_J.m`. The above functions are called by `fun_MAP.m`, which evaluates Equations 8, 9, 10, and 11. Comments within `fun_MAP.m` indicate which parts of Section 3 are being executed.

9. Bayesian inference for large problems with adjoint acceleration

Section 3.4 shows how to calculate the posterior pdf for small problems in which the linearized operator, $\mathcal{G}_k$, is sufficiently small to be stored as a matrix when discretized. We now introduce a method for tackling large problems that uses adjoint variables to avoid having to store a large dense matrix. At each step, k, of the Gauss-Newton method in Section 3.4, the model is linearized around $\boldsymbol{x}_k$ such that $\mathcal{Z}(\boldsymbol{x}) \simeq \mathcal{Z}(\boldsymbol{x}_k) + \mathcal{G}_k(\boldsymbol{x} - \boldsymbol{x}_k)$. We now focus on $\mathcal{G}_k \equiv \mathcal{S}\mathcal{A}_k$, where $\mathcal{A}_k \equiv (\mathcal{D}_{\boldsymbol{x}}^{\boldsymbol{u}})_k$ is the Jacobian of the model prediction, $\boldsymbol{u}(\boldsymbol{x})$, with respect to the model parameters, $\boldsymbol{x}$. This operator can be split such that, by the chain rule, $\mathcal{A}_k = \left[(\mathcal{D}_{\boldsymbol{u}}^{\mathscr{M}})^{-1}\mathcal{D}_{\boldsymbol{x}}^{\mathscr{M}}\right]_k$, where $\mathscr{M}$ represents a large problem, and $\mathcal{D}_{\boldsymbol{u}}^{\mathscr{M}}$ is the linearization of $\mathscr{M}$ w.r.t $\boldsymbol{u}$, and $\mathcal{D}_{\boldsymbol{x}}^{\mathscr{M}}$ is the linearization of $\mathscr{M}$ w.r.t $\boldsymbol{x}$. The inverse, $(\mathcal{D}_{\boldsymbol{u}}^{\mathscr{M}})^{-1}$, is dense, meaning that $\mathcal{G}_k$ is too large to be stored in computer memory.

To avoid this storage problem, we return to the expression for $(\mathcal{D}_{\boldsymbol{x}} \mathscr{J})_k$ in Equation 8a and, dropping the prior terms temporarily for ease of presentation, we introduce an adjoint variable $\boldsymbol{v}_k$ such that:

$$
\begin{aligned}
(\mathcal{D}_{\boldsymbol{x}} \mathscr{J})_k &= -\mathcal{G}_k^* \mathcal{C}_{\boldsymbol{u}^\star}^{-1}(\boldsymbol{u}^\star - \mathcal{G}_k \boldsymbol{x}_k) = -\mathcal{G}_k^* \mathcal{C}_{\boldsymbol{u}^\star}^{-1}(\boldsymbol{u}^\star - \mathcal{S}\boldsymbol{u}_k) && \text{18a.}\\
&= -\mathcal{A}_k^* \mathcal{S}^* \mathcal{C}_{\boldsymbol{u}^\star}^{-1}(\boldsymbol{u}^\star - \mathcal{S}\boldsymbol{u}_k) && \text{18b.}\\
&= -\left[\mathcal{D}_{\boldsymbol{x}}^{\mathscr{M}}\right]_k^* \left[(\mathcal{D}_{\boldsymbol{u}}^{\mathscr{M}})^{-1}\right]_k^* \mathcal{S}^* \mathcal{C}_{\boldsymbol{u}^\star}^{-1}(\boldsymbol{u}^\star - \mathcal{S}\boldsymbol{u}_k) && \text{18c.}\\
&= -\left[\mathcal{D}_{\boldsymbol{x}}^{\mathscr{M}}\right]_k^* \boldsymbol{v}_k && \text{18d.}\\
\text{where} \quad \left[\mathcal{D}_{\boldsymbol{u}}^{\mathscr{M}}\right]_k^* \boldsymbol{v}_k &= \mathcal{S}^* \mathcal{C}_{\boldsymbol{u}^\star}^{-1}(\boldsymbol{u}^\star - \mathcal{S}\boldsymbol{u}_k) && \text{18e.}
\end{aligned}
$$

We then set $\boldsymbol{x}_k$, solve the forward problem for $\boldsymbol{u}_k$, solve the adjoint problem for $\boldsymbol{v}_k$ with Equation 6e, compute the gradient $(\mathcal{D}_{\boldsymbol{x}} \mathscr{J})_k$ with 6d, re-introduce the gradients due to the prior terms, and then calculate $\boldsymbol{x}_{k+1}$ with Equation 8b. Even though the matrix $\left[\mathcal{D}_{\boldsymbol{u}}^{\mathscr{M}}\right]_k^*$ is sparse and can be small enough to store, it does not necessarily need to be stored because $\boldsymbol{v}_k$ can be approximated from matrix-vector products.

