New methods for infinite and high-dimensional approximate Bayesian computation

Guilherme Souza Rodrigues

Supervised by: Prof. Scott Sisson

Co-supervised by: Dr. Yanan Fan

A thesis in fulfilment of the requirements for the degree of
Doctor of Philosophy

UNSW
AUSTRALIA

School of Mathematics and Statistics
Faculty of Science

September 2017

THE UNIVERSITY OF NEW SOUTH WALES
Thesis/Dissertation Sheet

Surname or Family name: Rodrigues

First name: Guilherme Other name/s: Souza

Abbreviation for degree as given in the University calendar: PhD

School: School of Mathematics and Statistics Faculty: Faculty of Science

Title: New methods for infinite and high-dimensional approximate Bayesian computation

Abstract 350 words maximum: (PLEASE TYPE)

The remarkable complexity of modern applied problems often requires the use of probabilistic models where the likelihood is intractable – in the sense that it cannot be numerically evaluated, not even up to a normalizing constant. The statistical literature provides an extensive array of methods designed to bypass this constraint. Still, inference in this context remains computationally challenging, particularly for high-dimensional models. We focus on the important class of Approximation Bayesian Computation (ABC) methods.

Various state-of-the-art ABC techniques are combined to fit an intractable model that describes the epidemiological dynamics of multidrug-resistant tuberculosis. This study addresses a number of important biological questions in a principled manner, providing useful insights to this extraordinarily relevant research topic.

We propose a functional regression adjustment ABC procedure that permits the estimation of infinite-dimensional parameters, which effectively launches ABC into the non-parametric framework. Two likelihood-free algorithms are also introduced. The first exploits the principles of ABC and the so-called coverage property to recalibrate an auxiliary approximate posterior estimator. This approach further strengthens the links between ABC and indirect inference, allowing a more comprehensive use of the auxiliary estimator.

The second algorithm employs the ABC machinery to build approximate samplers for the intractable full conditional distributions. These samplers are then combined to form a likelihood-free approximate Gibbs sampler. The granular nature of our approach (that comes from breaking down the problem into small pieces) makes it suitable for highly-structured problems. We demonstrate this property by fitting an intractable and very high-dimensional state space model.

Declaration relating to disposition of project thesis/dissertation

I hereby grant to the University of New South Wales or its agents the right to archive and to make available my thesis or dissertation in whole or in part in the University libraries in all forms of media, now or here after known, subject to the provisions of the Copyright Act 1968. I retain all property rights, such as patent rights. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

I also authorise University Microfilms to use the 350 word abstract of my thesis in Dissertation Abstracts International (this is applicable to doctoral theses only).

Signature Witness Signature Date

The University recognises that there may be exceptional circumstances requiring restrictions on copying or conditions on use. Requests for restriction for a period of up to 2 years must be made in writing. Requests for a longer period of restriction may be considered in exceptional circumstances and require the approval of the Dean of Graduate Research.

FOR OFFICE USE ONLY Date of completion of requirements for Award:

ORIGINALITY STATEMENT

'I hereby declare that this submission is my own work and to the best of my knowledge it contains no materials previously published or written by another person, or substantial proportions of material which have been accepted for the award of any other degree or diploma at UNSW or any other educational institution, except where due acknowledgement is made in the thesis. Any contribution made to the research by others, with whom I have worked at UNSW or elsewhere, is explicitly acknowledged in the thesis. I also declare that the intellectual content of this thesis is the product of my own work, except to the extent that assistance from others in the project's design and conception or in style, presentation and linguistic expression is acknowledged.'

Signed .

Date ..

School of Mathematics and Statistics
The Red Centre, Centre Wing
Kensington Campus
UNSW Sydney, NSW 2051
Australia

Graduate Research School
Lvl 2 South Wing Rupert Myers Building
Gate 14 Barker Street Entrance
Kensington Campus
UNSW Sydney, NSW 2051
Australia

Acknowledgements

I feel extremely grateful to my supervisor Prof. Scott A. Sisson, who guided me brilliantly throughout these years. His solid knowledge of ABC, and Bayesian statistics more generally, ensured that the projects were always heading in a fruitful direction. Also impressive was his refined capacity to promote a working environment that inspired me to perform to the best of my abilities, respecting my (numerous) limitations and fostering my strengths. But what I learnt to admire the most was his goodwill and kindness that were above and beyond what one could hope for. Scott, thank you very, very much!

I was also extraordinarily fortunate to have had wonderful researches working with us along the way. David Nott greatly contributed to this thesis — his involvement was paramount to both the functional regression adjustment ABC (Chapter 3) and the approximate Gibbs sampler (Chapter 5) projects. I look forward, David, to meeting you personally somewhen in the future. Mark Tanaka and Andrew Francis were the precious experts in the biological aspects of the work on multidrug resistant Tuberculosis (Chapter 2). These are fantastic researches who I had the pleasure to have worked with. Dennis Prangle, in turn, had the clever insight to exploit the coverage property to recalibrate an ABC estimator (Chapter 4), not to mention his valuable inputs to the manuscript itself. Thank you for inviting us to be part of this beautiful project. This thesis was much improved in response to the valuable and well-informed comments and suggestions given by its examiners Jukka Corander and Richard Everitt. Thank you for taking the time to carefully review this work. I would also like to express my deep gratitude to my co-supervisor Yanan Fan for supporting me and, not less importantly, for being so wonderful to Thais.

My Brazilian sponsor, Coordination for the Improvement of Higher Education Personnel (CAPES), provided all the required funds that made this endeavour possible. This

acknowledgement extends, of course, to the Brazilian people, for whom I commit to keep working hard.

Words are unfit to express the immeasurable gratitude I feel for my wife, Thais, who has made huge sacrifices to accompany me to this wonderful and distant land called Australia. I'm also blessed enough to be son of a supportive and loving family that did everything within their reach to make sure I could pursue all of my dreams. To my family, Thais, Valdir, Aida, Luciano, Julia, Isabella, Leonardo, Joanna and Laura, this thesis is dedicated to you.

Life in Sydney was truly memorable thanks to my great mates (in no particular order) Boris, Eve, Alex, Ashish, Jaslene, Xin, Tom, Andi, Cecilia, Cassie, Yoshi, Reshma, Susannah, John, Olga, Wilder, Maria, Fernando, Astrid, Paloma, Father Tru, Danielle, Paolo, Rafael, Ademir, Renata, Sofia, Luiz, Cassine, Cormac, Yash, Varun, Alex, Santos, Kavindee, Josh, Matt, Ariane, Mike, Austin, Garima and Arvin. For my Brazilian friends who embraced this project long before it actually started — Laura, Marilia, Ana, João, Rogério, Petrus, Cristina, Luana, Amanda, Darley, Cibele, Antônio Eduardo, Ilvan, Kelly, Juliana, Larissa, Luiz and Gilmário, thank you from the bottom of my heart.

Abstract

The remarkable complexity of modern applied problems often requires the use of probabilistic models where the likelihood is intractable – in the sense that it cannot be numerically evaluated, not even up to a normalizing constant. The statistical literature provides an extensive array of methods designed to bypass this constraint. Still, inference in this context remains computationally challenging, particularly for high-dimensional models. We focus on the important class of Approximation Bayesian Computation (ABC) methods.

Various state-of-the-art ABC techniques are combined to fit an intractable model that describes the epidemiological dynamics of multidrug-resistant tuberculosis. This study addresses a number of important biological questions in a principled manner, providing useful insights to this extraordinarily relevant research topic.

We propose a functional regression adjustment ABC procedure that permits the estimation of infinite-dimensional parameters, which effectively launches ABC into the nonparametric framework. Two likelihood-free algorithms are also introduced. The first exploits the principles of ABC and the so-called coverage property to recalibrate an auxiliary approximate posterior estimator. This approach further strengthens the links between ABC and indirect inference, allowing a more comprehensive use of the auxiliary estimator.

The second algorithm employs the ABC machinery to build approximate samplers for the intractable full conditional distributions. These samplers are then combined to form a likelihood-free approximate Gibbs sampler. The granular nature of our approach (that comes from breaking down the problem into small pieces) makes it suitable for highly-structured problems. We demonstrate this property by fitting an intractable and very high-dimensional state space model.

Keywords

approximate Bayesian computation (ABC); Gaussian process prior; Gibbs sampler; hierarchical models; indirect inference; intractable state space models; likelihood-free inference; nonparametric density estimation; regression-adjustment.

Contents

Introduction

In statistical practice, the phenomenon under consideration is described by a probabilistic model, $p(\boldsymbol{X}|\boldsymbol{\theta})$, that assigns a probability to each possible outcome, $\boldsymbol{X}$, given a parameter vector $\boldsymbol{\theta}$. The so-called *frequentist* paradigm interprets $\boldsymbol{\theta}$ as an unknown but fixed quantity. In the *Bayesian* framework, in contrast, the (practitioner's) uncertainty about the model parameter is summarized by a prior distribution $\pi(\boldsymbol{\theta})$. Under the latter approach, Bayes' theorem then allows one to adequately process the information encapsulated in the observed data $\boldsymbol{X}_{\text{obs}}$. This learning machinery gracefully channels the available information to the posterior distribution, $\pi(\boldsymbol{\theta}|\boldsymbol{X}) = p(\boldsymbol{X}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})/p(\boldsymbol{X})$, from which inference is performed.

For most statistical models, the marginal likelihood $p(\boldsymbol{X})$ is a complex, multidimensional integral that cannot be directly evaluated, not even numerically – a distribution that cannot be evaluated is said to be *intractable*. The statistical literature has a rich and extensive array of methods for computing features of the posterior distribution (e.g. its mean and quantiles) without directly evaluating $p(\boldsymbol{X})$. These include but are not limited to *Markov Chain Monte Carlo* (MCMC) and *Sequential Monte Carlo* (SMC). This thesis focuses on the problem of statistical inference for models where not only the posterior distribution normalizing constant $p(\boldsymbol{X}_{\text{obs}})$ cannot be evaluated, but the likelihood $p(\boldsymbol{X}|\boldsymbol{\theta})$ itself is intractable. A number of strategies have been proposed to bypass this even harder constraint. *Composite likelihoods* (Lindsay, 1988), *Indirect inference* (Gourieroux et al., 1993), *pseudo-marginal methods* (Beaumont, 2003; Andrieu and Roberts, 2009) and *synthetic likelihoods* (Wood, 2010) are all well-established classes of algorithms in this context. We are particularly interested in the family of Approximate Bayesian Computation (ABC) methods (Beaumont et al., 2002).

In its simplest formulation, known as rejection-ABC, the ABC algorithm avoids direct evaluation of the likelihood by performing a sequence of three simple steps. First, a

candidate parameter $\boldsymbol{\theta}^* \sim \pi(\boldsymbol{\theta})$ is drawn from the prior. Then, a *synthetic* (or *pseudo*) sample, $\boldsymbol{X}^* \sim p(\boldsymbol{X}|\boldsymbol{\theta}^*)$ is generated from the model conditionally on $\boldsymbol{\theta}^*$. Finally, $\boldsymbol{\theta}^*$ is accepted as an approximate sample from the posterior distribution if, according to some distance measure, $S(\boldsymbol{X}^*) \approx S(\boldsymbol{X}_{\text{obs}})$, where $S(\cdot)$ is a function that collapses the data into a lower-dimensional set of summary statistics. This algorithm implements what is arguably the most obvious naive solution to the problem of estimation. Loosely speaking, to guess which context (the true parameter) resulted in a given outcome (the observed dataset), simply simulate a number of pairs (contexts, outcomes) and identify the contexts that successfully replicated the outcome at hand. On the computational level, however, ABC is remarkably inefficient. Notice that one needs to repeatedly draw full datasets from the model.

Several papers have proposed solutions that improved the efficiency of ABC. Beaumont et al. (2002), in the landmark paper *Approximate Bayesian Computation in Population Genetics*, introduced the regression-adjustment technique, which we extensively exploit and extend throughout this thesis. The idea involves using the synthetic samples to fit regression models (of the form $\boldsymbol{\theta}|\boldsymbol{s}$) to project the accepted particles into the space of the exact partial posterior distribution $p(\boldsymbol{\theta}|\boldsymbol{s}_{\text{obs}})$. To lessen the ABC approximation error, this procedure explicitly models the effect of accepting samples that are similar but not identical to the observed summary statistics. More details about regression adjustment is provided later in this thesis. Marjoram et al. (2003) embedded ABC into a Markov Chain Monte Carlo algorithm in an attempt to increase the probability of generating a candidate in the region of high posterior density. Sisson et al. (2007) used ABC to build an approximate Sequential Monte Carlo (SMC) algorithm that targets the posterior through a sequence of intermediate target distributions. Fearnhead and Prangle (2012) proposed running a preliminary ABC analysis to roughly identify the support of the posterior and, in a second stage, restrict the synthetic sampling to this more compact region. Prangle (2016) implemented a *lazy* ABC that allows the algorithm to abandon some synthetic samples that are likely to take too long to complete or to produce a synthetic sample that would be rejected with very high probability. A reweighing step ensures that no additional layer of approximation is induced. Nott et al. (2014) proposed the marginal adjustment algorithm that mitigates the approximation error by correcting the ABC posterior marginals. Precisely, they advocate estimating the marginals based on a reduced set

of summary statistics and combine these estimates in a way that preserves the correlation structure of the standard ABC estimate. Li et al. (2017) defined a copula-based formulation that extended the marginal adjustment algorithm, improving the methods capacity to capture bivariate correlations.

All these proposals have greatly expanded the reach of ABC to make room for ever more challenging applications. Nevertheless, we have identified three major limitations that are still in place. First, ABC was confined to the case of finite-dimensional parameter vectors, and was therefore missing out on the fascinating advances made in area of nonparametric Bayesian models. Second, to boost computational efficiency, most ABC approaches are not able to properly estimate the posterior correlation – when applying regression-adjustment, for instance, each component is usually modeled by an independent univariate regression model. The adjustment can, in principle, be based on a multivariate specification, but that creates technical complications, particularly for non-Gaussian error distributions, that heavily limits the practical use of such implementations. Third, estimating high-dimensional models (in the order of hundreds of parameters) is still extremely challenging. This thesis proposes a list of solutions that partially address each of these issues.

Chapter 1 contains a detailed review of ABC, along with a brief discussion on related likelihood-free methods. In Chapter 2, we combine a number of well-established ABC methods to make possible the estimation of an intractable epidemiological model. This chapter aims to give the reader a sense of the technical challenges that often arise when using ABC in a real, complex problem. But most importantly, this study sheds light on some important biological questions in regards to the evolution dynamics that lead to multidrug-resistant tuberculosis.

Chapter 3 introduces a Bayesian nonparametric method for hierarchical modeling on a set of related density functions, where grouped data in the form of samples from each density function are available. Borrowing strength across the groups is a major challenge in this context. To address this problem, we propose a hierarchically structured prior, defined over a set of univariate density functions, using convenient transformations of Gaussian processes. Inference is performed with a novel functional regression adjustment ABC. The performance of the proposed method is illustrated via simulation studies and an analysis of rural high school exam performance in Brazil.

In Chapter 4, a new likelihood-free algorithm for recalibrating an approximate posterior

estimate is proposed. We show that it provides an useful additional layer of variation reduction to ABC methods. In other non-ABC analyses, it adapts the ABC machinery to correct deficiencies of an auxiliary estimation method. This work extends and strengthens the links between ABC and indirect inference algorithms, allowing a more extensive use of a misspecified auxiliary model. Illustrative examples are provided, along with a simulation study that investigates the effects of recalibration under various conditions. The proposed technique is applied to the estimation of the model parameters in a stereological extreme value problem.

We introduce, in Chapter 5, a Likelihood-free approximate Gibbs sampler. Precisely, we propose using ABC to approximate the so-called full conditional distributions, which are then embedded in an otherwise intractable Gibbs sampler. By breaking the problem (of estimating the full posterior) into smaller pieces (the full conditional distributions), we are able to fit substantially more challenging models than would be possible using vanilla ABC methods alone. We present two simulated toy examples that illustrate the cases where the regression models used to build the approximations are and aren't well specified. To analyze a dataset of *Airbnb* rental prices we also implement an intractable high-dimensional multivariate non-linear state space model.

Finally, we summarize the benefits and limitations of our contributions in the discussion section, where we also identify promising avenues for further exploration.

Chapter 1

Background

Approximate Bayesian Computation methods are a family of simulation algorithms designed to produce samples from an approximate posterior distribution. Such methods were proposed as an alternative to standard Monte Carlo techniques to make statistical inference possible for a class of applied problems where the likelihood function cannot be evaluated. This chapter summarizes some of the most influential ABC approaches, particularly those directly related to the content introduced in the following chapters. Section 1.1 considers a simple introductory example of estimation via ABC. Then, in Section 1.2, we explore more deeply the nature of each of the input elements that define an ABC sampler (such as the summary statistics), taking the opportunity to provide some guidelines in regards to the question of how to specify them in practice. To conclude, several ABC approaches are reviewed in Section 1.3.

1.1 Introductory example

A rejection-ABC algorithm was previously described in the Introduction section. We now turn our attention to a slightly more sophisticated version of ABC, known as ABC importance sampling. The underlying principle is common to both implementations – that is, to avoid direct evaluation of the likelihood by matching synthetic samples to an observed dataset. But rather than simply rejecting or accepting a candidate parameter, the latter algorithm generalizes the former by assigning an importance weight to each approximate posterior sample.

The importance weights are defined by three interrelated components. First, a function

to compute *summary statistics* from data, $S(\boldsymbol{X})$, is chosen to reduce the dimension of the objects being matched. Second, a *distance metric*, $D(\boldsymbol{s}, \boldsymbol{s}_{\text{obs}}) = \|\boldsymbol{s} - \boldsymbol{s}_{\text{obs}}\|$, is employed to measure the degree of similarity between a synthetic sample $\boldsymbol{s}$ and the observed summary statistics $\boldsymbol{s}_{\text{obs}}$. Third, a *weighting function*, $K_h(d)$, which is controlled by a bandwidth parameter h, sets the importance weights according to $D(\boldsymbol{s}, \boldsymbol{s}_{\text{obs}})$ – the bigger the distance the smaller the weight.

Algorithm 1 produces weighted samples from the ABC posterior approximation given by

$$\pi_{ABC}(\boldsymbol{\theta}|\boldsymbol{s}_{\text{obs}}) = \int K_h(\|S(\boldsymbol{x}) - \boldsymbol{s}_{\text{obs}}\|) p(\boldsymbol{x}|\boldsymbol{\theta}) \pi(\boldsymbol{\theta}) d\boldsymbol{x}. \tag{1.1}$$

It follows from Eq. (1.1) that if the analyst assigns weight proportional to one when the observed summary statistics are perfectly reproduced ($\|S(\boldsymbol{x}) - \boldsymbol{s}_{\text{obs}}\| = 0$) and zero otherwise, then Algorithm 1 generates exact samples from the partial posterior distribution $\pi(\boldsymbol{\theta}|\boldsymbol{s}_{\text{obs}})$. However, this choice makes the sampler hopelessly inefficient, as the probability of generating an acceptable match becomes prohibitively small. In general, specifying those three components is a delicate, non-trivial task.

Algorithm 1 ABC Importance Sampling (vanilla version)

Inputs:

- A target posterior density $\pi(\boldsymbol{\theta}|\boldsymbol{X}_{\text{obs}}) \propto p(\boldsymbol{X}_{\text{obs}}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$, consisting of a prior distribution $\pi(\boldsymbol{\theta})$ and a procedure for generating data under the model $p(\boldsymbol{X}_{\text{obs}}|\boldsymbol{\theta})$.
- An integer $\tilde{N} > 0$.
- An observed vector of summary statistics $\boldsymbol{s}_{\text{obs}} = S(\boldsymbol{X}_{\text{obs}})$.
- A kernel function $K_h(u)$ and scale parameter $h > 0$.

Sampling:
For $i = 1, \ldots, \tilde{N}$:

1. Generate $\boldsymbol{\theta}^{(i)} \sim \pi(\boldsymbol{\theta})$ from the prior.
2. Generate $\boldsymbol{X}^{(i)} \sim p(\boldsymbol{X}|\boldsymbol{\theta}^{(i)})$ from the likelihood.
3. Compute the summary statistics $\boldsymbol{s}^{(i)} = S(\boldsymbol{X}^{(i)})$.
4. Assign $\boldsymbol{\theta}^{(i)}$ the weight $w^{(i)} \propto K_h(\|\boldsymbol{s}^{(i)} - \boldsymbol{s}_{\text{obs}}\|)$.

Output: A set of weighted parameter vectors $\{(\boldsymbol{\theta}^{(i)}, w^{(i)})\}_{i=1}^{\tilde{N}} \sim \pi_{ABC}(\boldsymbol{\theta}|\boldsymbol{s}_{\text{obs}})$.

Hereafter we explore – in a graphical, intuitive level – a simulated toy example to give the reader a gentle introduction to the ABC machinery, with main focus on its operational aspects. For the sake of illustration, we consider a simple model characterized by a

sequence of *dependent* stochastic events. Precisely, each observation, x_i, $i = 1, \ldots, N$, is obtained by (a) drawing a sample from a normal distribution with mean θ and variance 1; and (b) if the value obtained in (a) is negative, multiplying it by -1 with probability $p = 0.5$. The fundamental feature being that the action, or lack of thereof, in step (b) is conditional on what happened in (a). Alternatively, this model can described by the following equations:

$$x_i = \begin{cases} |y_i|, & \text{with probability } p = 0.5. \\ y_i, & \text{with probability } 1 - p = 0.5. \end{cases}$$

$$y_i \sim N(\theta, 1).$$

We assigned a standard Gaussian prior for θ, which completes the (prior predictive) data generative process under consideration. Conditioned on $\theta = -1$, which we take as the "true" parameter value, we simulated 50 samples from the model to act as the "observed" dataset. The density function $p(X|\theta = -1)$ is depicted in the bottom-right plot in Figure 1.1.

The likelihood can be easily evaluated here. However, in more elaborated constructions, where the number of probabilistic events are substantially higher, the conditional structure becomes overwhelmingly complex, rendering the likelihood intractable (Luciani et al., 2009; Tanaka et al., 2006). The epidemiological model analyzed in Chapter 2, for example, assumes that, in each time step, several random events can take place for each member of the population, depending on their current status – subjects can get infected, recover, die, etc. The number of distinct ways in which the population could have reached its final, observed state is extraordinarily large.

In this study, the univariate data is summarized by its sample mean, $\bar{x}$, and standard deviation, $\hat{\sigma}$. The Euclidean distance and the Epanechnikov kernel were chosen as the distance metric and the weighting function respectively. Algorithm 1 was then used to generate $\tilde{N} = 1000$ approximate samples from the ABC posterior distribution for a number of tolerances h.

The upper-left plot in Figure 1.1 show the objects generated by Steps 1 to 3 of Algorithm 1. Each parameter $\theta^{(1)}, \ldots, \theta^{(\tilde{N})}$ is associated to two points in the plot – one representing the mean and the other the standard deviation of the corresponding synthetic

sample. The relationship between the parameter of interest and the summary statistics forms an interesting pattern. When θ is large, the probability of generating a negative sample y becomes negligible, so the density function effectively becomes a single normal distribution centered at θ. In this region, the sample mean is sufficient for θ, while the standard deviation becomes uninformative – notice the cluster formed around the true standard deviation $\sigma = 1$. On the other side of the coordinate axis, when θ is, say, less then -3, then each generated data point is reflected with probability 0.5, turning the density function into a mixture of two normals with opposite means. In this case, $\bar{x}$ estimates $E(X|\theta) \approx 0$, carrying no useful information about θ. The standard deviation, however, assumes a linear, negative relationship with θ.

The vertical lines represent the "observed" summary statistics. We see that the model can only reproduce $\hat{\sigma} \approx 1.5$ if θ is in a neighborhood of -1. The plots in the right-hand-side explore the relationship between θ and the computed Euclidean distances $\|\boldsymbol{s}^{(i)} - \boldsymbol{s}_{\text{obs}}\|$. Again, the observed patterns suggests that the distance can only be small if $\theta \approx -1$.

Step 4 of Algorithm 1 is illustrated by the following plots of Figure 1.1. We first set h so that 25% of the candidate parameters are assigned positive weights. As the majority of the original samples were discarded, the clouds of points now cluster more closely around the vertical lines. As h is further reduced, the samples get even closer the space over which the exact posterior is supported. On the other hand, the number of accepted samples also shrinks, which prompts an increase in the Monte Carlo error. This trade-off, which is a well-known feature of ABC methods, is observed in the bottom-left plot of Figure 1.1.

1.2 Constructing an ABC sampler in practice

For any given model, there are numerous ways to setup an ABC sampler. For a more extensive introduction to the ABC literature, see Sisson et al. (2017a); Beaumont (2010); Csillery et al. (2010); Lopes and Beaumont (2010); Bertorelle et al. (2010); Sisson and Fan (2011); Turner and Van Zandt (2012); Blum et al. (2013). As illustrated in our toy example, defining an appropriate weighting system is paramount to alleviate the ABC approximation error.

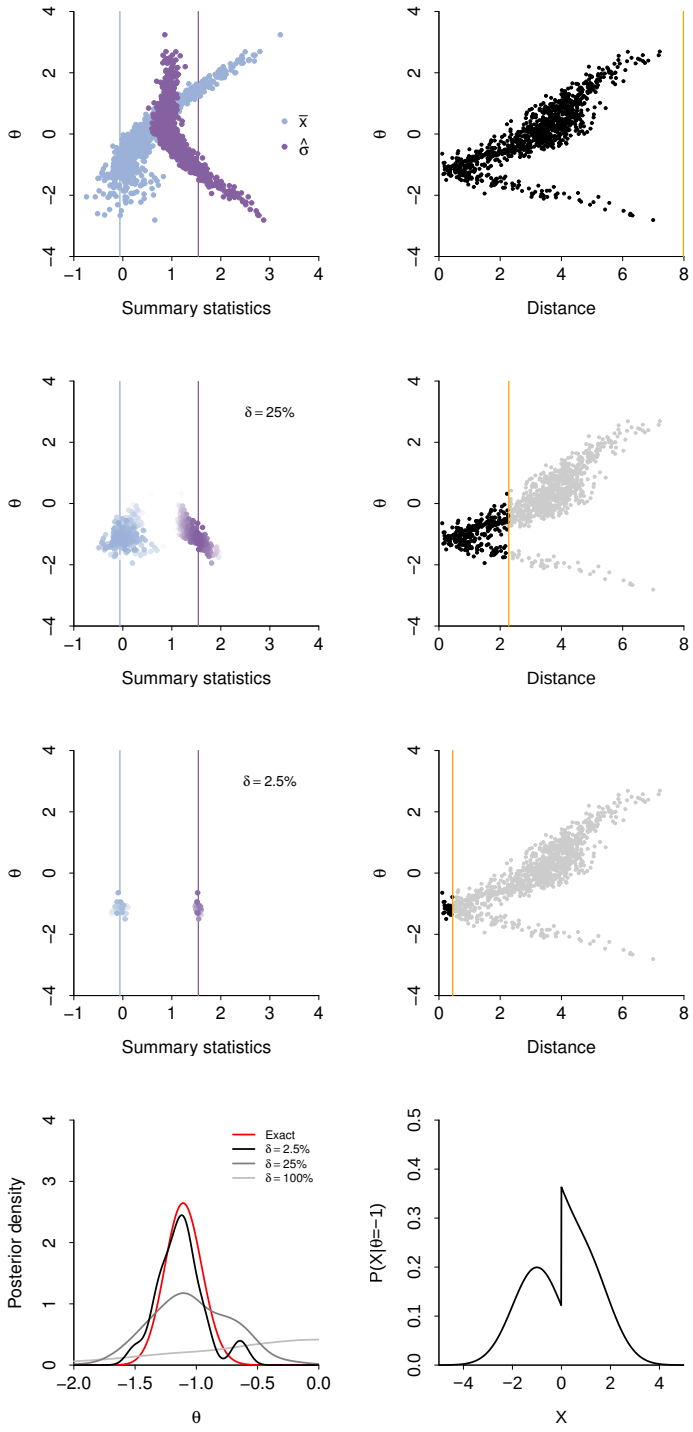


Figure 1.1: The upper-left plots show the synthetic (points) and the observed (vertical lines) summary statistics for different acceptance ratios δ. The color intensity reflects the importance weights (calculated from the Epanechnikov kernel function), with plain white representing zero weight. The corresponding plots in the right-hand-side present the Euclidean distances between synthetic and observed summary statistics. Samples in light gray were assigned weight zero (and therefore rejected). The orange line represents the tolerance h induced by δ. The bottom-left plot compares the exact (red) and the ABC approximate posteriors (in gray). The bottom-right plot shows the density function for data, conditioned on the "true" parameter value $\theta = -1$.

