Approximate Bayesian Tomography

By

Thomas Connell

A thesis submitted to Macquarie University
for the degree of Master of Research
Department of Earth and Planetary Sciences
14th October 2018

MACQUARIE
University
SYDNEY·AUSTRALIA

Except where acknowledged in the customary manner, the material presented in this thesis is, to the best of my knowledge, original and has not been submitted in whole or part for a degree in any university.

Thomas Connell

Acknowledgements

I would like to acknowledge the love and support of Melanie. Without it this work would not have been possible.

I would also like to thank the continuous encouragement from my parents Sean and Heather, who always pushed me to do what I wanted. Without such freedom I may never have ended up on the path I am now.

Abstract

Geophysical *inverse problems* define Earth's structure based on experiental observations. Their solution is an invaluable line of evidence to scrutinize hypotheses about our planet. Recently, to quantify uncertainty in the experimental observations and the *forward problem*, explore trade-offs between model parameters and constrain the solution with a priori information, probabilistic methods for inverse problems have been pioneered in geophysics. While traditional linear methods are weakened under these conditions, probabilistic methods based on a Bayesian formulation are well suited. There are, however, implicit assumptions in the construction and use of the likelihood function in Bayesian inference. Here I scrutinize these assumptions and advocate for the use of likelihood-free Bayesian inference, coined Approximate Bayesian Computation, as a method which can build upon the success of probabilistic tomography. Here I show that Approximate Bayesian Computation can utilize the information available in a dataset to drive rapid convergence to low misfit models while maintaining formal statistical guarantees. Efficient algorithms of this kind are required for probabilistic methods to tackle the scale and complexity of large inverse problems about Earths deep internal structure. By freeing inference from the limiting aspects of a likelihood function, Approximate Bayesian Computation promises to expand parameter inference to problems which were previously intractable.

This thesis is a 9 month study, undertaken from January 5 - October 14 2018.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Statement of Aims

This body of work is concerned with *inverse problems* (Tarantola, 2005; Aster et al., 2013; Menke, 2012; Kaipio and Somersalo, 2006; Biegler et al., 2010; Idier, 2013), also referred to as parameter inference or parameter estimation problems (Box and Tiao, 1973; Sprott, 2008; Casella and Berger, 1993; Cox, 2007). Parameter inference based on observable data is a pursuit which transcends scientific disciplines. For geophysics, it is the most important method which allows data to be transformed from surface observations into subsurface images. The effectiveness of traditional methods of geophysical parameter inference is limited when characterizing non-linear systems with trade-offs between parameters and inherit non-uniqueness. Under these conditions probabilistic parameter inference has demonstrated utility in imaging the subsurface (Tarantola, 2005). With that in mind, this project seeks to explore the geophysical applicability of a recently developed method of probabilistic parameter inference, known as Approximate Bayesian Computation (ABC), which has emerged from applied challenges in genetics (Tavare et al., 1997; Beaumont et al., 2002; Sunnåker et al., 2013). This project was started on January 5th, 2018 and concluded with the submission of this document on the 14th of October, 2018. All codes used in this project are written by the author and available at https://github.com/tomconnell/abc-toy-problems and https://github.com/tomconnell/approximate-bayesian-tomography.

1.2 Parameter inference

Parameter inference is the definition of unknown parameters, $\boldsymbol{\theta} = \{\theta_1, ..., \theta_N\}$, for a causative model, $\mathcal{M}$, based on observed data, $\boldsymbol{y} = \{y_1, ..., y_M\}$. For geophysics, we are interested in defining the physical properties of the subsurface (e.g. seismic velocity, density) from experimental data observed at the surface (e.g. seismic signals, gravitational acceleration). This search is alternatively referred to as the *inverse problem* (Tarantola, 2005; Aster et al., 2013; Menke, 2012):

$$\text{The inverse problem: } \boldsymbol{y} \rightarrow \boldsymbol{\theta} \tag{1.1}$$

Solving the inverse problem relies on the definition of deterministic physical models which link a parameterized subsurface model to simulated data., giving rise to the *forward problem*:

$$\text{The forward problem: } \boldsymbol{\theta'} \rightarrow \boldsymbol{y'} \tag{1.2}$$

The relationship between model parameters and resulting simulated data may be highly non-linear and is represented through the operator $\boldsymbol{g}$:

$$\boldsymbol{y} = \boldsymbol{g}(\boldsymbol{\theta}) \tag{1.3}$$

1.3 Probabilistic formulation

Geophysical inverse problems come with a unique set of challenges. Often the experimental data has significant levels of uncertainty, the number of unknown model parameters may far exceed the amount of constraining data and the uncertainty in the physical model (the forward problem) may itself be large. These conditions weaken traditional methods of parameter estimation.

Bayes' theorem, equation 1.4, is the basis for a probabilistic formulation to geophysical inverse problems (Tarantola and Valette, 1982; Mosegaard and Tarantola, 1995; Sambridge and Mosegaard, 2002; Mosegaard and Tarantola, 2002; Tarantola, 2005). The Bayesian formulation can quantify the uncertainty in the data and forward problem, scrutinize model non-uniqueness and constrain the solution with information obtained independent of the observed data. In this approach, geophysical inverse problems receive a full statistical treatment of uncertainty. The Bayesian method is fundamentally about embracing uncertainty and making use of all available information to update our state of knowledge. Bayes' theorem relies on the notion of conditional probability, i.e. $p(a|b)$ is the conditional probability of a given event b has occurred. Bayes' theorem is used to define the solution to the inverse problem as the PDF of the model parameters $\boldsymbol{\theta}$ conditioned on the observed data $\boldsymbol{y}$, $p(\boldsymbol{\theta}|\boldsymbol{y})$. This PDF is referred to as the *posterior*:

$$p(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{p(\boldsymbol{\theta})p(\boldsymbol{y}|\boldsymbol{\theta})}{p(\boldsymbol{y})} \tag{1.4}$$

Equation 1.4 states that the prior distribution, $p(\boldsymbol{\theta})$, multiplied by $p(\boldsymbol{y}|\boldsymbol{\theta})$, over a normalization constant, $p(\boldsymbol{y})$, is equal to the posterior PDF, $p(\boldsymbol{\theta}|\boldsymbol{y})$.

The prior, $p(\boldsymbol{\theta})$, is a user-defined probability distribution which incorporates what is already known about the model parameters, $\boldsymbol{\theta}$, independent of the experimental data, $\boldsymbol{y}$. When trying to understand subsurface structure with surface data there may be prior knowledge. For example, the thickness of layers in a vertical stack may be known to be distributed according to a exponential probability distribution (Mosegaard and Tarantola, 1995).

Once data $\boldsymbol{y}$ is observed, $p(\boldsymbol{y}|\boldsymbol{\theta})$ can be regarded as a function of $\boldsymbol{\theta}$ rather than $\boldsymbol{y}$ (now fixed) and therefore it describes the likelihood of $\boldsymbol{\theta}$ given $\boldsymbol{y}$. If for two possible sets of model parameters $\boldsymbol{\theta_1}$ and $\boldsymbol{\theta_2}$ we have $p(\boldsymbol{y}|\boldsymbol{\theta_1}) > p(\boldsymbol{y}|\boldsymbol{\theta_2})$ then $\boldsymbol{y}$ is more likely to occur under $\boldsymbol{\theta_1}$ than $\boldsymbol{\theta_2}$. In this case, $p(\boldsymbol{y}|\boldsymbol{\theta})$ is called the *likelihood* and commonly represented as $\mathcal{L}(\boldsymbol{\theta}|\boldsymbol{y})$. The likelihood is a central part of this thesis and is expanded upon in the next section.

The normalization constant $p(\boldsymbol{y}) = \int p(\boldsymbol{\theta})p(\boldsymbol{y}|\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta}$ is difficult to calculate and generally unconsidered. Statistics about the posterior such as maximum posterior value, mean, and standard deviation

can still be computed from an unnormalized posterior density. The consequence of leaving $p(\boldsymbol{y})$ unevaluated is simply that single point values of the posterior have no probabilistic interpretation. Instead relative values must be evaluated where ratios cancel the normalization constant. Hence, Bayes' theorem is practically applied in the form:

$$p(\boldsymbol{\theta}|\boldsymbol{y}) \propto p(\boldsymbol{\theta})p(\boldsymbol{y}|\boldsymbol{\theta}) \tag{1.5}$$

The posterior distribution, $p(\boldsymbol{\theta}|\boldsymbol{y})$, is the solution to a probabilistic formulation of an inverse problem. It is the complete state of knowledge about $\boldsymbol{\theta}$ given $\boldsymbol{y}$. Complete as it incorporates both what is already known, $p(\boldsymbol{\theta})$, as well as new experimental data, $\boldsymbol{y}$, through the likelihood. The probabilistic solution offers a philosophical shift compared to traditional solutions. A probability distribution over all model parameters is recovered offering in-built quantification of uncertainty, instead of a single 'optimal' set of model parameters.

The probabilistic solution to inverse problems outlined above has been successfully applied e.g. to seismic tomography (e.g. Sambridge (1999); Shapiro and Ritzwoller (2002); Trampert et al. (2004); Khan et al. (2011)) and in joint inversions of geophysical data (e.g. Khan et al. (2007); Moorkamp et al. (2010); Bodin et al. (2012); Shen et al. (2012); Afonso et al. (2013a,b, 2016)). Yet, for non-trivial problems there is no analytical solution for the posterior. A popular approach is to approximate the posterior via stochastic sampling, such as Monte Carlo or Markov chain Monte Carlo (MCMC). For subsurface models with many unknown model parameters, such as a 3D high-resolution grid, and where each posterior sample requires computationally expensive simulation from a deterministic physical model, highly efficient numerical sampling methods are required if the posterior is to be adequately defined within the current computational limits.

1.4 The likelihood

The likelihood is used after $\boldsymbol{y}$ is available to describe the plausibility of $\boldsymbol{\theta}$. Given this role, the probability $p(\boldsymbol{y}|\boldsymbol{\theta})$ in Bayes' theorem, equation 1.4, is practically applied as a conditional probability given $\boldsymbol{y}$, for varying values of $\boldsymbol{\theta}$ (Box and Tiao, 1973, p.10). Following Fisher (1922), this is referred to as the likelihood and is commonly denoted $\mathcal{L}(\boldsymbol{\theta}|\boldsymbol{y})$. This maps out a distribution of relative plausibility over $\boldsymbol{\theta}$ in describing the observed data $\boldsymbol{y}$. The likelihood is not necessarily a PDF and hence point values do not have a probabilistic interpretation. Only through comparison of relative values does the likelihood gain meaning (Box and Tiao, 1973, p.11).

A full derivation of the likelihood is given from first assumptions Technical figure 1 to the final forms which are commonly featured in geophysical inversions Technical figure 2.

For physical sciences a given observed data point, y_i, is considered equal to simulated data from a parameterized deterministic model, $\boldsymbol{g}(\boldsymbol{\theta})$, with additive statistical uncertainty from both the

measurement process, $e_i^{\mathcal{D}}$, and the modelization process, $e_i^{\mathcal{M}}$:

$$y_i = z_i + e_i^{\mathcal{D}} \tag{1.6}$$

$$z_i = g_i(\boldsymbol{\theta}) + e_i^{\mathcal{M}} \tag{1.7}$$

$$y_i = g_i(\boldsymbol{\theta}) + e_i^{\mathcal{M}} + e_i^{\mathcal{D}} \tag{1.8}$$

For the case of estimating $\boldsymbol{\theta} = \{\theta_1, ..., \theta_N\}$ given $\boldsymbol{y} = \{y_1, ..., y_M\}$, if it is assumed that the nature of the statistical uncertainty of measurement and modelization is independent and identically distributed from a Gaussian distribution then the likelihood takes the form (c.f. Gregory (2005, p.91-92)):

$$\begin{aligned} \mathcal{L}(\boldsymbol{\theta}|\boldsymbol{y}) &= p(\boldsymbol{y}|\boldsymbol{\theta}) = p(y_1, ..., y_m|\boldsymbol{\theta}) \\ &\propto \exp\left[\sum_{i=1}^{M} \frac{-(y_i - g_i(\boldsymbol{\theta}))^2}{2\left((\sigma_i^{\mathcal{M}})^2 + (\sigma_i^{\mathcal{D}})^2\right)} \right] \end{aligned} \tag{1.9}$$

Otherwise, if the uncertainty is Gaussian distributed but correlated, covariance matrices $C_{\mathcal{D}}$ and $C_{\mathcal{M}}$ are combined to quantify the total statistical uncertainty and define a likelihood (c.f. Tarantola (2005, p.35-36)):

$$\mathcal{L}(\boldsymbol{\theta}|\boldsymbol{y}) \propto \exp\left[-\frac{1}{2}(\boldsymbol{y} - \boldsymbol{g}(\boldsymbol{\theta}))^T (C_{\mathcal{M}} + C_{\mathcal{D}})^{-1} (\boldsymbol{y} - \boldsymbol{g}(\boldsymbol{\theta})) \right] \tag{1.10}$$

If more complicated models for uncertainty are determined or assumed then access to Bayesian parameter inference may close entirely if no closed form expression for the likelihood can be derived. For example, if the data uncertainty is irreducible to a simple standard distribution and the modelization errors are systematic and mesh-dependent (not uncommon in geophysics), then making the same derivations as is done in Technical figure 2 may not be possible. The problem would have an intractable likelihood. This issue is encountered in genetics, and lead to the development of the likelihood-free methods of Approximate Bayesian Computation. For example, the coalescent (Marjoram and Tavaré, 2006) describes how gene variants are passed down from a common ancestor, where coalescent events are approximated by an exponential distribution. This model is then overprinted by mutations, described by a Poisson distribution. As the number of samples of DNA expanded, it was not possible to formulate a likelihood to describe the time to the most recent common ancestor given the DNA set.

The accurate calculation of equation 1.9 and equation 1.10 is an important aspect of running algorithms which sample the posterior distribution. Often likelihood values can be extremely small, to the point where they approach and cross the limits of floating point arithmetic (arithmetic underflow). Crossing this limit would lead to a divergence from posterior sampling due to erroneous Monte Carlo or MCMC algorithm steps. Technical figure 3 expands on this topic. It covers the form in which the calculations take in order to be numerically stable.

Technical figure 1: General likelihood derivation

Gregory (2005, p.89-90) outline constructing a likelihood, $\mathcal{L}(\boldsymbol{\theta}|\boldsymbol{y})$, from starting assumptions. Firstly, a model for the data is assumed:

$$y_i = z_i + e_i \tag{1.11}$$

where e_i is the uncertainty in the data, and z_i is the data from the deterministic forward relation:

$$z_i = g_i(\boldsymbol{\theta}) \tag{1.12}$$

Both z_i and e_i are represented by distributions:

$$p(z_i|\boldsymbol{\theta}) = f_Z(z_i) \tag{1.13}$$

$$p(e_i|\boldsymbol{\theta}) = f_E(e_i) \tag{1.14}$$

We are interested in arriving in an equation for $p(y_i|\boldsymbol{\theta})$, which will define $\mathcal{L}(\boldsymbol{\theta}|\boldsymbol{y})$. However, the relationship in equation 1.11 specifies, y_i depends on both z_i and e_i. Hence, to arrive at $p(y_i|\boldsymbol{\theta})$ we must consider the distributions of equations 1.13 & 1.14. If we first consider the joint distribution $p(y_i, z_i, e_i|\boldsymbol{\theta})$ then our likelihood can be found by integrating out, 'marginalizing', $p(z_i|\boldsymbol{\theta})$ and $p(e_i|\boldsymbol{\theta})$ to leave a distribution for $p(y_i|\boldsymbol{\theta})$:

$$p(y_i|\boldsymbol{\theta}) = \int\int \mathrm{d}z_i\ \mathrm{d}e_i\ p(y_i, z_i, e_i|\boldsymbol{\theta}) \tag{1.15}$$

The definition of conditional probability allows equation 1.15 to be rewritten as:

$$p(y_i|\boldsymbol{\theta}) = \int\int \mathrm{d}z_i\ \mathrm{d}e_i\ p(z_i, e_i|\boldsymbol{\theta})\ p(y_i|z_i, e_i, \boldsymbol{\theta}) \tag{1.16}$$

If we assume z_i and e_i are independent:

$$p(y_i|\boldsymbol{\theta}) = \int\int \mathrm{d}z_i\ \mathrm{d}e_i\ p(z_i|\boldsymbol{\theta})\ p(e_i|\boldsymbol{\theta})\ p(y_i|z_i, e_i, \boldsymbol{\theta}) \tag{1.17}$$

From the relationship in equation 1.11, $y_i - z_i - e_i = 0$, $p(y_i|z_i, e_i, \boldsymbol{\theta})$ can be reasonably represented as a dirac-delta function such that:

$$p(y_i|z_i, e_i, \boldsymbol{\theta}) = \delta(y_i - z_i - e_i) \tag{1.18}$$

This is a natural definition, as the probability density of y_i given z_i and e_i should be focused when the exact relation of equation 1.11 is met. The dirac-delta function serves this role perfectly while still exhibiting properties of a PDF, non-negative and $\int_{-\infty}^{\infty} \delta = 1$. Considering equation 1.18, equation 1.17 becomes:

$$p(y_i|\boldsymbol{\theta}) = \int\int \mathrm{d}z_i\ \mathrm{d}e_i\ p(z_i|\boldsymbol{\theta})\ p(e_i|\boldsymbol{\theta})\ \delta(y_i - z_i - e_i) \tag{1.19}$$

Adopting the form from equations 1.13 & 1.14:

$$p(y_i|\boldsymbol{\theta}) = \int \mathrm{d}z_i\ f_Z(z_i) \int \mathrm{d}e_i\ f_E(e_i)\delta(y_i - z_i - e_i) \tag{1.20}$$

The dirac-delta function in the second integral of equation 1.20 serves to isolate the value/values of $f_E(e_i)$ when $e_i = y_i - z_i$. Hence, it is equivalent to consider:

$$\int \mathrm{d}e_i\ f_E(e_i)\delta(y_i - z_i - e_i) = f_E(y_i - z_i) \tag{1.21}$$

This leaves a general form for the likelihood $p(y_i|\boldsymbol{\theta})$ as:

$$p(y_i|\boldsymbol{\theta}) = \mathcal{L}(\boldsymbol{\theta}|y_i) = \int \mathrm{d}z_i\ f_Z(z_i)\ f_E(y_i - z_i) \tag{1.22}$$

However this general equation is not in a form which can be practically applied to inference problems. To achieve a practical form further assumptions must be made about the distributions $f_Z(z_i)$ and $f_E(e)$. This is done by assigning parametric distributions. As we will see in the next section, assigning Gaussian distributions will allow 1.22 to form a usable expression which will be our applied-form likelihood.

The likelihood for a single dataset can be built upon to form a joint likelihood. As long as the two datasets are independent then equations 1.9 and 1.10 can be expanded for two datasets, here $\boldsymbol{y}^a$ and $\boldsymbol{y}^b$. Consider the joint form of equation 1.10:

$$\mathcal{L}(\boldsymbol{\theta}|\boldsymbol{y}^a, \boldsymbol{y}^b) = \mathcal{L}(\boldsymbol{\theta}|\boldsymbol{y}^a)\ \mathcal{L}(\boldsymbol{\theta}|\boldsymbol{y}^b) \propto \exp\left[-\frac{1}{2}(\boldsymbol{y}^a - \boldsymbol{g}^a(\boldsymbol{\theta}))^T (C_{\mathcal{M}^a} + C_{\mathcal{D}^a})^{-1} (\boldsymbol{y}^a - \boldsymbol{g}^a(\boldsymbol{\theta})) \right]$$
$$\exp\left[-\frac{1}{2}(\boldsymbol{y}^b - \boldsymbol{g}^b(\boldsymbol{\theta}))^T (C_{\mathcal{M}^b} + C_{\mathcal{D}^b})^{-1} (\boldsymbol{y}^b - \boldsymbol{g}^b(\boldsymbol{\theta})) \right] \tag{1.23}$$

As a result of the different scale of various dataset residuals, $(\boldsymbol{y} - \boldsymbol{g}(\boldsymbol{\theta}))^T (C_{\mathcal{M}} + C_{\mathcal{D}})^{-1} (\boldsymbol{y} - \boldsymbol{g}(\boldsymbol{\theta}))$, the joint likelihood equation 1.23 can be dominated by a dataset with significantly more data points or larger residual values. One method which naturally accounts for this problem is the trans-dimensional formulation (Sambridge et al., 2006). Another solution is a weighting scheme to balance the input of the different datasets. Some common forms of weighting schemes for two datasets are:

$$\text{residual}^a, \frac{1}{c}\ \text{residual}^b \tag{1.24}$$

$$c\ \text{residual}^a, 1 - c\ \text{residual}^b \tag{1.25}$$

where c is the weighting factor to be defined. The most common scheme for defining the weighting factors is ad-hoc. It relies on running synthetic tests to determine when the solution is giving a balance between the two datasets. This approach is likely to be biased by our expectations of synthetic tests. The resulting parameter uncertainties may not truly reflect the uncertainty associated with the combination of the two complimentary datasets.

Technical figure 2: Applied likelihood derivation

It is assumed that both y_i and z_i contain statistical uncertainty. For y_i the source of statistical uncertainty is a result of the measurement process. While for z_i the source of uncertainty is a result of modelization. It is necessary to distinguish between these two sources of uncertainty. The data uncertainty arising from the measurement process will be denoted $e_i^{\mathcal{D}}$, while modelization uncertainty will be denoted $e_i^{\mathcal{M}}$. As such:

$$y_i = z_i + e_i^{\mathcal{D}} \tag{1.26}$$

$$z_i = g_i(\boldsymbol{\theta}) + e_i^{\mathcal{M}} \tag{1.27}$$

$$\therefore y_i = g_i(\boldsymbol{\theta}) + e_i^{\mathcal{M}} + e_i^{\mathcal{D}} \tag{1.28}$$

$e_i^{\mathcal{M}}$ and $e_i^{\mathcal{D}}$ are assumed to be uncorrelated. If it is assumed that $e_i^{\mathcal{D}}$ is described as independent and identically distributed from a Gaussian with a predetermined standard deviation $\sigma_i^{\mathcal{M}}$:

$$p(z_i|\boldsymbol{\theta}) = f_Z(z_i) = \frac{1}{\sqrt{2\pi(\sigma_i^{\mathcal{M}})^2}} \exp\left[\frac{-(e_i^{\mathcal{M}})^2}{2(\sigma_i^{\mathcal{M}\,2})^2}\right] \tag{1.29}$$

Likewise, if $e_i^{\mathcal{D}}$ is assumed to be i.i.d from a Gaussian with standard deviation $\sigma_i^{\mathcal{D}}$ then:

$$p(e_i|\boldsymbol{\theta}) = f_E(e_i) = f_E(y_i - z_i) = \frac{1}{\sqrt{2\pi(\sigma_i^{\mathcal{D}})^2}} \exp\left[\frac{-(e_i^{\mathcal{D}})^2}{2(\sigma_i^{\mathcal{D}})^2}\right] \tag{1.30}$$

As a result of defining parametric distributions for $f_Z(z_i)$ and $f_E(y_i - z_i)$ the general definition of $\mathcal{L}(\boldsymbol{\theta}|y_i)$, equation 1.22, can be evaluated. This is the convolution of the two distributions:

$$\mathcal{L}(\boldsymbol{\theta}|y_i) = \frac{1}{\sqrt{2\pi}\sqrt{(\sigma_i^{\mathcal{M}})^2 + (\sigma_i^{\mathcal{D}})^2}} \exp\left[\frac{-(y_i - g_i(\boldsymbol{\theta}))^2}{2\left((\sigma_i^{\mathcal{M}})^2 + (\sigma_i^{\mathcal{D}})^2\right)}\right] \tag{1.31}$$

Many of the assumptions made about the nature of statistical uncertainty form scientifically testable hypotheses. For example, it is possible to determine whether the noise from the measurement process conforms to an i.i.d Gaussian distribution, or whether there is some degree of correlation. Likewise, the data can be probed to determine whether the noise conforms to a Gaussian, or perhaps some other distribution is more appropriate.

Equation 1.31 defines the likelihood for a single data point. How then to move forward with a full dataset?

If you have a set of data $\boldsymbol{y} = \{y_1, ..., y_M\}$, where each y_i term is independent then, considering uncertainty in the data and model as i.i.d Gaussian:

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta}|\boldsymbol{y}) &= p(y_1, ..., y_M|\boldsymbol{\theta}) \\ &= (2\pi)^{M/2}\Big(\prod_{i=1}^{M}\left((\sigma_i^{\mathcal{M}})^2 + (\sigma_i^{\mathcal{D}})^2\right)^{1/2}\Big) \\ &\exp\left[\sum_{i=1}^{M}\frac{-(y_i - g_i(\boldsymbol{\theta}))^2}{2\left((\sigma_i^{\mathcal{M}})^2 + (\sigma_i^{\mathcal{D}})^2\right)}\right]\end{aligned} \tag{1.32}$$

Equation 1.32 is the likelihood $\mathcal{L}(\boldsymbol{\theta}|\boldsymbol{y})$ for i.i.d Gaussian uncertainty.