For mildly nonlinear problems (see Section 3.5), the posterior pdf is approximately Gaussian: $\pi^\bullet(\boldsymbol{x}|\boldsymbol{u}^\star, \mathcal{Z}) \simeq \mathcal{N}(\boldsymbol{x}_k, \mathcal{C}_{\boldsymbol{x}_k})$ at the final iteration. A computationally efficient way to approximate the inverse Hessian $\mathcal{C}_{\boldsymbol{x}_k}$ is to use a quasi-Newton method (e.g BFGS) using a matrix-vector product implementation as described in Ghattas & Willcox (2021), Tarantola (2005, §6.22.8).

10. NOMENCLATURE

u	generic state variable
u	velocity
p	pressure
x	parameter
ε	data uncertainty
v	adjoint velocity
q	adjoint pressure
ρ	density
$\boldsymbol{u}$	velocity vector
$\boldsymbol{y}$	direct variables; $\boldsymbol{y} \equiv (\boldsymbol{u}, p)$
$u^\star$	experimental data
u°	prior
$u^\bullet$	posterior
$\bar{u}$	expected value / mean
u^*	adjoint
π	probability density
ρ	un-normalized probability density
p	probability of an event
Z	normalizing constant
μ	measure
$\mathcal{N}(\bar{u}, \mathcal{C})$	Gaussian distribution with expected value $\bar{u}$ and covariance $\mathcal{C}$
$\boldsymbol{U}$	state space; $u \in \boldsymbol{U}$
$\boldsymbol{D}$	data space; $u^\star \in \boldsymbol{D}$
$\boldsymbol{M}$	model space; [check diff with statespace]
$\boldsymbol{\mathcal{X}}$	parameter space; $x \in \boldsymbol{\mathcal{X}}$
$\mathcal{A}$	generic linear operator
$\mathcal{Q}$	model mapping from $\boldsymbol{\mathcal{X}}$ to $\boldsymbol{U}$
$\mathcal{S}$	projection from $\boldsymbol{U}$ to $\boldsymbol{D}$
$\mathcal{Z}$	model mapping from $\boldsymbol{\mathcal{X}}$ to $\boldsymbol{D}$; $\mathcal{Z} \equiv \mathcal{S}\mathcal{Q}$
$\mathcal{G}$	linearized $\mathcal{Z}$
$\mathcal{C}$	covariance operator
$\mathcal{H}$	Hessian
$\mathcal{D}$	derivative operator
$\mathcal{I}$	identity operator
$\mathscr{J}$	total cost functional
$\mathscr{R}$	regularization cost functional
$\mathscr{U}$	data cost functional
$\mathscr{R}_x$	prior cost functional [poss redundant]
$\mathscr{M}$	boundary value problem cost functional

LITERATURE CITED

Bayes T. 1763. An essay towards solving a problem in the doctrine of chances

Bayes T, Price R, Molina E, Canton J, Deming WE. 1963. Facsimiles of two papers by Bayes : I. An essay toward solving a problem in the doctrine of chances. New York: Hafner

Bernoulli J. 1713. Ars conjectandi. Thurneysen Brothers, OCLC 7073795

Dashti M, Law KJH, Stuart AM, Voss J. 2013. MAP estimators and their consistency in Bayesian nonparametric inverse problems. *Inverse Problems* 29(9):095017

Gelman A. 2014. Bayesian Data Analysis. Taylor and Francis

Ghattas O, Willcox K. 2021. Learning physics-based models from data: perspectives from inverse

problems and model reduction. *Acta Numerica* 30:445–554
Jaynes E. 2003. Probability Theory: The Logic of Science. Cambridge University Press
Jaynes ET. 1984. Bayesian methods: General Background, In *Maximum Entropy and Bayesian Methods in Applied Statistics*, ed. JH Justice, pp. 1–25, Calgary: Cambridge University Press
Jeffreys H. 1983. Theory of probability. Clarendon Press ; Oxford University Press, 3rd ed.
Kuhn TS. 2009. The structure of scientific revolutions. University of Chicago Press, 3rd ed.
Laplace PM. 1820. Théorie analytique des probabilités. Paris: Courgier, V., 3rd ed.
MacKay DJC. 2003. Information Theory, Inference, and Learning Algorithms. Cambridge University Press
Sivia DS, Skilling J. 2006. Data analysis : a Bayesian tutorial. Oxford University Press
Stuart AM. 2010. Inverse problems: A Bayesian perspective. *Acta Numerica* 19:451–459
Tarantola A. 2005. Inverse Problem Theory. SIAM