1.2.1 Summary statistics

Ideally, a set of *sufficient* summary statistics should be adopted, so that all available information is retained – in the sense that $p(\boldsymbol{\theta}|\boldsymbol{X}_{\text{obs}}) = p(\boldsymbol{\theta}|\boldsymbol{s}_{\text{obs}})$. In addition, to increase the sampler's computational efficiency, the dimension of $\boldsymbol{s}_{\text{obs}}$ should be kept as low as possible. That mitigates the so-called curse of dimensionality (Blum, 2010), which is a major problem in the ABC context. This term refers to the practical difficulty in generating synthetic samples that are "close" to the observed data when more than a few summery statistics are considered. Or, from a more theoretical perspective, it means that the rate in which $\pi_{ABC}(\boldsymbol{\theta}|\boldsymbol{s}_{\text{obs}})$ converges to $\pi(\boldsymbol{\theta}|\boldsymbol{s}_{\text{obs}})$, as a function of h, decreases dramatically with the dimension of the summary statistic. In practice, fully accommodating these two desirable properties (sufficiency and low dimension) is not generally possible. The ABC literature was therefore compelled to propose various approaches to sensibly compact the data (see Blum et al., 2013).

In our introductory example, we summarized univariate datasets by their sample mean and standard deviation. This choice was *inspired by statistical analogies*. In Chapter 2.4.1 (Rodrigues et al., 2017a), we again follow this approach and condense the data, which is akin to a design matrix of a regression model, by the unique elements of

$$(\mathbf{1}|\mathbf{X})^{\top}(\mathbf{1}|\mathbf{X}),$$

where the vertical lines denote the addition of an extra column.

Domain knowledge is commonly abundant in statistical practice. Experts often know beforehand what features of the data are related to the model parameters (Luciani et al., 2009). In our toy example, because the negative samples change sign with a known probability, the proportion of negative data points, for instance, is potentially informative for θ.

Point estimators may still be available even when the likelihood is intractable. In Section 5.4, we define as summary statistics the L-moment estimates (Peters et al., 2016) of the parameters of the g-and-k distribution (Haynes, 1998). This choice results in a reliable and easy-to-handle (in post-processing techniques) low-dimensional vector.

The principles of *indirect inference* (Gourieroux et al., 1993; Gleim and Pigorsch, 2013) can also be employed in the ABC context (Drovandi et al., 2015; Martin et al., 2017;

Drovandi et al., 2017). This powerful method makes use of an auxiliary, misspecified, tractable model that offers a simplified representation of its intractable counterpart. The maximum likelihood estimate (MLE), or maximum a posteriori probability (MAP) estimate, of the auxiliary model is then used as a summary statistic in the ABC machinery. Interesting applications of this technique are presented in Subsections 4.3.1 and 4.4. In our introductory example of the previous section, one could assume that the likelihood is sufficiently well approximated by a mixture of two Gaussian. The summary statistics would then be defined as the MLE of the parameters of this misspecified model.

Non-parametrically representing the dataset is yet another way of deriving suitable summary statistics. For instance, one could exploit the whole Kernel Density Estimator (KDE), rather than just recording the mean and standard deviation. That is precisely what we do in Section 3.3.1.

Instead of using one of the described approaches to directly define the summary statistics, it is sometimes convenient to start with a large set of potentially informative features, and then, in a second stage, reduce the dimension of $\boldsymbol{s}$ using a *best subset selection* approach. This can be executed based on various criteria, including measures of sufficiency (Joyce and Marjoram, 2008), entropy of the posterior distribution (Nunes and Balding, 2010) and Akaike (AIC) and Bayesian Information Criteria (BIC) (Blum et al., 2013).

In an attempt to simplify the process of constructing appropriate summary statistics, Fearnhead and Prangle (2012) proposed a semi-automatic approach. The authors advocate using regression models and a preliminary ABC analysis to establish an estimator for the posterior mean, as a function of data. Then, in a second round of ABC, for each synthetic sample, they estimate the associated posterior mean, which is taken as the actual summary statistic.

1.2.2 Distance metric

Next, once the observed data has been collapse down into a low-dimensional vector, one needs to set a distance metric. The Euclidean distance (otherwise known as L_2 distance) is an obvious, but often suboptimal, option. In particular, when the summary statistics are on substantially different scales (under the *true* model), the distances $\|\boldsymbol{s}^{(i)} - \boldsymbol{s}_{\text{obs}}\|$ are dominated by the highest variability components of $\boldsymbol{s}$. In that case, even if sufficient summary statistics are in place, the algorithm will perform poorly.

An appealing remedial measure to this problem is to standardize the entries of the summary statistics. However, that requires some knowledge of the variability of $\boldsymbol{s}$, conditional on the true parameter $\boldsymbol{\theta}$, which is obviously not available in advance.

There are a few documented ways to implement the standardization (e.g., Erhardt and Sisson, 2016). The *abc* R function (Csilléry et al., 2012), for instance, divides, by default, the components of $\boldsymbol{s}$ by the corresponding Median Absolute Deviation (MAD), computed from the synthetic samples. Therefore, the standardization is based on the variability of the prior predictive distribution, which may be considerably different than the one of the true model.

Alternatively, a normalization may be performed implicitly, within the Mahalanobis distance metric defined as

$$D(\boldsymbol{s}, \boldsymbol{s}_{\text{obs}}) = \sqrt{(\boldsymbol{s} - \boldsymbol{s}_{\text{obs}})' \Sigma_s^{-1} (\boldsymbol{s} - \boldsymbol{s}_{\text{obs}})} \,,$$

where Σ_s denotes the covariance matrix of $\boldsymbol{s}|\boldsymbol{\theta}$. Compared to the approach based on the MAD, this has the advantage of taking into account the whole correlation structure. To implement the normalization, Luciani et al. (2009) estimated Σ_s using samples generated from $p(\boldsymbol{s}|\hat{\boldsymbol{\theta}})$, where $\hat{\boldsymbol{\theta}}$ was an initial estimate of $\boldsymbol{\theta}$. Erhardt and Sisson (2016) showed that this procedure outperformed the ABC implementation based on the Euclidean distance on a stereological application.

Gutmann et al. (2017) framed the problem of quantifying the similarity between two datasets as a classification problem – the idea involves resampling (or slicing) the observed summary statistics $\boldsymbol{s}_{\text{obs}}$ to create a collection of k replicates $\tilde{\boldsymbol{s}}^1_{\text{obs}}, \ldots, \tilde{\boldsymbol{s}}^k_{\text{obs}}$. Then, the the distance $D(\boldsymbol{s}^{(i)}, \boldsymbol{s}_{\text{obs}})$ is defined in terms of how accurately a (machine learning) classifier can distinguish $\tilde{\boldsymbol{s}}^1_{\text{obs}}, \ldots, \tilde{\boldsymbol{s}}^k_{\text{obs}}$ from the (resempled) synthetic samples $\tilde{\boldsymbol{s}}^{(i)1}, \ldots, \tilde{\boldsymbol{s}}^{(i)k}$.

In Chapter 3, due to the functional nature of the summary statistics, we used the sum of Kullback-Leibler (KL) divergences to measure the degree of separation (not distance in the strict sense) between sets of density functions.

1.2.3 Weighting function

Throughout this thesis, we followed Beaumont et al. (2002) and took the Kernel function to be the Epanechnikov kernel, defined as $K_h(D) \propto 1 - (D/h)^2$ if $D < h$ and zero otherwise.

This function has the convenient property of assigning zero weight to all approximate posterior samples for which the distance between the corresponding synthetic dataset and the observed summary statistic is greater than h. That means that unsatisfactory samples are promptly discarded, enhancing the computational efficiency (Fan and Zhang, 1999) and reducing storage requirements.

Several other kernel weighting functions are available in the statistical literature. Nevertheless, like other kernel density estimators, the ABC error is dominated by the choice of h, rather then the actual form of $K_h(D)$ (Blum, 2010). In practice, the bandwidth is more naturally specified indirectly, through the acceptance fraction δ, which determines the ratio of synthetic samples that are assigned positive weight.

For a fixed computational budget and ABC specification, finding the tolerance h that minimizes the ABC error is impractical in all but the most trivial cases. In this context, cross validation techniques allow the user to asses the effect of the choice of h on the quality of the ABC estimate (Csilléry et al., 2012). These procedures involve using the synthetic samples to compare several simulated ABC estimates (one for each of a number of test sets) with the corresponding "true" parameters (which are known by construction) in a standard leave-one-out approach.

Prangle et al. (2014) proposed a diagnostic to test whether the ABC posterior marginals are well calibrated, in that the coverage property (as defined in Cook et al., 2006; Fearnhead and Prangle, 2012; Prangle et al., 2014) is satisfied. If, for a given value of h the test rejects the hypothesis that the approximate marginals are properly calibrated, then the authors suggest further reducing the acceptance fraction (which comes at the expense of the Monte Carlo error). In Chapter 4, we present an algorithm that exploits the diagnostic output of such test to perform a *recalibration-adjustment* to the samples from an ABC posterior approximation.

Bortot et al. (2007) treated the bandwidth as an unknown model parameter (see also Ratmann et al., 2009), and then estimated it from the data within an ABC-MCMC sampler (which we briefly describe in Section 1.3). Sisson et al. (2007), in turn, introduced a Sequential Monte Carlo that allows the tolerance to be reduce gradually, through a sequence of ABC importance sampling stages (this technique is also considered in Section 1.3).

1.3 ABC algorithms

The family of approximate Bayesian computation methods is rapidly growing, with new schemes being proposed to boost algorithmic efficiency, reduce the approximation error and extend the reach of likelihood-free models. For an up to date overview on ABC methods, see Sisson et al. (2017a); Lintusaari et al. (2017).

For clarity of exposition only, we divide the ABC algorithms into three non-exclusive categories. The first being composed by post-precessing techniques where the posterior samples are adjusted in beneficial ways. The second class covers the strategies designed to sample the candidate parameters from a more appropriate proposal distribution – and therefore increase the probability of generating an acceptable match in the ABC machinery. Lastly, we consider the group of methods that, in different levels, address the curse of dimensionality problem, fostering the capacity of ABC methods to accommodate high dimensional models.

1.3.1 Post-processing adjustments

Based on distinct rationales, all methods described in this subsection share the same underlying concept of moving the approximate posterior samples in an attempt to improve the quality of the initial approximation.

We dedicate special attention to regression-adjustment techniques (Beaumont et al., 2002; Blum and François, 2010; Blum et al., 2013) as these are exploited or extended in each and every chapter of this thesis. To motivate their use, we revisit our introductory example. For the uninitiated reader, Beaumont (2010) and Csillery et al. (2010) describe the technique in a rather gentle and visual manner.

In the central-left plot in Figure 1.2, we re-reproduce (with rescaled axes) the weighted synthetic samples for an acceptance rate of $\delta = 25\%$. However, for ease of explanation, we only consider here the standard deviation summary statistic $\hat{\sigma}$ (and not $\bar{x}$). Our ultimate goal is to generate exact samples from the posterior distribution, which, as previously noticed, is restricted to the space depicted by the vertical line. Although no point falls exactly in that region, a suitable regression model (from which the orange lines are created) can be built to project the samples so that $\theta^{(i)}$ behaves as an approximate sample from $\pi(\theta|s_{\text{obs}})$, rather than an exact sample from $\pi(\theta|s^{(i)})$. This procedure therefore belongs to

the class of conditional density estimation methods.

Beaumont et al. (2002) assumed a local linear model of the form

$$\boldsymbol{\theta}^{(i)} = m(\boldsymbol{s}^{(i)}) + \boldsymbol{\epsilon}^{(i)}, \tag{1.2}$$

where $m(\boldsymbol{s}^{(i)}) = \mathbb{E}(\boldsymbol{\theta}|\boldsymbol{s}^{(i)}) = \boldsymbol{\alpha} + \boldsymbol{\beta}'\boldsymbol{s}^{(i)}$, and the $\boldsymbol{\epsilon}^{(i)}$ are independent zero-mean random variables with common variance. The regression model parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are estimated by the weighted least squares method – that is, they are obtained by minimizing $\sum_{i=1}^{\tilde{N}} w^{(i)} \| m(\boldsymbol{s}^{(i)}) - \boldsymbol{\theta}^{(i)} \|$.

The approximate posterior samples are then adjusted to

$$\tilde{\boldsymbol{\theta}}^{(i)} = \hat{m}(\boldsymbol{s}_{\text{obs}}) + (\boldsymbol{\theta}^{(i)} - \hat{m}(\boldsymbol{s}^{(i)})) \tag{1.3}$$

$$= \boldsymbol{\theta}^{(i)} + \hat{\boldsymbol{\beta}}'(\boldsymbol{s}_{\text{obs}} - \boldsymbol{s}^{(i)}), \tag{1.4}$$

where $\hat{m}(\boldsymbol{s}_{\text{obs}}) = \hat{\mathbb{E}}(\boldsymbol{\theta}|\boldsymbol{s}_{\text{obs}}) = \hat{\boldsymbol{\alpha}} + \hat{\boldsymbol{\beta}}'\boldsymbol{s}_{\text{obs}}$ is the estimated posterior mean, and the added term represents the ith empirical residual. The geometrical interpretation of the orange lines in Figure 1.2 are more easily understood through Eq. (1.4) – the correction depends on the local linear model only though the estimated slope parameter $\hat{\boldsymbol{\beta}}'$. If the regression model in Eq. (1.2) perfectly describes the underlying probabilistic relationship (in the region such that $D(\boldsymbol{s}, \boldsymbol{s}_{\text{obs}}) < h$), then the ABC sampler produces exact samples from $p(\boldsymbol{\theta}|\boldsymbol{s}_{\text{obs}})$.

The model assumptions are generally not fully satisfied in practice. Figure 1.2 illustrates the role played by the acceptance rate in this context. Typically, when δ is large, the regression model does not provide an adequate representation of the conditional distributions, leading to substantial systematic biases (upper-left plot). On the other end, when δ is very low and the points are already close to $\boldsymbol{s}_{\text{obs}}$, the projection barely affects the samples (the regression model nearly degenerates into a constant model, as seen in the bottom-left plot). This behavior is also observed in Section 4.3.2, where we estimate the parameters of a "twisted normal" distribution. The plots in the right-hand-side in Figure 1.2 show the adjustment effect in each case.

Blum and François (2010) extended the local linear model considered above in two meaningful ways. First, they proposed an heteroscedastic construction to explicitly account for unequal error variances. Second, they suggested using neural networks to esti-

mate non-linear conditional mean $m(\cdot)$ and variance $\sigma^2(\cdot)$ functions. Adding to the effort of making regression-adjustment more robust and suitable for a wider class of intractable models, Blum et al. (2013) recommended using ridge regression to perform the projections, which avoids the risk of over-adjusting the posterior samples in the direction of uninformative summary statistics. Chapter 3 (Rodrigues et al., 2016), in turn, extends regression-adjustment to allow the estimation of infinite-dimensional parameters. This requires fitting functional regression models (Ramsay and Silverman, 2005) in which the dependent and the independent variables represent functional objects, rather than vectors.

Gutmann and Corander (2016) recently introduced a method that uses Bayesian optimisation techniques to make ABC more efficient. In this approach, the relationship between the parameters $\boldsymbol{\theta}$ and the distance metric $D(\boldsymbol{s}, \boldsymbol{s}_{\text{obs}})$ is modeled. Therefore, while regression-adjustment focuses on the information contained in the left-hand-side plots of Figure 1.1, the Bayesian optimization for likelihood-free inference (BOLFI) exploits the relationship shown in the right-hand-side plots.

Marginal adjustment (Nott et al., 2014) is yet another useful method that allows the user to improve the precision of the marginal estimates while retaining the correlation structure of a standard ABC estimator. The idea is to use ABC to estimate each posterior marginal separately, using summary statistics that only need to be informative for the individual parameter under consideration. These marginal estimates will be more accurate simply because the lower dimensionality of the summary statistics reduces the curse of dimensionality ABC error. The ABC samples (based on the full set of summary statistics) are then shifted to match the individually fitted approximate marginals.

The effect of post-processing adjustments on the quality of the approximation is hard to asses in practice. We propose in Chapter 4 a post-processing technique that *recalibrates* an ABC estimator (or in fact any other Bayesian marginal posterior estimator), so that the new estimate approximately satisfies the coverage property.

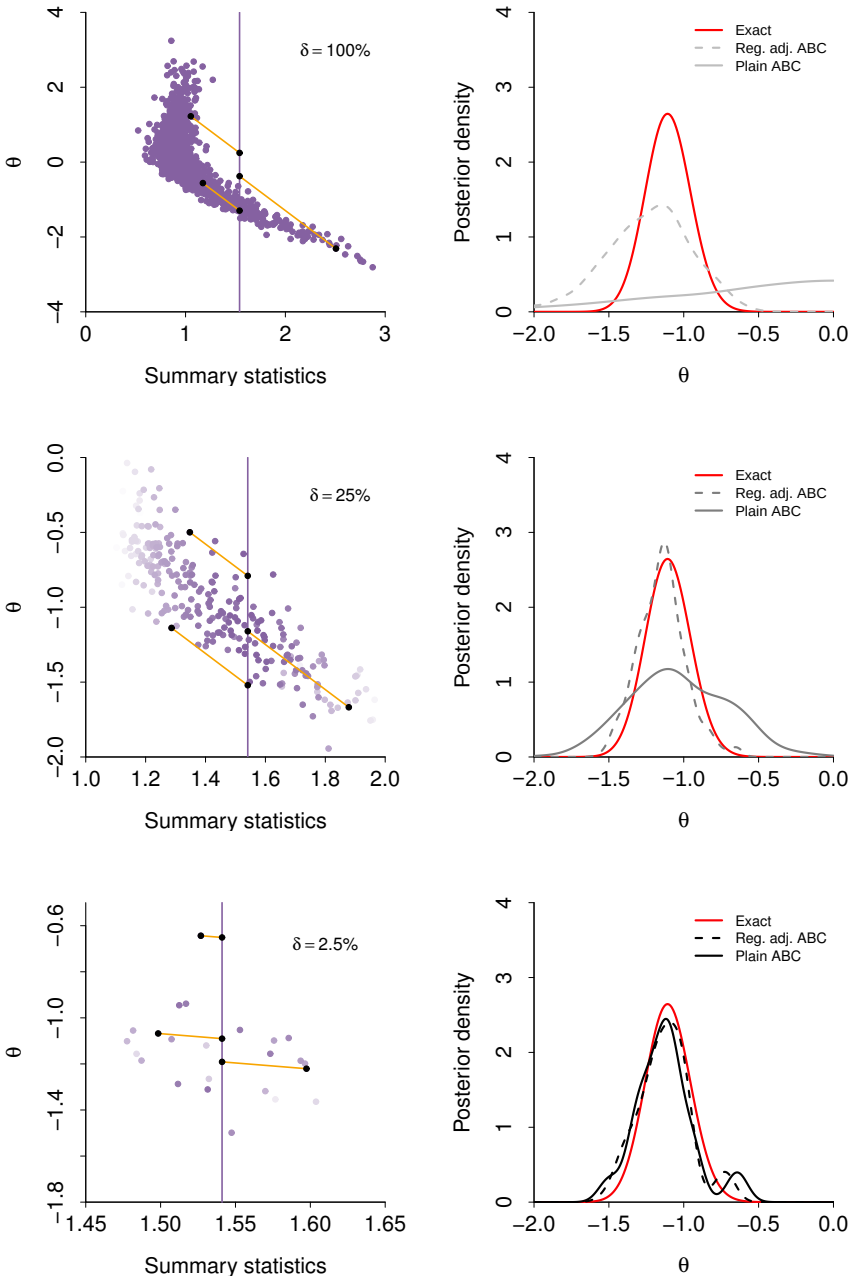


Figure 1.2: The right-hand-side plots illustrate the mechanics of regression adjustment for different acceptance ratios. The orange lines represent the projection induced by the assumed linear model. Notice the different axis scales and the connection to Figure 1.1. Plots in the right-hand-side overlays the ABC approximations before and after applying the linear regression adjustment.

1.3.2 Methods for improving the proposal distribution

As previously mentioned, one of the main sources of error in ABC is that most of the parameter values generate synthetic samples that are substantially different then the observed summary statistics, particularly if the prior is diffuse with respect to the posterior. Several families of methods have been created to tackle this particular problem.

Marjoram et al. (2003), for example, proposed the MCMC ABC algorithm. Similarly to regular MCMC, in each iteration, a candidate parameter $\boldsymbol{\theta}^*$ is drawn from a proposal distribution $q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(i-1)})$, conditional on the current parameter value $\boldsymbol{\theta}^{(i-1)}$. However, to avoid direct evaluation of the likelihood, a synthetic dataset $\boldsymbol{s}^*$ is drawn from $p(\boldsymbol{s}|\boldsymbol{\theta}^*)$. Then, if $D(\boldsymbol{s}^*, \boldsymbol{s}_{\text{obs}}) < h$, $\boldsymbol{\theta}^*$ is accepted (i.e. $\boldsymbol{\theta}^{(i)} = \boldsymbol{\theta}^*$) with a Metropolis-Hastings probability. Otherwise the chain remains unchanged ($\boldsymbol{\theta}^{(i)} = \boldsymbol{\theta}^{(i-1)}$). The MCMC ABC algorithm indeed targets the partial posterior distribution (Marjoram et al., 2003).

After convergence, the chain is expected to stay in the region of high posterior density, and therefore increase the probability of generating an acceptable match (as compared to rejection ABC). However, it easily gets "stuck" in the tails of the posterior distribution, as the condition $D(\boldsymbol{s}^*, \boldsymbol{s}_{\text{obs}}) < h$ will hardly be satisfied for an $\boldsymbol{s}^*$ sampled from a parameter in this region. The added synthetic simulation step also hampers the efficiency of the algorithm, given that numerous synthetic datasets may need to be generated in each MCMC iteration.

Fearnhead and Prangle (2012) suggested using a preliminary ABC analysis to roughly learn the region $A(\boldsymbol{\theta})$ where the posterior has non-negligible mass. Then, in a second round of ABC, the simulated samples are drawn from the truncated distribution $\pi(\boldsymbol{\theta})I(\boldsymbol{\theta} \in A)$, which avoids time being wasted on samples that would almost certainly be rejected anyway.

The importance sampling ABC algorithm can be slightly adapted to allow the parameters to be generated from a more appropriate proposal distribution $q(\boldsymbol{\theta})$ (this implementation is detailed in Section 2.4). The sample weights are then set in a way that preserves $p(\boldsymbol{\theta}|\boldsymbol{s}_{\text{obs}})$ as the target distribution. As with standard importance sampling, $q(\boldsymbol{\theta})$ should be as close as possible to the unknown target distribution. In practice, however, finding a suitable $q(\boldsymbol{\theta})$ is not always easy. Considering that, Sisson et al. (2007) introduced the sequential Monte Carlo ABC (SMC ABC) that provides a structured framework to address this problem. In their scheme, a sequence of importance samplers is constructed in a way

in which the proposal distribution gradually improves, becoming systematically closer to the $p(\boldsymbol{\theta}|\boldsymbol{s}_{\text{obs}})$. A graphical illustration of this technique is available in Lintusaari et al. (2017).

1.3.3 Mitigating the curse of dimensionality

All strategies described so far in this chapter are useful on their own or combined to other methods. Nevertheless, even if we could draw candidate parameters from the best possible proposal distribution (that is, the unknown posterior distribution itself), the curse of dimensionality would still be present, imposing serious constraints in high-dimensional models. Recognizing this fact, Prangle et al. (2016) proposed a clever sampling procedure (named RE-ABC) that exploits *rare event* methods to estimate the likelihood function, using this within the pseudo-marginal algorithm of Andrieu and Roberts (2009).

The concept of employing (auxiliary) statistical models to approximate the intractable (target) model of interest has been developed in various forms. All these methods empirically learn about (and correct for) the stochastic difference between the auxiliary and the target models, otherwise one would be merely performing inference based on the simplified (tractable but inadequate) representation of the phenomenon under consideration. Examples of such model-based approximations include Bayesian indirect inference (Drovandi et al., 2015, 2017), variational Bayes (Tran et al., 2017), synthetic likelihoods (Wood, 2010; Ong et al., 2016), Gaussian mixture models (Bonassi et al., 2011), Gaussian processes (Gutmann and Corander, 2016), Gaussian copula models (Li et al., 2017) and regression density estimation (Fan et al., 2013).

Other techniques have been specifically designed to exploit the conditional architecture of the likelihood. The reasoning here being that, whenever the intractable model can be factorized into lower dimensional components, the summary statistics vector should also allow a corresponding segmentation. For instance, in the estimation of a Bayesian hierarchical model with $\boldsymbol{\theta} = (\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(G)}, \boldsymbol{\alpha})$, where $\boldsymbol{\alpha}$ denotes the hyperparameters and $\boldsymbol{\theta}^{(g)}$ are the parameters associated to group g, Bazin et al. (2010) notice that the posterior distribution decomposes as $p(\boldsymbol{\theta}, \boldsymbol{\alpha}|\boldsymbol{X}) = p(\boldsymbol{\alpha}|\boldsymbol{X}) \prod_{g=1}^{G} p(\boldsymbol{\theta}^{(g)}|\boldsymbol{\alpha}, \boldsymbol{X})$. Therefore, one may summarize $\boldsymbol{X}$ into different groups of summary statistics, one for each component of $\boldsymbol{\theta}$, so that the summary statistics comparisons are made in terms of lower dimensional vectors. See also the expectation-propagation scheme by Barthelmé and Chopin (2014).

Kousathanas et al. (2016) reformulated the MCMC ABC approach to increase the algorithm's acceptance rate. Their method combines two related innovations. First, they suggest using a transition kernel that updates one parameter at a time. The proposal is then accepted or rejected depending on the distance computed over a (low-dimension) vector of summary statistics – possibly specified as in Fearnhead and Prangle (2012) – that only needs to be conditionally informative for the parameter being currently updated. The method outperformed MCMC ABC by a sizable margin in some simulation studies. Nevertheless, the cost of generating multiple synthetic datasets (some of which are unavoidably rejected) for each successful (individual) parameter update creates a heavy computational burden.

In Chapter 5 we propose an approximate Gibbs sampler that combines both concepts described in this section, namely, using auxiliary approximate models and decomposing the posterior distribution in a beneficial manner. As in Kousathanas et al. (2016), we also update one (or a few) parameter at a time, but the algorithms behave remarkably differently. In our formulation, the ABC machinery is employed to approximate the full conditional distributions, which then serve as the MCMC transition kernel. An important advantage of our method is that once the approximations have been established, the MCMC does not require simulation of new synthetic datasets.