For the case of correlated Gaussian uncertainty vectors of data and model parameters can be represented by a multivariate Gaussian distributions, defined with covariance matrices $\boldsymbol{C}_{\mathcal{D}}$ and $\boldsymbol{C}_{\mathcal{M}}$ respectively, such that:

$$f_Z(z) \propto \exp\left[-\frac{1}{2}(\boldsymbol{y} - \boldsymbol{g}(\boldsymbol{\theta}))^T \boldsymbol{C}_{\mathcal{M}}^{-1}(\boldsymbol{y} - \boldsymbol{g}(\boldsymbol{\theta}))\right] \tag{1.33}$$

$$f_E(e) \propto \exp\left[-\frac{1}{2}(\boldsymbol{y} - \boldsymbol{g}(\boldsymbol{\theta}))^T \boldsymbol{C}_{\mathcal{D}}^{-1}(\boldsymbol{y} - \boldsymbol{g}(\boldsymbol{\theta}))\right] \tag{1.34}$$

Then the likelihood involves a combination of their covariance matrices:

$$\mathcal{L}(\boldsymbol{\theta}|\boldsymbol{y}) \propto \exp\left[-\frac{1}{2}(\boldsymbol{y} - \boldsymbol{g}(\boldsymbol{\theta}))^T (\boldsymbol{C}_{\mathcal{M}} + \boldsymbol{C}_{\mathcal{D}})^{-1}(\boldsymbol{y} - \boldsymbol{g}(\boldsymbol{\theta}))\right] \tag{1.35}$$

Equation 1.32 and 1.35 represent commonly implemented forms of the likelihood. It is instructive to understand the origin of the likelihood function as often it is overlooked in applied studies. Of particular importance are the explicit assumptions which are necessary steps in construction. Assumptions include:

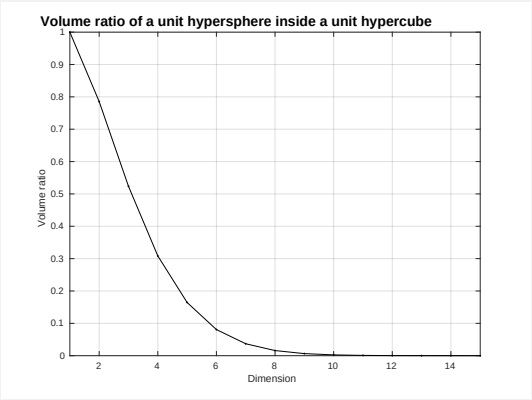
1. The form of the model, equation 1.28. The uncertainty is additive
2. The i.i.d Gaussian nature of measurement and modelization statistical uncertainty, or
3. The correlated multivariate Gaussian nature of the measurement and modelization uncertainty

A shift away from these assumptions is likely to close access to an analytical likelihood function.

Technical figure 3: Stable computation of likelihood values

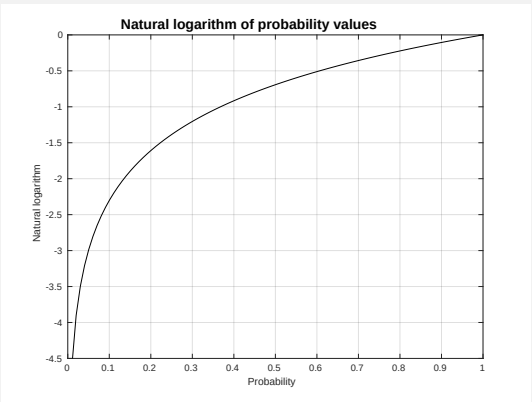
The stable computation of likelihood values is an essential component of running algorithms which resolve posterior distributions. This can be an issue as the value of the likelihood for a given set of parameters can be extraordinarily small, to the point where these values approach or cross the limits of what can be stored in a computers memory.

Consider a unit hyper sphere that is located in a unit hyper cube. In low dimensions the sphere inside a cube takes up a large % of the cubes volume. However, as the dimension increases this volume very rapidly diminishes. Figure 1.4 illustrates this relationship. The purpose of this example is to demonstrate that high dimensional spaces are very empty. Geophysical inverse problems suffer particularly from this dimensionality, our images have many unknown parameters. Hence problems arising as a result of dimensionality, often referred to as *the curse of dimensionality*, are of particular importance to geophysics.



For a likelihood distribution over a high dimensional parameter space, the majority of the space will be extraordinarily empty - i.e. with very low probability. Considering probability is a value between 1 and 0, that means a lot of parameter space will have a likelihood very close to 0. Given this circumstance it is essential to be able to accurately compute and compare extremely small probability values. The risk of improper evaluation and comparison is erroneous steps in algorithms and a divergence from the posterior the algorithm is supposed to be sampling. In short your solution will be wrong.

How to overcome this issue? The solution is to compute the natural logarithm of the likelihood. Figure 1.4 shows taking the natural logarithm of probability values. The result is very small probability values become very large negative numbers.



These very large negative numbers transform a very small value, close to or crossing the limits of arithmetic underflow, into a large number which can be easily stored.

For a Gaussian likelihood, taking the natural logarithm of an unnormalized likelihood is in effect removing the exponential term. Consider the likelihood as in equation 1.10. Repeated here

$$\mathcal{L}(\boldsymbol{\theta}|\boldsymbol{y}) \propto \exp\left[-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{g}(\boldsymbol{\theta}))^T (C_{\mathcal{M}}+C_{\mathcal{D}})^{-1}(\boldsymbol{y}-\boldsymbol{g}(\boldsymbol{\theta}))\right] \quad (1.36)$$

For $l(\boldsymbol{\theta}|\boldsymbol{y}) = \ln(\mathcal{L}(\boldsymbol{\theta}|\boldsymbol{y}))$

$$l(\boldsymbol{\theta}|\boldsymbol{y}) \propto -\frac{1}{2}(\boldsymbol{y}-\boldsymbol{g}(\boldsymbol{\theta}))^T (C_{\mathcal{M}}+C_{\mathcal{D}})^{-1}(\boldsymbol{y}-\boldsymbol{g}(\boldsymbol{\theta})) \quad (1.37)$$

Reference to the negative log-likelihood would then mean

$$-l(\boldsymbol{\theta}|\boldsymbol{y}) \propto \frac{1}{2}(\boldsymbol{y}-\boldsymbol{g}(\boldsymbol{\theta}))^T (C_{\mathcal{M}}+C_{\mathcal{D}})^{-1}(\boldsymbol{y}-\boldsymbol{g}(\boldsymbol{\theta})) \quad (1.38)$$

The negative log-likelihood is then a large positive number for very small probability values. Often minimization of an equation in the form of equation 1.38 is the goal of inversion schemes. Herein lies a connection between the likelihood function and linearized schemes, the result of a successful linearized inversion represents the minimum of the negative log-likelihood, $-l(\boldsymbol{\theta}|\boldsymbol{y})$. Which is also the maximum likelihood value. However our goal is generally broader. To explore the joint posterior distribution, a function of both likelihood and prior. To achieve this goal in high dimensional spaces practitioners often use Markov chain Monte Carlo (MCMC) sampling. So how does our shift to working with the natural logarithm of the likelihood impact this algorithm?

For a MCMC algorithm with stationary distribution which is the posterior $p(\boldsymbol{\theta}|\boldsymbol{y})$, the key ingredient is evaluating the Metropolis-Hastings acceptance ratio, α

$$\alpha = \frac{\mathcal{L}(\boldsymbol{\theta}^*|\boldsymbol{y})\; p(\boldsymbol{\theta}^*)}{\mathcal{L}(\boldsymbol{\theta}|\boldsymbol{y})\; p(\boldsymbol{\theta})} \quad (1.39)$$

Where * represents a candidate move from the proposal distribution $q(\boldsymbol{\theta}, \cdot)$. As long as the proposal distribution is a symmetric distribution the ratio $q(\boldsymbol{\theta}^*, \boldsymbol{\theta})/q(\boldsymbol{\theta}, \boldsymbol{\theta}^*)$ does not need to be computed as part of α.

Hence, the ratio of posterior values needs to be computed at each time step over the parameter space. The raw posterior values will be prone to numerical underflow due to the limits of floating point arithmetic, hence the log values for the likelihood must be used in this calculation. How does the form of α get adjusted within a computer program to ensure stable and reliable computation?

This shift represents a major divergence between how the theory is established in equation form, and the practical application in computer code.

Consider we have access to the log-likelihood and log-prior. Recall the logarithm rules where $\ln(xy) = \ln(x) + \ln(y)$ and $\ln(x/y) = \ln(x) - \ln(y)$. Then

$$\alpha = \exp\left[\left(l(\boldsymbol{\theta}^*|\boldsymbol{y}) - l(\boldsymbol{\theta}|\boldsymbol{y})\right) + \left(\ln(p(\boldsymbol{\theta}^*)) - \ln(p(\boldsymbol{\theta}))\right)\right] \quad (1.40)$$

The resulting acceptance-ratio computed here will be equivalent to equation 1.39, however, now the probability densities required in the computation will be restricted to large numbers, stable in their computation and comparison. The end result, a reliable MCMC algorithm.

1.5 Approximate Bayesian Computation

The likelihood, while underpinning a system which is formally strong and practically useful for geophysics, does suffer some weaknesses which are implicit in its construction and application. The likelihood limits the number of models (combined deterministic forward and uncertainty) which can be considered and applied as it requires a closed form expression. The likelihood also mixes and dilutes the available information about the data fit and model into a single metric. While necessary in the likelihood framework, this may not be the best way to drive an inversion scheme. Other disciplines have found Approximate Bayesian Computation offers an alternative to traditional likelihood machinery which can utilize more of what is known about the structure, physics and nature of the parameter inference problems at hand (Tavare et al., 1997; Ratmann et al., 2009; Vrugt and Sadegh, 2013). It opens inference to a broader range of models and likewise offers formal statistical guarantees (Sunnåker et al., 2013).

Likelihood-free methods for Bayesian inference have been developed in response to parameter estimation problems where it is not possible to formulate or justify a closed form likelihood (e.g. Tavare et al. (1997); Fu and Li (1997); Weiss and von Haeseler (1998); Pritchard et al. (1999); Beaumont et al. (2002); Marjoram et al. (2003)). Instead, likelihood-free methods target the same posterior distribution without evaluating a likelihood. The algorithms and methods of likelihood-free Bayesian inference have been termed *Approximate Bayesian Computation* (ABC). ABC simply requires the ability to simulate data given model parameters. Problems ripe for attack by ABC are common in science due to the breadth of models which have been developed to describe natural systems. It is frequently the case that the uncertainty associated with the data can be simulated rapidly but explicit formulas for probability distributions are difficult to formulate, expensive to evaluate, impossible to justify, or do not exist. Hence, traditional likelihood machinery becomes infeasible. ABC is a means to overcome these issues, it is backed by a sound theoretical underpinning and is subject to a rapidly expanding set of literature (Ratmann et al., 2009; Blum et al., 2010; Vrugt and Sadegh, 2013; Sunnåker et al., 2013; Blum et al., 2013; Sadegh and Vrugt, 2014; Pudlo et al., 2015; Meeds et al., 2015; Lintusaari et al., 2016; Gutmann et al., 2016; Sisson et al., 2016; Li et al., 2017).

The premise of ABC begins with the notion that a set of model parameters, $\boldsymbol{\theta}$, is a sample from the posterior if the observed data, $\boldsymbol{y}$, and a simulated dataset, $\boldsymbol{y}^*$, are equal. Algorithm 1 generates i.i.d samples from the posterior $p(\boldsymbol{\theta}|\boldsymbol{y})$ (cf. Marjoram et al. (2003)).

Algorithm 1

1. Generate $\boldsymbol{\theta}^*$ as a random sample from $p(\boldsymbol{\theta})$
2. Simulate $\boldsymbol{y}^*$ from the forward operator $\boldsymbol{g}_s(\boldsymbol{\theta}^*)$
3. Accept $\boldsymbol{\theta}^*$ as a posterior sample if $\boldsymbol{y}^* = \boldsymbol{y}$
4. Repeat

The operator $\boldsymbol{g}_s$ in Algorithm 1, and all of likelihood-free inference, differs from the forward operators traditionally defined in geophysics. Forward operators in geophysics are *deterministic*. That is, for a given set of model parameters, $\boldsymbol{\theta}$, the output of the forward operator $\boldsymbol{g}(\boldsymbol{\theta})$ will always be the same. The uncertainty in both data and model, are then later built into the construction of a likelihood. The ABC forward operator which simulates data, $\boldsymbol{g}_s$, is *stochastic*. That is, the uncertainty from both data and modelization is built into the simulation process. As a result of this modification, algorithm 1 does not require the formulation or evaluation of a likelihood. The ABC forward operator $\boldsymbol{g}_s$ allows a practitioner to build in whatever uncertainty is justified for the scientific problem at hand.

Algorithm 1 is acceptable for basic problems where the datasets $\boldsymbol{y}$ and $\boldsymbol{y}^*$ are a discrete probability function, e.g. integers (Tavare et al., 1997; Fu and Li, 1997). However, for large and continuous datasets the probability of generating a sample where the acceptance criteria, $\boldsymbol{y}^* = \boldsymbol{y}$, is met diminishes to levels unacceptable for parameter inference. ABC relaxes the problematic requirement of equality by accepting samples when the distance between $\boldsymbol{y}^*$ and $\boldsymbol{y}$, $\mathrm{d}(\boldsymbol{y}, \boldsymbol{y}^*)$, is less than a tolerance value ϵ (Weiss and von Haeseler, 1998). This method, algorithm 2, does not target our true posterior of interest, but instead, the ABC posterior $p(\boldsymbol{\theta}|\mathrm{d}(\boldsymbol{y}, \boldsymbol{y}^*) \leq \epsilon)$ which approximates the true posterior.

Algorithm 2

1. Generate $\boldsymbol{\theta}^*$ as a random sample from $p(\boldsymbol{\theta})$
2. Simulate $\boldsymbol{y}$ from the forward operator $\boldsymbol{g}_s(\boldsymbol{\theta}^*)$
3. Accept $\boldsymbol{\theta}^*$ as a posterior sample if $\mathrm{d}(\boldsymbol{y}, \boldsymbol{y}^*) \leq \epsilon$
4. Repeat

Algorithm 2 requires a user specified metric, d, which defines the distance between datasets. Common choices are the absolute-value norm, Euclidean distance and Mahalanobis distance. Likewise, the value for the tolerance ϵ must be user specified. As $\epsilon \to \infty$, the sampled target distribution of algorithm 2 $p(\boldsymbol{\theta}|\mathrm{d}(\boldsymbol{y}, \boldsymbol{y}^*) \leq \epsilon) \to p(\boldsymbol{\theta})$. Algorithm 2 simply recovers the prior distribution. However, as $\epsilon \to 0$ the sampled distribution $p(\boldsymbol{\theta}|\mathrm{d}(\boldsymbol{y}, \boldsymbol{y}^*) \leq \epsilon) \to p(\boldsymbol{\theta}|\boldsymbol{y})$. The exact posterior is recovered. Encoded in the value of ϵ is a trade-off between acceptance rate and accuracy. As ϵ increases so does the efficiency of the sampler, but the accuracy of the recovered distribution is increasingly eroded (Sisson and Fan, 2010). Conversely, as ϵ decreases the accuracy of the recovered distribution converges to the true posterior, but the acceptance rate approaches computationally infeasible levels. This trade-off for computational efficiency at the cost of accuracy is where ABC derives the name *Approximate* Bayesian Computation, as the samplers recover a distribution which approximates the true posterior.

Rejection sampler inference in the form of algorithm 2 can run into efficiency problems as the dimensionality (size) of the datasets grow. In this case the probability of sampling $\mathrm{d}(\boldsymbol{y}, \boldsymbol{y}^*) \leq \epsilon$ diminishes as the size of the dataset grows. Since first application (Tavare et al., 1997) likelihood-free methods have adopted the use of low-dimensional summary statistics about the data in the evaluation of distance, d. A metric over a vector of summary statistics is evaluated, $\mathrm{d}(\boldsymbol{S}(\boldsymbol{y}), \boldsymbol{S}(\boldsymbol{y}^*))$. This leads to

the most common form of an ABC rejection sampling scheme of algorithm 3 (Pritchard et al., 1999). This rejection sampler generates i.i.d samples from the distribution $p(\boldsymbol{\theta}|\mathrm{d}(\boldsymbol{S}(\boldsymbol{y}),\boldsymbol{S}(\boldsymbol{y}^*)) \leq \epsilon)$.

Algorithm 3

1. Generate $\boldsymbol{\theta}^*$ as a random sample from $p(\boldsymbol{\theta})$
2. Simulate $\boldsymbol{y}^*$ from the forward operator $\boldsymbol{g}_s(\boldsymbol{\theta}^*)$
3. Compute summary statistics $\boldsymbol{S}(\boldsymbol{y})$ and $\boldsymbol{S}(\boldsymbol{y}^*)$
4. Accept $\boldsymbol{\theta}^*$ as a posterior sample if $\mathrm{d}(\boldsymbol{S}(\boldsymbol{y}),\boldsymbol{S}(\boldsymbol{y}^*)) \leq \epsilon$
5. Repeat

In the first application of likelihood-free inference, Tavare et al. (1997) replace the full sequences of DNA with the number of sites which differ between DNA samples. The idea being that as long as the statistics used are *sufficient* there is no information loss for parameter inference. As a result of sufficiency the posterior computed with statistics will be equivalent to the the posterior computed by the full dataset, i.e $p(\boldsymbol{\theta}|\boldsymbol{S}(\boldsymbol{y})) = p(\boldsymbol{\theta}|\boldsymbol{y})$. It is often the case that no truly sufficient statistics exist for problems of scientific interest. Instead practitioners settle for a reasonably sufficient low-dimensional set of summary statistics. This lack of sufficiency introduces a second bias, tolerance being the first, into the ABC-posterior $p(\boldsymbol{\theta}|\mathrm{d}(\boldsymbol{S}(\boldsymbol{y}),\boldsymbol{S}(\boldsymbol{y}^*)) \leq \epsilon)$ due to the information lost by summarizing the full data set. It is, however, thanks to summary statistics that likelihood-free inference owes its origin and can proceed. Summary statistics have allowed parameter inference to be applied to problems as challenging as noisy near-chaotic ecology populations (Wood, 2010) and have been shown to offer superior power to diagnose model insufficiency (Ratmann et al., 2009; Vrugt and Sadegh, 2013).

Seminal developments in ABC have considered algorithms and equations in the form which have been introduced so far (Fu and Li, 1997; Pritchard et al., 1999; Beaumont et al., 2002; Marjoram et al., 2003). However, ABC can be cast in another mathematical light which enables straight forward implementation into sampling routines such as MCMC. The stochastic forward simulations from ABC, $\boldsymbol{y}^*$, are viewed as an auxiliary parameter which is introduced into the posterior to facilitate computation (Sisson and Fan, 2010). This process changes the computed posterior from the traditional $p(\boldsymbol{\theta}|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})$ to the ABC posterior:

$$p_{ABC}(\boldsymbol{\theta},\boldsymbol{y}^*|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{y}^*,\boldsymbol{\theta})p(\boldsymbol{y}^*|\boldsymbol{\theta})p(\boldsymbol{\theta}) \tag{1.41}$$

where $\boldsymbol{y}^*$ is viewed as a realization from the density $p(\boldsymbol{y}^*|\boldsymbol{\theta})$. $p(\boldsymbol{y}|\boldsymbol{y}^*,\boldsymbol{\theta})$ is introduced to serve the same role as the accept/reject step in Algorithms 2 and 3. $p(\boldsymbol{y}|\boldsymbol{y}^*,\boldsymbol{\theta})$ is chosen to weight the posterior with high values when the observed and simulated datasets are close. However, the form of equation 1.41 allows the mathematics of kernel densities to be introduced into $p(\boldsymbol{y}|\boldsymbol{y}^*,\boldsymbol{\theta})$ such that (Sisson and Fan, 2010):

$$p(\boldsymbol{y}|\boldsymbol{y}^*,\boldsymbol{\theta}) = \frac{1}{\epsilon}K\left(\frac{\mathrm{d}(\boldsymbol{y},\boldsymbol{y}^*)}{\epsilon}\right) \tag{1.42}$$

Where K is some standard kernel, and the tolerance ϵ serves as the kernel bandwidth.

Thinking of equation 1.41 in the form of a rejection sampler still allows a straightforward understanding.

Algorithm 4

1. Generate $\boldsymbol{\theta}^*$ as a random sample from $p(\boldsymbol{\theta})$
2. Simulate $\boldsymbol{y}^*$ from the forward operator $\boldsymbol{g}_s(\boldsymbol{\theta}^*)$
3. The distance between simulated and observed datasets is evaluated
4. The parameter set is accepted with probability equal to $p(\boldsymbol{y}|\boldsymbol{y}^*, \boldsymbol{\theta})$
5. Repeat

Algorithm 4 is a more general version of algorithm 2. They are equivalent when K is a uniform distribution $\mathcal{U}(0, 1)$.

While equation 1.41 targets the joint posterior density of simulations and model parameters, marginalization to our posterior of interest, $p_{abc}(\boldsymbol{\theta}|\boldsymbol{y})$, is done numerically by discarding the simulations recovering:

$$p_{ABC}(\boldsymbol{\theta}|\boldsymbol{y}) \propto p(\boldsymbol{\theta}) \int_{\boldsymbol{y}^*} p(\boldsymbol{y}|\boldsymbol{y}^*, \boldsymbol{\theta}) p(\boldsymbol{y}^*|\boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{y}^* \tag{1.43}$$

Equation 1.43 can be adjusted to rely on summary statistics. The density $p(\boldsymbol{S}(\boldsymbol{y}^*)|\boldsymbol{\theta})$ is introduced as the density implied from taking summary statistics about $p(\boldsymbol{y}^*|\boldsymbol{\theta})$ and $p(\boldsymbol{S}(\boldsymbol{y})|\boldsymbol{S}(\boldsymbol{y}^*), \boldsymbol{\theta})$ is a kernel over the distance between summary statistics. Our approximate Bayesian posterior distribution becomes:

$$p_{ABC}(\boldsymbol{\theta}|\boldsymbol{S}(\boldsymbol{y})) \propto p(\boldsymbol{\theta}) \int_{\boldsymbol{S}(\boldsymbol{y}^*)} p(\boldsymbol{S}(\boldsymbol{y})|\boldsymbol{S}(\boldsymbol{y}^*), \boldsymbol{\theta}) \, p(\boldsymbol{S}(\boldsymbol{y}^*)|\boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{S}(\boldsymbol{y}^*) \tag{1.44}$$

Marjoram et al. (2003) first demonstrated a Markov chain Monte Carlo (MCMC) scheme to sample the ABC posterior distribution. A simple MCMC algorithm with a Metropolis-Hastings (MH) acceptance probability is demonstrated in algorithm 5. The M-H acceptance probability α to recover the ABC posterior is:

$$\alpha_{ABC} = \frac{p(\boldsymbol{S}(\boldsymbol{y})|\boldsymbol{S}(\boldsymbol{y}^*), \boldsymbol{\theta}) \, p(\boldsymbol{\theta}^*) \, q(\boldsymbol{\theta}^*, \boldsymbol{\theta_{t-1}})}{p(\boldsymbol{S}(\boldsymbol{y})|\boldsymbol{S}(\boldsymbol{y_{t-1}}), \boldsymbol{\theta}) \, p(\boldsymbol{\theta_{t-1}}) \, q(\boldsymbol{\theta_{t-1}}, \boldsymbol{\theta}^*)} \tag{1.45}$$

Where $\boldsymbol{t}$ denotes the time step in the chain, and $\boldsymbol{\theta}^*$ is used to represent a candidate move from the proposal distribution $q(\boldsymbol{\theta_{t-1}}, \cdot)$.

Algorithm 5

1. Start from an initial state $\boldsymbol{\theta^0}$ and select a proposal distribution $q(\cdot, \cdot)$
2. At each step where the current state is $\boldsymbol{\theta_{t-1}}$, propose a candidate move $\boldsymbol{\theta^*}$ from the distribution $q(\boldsymbol{\theta_{t-1}}, \cdot)$
3. If the candidate state is better than the previous state, $\alpha_{ABC} > 1$, then the candidate state is accepted unconditionally meaning $\boldsymbol{\theta_t} = \boldsymbol{\theta^*}$
4. If the candidate move is not better in the above sense, then $\boldsymbol{\theta^*}$ is accepted with probability equal to α_{ABC}
5. If the candidate move is not accepted, then the chain remains in its current state, meaning $\boldsymbol{\theta_t} = \boldsymbol{\theta_{t-1}}$
6. Repeat the simulation steps 2-5 until enough values have been generated

ABC originated from applied problems where a likelihood was not available. As we saw in the last section, and expanded upon in technical figure 2, geophysics has been able to leverage likelihoods, for example, in the form of equations 1.9 and 1.10. This access to a likelihood is the direct result of the assumptions about the statistical properties of the measurement and modelization uncertainty. However, under more complex models for uncertainty, an analytical formula for the likelihood may not be accessible. This is where ABC algorithms have traditionally been able to step in, bypassing a likelihood, and in the process opening parameter inference a range of more complex models.

The main postulate of this thesis is that ABC can offer improvements over some limiting aspects of traditional likelihood based Bayesian inference. This improvement may be focused in several different areas. For instance, ABC does not require simple distributions (e.g. Gaussian) for modelling uncertainties. This may allow, for example, realistic measurement and modelization uncertainties to be incorporated into the inversion scheme. ABC can uniquely investigate the adequacy of the model in absolute terms against the data, as opposed to relative to the performance of other models (ABCμ of Ratmann et al. (2009)). This allows fundamental deficiencies in the model to be exposed. While the closed form expression for the likelihood requires the mixing and dilution of information into a single metric, this is not required of ABC. ABC may be capable of considering the full scope of information available within a simulated dataset to drive a diagnostic inversion. In short, the freedom of ABC provides an alternative to the current paradigm of inversion schemes. In the same way approximate numerical methods allow us to tackle a greater number of problems than otherwise possible with analytical solutions.

Many of the above possibilities are outside the scope of an 9 month Masters' project. However these will serve as guiding aims and potential for future exploration. My goal here is to simply demonstrate some advantage of a geophysical ABC scheme, which will motivate future investigation.

Chapter 2

ABC

This chapter will outline a series of illustrative ABC experiments. These examples demonstrate the core concepts introduced in section 1.5, demonstrate ABC can sample from complex distributions and build a boutique code base which is the foundations of this project. The code for the figures in this section can be found at https://github.com/tomconnell/abc-toy-problems.

2.1 Toy problem 1: 1D Gaussian

As a first example consider we have observed $n = 100$ realizations from a Gaussian distribution, making our model: $\boldsymbol{g}_s = p(\boldsymbol{y}^*|\boldsymbol{\theta}) \sim \mathcal{N}(\mu, \sigma)$, with unknown model parameters $\boldsymbol{\theta} = [\mu, \sigma]$. Given that it is easy to simulate this model, it is possible to leverage ABC algorithms to estimate the unknown model parameters. A synthetic dataset for this problem is created with $\mu = 5$ and $\sigma = 2$. The summary statistics sample mean, $\bar{\mu}$, and sample standard deviation, $\bar{\sigma}$, are used, $\boldsymbol{S}(\boldsymbol{y}) = [\bar{\mu}, \bar{\sigma}]$. These are sufficient statistics for the unknown model parameters. Figure 2.1 plots the ABC posterior obtained from using algorithm 3, the traditional form of an ABC rejection sampler, compared to the analytical likelihood. The distance between simulated summary statistics and observed summary statistics is evaluated marginally for simplicity. Throughout this thesis I do not consider the joint distribution and correlation structure present in the distance between summaries. Our distance measure for this example takes the form $\mathrm{d}(\boldsymbol{S}(\boldsymbol{y}), \boldsymbol{S}(\boldsymbol{y}^*)) = |S_1(\boldsymbol{y}^*) - S_1(\boldsymbol{y})| + |S_2(\boldsymbol{y}^*) - S_2(\boldsymbol{y})|$. A uniform prior is used to give equal probability to a bounded area, $p(\mu) = \mathcal{U}(0, 10)$ and $p(\sigma) = \mathcal{U}(0, 10)$. Figure 2.2 explores the impact of varying the tolerance for this problem.

Figure 2.1 demonstrates that with sufficient statistics and a low tolerance ABC can accurately resolve the posterior using only the ability to simulate data. However, as figure 2.2 demonstrates, high tolerances erode posterior accuracy and uncertainty is over-estimated. However, it is true that increasing the tolerance increases the acceptance rate and hence relaxes computational resources. In this case the acceptance rate with $\epsilon = 0.1$ was 0.02%, while the acceptance rate with $\epsilon = 1$ was 2.02%. Under a model which is expensive to simulate from, walking the tightrope between accuracy and efficiency becomes important and needs to be carefully examined. In spite of shortcomings, the strengths of the ABC rejection sampler has led to significant scientific results in genetics (Fu and Li, 1997; Weiss and von Haeseler, 1998; Pritchard et al., 1999).

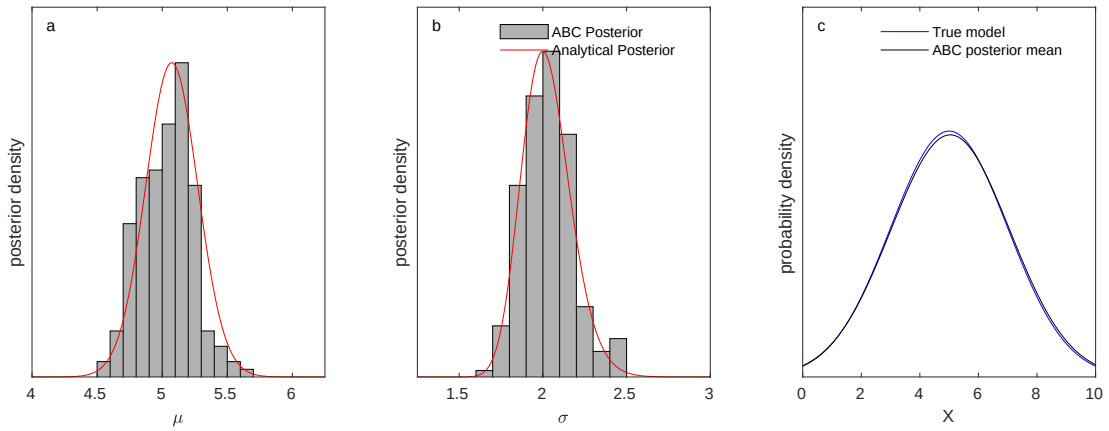


FIGURE 2.1: Posterior comparison between ABC and traditional likelihood inference for estimating the parameters, $\boldsymbol{\theta} = [\mu, \sigma]$, for a Gaussian model given $n = 100$ observations, $\boldsymbol{y}$. The ABC algorithm, algorithm 3, uses 1 million repetitions and a tolerance $\epsilon = 0.1$. The likelihood takes the form $\mathcal{L}(\boldsymbol{\theta}|\boldsymbol{Y}) = (2\pi\sigma^2)^{-n/2} \exp\Big[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2\Big]$. (a) Marginal posterior compared to marginal ABC posterior for unknown parameter μ. (b) same as (a) but for unknown parameter σ. (c) Mean of the joint ABC posterior compared to true model, $\mathcal{N}(5, 2)$.

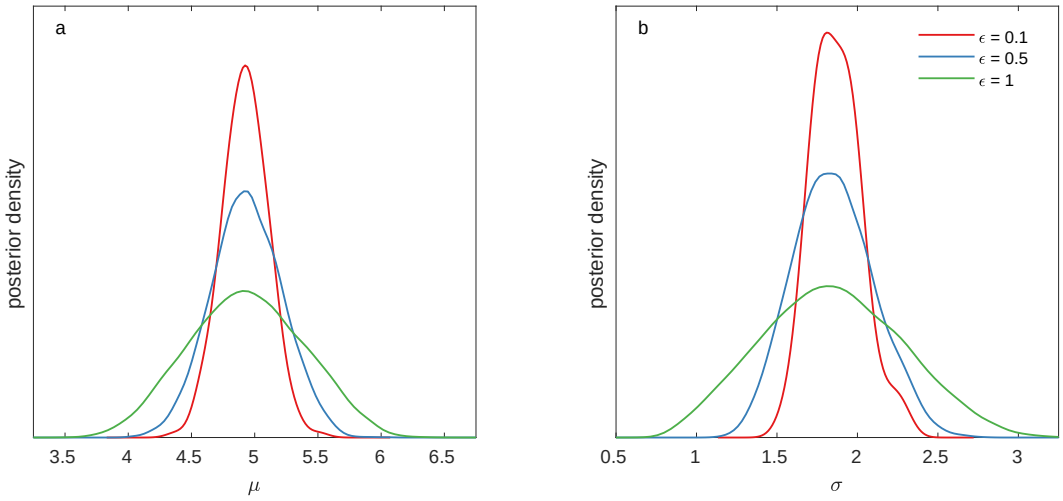


FIGURE 2.2: The effect of varying the tolerance ϵ when estimating $\boldsymbol{\theta} = [\mu, \sigma]$ to a Gaussian model given given $n = 100$ observations, $\boldsymbol{y}$. Three tolerances are considered, $\epsilon = 0.1$, $\epsilon = 0.5$, $\epsilon = 1$. (a) Kernel density estimate of the marginal ABC posterior for unknown parameter μ. The true value is $\mu = 5$ (b) same as (a) but for unknown parameter σ. The true value is $\sigma = 2$

.

2.2 Toy problem 2: Linear regression

As a second example consider we have observed some data, $\boldsymbol{y}$, from the linear model $\boldsymbol{g}(\boldsymbol{\theta}) = m\boldsymbol{x} + b$ and there is some stochasticity in the measurement process such that $\boldsymbol{g_s}(\boldsymbol{\theta}) = \boldsymbol{g}(\boldsymbol{\theta}) + \mathcal{N}(0, \sigma^2)$. Our unknown parameters are $\boldsymbol{\theta} = [m, b]$, while σ is known. In this case MCMC, which uses local transitions, will be needed to improve acceptance rates (Gilks et al., 1995). MCMC will also be needed when the search spaces are high dimensional, with many unknown parameters, or the posterior is a long way from the prior. For this case we can call upon ABC-MCMC in the form of algorithm 5 (Marjoram et al., 2003; Sisson and Fan, 2010).

Algorithm 5 samples the ABC posterior, equation 1.44. The algorithm relies on evaluating the MH acceptance probability, equation 1.45. The current examples use Gaussian proposal distributions, $q(\cdot, \cdot)$. As such the proposal distributions cancel out in equation 1.45. That leaves the prior and the weighting kernel to be evaluated at each time step in the Markov chain. The weighting kernel takes the form of equation 1.42. In this case a uniform weighting kernel, K_U, is implemented. The interpretation of K_U is the same as the accept/reject step in the rejection sampler:

$$K_U = \begin{cases} 1 & \text{if } \frac{\mathrm{d}(S_i(\boldsymbol{y}^*), S_i(\boldsymbol{y}))}{\epsilon_i} \leq 1 \\ 0 & \text{if } \frac{\mathrm{d}(S_i(\boldsymbol{y}^*) - S_i(\boldsymbol{y}))}{\epsilon_i} > 1 \end{cases} \tag{2.1}$$

K_U hence forms an indicator function which is equal to 1 when the distance between statistics, which are fit marginally, is less than the tolerance. Given, $\boldsymbol{S} = \{S_1, \ldots, S_O\}$, the weighting kernel takes the form:

$$p(\boldsymbol{S}(\boldsymbol{y})|\boldsymbol{S}(\boldsymbol{y}^*), \boldsymbol{\theta}) = \prod_{i=1}^{O} K_U\left(\frac{\mathrm{d}(S_i(\boldsymbol{y}), S_i(\boldsymbol{y}^*)}{\epsilon_i}\right) \tag{2.2}$$

As with the 1D Gaussian example it would be possible to use summaries which have a one-to-one correspondence to the unknown parameters. For a given simulation, a linear model could be fit to the simulated data, i.e. the slope and intercept of the fit model could then be used as summaries. However, to demonstrate that this is not necessary, the sample mean, $\bar{\mu}$, and sample standard deviation, $\bar{\sigma}$, are used as summary statistics (Vrugt and Sadegh, 2013). The slope is restricted to a positive value to make these statistics sufficient.

Figure 2.3 shows ABC-MCMC applied to the linear regression problem.

The next section continues to introduce algorithmic consideration into ABC, building toward a geophysically relevant parameter estimation method.

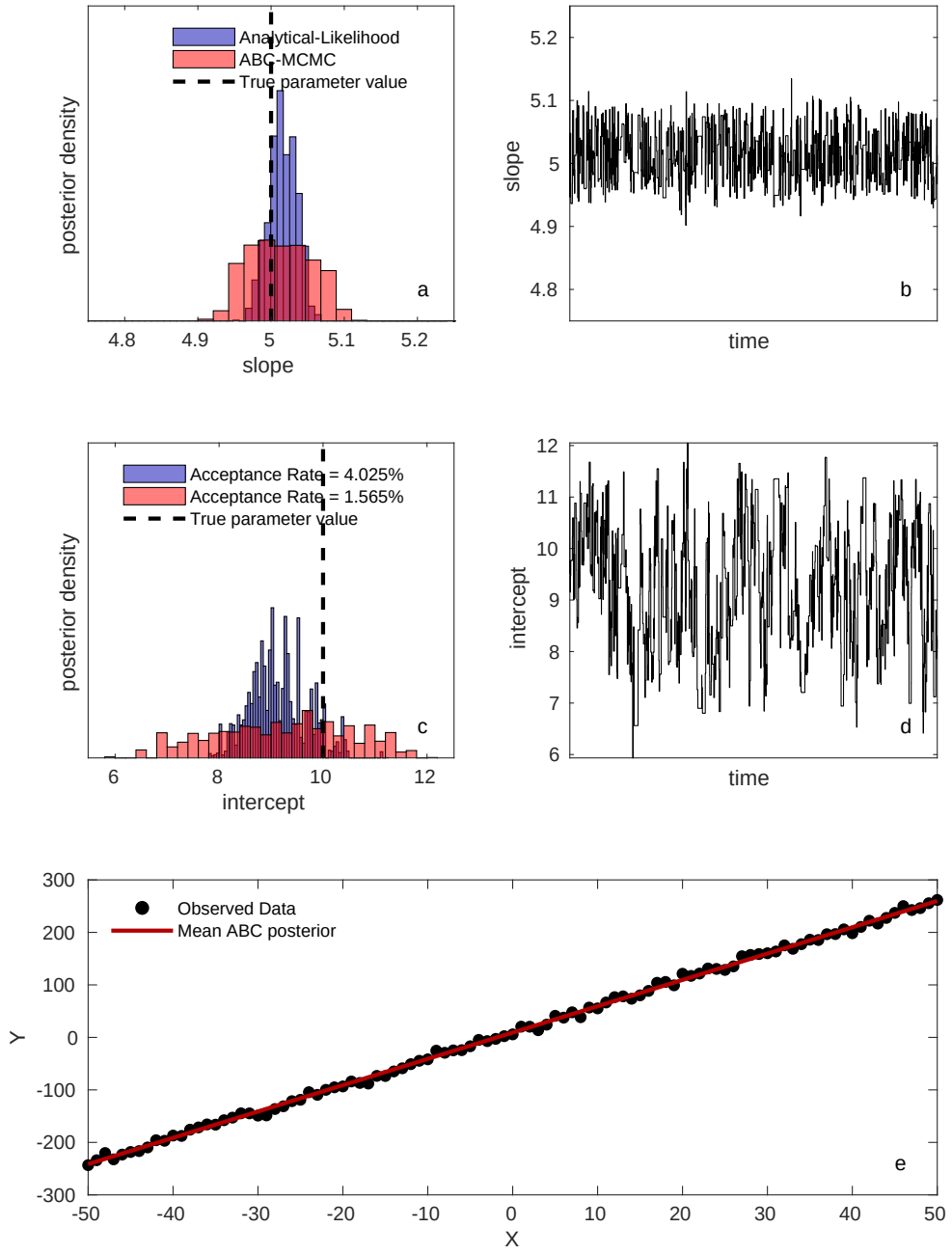


FIGURE 2.3: Linear regression with ABC. The ABC posterior is also compared to traditional likelihood inference. MCMC is used to sample both posteriors. For ABC the tolerance is $\epsilon = [2,2]$. The proposal distribution is $q = \mathcal{N}([0,0], I)$. The Markov chain length is $20,000$. (a) The marginal ABC posterior and marginal analytical posterior compared for unknown parameter m. (b) Plot of the ABC-MCMC Markov chain through time for m. (c) Same as (a) except for unknown parameter b. (d) Same as (b) except for unknown parameter b. (e) Comparison of the mean ABC posterior model and the observed data generated with $m = 5$, $b = 10$ and $\sigma = 5$.

2.3 Toy problem 3: Bivariate Gaussian

As a third example, consider we have observed $n = 100$ realizations from a bivariate Gaussian distribution $[X,\ Y] = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\mu} = \begin{bmatrix} \mu_X,\ \mu_Y \end{bmatrix}^T$ and $\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{bmatrix}$ where both $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ are unknown, $\boldsymbol{\theta} = [\boldsymbol{\mu}, \boldsymbol{\Sigma}]$. The causative model has parameters $\boldsymbol{\mu} = [2.5,\ 7.5]$, $\sigma_X = 4$, $\sigma_Y = 6$ and $\rho = 0.9$. Figure 2.4(a) plots the causative model and the $n = 100$ realizations which constitute the observed data. This example has 5 unknown parameters in total and ABC-MCMC is used for efficient sampling. The weighting kernel takes the same form as equation 2.2 and summaries with a one-to-one correspondence to the unknown parameters are used, the marginal sample means, $\bar{\mu_X}$ and $\bar{\mu_Y}$, sample standard deviations, $\bar{\sigma_X}$ and $\bar{\sigma_Y}$, and sample covariance, $\bar{\text{cov}}(X, Y)$. The prior distribution, $p(\boldsymbol{\theta}) = \mathcal{U}(0, 10)$, is used to provide equal probability to a bounded area for all unknowns beside correlation, ρ, which is bounded as $p(\rho) = \mathcal{U}(-1, 1)$. Figure 2.4(b) shows the results of the first 10,000 time steps from an ABC-MCMC. This figure shows the Markov chain does not make an accepted proposal during the first 4,500 time steps.

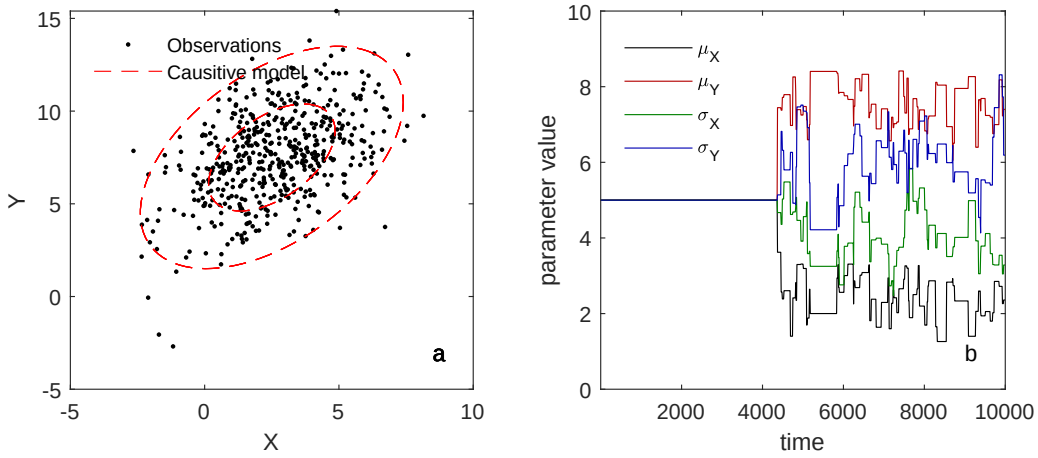


FIGURE 2.4: (a) Observed data from the underlying causative model, whose parameters will be the inference target. (b) First 10 000 iterations of an ABC-MCMC scheme with a uniform weighting kernel. The algorithm initially struggles to find a region of non-zero probability from a random starting position within the parameter space.

While ABC-MCMC and the uniform weighting kernel K_U offer improved acceptance rates compared to a rejection scheme, the algorithm can run into an initialization problem, as highlighted in figure 2.4(b). This is a problem which occurs when using a weighting kernel $p(\boldsymbol{y}|\boldsymbol{y}^*, \boldsymbol{\theta})$ with compact support. This problem originates when the chains starting position is far from a region where the weighting kernel offers support, within the bounds of $\pm\epsilon$. The chain then cannot make efficient local transitions via the proposal distribution to the region of high posterior density. There are several strategies to overcome this problem. For example, repeated simulations can be made from the prior until a starting position is found which is within $\pm\epsilon$ (Sisson and Fan, 2010). Alternatively, Bortot

et al. (2007) augment the posterior to $p_{ABC}(\boldsymbol{\theta}, \boldsymbol{y}^*, \epsilon|\boldsymbol{y})$, introducing ideas akin to simulated annealing or simulated tempering to ABC via a variable ϵ. Even if the initialization is overcome, the acceptance rate under K_U with low ϵ will be low when moving through regions of very diffuse posterior density. Here, I circumvent the initialization problem which occurs in weighting kernels with compact support by favoring a weighting kernel $p(\boldsymbol{y}|\boldsymbol{y}^*, \boldsymbol{\theta})$ with infinite support. That is, the weight diminishes with distance, however it never reaches zero. There are many kernels which have this required property. Here I use a weighting kernel based on the Gaussian distribution, K_G, figure 2.5(a). This choice is advantageous as the log-distance can be evaluated for stable computation in sampling algorithms. This mirrors the stable implementation of likelihood values, expanded upon in Technical figure 3. This shift to K_G will allow the algorithm, if in an area of low posterior density, to make efficient local transitions toward an area of high probability density.

For an observed dataset $\boldsymbol{S}(\boldsymbol{y})$ and simulated dataset $\boldsymbol{S}(\boldsymbol{y}^*)$, where $\boldsymbol{S} = \{S_1, \ldots, S_O\}$, the Gaussian weighting kernel is computed as:

$$p(\boldsymbol{S}(\boldsymbol{y})|\boldsymbol{S}(\boldsymbol{y}^*), \boldsymbol{\theta}) = K_G\big(\mathrm{d}(\boldsymbol{S}(\boldsymbol{y}), \boldsymbol{S}(\boldsymbol{y}^*))\big) \propto \prod_{i=1}^{O} \exp\Big[-\frac{1}{2}\Big(\frac{\mathrm{d}(S_i(\boldsymbol{y}), S_i(\boldsymbol{y}^*))}{\epsilon_i}\Big)^2\Big] \tag{2.3}$$

Figure 2.5 compares K_U and K_G over the first 10 000 time steps in an ABC-MCMC algorithm targeting the parameters of a bivariate Gaussian distribution, $\boldsymbol{\theta} = \Big[\boldsymbol{\mu}\ \boldsymbol{\Sigma}\Big]$. K_G does not suffer from the same initialization problem as K_U, figure 2.5(d)(f)(h)(j))(l). K_G also has an improved acceptance rate and better mixing when compared to K_U. Figure 2.6 demonstrates that with time both kernels converge to the same posterior after 1,000,000 time steps.

The posterior sampled in both figure 2.5 and 2.6 shows an insensitivity to ρ. This is a direct result of the insufficiency of the selected summary statistics for that parameter. This reinforces that while a smaller ϵ improves the approximation of $p(\boldsymbol{\theta}|\mathrm{d}(S(\boldsymbol{y}), S(\boldsymbol{y}^*)) < \epsilon)$ to $p(\boldsymbol{\theta}|S(\boldsymbol{y}))$, if the summary statistics are not informative such that $p(\boldsymbol{\theta}|S(\boldsymbol{y})) \approx p(\boldsymbol{\theta}|\boldsymbol{y})$, then the parameter inference will be inaccurate. In this case the statistic $\mathrm{cov}(X, Y)$ which should be sensitive to ρ, remains overwhelmed by σ_X and σ_Y. Here the sample correlation, $\bar{\rho}$, would have provided a better summary statistic choice.

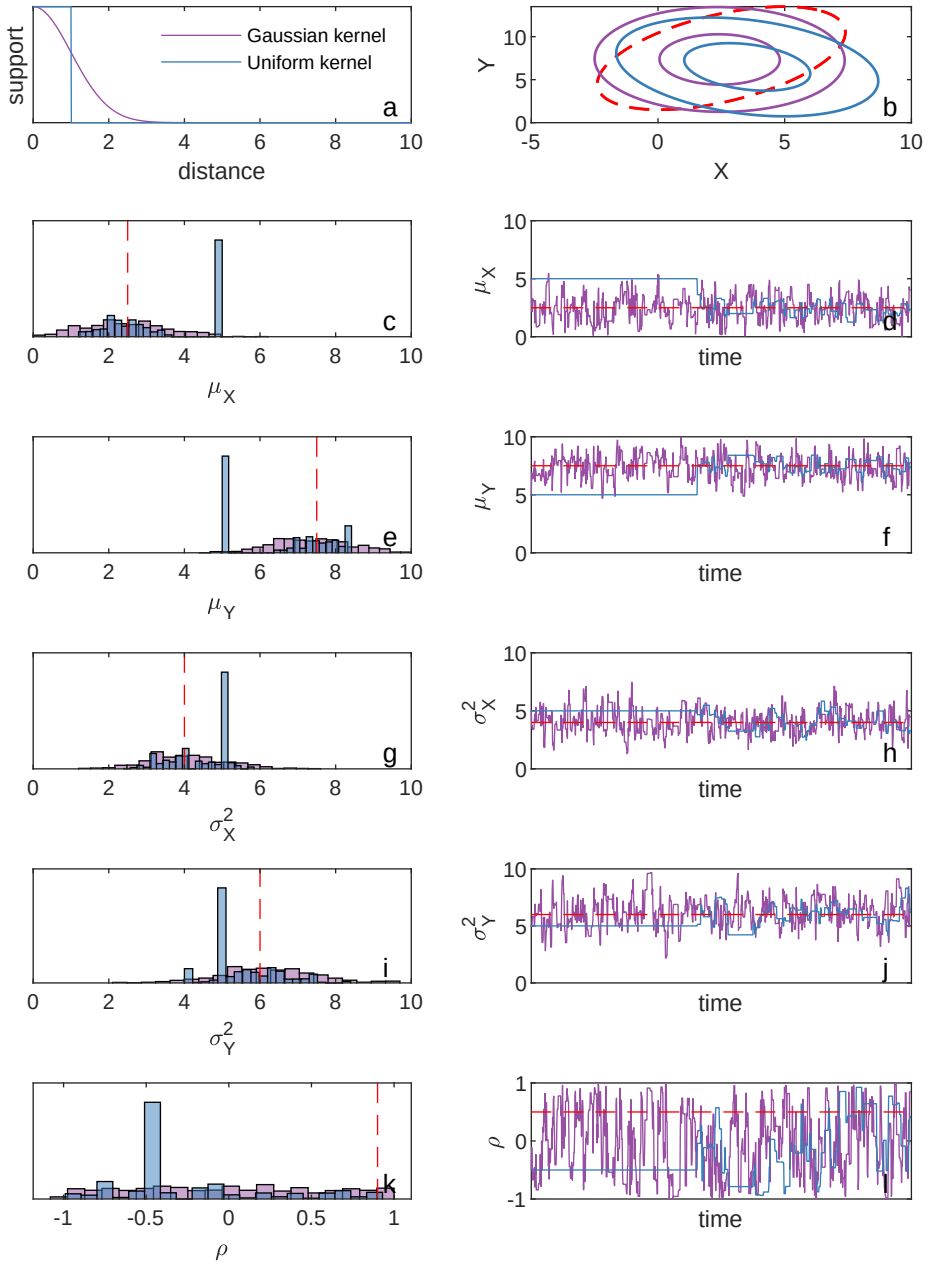


FIGURE 2.5: Comparison between ABC-MCMC with a weighting kernel based on a uniform distribution, K_U, with compact support, and a Gaussian distribution, K_G, with infinite support. Both Markov chains target the ABC posterior for the parameters to a bivariate Gaussian distribution, $\boldsymbol{\theta} = [\boldsymbol{\mu}, \boldsymbol{\Sigma}]$, based on the observed data in figure 2.4(a). The comparison in this figure is limited to the first 10,000 time steps of the respective Markov chains. (a) Plots the support offered by K_U and K_G over distance, d($\boldsymbol{y}$,$\boldsymbol{y}^*$), for $\epsilon = 1$. (b) plots the causative model, red, compared to the mean marginal posterior model for the first 10,000 time steps of the Markov chain based on K_U and K_G. (c),(e),(g),(i),(k) plots the marginal posterior sampled by the Markov chain over the first 10,000 time steps for both K_U and K_G. (d),(f),(h),(j),(l) plots the Markov chain position through time for the first 10,000 time steps. This figure highlights how ABC based on K_G overcomes the initialization problem, seen in the Markov chain traces for K_U, and offers improved acceptance rates and mixing relative to K_U.