As discussed above, ABC undoubtedly faces considerable challenges, and there is much to be done. This is nevertheless a highly active area of statistical research, with numerous opportunities to further develop and strengthen the existing methodologies. In the following chapters we make new contributions in the areas of a) multidrug-resistant Tuberculosis (MDR-TB), b) nonparametric density estimation, intractable hierarchical models, infinite dimensional regression-adjustment, c) post-processing ABC techniques, d) approximate Gibbs sampling for high dimensional ABC.

Chapter 2

Inferences on the acquisition of multidrug resistance in *Mycobacterium tuberculosis* using molecular epidemiological data [1]

2.1 Introduction

Tuberculosis (TB) is a lung disease caused by the bacterium *Mycobacterium tuberculosis* which kills around 1.5 million people each year and remains a serious challenge for global public health (WHO, 2015). Antibiotic drugs for treating TB have been available since the mid 20th century, and currently implemented strategies for TB control rely on the efficacy of these drugs. Treatment of TB involves combination therapy – in which multiple drugs are administered together in part to improve killing efficacy. The "first-line" drugs used in combination to treat tuberculosis are rifampicin, isoniazid, pyrazinamide, ethambutol and streptomycin.

As with most other pathogens, resistance to antibiotic drugs has rapidly evolved in *M. tuberculosis*. Streptomycin was the first of the first-line drugs to be developed and deployed in 1943, but resistance was observed before the end of that decade (Mitchison,

[1] Published as: Rodrigues, G. S., Francis, A. R., Sisson, S. A., and Tanaka, M. M. (2017). Inferences on the acquisition of multidrug resistance in Mycobacterium turbeculosis using molecular epidemiolgical data. In Sisson, S. A., Fan, Y., and Beaumont, M. A. (Eds.), *Handbook of Approximate Bayesian Computation*, in press, Chapman and Hall/CRC Press.

1951; Gillespie, 2002). Of particular concern is the rise of bacterial strains resistant to multiple drugs, as cases caused by them are difficult to treat successfully. Multidrug resistance (MDR) is defined as resistance to both rifampicin and isoniazid. These are the two most effective drugs against tuberculosis (when the strain is not resistant). Currently, 3.3% of new TB cases are multi-drug resistant (WHO, 2015). The occurrence of MDR-TB strains that have additional resistance (called extensively drug resistant, XDR and totally drug resistant, TDR) are particularly problematic and have the potential to cause large outbreaks that are difficult to control (Gandhi et al., 2006). A better understanding of how multiple drug resistance evolves would aid efforts to contain resistance and control tuberculosis.

Genetic studies have established that many independent mutation events have led to resistance (Ramaswamy and Musser, 1998). Although this suggests that mutation of genes is an important source of resistance, model-based analysis of molecular data has revealed that among resistant cases, most are due to the transmission of already resistant bacteria (Luciani et al., 2009). It is therefore of interest to investigate whether or not this finding also holds for multi-drug resistant tuberculosis.

The rates at which resistance evolves against different drugs vary. For instance, isoniazid resistance is known to be acquired faster than rifampicin resistance (Ford et al., 2013; Gillespie, 2002; Nachega and Chaisson, 2003). The rates of mutation to resistance per cell generation are low in absolute value; for example for isoniazid the rate is around 3×10^{-8} and for rifampicin it is around 2×10^{-10} (David, 1970; Gillespie, 2002), although there is a high degree of variation across different lineages of *M. tuberculosis* (Ford et al., 2013). One might therefore expect that double resistance of these drugs (MDR) evolves at an exceedingly low rate (Nachega and Chaisson, 2003). However, MDR strains often occur at appreciable frequencies (Zhao et al., 2012; Anderson et al., 2014) and a recent study has presented a theoretical model showing how double resistance can evolve rapidly within hosts (Colijn et al., 2011). It would be useful to establish whether such fast direct acquisition of double resistance can be detected in bacterial isolates from epidemiological studies.

To characterise patterns of TB transmission and drug resistance in a given geographic region, bacterial isolates from TB patients are often genotyped using molecular markers known as variable numbers of tandem repeats (VNTRs) which are repeated genetic se-

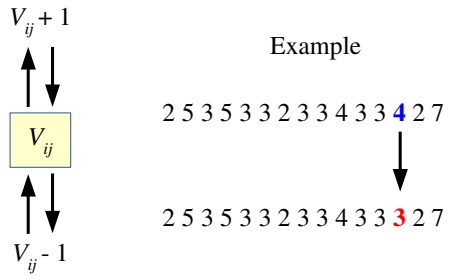


Figure 2.1: VNTR loci mutate in a stepwise manner so that the number of repeat units at a locus increases or decreases. In our analysis we assume that when mutation occurs at a locus j in genotype i, the repeat number V_{ij} increases or decreases by a single copy. We further assume that a single unit (repeat number of 1) is an absorbing boundary. The hypothetical example shows how mutation at locus number 13 creates a new VNTR genotype.

quences that exhibit variation across isolates. The source of this variation is mutation at the VNTR genetic loci which leads to the expansion or contraction of repeat numbers at those loci (Figure 2.1). A scheme for discriminating effectively among a set of isolates involves considering repeat numbers at multiple VNTR sites. This molecular typing scheme is called multi-locus VNTR analysis (MLVA); in the context of tuberculosis epidemiology it is often known as mycobacterial interspersed repetitive units-VNTR (MIRU-VNTR) (Mazars et al., 2001; Supply et al., 2006). Typing techniques such as MLVA have been useful for tracking particular strains and understanding how drug resistance evolves and disseminates at the epidemiological level (Monteserin et al., 2013; Anderson et al., 2014).

Here, we investigate the rates of drug resistance acquisition in a natural population using molecular epidemiological data from Bolivia (Monteserin et al., 2013). First, we study the rate of direct acquisition of double resistance from the double sensitive state within patients and compare it to the rates of evolution to single resistance. In particular, we address whether or not double resistance can evolve directly from a double sensitive state within a given host. Second, we aim to understand whether the differences in mutation rates to rifampicin and isoniazid resistance translate to the epidemiological scale. Third, we estimate the proportion of MDR TB cases that are due to the transmission of MDR strains compared to acquisition of resistance through evolution. To address these problems we develop a model of TB transmission in which we track the evolution of resistance to two drugs and the evolution of VNTR loci. However, the available data (see Section 2.2) is incomplete, in that it is recorded only for a fraction of the population and at a single point in time. The likelihood function induced by the proposed model is computationally prohibitive to evaluate and accordingly impractical to work with directly. We therefore

approach statistical inference using approximate Bayesian computation techniques.

2.2 Data

The data set we use is taken from a study of tuberculosis in Bolivia (Monteserin et al., 2013). Bolivia has a population of 11 million people and a TB incidence of 120 per 100,000 per year. This rate is comparable to the global incidence of TB (133 per 100,000 per year) and to the rate in Peru, but is 3–6 times the TB incidence in neighbouring countries Brazil, Paraguay, Uruguay, Argentina and Chile (WHO, 2015). In the molecular epidemiological study, the investigators genotyped 100 isolates collected in 2010, which represented an estimated 1.1% of the cases in Bolivia at the time of the study (Monteserin et al., 2013). Each isolate was tested for drug sensitivity to five drugs. Here, we focus on resistance against the two drugs isoniazid and rifampicin used to define multidrug resistance. Of the 100 isolates, 14 were found to be MDR, that is, resistant to both of these drugs, 78 were sensitive to both drugs and the remaining 8 were resistant to isoniazid but sensitive to rifampicin. No isolates were resistant to rifampicin while being sensitive to isoniazid.

In addition to these drug resistance profiles, each isolate was genotyped using 15 VNTR loci. For example, an isolate in the data set, which was resistant to isoniazid but sensitive to rifampicin, had the following 15 repeat numbers for its 15 VNTR loci: 143533233433527, which together constitute its genotype. Variation in these genotypes occurs through a process of mutation in which repeat numbers increase or decrease (see Figure 2.1).

Let g be the number of distinct genotypes present in a sample, and label the resistance profiles by (0, INH, RIF, MDR), where 0 denotes sensitivity to both drugs, INH denotes resistance to isoniazid and sensitivity to rifampicin, RIF denotes resistance to rifampicin and sensitivity to isoniazid, and MDR denotes resistance to both drugs. The observed data $\mathbf{X}_{obs}$ are then a $g \times 4$ matrix of counts, such that each row gives the distribution of isolates across the four resistance profiles for a given genotype and each column gives the distribution of isolates across genotypes for a given resistance profile. The sum of entries in a particular row is the number of isolates with that genotype, while the sum of entries in a particular column is the number of isolates with that resistance profile. The data set also includes a $g \times 15$ matrix of repeat numbers from the VNTR genotyping.

The Bolivian data set is displayed in full in Table 2.1, which shows all $g = 66$ distinct

genotypes and classifies all 100 isolates according to genotype and resistance profile. The $\mathbf{X}_{obs}$ matrix is formed by combining the 0, INH, RIF and MDR columns.

Genotype	**0**	INH	RIF	MDR	**Genotype**	**0**	INH	RIF	MDR
253533233433427	4	0	0	0	243413342212437	1	0	0	0
253533233433327	3	0	0	0	233312442212437	1	0	0	0
253533233433527	11	1	0	0	233313441212437	1	0	0	0
253533233433525	3	0	0	0	233413442212248	1	0	0	0
143533233433527	2	1	0	0	233413442212249	1	0	0	0
253333244232232	1	0	0	0	233213442212349	1	0	0	0
25333324423-232	1	0	0	0	231413542212335	1	0	0	0
254333243232342	0	2	0	2	232433242212436	1	0	0	0
263532232423139	3	0	0	0	234413442212436	1	0	0	0
223413442212437	2	0	0	0	434433452212427	1	0	0	0
233413542212347	0	1	0	2	256432342122237	2	1	0	0
244333244232332	0	0	0	1	256433342123236	2	0	0	0
244333244232322	1	0	0	0	247432342122136	1	0	0	0
245333244242332	1	0	0	0	268432252122227	0	1	0	0
254333244232232	0	0	0	1	268632252122227	1	0	0	0
254333244232332	1	0	0	0	221313352122338	0	0	0	1
2543333244242332	1	0	0	0	263532233423148	1	0	0	0
253333244242232	1	0	0	0	360332233423138	1	0	0	0
252333243232232	0	0	0	1	263513233523344	1	0	0	0
252333243232332	0	0	0	1	253523233433527	1	0	0	0
251333243242332	1	0	0	0	253533232433527	1	0	0	1
252333243262222	0	0	0	1	253533232433427	1	0	0	0
244233234222322	1	0	0	0	253523133433527	1	0	0	0
233373242232325	1	0	0	0	253533133433527	1	0	0	0
252343242232524	1	0	0	0	353533233433427	0	0	0	1
25233234423251a	1	0	0	0	253533233433837	1	0	0	0
35234234423251a	1	0	0	0	253533233433237	1	0	0	0
233413442212338	0	1	0	0	254533233433537	0	0	0	1
233413442212335	1	0	0	0	253533233433536	1	0	0	0
233413442212337	1	0	0	0	252533233433428	1	0	0	0
21341344221233a	1	0	0	0	253534233433325	1	0	0	0
213413442212327	0	0	0	1	243533232433737	1	0	0	0
233413442212437	3	0	0	0	242433433433436	1	0	0	0

Table 2.1: Molecular data set compiled from Monteserin et al. (2013). All isolates were classified according to their genotype and resistance profile. The symbol "a" represents 10 repeat units and "-" represents missing data. The entries in the four columns sum to the total number of isolates, 100.

2.3 Model

In this Section we introduce a model that incorporates both VNTR-based genotyping and drug resistance states. The dynamic variables of the model correspond to numbers of cases of untreated and treated tuberculosis, their resistance states and VNTR genotypes associated with these infections in the population. We will now briefly describe processes involved in the model, and provide further details in the following Subsections.

An untreated case of TB can become detected and treated, and treatment involves a combination of drugs including the two in question. Drug sensitive strains can acquire resistance under treatment with some probability and thereby change their resistance

state. Treated and untreated cases can infect susceptible individuals and convert them to untreated cases. We disregard latent infections for simplicity (although latency is an important feature of the natural history of tuberculosis), and focus on active infections which are the larger source of new infections. Treated and untreated individuals can also recover or die. Treated individuals enjoy an additional probability of recovery that depends on the efficacy of the drugs, which in turn depends on the sensitivity or resistance of the infecting strain. Treated and untreated cases are also associated with a VNTR genotype, and this genotype evolves over time according to a stepwise mutation process for each locus. Figure 2.2 shows the broad structure of the model with respect to treatment and resistance states, while suppressing details of transmission, recovery, death and mutation of the VNTR loci.

At the end of the period of evolution, a simple random sample of 100 isolates is taken without replacement from the population, which matches the sample size of the Bolivian dataset. This provides a full description of the generative process for the observable data.

Let G be the number of distinct genotypes in the population (the number of distinct genotypes in the *sample* is g) and L be the number of VNTR loci used in the genotyping scheme. For the Bolivian dataset $L = 15$. In the model, the variable G is unknown and varies dynamically. We maintain three matrices which change through time: a $G \times L$ matrix, $\mathbf{V}$, which describes the VNTR genotypes; a $G \times 4$ matrix, $\mathbf{U}$, which describes the numbers of *untreated* cases of tuberculosis classified according to VNTR genotype and resistance state; and a $G \times 4$ matrix $\mathbf{T}$ which describes the numbers of *treated* cases of tuberculosis, again classified according to VNTR genotype and resistance state. It will be useful to define a $G \times 4$ matrix, $\mathbf{W}$, whose entries are the total numbers of both treated and untreated cases: $\mathbf{W} = \mathbf{U} + \mathbf{T}$.

As it will be helpful to be able to pick out columns of these matrices, we adopt notation for the standard basis vectors of $\mathbb{R}^n$. Let e_i denote the i-th basis (column) vector, so that $e_i = (0, \cdots, 0, 1, 0 \cdots, 0)^\top$, with the 1 in the i-th position. This allows us, for instance, to write the columns of the matrix $\mathbf{T}$ corresponding to each resistance state as $\mathbf{T}_0 = \mathbf{T}\, e_1$, $\mathbf{T}_{\text{INH}} = \mathbf{T}\, e_2$, $\mathbf{T}_{\text{RIF}} = \mathbf{T}\, e_3$ and $\mathbf{T}_{\text{MDR}} = \mathbf{T}\, e_4$, with similar notation for other matrices (note that the dimension of the e_i is left open but inferred from the matrix multiplication; in this case they are in $\mathbb{R}^4$).

Further, writing $\mathbf{1}_i$ for the column vector in $\mathbb{R}^i$ whose entries are all 1, then the product

$\mathbf{T}\,\mathbf{1}_4$ is a $G \times 1$ column vector whose entries are the numbers of individual cases of each VNTR genotype in the treated population, and the product $\mathbf{1}_G^\top \mathbf{T}\,\mathbf{1}_4$ is the sum of all the entries in $\mathbf{T}$ (the size of the treated population). Thus, we can write the size of the susceptible population, S, as

$$S = N - \mathbf{1}_G^\top \mathbf{W}\,\mathbf{1}_4,$$

where $\mathbf{1}_G^\top \mathbf{W}\,\mathbf{1}_4$ is the size of the infected population, and where N is the total population size which remains constant. We treat N as modelling the set of all individuals who come in contact with infectious cases and so we exclude individuals who either do not encounter infectious cases or are otherwise protected from infection. This variable therefore may be smaller than the actual population size.

The components of each vector $\mathbf{T}_k$ for $k = 0$, INH, RIF or MDR, are integers, representing the number of individual cases for each genotype. In the schematic diagram of the model in Figure 2.2, we use $T_k = \mathbf{1}_G^\top \mathbf{T}_k$ to represent the total population number of treated individuals with resistance state k, with similar notation $U_k = \mathbf{1}_G^\top \mathbf{U}_k$ to represent the untreated populations. The matrix notation is gathered and shown in Table 2.2.

The arrows between populations in Figure 2.2 represent the directional rates of detection and treatment τ and acquisition of resistance to each drug or set of drugs, so that ρ_{INH} and ρ_{RIF} represent rates of acquisition of resistance to isoniazid and rifampicin respectively and ρ_{MDR} the rate of double acquisition.

Symbol	Meaning
$\mathbf{V}$	$G \times L$ matrix describing the VNTR genotypes.
$\mathbf{U}$, $\mathbf{T}$, $\mathbf{W}$	$G \times 4$ matrices of untreated, treated and total cases respectively, with columns corresponding to resistance profiles.
$\mathbf{U}_k$, $\mathbf{T}_k$, $\mathbf{W}_k$	$G \times 1$ column vector for resistance profile k of untreated, treated and total cases respectively.
U_k, T_k, W_k	Total population sizes of untreated, treated and total with resistance profile k.
$\mathbf{U}_{i,k}$, $\mathbf{T}_{i,k}$, $\mathbf{W}_{i,k}$	(i,k) entries of the matrices $\mathbf{U}$, $\mathbf{T}$, $\mathbf{W}$: the number of cases in each category with genotype i and resistance profile k.
$\mathbf{1}_i$	$i \times 1$ column vector whose entries are all 1.
e_i	Column vector whose entries are 0 except for 1 in position i. Dimension determined by context.

Table 2.2: Summary of linear algebra notation.

In this model time is discrete, and during each time step the following events takes place in sequence.

1. Disease transmission giving rise to new cases;
2. Natural recovery, cure or death of cases;
3. Detection of cases which are then treated with drugs;
4. Conversion among resistant profiles in treated cases due to acquisition of resistance;
5. Mutation of the genetic marker (multiple VNTR loci).

The remainder of this Section provides details of how each of these events are modelled. Readers wishing to focus on the statistical aspects of the ABC inference can skip these subsections and go directly to Section 2.4.

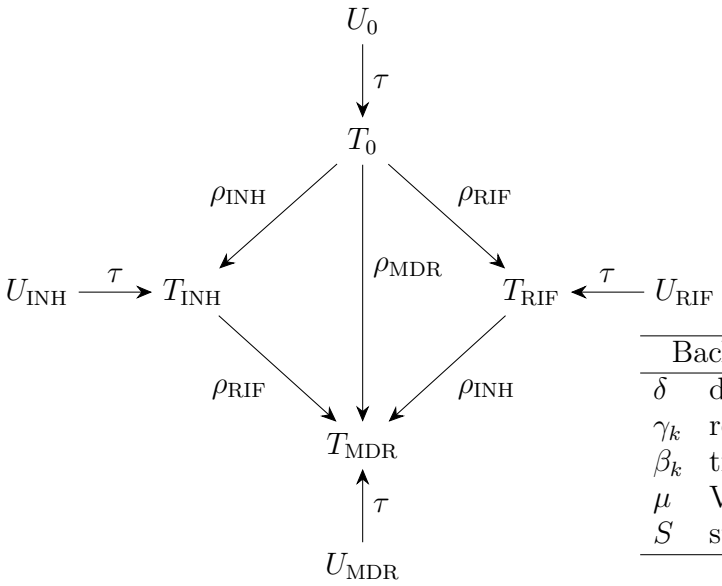
We regard the above process as a discrete-time stochastic model rather than a discrete-time approximation of a continuous-time stochastic process with rates approximating probabilities, although the latter interpretation becomes more appropriate as the time step length decreases. Here, rates of events will be treated as probabilities, which again is appropriate when time steps are short. The rate parameters are measured in years but we make time steps 1/12 of a year.

A summary of all model parameters, both fixed and to be estimated, and their meanings, are provided in Table 2.3.

2.3.1 New infections

In our model, new infections occur by mass action. The per capita rate at which a susceptible individual becomes infected by a case with resistance profile k is given by β_k/N times the number of infected cases in state k. The transmission parameters β_k are scaled by $1/N$ for convenience since realistic values of β_k/N are typically very small, and this ensures that the β_k are on the natural "per person per unit time" scale.

The acquisition of resistance to antibiotics often comes at a cost to the fitness of the bacterium, and we implement this fitness cost by assuming the transmission rate is lower for cases that carry resistance. Specifically, we assume a cost of "c per drug", so that if



Background parameters	
δ	death rate
γ_k	recovery rate
β_k	transmission rate
μ	VNTR mutation rate
S	susceptible population

Figure 2.2: Model structure for numbers of untreated (U_k) and treated (T_k) cases and per capita rates of conversion (within-host substitution) among resistance classes. Rates are ρ_{INH} and ρ_{RIF} for acquisition of resistance to isoniazid and rifampicin respectively, and ρ_{MDR} for single step acquisition of resistance to both drugs. Detection (and treatment) of cases is shown with arrows labelled with τ. Background parameters are shown in the table to the right, with rates per capita per unit time, and resistance states $k = 0$, INH, RIF, MDR. The mutation process of the VNTR locus is described in Section 2.3.5.

β_0 is the transmission rate of sensitive cases, then cases resistant to one drug transmit at rate $\beta_{\text{INH}} = \beta_{\text{RIF}} = (1-c)\beta_0$ and cases resistant to two drugs at $\beta_{\text{MDR}} = (1-c)^2\beta_0$. Here, $c = 0.1$ is considered known and fixed based on previous analyses of molecular epidemiological data (Luciani et al., 2009).

We now construct an expression for the average transmission probability across the infected population. The matrix $\mathbf{W}$ records all infected cases with different resistance states in each column, and the individuals corresponding to these columns have different transmission rates $\beta = (\beta_0, \beta_{\text{INH}}, \beta_{\text{RIF}}, \beta_{\text{MDR}})^\top$. If we write D_β for the diagonal matrix whose entries are from β, then the matrix $\mathbf{W}\,D_\beta$ is the infected population matrix $\mathbf{W}$ whose columns have been scaled by the entries of β (the relevant transmission rates). The expression

$$p = \frac{1}{N}\mathbf{1}_G^{\mathrm{T}}\,\mathbf{W}\,D_\beta\,\mathbf{1}_4$$

then gives the average transmission rate per susceptible individual. Since the population size N is usually large and the time steps are short, the value for p will nearly always be small. Accordingly, and to ensure that it does not exceed 1, we model the probability of transmission per susceptible individual as $\tilde{p} = \min\{1, p\}$.

Symbol	Meaning	Fixed value
δ	rate of death and natural recovery	0.52
γ_0	cure rate for resistance profile 0, when treated	0.5
γ_{INH}	cure rate for resistance profile INH, when treated	0.25
γ_{RIF}	cure rate for resistance profile RIF, when treated	0.25
γ_{MDR}	cure rate for resistance profile MDR, when treated	0.05
N	total susceptible population size in absence of disease	10^4
τ	treatment and detection rate	0.5
c	cost of resistance	0.1
Symbol	**Meaning**	**Prior**
β_0	transmission rate for resistance profile 0	Gamma*
μ	mutation rate of VNTR per locus per unit time	$U(0,1)$
ρ_{INH}	rate of acquisition of resistance to INH	$U(0,1)$
ρ_{RIF}	rate of acquisition of resistance to RIF	$U(0,1)$
ρ_{MDR}	rate of acquisition of resistance to INH and RIF	$U(0,1)$

Table 2.3: Summary of model parameters. The top set of parameters are given fixed values, whereas the bottom set of parameters are allocated prior distributions and estimated using ABC. Fixed values and priors are justified in Section 2.4.2. Rates are in units of per capita per year, but the time unit is set to 1/12 year in simulations. *Specifically, β_0 is assumed to follow a (shifted) Gamma prior defined as $\beta_0 - 0.68 \sim \text{Gamma}(\text{shape} = 2, \text{rate} = 0.73)$. See Section 2.4.2 for further details.

At each time step the number B of new infections is a random variable distributed as

$$B \sim \text{Binomial}(S, \tilde{p}).$$

These B new infections are then allocated across VNTR genotypes and resistance profiles according to the proportions represented by the matrix $\mathbf{W}\, D_\beta$. That is, a multinomial random sample distributes B according to the existing infected population and their relative transmission rates, so that the resulting allocation is a $G \times 4$ matrix Δ_β. Finally, as new infections are all assumed to be initially undetected, they are allocated to the untreated subpopulation, so that the matrix $\mathbf{U}$ is updated to $\mathbf{U} \to \mathbf{U} + \Delta_\beta$.

2.3.2 Cure, recovery and death

Infected individuals who are untreated (the population represented by the counts $\mathbf{U}$) recover or die at rate $\delta = \delta_r + \delta_d$ per case per time unit, where $\delta_r > 0$ is the rate of recovery and $\delta_d > 0$ is the rate of death due to any cause. The rate of cure due to successful treatment may vary according to resistance profile, so this rate is given by γ_k

for $k = 0$, INH, RIF, MDR. The number of cures, recoveries and deaths in a time step is given by

$$R \sim \text{Binomial}(U, \delta)$$

for the untreated population, where $U = \mathbf{1}_G^\top \mathbf{U}\, \mathbf{1}_4$ is the total number of all untreated cases, and

$$C_k \sim \text{Binomial}(T_k, \delta + \gamma_k),$$

for the treated population, where T_k is the number of treated individuals with resistance profile k (as defined at the start of this Section). The R *untreated* recovered individuals are distributed across both VNTR genotypes and resistance profiles with a multinomial distribution according to the counts given in $\mathbf{U}$. These are recorded in the $G \times 4$ update matrix Δ_δ (so that the sum of the entries in Δ_δ is $R = \mathbf{1}_G^\top \Delta_\delta\, \mathbf{1}_4$). Similarly, the C_k *treated* recovered individuals of resistance profile k are distributed across the VNTR genotypes according to the distribution observed in $\mathbf{T}_k$. These recovered counts for *all* resistance profiles are recorded in the G update matrix $\Delta_{\delta+\gamma}$, which is constructed from the column vectors of recovered treated counts for profile k in the order $k = 0$, INH, RIF and MDR. The matrices $\mathbf{U}$ and $\mathbf{T}$ are then updated to $\mathbf{U} \to \mathbf{U} - \Delta_\delta$ and $\mathbf{T} \to \mathbf{T} - \Delta_{\delta+\gamma}$ respectively. If the last instance of any genotype is removed by cure, recovery or death, the matrices $\mathbf{U}, \mathbf{T}, \mathbf{V}$ are adjusted by removing the rows corresponding to those genotypes, and the number of genotypes is updated with $G \to G - 1$.

Similarly to the case of new infections (Section 2.3.1), we assume that the recovery rate due to treatment depends only on the number of drugs the infecting strain is resistant to. Specifically, this implies that $\gamma_{\text{INH}} = \gamma_{\text{RIF}}$.

2.3.3 Detection and treatment

In this model, the detection of cases and the commencement of treatment are combined as a single process. Detected cases are transferred from the untreated class to the treated class. We denote this combined detection and treatment rate, per case, per unit time, as $\tau > 0$. With this rate we draw D individuals to transfer between untreated and treated populations, where

$$D \sim \text{Binomial}(U, \tau).$$

These D individuals are then allocated across VNTR genotypes and resistance profiles according to the observed distribution of untreated cases, $\mathbf{U}$. As before, this results in a $G \times 4$ update matrix Δ_τ, which we use to update $\mathbf{U} \to \mathbf{U} - \Delta_\tau$ and $\mathbf{T} \to \mathbf{T} + \Delta_\tau$.