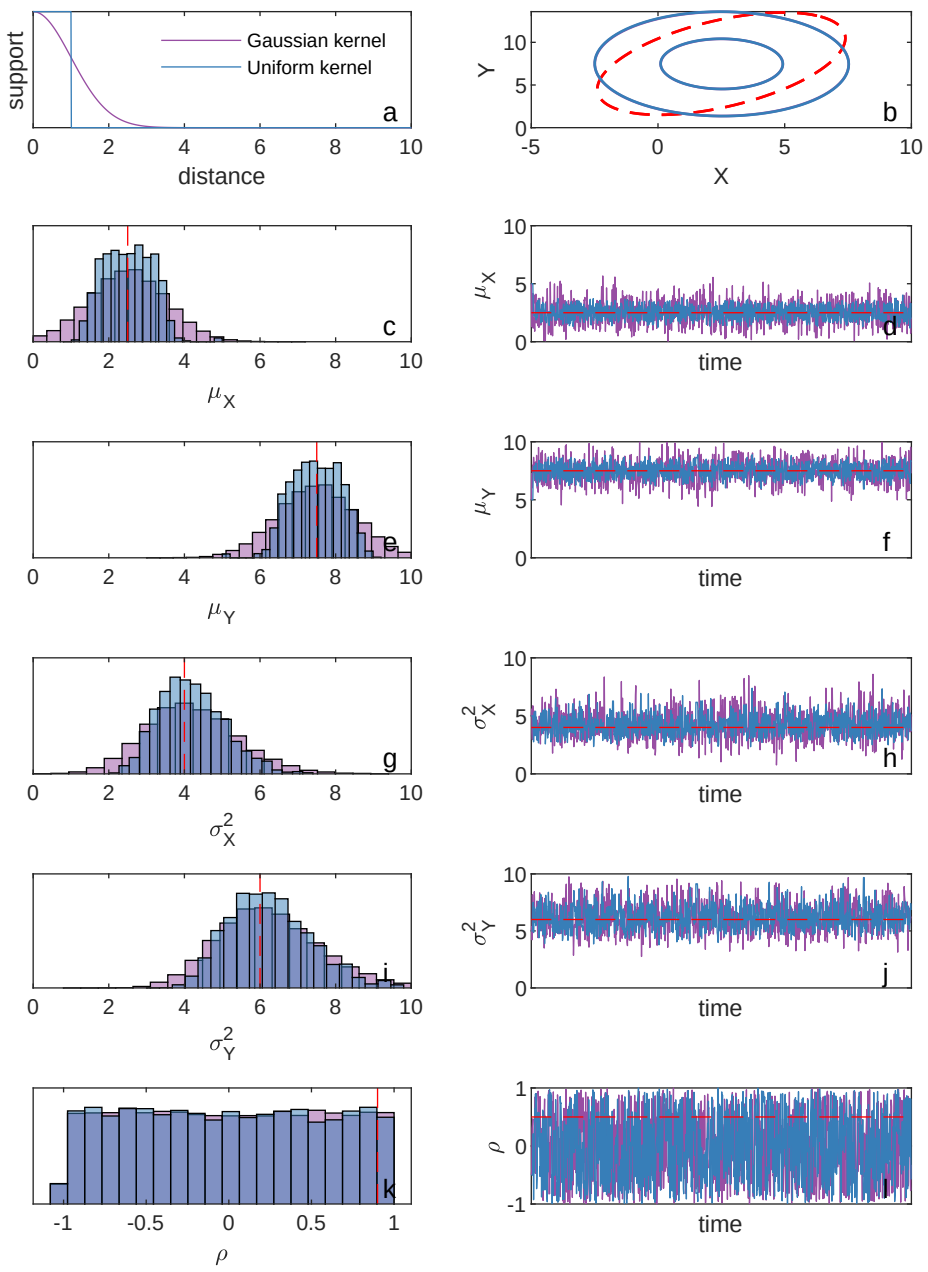


FIGURE 2.6: This figure is a re-creation of figure 2.5 for 1,000,000 time steps. It is a comparison between ABC-MCMC with a weighting kernel based on a uniform distribution, K_U, with compact support, and a Gaussian distribution, K_G, with infinite support. Both Markov chains target the ABC posterior for the parameters to a bivariate Gaussian distribution, $\boldsymbol{\theta} = [\boldsymbol{\mu}, \boldsymbol{\Sigma}]$, based on the observed data in figure 2.4(a). (a) Plots the support offered by K_U and K_G over distance, d($\boldsymbol{y}$,$\boldsymbol{y}^*$), for $\epsilon = 1$. (b) plots the causative model, red, compared to the mean marginal posterior model for the full 1,000,000 time steps of the Markov chain based on K_U and K_G. (c),(e),(g),(i),(k) plots the marginal posterior sampled by the Markov chain over the full 1,000,000 time steps for both K_U and K_G. (d),(f),(h),(j),(l) plots the Markov chain position through time for the full 1,000,000 time steps. This figure highlights how the two methods converge to the same solution, in contrast to figure 2.5. The difference in the marginal posterior of K_U and K_G is a result of the wider support offered by the Gaussian kernel compared to the equivalent uniform kernel (a).

2.4 Toy problem 4: Banana distribution

As a fourth example we design a more challenging problem, one for which there are no immediate or obvious summary statistics. Consider we have $n = 1000$ observations from a 'banana' distribution, Figure 2.7. This is a standard distribution to test the performance of MCMC as it is moderately non-linear and the distribution can be analytically defined (Haario et al., 1999). To define the banana distribution, we start with a 2-dimensional Gaussian distribution:

$$\begin{bmatrix} x & y \end{bmatrix} = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}),\ \boldsymbol{\mu} = \begin{bmatrix} \mu_X, & \mu_Y \end{bmatrix}^T,\ \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{bmatrix} \tag{2.4}$$

The Gaussian co-ordinates x and y are then twisted to produce a more nonlinear target, X and Y, using:

$$X = b_1 x \tag{2.5}$$

$$Y = y/b_1 - b_2(b_1^2 x^2 + b_1^2) \tag{2.6}$$

In total the parameters which define this model are $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$ and the 'bananity' parameters $\begin{bmatrix} b_1 & b_2 \end{bmatrix}$. Here the unknown target parameters are $\boldsymbol{\mu} = \begin{bmatrix} \mu_X & \mu_Y \end{bmatrix}^T$, $\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{bmatrix}$ while $\begin{bmatrix} b_1 & b_2 \end{bmatrix} = \begin{bmatrix} 1 & 1 \end{bmatrix}$.

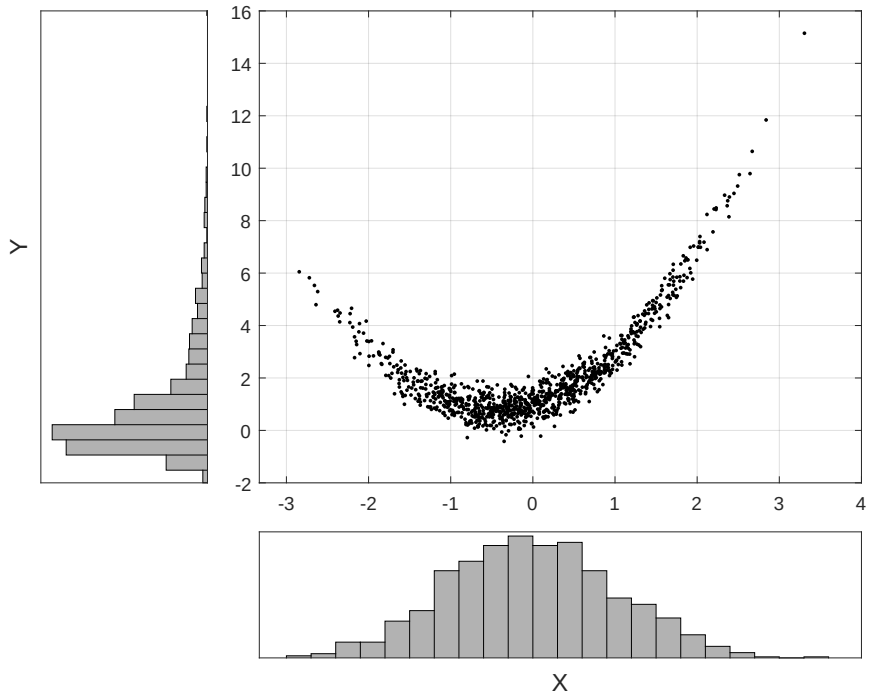


FIGURE 2.7: $n = 1000$ observations from the 'banana' distribution, analytically defined by equations 2.4, 2.5, 2.6. This problem is more challenging as there is no apparent sufficient statistics to facilitate ABC estimation of the unknown parameters.

To apply ABC inference a set of reasonably sufficient statistics is required. In the previous examples it has been possible to use the sample values for each of the unknown parameters, however, this proved to be insufficient in a trial run. In the ABC literature summary statistic selection is the focus of technique development and active research. Blum et al. (2013) and Prangle (2017) offer comprehensive reviews of this research area. The statistics have the same role in ABC as the data does in a traditional likelihood. Hence, there is no need for certain statistics to relate to specific parameters any more than there is the need for certain data points to relate to specific parameters. The key is in identifying a set of statistics which is sensitive to the important and repeatable features of the data, and insensitive to the transient or noisy components. Wood (2010) identifies marginal distribution statistics as being a useful first step. The marginal distribution for our banana observations, X and Y, are plotted in figure 2.7. Here we can see X is a regular Gaussian distribution, while Y appears as a skewed distribution. This motivates the choice of marginal statistics for X as the sample mean, $\bar{\mu_X}$, and sample standard deviation, $\bar{\sigma_X}$. While marginal statistics for Y are sample skewness, $\bar{\gamma_Y}$, as well as $\bar{\mu_Y}$ and $\bar{\sigma_Y}$. However, this does not capture the 'shape' of the joint distribution in 2D dimensions. To capture the bananity of the joint distribution a second degree polynomial is fit to the data, $Y = aX^2 + bX + c$. The co-efficient values a, b and c are then used as the summary statistics. For this example we consider this set, $\left[\bar{\mu_X}\ \bar{\sigma_X}\ \bar{\gamma_Y}\ \bar{\mu_Y}\ \bar{\sigma_Y}\ a\ b\ c\right]^T$, to constitute reasonably sufficient summary statistics for the banana distribution. However, these are not truly *sufficient*, and the effect of using them will be the introduction of some degree of bias into the ABC posterior, relative to the posterior under access to a closed form expression for the likelihood.

In our applications so far the distance metric has been a vector composed of the absolute distance between observed and simulated summary statistics, hence:

$$\mathrm{d}_i(S_i(\boldsymbol{y}), S_i(\boldsymbol{y}^*)) = |S_i(\boldsymbol{y}) - S_i(\boldsymbol{y}^*)| \tag{2.7}$$

is the marginal fit for a given statistic. In previous examples it has not been necessary to consider the scale the chosen statistics vary over and their sensitivity to the unknown parameters. For the most part all statistics have been equally sensitive and varied over a similar scale. Using a rejection scheme, it is possible to first pre-compute all simulated data by sampling the prior, then compute each marginal distance, and before computing the accept/reject step, normalize the set marginal distances for each statistic. However, it is not clear how to best account for this scale and sensitivity when using ABC-MCMC. One solution would be to use variable tolerances which account for this scale and sensitivity, establishing $\epsilon = \{\epsilon_1, \ldots, \epsilon_O\}$. However, this is not a user friendly solution as implementation would invariably involve many repetitions as ϵ is tuned. Other authors, Ratmann et al. (2010), have normalized the marginal weighting term, the weight for each individual statistic, to one. I attempt a transformation in a similar spirit by approximating a term σ_{S_i} which will normalize the spread of distance for each statistic to one. In algorithms distance is replaced by a normalized-distance:

$$\hat{d}_i(S_i(\boldsymbol{y}), S_i(\boldsymbol{y}^*)) = \frac{|S_i(\boldsymbol{y}) - S_i(\boldsymbol{y}^*)|}{\sigma_{S_i}} \tag{2.8}$$

A good sampling algorithm would temper $\boldsymbol{\sigma_S} = \{\sigma_{S_1}, \ldots, \sigma_{S_O}\}$ on the fly. Without such an algorithm $\boldsymbol{\sigma_S}$ must be predefined. A Monte Carlo approximation based on $k = 10,000$ simulations from the prior is used and given the observed data. Each marginal term σ_{S_i} is then defined as the sample standard deviation of the distance, $\mathrm{d}(\boldsymbol{S}(\boldsymbol{y}), \boldsymbol{S}(\boldsymbol{y}^*))$:

$$\sigma_{S_i} = \sqrt{\frac{\sum_{j=1}^{k} (\mathrm{d}_j)^2}{k-1}} \tag{2.9}$$

This does not preclude varying ϵ, however, it does ensure that each marginal ϵ_i is of a simular magnitude.

Figure 2.8 shows the results of an ABC-MCMC algorithm targeting the banana distribution. This leverages a Gaussian kernel, summary statistics as described and uses a normalized distance metric. For this problem only the Gaussian parameters, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are unknown.

For this example I plot the median marginal posterior model. That is, the model which is defined by taking the median of the Markov chain for each individual parameter. One example is figure 2.8(f), here the red dashed lines mark the 50% and 95% confidence intervals of the causative model the data was generated from, while the solid black lines mark the 50% and 95% confidence intervals of the median marginal posterior model. Often the marginal parameter densities are considered as the principal result from Bayesian parameter inference. However, the marginal posterior distributions hide correlations between parameter which may be present in the joint posterior, which is the distribution the Markov chain sampled. To understand the joint banana posterior I plot the correlations between each parameter, figure 2.9, from the Markov chain generated in figure 2.8. This shows there are limited correlations between parameters. As a result, extracting the marginal median is an accurate representation of the median model under the joint distribution. An accurate estimate of the uncertainty can also be found by taking the marginal standard deviation. If there was significant correlation between some θ_1 and θ_2 then it would not make sense to evaluate θ_2 marginally with $\theta_2 = \mu \pm \sigma$. Instead σ would need to be evaluated with respect to a given value for θ_1, in order to have a proper understanding of the uncertainty in θ_2 present in the joint distribution.

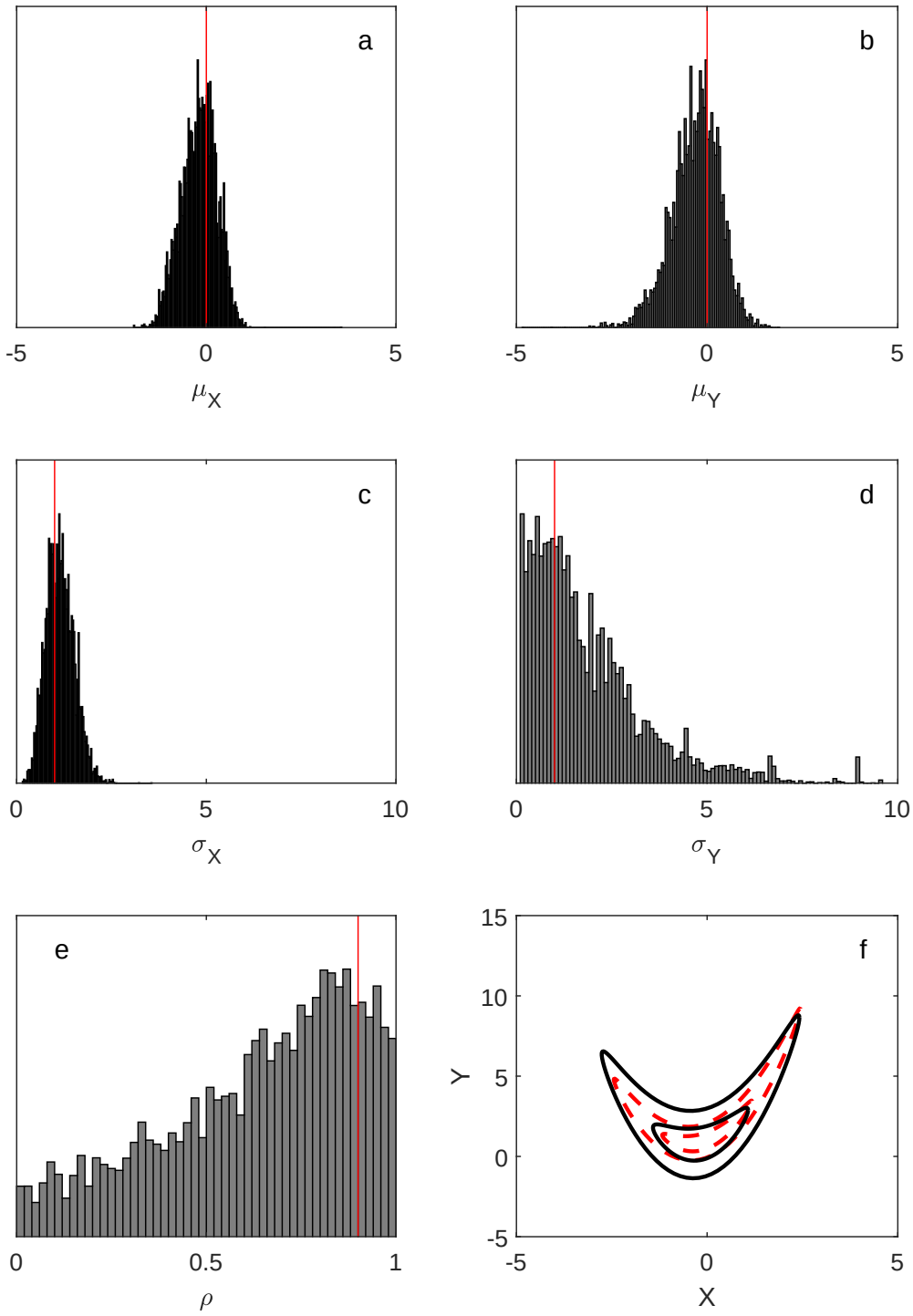


FIGURE 2.8: The marginal posterior distributions for banana parameter inference, as well as the median posterior model (f), in black, compared to the causative model, red dashed lines. Red lines mark the true parameter values on the marginal posterior plots. I justify the legitimacy of extracting the median marginal model as there are limited correlations between the unknown parameters, figure 2.9.

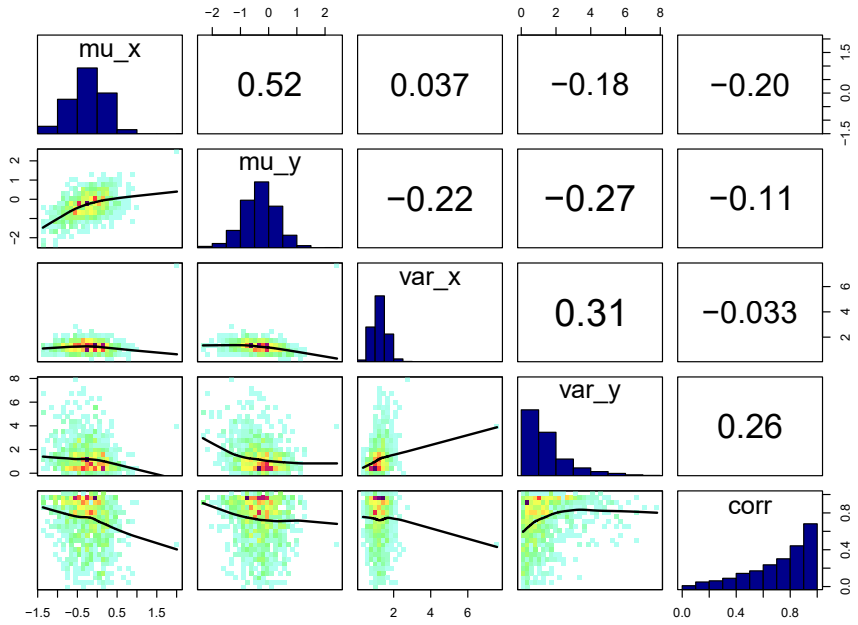


FIGURE 2.9: The panels visualize correlations between parameters in the posterior sampled by the Markov chain in figure 2.8. The diagonal shows the marginal distributions which are also plotted in figure 2.8(a)(b)(c)(d)(e). The lower triangle shows the correlation density between the parameter on the diagonal (red marks higher density) and a polynomial fit to the correlation (black line). The upper triangle shows Spearman's rank correlation coefficients for the correlations in the lower triangle. There is limited correlation between the parameters, as indicated by the low Spearman's rank values and broadly distributed correlation density. This is used as justification for plotting statistics of the posterior marginal distributions as representative of the joint posterior distribution.

The performance of the MCMC algorithm as described by algorithm 5 depends on the users ability to choose a suitable proposal distribution $q(\cdot,\cdot)$. If the proposal distribution is limited to a multivariate Gaussian distribution, then a suitable covariance matrix must be chosen. If the variance for each parameter is too large then the probability of accepting a candidate move will be low, as each step will be erratic and far from the current location. Consider that as the variance for $q(\cdot,\cdot) \to \infty$ we effectively have a Monte Carlo scheme. However, if the variance is too small then the acceptance rate will be very high but the algorithm will fail to explore the full parameter space. We desire both a reasonable acceptance rate and a good exploration of the parameter space. The variance is selected to explore the posterior distribution efficiently. It is very difficult to set the correlation between parameters, the off-diagonal terms in the covariance matrix, a priori. It is also time consuming and difficult to tune the correlation values. Generally, it is simplest to leave the correlations set to zero.

However, such a system is not ideal. Selecting an optimal proposal distribution is non-trivial problem. Here we can turn to theoretical developments to guide a better proposal distribution. Gelman and Roberts (1996) show that when the target distribution is Gaussian, an efficient sampler can be constructed by scaling the proposal covariance to $2.4^2/d$, where d is the number of unknown parameters. In the previous sections it has been best to ignore the correlations between parameters.

However, these correlations can be important to consider, if there are strong parameter correlations in the posterior then sampling acceptance rate will be lowered and convergence undermined. Ideally the algorithm should be able to learn about the posterior on the fly, and if there are correlations between parameters, adjust accordingly. With this in mind Haario et al. (2001) developed Adaptive Metropolis (AM). AM tunes the proposal distribution to the posterior distribution by using the history of the chain generated up to the current time. After some designated time, t, the covariance matrix of the proposal distribution, $\mathbf{\Sigma_q}$, is set to the covariance of the Markov chain.

$$\mathbf{\Sigma_q} = \text{cov}(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_t) s_{AM} + I\upsilon \tag{2.10}$$

Where s_{AM} is the scaling factor, generally $2.4^2/d$, and υ is a small positive number which prevents the covariance from becoming singular. AM sampler efficiency is compared to MH in table 2.1.

Figure 2.11 plots the effect AM has on a sub-optimal choice for a proposal distribution when targeting the ABC posterior for the Gaussian parameters to the banana distribution.

In this example a 4-stage adaptive scheme is used over the length of the chain. Figure 2.10 plots this scheme. Figure 2.11 plots the proposal distribution before the first adaption, red, and after the 4th adaption, blue.

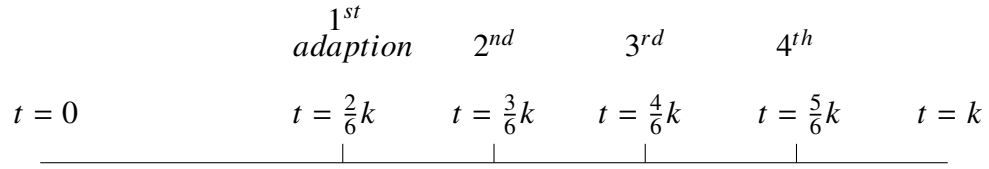


FIGURE 2.10: 4-stage adaptive scheme for a Markov chain of length k

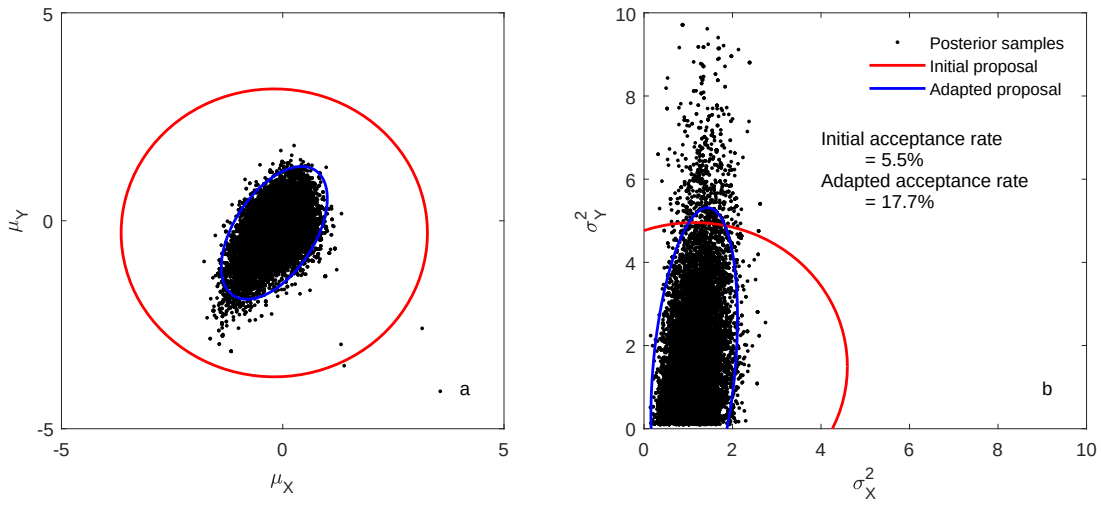


FIGURE 2.11: Adapted proposal distribution compared to a suboptimal initial proposal when targeting the ABC posterior for the Gaussian parameters to the banana distribution. This suboptimal choice was the proposal distribution of the MCMC algorithm in figure 2.8.

Delayed Rejection (DR) can also be implemented to improve sampling efficiency (Mira, 2001). Under DR, when a proposed move is rejected, instead of advancing a time step and retaining the same position, another proposed candidate move is considered. The acceptance probability of the second proposal differs from the MH acceptance probability in order to retain time reversibility of the Markov chain, ensuring that the chain converges to the desired stationary distribution. When a Markov chain remains in the same position over time, the estimates obtained by averaging along the chain become less accurate. Increases to the autocorrelation of the chain increase the variance of estimates based on the chain (Mira, 2001). In this way DR gives more reliable estimates than a standard MH algorithm. A two-stage DR algorithm, as is applied in this text, operates as follows.

At the first stage the acceptance probability is the Metropolis-Hastings acceptance probability:

$$\alpha_1(\theta, \theta^*) = \min\left\{1, \frac{\pi(\theta^*)q_1(\theta^*, \theta)}{\pi(\theta)q_1(\theta, \theta^*)}\right\} \tag{2.11}$$

where π is the target distribution of the algorithm and q_1 is the proposal distribution. If the proposed move to θ^* is rejected then a second candidate move, θ^{**}, can be considered from the second proposal distribution q_2. The acceptance probability for θ^{**} is:

$$\alpha_2(\theta, \theta^*, \theta^{**}) = \min\left\{1, \frac{\pi(\theta^{**})q_1(\theta^{**}, \theta)q_2(\theta^{**}, \theta^*, \theta)[1 - \alpha_1(\theta^{**}, \theta^*)]}{\pi(\theta)q_1(\theta, \theta^*)q_2(\theta, \theta^*, \theta^{**})[1 - \alpha_1(\theta, \theta^*)]}\right\} \tag{2.12}$$

Here I limit implementation to a two-stage DR scheme and use a q_2 with a smaller covariance matrix. q_1 is downscaled to give q_2 by the relation, $q_2 = q_1/s_{DR}$. DR is compared to MH and AM in table 2.1.

Both Adaptive Metropolis and Delayed Rejection can be implemented together to give a Delayed Rejection Adaptive Metropolis scheme (Haario et al., 2006). There are many implementation possibilities given the mixing of AM and DR. A straight forward implementation is favoured, as in Laine

(2008), which retains the principal advantages of each method to give an efficient adaptive algorithm. The first stage proposal is adapted to the chain generated so far, and one scaled down second stage proposal is considered. The second stage proposal inherits the adapted covariance matrix of the first stage proposal, only the variances are reduced. DRAM is compared to MH, AM and DR in table 2.1.

When running a 'one-size fits all' generic MCMC algorithm, the standard procedure is to propose a candidate move which updates each and every parameter value. If we are at the position $\boldsymbol{\theta_{t-1}}$ in the parameter space, then the proposal distribution suggests a candidate move:

$$\boldsymbol{\theta_t} = \mathcal{N}(\boldsymbol{\theta_{t-1}}, \boldsymbol{\Sigma_q}) \tag{2.13}$$

However, this is not the only option. Another choice is to update a single parameter per time step. The choice to update one, or all parameters are end members on a spectrum of updating some sub-set of all parameters. This technique is referred to as parameter blocking (Roberts and Sahu, 1997; Sargent et al., 2000). Much like the variance of $\boldsymbol{\Sigma_q}$, there is a trade-off between acceptance rate and speed of parameter space exploration encoded into this decision. Table 2.1 compares the acceptance rate of a MH algorithm where all parameters are updated at once, and an MH algorithm where a single parameter is updated per time step. A significant increase in acceptance rate is made through this change. However, while acceptance rate is increased, the speed of parameter space exploration is slowed. Effective parameter blocking requires problem specific implementation and tuning.

Sampling method	Acceptance rate (%)
Metropolis-Hastings (MH)	5.43%
Adaptive Metropolis (AM)	13.35%
Delayed Rejection (DR)	23.04%
Delayed Rejection Adaptive Metropolis (DRAM)	36.42%
Single parameter update	43.7%

TABLE 2.1: Comparison of sampler efficiency estimating the Gaussian parameters to the banana distribution. The sampled posterior is the same for each method. The AM and DRAM acceptance rate is taken over the full length of the chain. Each chain is of length t = 1,000,000.

While ABC can target the underlying Gaussian parameters which control the resulting banana distribution, there is also access to the probability density of a banana distribution for a fixed set of Gaussian parameters. For the causative model, causative in the sense that it produced the data in figure 2.7, there is access to the probability density in $X - Y$, equations 2.5 and 2.6. Given this access it is possible to run a Markov chain to discover the causative distribution. Figure 2.12(a) plots a Markov chain with the stationary distribution of the causative model, black dots, as well as the 50% and 95% confidence intervals of the causative model. The Markov chain can be used for kernel density estimation of the causative model, figure 2.12(b). It is then possible to compare the estimation of the causative model which results from estimating the underlying Gaussian parameters

from observed data using ABC, and estimation of the causative model through access to its closed form density using MCMC. Figure 2.13 compares the estimation of the causative model based on a standard MH algorithm (a), AM (b), DR (c), DRAM (d), and ABC-MH (e), ABC-AM (f), ABC-DR (g), ABC-DRAM (h).

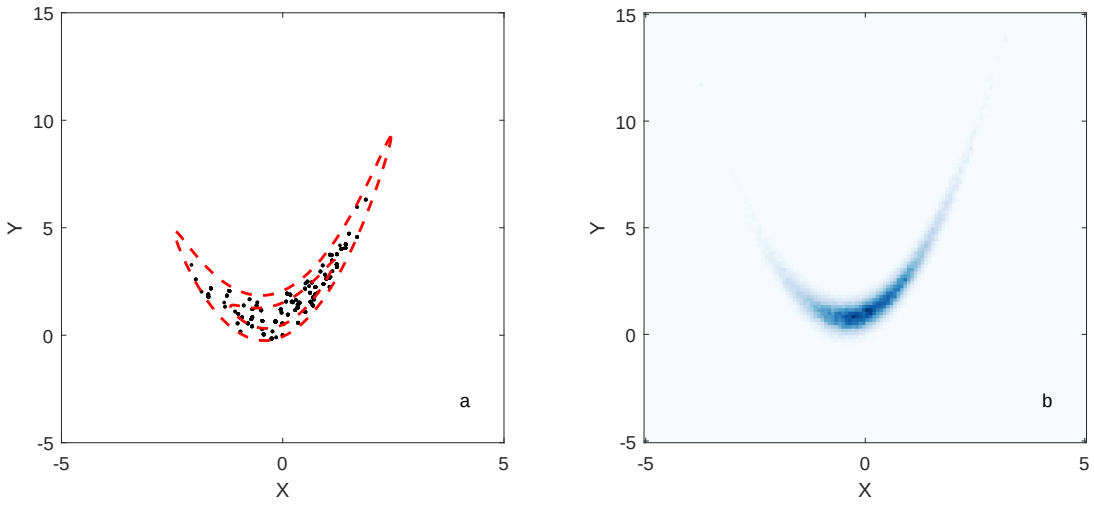


FIGURE 2.12: A Metropolis-Hastings algorithm sampling the causative model. (a) The MCMC samples, black dots, to the causative model. Red dashed lines mark the 50% and 95% confidence intervals of the causative model. (b) kernel density estimation based on the MCMC samples. Kernel density estimation in 2d is based on the diffusion algorithm of Botev et al. (2010).

Throughout this example a strong trend has emerged. The acceptance rate of a Markov chain based algorithm targeting the ABC posterior is strongly tied to the tolerance. This shows the effect of tolerance in a rejection scheme is persistent, and generalizes to more advanced samplers. Increasing the tolerance increases the acceptance rate, while decreasing the tolerance decreases the acceptance rate. With this in mind, the improvements to the acceptance rate of ABC-MCMC offered by AM, DR, single parameter updates and blocking are very important for sampling the most accurate posterior possible. With each improvement to acceptance rate the tolerance can be lowered while retaining a reasonable acceptance rate, around 5-30%. So far all comparisons have been done using a tolerance $\epsilon = \left[0.625\ 0.625\ 0.625\ 0.625\ 0.625\ 0.125\ 0.625\ 0.625\right]^T$ for the statistic set $\left[\bar{\mu_X}\ \bar{\sigma_X}\ \bar{\gamma_Y}\ \bar{\mu_Y}\ \bar{\sigma_Y}\ a\ b\ c\right]^T$. This tolerance is used to keep ABC-MH, the worst performing sampler, at a reasonable acceptance rate (5%). However, using ABC-DRAM with single parameter updates it is possible to lower the tolerance to $\epsilon = \left[0.01\ 0.01\ 0.01\ 0.01\ 0.002\ 0.01\ 0.01\ 0.01\right]^T$ while still maintaining a reasonable acceptance rate. Figure 2.14 plots this scenario. With ABC-DRAM and single parameter updates the acceptance rate is 7% under this tolerance. While ABC-MCMC sampling, with all parameter updates, becomes computationally infeasible with an acceptance rate <0.3%.

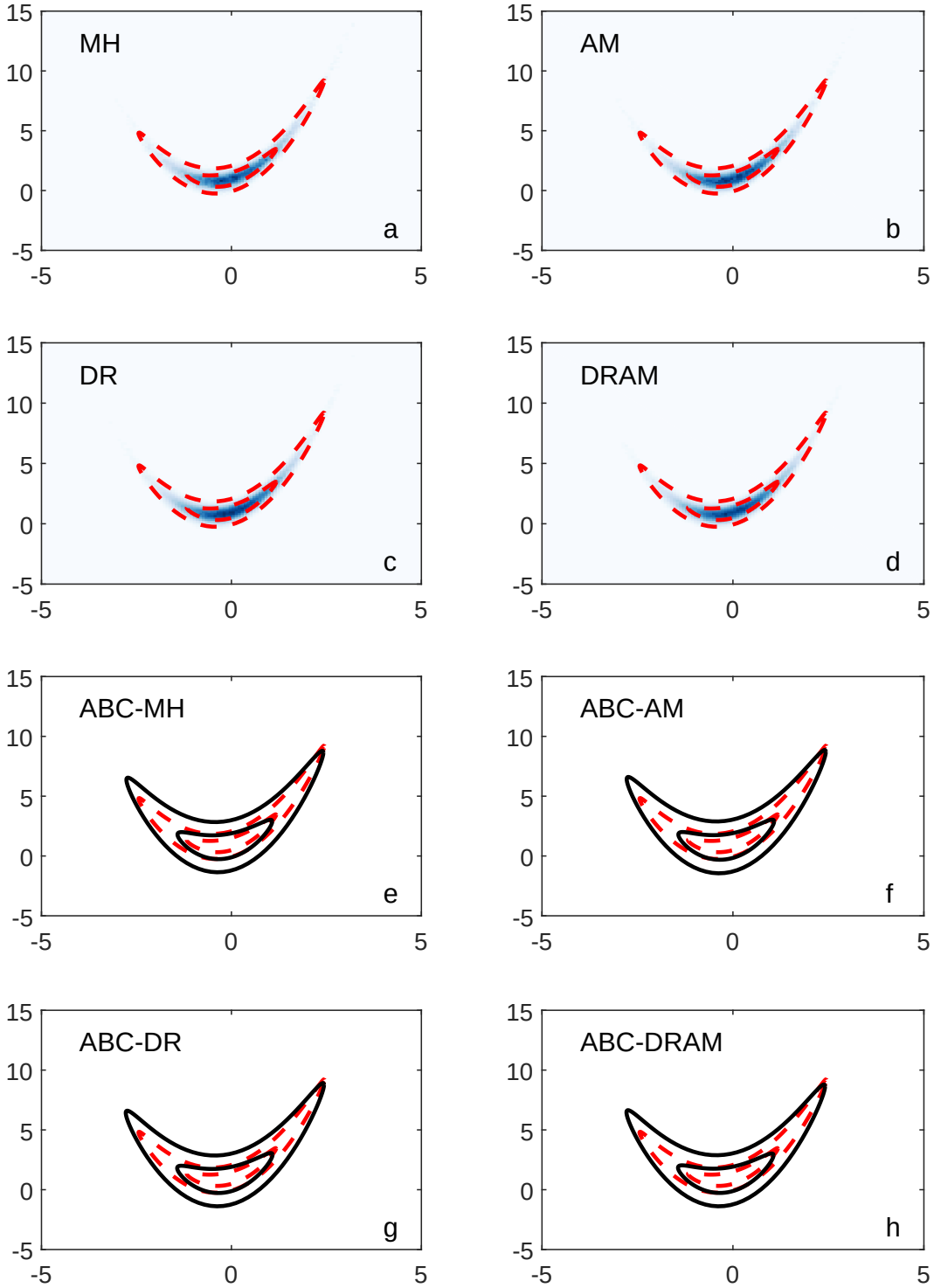


FIGURE 2.13: Comparison of estimation of the causative banana model based on access to the analytical formula, given fixed values for all parameters, (a,b,c,d) and no access to the analytical formula (e,f,g,h), simply simulation to find the Gaussian parameters.

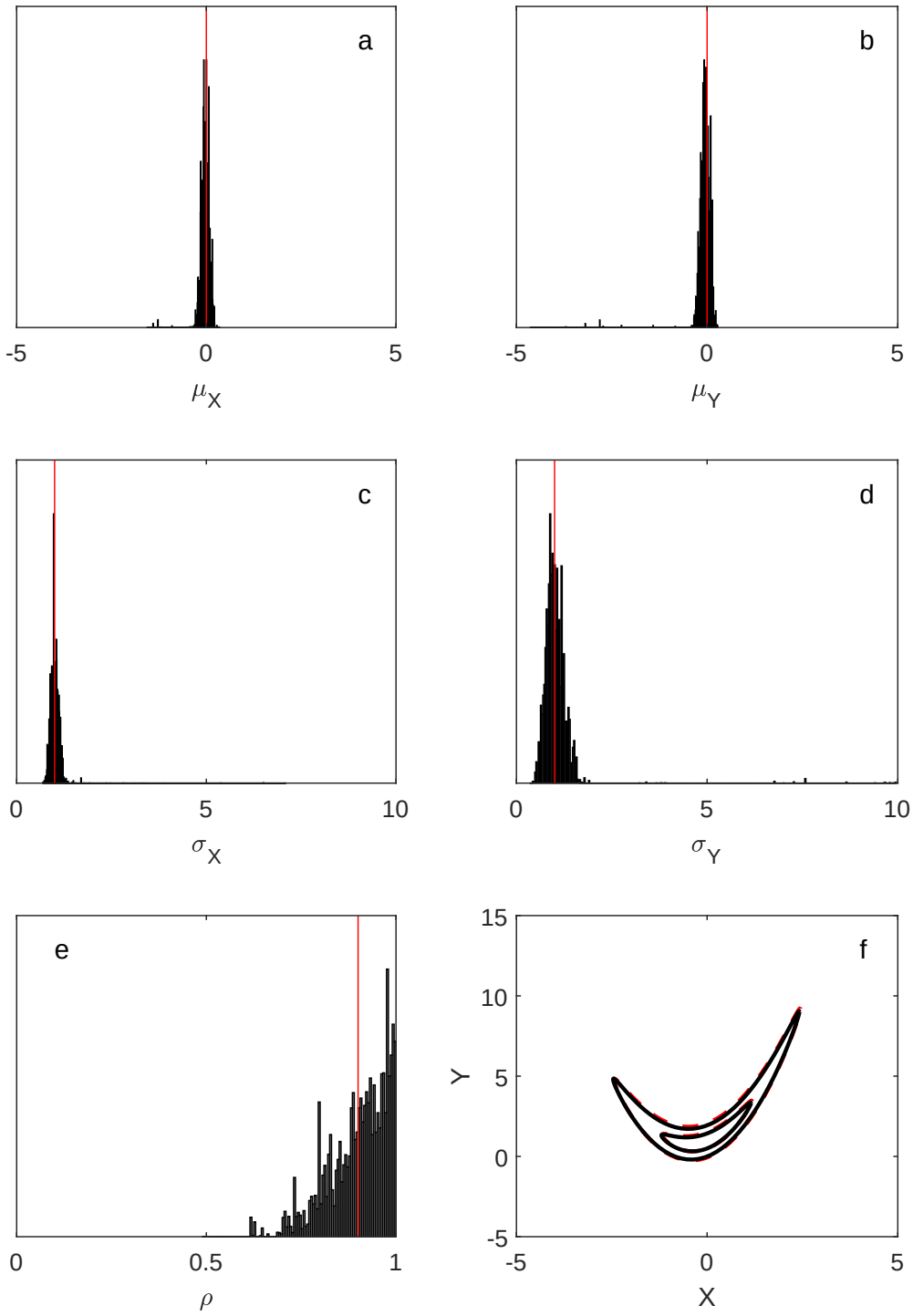


FIGURE 2.14: Banana parameter estimation based on ABC-DRAM with single parameter updates. The tolerance is lowered, see text for tolerance values. By lowering the tolerance inference is more accurate. Compared to 2.8 the marginal posteriors are more tightly clustered around the true solution and the median marginal model, black lines of (f) fits the causative model, red dashed lines, closely.

Chapter 3

Synthetic geophysical experiments

This chapter will outline a series of experiments which will test the main postulate of this thesis, that ABC can offer some improvement over traditional likelihood based Bayesian inference for geophysics. There are many potential avenues to pursue 'improvement'. For example, ABC opens parameter inference to models where there is no closed form expression for the likelihood. This problem can stem from joint probabilistic formulations of multiple datasets with dependent, non-Gaussian uncertainties. In this work, however, I focus on other advantageous aspects of ABC, namely its diagnostic potential. Diagnostic parameter inference, which makes use of the information about the model contained in the simulated data, has been developed and applied (Ratmann et al., 2009; Vrugt and Sadegh, 2013). Here diagnostic parameter inference for geophysics is explored with the goal of improving the convergence rate of an inversion. The benchmark for this test will be a MCMC algorithm targeting a posterior distribution with an analytically-defined likelihood. Currently, there are many geophysical parameter inference problems where a probabilistic approach is required due to a trade-off between parameters, a non-unique parameter space or the availability of prior information. One example is an inversion for the thermal-chemical state of the lithosphere based on multiple different sets of constraining data (Afonso et al., 2013a,b). Currently, exploring this 3D structure with high resolution is limited as the problem must be kept within the computational limits of the posterior sampling algorithm. If high resolution, many unknown parameters, and large scale (e.g. the whole mantle) inversions are to be implemented on problems which require a probabilistic formulation, then methods are required which will efficiently find and define the set of low misfit models. Here a diagnostic ABC inversion may be able to improve the search for high probability (low misfit) models by taking into account more of the information available within the simulated dataset.

As with the previous chapter, the code to produce all figures in this section can be found at https://github.com/tomconnell/approximate-bayesian-tomography.

3.1 Crustal density inversion

As a first experiment, I consider an inversion for crustal density (ρ) with a vertical gravity anomaly dataset (Δg) for a 2D discretized subsurface (Blakely, 1996, p.184-195,378). The dimensionality of the parameter space is kept modest, an 8x4 grid, with an observed data point above each column. The grid is defined over a 160 km by 40 km area. Density is bounded between 2-3.5 g/cm^3, the limits for which Brocher (2005) define an empirical relationship between density and compressional-wave

velocity (V_p). The 'true model', Figure 3.1, which will be the target of our inversion scheme, is kept smooth to allow a prior term, $p(\boldsymbol{\theta})$, to be set for smoothness which will limit the inversion to a unique solution. The prior term is:

$$-\log\big(p(\boldsymbol{\theta})\big) = \sum_{i=1}^{N}\Big(\sum_{j}(\rho_i - \rho_j)^2\Big) \tag{3.1}$$

where j describes all blocks in immediate contact with the given block, i. The edge effect for the 2D subsurface grid is compensated by adding the vertical gravity anomaly which will result from extending the grid by a width of one on both sides, tripling the total domain width, with a density which is the average of the parameter space, 2.75 g/cm^3. The model is assumed to contain no modelization uncertainty and the data is known to be contaminated with noise defined by $\sigma^{\mathcal{D}} = \mathcal{N}(0, \sqrt{2})$.

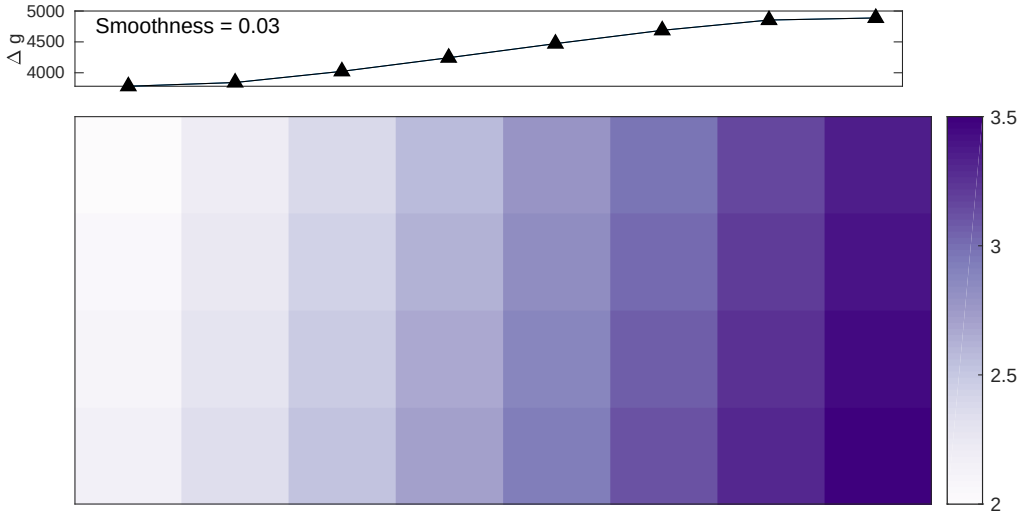


FIGURE 3.1: The 'observed data', vertical component of gravity (Δg), and 'true model', a 2D density (g/cm^3) slice, which will be the target of our synthetic geophysical experiments to compare ABC to likelihood based Bayesian inference. A smoothness value, -log($p(\boldsymbol{\theta})$), equation 3.1, for the true model is plotted for reference to later solutions.

The ABC approach will be benchmarked against a MCMC algorithm targeting an analytically defined posterior. Given there is no modelization uncertainty and $\sigma^{\mathcal{D}}$ is known, the negative log-likelihood can be defined by, equation 1.9:

$$-l(\boldsymbol{\theta}^*|\boldsymbol{y}) = \frac{\sum_{i=1}^{M}(y_i - y_i^*)^2}{(\sigma^{\mathcal{D}})^2} \tag{3.2}$$

For this example, only MCMC is considered (i.e. no AM, DR or DRAM). This maintains a consistent sampling benchmark for comparison to the ABC scheme. The proposal distribution for both MCMC sampling of the analytical distribution and ABC-MCMC is held constant for all runs as $q(\cdot,\cdot) = \mathcal{N}(0, I \cdot 250)$. The parameter space is subdivided into eight 2x2 blocks. At each time step a random block is selected in the subsurface and updated via $q(\cdot,\cdot)$. As demonstrated in table 2.1, the number

of parameters updated at each time step impacts the acceptance rate and the rate of space exploration of the chain. To keep inference comparable, both the analytical scheme and ABC scheme will update 4 parameters per time step via the proposal distribution. Likewise, for each experiment considered in this section the Markov chain starting position is the same for both the analytical MCMC and ABC-MCMC. The starting position for both is set by sampling a uniform distribution, $\mathcal{U}(2, 3.5)$, to define each unknown parameter.