2.3.4 Acquisition of drug resistance

Individual treated cases are able to convert from one resistance profile to another through adaptive evolution. That is, under drug treatment, natural selection acts to favour increasing levels of resistance. As a result of this process, individuals may move from the $k = 0$ resistance profile (sensitive to both drugs) to one of the other three resistance profiles: INH, RIF, or MDR (resistance to one or both drugs). Individuals may also move from resistance to exactly one of the drugs (INH or RIF) to the multiple drug resistance profile MDR. We respectively denote the rate of acquisition of resistance to INH or RIF by ρ_{INH} and ρ_{RIF}, and denote the rate of acquisition of resistance from individuals in the sensitive population to both drugs simultaneously by ρ_{MDR}. These conversions and rates are illustrated schematically in Figure 2.2.

To model resistance acquisition, we select individuals to move between resistance profiles in the treated population i.e. between columns in the matrix $\mathbf{T}$. Acquiring resistance to the drug rifampicin will result in individuals moving from the column $\mathbf{T}_0$ to $\mathbf{T}_{\text{RIF}}$, and from $\mathbf{T}_{\text{INH}}$ to $\mathbf{T}_{\text{MDR}}$ at a rate ρ_{RIF}. Similarly, acquiring resistance to the drug isoniazid results in individuals moving from the column $\mathbf{T}_0$ to $\mathbf{T}_{\text{INH}}$, and from $\mathbf{T}_{\text{RIF}}$ to $\mathbf{T}_{\text{MDR}}$ at a rate ρ_{RIF}. Simultaneous acquisition of resistance to both drugs moves individuals from the column $\mathbf{T}_0$ to $\mathbf{T}_{\text{MDR}}$ at the rate ρ_{MDR}. These movements occur between columns but not across rows (infections do not change VNTR genotypes through this process).

Mechanistically, we can obtain the number of cases of genotype i transitioning from resistance profile k to resistance profile k', denoted $A_{i,k\to k'}$, as

$$A_{i,0\to *} \sim \text{Multinomial}(\mathbf{T}_{i,0}, \rho_{0\to *})$$

$$A_{i,\text{INH}\to\text{MDR}} \sim \text{Binomial}(\mathbf{T}_{i,\text{INH}}, \rho_{\text{RIF}})$$

$$A_{i,\text{RIF}\to\text{MDR}} \sim \text{Binomial}(\mathbf{T}_{i,\text{RIF}}, \rho_{\text{INH}})$$

where $A_{i,0\to *} = (A_{i,0\to\text{INH}}, A_{i,0\to\text{RIF}}, A_{i,0\to\text{MDR}}, A_{i,0\to 0})^\top$ is the vector of cases transitioning from sensitivity, $\mathbf{T}_{i,k}$ is the entry of the matrix $\mathbf{T}$ corresponding to the genotype i

and resistance profile k (see Table 2.2), and $\rho_{0\to *} = (\rho_{\text{INH}}, \rho_{\text{RIF}}, \rho_{\text{MDR}}, 1 - \sum_k \rho_k)^\top$ is the vector of probabilities of these events.

If we denote $\Delta_{k\to k'}$ as column vectors of counts of movements from resistance profile k to k' across all G genotypes, we can then construct the overall $G \times 4$ update matrix

$$\Delta_\rho = (\Delta_0 \mid \Delta_{\text{INH}} \mid \Delta_{\text{RIF}} \mid \Delta_{\text{MDR}})\,.$$

from the column vectors Δ_k, which denote the total population change for resistance profile k, where

$$\begin{aligned}
\Delta_0 &= -(\Delta_{0\to\text{INH}} + \Delta_{0\to\text{RIF}} + \Delta_{0\to\text{MDR}}) \\
\Delta_{\text{INH}} &= \Delta_{0\to\text{INH}} - \Delta_{\text{RIF}\to\text{MDR}} \\
\Delta_{\text{RIF}} &= \Delta_{0\to\text{RIF}}\Delta_{\text{INH}\to\text{MDR}} \\
\Delta_{\text{MDR}} &= \Delta_{0\to\text{MDR}} + \Delta_{\text{RIF}\to\text{MDR}} + \Delta_{\text{INH}\to\text{MDR}}.
\end{aligned}$$

The population of treated cases is then updated to $\mathbf{T} \to \mathbf{T} + \Delta_\rho$.

2.3.5 Mutation of the marker

The set of $L = 15$ VNTR loci constitute the genetic marker used to genotype bacterial isolates (see Section 2.2). Each genotype is a list of numbers of tandem repeat units at the L loci. The states of all VNTRs in the infected population are given by the $G \times L$ matrix $\mathbf{V}$ with elements V_{ij} describing the repeat number of locus j in genotype i. Each locus mutates through a stepwise mutation process at rate μ per locus per case per unit time. When mutation occurs, the repeat number V_{ij} at a locus j of genotype i changes by $+1$ or -1, each with probability 0.5. A repeat number of 1 is treated as an absorbing boundary (i.e. there is zero probability of the repeat number increasing from 1 to 2) because at state 1 there is no longer a genetic sequence that is tandemly repeated and no mechanism such as replication slippage acts to expand it from 1 to 2.

Mutation of the marker has the effect of moving cases between the rows of the matrix $\mathbf{W}$. We first identify the number of mutation events in the population, M, where $M \sim \text{Binomial}(S, \mu)$ and $S = N - \mathbf{1}_G^{\mathrm{T}}\,\mathbf{W}\,\mathbf{1}_4$ is the size of the susceptible population (see Section 2.3). The M cases are then distributed across the population of VNTR genotypes

and resistance profiles, according to the entries of the matrices $\mathbf{T}$ and $\mathbf{U}$. Each individual case undergoing mutation corresponds to a specific entry in either $\mathbf{T}$ or $\mathbf{U}$. This entry is described by its VNTR genotype $\mathbf{V}_i = (V_{i,1}, \ldots, V_{i,L})$ where $L = 15$ for the Bolivian data, and its resistance profile, $k = 0$, INH, RIF, MDR. The result of the mutation is a change to the VNTR genotype, which is represented by a change in the repeat number at a single locus, V_{ij}, by ± 1. This may or may not result in a VNTR genotype that is already present in the population.

If the new VNTR genotype already appears as a row in the matrix $\mathbf{V}$ as an existing type in the data, then there is no change to $\mathbf{V}$. The matrix $\mathbf{T}$ or $\mathbf{U}$ on the other hand is changed by subtracting 1 from one entry and adding one to another entry in the same column (the resistance profile, k, does not change). In matrix terms, supposing the change is to a treated case, this can be described by updating $\mathbf{T} \to \mathbf{T} - e_{i,j} + e_{i,k}$, where $e_{i,j}$ is the matrix whose entries are zero except for a 1 in the (i, j)-th position, and where the VNTR genotype changes from row j to row k.

If the new VNTR genotype does not already appear in the population, then the matrix $\mathbf{V}$ is expanded to include a new row describing the new genotype, so that $\mathbf{V}$ becomes a $(G + 1) \times 15$ matrix. The update for $\mathbf{T}$ or $\mathbf{U}$ is the same as described above, except that now both matrices are $(G + 1) \times 4$ dimensional. Subsequent to this update we increment $G \to G + 1$. If mutation of a VNTR genotype removes the last instance of the original genotype from $\mathbf{U}$ and $\mathbf{T}$ the corresponding rows of matrices $\mathbf{V}$, $\mathbf{U}$ and $\mathbf{T}$ are deleted, requiring the update $G \to G - 1$.

2.3.6 Initial conditions of the model

The model covers the period from when drugs are introduced at time $t = 0$ to when sampling occurs. Since the main first-line anti-tuberculosis drugs were discovered/developed in the 1940s to early 1960s, we assumed treatment commenced around 1960 and ran the simulation for a period of 50 years. We assumed that both drugs, isoniazid and rifampicin, were introduced at the same time and are administered together in combination therapy. The standard course of treatment includes both drugs along with other first-line drugs (WHO, 2015).

We assume that at the start of the process all cases are sensitive to both drugs and that the number of cases is at equilibrium in the absence of treatment and resistance.

To compute this equilibrium state, we consider the differential equation describing the deterministic version of the model ignoring VNTR genotypes. Namely,

$$\frac{dU}{dt} = (\beta/N)SU - \delta U$$

where $S = N - U$ and t indicates time. Setting dU/dt to zero and solving for the dynamic variables we obtain equilibrium values of

$$\hat{U} = N\left(1 - \frac{\delta}{\beta_0}\right) \qquad \text{and} \qquad \hat{S} = \frac{\delta N}{\beta}$$

for $U > 0$.

The basic reproduction number of a pathogen R_0 is defined to be the average number of new infectious cases caused by a single infection in a completely susceptible population. In our model, before there is any treatment, assuming all cases are doubly susceptible, a single case on average persists for $1/\delta$ years and generates $S\beta_0/N$ new cases per unit time but since $S = N$ in a wholly susceptible population then $R_0 = \beta_0/\delta$.

All cases are initially untreated and sensitive. From time $t = 0$ treatment in the population commences. To reintroduce into the model genetic variation at the marker loci, the initial distribution of genotype clusters is a random sample drawn from the infinite alleles model from population genetic theory (Ewens, 1972; Hubbell, 2001; Luciani et al., 2008). The infinite alleles model depends on a single parameter, the diversity parameter, which we set to $2\hat{U}\mu L$ where $\hat{U}$ is the number of cases, taken from the equilibrium value described above, μ is the mutation rate per VNTR locus and L is the number of VNTR loci used in genotyping isolates. To initialise the multi-locus VNTR genotypes, each genotype is a sequence of random integers, of length L, with each VNTR number V_{ij} drawn from a discrete uniform distribution over $\{1, \ldots, 10\}$. Although the initial distribution of genotype clusters is set under the infinite alleles model, the mutation process for VNTRs brings the distribution in line with the stepwise model over time.

The initial conditions are a function of the parameters which are set according to the priors specified in Section 2.4.2.

2.4 Inference with approximate Bayesian computation

For the model in Section 2.3, when the data are only observed at a single point in time, the cost of evaluating the likelihood function is computationally prohibitive. This results from the "incomplete" nature of the observed data (see Section 2.2) in the sense that we only have access to a snapshot of the population, via the observed sample, at the time the study was conducted, with no direct measurements of the system as it progressed. Computing the likelihood then requires integrating over all potential trajectories the population could have gone through before reaching its final, observed state.

As such we adopt approximate Bayesian computation (ABC) methods as a means of performing Bayesian statistical inference for the unknown model parameters $\theta = (\beta_0, \mu, \rho_{\text{INH}}, \rho_{\text{RIF}}, \rho_{\text{MDR}})^\top$. As observed in other chapters in this thesis, the ABC approximation to the true posterior distribution is given by

$$\pi_{ABC}(\theta|s_{obs}) \propto \pi(\theta) \int K_h(\|s - s_{obs}\|) p(s|\theta) ds,$$

where $\pi(\theta)$ is the prior distribution, $s = S(\mathbf{X})$ is a vector of summary statistics with $s_{obs} = S(\mathbf{X}_{obs})$, $p(s|\theta)$ is the computationally intractable likelihood function for the summary statistics s, and $K_h(u) = K(u/h)/h$ is a standard smoothing kernel with scale parameter $h > 0$. In the following analyses we used the uniform kernel on $[-h, h]$ for $K_h(u)$. The quality of the ABC approximation depends on the information loss in the summary statistics s over the full dataset $\mathbf{X}$, and the size of the kernel scale parameter h with smaller h producing greater accuracy and increased computational cost. Choice of both s and h are typically driven by the amount of expert knowledge and computation available for the analysis.

For the present analysis we implement a version of a simple ABC importance sampling algorithm, as outlined in the box. Given a suitable importance sampling distribution $q(\theta)$, the algorithm produces a set of weighted samples from the ABC approximation to the true posterior $(\theta^{(1)}, w^{(1)}), \ldots, (\theta^{(\tilde{N})}, w^{(\tilde{N})}) \sim \pi_{ABC}(\theta|s_{obs})$. As with standard importance sampling, suitable choice of $q(\theta)$ is important to avoid high variance in the importance weights, and also to avoid needlessly generating datasets $s = S(\mathbf{X}^{(i)})$, $\mathbf{X}^{(i)} \sim p(\mathbf{X}|\theta)$ for which $s^{(i)}$ and s_{obs} will never be close.

Algorithm 2 ABC Importance Sampling

- A target posterior density $\pi(\theta|\mathbf{X}_{obs}) \propto p(\mathbf{X}_{obs}|\theta)\pi(\theta)$, consisting of a prior distribution $\pi(\theta)$ and a procedure for generating data under the model $p(\mathbf{X}_{obs}|\theta)$.
- A proposal density $q(\theta)$, with $q(\theta) > 0$ if $\pi(\theta|\mathbf{X}_{obs}) > 0$.
- An integer $\tilde{N} > 0$.
- An observed vector of summary statistics $s_{obs} = S(\mathbf{X}_{obs})$.
- A kernel function $K_h(u)$ and scale parameter $h > 0$.

Sampling:
For $i = 1, \ldots, \tilde{N}$:

1. Generate $\theta^{(i)} \sim q(\theta)$ from sampling density q.
2. Generate $\mathbf{X}^{(i)} \sim p(\mathbf{X}|\theta^{(i)})$ from the likelihood.
3. Compute the summary statistics $s^{(i)} = S(\mathbf{X}^{(i)})$.
4. Assign $\theta^{(i)}$ the weight $w^{(i)} \propto K_h(\|s^{(i)} - s_{obs}\|)\pi(\theta^{(i)})/q(\theta^{(i)})$.

Output:
A set of weighted parameter vectors $\{(\theta^{(i)}, w^{(i)})\}_{i=1}^{\tilde{N}} \sim \pi_{ABC}(\theta|s_{obs})$.

To determine a suitable importance sampling distribution $q(\theta)$ we adopt a two stage procedure, following the approach of Fearnhead and Prangle (2012). In the first stage we perform a pilot ABC analysis using a sampling distribution that is diffuse enough to easily encompass the ABC posterior approximation obtained for a moderate value of the kernel scale parameter h. We specified $q(\theta) \propto \pi(\theta)I(\theta \in A)$ which is proportional to the prior, but restricted to the hyper-rectangle A. Here, A is constructed as the smallest credible hyper-rectangle that we believe contains the ABC posterior approximation. As such, this $q(\theta)$ will identify the general region in which $\pi_{ABC}(\theta|s_{obs})$ is located. Specifically, for $\theta = (\beta_0, \mu, \rho_{\text{INH}}, \rho_{\text{RIF}}, \rho_{\text{MDR}})^\top$ we adopt $q(\theta) = \tilde{\pi}_{15}(\beta_0) \times U(0, .005) \times U(0, .01) \times U(0, .005) \times U(0, .001)$, where $\tilde{\pi}_{15}(\beta_0)$ is the prior $\pi(\beta_0)$ for β_0 specified in Section 2.4.2, but truncated to exclude density above the point $\beta_0 = 15$.

For posterior distributions with strong dependence between parameters, defining $q(\theta)$ over such a hyper-rectangle may be inefficient as it will cover many regions of effectively zero posterior density. Accordingly we construct the sampling distribution for the second stage, with the lowest value of h, as a kernel density estimate of the previous ABC estimate of the posterior distribution: $q(\theta) = \sum_i w^{(i)} L(\theta|\theta^{(i)})$, where L is a suitable kernel density (not to be confused with the kernel K_h). This approach follows the ideas behind the

sequential Monte Carlo-based ABC samplers of Sisson et al. (2007) and others. At each stage the kernel scale parameter h is decreased, and determined as the value which results in $\sim$ 2,000 posterior samples with non-zero weights, for the given computational budget.

To ensure greater efficiency at each stage we also performed a non-linear regression adjustment using a neural network with a single hidden layer (see Blum and François, 2010; Csilléry et al., 2012; Beaumont et al., 2002), as implemented in the *R* package `abc`. The adjustment used logistic transformations for the response.

For samples drawn from the final importance sampling distribution $q(\theta)$, the data generation procedure took on average $\sim$ 40 seconds in *R*. This is computationally expensive from an ABC context, and could be reduced by recoding the simulator in a compiled language such as *C*, or by adapting the "lazy ABC" ideas of Prangle (2016) to terminate early those simulations that are likely to be rejected. In this implementation we performed importance sampling from each distribution $q(\theta)$ in parallel on multiple nodes of a computational cluster.

2.4.1 Summary statistics

Considering the matrix structure of the observed data $\mathbf{X}_{obs}$ (see Section 2.2), we determine the information content in $\mathbf{X}$ as if it was the design matrix of a regression model and summarise it accordingly. Specifically, we define the summary statistics $s = S(\mathbf{X})$ to be the upper-triangular elements of the matrix

$$(\mathbf{1}_g|\mathbf{X})^\top(\mathbf{1}_g|\mathbf{X}),$$

where the vertical lines denote the addition of an extra column of ones. The added columns of ones enriches the set of summary statistics by including the row and column totals of $\mathbf{X}$. Alternatively, these summary statistics can be described as:

i) g: the number of distinct genotypes in the sample.

ii) n_k: the number of isolates with resistance profile $k = 0$, INH, RIF and MDR.

iii) $c_{k,k'} = (\mathbf{X}_k)^\top \mathbf{X}_{k'}$: the dot product between the resistance profiles of k and k' within $\mathbf{X}$.

Note that these summary statistics are over specified in that $n_0 + n_{\text{INH}} + n_{\text{RIF}} + n_{\text{MDR}}$

equals the total number of isolates sampled from the population, which is known and equal to the number of isolates in the observed data sample (100 for the Bolivian data). Accordingly, and without loss of generality, we remove n_{MDR} as a summary statistic to avoid collinearity. In combination, this set of 14 summary statistics efficiently encapsulates the available information about the covariance structure of the original dataset $\mathbf{X}$, the distribution of the isolates among the different resistance profiles and the degree of diversity of isolates within the sample.

For the Bolivian dataset, there are $g = 68$ distinct genotypes, $n_0 = 78$ sensitive isolates, $n_{\text{INH}} = 8$ isolates resistant to isoniazid only, $n_{\text{RIF}} = 0$ isolates resistant to rifampicin only and $n_{\text{MDR}} = 16$ doubly resistant isolates (see Table 2.1). The remaining statistics, $c_{k,k'}$, are computed as:

	0	INH	RIF	MDR
0	232	15	0	1
INH	–	10	0	6
RIF	–	–	0	0
MDR	–	–	–	18

Finally, in order to reduce the impact of summary statistics operating on different scales, we compare simulated and observed summary statistics within the kernel $K_h(\|s - s_{obs}\|)$ via the $L_{\frac{1}{2}}$ norm

$$\|s - s_{obs}\| = \|S(\mathbf{X}) - S(\mathbf{X}_{obs})\| = \left(\sum_{j=1}^{\dim(s)} [S(\mathbf{X})_j - S(\mathbf{X}_{obs})_j]^{\frac{1}{2}} \right)^2,$$

where $\dim(s) = 14$ is the number of summary statistics. Alternative approaches could rescale the statistics via an appropriate covariance matrix (e.g. Luciani et al., 2009; Erhardt and Sisson, 2016) or use other norms, however the results in the following Section proved to be robust to more structured comparisons, so we did not pursue this further. In particular the following results were robust to these choices because of the use of a good (non-linear) regression adjustment, which greatly improves the ABC posterior approximation, and which has a larger impact on this approximation than the choice of metric $\|\cdot\|$.

2.4.2 Parameter specifications and prior distributions

Of the 13 model parameters (see Table 2.3), eight of these are known well enough for the purposes of our analysis to fix their values. Namely, the parameters $(\delta, \gamma_0, \gamma_{\text{INH}}, \gamma_{\text{RIF}}, \gamma_{\text{MDR}}, N, \tau, c)^\top$ are set to these fixed values. We justify our choices for these values below. The remaining five parameters $\theta = (\beta_0, \mu, \rho_{\text{INH}}, \rho_{\text{RIF}}, \rho_{\text{MDR}})^\top$ are to be estimated, and require a prior distribution specification.

The rate of death or recovery, δ, is fixed and set to be $\delta = 0.52$ per case per year following Dye and Espinal (2001) and Cohen and Murray (2004). Similarly, following Dye and Espinal (2001), untreated individuals are detected and treated at rate $\tau = 0.5$ per case per year. The rates of recovery due to treatment, γ_k, for resistance profiles $k = 0$, INH, RIF and MDR, can be written in terms of the probability of treatment success

$$p_k = \frac{\delta_r + \gamma_k}{\delta_d + \delta_r + \gamma_k}.$$

We set the cure rates to be $\gamma_0 = 0.5, \gamma_{\text{INH}} = \gamma_{\text{RIF}} = 0.25, \gamma_{\text{MDR}} = 0.05$, which, by using $\delta_r = 0.2$ (Dye and Espinal, 2001; Cohen and Murray, 2004), corresponds to treatment success probabilities of approximately $p_0 = 0.69$, $p_{\text{INH}} = p_{\text{RIF}} = 0.58$ and $p_{\text{MDR}} = 0.44$. These values are within the supported ranges in the literature, namely, $p_0 = 0.45 - 0.75$, $p_{\text{INH}} = p_{\text{RIF}} = 0.3 - 0.6$ and $p_{\text{MDR}} = 0.05 - 45$ (Blower and Chou, 2004). We chose higher values within these ranges since Blower and Chou (2004) explored a wide range of possibilities in models including epidemiologically pessimistic scenarios.

In a whole-genome sequencing study based on a dataset involving a Russian population, Casali et al. (2014) state that the fitness cost of drug resistance is negligible for *certain* bacterial strains (see also Pym et al., 2002). Nevertheless, the *overall* fitness cost, c, is expected to be positive and to vary geographically, being ultimately determined by which strains are predominantly circulating in the region. In this study, c was fixed and set to be $c = 0.1$, according to estimates by Luciani et al. (2009).

To set the total population size N we first observe that because the sample of 100 isolates represents $\sim 1.1\%$ of the population, this implies that the infected population is 9091. We expect that the number of susceptible individuals who are exposed to disease is somewhat higher than this. Accordingly, we assumed that the total size of the population susceptible to tuberculosis is $N = 10,000$. Larger total population sizes can be used, at

the price of greater computational overheads for generating data under the model.

Previous work estimated rates of resistance acquisition by mutation to be around $0.0025 - 0.02$ per case per year (Luciani et al., 2009). The rate of mutation of the VNTR loci in *M. tuberculosis* was estimated to be around 10^{-3} per locus per case per year (Reyes and Tanaka, 2010; Aandahl et al., 2012; Ragheb et al., 2013) but lower estimates have also been found (Wirth et al., 2008; Supply et al., 2011). All of these mutation rates are much lower than 1. We treat these mutation rate parameters as probabilities and conservatively set the standard uniform distribution as a wide prior on each parameter. That is, for the acquisition of resistance to isoniazid or rifampicin (or both), we specify priors for the rates of resistance acquisition as $\rho_{\text{INH}}, \rho_{\text{RIF}}, \rho_{\text{MDR}} \sim U(0,1)$. Similarly, for the mutation rate of the VNTR molecular marker, μ, we use the prior $\mu \sim U(0,1)$.

The transmission parameter for doubly sensitive strains β_0 is given the shifted gamma prior

$$\beta_0 - 0.68 \sim \text{Gamma}(\text{shape} = 2, \text{rate} = 0.73)$$

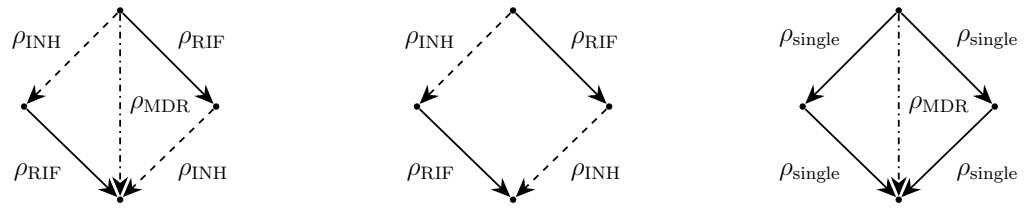
where the parameters are chosen such that the resulting prior distribution of the basic reproduction number R_0 closely resembles the distribution obtained in a numerical analysis of tuberculosis dynamics by Blower et al. (1995). Note that the prior on β_0 is shifted in order ensure the realistic condition that $R_0 > 1$. A value of R_0 lower than unity would lead to extinction of *M. tuberculosis*.

We reiterate that we interpret the rate parameters as probabilities per time step and handle the parameters so that their values remain in (0,1). This approximation increases in accuracy as the time unit decreases. Here we divide the natural time unit of one year into new units of 1/12 year per time step.

2.5 Competing models of resistance acquisition

We estimate the rates of acquisition of drug resistance to rifampicin and isoniazid by fitting the model described in Section 2.3 to the Bolivian data (Monteserin et al., 2013) with the ABC method described in Section 2.4. Additionally, by constraining particular resistance-acquisition parameters ρ_k to produce meaningful submodels of the full model, we are able to examine two specific biological questions. The relationships between the two submodels and the full model are illustrated in Figure 2.3. First, we ask whether

it is possible for multidrug resistance to evolve directly from doubly sensitive bacteria or whether this direct conversion does not occur (i.e., $\rho_{\text{MDR}} = 0$: Submodel 1). Second, we ask whether differences between rates of mutation to rifampicin and isoniazid resistance are apparent at the epidemiological scale (i.e., $\rho_{\text{INH}} = \rho_{\text{RIF}} = \rho_{\text{single}}$: Submodel 2).



(a) Full model (b) Submodel 1 (c) Submodel 2

Figure 2.3: Three candidate models of acquisition of multiple drug resistance. (a) The full model: two different rates of conversion leading to acquisition of resistance and a rate of conversion from resistance profile 0 to resistance profile MDR. This model is also shown in Figure 2.2. (b) Submodel 1: no direct conversion from resistance profile 0 to resistance profile MDR ($\rho_{\text{MDR}} = 0$). (c) Submodel 2: same rate of conversion for the two drugs ($\rho_{\text{INH}} = \rho_{\text{RIF}} = \rho_{\text{single}}$).

Figure 2.4 illustrates the ABC marginal posterior density estimates of each parameter under the three different sets of model assumptions. Under the full model there is a clear visual difference between the rates of mutation of rifampicin and isoniazid resistance, with the latter occurring at a much higher rate. In contrast, the rate of simultaneous resistance acquisition appears to be higher than that for rifampicin alone. When eliminating the possibility of simultaneous acquisition of multiple drug resistance $\rho_{\text{MDR}} = 0$ (Submodel 1), ρ_{INH} and ρ_{RIF} both increase, relative to the full model, to compensate for the imposed restriction when fitting to the observed data (Figure 2.4b). Similarly, when we fix the identity $\rho_{\text{INH}} = \rho_{\text{RIF}} = \rho_{\text{single}}$ (Submodel 2) to impose a single rate of resistance acquisition, the posterior density of this parameter moves to intermediate values compared to the two distinct rates of acquisition estimated under the full model (Figure 2.4c). ABC marginal posterior means and highest posterior density (HPD) credible intervals for all models are reported in Table 2.4.

2.5.1 Can resistance to both drugs be acquired simultaneously?

To determine whether resistance to both drugs can evolve directly from a double sensitive strain within an infection, we compare Submodel 1 ($\rho_{\text{MDR}} = 0$) against the full model.

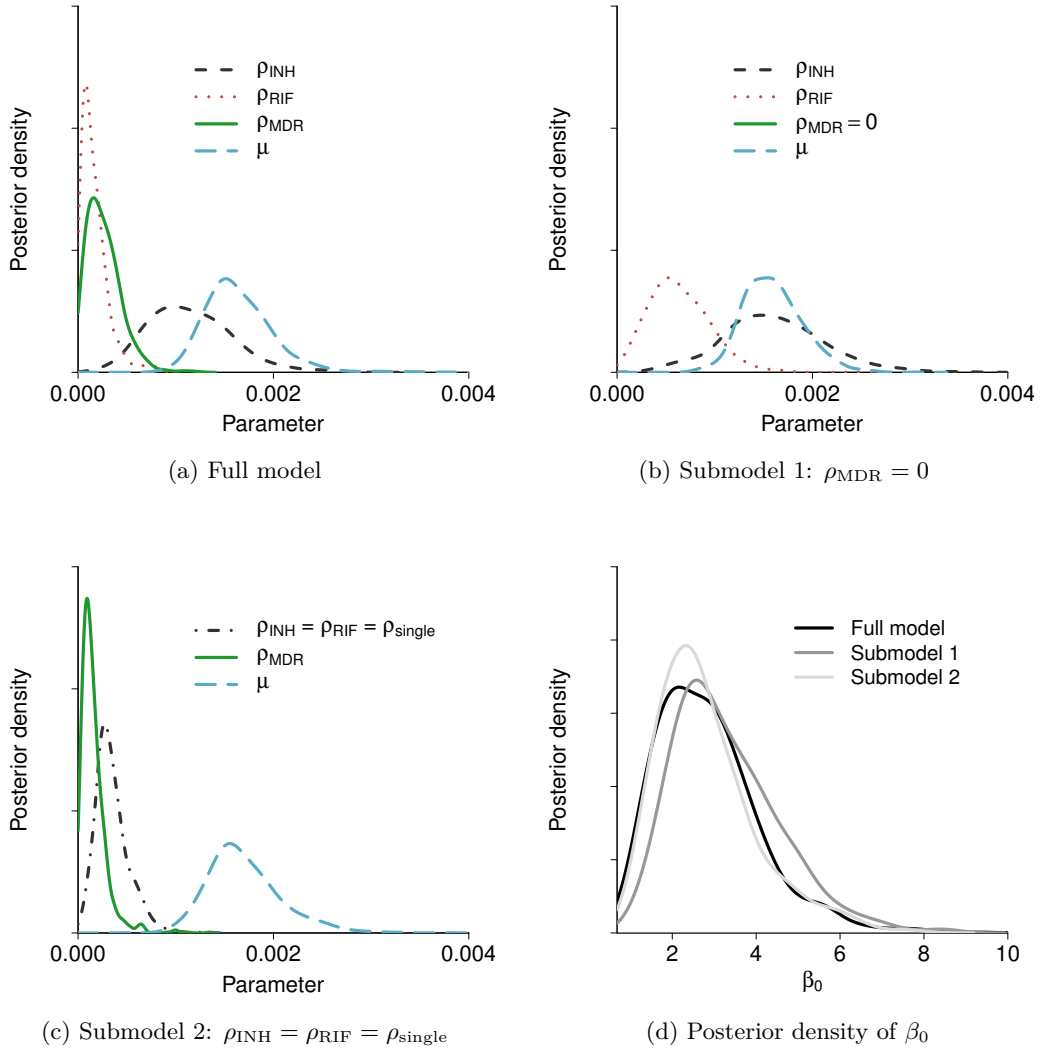
	ρ_{INH}	ρ_{RIF}	ρ_{MDR}	μ	β_0
Full model					
Posterior mean	1.14×10^{-3}	1.67×10^{-4}	2.62×10^{-4}	1.64×10^{-3}	2.85
CI lower limit	3.40×10^{-4}	3.82×10^{-6}	3.93×10^{-6}	1.11×10^{-3}	0.97
CI upper limit	1.94×10^{-3}	4.28×10^{-4}	5.81×10^{-4}	2.40×10^{-3}	5.33
Submodel 1					
Posterior mean	1.60×10^{-3}	6.37×10^{-4}	–	1.59×10^{-3}	3.29
CI lower limit	4.55×10^{-4}	1.27×10^{-4}	–	1.03×10^{-3}	1.20
CI upper limit	2.49×10^{-3}	1.24×10^{-3}	–	2.19×10^{-3}	5.78
Submodel 2					
Posterior mean	3.46×10^{-4}	–	1.56×10^{-4}	1.70×10^{-3}	2.81
CI lower limit	7.26×10^{-5}	–	6.62×10^{-7}	1.10×10^{-3}	0.86
CI upper limit	6.90×10^{-4}	–	3.76×10^{-4}	2.54×10^{-3}	5.20

Table 2.4: ABC posterior means with lower and upper limits of the 95% HPD (highest posterior density) credibile intervals for each parameter of each fitted model.

Formal standard Bayesian model comparison typically occurs through Bayes factors. In the ABC framework this task is complicated by the need to perform ABC with summary statistics that are informative for the model indicator parameter, in addition to those informative for the model specific parameters. Such summary statistics can not only be difficult to identify, but the resulting composite vector of summary statistics can be high dimensional, which may then produce more inaccurate inference than if each model was analysed independently. See e.g. Robert et al. (2011); Marin et al. (2014) and Marin et al. (2017, this volume) for a discussion of these issues. A useful alternative is to consider posterior predictive checks or related goodness-of-fit tests (e.g. Thornton and Andolfatto, 2006; Csillery et al., 2010; Aandahl et al., 2012; Prangle et al., 2014).

Figure 2.5 shows the posterior predictive distribution of the summary statistics ($n_0, n_{\text{INH}}+n_{\text{RIF}}, n_{\text{MDR}}$) described in Section 2.4.1, for the full model (panel (a)) and Submodel 1 (panel (b)), where a darker intensity indicates higher density. This predictive distribution graphically illustrates each model's ability to generate the observed summary statistics (78, 8, 16), indicated by the asterisks, which represent the number of individuals in the sample sensitive to both drugs (n_0), resistant to a single drug ($n_{\text{INH}} + n_{\text{RIF}}$) and resistant to both drugs (n_{MDR}).

The predictive distributions for each model are diffuse, particularly for the full model. This variability is expected given that the sample size is small (100 isolates) and that the evolution of drug resistance from sensitivity is a relatively rare stochastic event. In the



(a) Full model

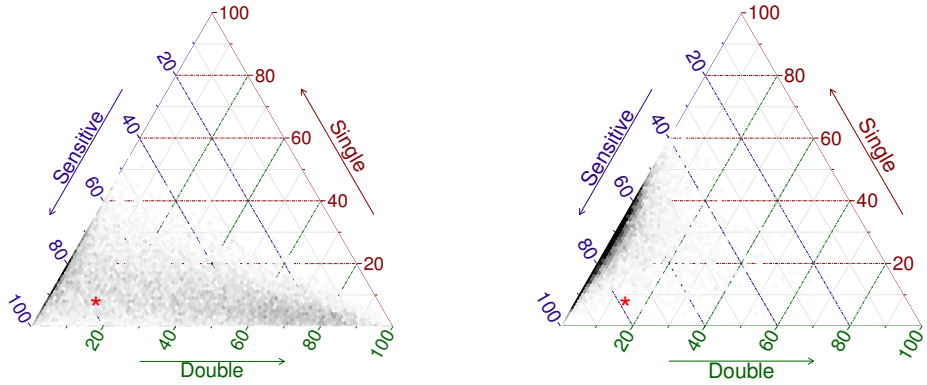
(b) Submodel 1: $\rho_{\text{MDR}} = 0$

(c) Submodel 2: $\rho_{\text{INH}} = \rho_{\text{RIF}} = \rho_{\text{single}}$

(d) Posterior density of β_0

Figure 2.4: Estimated ABC marginal posterior densities for each estimated parameter under (a) the full model, (b) Submodel 1 ($\rho_{\text{MDR}} = 0$), and (c) Submodel 2 ($\rho_{\text{INH}} = \rho_{\text{RIF}} = \rho_{\text{single}}$). Panel (d) shows the estimated ABC marginal posterior density of the transmission rate β_0 of the sensitive strain for each model structure.

case of Submodel 1 (Figure 2.5 panel (b)) where we impose the condition $\rho_{\text{MDR}} = 0$, the density of samples is shifted away from the bottom-right corner which represents double resistance. This pattern is due to the lack of the direct route to multidrug resistance. The observed data (asterisk) is in a region of low posterior predictive density under Submodel 1, and so we conclude that this model is not particularly supported by the data. In contrast, the observed data lie more clearly within a moderately high density region of the posterior predictive under the full model (Figure 2.5 panel (a)). This analysis therefore suggests that of the two competing hypotheses, it is more likely that resistance to both drugs can be acquired simultaneously ($\rho_{\text{MDR}} > 0$) than otherwise. Note, however, that



(a) Full model

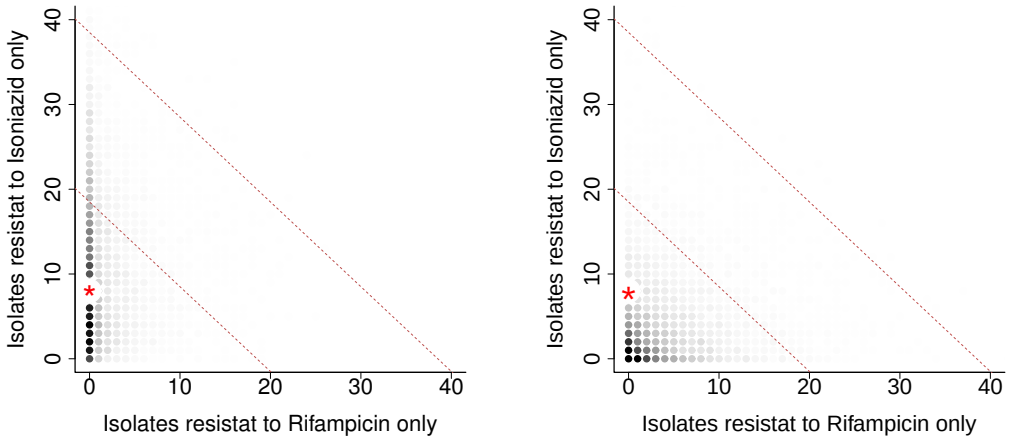
(b) Submodel 1: $\rho_{\text{MDR}} = 0$

Figure 2.5: Posterior predictive distribution of $(n_0, n_{\text{INH}} + n_{\text{RIF}}, n_{\text{MDR}})$ under the full model (panel (a)) and Submodel 1 (panel (b)). Darker intensity indicates higher posterior density. The asterisk (*) indicates the observed data (78, 8, 16).

this direct route is not the only possible path to double resistance, which can still occur in stages through single resistance.

2.5.2 Is resistance to both drugs acquired at equal rates?

In order to determine whether the rates of acquisition of resistance to the two drugs are equal ($\rho_{\text{INH}} = \rho_{\text{RIF}}$), we compare Submodel 2 against the full model. Figure 2.6 depicts the posterior predictive distribution of $(n_{\text{INH}}, n_{\text{RIF}})$ under each model – the number of cases resistant only to isoniazid (n_{INH}) and the number of cases resistant only to rifampicin (n_{RIF}) in the sample. The observed values of these summary statistics are $n_{\text{INH}} = 8$ for isoniazid and $n_{\text{RIF}} = 0$ for rifampicin, illustrated as the asterisk in Figure 2.6. As Submodel 2 does not favor any drug over the other, the predictive surface is symmetric with respect to the line $n_{\text{INH}} = n_{\text{RIF}}$. The extra flexibility provided by the full model shifts the predictive distribution towards the observed data. While the distribution under the full model comfortably accommodates the empirical point in a high density region, the predictive distribution under Submodel 2 is much more diffuse. This indicates that while the observed data is not unsupported under Submodel 2, it is far more likely to be observed under the full model. As a result, we conclude that the evidence favours the drugs being acquired at different rates; specifically, isoniazid resistance evolves faster than rifampicin resistance.



(a) Full model

(b) Submodel 2: $\rho_{\text{INH}} = \rho_{\text{RIF}}$

Figure 2.6: Posterior predictive distribution of $(n_{\text{INH}}, n_{\text{RIF}})$ under the full model (panel (a)) and Submodel 2 (panel (b)). Darker intensity indicates higher posterior predictive density. The asterisk (*) indicates the observed data (8, 0).

2.5.3 The relative contribution of transmission and treatment failure to MDR-TB

In addition to estimating the rates of acquisition of drug resistance and assessing whether rates differ, we may also consider where doubly resistant cases come from. That is, estimation of the relative contribution to multidrug resistant cases of transmission of existing MDR-TB strains compared to treatment failure leading to evolution of multidrug resistance. The posterior predicted samples generated under the full model provide a clear portrait of the relative contribution of the different paths to achieving double resistance (see e.g. Luciani et al., 2009) for an additional illustration of this procedure.

Table 2.5 shows the means, medians and the 95% HPD credible intervals for the predicted proportion of cases of double resistance from each potential source. These proportions are obtained conditionally on there being at least one case of double resistance in the predictive sample. Simulated samples of this nature account for 99.67% of all predictive samples. The predictive distributions of the proportions are highly asymmetric (not shown), making the median a more reliable point estimate than the mean.

In the overwhelming majority of posterior predictive samples, direct *transmission* was the main source of acquisition of double resistance, followed by *conversion* in a single step directly from a sensitive profile (from profile 0 to MDR) and conversion in two steps via a state of resistance to a single drug (from profile 0 to INH to MDR, or from 0 to RIF to

MDR). This analysis corroborates the finding from Section 2.5.1 that ρ_{MDR} is most likely positive, and furthermore that this path is likely to be of even greater importance than conversion in two steps.

Source	Median	Mean	95% Credible Interval
Transmission	0.9975	0.9655	(0.7826, 0.9999)
Conversion in one step	0.0023	0.0284	(0.0000, 0.1667)
Conversion in two steps	0.0000	0.0060	(0.0000, 0.0073)

Table 2.5: Contributions to MDR-TB from alternative sources. This table contains the posterior medians and means and lower and upper limits of the 95% HPD credibility intervals for the proportion of double resistance cases originating from each possible source.

2.6 Conclusions

In this chapter we have estimated epidemiological parameters describing the acquisition of multi-drug resistance in *M. tuberculosis* from molecular epidemiological data (Monteserin et al., 2013) using approximate Bayesian computation. The underlying model is intended to capture essential processes that give rise to the data, namely, transmission of the disease, recovery or death, and within-host evolution giving rise to drug resistance and new genotypes at the molecular marker loci. From this analysis we may draw three major biological conclusions about the manner in which drug resistance arises.

First, there is an asymmetry in the acquisition of resistance to isoniazid and rifampicin. Specifically, isoniazid resistance occurs approximately an order of magnitude more frequently than resistance against rifampicin (see Table 2.4). This asymmetry in rates is consistent with *in vitro* (that is, through laboratory experiments) microbiological estimates of mutation rates per cell generation which find around 1 to 2 orders of magnitude difference between the two rates (David, 1970; Ford et al., 2013).

Second, the analysis supports the occurrence of *direct conversion* from doubly drug sensitive to doubly resistant (MDR) infections. This may be initially unintuitive because under mutation alone, if mutation occurs at rate ρ per gene per unit time, the rate of appearance of double mutants is ρ^2, which would be vanishingly small if ρ is low. However, using a mathematical model, Colijn et al. (2011) argued that direct conversion can occur surprisingly fast because resistant cells are sometimes present at low frequencies in a within-host population even before treatment commences. Our analysis of data at the

epidemiological level is consistent with that theoretical result. This direct conversion to double resistance is epidemiologically important as it accelerates the accumulation of resistance, in that resistance evolution does not have to take place sequentially. Once double resistant mutants appear, transmission of these mutants further increases their prevalence in the population.

Third, the overwhelming majority of cases of multidrug resistant tuberculosis come from transmission of already multidrug resistant strains (see Table 2.5), a finding that is consistent with those of Luciani et al. (2009). This large contribution of transmission occurs despite the 10% transmission cost of each resistance which results in a $\sim 20\%$ cost for MDR-TB. This implies that in controlling drug resistance, although there is widespread concern about treatment failure leading to rising resistance, most resistant cases may be due to transmission. Therefore, although it is important to support treatment adherence, public health efforts may benefit from focusing more on preventing disease transmission. That is, control measures that reduce the incidence of new cases are likely to help reduce MDR-TB.

By developing epidemiological models with evolutionary processes we have been able to estimate parameters describing how drug resistance – particularly multidrug resistance – emerges in *M. tuberculosis*. Although there is existing knowledge of rates of mutation to resistant states *in vitro*, there is a need to assess the extent to which those rates translate to the epidemiological level. Large scale molecular epidemiological models, such as those presented here, are highly complex and multidimensional, and as such, likelihood-based analyses are not straightforward mathematically or computationally. In such cases, approximate Bayesian computation methods present a practical and viable approach to making statistical inferences, particularly as continually advancing molecular technologies require dynamical models to be extended and refined.

Chapter 3

Functional regression approximate Bayesian computation for Gaussian process density estimation [1]

3.1 Introduction

We introduce a new statistical procedure for hierarchical modelling on a set of related densities $f_i(x)$ for $i = 1, \ldots, g$, based on random samples $\{X_{ij}\}$ such that $X_{ij} \sim f_i(x)$ for $j = 1, \ldots, n_i$. The benefits of hierarchical modelling are well known in the parametric context, where the same model is fitted to different but related datasets or groups, and where model parameters are allowed to vary across groups (Gelman et al., 2004). A hierarchical prior may be defined on the group varying parameters, and the data used to determine how much pooling of information to perform across groups. The problem considered here is the nonparametric equivalent of this. Here, the nonparametric density functions $f_i(x)$ are thought to be related and we aim to share information hierarchically to improve estimation of each density, especially for those densities $f_i(x)$ for which the corresponding sample size n_i is small.

Bayesian nonparametric methods have made enormous advances over the last two

[1] Published as: Rodrigues, G. S., Nott, David J. and Sisson, S. A. (2016), "Functional regression approximate Bayesian computation for Gaussian process density estimation." *Computational Statistics and Data Analysis*, 103, 229-241.

decades. The Dirichlet process (DP) (Ferguson, 1973) has played a central role in this development. Methods based on Dirichlet process mixture (DPM) models, where a mixing distribution is given a Dirichlet process prior, are a standard approach to flexible Bayesian density estimation (Lo, 1984; West et al., 1994; Escobar and West, 1995). These methods also have extensions to grouped data and the estimation of a set of related density functions, the problem considered here. Important methods for grouped data include the analysis of densities model of Tomlinson and Escobar (1999) which uses a DPM model for each density in which the base measure for the mixing densities is the same and given a DPM prior; the hierarchical Dirichlet process (Teh et al., 2006) where mixing distributions are given DP priors with a common base measure which is given a DP prior; Dirichlet process mixture of ANOVA models (De Iorio et al., 2004) where the atoms in a Dirichlet process are modelled as dependent on a covariate following an ANOVA type dependence structure; and the hierarchical model of Muller et al. (2004) in which distributions are modelled as a mixture of a common and group specific component, with these components being given DP mixture priors.

One alternative to DP based methods in Bayesian nonparametrics is the use of Gaussian processes (Leonard, 1978; Thorburn, 1986; Lenk, 1988, 1991; Tokdar, 2007; Tokdar and Ghosh, 2007). However, to the best of our knowledge hierarchical versions of Gaussian process priors for grouped data situations have not been developed in the literature (for the case of non-grouped data, see Flaxman et al., 2016). Gaussian process methods can be attractive because the parameters in the resulting priors allow very easy expression of relevant prior information, such as smoothness of the densities and, in the hierarchical setting, the extent of sharing information between groups. For the DP mixture based methods, on the other hand, generally prior information must be expressed through a prior on a mixing distribution, through which it is difficult to adequately express similar prior beliefs. One reason that Gaussian process density estimation methods are not more popular is perhaps the computational difficulty. Such difficulties are even more acute in the hierarchical setting involving grouped data.

Here we introduce a hierarchical Gaussian process (HGP) prior, formulated as a univariate hierarchical extension of the multivariate prior proposed by Adams et al. (2009), and discuss tractable methods for computation with this prior. Our construction, besides being able to handle an arbitrary number of hierarchy levels, provides a convenient way

of expressing prior beliefs regarding both the degree of similarity between the densities and the nature of their characterising features, such as smoothness and support. We also establish a remarkably different approach for making inference. Instead of relying on Markov chain Monte Carlo (MCMC) methods to draw samples from the posterior distribution, as is commonly implemented in other approaches to Gaussian process density estimation (Adams et al., 2009) which can suffer from poor performance, we alternatively introduce an approximate Bayesian computation (ABC) (Beaumont et al., 2002) functional regression-adjustment to draw approximate samples from the posterior. The use of ABC to estimate functional objects in itself represents an important contribution to the ABC literature. Moreover, this approach provides a great practical advantage in terms of flexibility, as, for most cases, changes to the prior specification are operationally straightforward to accommodate and do not require the derivation and coding of a new MCMC sampler.

In Section 3.2, we introduce the hierarchical Gaussian process prior and present an algorithm for sampling data from this prior. Section 3.3 describes the inferential strategy for estimating the functional parameters. Performance of the proposed density estimator is investigated in the simulation studies in Sections 3.4 and 3.5. Finally, in Section 3.6, we use the proposed estimator to compare rural high school exam performance across states in Brazil.

3.2 The hierarchical Gaussian process prior

We specify a hierarchical Gaussian process prior on the set of densities $f_i(x)$, $i = 1, \ldots, g$, as follows:

$$f_i(x) = \frac{L(Z_i(x))b(x|\phi)}{c_i(\phi, Z_i)} \tag{3.1a}$$

$$Z_i(x) \sim \mathcal{GP}(\mu(x), k(x, x'|\theta_Z)) \tag{3.1b}$$

$$\mu(x) \sim \mathcal{GP}(m(x), k(x, x'|\theta_\mu)) \tag{3.1c}$$

$$(\theta_Z, \theta_\mu) \sim \pi(\cdot). \tag{3.1d}$$

Here $L(z) = 1/(1 + \exp(-z))$ denotes the logistic function and $\mathcal{GP}(\mu(x), k(x, x'|\theta))$ represents a Gaussian process with mean function $\mu(x)$ and covariance function $k(x, x'|\theta)$

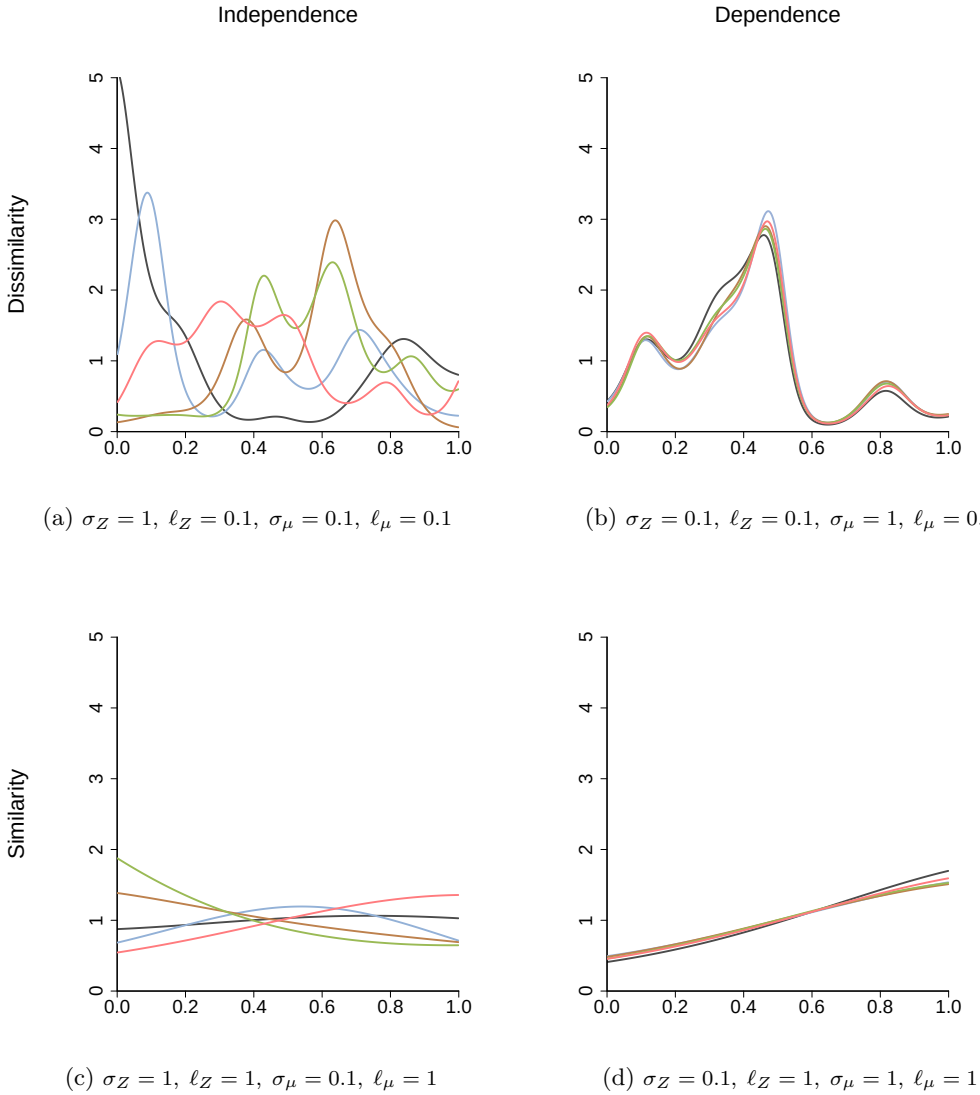
with parameters θ. $b(x|\phi)$ is an arbitrarily chosen parametric base density with hyperparameters ϕ. It can be regarded as the modeller's 'best guess' of the densities $f_i(x)$ and the 'center' of the prior distribution – the term $L(Z_i(x))$ acts by 'deforming' $b(x)$ to create new densities. The function $m(x)$ is a conveniently chosen function discussed further below, $\pi(\cdot)$ is a hyperprior for the parameters of each covariance function, and $c_i(\phi, Z_i) = \int L(Z_i(x))b(x|\phi)dx$ is the normalising constant of $f_i(x)$.

Gaussian processes are a common tool for modelling functional data and can be understood as an infinite-dimensional generalization of the Gaussian distribution. A detailed review of Gaussian processes is found in Rasmussen and Williams (2006). See also Shi and Choi (2011) for a comprehensive treatment of Gaussian processes in functional data analysis.

Equation (3.1a) defines a deterministic map from auxiliary functions $Z_i(x)$ to the normalised densities $f_i(x)$. Operating on the transformed space avoids difficulties otherwise implied by the intrinsic properties of density functions – namely, that they be non-negative and integrate to one. Under this prior, all marginal densities are considered potentially related, and bound by a common Gaussian process mean function $\mu(x)$ (Equation 3.1b), which is itself unknown. The above formulation, though simple, will be shown to be highly adaptable, and capable of adequately describing a wide range of possible prior beliefs. For ease of presentation only two levels of hierarchy are described above, but it is straightforward to incorporate multiple hierarchical levels, as is illustrated in the analysis of Brazilian school exam performance in Section 3.6.

For this article, we adopt as covariance function the squared exponential kernel $k(x, x') = \sigma^2 \exp\left(-(x - x')^2/(2\ell^2)\right)$, although other options are easily accommodated. The hyperparameters $(\theta_Z, \theta_\mu) = (\sigma_Z, \ell_Z, \sigma_\mu, \ell_\mu)$ play a crucial role in the behaviour of $f_i(x)$. Figure 1 shows $g = 5$ densities drawn from the proposed prior under various parameter settings for $(\sigma_Z, \ell_Z, \sigma_\mu, \ell_\mu)$. The standard uniform distribution was chosen as the base density $b(x|\phi)$ and we set $m(x) = -10$ (the motivation for this choice will be made clear in Section 3.3).