A set of summary statistics to describe the data in Figure 3.1 is needed to proceed with ABC. A comprehensive description of the data is defined by the coefficients to a linear model, m and b, the sample mean, $\bar{\mu_{\Delta g}}$, sample standard deviation, $\bar{\sigma_{\Delta g}}$, and a fail-safe residual term, $\sum R$, for the difference between a simulation and the observed data for a given $\boldsymbol{\theta^*}$, $R = |\boldsymbol{y} - \boldsymbol{y^*}|$. Initial testing showed the set of statistics $\left[\bar{\mu_{\Delta g}}\ \bar{\sigma_{\Delta g}}\ m\ b\ \sum R\right]^T$ comprehensively describes $\boldsymbol{y}$. As with section 2.4, it is convenient to normalize the spread of each marginal distance term, $\mathrm{d}_i(S_i(\boldsymbol{y}), S_i(\boldsymbol{y^*}))$, equation 2.7, to one. This accounts for the different scales and sensitivities of the summary statistics to the unknown parameters. The normalization term $\boldsymbol{\sigma_S}$ is approximated from 10,000 Monte Carlo simulations over parameter range 2-3.5 g/cm^3. The tolerance is closely related to the acceptance rate of the algorithm. Given the tolerance is tightly tied to the acceptance rate, the tolerance is tuned to balance accuracy of the posterior and acceptance rate.

Sisson and Fan (2010) highlight that acceptance rate and mixing can be improved by increasing the number of simulated datasets, $\boldsymbol{y^*}$, for each set of model parameters, $\boldsymbol{\theta^*}$. This technique is adopted in applications with large uncertainty in $\boldsymbol{g_s}$ (Ratmann et al., 2009; Wood, 2010). It has a stabilizing impact when the uncertainty in each simulation is large. The set of simulated data make up a Monte Carlo approximation to the true value of the ABC likelihood approximation ($p(\boldsymbol{y}|\boldsymbol{y^*}, \boldsymbol{\theta})$) which would be found as the number of simulated datasets tends to ∞. The estimate of the ABC likelihood approximation for a given set of parameters becomes less noisy as the number of simulated datasets increases. This technique is used for $k = 100$ simulations from $\boldsymbol{g_s}$ in the ABC gravity inversion. Since the uncertainty associated with the measurement process is additive and fast to simulate, there is little computational cost in increasing the number of simulated datasets, $\boldsymbol{y^*} \sim \boldsymbol{g}_s = \boldsymbol{g}(\boldsymbol{\theta}) + e^{\mathcal{D}}$, as the deterministic physical model, $\boldsymbol{g}(\boldsymbol{\theta})$, does not need to be recomputed for each new realization. The distance is then computed based the sample mean of $\boldsymbol{y^*_{1:k}}$, $\boldsymbol{d} = |\boldsymbol{S}(\boldsymbol{y}) - \boldsymbol{S}(\bar{\boldsymbol{\mu}_{\boldsymbol{y}^*}})|$.

The diagnostic ABC scheme is free to open the likelihood, consider the information available, and use that to make an informed next step in the Markov chain. The ABC scheme moves away from randomly selecting 2x2 blocks in the subsurface to update, as is done in the analytical MCMC scheme. Instead, the update is selected to ensure that it is occurring in a location where the current fit of the simulated data to the observed data is poor. A discrete PDF is defined over each data point which is proportional to the misfit:

$$p(\text{update}) \propto \text{misfit} = (\boldsymbol{y} - \boldsymbol{y^*})^2 \tag{3.3}$$

A random sample from this PDF is then drawn at each time step to determine where the parameter

updates will occur. Once a data point is selected the 4 parameters directly below are updated. The effectiveness of dynamically selecting and localizing the updates relies on our physical intuition about the relationship between the unknown model parameters and the resulting data. In this case, each simulated data point is most sensitive to the parameters which are directly below it, and by focusing model updates on regions which are poorly fitting, the most is made from each move during the initial phase of the algorithm where local transitions are made in search of an area of high posterior density. This change has ideas similar to the way gradient-based linear optimization methods work, however, here I simply rely on physical intuition within a Bayesian framework.

As a next step in defining a diagnostic ABC inversion, the proposal distribution is directed to bias the parameter updates in the direction which the simulated data needs to move to lower the misfit. For example, if a data point is selected based on sampling the discrete PDF equation 3.3 and it is found the simulated data point is too large, i.e. the value $y^*_{\Delta g} - y_{\Delta g}$ is positive, then the proposal for the column of parameters at this update step will be biased towards a density lower than the current parameter value. In this application a truncated Gaussian distribution, figure 3.2, is used to bias the proposal updates. For the example where the currently selected data point is too large, the update can be biased toward a lower density by sampling a truncated Gaussian distribution, centered at the current chain location, and truncated at some position greater than the current chain location. A truncated Gaussian is convenient to implement as detailed balance is satisfied without computing the ratio $\frac{q(\boldsymbol{\theta}^*,\boldsymbol{\theta}_{t-1})}{q(\boldsymbol{\theta}_{t-1},\boldsymbol{\theta}^*)}$ at each time step in the Markov chain.

The analytical definition for the truncated normal distribution, $f(x)$, used to direct the parameter updates is:

$$f(x) = \begin{cases} 0 & x < a \\ \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(- \frac{(x-\mu)^2}{2\sigma^2} \right) & a \leq x \leq b \\ 0 & x > b \end{cases} \tag{3.4}$$

Here a is the lower cut-off and b is the upper cut-off. In this application the negatively truncated distribution has $a = -\frac{1}{2}\sigma$ and $b = \infty$. While the positively truncated distribution has $a = -\infty$ and $b = \frac{1}{2}\sigma$. For our application μ is the current chain location and $\sigma = 250$.

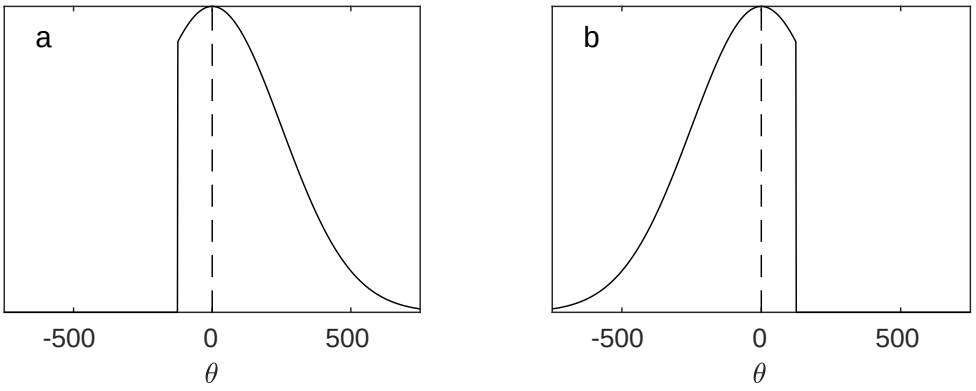


FIGURE 3.2: The proposal distributions which are used to either positively or negatively bias the parameter update. The negatively truncated Gaussian distribution (a) is used when the parameter values are deemed too low. The positively truncated Gaussian distribution (b) is used when the parameter values are deemed too high. Equation 3.4 defines the truncated Gaussian distribution. During application the distributions are centered on the current chain location and $\sigma = 250$. The cut-off a and b are selected to set the bias at ~70% to ~30%.

Figure 3.3 plots a comparison between the misfit, equation 3.3, for the analytical scheme described above and the diagnostic ABC scheme during the initial phase (first 5,000 time steps) of their respective Markov chains. Exploiting the diagnostic properties of ABC by dynamically selecting and directing the parameter updates increases the rate at which the ABC algorithm moves to a low misfit model in comparison to the likelihood based MCMC scheme. The integrated auto-correlation time (τ), figure 3.3, estimates the asymptotic variance of Markov chain based estimates relative to a sampler which draws i.i.d samples of the posterior. This result can be interpreted as the diagnostic ABC algorithm drawing more i.i.d samples from the ABC posterior relative to the MCMC sampling of the analytically-defined posterior. Figure 3.4 and 3.5 plot the solution obtained, mean marginal model, for each full Markov chain (100,000 time steps).

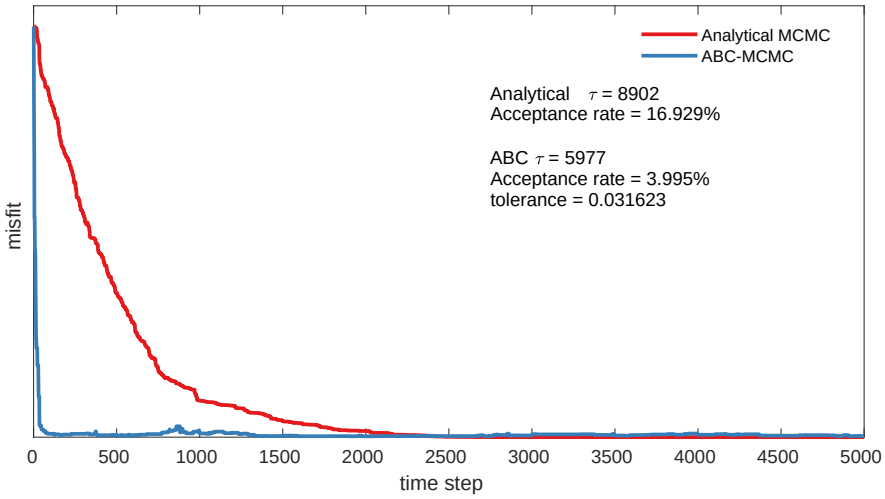


FIGURE 3.3: The misfit, l2-norm, equation 3.3, for the chain state during the initial phase (first 5,000 time steps) for an analytically defined posterior compared to diagnostic ABC. The 'true model' and observed data is plotted in figure 3.1. Both posteriors, analytical and ABC, use a parameter space bound between 2-3.5 g/cm^3 and a prior smoothness defined by equation 3.1. The speed increase in finding low misfit models is a result of opening the likelihood and using the information contained in each iteration to select and direct the next update. The details of the full chain run, 100,000 time steps, are also displayed on the figure. τ is the integrated auto-correlation time.

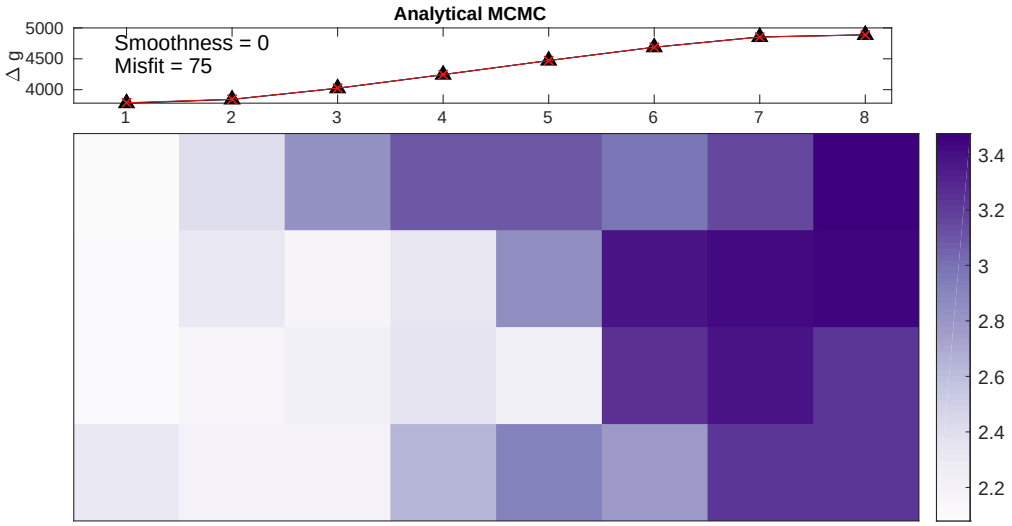


FIGURE 3.4: The mean of the marginal posterior, $p(\boldsymbol{\theta}|\boldsymbol{y})$, defined by the prior equation 3.1 and likelihood 3.2 targeting the 'true model' and observed data of figure 3.1. The simulated data generated by this 'solution' is plotted, red, compared to the observed data, black. The smoothness, equation 3.1, and misfit, equation 3.3, for this model are displayed alongside the data. The misfit during the initial phase for this chain is plotted in figure 3.3. This model can be compared to the equivalent for the diagnostic ABC inversion, figure 3.5.

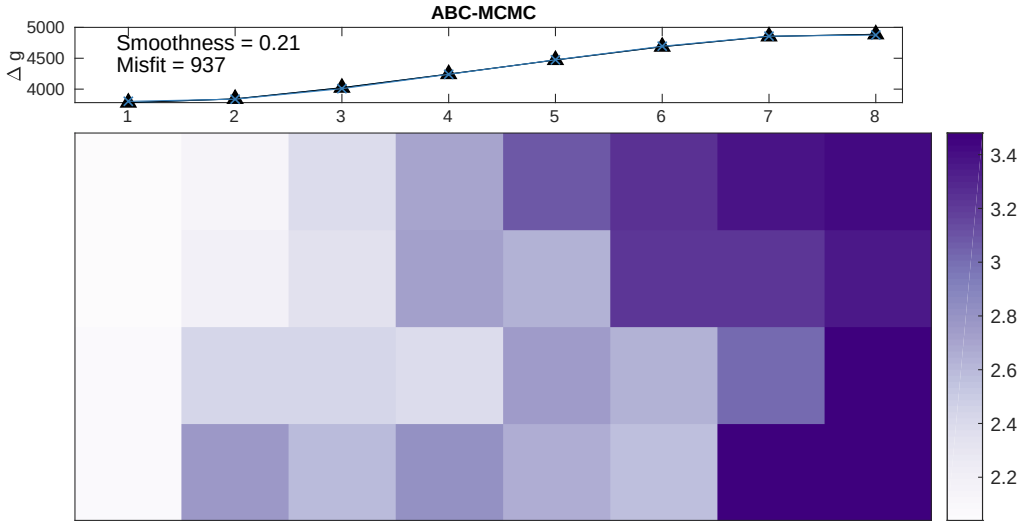


FIGURE 3.5: The mean of the marginal ABC posterior, $p_{ABC}(\boldsymbol{\theta}|\boldsymbol{S}(\boldsymbol{y}))$, targeting the 'true model' and observed data of figure 3.1. The simulated data generated by this 'solution' is plotted, blue, compared to the observed data, black. The smoothness, equation 3.1, and misfit, equation 3.3, for this model are displayed alongside the data. The misfit during the initial phase for this chain is plotted in figure 3.3. This model can be compared to the equivalent for the analytically defined posterior, figure 3.4.

3.2 Crustal density joint inversion

As a second experiment I extend the inversion for crustal density (ρ) for a 2D discretized subsurface to a joint inversion with vertical gravity anomaly (Δg) and travel times (Δt) for compressional-waves V_p. Both datasets share the same underlying parameter space in ρ. The calculation for V_p is made through the empirical relationship for V_p as a function of ρ defined by Brocher (2005):

$$V_p\ (km/sec) = 0.8228\rho^5 - 9.1819\rho^4 + 37.0.83\rho^3 - 63.064\rho^2 + 39.128\rho \tag{3.5}$$

This relationship is valid for crustal rocks with ρ in the range of 2-3.5 g/cm^3. A uniform prior distribution bounds the parameter space, $p(\boldsymbol{\theta}) = \mathcal{U}(2, 3.5)$. As with the first example the 'true model' is kept smooth, figure 3.6. The parameter space is expanded to a 20x6 grid over the same 160 km by 40 km area. There is a Δg measurement made above each of the 20 grid columns. Both the Δg and Δt datasets are contaminated with measurement uncertainty defined by $\sigma^{\mathcal{D}} = \mathcal{N}(0, \sqrt{2})$. As before, our ABC-tomography algorithm will be compared to a MCMC algorithm targeting an analytically-defined posterior. The likelihood for this problem is based on 1.23, with a negative log-likelihood defined as:

$$-l(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{c(\boldsymbol{y}_{\Delta g} - \boldsymbol{y}^*_{\Delta g})^2}{(\sigma^{\mathcal{D}})^2} + \frac{(\boldsymbol{y}_{V_p} - \boldsymbol{y}^*_{V_p})^2}{(\sigma^{\mathcal{D}})^2} \tag{3.6}$$

where c is a constant defined so as the contribution to the negative log-likelihood of both datasets is of the same order of magnitude. The proposal distribution is held as $q(\cdot,\cdot) = N(0, I \cdot 250)$ however, given

the increased parameter space the subsurface is divided into 20 2x3 blocks. A random block is selected for update at each time step in the analytical scheme. The starting position is initialized from $\mathcal{U}(2, 3.5)$ and once again held constant between both the analytical Markov chain and ABC-tomography.

I once again seek a comprehensive set of summary statistics to describe $\boldsymbol{y_{\Delta g}}$ and $\boldsymbol{y_{V_p}}$. The set $\left[\bar{\mu_{\Delta g}}\ \bar{\sigma_{\Delta g}}\ m\ b\ \sum R\right]^T$ is used for $\boldsymbol{y_{\Delta g}}$. Given $\boldsymbol{y_{V_p}}$ is a vector of arrival times for all sources (12 sources in this application), instead of fitting statistics to the larger $\boldsymbol{y_{V_p}}$, $\boldsymbol{y_{V_p}}$ is decomposed to the set of data for each source, 1 to l, where l is the number of sources. The arrivals from each individual source are then described by $\left[m\ b\ \sum R\right]^T$. The weighting kernel for $\boldsymbol{y_{V_p}}$ then becomes:

$$p(\boldsymbol{y_{V_p}}|\boldsymbol{y^*_{V_p}}, \boldsymbol{\theta^*}) = \prod_{i=1}^{l} p(\boldsymbol{y_i}|\boldsymbol{y_i^*}, \boldsymbol{\theta^*}) \tag{3.7}$$

A normalization term $\boldsymbol{\sigma_S}$ is approximated for both datasets from 10,000 Monte Carlo simulations over the prior parameter range 2-3.5 g/cm^3.

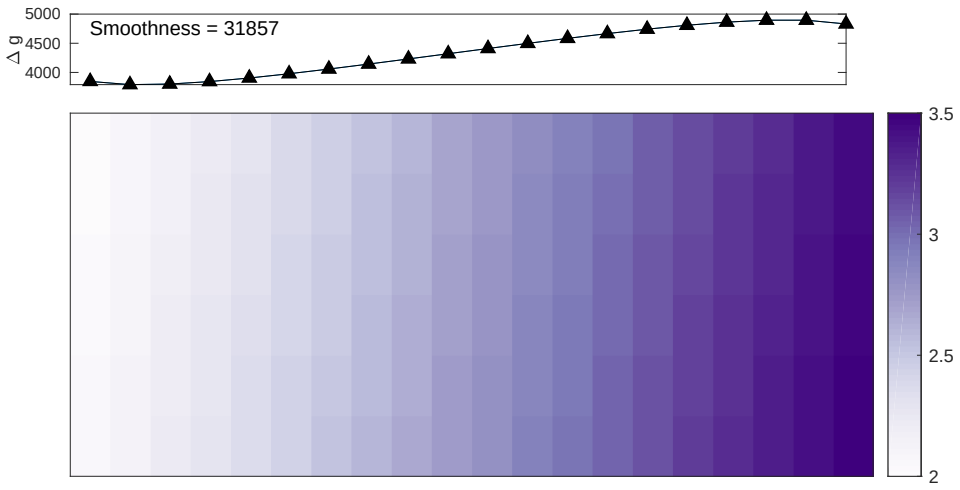


FIGURE 3.6: The 'observed data', vertical component of gravity (Δg), and 'true model', a 2D density (g/cm^3) slice, which will be the target of our synthetic geophysical experiments to compare ABC to likelihood based Bayesian inference. A smoothness value, $\log(p(\boldsymbol{\theta}))$, equation 3.1, for the true model is plotted for reference to later solutions.

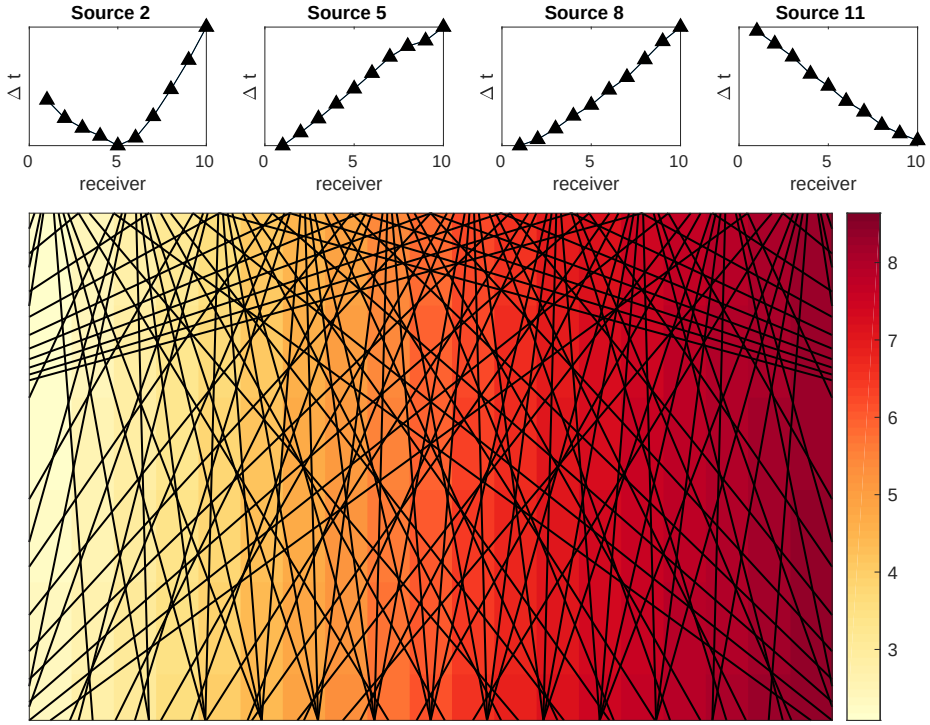


FIGURE 3.7: The 'true model', a 2D Vp (km/s) slice, converted from ρ by the empirical relationship defined by equation 3.5. The black lines across the model define the rays from 12 sources on their path to 10 receivers. The arrival time Δt of these rays at the receivers defines the 'observed data'. A sample of observed data is plotted for 4 sources. The sources are numbered 1-12 starting from the top left and going to the top right, with 4 sources per side.

The ABC-tomography inversion adopts the same diagnostic scheme used in the previous section. At each time step in the chain, a discrete PDF is defined by the misfit, equation 3.3, between simulated data based on the current chain location, $\boldsymbol{y}^* \sim p(\boldsymbol{y}^*|\boldsymbol{\theta_{t-1}})$, and observed data, $\boldsymbol{y}$. A random sample is then drawn, distributed according to this discrete PDF, to determine a column of parameters to update. The proposal distribution is then biased to direct the parameter updates based on whether the location sampled from the discrete PDF is deemed to be a poor fit due to the parameter values being too high, $y^*_{\Delta g} - y_{\Delta g} > 0$, or too low, $y^*_{\Delta g} - y_{\Delta g} < 0$. The biased update is encoded by sampling a positively or negatively truncated Gaussian distribution, figure 3.2, as the proposal distribution, $q(\cdot, \cdot)$.

Figure 3.8 plots a comparison between the normalized misfit across both datasets, Δg and Δt, for the analytical scheme, red, and diagnostic ABC-tomography, blue, during the initial phase of their respective Markov chains. This demonstrates a diagnostic ABC joint inversion can improve convergence to a low misfit model. While acceptance rate is particularly low compared to sampling of the analytically defined posterior, each move in the ABC scheme is more informed and as a result higher quality (i.e the misfit is lower). In light of this there is also scope for the joint implementation of Approximate and analytical Bayesian inference. Figure 3.9 and 3.10 plot the mean marginal model

for the full chain length (100,000 time steps). Figure 4.1 (Supplementary material) plots a comparison between the sampled marginal ABC posterior and marginal analytical-posterior for a sub-set of the parameter space. Likewise, figure 4.2 plots a comparison of the Markov chain trace for the diagnostic ABC scheme and the analytical posterior for a sub-set of the parameter space. The over-estimation of uncertainty in the ABC posterior (figure 4.1 and 4.2) is a result of approximation introduced by the tolerance. This can be seen by comparing the marginal ABC posterior for a tolerance of $\epsilon = \vec{1} \cdot 0.1$, figure 4.1, to the ABC posterior for a tolerance of $\epsilon = \vec{1} \cdot 0.05$, figure 4.3, and a tolerance of $\epsilon = \vec{1} \cdot 0.025$, figure 4.4. As ϵ is decreased the posterior density becomes tightly clustered around the true parameter value. However, as tolerance is decreased the sampler efficiency is increasingly eroded, with lowered acceptance rates and increased τ, figure 4.5 and 4.6. As discussed in chapter 2, this trade-off can be addressed with improved sampling routines such as AM, DR and DRAM. ABC samplers are an active area and efficient new algorithms such as Hamiltonian ABC (Meeds et al., 2015) offer potential to sample increasingly accurate posteriors.

As an alternative to using a truncated Gaussian distribution to bias the parameter update, it is also possible to use a skew Gaussian distribution. This is applied in the Supplementary material, figure 4.8, 4.9 and 4.10. The analytical definition for the skew normal distribution, $f(x)$, used to direct the parameter updates is:

$$\begin{aligned} \phi(x) &= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2} \right) \\ \Phi(x) &= \frac{1}{2}\left(1 + \operatorname{erf}\left(\frac{x}{\sqrt{2}}\right)\right) \\ f(x) &= \frac{2}{\omega}\phi\left(\frac{x-\xi}{\omega}\right)\Phi\left(\alpha\left(\frac{x-\xi}{\omega}\right)\right) \end{aligned} \tag{3.8}$$

here erf is the error function, ξ is the location, ω is the scale and α is a shape parameter. In this application ξ is the current chain location, ω is held at 250, and α is either 1, for a positively skewed distribution, or -1, for a negatively skewed distribution. It should be noted that when using the skew normal distribution the ratio $\frac{q(\boldsymbol{\theta}^*, \boldsymbol{\theta}_{t-1})}{q(\boldsymbol{\theta}_{t-1}, \boldsymbol{\theta}^*)}$ must be computed at each time step to ensure detailed balance is maintained.

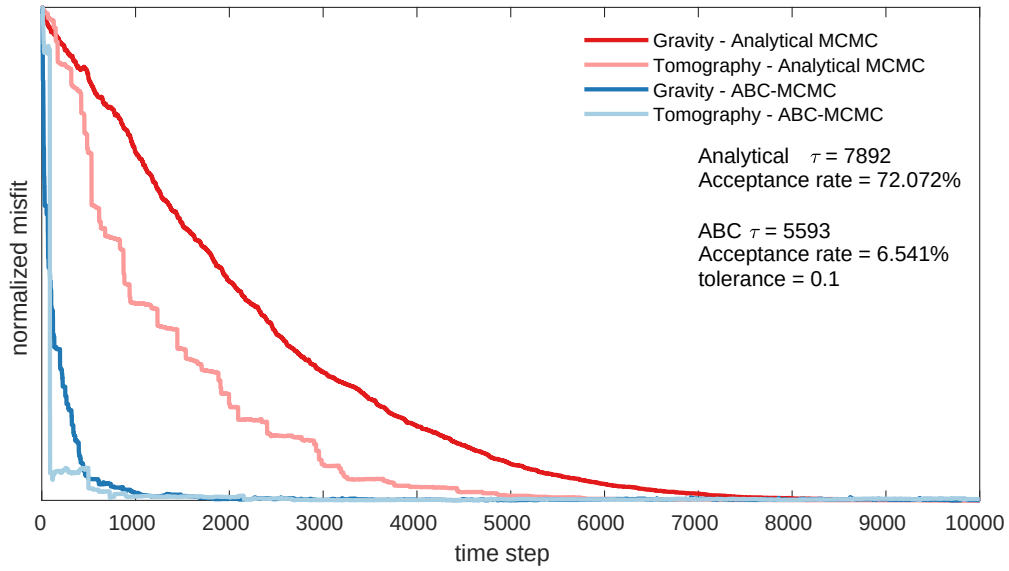


FIGURE 3.8: Truncated Gaussian scheme. The normalized misfit of the chain state during the initial phase (first 10,000 steps) for both datasets, Δg and Δt, and for both methods, an analytically defined posterior sampled by MCMC and ABC-tomography. The 'true model' and observed data is plotted in figure 3.6 and 3.7. The increase in convergence to a low misfit model under ABC-tomography is a result of dynamically selecting, localizing and directly the update at each time step in the Markov chain. The details of the full chain run (100,000 time steps), are also displayed on the figure. τ is the integrated auto-correlation time. The mean marginal models from ABC-tomography using truncated Gaussian updates is plotted in figure 3.9 and 3.10.

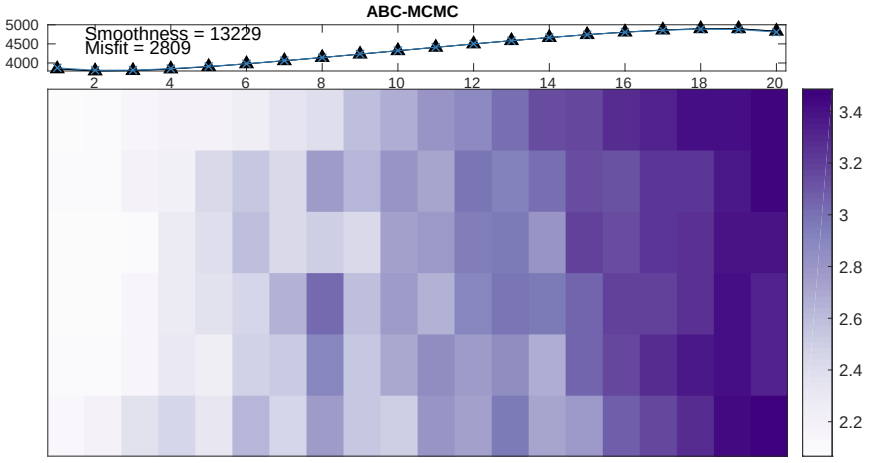


FIGURE 3.9: Truncated Gaussian scheme. The mean of the marginal ABC posterior, $p_{ABC}(\boldsymbol{\theta}|\boldsymbol{S}(\boldsymbol{y}))$, targeting the 'true model' and observed data, figure 3.6. The simulated data generated by this 'solution', is plotted, blue, compared to the observed data, black. The misfit during the initial phase for this chain is plotted in figure 3.8. The corresponding mean marginal V_p model is plotted in figure 3.10.

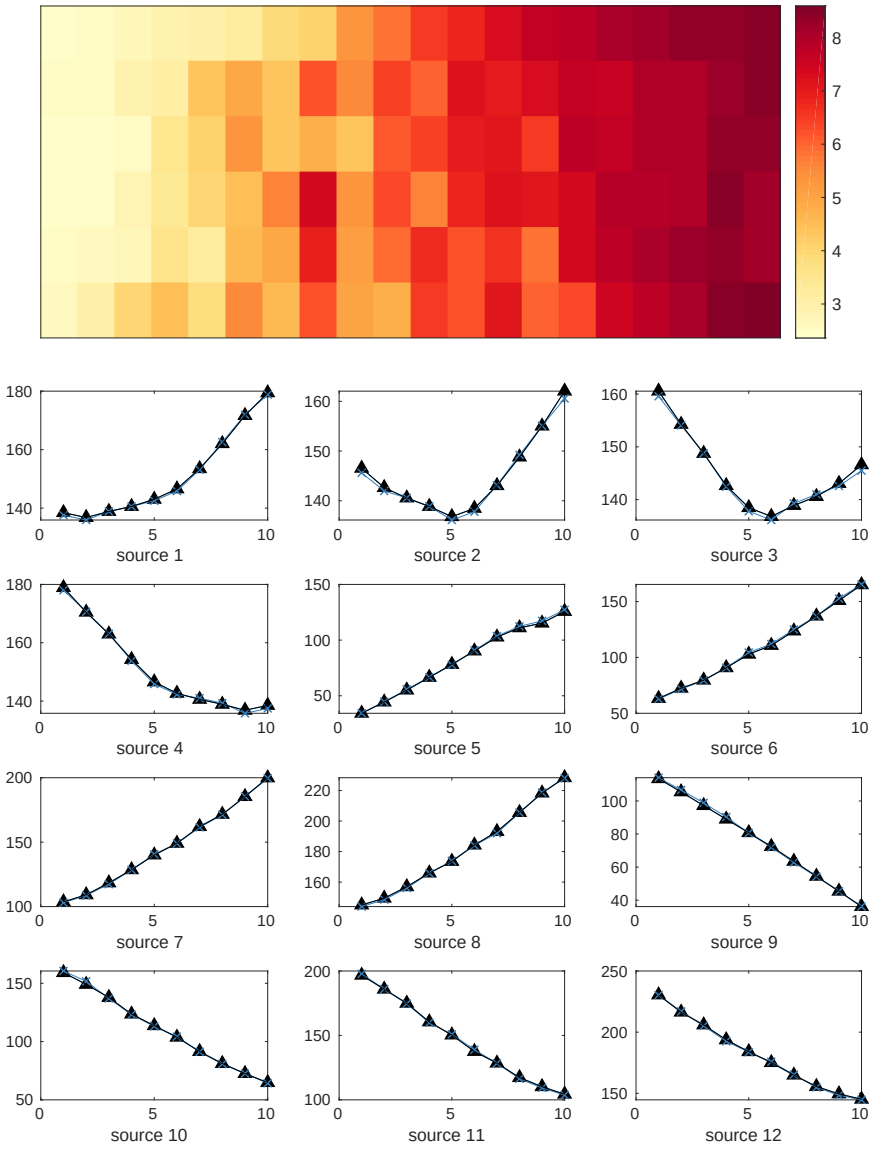


FIGURE 3.10: Truncated Gaussian scheme. The mean of the marginal ABC posterior, $p_{ABC}(\boldsymbol{\theta}|\boldsymbol{S}(\boldsymbol{y}))$, targeting the 'true model' and observed data of figure 3.7. The simulated data generated by this 'solution', is plotted, blue, compared to the observed data, black. The misfit during the initial phase for this chain is plotted in figure 3.8. The corresponding mean marginal ρ model is plotted in figure 3.9.

Chapter 4

Conclusion

This thesis explored some potential advantages offered by Approximate Bayesian Computation (ABC) in the context of geophysical inversions. Chapter 3 illustrated how ABC can use the information available by opening the likelihood to drive a more diagnostic inversion scheme. This diagnostic inversion improves in comparison to traditional likelihood machinery by ensuring each model update has impact by providing information on where the model is inadequate and how it can improve to match the observed data. A specific result is rapid convergence to low misfit models while retaining the same formal statistical guarantees offered by probabilistic formulations of geophysical inverse problems. Algorithms which rapidly find and explore low misfit models are necessary if we wish to push the limits in the resolution and scale of solvable geophysical inverse problems which fundamentally require a probabilistic approach. This thesis represents the first attempt at using ABC in joint geophysical inversions, providing an initial connection to the rapidly expanding field of likelihood-free methods. This value is highlighted by the new likelihood-free methods consistently emerging which may benefit geophysics (Papamakarios and Murray, 2016; Song et al., 2017).

While the results presented in this exploratory analysis are overall positive, more comprehensive studies are required to fully uncover the potential of ABC in geophysics. Firstly, the range of tests for the diagnostic ABC inversion needs to be expanded to larger and more realistic Earth models. The intrinsic trade-off between tolerance, sampler efficiency and accuracy of the posterior PDF needs to be addressed in more detail if ABC is to become a robust practical option. Further exploration is also needed in how to best drive a diagnostic scheme given the available information, while still retaining formal statistical guarantees. The potential for comparative improvement is not restricted to this angle. It is worthwhile assessing the impact of simplifying assumptions about the nature of modelization and data uncertainty, compared with its true nature. There is also scope for assessing the model adequacy given the data, as opposed to relative adequacy against other models.

This thesis demonstrates how ABC can be used to build upon the progress made in probabilistic tomography.

References

Afonso, J. C., Fullea, J., Griffin, W. L., Yang, Y., Jones, A. G., D Connolly, J. A., and O'Reilly, S. Y. (2013a). 3-D multiobservable probabilistic inversion for the compositional and thermal structure of the lithosphere and upper mantle. I: A priori petrological information and geophysical observables. *Journal of Geophysical Research: Solid Earth*, 118(5):2586–2617.

Afonso, J. C., Fullea, J., Yang, Y., Connolly, J. A. D., and Jones, A. G. (2013b). 3-D multi-observable probabilistic inversion for the compositional and thermal structure of the lithosphere and upper mantle. II: General methodology and resolution analysis. *Journal of Geophysical Research: Solid Earth*, 118(4):1650–1676.

Afonso, J. C., Rawlinson, N., Yang, Y., Schutt, D. L., Jones, A. G., Fullea, J., and Griffin, W. L. (2016). 3-D multiobservable probabilistic inversion for the compositional and thermal structure of the lithosphere and upper mantle: III. Thermochemical tomography in the Western-Central US. *Journal of Geophysical Research: Solid Earth*, 121(10):7337–7370.

Aster, R. C., Borchers, B., and Thurber, C. H. (2013). *Parameter Estimation and Inverse Problems*.

Beaumont, M. A., Zhang, W., and Balding, D. J. (2002). Approximate Bayesian Computation in Population Genetics. *Genetics*, 162(4):2025–2035.

Biegler, L., Biros, G., Ghattas, O., Heinkenschloss, M., Keyes, D., Mallick, B., Marzouk, Y., Tenorio, L., van Bloemen Waanders, B., and Willcox, K. (2010). *Large-Scale Inverse Problems and Quantification of Uncertainty*.

Blakely, R. J. (1996). *Potential theory in gravity and magnetic applications*. Cambridge University Press.

Blum, M. G. B., François, O., Blum, M. G. B., and François, O. (2010). Non-linear regression models for Approximate Bayesian Computation. *Stat Comput*, 20:63–73.

Blum, M. G. B., Nunes, M. A., Prangle, D., and Sisson, S. A. (2013). A Comparative Review of Dimension Reduction Methods in Approximate Bayesian Computation. *Statistical Science*, 28(2):189–208.

Bodin, T., Sambridge, M., Tkalćić, H., Arroucau, P., Gallagher, K., and Rawlinson, N. (2012). Transdimensional inversion of receiver functions and surface wave dispersion. *Journal of Geophysical Research: Solid Earth*.

Bortot, P., Coles, S. G., and Sisson, S. A. (2007). Inference for stereological extremes. *Journal of the American Statistical Association*, 102(477):84–92.

Botev, Z. I., Grotowski, J. F., and Kroese, D. P. (2010). Kernel density estimation via diffusion. *The Annals of Statistics*.

Box, G. E. P. and Tiao, G. C. (1973). *Bayesian Inference in Statistical Analysis*.

Brocher, T. M. (2005). Empirical relations between elastic wavespeeds and density in the Earth's crust. *Bulletin of the Seismological Society of America*.

Casella, G. and Berger, G. L. (1993). *Statistical inference*.

Cox, D. (2007). *Principles of statistical inference*.

Fisher, R. A. (1922). On the Mathematical Foundations of Theoretical Statistics. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*.

Fu, Y. X. and Li, W. H. (1997). Estimating the age of the common ancestor of a sample of DNA sequences. *Molecular Biology and Evolution*, 14(2):195–199.

Gelman, A. and Roberts, G. (1996). Efficient Metropolis jumping rules. *Bayesian statistics*, 5:599–608.

Gilks, W. R., Richardson, S., and Spiegelhalter, D. (1995). *Markov chain Monte Carlo in practice*. CRC press.

Gregory, P. (2005). *Bayesian Logical Data Analysis for the Physical Sciences: A Comparative Approach with Mathematica®Support*. Cambridge University Press.

Gutmann, M. U., Corander, J., and Others (2016). Bayesian optimization for likelihood-free inference of simulator-based statistical models. *Journal of Machine Learning Research*.

Haario, H., Laine, M., Mira, A., and Saksman, E. (2006). DRAM: Efficient adaptive MCMC. *Statistics and Computing*, 16(4):339–354.

Haario, H., Saksman, E., and Tamminen, J. (1999). Adaptive proposal distribution for random walk Metropolis algorithm. *Computational Statistics*.

Haario, H., Saksman, E., and Tamminen, J. (2001). An Adaptive Metropolis Algorithm. *Bernoulli*, 7(2):223.

Idier, J. (2013). *Bayesian approach to inverse problems*. John Wiley \& Sons.

Kaipio, J. and Somersalo, E. (2006). *Statistical and computational inverse problems*. Springer Science & Business Media.

Khan, A., Connolly, J. A. D., Maclennan, J., and Mosegaard, K. (2007). Joint inversion of seismic and gravity data for lunar composition and thermal state. *Geophysical Journal International*, 168(1):243–258.

Khan, A., Zunino, A., and Deschamps, F. (2011). The thermo-chemical and physical structure beneath the North American continent from Bayesian inversion of surface-wave phase velocities. *Journal of Geophysical Research: Solid Earth*.

Laine, M. (2008). *Adaptive MCMC Methods with Applications in Environmental and Geophysical Models*. PhD thesis, Lapeenranta University of Technology.

Li, J., Nott, D. J., Fan, Y., and Sisson, S. A. (2017). Extending approximate Bayesian computation methods to high dimensions via a Gaussian copula model. *Computational Statistics and Data Analysis*, 106:77–89.

Lintusaari, J., Gutmann, M. U., Dutta, R., Kaski, S., and Corander, J. (2016). Fundamentals and Recent Developments in Approximate Bayesian Computation. *Syst. Biol*, 00(0):1–17.

Marjoram, P., Molitor, J., Plagnol, V., and Tavare, S. (2003). Markov chain Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences of the United States of America*, 100(26):15324–8.

Marjoram, P. and Tavaré, S. (2006). Modern computational approaches for analysing molecular genetic variation data.

Meeds, E., Leenders, R., and Welling, M. (2015). Hamiltonian abc. *arXiv preprint arXiv:1503.01916*.

Menke, W. (2012). *Geophysical Data Analysis: Discrete Inverse Theory*.

Mira, A. (2001). On Metropolis-Hastings algorithms with delayed rejection. *Metron*.

Moorkamp, M., Jones, A. G., and Fishwick, S. (2010). Joint inversion of receiver functions, surface wave dispersion, and magnetotelluric data. *Journal of Geophysical Research: Solid Earth*.

Mosegaard, K. and Tarantola, A. (1995). Monte Carlo sampling of solutions to inverse problems. *Journal of Geophysical Research*.

Mosegaard, K. and Tarantola, A. (2002). Probabilistic approach to inverse problems. *International Geophysics*, 81:237–265.

Papamakarios, G. and Murray, I. (2016). Fast epsilon-free inference of simulation models with Bayesian conditional density estimation. In *Advances in Neural Information Processing Systems*, pages 1028–1036.

Prangle, D. (2017). Summary statistics in approximate Bayesian computation. In *Handbook of approximate Bayesian computation*, page 320.

Pritchard, J. K., Seielstad, M. T., Perez-Lezaun, A., and Feldman, M. W. (1999). Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Molecular Biology and Evolution*, 16(12):1791–1798.

Pudlo, P., Marin, J. M., Estoup, A., Cornuet, J. M., Gautier, M., and Robert, C. P. (2015). Reliable ABC model choice via random forests. *Bioinformatics*, 32(6):859–866.

Ratmann, O., Andrieu, C., Wiuf, C., and Richardson, S. (2009). Model criticism based on likelihood-free inference, with an application to protein network evolution. *Proceedings of the National Academy of Sciences*, 106(26):10576–10581.

Ratmann, O., Andrieu, C., Wiuf, C., and Richardson, S. (2010). Reply to Robert et al.: Model criticism informs model choice and model comparison. *Proceedings of the National Academy of Sciences*, 107(3):E6–E7.

Roberts, G. O. and Sahu, S. K. (1997). Updating schemes, correlation structure, blocking and parameterization for the gibbs sampler. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*.

Sadegh, M. and Vrugt, J. A. (2014). Approximate Bayesian Computation using Markov Chain Monte Carlo simulation. *Water Resources Research*, 10(2):6767–6787.

Sambridge, M. (1999). Geophysical inversion with a neighbourhood algorithm–II. Appraising the ensemble. *Geophysical Journal International*.

Sambridge, M., Gallagher, K., Jackson, A., and Rickwood, P. (2006). Trans-dimensional inverse problems, model comparison and the evidence. *Geophysical Journal International*.

Sambridge, M. and Mosegaard, K. (2002). Monte Carlo methods in geophysical inverse problems. *Reviews of Geophysics*.

Sargent, D. J., Hodges, J. S., and Carlin, B. P. (2000). Structured Markov chain Monte Carlo. *Journal of Computational and Graphical Statistics*.

Shapiro, N. M. and Ritzwoller, M. H. (2002). Monte-Carlo inversion for a global shear-velocity model of the crust and upper mantle. *Geophysical Journal International*.

Shen, W., Ritzwoller, M. H., Schulte-Pelkum, V., and Lin, F.-C. (2012). Joint inversion of surface wave dispersion and receiver functions: a Bayesian Monte-Carlo approach. *Geophysical Journal International*.

Sisson, S. a. and Fan, Y. (2010). Likelihood-free Markov chain Monte Carlo. *Handbook of Markov chain Monte Carlo*, (Mcmc).

Sisson, S. A., Fan, Y., and Beaumont, M. (2016). Handbook of Approximate Bayesian Computation.

Song, J., Zhao, S., and Ermon, S. (2017). A-nice-mc: Adversarial training for mcmc. In *Advances in Neural Information Processing Systems*, pages 5140–5150.

Sprott, D. A. (2008). *Statistical Inference in Science*. Springer Science & Business Media.

Sunnåker, M., Busetto, A. G., Numminen, E., Corander, J., Foll, M., and Dessimoz, C. (2013). Approximate Bayesian Computation. *PLoS Computational Biology*, 9(1).

Tarantola, A. (2005). *Inverse problem theory and methods for model parameter estimation*. SIAM.

Tarantola, A. and Valette, B. (1982). Inverse Problems = Quest for Information. *Journal of Geophysics*, 50(3):159–170.

Tavare, S., Balding, D. J., Griffiths, R. C., and Donnelly, P. (1997). Inferring Coalescence Times from DNA Sequence Data. *Genetics*, 145(2):505–518.

Trampert, J., Deschamps, F., Resovsky, J., and Yuen, D. (2004). Probabilistic tomography maps chemical heterogeneities throughout the lower mantle. *Science*.

Vrugt, J. A. and Sadegh, M. (2013). Toward diagnostic model calibration and evaluation: Approximate Bayesian computation. *Water Resources Research*, 49(7):4335–4345.

Weiss, G. and von Haeseler, A. (1998). Inference of population history using a likelihood approach. *Genetics*, 149(July):1539–1546.

Wood, S. N. (2010). Statistical inference for noisy nonlinear ecological dynamic systems. *Nature*, 466(7310):1102–1104.

Supplementary material

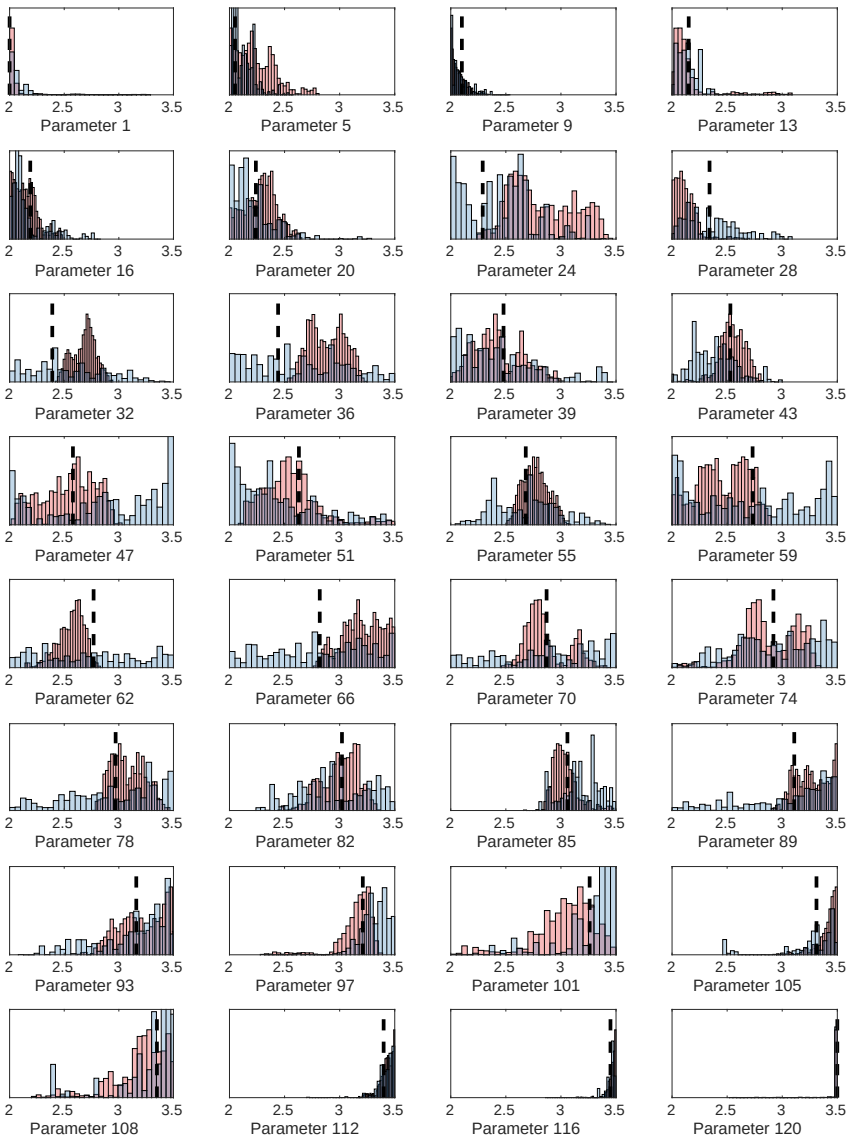


FIGURE 4.1: A comparison of the marginal ABC posterior (blue), sampled by the diagnostic ABC joint inversion, and the analytically defined posterior (red), sampled by MCMC, for a subset of the parameter space. This is the PDF which underlies the experiment in section 3.2 and figures 3.8, 3.9 and 3.10. The tolerance here is $\epsilon = \vec{1} \cdot 0.1$. The black dashed lines marks the 'true' parameter values from figures 3.6 and 3.7. This ABC posterior can be compared to the ABC posterior for the same problem with $\epsilon = \vec{1} \cdot 0.05$, figure 4.3 and $\epsilon = \vec{1} \cdot 0.025$, figure 4.4. These comparisons highlight how uncertainty is overestimated in the ABC posterior documented in this figure as a result of the approximation introduced by the tolerance. The parameters are numbered down column, starting with '1' in the top left corner of figure 3.6/3.7.