While the ℓ's are length-scale parameters that determine the speed with which the covariance decays as a function of the distance $|x - x'|$, the σ parameters dictate how similar the densities are to $b(x|\phi)$, with larger σ indicating less similarity. The ratio σ_Z/σ_μ defines how closely related the different densities are: if the ratio is large, the densities are nearly independent (Figures 3.1a and 3.1c). On the other hand, if the ratio



(a) $\sigma_Z = 1,\ \ell_Z = 0.1,\ \sigma_\mu = 0.1,\ \ell_\mu = 0.1$

(b) $\sigma_Z = 0.1,\ \ell_Z = 0.1,\ \sigma_\mu = 1,\ \ell_\mu = 0.1$

(c) $\sigma_Z = 1,\ \ell_Z = 1,\ \sigma_\mu = 0.1,\ \ell_\mu = 1$

(d) $\sigma_Z = 0.1,\ \ell_Z = 1,\ \sigma_\mu = 1,\ \ell_\mu = 1$

Figure 3.1: Samples from the prior distribution based on the squared exponential function with $g = 5$, under varying prior conditions. The above panels show: (a) independence and dissimilarity to the base density ($\sigma_Z = 1, \ell_Z = 0.1, \sigma_\mu = 0.1, \ell_\mu = 0.1$); (b) dependence and dissimilarity to the base density ($\sigma_Z = 0.1, \ell_Z = 0.1, \sigma_\mu = 1, \ell_\mu = 0.1$); (c) independence and similarity to the base density ($\sigma_Z = 1, \ell_Z = 1, \sigma_\mu = 0.1, \ell_\mu = 1$), and (d) dependence and similarity to the base density ($\sigma_Z = 0.1, \ell_Z = 1, \sigma_\mu = 1, \ell_\mu = 1$).

is small, most of the variability is due to the variation of the common mean, which results in very similar marginal densities (Figures 3.1b and 3.1d).

3.2.1 Sampling from the prior

Approximate draws from this prior may be naturally obtained by sequentially sampling the unknown parameters from the top (Equation (3.1d)) to the bottom (Equation (3.1a)) of the hierarchy. It is possible to derive an exact sampler using an extension of Algorithm 3.1 in Adams et al. (2009). However, for a faster simulator, we favor a finite-dimensional ap-

proximation as in Tokdar (2007), which results in the so-called surrogate prior. Therefore, while the prior introduced in Equation (3.1) is defined over infinite-dimensional parameters, our inference scheme is based on a finite approximation, which differs from the implementation in Adams et al. (2009) and Muller et al. (2004).

In particular, we first sample the hyperparameters $(\theta_Z, \theta_\mu) \sim \pi(\cdot)$ from (3.1d). Next, define x_{low} and x_{high} to be the lower and upper β-quantiles of $b(x|\phi)$, where β is chosen to be small, and construct a regular grid of k points on the range $[x_{\text{low}}, x_{\text{high}}]$, i.e., $\psi_1 < \ldots < \psi_k$, with $\psi_j = x_{\text{low}} + (x_{\text{high}} - x_{\text{low}})(j-1)/(k-1)$, for $j = 1, \ldots, k$.

It follows from the Gaussian process definition that $\mu^\psi = (\mu(\psi_1), \ldots, \mu(\psi_k))^\top$ follows a k-variate normal distribution with mean $m^\psi = (m(\psi_1), \ldots, m(\psi_k))^\top$ and covariance matrix Σ, with typical element $\Sigma_{ij} = k(\psi_i, \psi_j|\theta_\mu)$. Sampling μ^ψ is therefore trivial. Similarly, we obtain $Z_i^\psi = (Z_i(\psi_1), \ldots, Z_i(\psi_k))^\top$, $i = 1, \ldots, g$, as draws from the multivariate normal distribution $N(\mu^\psi, \Lambda)$, with covariance matrix defined as $\Lambda_{ij} = k(\psi_i, \psi_j|\theta_Z)$. The normalising constants $c_i(\phi, Z_i)$ may then be estimated using numerical integration. Through (3.1a), this uniquely determines the corresponding densities $f_1^\psi, \ldots, f_g^\psi$.

Each vector Z_i^ψ and f_i^ψ represents a discretized observation of the corresponding underlying function. Continuous approximations for $Z_i(x)$ and $f_i(x)$ can then be obtained by fitting B-splines to each of these vectors via least squares, resulting in $\tilde{Z}_i(x)$ and $\tilde{f}_i(x)$, respectively. The quality of this approximation, including the degree to which $\int \tilde{f}_i(x)dx \approx 1$, is controlled by the number of points in the grid, k, and the number of basis functions used. In general, the approximation errors introduced by the B-spline approximation tend to be dominated by the ABC approximation error, however it is good practice as part of the inference procedure to investigate the quality of the B-spline approximation to the synthetic KDEs, and increase the number of grid points if required. Finally, data X_{ij} may be generated from $\tilde{f}_i(x)$ using a rejection sampling algorithm with $b(x|\phi)$ as the proposal distribution. A candidate x^* is accepted with probability $\tilde{f}_i(x^*)/[Mb(x^*|\phi)]$, where $M = \max\{\tilde{f}_i(x)/b(x|\phi)\}$. The value of M is unknown, but may be determined by numerical search.

In practice, choosing an appropriate base density is important to promote a reasonable algorithmic efficiency. In this regard, perhaps more important than the actual shape of this density, making sure that the support of $b(x|\phi)$ is similar to the support of the posterior predictive distribution is particularly critical.

3.3 An approximate Bayesian inferential procedure

The resulting posterior distribution is given as

$$\begin{aligned} p(\mathbf{Z}, \theta|\mathcal{D}) &= \frac{p(\mathbf{Z}, \theta)p(\mathcal{D}|\mathbf{Z})}{\int\int p(\mathbf{Z}', \theta')p(\mathcal{D}|\mathbf{Z}')d\mathbf{Z}'d\theta'} \\ &\propto \pi(\theta_Z, \theta_\mu)p(\mu(x)|\theta_\mu)\prod_{i=1}^{g} p(Z_i(x)|\mu(x), \theta_Z)\prod_{j=1}^{n_i}\frac{L(Z_i(x_{ij}))b(x_{ij}|\phi)}{c_i(\phi, Z_i)}, \end{aligned}$$

where p denotes a joint or conditional distribution, $\theta = (\theta_Z, \theta_\mu)$, $\mathbf{Z}$ compactly denotes the set of functions $Z_1(x), \ldots, Z_g(x)$ and $\mathcal{D}$ represents the observed data. The above posterior distribution is computationally difficult to work with directly. As an alternative we develop an approximate Bayesian computation procedure. ABC methods have been extensively developed to draw samples from an approximation to the posterior distribution, based only on the ability to sample data from the model, without the need to evaluate the posterior directly. See e.g. Beaumont et al. (2002); Blum et al. (2013); Nott et al. (2014) for more details on ABC methods. As part of this inference, we extend the ideas behind the ABC regression-adjustment (Beaumont et al., 2002) to the functional setting.

Samples from the posterior distribution can be used to estimate any quantity of interest, such as the mean or quantiles of $f_i(x)$. Here, the posterior mean $\mathbb{E}[f_i(x)|\mathcal{D}]$, which is a function itself, is a Bayes estimator for $f_i(x)$. Note that while any given sample from $p(Z_i(x)|\mathcal{D})$ corresponds to a unique sample from $p(f_i(x)|\mathcal{D})$, the reverse does not hold.

3.3.1 ABC for Gaussian process density estimation

Approximate Bayesian computation methods implement an approximate Bayesian inference based on the matching of simulated and observed summary statistics. A typical rejection-sampling ABC algorithm applied to this particular context would be as follows:

1. Draw a sample (a set of density functions) from the hierarchical Gaussian process prior using the method described in Section 3.2.1, and generate synthetic data from these densities of the same size as the observed data.

2. Summarise the synthetic data using a set of summary statistics. We adopt the well known kernel density estimator, computed for each group, $i = 1, \ldots, g$.

3. If the resulting summary statistic is, by an appropriate metric, similar to the sum-

mary statistic computed over the observed data, accept the densities generated in step 1. Otherwise these are rejected.

4. Repeat steps 1–3 to produce m accepted samples.

The popularity and computational simplicity of the kernel density estimator makes it a suitable candidate to play the role of the summary statistic in the ABC procedure. In particular, we employ a Gaussian kernel and set the bandwidth as $h = \hat{\sigma}(4/3n)^{-1/5}$ where $\hat{\sigma}$ denotes the sample standard deviation (Scott, 1992). We denote by $K_{il}(\psi_j)$ the resulting kernel density estimate obtained from the data in group i in synthetic dataset l evaluated at ψ_j, for $i = 1, \ldots, g$, $l = 1, \ldots, m$. Alternative density estimators can be adopted, providing they are computationally fast.

Step 3 above requires the specification of a measure of similarity between the synthetic and observed data. Here, a natural and convenient option is to define the divergence of synthetic dataset l from the observed data as

$$D_l = \sum_{i=1}^{g} \sum_{j=1}^{k} \Big(\log(K_i^{obs}(\psi_j)) - \log(K_{il}(\psi_j)) \Big) K_i^{obs}(\psi_j), \tag{3.2}$$

where $K_i^{obs}(\psi_j)$ denotes the kernel density estimator described above applied to the observed data in group i, evaluated at ψ_j. Therefore, the contribution of group i is proportional to a quadrature estimate of the Kullback-Leibler divergence of K_{il} from its (observed) counterpart K_i^{obs}. The Kullback-Leibler measure imposes heavier penalties for differences occurring in regions of higher estimated likelihood, which positively contributes to its performance. Other valid divergence measures could be formulated based, for example, on integrated squared errors or on the Kolmogorov-Smirnoff statistic.

In order to produce more accurate posterior samples, we follow Beaumont et al. (2002) and weight the samples according to how well the synthetic data reproduces the observed data through (3.2) using the Epanechnikov kernel $w(D) = 1 - (D/\delta)^2$ for $D \leq \delta$ and 0 otherwise. Here δ is a threshold which determines how close the synthetic and observed summary statistics must be before the sample densities are accepted as an approximate draw from the posterior distribution in step 3. Its value is often determined in terms of a quantile of the obtained divergences of a large number of samples $\{D_l\}$ (Beaumont et al., 2002; Blum et al., 2013).

Notice that the Kullback-Leibler divergence between synthetic and observed (func-

tional) summary statistics is not affected by the choice of k. Therefore, increasing the number of grid points does not lead to the well-known curse of dimensionality problem, even though it does add to the method's computational burden, given that the functions then need to be evaluated at more locations.

3.3.2 A functional regression-adjustment

As with all ABC algorithms, the approximation error associated with the ABC method described in Section 3.3.1 can be substantial, unless the threshold δ is small. However, if δ is too small, then the acceptance rate of the algorithm will be prohibitively low. Beaumont et al. (2002) proposed the regression-adjustment as a simple approach to adjust the posterior samples obtained with $\delta > 0$, so that $\delta \approx 0$. In the simple parametric regression model considered, a (local) linear relationship

$$\vartheta = \alpha + \beta^\top (s - s_{obs}) + \epsilon$$

is assumed to hold between the model parameters ϑ and the vector of simulated and observed summary statistics, s and s_{obs}, where ϵ denotes a zero mean error. In the present setting, s and s_{obs} respectively represent the kernel density estimator summaries K_i and K_i^{obs} for each group i. If this regression model approximately holds, then

$$\vartheta^* = \vartheta - \hat{\beta}^\top (s - s_{obs}),$$

where $\hat{\beta}$ is the least-squares estimate of β, is an approximate draw from the posterior with $s = s_{obs}$ (i.e. $\delta = 0$.) This expression for ϑ^* is obtained as the fitted regression mean at s_{obs}, which is $\hat{\alpha}$, plus the empirical residual $\vartheta - \hat{\alpha} - \hat{\beta}^\top (s - s_{obs})$. See Beaumont et al. (2002); Blum and François (2010); Blum et al. (2013) for further discussion of standard regression-adjustment methods. In the present setting, we may similarly perform a regression-adjustment, but using a functional regression model (Ramsay and Silverman, 2005) due to the functional nature of the parameters and summary statistics (K_i and K_i^{obs}).

From the definition of the logistic function, $\log(L(z)) = -\log(1 + \exp(-z)) \approx z$, for negative values of z. As a specific example, $\log(L(-5)) = -5.0067$. Therefore, Equation

(3.1a) can be rewritten as

$$Z_i(x) \approx \log\left(\frac{f_i(x)c_i(\phi, Z_i)}{b(x|\phi)}\right). \tag{3.3}$$

This approximation is accurate up to the fourth decimal place for $Z_i(x) \leq -10, \forall x$, which can be easily achieved through appropriate choice of $m(x)$. In this article we set $m(x) = -10$. In this operating region, Equation (3.3) implies that changes to the choice of $m(x)$ (provided it remains a constant function) only affect the normalising constants $c_i(\phi, Z_i)$, with no impact on the densities $f_i(x)$. A useful, additional advantage of using a negative mean function for $\mu(x)$ is that this prior does not give strong support to unrealistic, flat-peaked densities, which can happen if one sets $m(x) = 0$. Equation (3.3) forms the basis to appropriately specify a functional regression model. For group $i = 1, \ldots, g$ and synthetic dataset l, we write

$$\tilde{Z}_{il}(x) = \text{offset} + \gamma_0^i(x) + \gamma_1^i(x)\log(\tilde{K}_{il}(x)) + \gamma_2^i(x)\log(\tilde{K}_l^p(x)) + \epsilon^l(x), \tag{3.4}$$

where $\tilde{Z}_{il}(x)$ denotes a B-spline fitted to $Z_{il}(\psi_1), \ldots, Z_{il}(\psi_k)$, $\gamma_0^i(x), \gamma_1^i(x)$ and $\gamma_2^i(x)$ are functional regression parameters to be estimated, and $\epsilon^l(x)$ is a realisation of a zero-mean functional error term $\epsilon(x)$. Recall that $\tilde{Z}_{il}(x)$ is the l-th synthetic realisation of the random function $\tilde{Z}_i(x)$, which in turn represents the B-spline approximation of $Z_i(x)$ (the Gaussian process associated to group i). The offset term is given by $\log\left(\frac{b(x|\phi)}{c_i(\phi, \tilde{Z}_{il})}\right)$ and is computed through (3.1a), $\tilde{K}_{il}(x)$ denotes a B-spline fitted to $K_{il}(\psi_1), \ldots, K_{il}(\psi_k)$, and $\log(\tilde{K}_l^p(x))$ is the KDE computed over the pooled data. With this construction, while $\log(\tilde{K}_{il}(x))$ conveys information from the i-th group itself, $\log(\tilde{K}_l^p(x))$ explores what has been observed in the combined data, irrespective of the group classification. Accordingly, the functional term $\gamma_2^i(x)$ defines how much strength should be borrowed, and from where on the density as a function of x.

From (3.3) and (3.4) it can be seen that the data relates to $\tilde{Z}_i(x)$ through the function

$$\log\left([\tilde{K}_i(x)]^{\gamma_1(x)}[\tilde{K}^p(x)]^{\gamma_2(x)}\right).$$

This means we are ultimately using the logarithm of a weighted geometric mean between the group-specific and pooled kernel densities to model the dependent functional object.

By no means we are suggesting that our final estimate is simply a weighted geometric average – there are other terms in the model.

Based on (3.4), approximate samples from the marginal posterior distribution $p(Z_i(x)|\mathcal{D})$ can then be obtained by performing the functional regression adjustment

$$\tilde{Z}^*_{il}(x) = \tilde{Z}_{il}(x) - \hat{\gamma}^i_1(x)[\log(\tilde{K}_{il}(x)) - \log(\tilde{K}^{obs}_i(x))] - \hat{\gamma}^i_2(x)[\log(\tilde{K}^p_l(x)) - \log(\tilde{K}^{p,\,obs}(x))], \tag{3.5}$$

where $\hat{\gamma}^i_1(x)$ and $\hat{\gamma}^i_2(x)$ are least-squares estimates of $\gamma^i_1(x)$ and $\gamma^i_2(x)$ (obtained using the `R` package `fda`, Ramsay et al. (2013)), and where $\tilde{K}^{obs}_i(x)$ and $\tilde{K}^{p,\,obs}(x)$ are the equivalent summary statistics obtained from the observed data.

Note that the right-hand side of the proposed functional regression model (3.4) is only defined for strictly positive kernel estimates. If necessary, this can be artificially satisfied by adding an arbitrary tiny constant to the kernel estimates. This modified summary statistic ensures the model is well-defined and may provide extra stability.

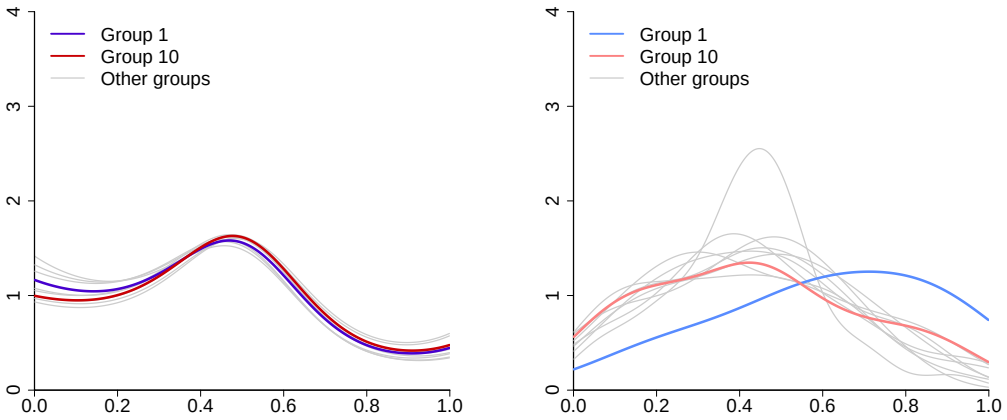
Finally, while the above regression models are intuitive and seem sensible, as with the parametric regression-adjustment (Beaumont et al., 2002) they remain only a model for the perhaps complex relationships that exist between the density estimates and the summary statistics. Given that the regression model is fitted to those samples for which $D_l < \delta$, this means that the local-linearity assumption of (3.4) may be adequate for many situations providing that δ is sufficiently small. As for the parametric regression-adjustment, more complex models can be developed if required (Blum and François, 2010).

3.4 A simulated example

In this section we present a simulated analysis to illustrate some important features of the ABC density estimator. We construct $g = 10$ groups, with sample sizes $5, 20, 35, \ldots, 140$. The base density is taken to be the standard uniform distribution and the prior for the parameters of the squared exponential covariance functions is $\sigma_Z, \sigma_\mu, \ell_Z, \ell_\mu \sim \text{Gamma}(3, 5)$, which is fairly uninformative. The gamma distribution provides good *a priori* control over the densities' features, as compared to e.g. the single-parameter half-Cauchy distribution, and can be easily specified to avoid placing too much mass in unrealistic regions of the parameter space. In particular, very low values of the length-scale parameters ℓ (see Figure

3.1) induce remarkably erratic densities (with many modes) that would be unreasonable in many applications. Further, the inclusion of such erratic densities in the functional regression adjustment can affect the quality of the resulting ABC posterior approximation. While this effect can be minimized by increasing the number of samples generated from the prior in order to produce more samples closer to the observed density, the gamma prior provides a good balance of modelling flexibility and computational efficiency.

In this example, the 'true' densities are a random realization from the prior, with the exception that a moderate degree of similarity between the densities was specified by fixing $\sigma_Z = 0.5$ and $\sigma_\mu = 1$. The 'observed' data were then sampled from the true densities using the procedure described in Subsection 3.2.1. We generate $50,000$ samples from the prior distribution and accept the $m = 5,000$ samples closest to the observed data. We use $k = 100$ grid points and fit the B-spline functions using 50 basis terms.

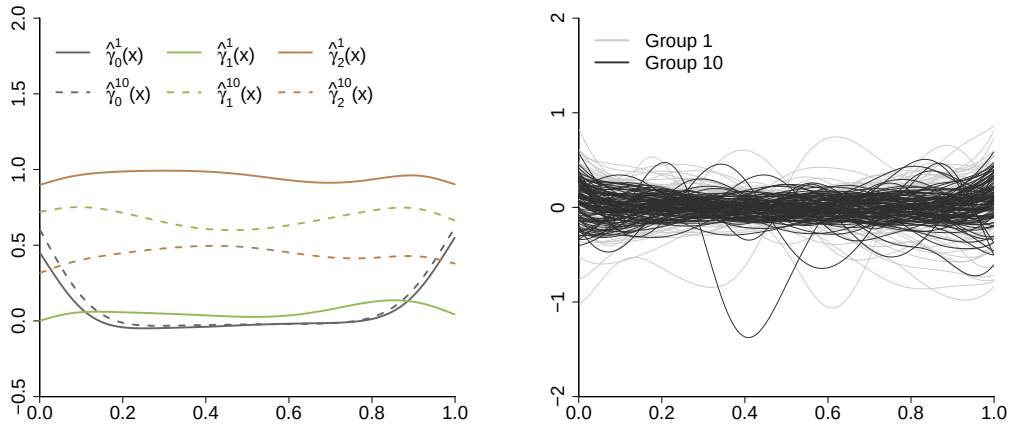


(a) True densities

(b) Kernel density estimates

Figure 3.2: (a) True densities and (b) kernel density estimates of the simulated data from $g = 10$ groups. Group 1 (5 datapoints) and group 10 (140 datapoints) are highlighted.

We focus our attention on groups 1 and 10, as they have the smallest and largest sample sizes, respectively. Figure 3.2 shows the true densities (left panel) and the corresponding kernel density estimates (right panel) of each of the groups, with the latter also representing a first crude estimation of the respective densities. Although the density estimate for group 10 looks reasonable, the estimation for group 1 is clearly unsatisfactory due to the low sample size. The well known boundary-bias of this kernel density estimator (Wand et al., 1991; Chen, 1999; Jones, 1993; Jones and Henderson, 2007; Dai and Sperlich, 2010; Geenens, 2014) is clearly visible as a downward dip in the density estimates near the



(a) Estimated regression coefficients

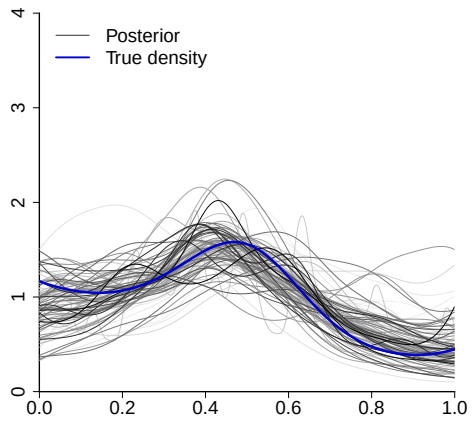
(b) Residual plot

Figure 3.3: (a) Least-squares estimates of the functional regression coefficients $\gamma_0(x), \gamma_1(x)$ and $\gamma_2(x)$ for groups 1 (solid lines) and 10 (dashed lines); (b) Least-squares functional residuals for groups 1 (black lines) and 10 (grey lines).

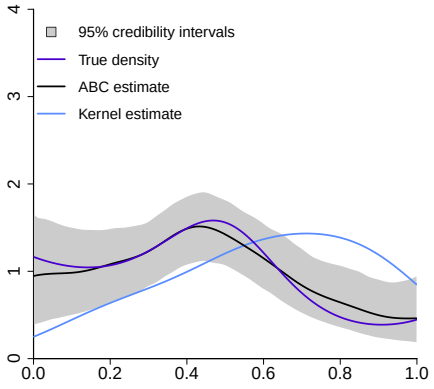
borders of the support.

Figure 3.3 (left panel) shows the estimated functional regression coefficients for groups 1 (solid lines) and 10 (dashed lines). The intercepts γ_0^1 and γ_0^{10} (black lines) are effectively zero except for upward ticks near the boundaries. In effect, the regression model is both identifying and attempting to correct for the boundary-bias in the density estimates. The other regression parameters describe the contribution of $\log(\tilde{K}_i^{obs}(x))$ and $\log(\tilde{K}^{p,\,obs}(x))$ in estimating $\tilde{Z}_i(x)$. For Group 1, virtually all relevant information comes from the kernel estimates of the pooled data, as $\hat{\gamma}_1^1$ (solid grey line) is near zero over the entire range. For Group 10, the situation is reversed – most of the information comes from its own kernel density ($\hat{\gamma}_1^{10}(x)$ is large), although it still borrows strength from other groups as $\hat{\gamma}_2^{10}(x)$ (light grey dashed line) is far from zero everywhere. All parameter estimates are not constant, indicating that the relationship varies in different regions of the data-space. For groups 2 to 9 (plots not shown), a steady transition is observed between the results of groups 1 and 10.

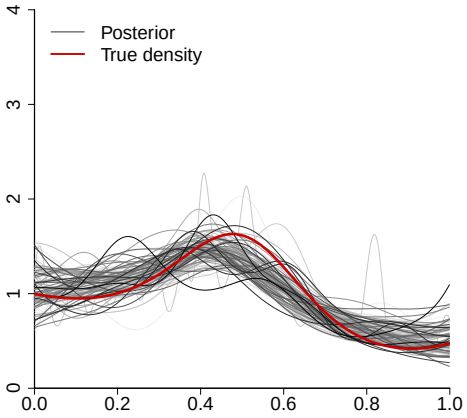
The right panel of Figure 3.3 shows the plot of the estimated functional residuals given by $\hat{\varepsilon}_{il} = \tilde{Z}_{il}(x) - \hat{Z}_{il}(x)$, for $i = 1, 10$, where $\hat{Z}_{il}(x)$ denotes the fitted regression model. For both groups, the residuals are centred on zero everywhere, with greater variability near the boundary as expected. In addition, the residuals are more variable for group 1 than for group 10 due to the differences in the sample sizes.



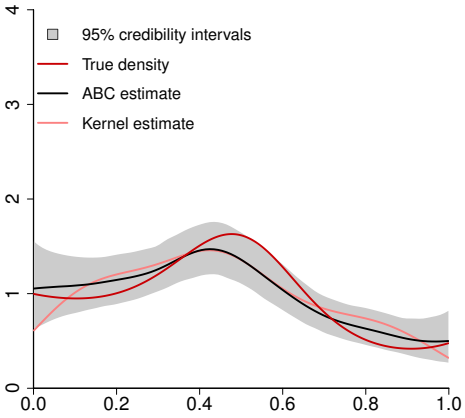
(a) Samples from the posterior - Group 1



(b) 95% credibility intervals - Group 1



(c) Samples from the posterior - Group 10



(d) 95% credibility intervals - Group 10

Figure 3.4: Left panels: Samples from the posterior distribution for groups 1 and 10 (grey lines), with a darker line indicating greater sample weight. True density is indicated by the coloured line. Right panels: Comparison of the true densities, the initial kernel density estimate and the ABC posterior mean $\mathbb{E}[f_i(x)|\mathcal{D}]$ with pointwise 95% central credible intervals, for groups 1 and 10.

The left panels in Figure 3.4 illustrate 100 randomly chosen approximate samples from the posterior. The darker lines indicate larger sample weights. It is clear that the samples cluster around the true densities, with the variability for group 1 (small sample size) being greater than that of group 10. The right panels in Figure 3.4 depict pointwise 95% central credibility intervals of the posterior distribution, the posterior mean $\mathbb{E}[f_i(x)|\mathcal{D}]$, the initial group kernel density estimate and the true density function. As expected, the ABC posterior mean strongly outperforms the initial kernel density estimate. For group 1, the hierarchical sharing of information between density estimates has produced a substantially improved and more accurate estimate. Even for group 10, which has the most data, the

posterior mean is more accurate, and less biased at the boundaries than the initial kernel density estimate.

3.5 Model comparison

We now investigate the performance of the proposed estimation procedure via a simulation study. Because the posterior distribution is inaccessible when using the hierarchical Gaussian process prior, it is not possible to directly compare the ABC estimates with their exact counterparts. To overcome this hurdle, we perform an indirect comparison with the aid of an alternative prior distribution for which exact results are available. Here we choose to compare our HGP prior with the DPM prior introduced by Muller et al. (2004), as the latter is also suitable for hierarchical modelling and a version of its MCMC sampler is fully implemented in the `R` package `DPpackage` (Jara et al., 2011). Precisely, we consider the following prior specifications:

Dirichlet process mixture prior	**Hierarchical Gaussian process prior**
$X_{ij} \sim f_i(x)$	$X_{ij} \sim f_i(x)$
$f_i(x) = \varepsilon\, h_{g+1}(x) + (1-\varepsilon)\, h_i(x)$	$f_i(x) \propto L(Z_i(x))\; b(x)$
$h_i(x) = \int \mathrm{N}(\mu, 1/3)\; dG_i(\mu)$	$Z_i(x) \sim \mathcal{GP}(\mu(x), k(x, x'\|1-\varepsilon, 1/\sqrt{2}))$
$G_i \sim \mathcal{DP}(M = 5, b(x))$	$\mu(x) \sim \mathcal{GP}(-10, k(x, x'\|\varepsilon, 1/\sqrt{2}))$
$b(x) = \mathrm{N}(0, 1)$	$b(x) = \mathrm{N}(0, 1.1)$
$\varepsilon = \mathrm{U}(0, 1)$	$\varepsilon \sim \mathrm{U}(0, 1),$

where $\mathcal{DP}(M, b(x))$ denotes a Dirichlet process with base probability distribution $b(x)$ and concentration parameter M. This exact specification produces structurally distinct but visually equivalent priors.

Notice that while the DPM prior writes the densities $f_i(x)$ as a mixture of infinite mixtures of normal distributions, the HGP prior builds them by bending the base density using a Gaussian processes-based transformation. In either case, the ε parameter dictates the degree of similarity between the groups, with $\varepsilon = 0$ implying conditional (on the hyperparameters) independence and $\varepsilon = 1$ forcing identical distributions. Therefore, letting $\varepsilon \sim \mathrm{U}(0, 1)$ works as an intermediate assumption in-between independence and

equal populations.

For a range of values of ε in $0, 0.1, \ldots, 1$, we simulated 500 realisations for each of those two generative processes. Every realisation constitutes a set of related group densities, treated hereafter as the 'true' densities, and a corresponding observed dataset, based on which we would like to make inference. As in Section 3.4, we consider 10 groups within each replicate, with sample sizes $5, 20, 35, \ldots, 140$.

Six methods are compared with respect to their ability to recover the true densities:

1. **KDE** The Kernel density estimator fitted to each group independently. No strength is borrowed across the groups.

2. **Pooled-KDE** All groups are assumed identical and a single KDE is fitted to the pooled data. This is the second covariate of the functional regression model in Equation 3.4.

3. **ABC-REJ** The rejection ABC algorithm introduced in Subsection 3.3.1. No adjustment is performed.

4. **ABC-REG** The functional regression adjustment ABC method described in Subsection 3.3.2 and illustrated in Section 3.4.

5. **ABC-REG2** Functional regression adjustment ABC based on an alternative model structure, detailed below.

6. **DPM** The Dirichlet Process Mixture estimator. Assuming the DPM prior shown above, MCMC samples from the posterior distribution are generated using the `R` function *HDPMdensity*.

With the functional regression model defined in Equation (3.4), the relative importance of each covariate, expressed by the estimated regression coefficients, only depends on the observed data through its sample size and the weights assigned to synthetic samples (see the last paragraph of Subsection 3.3.1). However, intuitively, we would like the amount of borrowing to increase, for example, if the observed data suggested that the groups are alike, according to some measure. With this in mind, one possible way to formally incorporate this feature into the adjustment is by the inclusion of interaction terms. In

particular, we additionally consider the alternative model (ABC-REG2):

$$\tilde{Z}_{il}(x) = \text{offset} + \gamma_0^i(x) + \gamma_1^i(x)\log(\tilde{K}_{il}(x)) + \gamma_2^i(x)\log(\tilde{K}_l^p(x)) + \gamma_3^i(x)\log(\tilde{K}_{il}(x))S_l + \gamma_4^i(x)\log(\tilde{K}_l^p(x))S_l + \epsilon^l(x),$$

where S_l is a measure of divergence between the groups, defined as

$$S_l = \sum_{i=1}^{g}\sum_{j=1}^{k}\Big(\log(\tilde{K}_l^p(\psi_j)) - \log(\tilde{K}_{il}(\psi_j))\Big)\tilde{K}_l^p(\psi_j).$$

Under this construction, the effect of $\log(\tilde{K}_l^p(x))$ on $\tilde{Z}_{il}(x)$ is given by $\gamma_2^i(x) + \gamma_4^i(x)S_l$, which enables the model to further exploit the available data.

Figure 3.5 summarizes the resulting performance of the six methods. For each fixed ε (horizontal axis), the performance of method m (vertical axis) is measured by the mean divergence D_m, defined as

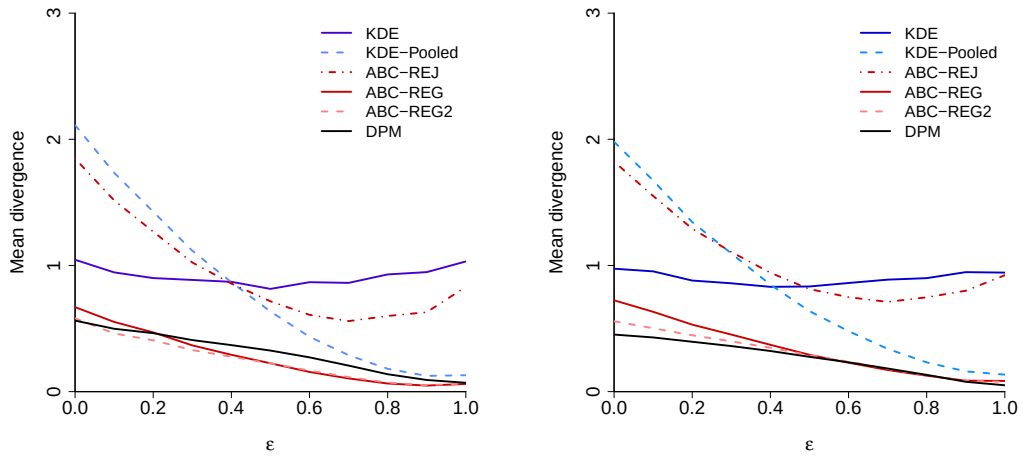
$$D_m = \frac{1}{500}\sum_{r=1}^{500}\sum_{i=1}^{g}\sum_{j=1}^{k}\Big(\log(f_{ir}(\psi_j)) - \log(\hat{f}_{ir}^m(\psi_j))\Big)f_{ir}(\psi_j),$$

where $f_{ir}(\psi_j)$ and $\hat{f}_{ir}^m(\psi_j)$ denote the true and the estimated (by method m) density of group i, for the r-th replicated dataset, calculated at ψ_j.

The mean divergence curve for the KDE method is nearly flat, which is in accordance with the fact that the method is not sensitive to possible association between groups. The slight concave shape reflects the greater smoothness of the true densities around $\varepsilon = 0.5$. As expected, the performance of the pooled-KDE estimator improves as the distance between groups is reduced ($\varepsilon \to 1$).

For the ABC approaches (based on sampling 50,000 synthetic datasets, and retaining the best 5,000), it is clear that the simple rejection algorithm is worse than the others, even the KDE estimators. For moderate to high levels of association ($\varepsilon > 0.5$) the two regression adjustment ABC methods are nearly indistinguishable. This is arguably the region where the hierarchical methods are more appealing. That said, when we move towards the left border of the horizontal axis, we observe an improvement arising from the inclusion of the interaction terms – the enhanced model is slightly more capable of determining the amount of shrinkage.

The DPM approach is an exact sampler that is specifically derived to estimate Gaussian



(a) Data generated from the HGP prior (b) Data generated from the DPM prior

Figure 3.5: Mean divergence over 500 replicates between the true and the estimated densities for each method, as a function of increasing (from left to right) levels of similarity among groups. Results are based on data simulated from (a) the HGP prior, and (b) the DPM prior. Blue and red lines denote the KDE estimators and ABC-based estimators respectively. The black line illustrates the DPM estimator.

mixture-based densities with no error. So from this perspective, it can be considered the gold standard for this analysis. The ABC estimator was narrowly better than the DPM estimator when the observed data was generated from the HGP prior, with the opposite result when the observed data was sampled from the DPM prior. This suggests that the slightly different performances resulted from the distinct prior structures of each model, and not from approximation errors. Viewed from this perspective, the ABC approach has performed remarkably well.

A major operational advantage of the ABC approach is that it is both generic, and highly flexible. Specifically, that the methodology remains the same regardless of modifications to the prior structure. For example, in the analysis of Section 3.4 if we wanted to use an infinite mixture of beta distributions as the model in the DPM prior but code was only available to fit an infinite mixture of normals model then the effort required to modify the sampling scheme would be considerable. In contrast, the ABC-based approach can easily vary the base density with only very minor modifications.

3.6 An analysis of high school exam performance in Brazil

Appropriate assessment of school performance is paramount to efficiently manage large educational systems. In 1998 the Brazilian government introduced the high school national

exam, Enem (*Exame Nacional do Ensino Medio*), which annually evaluates high school students in all states of the country. Enem is used by the Department of Education to set the education agenda and strategically define public policies, and by Federal universities and other educational institutes as part of their student selection criteria. According to the Organisation for Economic Co-operation and Development, students who attend schools in urban areas tend to perform at higher levels than other students (OECD, 2013). They conclude that beyond socio-economic reasons, schools in urban areas are generally larger and benefit from better resources and greater autonomy.

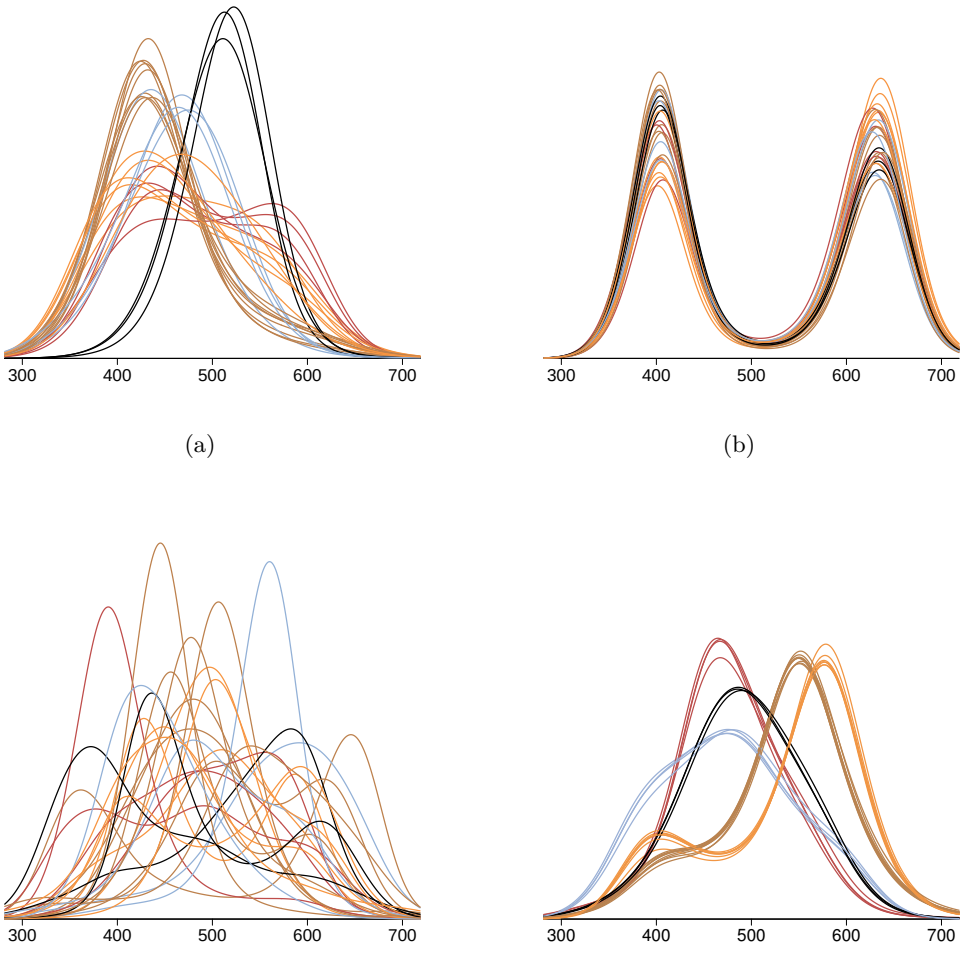


Figure 3.6: Samples from the two-level prior distribution based on the squared exponential covariance function with $g_1 = 27$ state groups and $g_2 = 5$ regional groups under varying parameter conditions. The above panels show: (a) moderate state and regional dissimilarity; (b) moderate state dissimilarity and regional similarity; (c) strong state and regional dissimilarity, and (d) strong state similarity and regional dissimilarity. Density colours indicate regional membership.

We analyse data extracted from the 2012 Enem dataset to evaluate rural school performance. The observational units are the rural schools themselves, each of which is measured by the mean grade of its students. Interest is in estimating and comparing the grade densities of each of the 27 states. Only schools with more than 10 student attending the exam are considered. The number of recorded rural schools in each state varies from 3 in Roraima, to 101 in Sao Paulo. The data are available from `http://portal.inep.gov.br/basica-levantamentos-acessar`.

The 27 states are divided into 5 regions – North, Northeast, West Central, Southeast and South. We expect that there may be some degree of similarity between states in the same region, and so a reasonable prior should allow us to express this formally. For this purpose, we can extend the prior described in (3.1a)–(3.1d) to include an additional level. More precisely,

$$\begin{aligned} f_i(x) &= \frac{L(Z_i^1(x))b(x|\phi)}{c_i(\phi, Z_i^1)} \\ Z_i^1(x) &\sim \mathcal{GP}(Z^2_{\delta(i)}(x), k(x, x'|\sigma_1, \ell_1)) \\ Z_j^2(x) &\sim \mathcal{GP}(Z^3(x), k(x, x'|\sigma_2, \ell_2)) \\ Z^3(x) &\sim \mathcal{GP}(m(x), k(x, x'|\sigma_3, \ell_3)), \end{aligned}$$

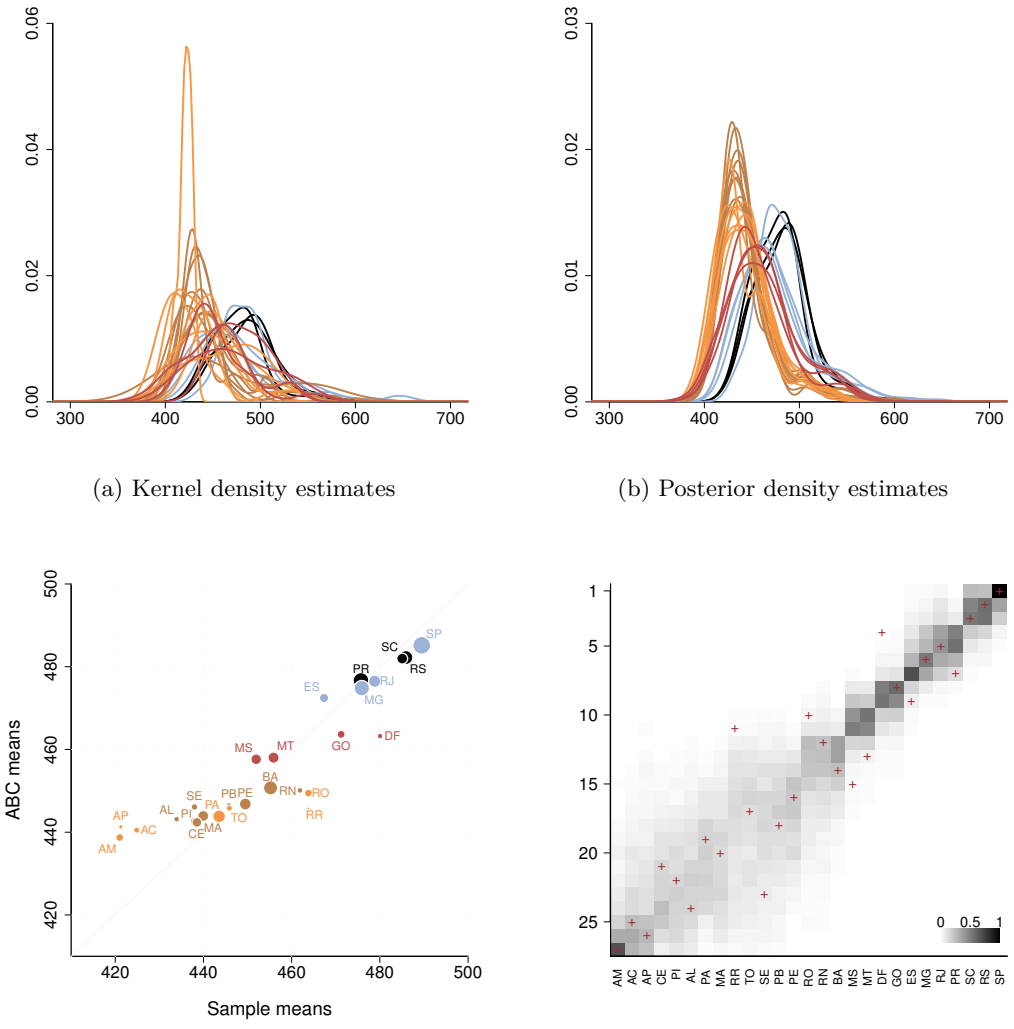
for $i = 1, \ldots, g_1 = 27$ and $j = 1, \ldots, g_2 = 5$, and where $\delta(i)$ indicates the region to which state i belongs. Figure 3.6 illustrates samples from this prior under varying parameter conditions, where the different colours indicate regional membership. Panels (a) and (d) represent the case where there are differences between regions, but where the states within each regions are fairly similar. This flexibility would be effectively impossible without an extra level of prior hierarchy.

To handle this extended prior, we modify the functional regression model (3.4) appropriately by including a regional level predictor. Specifically, the functional regression model for state i, located in region j, is given by

$$\begin{aligned} \tilde{Z}_{il}(x) = \text{offset} + \gamma_0^i(x) + \gamma_1^i(x)\log(\tilde{K}_{il}(x)) + \\ \gamma_2^i(x)\log(\tilde{K}_l^{R_j}(x)) + \gamma_3^i(x)\log(\tilde{K}_l^p(x)) + \epsilon^l(x), \end{aligned}$$

for $l = 1, \ldots, m$, where $\log(\tilde{K}_l^{R_j}(x))$ and $\log(\tilde{K}_l^p(x))$ denote the KDE for region j and the

KDE for the entire country, respectively. The functional regressors $\gamma_1^i(x), \gamma_2^i(x)$ and $\gamma_3^i(x)$ correspond to the Gaussian processes in the three levels of the hierarchy: state, region and country. The functional regression-adjustment is then performed by modifying (3.5) in the obvious way.



(a) Kernel density estimates

(b) Posterior density estimates

(c) Sample and ABC means

(d) Posterior ranking

Figure 3.7: (a) Independent kernel density estimates for each of the $g_1 = 27$ states in the Brazilian Enem analysis. Line colour indicates region membership. (b) ABC posterior mean density estimates for each state. (c) Sample means versus posterior means for each state. Circle area is proportional to the number of observations in each state. (d) Posterior rank of each state according to the mean of the posterior density estimates. Crosses indicate the rank of each state according to the sample mean (x-axis in panel (c)).

In the following, we take the base density $b(x|\phi)$ to be $N(500, 5000)$, roughly representing our prior knowledge about the location and variability of the ungrouped data, and set the prior for the parameters of the covariance functions of the Gaussian processes to be $\sigma_h \sim \text{Gamma}(3, 5)$ and $\ell_h \sim \text{Gamma}(5, 0.05)$, for $h = 1, 2, 3$. We use $k = 200$ grid points

and 150 B-spline basis functions. We generate $10,000$ samples from the prior distribution, and accept the $m = 1,000$ samples closest to the observed data.

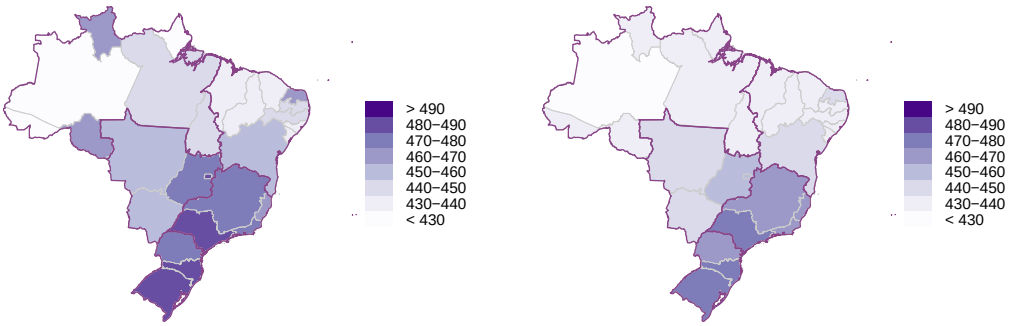
Figure 3.7(a) illustrates independent kernel density estimates of the school's mean grades for each of the 27 states, with line colour indicating membership in the 5 regions. While there is some inter-state variability, there appear to be some visible similarities of density estimates for states in the same region. These similarities become clearer in Figure 3.7(b), which shows the regression-adjusted posterior mean density estimates (notice the different plot scale). These densities indicate a strong regional clustering, and some obvious sharing of information between states within each region.

Similarly to the example in Section 3.4, an analysis of the fitted regression coefficients (not shown) indicates that the regressors have contrasting effects on each state. For example, for the heavily populated state of Sao Paulo (SP), the posterior samples are mostly determined by the kernel density estimate for this state. Alternatively, for small states in the Northeast region, such as Alagoas (AL) and Paraiba (PB), the posterior samples are predominantly affected by the average of the kernel estimates of the states within this region. Finally, for the Federal District (DF) and Acre (AC), although not the dominant regression term, the overall country kernel estimate plays a substantial role. It is clear that the model is capable of borrowing strength with the appropriate intensity from each source for each group.

Figure 3.7(c) plots the sample mean for each state against the mean of the mean posterior density estimate, $\mathbb{E}[f_i(x)|\mathcal{D}]$. Here the area of each point is proportional to the sample size within each state. An obvious shrinkage effect can be observed, particularly in states with fewer schools. In addition, the evident regional clusters (indicated by colouring) suggest that geographical forces critically affect the system.

While not making full use of the posterior density estimates, a simple way to construct a state ranking system could be via the means of these posterior densities. Figure 3.7(d) illustrates the posterior distribution of the rank of each state, according to the mean of the posterior density estimate, with a darker shade indicating greater posterior mass. The crosses mark each state's rank according to the independent sample means (x-coordinate values in Figure 3.7(c)). From the plot, we see that Sao Paulo (SP) is most likely in the top 4 states (probability > 0.99), and that the Federal District (DF) is likely ranked too high according to the independent sample means, and Mato Grosso do Sul (MS) too low.

The posterior rank uncertainty is large for states in the lower part of the ranking. For example, while Amapa (AP) and Roraima (RR) have substantially different sample means (421.27 and 463.76) the posterior probability of the mean of Amapa being greater than the mean of Roraima is still considerable at 0.20.



(a) Sample means (b) ABC posterior means

Figure 3.8: Geographical maps for comparing (a) the sample and (b) posterior density means.

Finally, Figure 3.8 illustrates the spatial distribution of the sample means (panel (a)) and the posterior means (panel (b)). The effect of the hierarchical model is apparent in panel (b), as there is clear evidence of smoothing both within and between regions. There is an obvious north-south effect in the mean performance of schools in the Enem examination. Note that while comparison are made here in terms of means, any other feature could equally be considered, including the densities variability and shape.

3.7 Discussion

The hierarchical Gaussian process prior introduced in this paper permits, in a very simple way, the characterisation of prior beliefs about a set of univariate densities. The use of ABC methods allows the sampling of approximate draws from the posterior distribution at a moderate computational cost. This cost is almost certainly smaller than could be realistically obtained by developing a direct MCMC sampler along the lines of, say, Adams et al. (2009). In their implementation, the samples are generated sequentially, following an algorithm that realises the Gaussian process at certain random places where the function needs to be known in order to run their rejection sampler. The drawback of this approach is

that as the number of proposed samples increases, so does the dimension of the associated covariance matrix. Repeatedly inverting this matrix throughout the MCMC iterations is computationally demanding, not only due to its magnitude but also because of the high correlation induced by the squared exponential kernel $k(x, x')$. Therefore, while it is not strictly necessary to use ABC for fitting the models under consideration, our approximate strategy naturally avoids some of the technical hurdles observed in such an exact algorithm.

Nevertheless, ABC is a computationally expensive simulation method – the simulations in Section 3.5 took ≈ 1 second per accepted sample (39% for generating the synthetic densities through the Gaussian processes, 53% for sampling the synthetic data, 8% for computing summary statistics), with the functional regression adjustment for 5,000 accepted samples taking ≈ 80 seconds. In comparison, the `DPpackage`, which is internally written in Fortran, takes $\approx$100 seconds to draw 5,000 samples from the posterior distribution (without burn-in or thinning). However, when multiple datasets are to be analysed (as in Section 3.5), we only need to sample from the prior once and the ABC constructions become much faster.

One advantage of the approach presented here over hierarchical Dirichlet process models is that the parameters in our hierarchical prior are easily interpretable, and allow very direct control over the features of the prior. With Dirichlet process based methods, which build elaborate hierarchies involving mixing distributions, it can be challenging to be expressive of prior information. In addition, in our framework changes to the prior specification require minimal code updating. Except for generating data from the prior (first step of the ABC rejection algorithm), the same code is suitable for different base densities (betas, Normals, Gammas, etc.) and hyperparameter distributions – we are not restricted to probabilistically convenient (e.g. conjugate) priors.

While we only consider univariate densities in this paper, in principle our ABC method can be extended to the multivariate case by developing a multivariate functional regression adjustment. However, the computational cost would rapidly grow with the number of model dimensions (and grid points), likely making the procedure impractical for more than low dimensional models. A further extension could involve the use of variational Bayes methods, which are a collection of fast approximate inference approaches. Variational Bayes methods have had great success in fitting Gaussian process models, and have recently been considered as a way of implementing ABC techniques themselves (Tran et al., 2017).

In the present context they could additionally be used as a way of approximating the intractable normalising constants $c_i(\phi, Z_i)$.

In this work, regular grids were employed to discretise functional objects, with the number of grid points and B-splines basis functions being determined somewhat informally, based on graphical assessments – the error prompted by the use of B-splines to approximate known functions can be easily spotted with a simple comparative plot. More sophisticated approaches could also be considered. One popular strategy, for example, is to place a knot at every fixed number of observations. Even better would be to extend (in future work) the inferential process to allow the method to automatically learn the optimal configuration from the observed data, as in Tokdar and Ghosh (2007) and Fan et al. (2010).