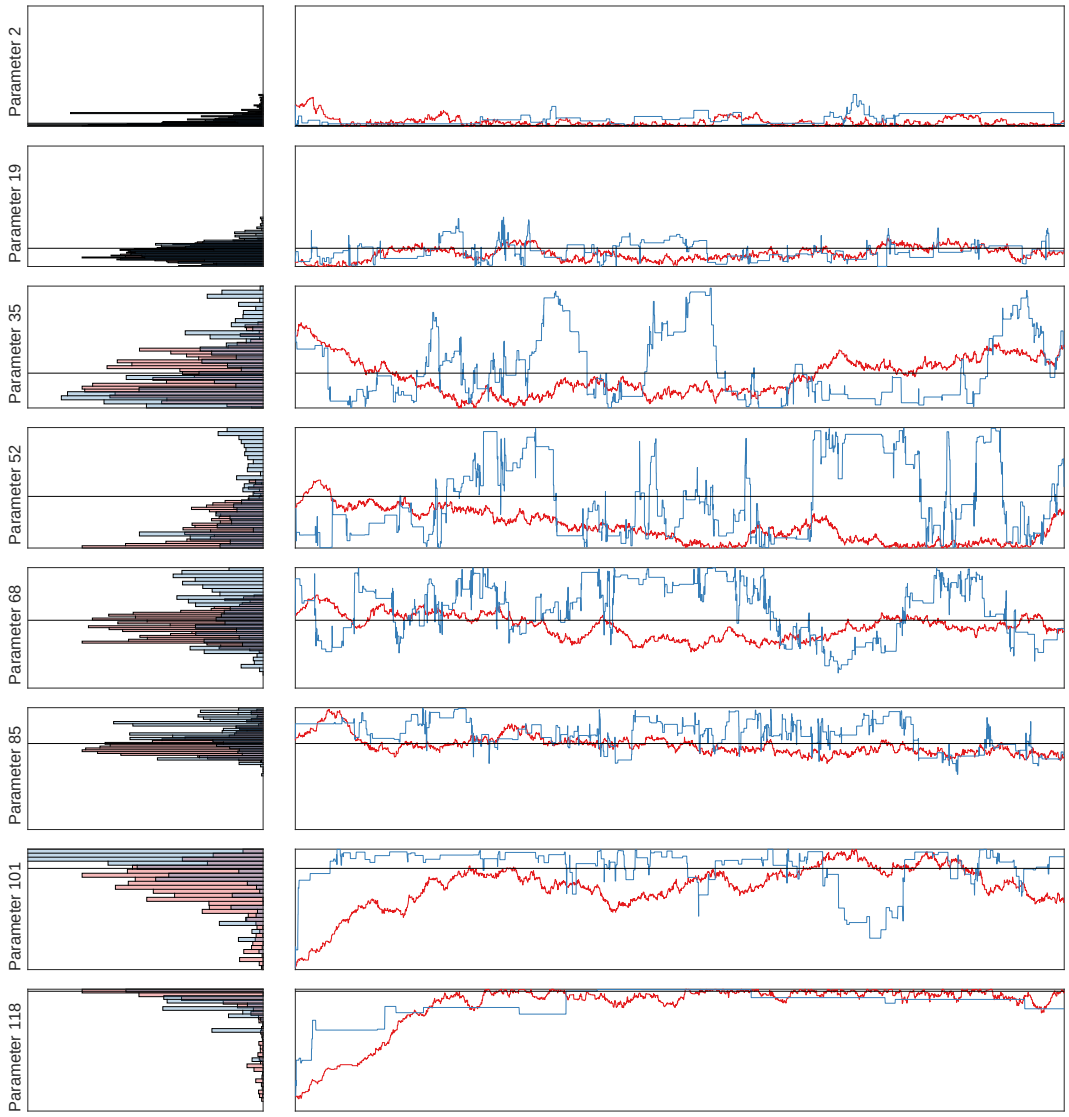


FIGURE 4.2: A comparison of Markov chain traces for the sampling of the ABC (blue) and analytically-defined (red) posteriors undertaken in section 3.2. The Markov chains are both of length 100,000. This figure is limited to a subset of 8 parameters out of the 120 total unknowns. The marginal histograms associated with each chain are also displayed. The black line marks the 'true' parameter values from figures 3.6 and 3.7. The parameters are numbered down column, starting with '1' in the top left corner of figure 3.6/3.7.

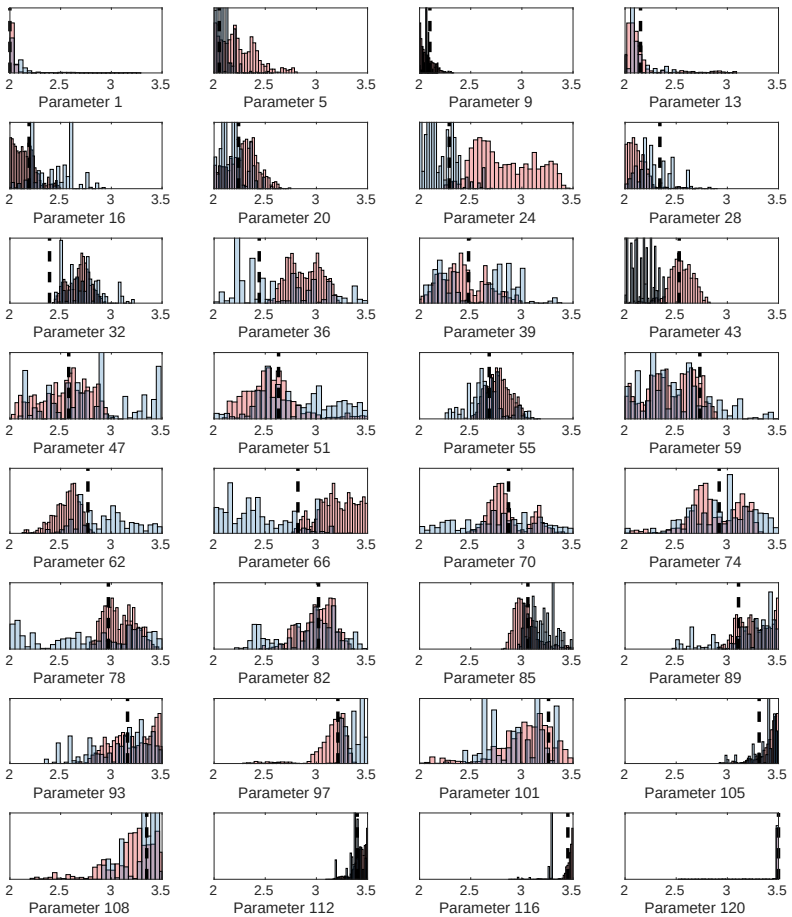


FIGURE 4.3: A comparison of the marginal ABC posterior (blue), sampled by the diagnostic ABC joint inversion, and the analytically defined posterior (red), for a repetition of the experiment in section 3.2 with a lower tolerance in the ABC likelihood approximation $p(\boldsymbol{S}(\boldsymbol{y})|\boldsymbol{S}(\boldsymbol{y}^*, \boldsymbol{\theta}))$. The tolerance here is $\epsilon = \vec{1} \cdot 0.05$, half the tolerance of the original experiment documented in section 3.2. The black dashed lines marks the 'true' parameter values from figures 3.6 and 3.7. This ABC posterior can be compared to the ABC posterior for the original experiment, figure 4.3 and an ABC posterior with $\epsilon = \vec{1} \cdot 0.025$, figure 4.4. The sampling acceptance rate of the ABC posterior documented in this figure is 2.635%, and the auto-correlation time (τ) is 7892, figure 4.6. The ABC posterior density clusters closer around the true parameter value than the ABC posterior with a tolerance of $\epsilon = \vec{1} \cdot 0.1$. These repetitions demonstrate that decreasing tolerance improves the accuracy of the ABC posterior at the expense of sampling efficiency.

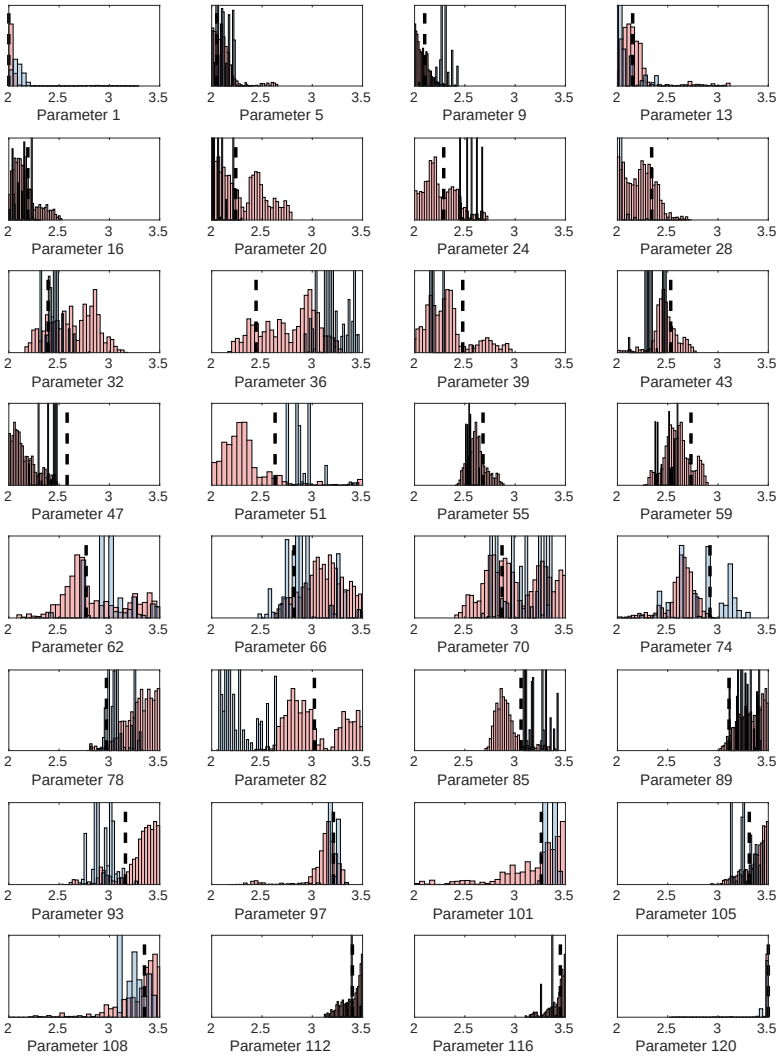


FIGURE 4.4: A comparison of the marginal ABC posterior (blue), sampled by the diagnostic ABC joint inversion, and the analytically defined posterior (red), for a repetition of the experiment in section 3.2 with a lower tolerance in the ABC likelihood approximation $p(\boldsymbol{S}(\boldsymbol{y})|\boldsymbol{S}(\boldsymbol{y}^*, \boldsymbol{\theta}))$. The tolerance here is $\epsilon = \vec{1} \cdot 0.025$, a quarter the tolerance of the original experiment documented in section 3.2. The black dashed lines marks the 'true' parameter values from figures 3.6 and 3.7. This ABC posterior can be compared to the ABC posterior for the original experiment, figure 4.3 and an ABC posterior with $\epsilon = \vec{1} \cdot 0.05$, figure 4.3. The sampling acceptance rate of the ABC posterior documented in this figure is 0.428%, and the auto-correlation time (τ) is 8175, figure 4.6. The ABC posterior density clusters closer around the true parameter value than the ABC posterior with a tolerance of $\epsilon = \vec{1} \cdot 0.1$ and $\epsilon = \vec{1} \cdot 0.05$. This repetitions demonstrate that decreasing tolerance improves the accuracy of the ABC posterior at the expense of sampling efficiency.

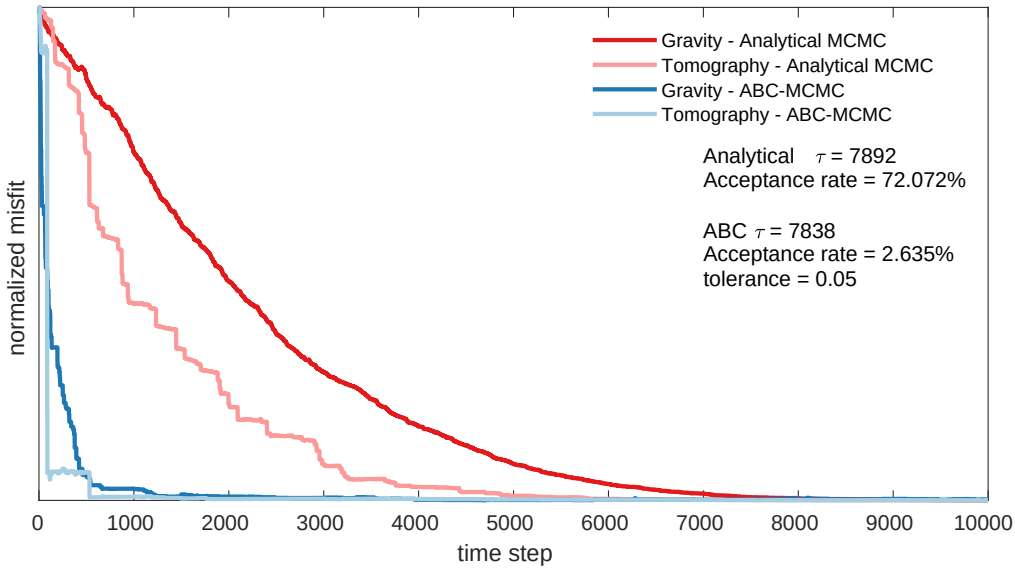


FIGURE 4.5: $\epsilon = \vec{1} \cdot 0.05$. The normalized misfit of the chain state during the initial phase (first 10,000 steps) for a repetition of the experiment in section 3.2 with a lower tolerance. The tolerance for this example is $\epsilon = \vec{1} \cdot 0.05$. This highlights that despite a decrease in sampler efficiency, lower acceptance rate and higher integrated auto-correlation time (τ), the diagnostic ABC scheme still converges to a low misfit model faster than an analytical scheme. The posterior density sampled by the complete Markov chain (100,000 time steps) for a sub-set of the parameter space is plotted in figure 4.3.

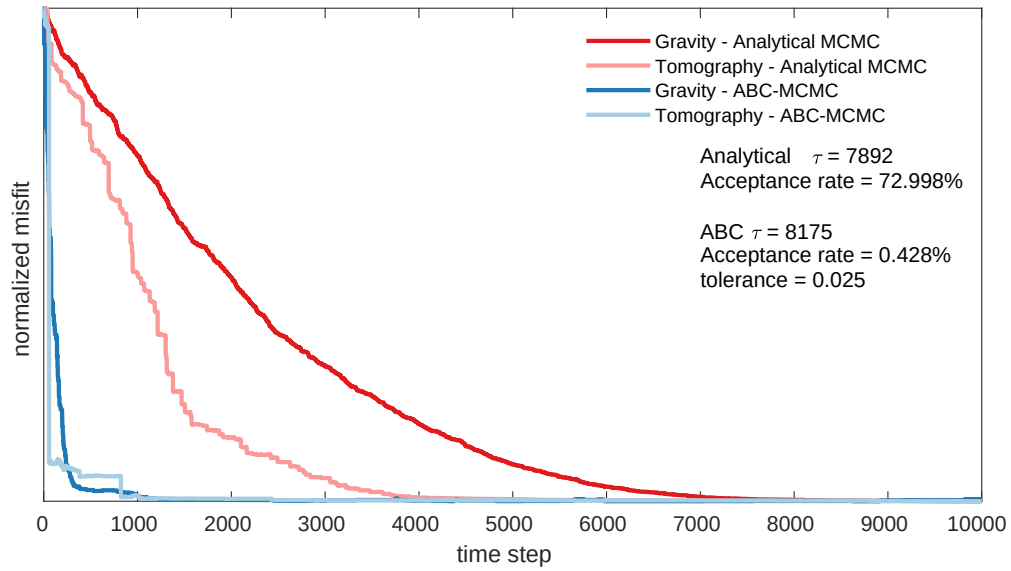


FIGURE 4.6: $\epsilon = \vec{1} \cdot 0.025$. The normalized misfit of the chain state during the initial phase (first 10,000 steps) for a repetition of the experiment in section 3.2 with a lower tolerance. The tolerance for this example is $\epsilon = \vec{1} \cdot 0.025$. This highlights that despite a decrease in sampler efficiency, lower acceptance rate and higher integrated auto-correlation time (τ), the diagnostic ABC scheme still converges to a low misfit model faster than an analytical scheme. The posterior density sampled by the complete Markov chain (100,000 time steps) for a sub-set of the parameter space is plotted in figure 4.4.

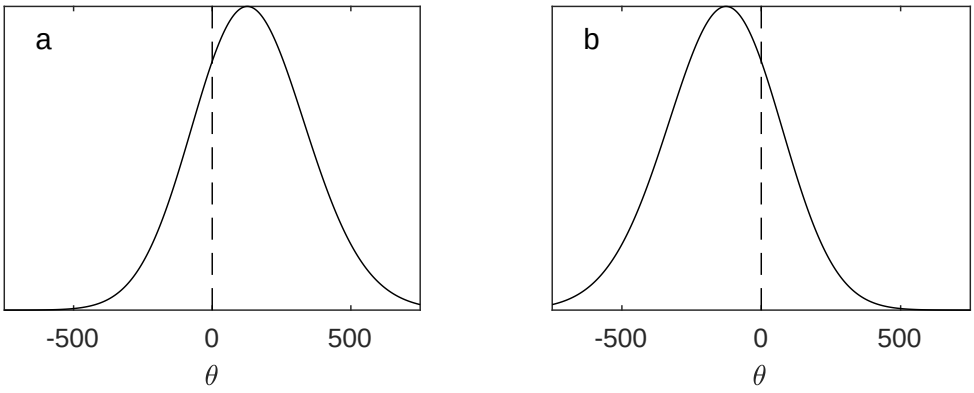


FIGURE 4.7: The skew Gaussian proposal distributions which can be used to positively or negatively bias the parameter update as an alternative to the truncated Gaussian distribution, figure 3.2. The positively skewed distribution (a) is used when the parameter values are deemed too low. The negatively skewed distribution (b) is used when the parameter values are deemed too high. During application the distributions are centered on the current chain and $\omega = 250$.

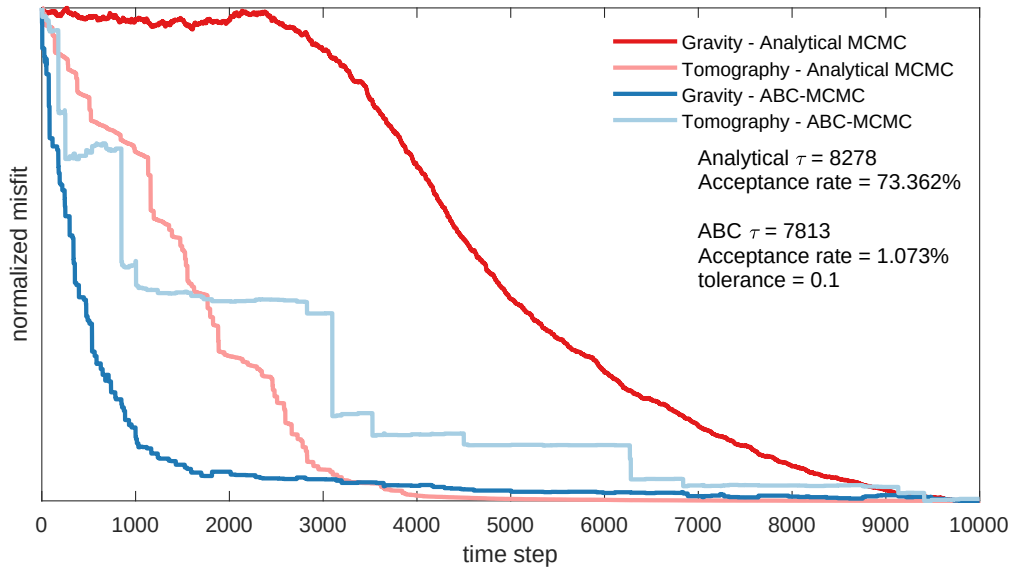


FIGURE 4.8: Skew Gaussian scheme. The normalized misfit of the chain state during the initial phase (first 10,000 steps) for both datasets, Δg and Δt, and for both methods, an analytically defined posterior sampled by MCMC and ABC-tomography. The 'true model' and observed data is plotted in figure 3.6 and 3.7. The increase in the rate of convergence to a low misfit model under ABC-tomography is a result of dynamically selecting, localizing and directly the update at each time step in the Markov chain. The details of the full chain run (100,000 time steps), are also displayed on the figure. τ is the integrated auto-correlation time. The mean marginal models from ABC-tomography using skew Gaussian updates is plotted in figure 4.9 and 4.10.

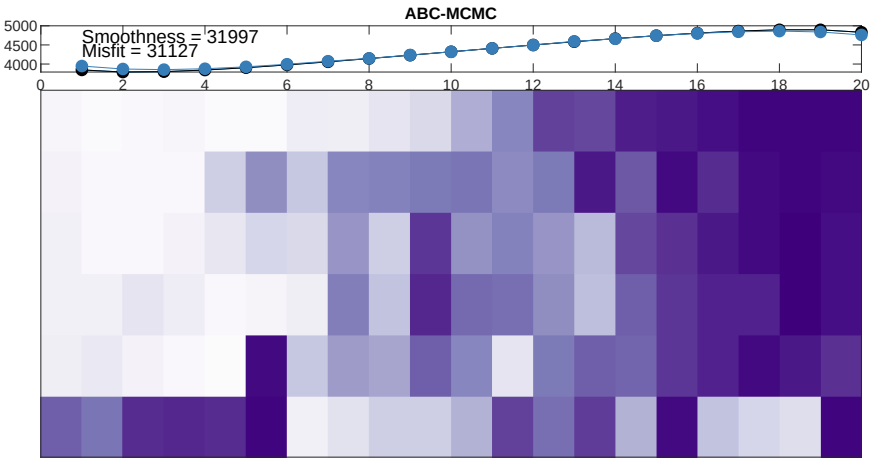


FIGURE 4.9: Skew Gaussian scheme. The mean of the marginal ABC posterior, $p_{ABC}(\boldsymbol{\theta}|\boldsymbol{S}(\boldsymbol{y}))$, targeting the 'true model' and observed data of figure 3.6. The simulated data generated by this 'solution', is plotted, blue, compared to the observed data, black. The misfit during the initial phase for this chain is plotted in figure 4.8. The corresponding mean marginal V_p model is plotted in figure 4.10.

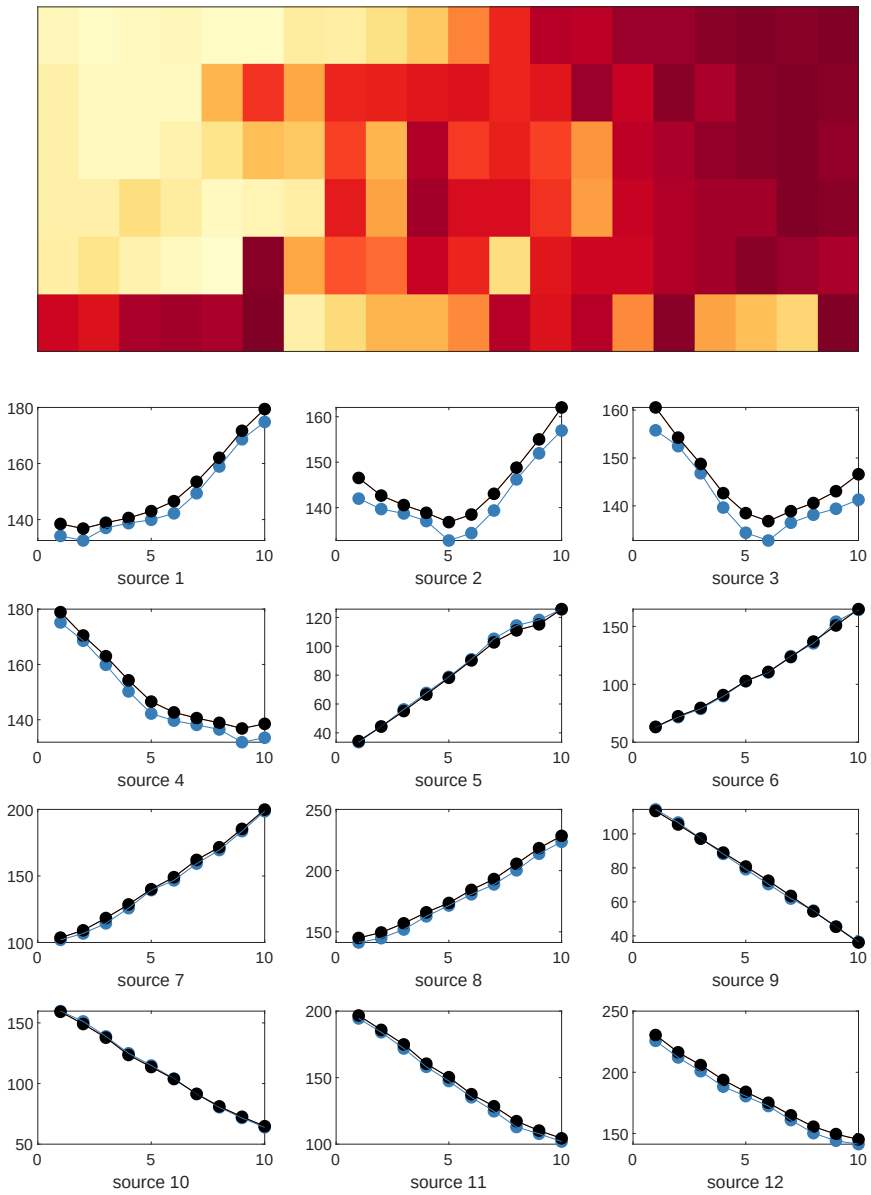


FIGURE 4.10: Skew Gaussian scheme. The mean of the marginal ABC posterior, $p_{ABC}(\boldsymbol{\theta}|\boldsymbol{S}(\boldsymbol{y}))$, targeting the 'true model' and observed data of figure 3.7. The simulated data generated by this 'solution', is plotted, blue, compared to the observed data, black. The misfit during the initial phase for this chain is plotted in figure 4.8. The corresponding mean marginal ρ model is plotted in figure 4.9.