The construction of a nonparametric extension to the standard ABC regression-adjustment is a novel contribution to the ABC literature. However, ABC is an approximate inferential procedure which heavily relies on the information content of its summary statistics and in the accuracy of the regression-adjustment model. Fortunately, there is a range of credible candidates for summarizing the data – kernel density estimation methods have been extensively developed and investigated by numerous authors. Further, regression-adjustment models are well understood in the ABC literature as a method to improve the conditional density estimation of the posterior distribution (Blum and François, 2010).

Both the simulation study and the analysis of high school exam performance in Brazil highlights the capability of the ABC procedure to borrow strength across multiple groups, to appropriately produce shrinkage where needed (with small sample sizes), and to suitably handle the boundary-bias problem of the kernel density estimator summary statistic. Ultimately, the hierarchical Gaussian process density model presents an enhanced approach over existing methodologies, and provides a valuable tool in the practitioners toolbox.

Chapter 4

Recalibration: A post-processing method for approximate Bayesian computation [1]

4.1 Introduction

Approximate Bayesian Computation (ABC) refers to a class of algorithms designed to sample from an approximation to the posterior distribution without directly evaluating the likelihood function. These techniques have expanded the reach of statistical inference to a range of problems where the likelihood function is computationally intractable, in that it is prohibitively expensive or even impossible to evaluate. Instead, inference is based on the ability to simulate data from the model of interest (e.g. Beaumont et al., 2002; Fearnhead and Prangle, 2012; Sisson et al., 2017b).

Consider the usual Bayesian setting with a parameter vector $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_d)^\top$, a prior $\pi(\boldsymbol{\theta})$, and a model for data $\boldsymbol{y}$, $p(\boldsymbol{y}|\boldsymbol{\theta})$. Let $\boldsymbol{y}_{\text{obs}}$ denote the observed data. In its simplest implementation, ABC repeatedly executes two steps: sampling $(\boldsymbol{\theta}, \boldsymbol{y})$ from the (prior predictive) generative process $\pi(\boldsymbol{\theta})p(\boldsymbol{y}|\boldsymbol{\theta})$, and accepting $\boldsymbol{\theta}$ if $\boldsymbol{y} \approx \boldsymbol{y}_{\text{obs}}$ according to some distance measure. This second step is commonly implemented in an importance sampling framework whereby a weight $w(\boldsymbol{\theta})$ is attached to $\boldsymbol{\theta}$ of the form $w(\boldsymbol{\theta}) \propto K_h(\|\boldsymbol{s} - \boldsymbol{s}_{\text{obs}}\|)$, where $\boldsymbol{s} = S(\boldsymbol{y})$ maps $\boldsymbol{y}$ to a low dimensional vector of summary statistics, $\boldsymbol{s}_{\text{obs}} =$

[1] Submitted for publication as: Rodrigues, G. S., Prangle, D., Sisson, S. A. (2017), "Recalibration: A post-processing method for approximate Bayesian computation."

$S(\boldsymbol{y}_{\text{obs}})$, and K_h is a smoothing kernel with scale parameter $h \geq 0$. The idealised algorithm where only exact matches $\boldsymbol{y} = \boldsymbol{y}_{\text{obs}}$ are accepted ($h = 0$) would produce samples from the exact posterior $\pi(\boldsymbol{\theta}|\boldsymbol{y})$ (or more generally the partial posterior $\pi(\boldsymbol{\theta}|\boldsymbol{s}_{\text{obs}})$, if matching $\boldsymbol{s} = \boldsymbol{s}_{\text{obs}}$). In practice, approximate matches based on weights $w(\boldsymbol{\theta})$ are retained to sidestep the impossibility of exactly matching simulated and observed data in all but the simplest settings. However this necessity accordingly introduces an approximation error to the ABC posterior approximation. In general, the ABC posterior approximation can be expressed as

$$\pi_{ABC}(\boldsymbol{\theta}|\boldsymbol{s}_{\text{obs}}) = \int K_h(\|S(\boldsymbol{y}) - \boldsymbol{s}_{\text{obs}}\|)p(\boldsymbol{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{y}. \tag{4.1}$$

See e.g. Sisson et al. (2017b) for further details.

A number of post-processing techniques have been proposed to correct this approximation error once samples from the ABC posterior approximation have been obtained, resulting in an estimate $\hat{\pi}_{ABC}(\boldsymbol{\theta}|\boldsymbol{s}_{\text{obs}})$ which better approximates the true (partial) posterior $\pi(\boldsymbol{\theta}|\boldsymbol{s}_{\text{obs}})$ than (4.1). Beaumont et al. (2002) introduced a regression-adjustment approach, in which the ABC samples are corrected with the aid of a local linear regression model for $\boldsymbol{\theta}|\boldsymbol{s} - \boldsymbol{s}_{\text{obs}}$, fitted to the $(\boldsymbol{\theta}, \boldsymbol{s})$ samples from (4.1). Various extensions to this technique include non-linear, heteroscedastic regression (Blum and François, 2010), and ridge regression adjustments (Blum et al., 2013). However, there is some evidence emerging to suggest that regression-adjustments tend to overcorrect and produce approximate posteriors that are too precise, leading to nominal credible intervals with coverage much higher than should occur under $\pi(\boldsymbol{\theta}|\boldsymbol{s}_{\text{obs}})$ (Marin et al., 2016; Frazier et al., 2017). From the perspective of marginal density estimation, Nott et al. (2014) (see also Li et al., 2017) developed a marginal-adjustment which replaces low-dimensional marginal distributions of (4.1) by more accurate marginal distributions estimated using smaller numbers of summary statistics than in $\boldsymbol{s}$. This exploits the fact that ABC methods are known to perform poorly for larger numbers of summary statistics due to the curse of dimensionality in the comparison $\|\boldsymbol{s} - \boldsymbol{s}_{\text{obs}}\|$, however this approach requires the identification of subsets of summary statistics that are informative for each margin, which may not be easily available.

In this paper we introduce a novel *recalibration* post-processing method for improving

the accuracy of the ABC posterior approximation that avoids the problems of existing post-processing techniques. It is based on the ideas in Prangle et al. (2014), who derive a diagnostic tool for ABC based on the so-called *coverage property* (Cook et al., 2006; Fearnhead and Prangle, 2012; Prangle et al., 2014), which tests whether for a given $h > 0$ the estimated marginals of $\pi_{ABC}(\boldsymbol{\theta}|\boldsymbol{s}_{\text{obs}})$ (or $\hat{\pi}_{ABC}(\boldsymbol{\theta}|\boldsymbol{s}_{\text{obs}})$) are well "calibrated". Calibration requires that estimated credible intervals have the correct probabilities of containing the true parameter values. If calibration does not hold, Prangle et al. (2014) suggest reducing h until it does hold. However, this is not always feasible, particularly as reducing h increases the Monte Carlo error of the Monte Carlo sample approximation of (4.1) for a fixed computational budget.

Our approach extends the ideas in Prangle et al. (2014) to develop a post-processing recalibration adjustment that aims to produce an approximation $\hat{\pi}_{ABC}(\boldsymbol{\theta}|\boldsymbol{s}_{\text{obs}})$ that is well calibrated. Our method achieves this approximately and, as a result, the coverage problems associated with the regression adjustment (Marin et al., 2016) can be mitigated by construction. Recalibration can be applied directly to samples from $\pi_{ABC}(\boldsymbol{\theta}|\boldsymbol{s}_{\text{obs}})$, or to improve the output from other post-processing adjustments. Recalibration is related to indirect inference – a technique in which inference is performed with the aid of an auxiliary misspecified model (Gourieroux et al., 1993). The use of indirect inference in the ABC framework has been previously explored by Drovandi et al. (2015, 2017).

Our strategy also relates to several procedures that attempt to correct the bias in an initial interval estimate based on Monte Carlo simulation. Menéndez et al. (2014), for instance, propose drawing pseudo-samples $\boldsymbol{y}^{(1)}, \ldots, \boldsymbol{y}^{(N)} \sim p(\boldsymbol{y}|\tilde{\theta})$, from which synthetic credibility (or confidence) intervals are obtained. These are then used to adjust the original interval, calculated over the observed dataset. In their procedure, it is assumed that a consistent estimator of θ, $\tilde{\theta}$, is available, which is not often the case, especially when the likelihood is intractable. Moreover, the correction is carried out for an interval defined by a single, pre-specified confidence level α. In our approach, in contrast, the whole multivariate posterior density function is simultaneously corrected. Taking into account the correlation structure of the parameter of interest is another important advantage of our recalibration technique, considering that, in Menéndez et al. (2014), the correction is performed independently for each margin of $\boldsymbol{\theta}$.

In the frequentist framework, several bootstrap-based methods have been proposed

to improve the coverage accuracy. In an influential paper, Hall (1986) introduced the general class of *iterative bootstrap* techniques, which, according to the author, involves 'simulations of simulations'. Further theoretical justifications for its use were given in Hall and Martin (1988) and Martin (1990), although in its original formulation these methods were highly computationally intensive. Several solutions were given to address this overwhelming computational burden, including the works of Lee and Young (1995); Davidson and MacKinnon (2002, 2007).

Efron (1987), in turn, introduced parametric and non-parametric bootstrap methods for automatically lessening the coverage error, providing a convenient way of dealing with potentially misleadingly maximum likelihood-based confidence intervals. Beran (1987) contributed to the growing toolbox of resempling methods with the introduction of the useful 'prepivoting' procedure, which was later shown to be become even more efficient when using weighted bootstrap iterations (Lee and Young, 2003). All of these frequentist algorithms were not designed to improve the accuracy of an approximate posterior distribution, and therefore do not offer genuine alternatives to the solution proposed hereafter.

We introduce our recalibration approach in Section 4.2. We demonstrate its performance in two simulation studies in Section 4.3, using a Gaussian auxiliary posterior estimator for inference on a sum of lognormals distribution, and a standard ABC analysis of a "twisted normal" model. Section 4.4 revisits the analysis of Erhardt and Sisson (2016) in a real stereological extremes problem and shows that the recalibration adjustment can correct the bias of their regression-adjustment ABC implementation. We conclude with a discussion of the merits and limitations of recalibration in Section 4.5, including the possibility of correcting approximate Bayesian inference methods beyond ABC.

4.2 Recalibration

4.2.1 Motivation

Our recalibration post-processing procedure is based on the *coverage* property. An $\alpha\%$ credible region for a parameter $\boldsymbol{\theta}$ is a region R with the property that $\Pr(\boldsymbol{\theta} \in R|\boldsymbol{y}_{\text{obs}}) = \alpha/100$. Loosely, the coverage property asserts that for data $\boldsymbol{y}_0$ generated under the model for a known parameter value $\boldsymbol{\theta}_0 = (\theta_{0,1}, \ldots, \theta_{0,d})^\top$, so that $\boldsymbol{y}_0 \sim p(\boldsymbol{y}|\boldsymbol{\theta}_0)$, credible intervals constructed from the posterior $\pi(\theta|\boldsymbol{y}_0)$ will have the claimed probability of containing $\boldsymbol{\theta}_0$.

Coverage has been previously examined in the ABC literature. Most commonly it has been used to validate analyses (e.g. Wegmann et al., 2009, 2010; Aeschbacher et al., 2012), with Prangle et al. (2014) extending coverage ideas to develop testable diagnostics to determine whether the marginals of $\pi_{ABC}(\boldsymbol{\theta}|\boldsymbol{s}_{\text{obs}})$ are different to those of $\pi(\boldsymbol{\theta}|\boldsymbol{s}_{\text{obs}})$, and similarly whether estimated model probabilities under ABC are different to the true posterior model probabilities given $\boldsymbol{s}_{\text{obs}}$ in a multi-model analysis. Coverage is identified as a desirable property of ABC posterior distributions by Fearnhead and Prangle (2012), who also introduce 'noisy ABC' which automatically satisfies the coverage property, and Menéndez et al. (2014) use related ideas to correct bias in ABC credible intervals. Finally, the failure of regression adjustment techniques to produce ABC approximations $\hat{\pi}(\theta|\boldsymbol{s}_{\text{obs}})$ that satisfy the coverage property, is being used as evidence that they are producing poor approximations (Marin et al., 2016; Frazier et al., 2017).

Our recalibration adjustment is closely linked to the diagnostic techniques of Prangle et al. (2014). Let $F_{\boldsymbol{s}}(\boldsymbol{\theta})$ be the distribution function of $\pi(\boldsymbol{\theta}|\boldsymbol{s})$, the partial posterior for $\boldsymbol{\theta}$ given some summary dataset $\boldsymbol{s}$, and $F_{j,\boldsymbol{s}}(\theta_j)$ be the j-th associated marginal distribution function, for $j = 1, \ldots, d$. Our interest is sampling from $F_{\boldsymbol{s}_{\text{obs}}}(\boldsymbol{\theta})$, the partial posterior distribution given the observed data summary $\boldsymbol{s}_{\text{obs}}$.

For some choice of parameter $\boldsymbol{\theta}_0$, and generated dataset $\boldsymbol{s}_0 = S(\boldsymbol{y}_0)$ with $\boldsymbol{y}_0 \sim p(\boldsymbol{y}|\boldsymbol{\theta}_0)$, Prangle et al. (2014) demonstrated that the location of the j-th marginal parameter $\theta_{0,j}$ in the j-th marginal posterior distribution of $\pi(\boldsymbol{\theta}|\boldsymbol{s}_0)$, as measured by $p_j = F_{j,\boldsymbol{s}_0}(\theta_{0,j}) := \Pr(\theta_j < \theta_{0,j}|\boldsymbol{s}_0)$ will give $p_j \sim U(0,1)$ for $j = 1, \ldots, d$. This then allows for the basis of a test for whether $\tilde{F}_{j,\boldsymbol{s}_{\text{obs}}}(\theta_j)$, the j-th marginal distribution function of the ABC posterior approximation $\pi_{ABC}(\boldsymbol{\theta}|\boldsymbol{s}_{\text{obs}})$, is the same as the true marginal distribution function, i.e. whether $\tilde{F}_{j,\boldsymbol{s}_{\text{obs}}}(\theta_j) = F_{j,\boldsymbol{s}_{\text{obs}}}(\theta_j)$.

This test proceeds by generating $(\boldsymbol{\theta}^{(i)}, s^{(i)})$ pairs, $i = 1, \ldots, N$, from $\boldsymbol{\theta}^{(i)} \sim \pi(\boldsymbol{\theta})$ (or other suitable distribution) and $\boldsymbol{s}^{(i)} = S(\boldsymbol{y}^{(i)})$, $\boldsymbol{y}^{(i)} \sim p(\boldsymbol{y}|\boldsymbol{\theta}^{(i)})$, and constructing the ABC posterior approximation $\pi_{ABC}(\boldsymbol{\theta}|\boldsymbol{s}^{(i)})$ for each $\boldsymbol{s}^{(i)} \in A(\boldsymbol{s}_{\text{obs}})$, where $A(\boldsymbol{s}_{\text{obs}})$ is some set centred around $\boldsymbol{s}_{\text{obs}}$. Then, for each $\boldsymbol{s}^{(i)} \in A(\boldsymbol{s}_{\text{obs}})$, the statistics $p_j^{(1)}, \ldots, p_j^{(N)}$, where $p_j^{(i)} = \tilde{F}_{j,\boldsymbol{s}^{(i)}}(\theta_j^{(i)})$, will only be distributed as $U(0,1)$ if $\tilde{F}_{j,\boldsymbol{s}^{(i)}}(\theta_j) = F_{j,\boldsymbol{s}^{(i)}}(\theta_j)$, which can be determined via standard tests of uniformity for each margin $j = 1, \ldots, d$. If this test is satisfied, then it can be inferred that the marginal distributions of $\tilde{F}_{\boldsymbol{s}_{\text{obs}}}(\boldsymbol{\theta})$ are approximately those of $F_{\boldsymbol{s}_{\text{obs}}}(\boldsymbol{\theta})$ and that, marginally at least, the ABC posterior

approximation $\pi_{ABC}(\boldsymbol{\theta}|\boldsymbol{s}_{\text{obs}})$ is a good approximation of $\pi(\boldsymbol{\theta}|\boldsymbol{s}_{\text{obs}})$. (Note that in practice, $\pi_{ABC}(\boldsymbol{\theta}|\boldsymbol{s})$ and $\tilde{F}_{j,\boldsymbol{s}}$ are constructed from weighted samples.)

We now extend this idea. However, rather than merely testing whether there are significant marginal deviations between $\tilde{F}_{\boldsymbol{s}_{\text{obs}}}(\boldsymbol{\theta})$ and $F_{\boldsymbol{s}_{\text{obs}}}(\boldsymbol{\theta})$, we use the measured differences to adjust those samples $\boldsymbol{\theta}$ from $\pi_{ABC}(\boldsymbol{\theta}|\boldsymbol{s}_{\text{obs}})$ so that $\hat{\tilde{F}}_{j,\boldsymbol{s}_{\text{obs}}}(\boldsymbol{\theta}) \approx F_{j,\boldsymbol{s}_{\text{obs}}}(\boldsymbol{\theta})$ is a good approximation (where $\hat{\tilde{F}}_{j,\boldsymbol{s}_{\text{obs}}}(\boldsymbol{\theta})$ is the j-th marginal distribution function of the adjusted samples). That is, that the resulting post-processed approximation $\hat{\pi}(\boldsymbol{\theta}|\boldsymbol{s}_{\text{obs}})$, approximately satisfies the coverage property, and is accordingly approximately well calibrated.

4.2.2 Method

So far we have assumed that $\tilde{F}_{j,\boldsymbol{s}}(\theta_j)$, the j-th marginal distribution of $\tilde{F}_{\boldsymbol{s}}(\boldsymbol{\theta})$, is the j-th marginal distribution function of the ABC posterior approximation $\pi_{ABC}(\boldsymbol{\theta}|\boldsymbol{s})$. However, all that is required to implement the recalibration adjustment is that some approximate method for inferring the posterior marginal distribution functions is available. Such approximate methods arise from adopting auxiliary models which approximate $\pi(\boldsymbol{\theta}|\boldsymbol{s})$ with different posterior forms, such as those obtained under the Bayesian indirect inference framework (Drovandi et al., 2017, 2015), variational Bayes (Tran et al., 2017), regression density estimation (Fan et al., 2013) and expectation-propagation (exponential family) based approximations (Barthelmé and Chopin, 2014). We now suppose that $\tilde{F}_{\boldsymbol{s}}(\boldsymbol{\theta})$ and the associated marginal distribution functions $\tilde{F}_{j,\boldsymbol{s}}(\theta_j)$, $j = 1, \ldots, d$, are available as approximations to $F_{\boldsymbol{s}}(\boldsymbol{\theta})$ and $F_{j,\boldsymbol{s}}(\theta_j)$, based on some auxiliary model, which may include the standard ABC posterior approximation $\pi_{ABC}(\boldsymbol{\theta}|s)$. Note that the recalibration adjustment will only make use of the marginal distribution functions $\tilde{F}_{j,\boldsymbol{s}}(\theta_j)$, and not the joint distribution function $\tilde{F}_{\boldsymbol{s}}(\boldsymbol{\theta})$, and that these approximate marginal distribution functions are assumed to have a well defined inverse, $\tilde{F}^{-1}_{j,\boldsymbol{s}}(\cdot)$.

In order to state the recalibration adjustment, first define

$$G_{\boldsymbol{s}}(\boldsymbol{p}) = F_{\boldsymbol{s}}[(\tilde{F}^{-1}_{1,\boldsymbol{s}}(p_1), \ldots, \tilde{F}^{-1}_{d,\boldsymbol{s}}(p_d))^\top]$$

where $\boldsymbol{p} = (p_1, p_2, \ldots, p_d)^\top \in [0,1]^d$ (where d is the number of parameters). The function $G_{\boldsymbol{s}}(\boldsymbol{p})$ incorporates the posterior dependence structure of $\pi(\boldsymbol{\theta}|\boldsymbol{s})$, through $F_{\boldsymbol{s}}(\cdot)$, but it also provides a connection between the true (through $F_{\boldsymbol{s}}(\boldsymbol{\theta})$) and the estimated marginal

posterior quantile functions $\tilde{F}^{-1}_{j,\boldsymbol{s}}(p_j)$. We now provide several simple results on $G_{\boldsymbol{s}}(\boldsymbol{p})$ which will be useful to establish the recalibration adjustment.

Result 1 Suppose a random variable $\boldsymbol{P} = (P_1, \ldots, P_d)^\top$ has distribution $G_{\boldsymbol{s}}(\boldsymbol{p})$. Then $P_j|\boldsymbol{s} \sim U(0,1)$ for $j = 1, \ldots, d$, if and only if the estimated marginal posteriors $\tilde{F}_{j,\boldsymbol{s}}(\cdot)$ equal the true marginal posteriors $F_{j,\boldsymbol{s}}(\cdot)$.

Proof. First suppose that $\tilde{F}_{j,\boldsymbol{s}}(\cdot) = F_{j,\boldsymbol{s}}(\cdot)$. Then the j-th marginal distribution function of $G_{\boldsymbol{s}}(\boldsymbol{p})$ is $F_{j,\boldsymbol{s}}[\tilde{F}^{-1}_{j,\boldsymbol{s}}(p_j)] = p_j$, which is a $U(0,1)$ distribution. Next suppose that the j-th marginal distribution of $G_{\boldsymbol{s}}(\boldsymbol{p})$ is a $U(0,1)$ distribution. Then $F_{j,\boldsymbol{s}}[\tilde{F}^{-1}_{j,\boldsymbol{s}}(p_j)] = p_j$. Let $q_j = \tilde{F}^{-1}_{j,\boldsymbol{s}}(p_j)$. Then we have $F_{j,\boldsymbol{s}}(q_j) = \tilde{F}_{j,\boldsymbol{s}}(q_j)$ as required. □

Result 1 states that $\boldsymbol{P} \sim G_{\boldsymbol{s}}(\boldsymbol{p})$ is *marginally* uniform if and only if $\tilde{F}_{j,\boldsymbol{s}}(\cdot) = F_{j,\boldsymbol{s}}(\cdot)$, for $j = 1, \ldots, d$, but does not comment on its dependence structure. Prangle et al. (2014) exploited a variant of this result to test whether the marginal distributions of $\pi_{ABC}(\boldsymbol{\theta}|\boldsymbol{s}_{\text{obs}})$ were equal to those of $\pi(\boldsymbol{\theta}|\boldsymbol{s}_{\text{obs}})$ by testing for uniformity of realised P_i values, as described in Section 4.2.1.

Result 2 Suppose that the random variable $\boldsymbol{P} = (P_1, \ldots, P_d)^\top$ has distribution function $G_{\boldsymbol{s}}(\boldsymbol{p})$. Then conditional on $\boldsymbol{s}$, $(\tilde{F}^{-1}_{1,\boldsymbol{s}}(P_1), \ldots, \tilde{F}^{-1}_{d,\boldsymbol{s}}(P_d))^\top$ has distribution $F_{\boldsymbol{s}}(\boldsymbol{\theta})$.

Proof.

$$\begin{aligned}
& \Pr(P_1 \leq p_1, \ldots, P_d \leq p_d|\boldsymbol{s}) = F_{\boldsymbol{s}}[(\tilde{F}^{-1}_{1,\boldsymbol{s}}(p_1), \ldots, \tilde{F}^{-1}_{d,\boldsymbol{s}}(p_d))^\top] \\
\Rightarrow \quad & \Pr(P_1 \leq \tilde{F}_{1,\boldsymbol{s}}(\theta_1), \ldots, P_d \leq \tilde{F}_{d,\boldsymbol{s}}(\theta_d)|\boldsymbol{s}) = F_{\boldsymbol{s}}((\theta_1, \ldots, \theta_d)^\top) \\
\Rightarrow \quad & \Pr(\tilde{F}^{-1}_{1,\boldsymbol{s}}(P_1) \leq \theta_1, \ldots, \tilde{F}^{-1}_{d,\boldsymbol{s}}(P_d) \leq \theta_d|\boldsymbol{s}) = F_{\boldsymbol{s}}(\boldsymbol{\theta})
\end{aligned}$$

as required. □

Result 2 provides a straightforward way to use an observation from $G_{\boldsymbol{s}}(\boldsymbol{p})$ to generate a sample from $F_{\boldsymbol{s}}(\boldsymbol{\theta})$. Result 3 below provides the converse – a way to use an observation from $F_{\boldsymbol{s}}(\boldsymbol{\theta})$ to generate a sample from $G_{\boldsymbol{s}}(\boldsymbol{p})$.

Result 3 Suppose that the random variable $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d)^\top$ has distribution function $F_{\boldsymbol{s}}(\boldsymbol{\theta})$. Then conditional on $\boldsymbol{s}$, $(\tilde{F}_{1,\boldsymbol{s}}(\theta_1), \ldots, \tilde{F}_{d,\boldsymbol{s}}(\theta_d))^\top$ has distribution $G_{\boldsymbol{s}}(\boldsymbol{p})$.

Proof.

$$\begin{aligned}\Pr(\tilde{F}_{1,\boldsymbol{s}}(\theta_1) \leq p_1, \ldots, \tilde{F}_{d,\boldsymbol{s}}(\theta_d) \leq p_d|\boldsymbol{s}) &= \Pr(\theta_1 \leq \tilde{F}^{-1}_{1,\boldsymbol{s}}(p_1), \ldots, \theta_d \leq \tilde{F}^{-1}_{d,\boldsymbol{s}}(p_d)|\boldsymbol{s}) \\ &= F_{\boldsymbol{s}}[(\tilde{F}^{-1}_{1,\boldsymbol{s}}(p_1), \ldots, \tilde{F}^{-1}_{d,\boldsymbol{s}}(p_d))^\top]\end{aligned}$$

as required. □

These results may be combined in a procedure to recalibrate the ABC posterior approximation. For simplicity of presentation, we first focus on the recalibration of samples drawn from $\pi_{ABC}(\boldsymbol{\theta}|\boldsymbol{s}_{\text{obs}})$ (or $\hat{\pi}_{ABC}(\boldsymbol{\theta}|\boldsymbol{s}_{\text{obs}})$) under the standard ABC implementation. Following this, in Section 4.2.3 we describe how recalibration can also be implemented using an auxiliary estimator.

A standard ABC posterior simulation algorithm, complete with the recalibration procedure, is outlined in Algorithm 3. More sophisticated versions of ABC algorithms could be used. In Algorithm 3, simulation from $\pi_{ABC}(\boldsymbol{\theta}|\boldsymbol{s}_{\text{obs}})$ begins by drawing N parameter and summary statistic pairs $\{(\boldsymbol{\theta}^{(i)}, \boldsymbol{s}^{(i)})\}_{i=1}^N$ from $\boldsymbol{\theta}^{(i)} \sim \pi(\boldsymbol{\theta})$ and $\boldsymbol{s}^{(i)} = S(\boldsymbol{y}^{(i)})$ where $\boldsymbol{y}^{(i)} \sim p(\boldsymbol{y}|\boldsymbol{\theta}^{(i)})$. These samples are then used to approximate $\pi(\boldsymbol{\theta}|\boldsymbol{s}_{\text{obs}})$ by weighting them by $w^{(i)} \propto K_h(\|\boldsymbol{s}^{(i)} - \boldsymbol{s}_{\text{obs}}\|)$. From this posterior approximation, the marginal distribution functions $\tilde{F}_{j,\boldsymbol{s}_{\text{obs}}}(\theta_j)$ based on $\boldsymbol{s}_{\text{obs}}$ can be constructed by e.g. the empirical cdf or by smoothed versions of such.

For each of these samples $\boldsymbol{\theta}^{(i)}|w^{(i)} > 0$ used, an individual recalibration adjustment is performed. Firstly, samples are first drawn from the ABC posterior $\pi_{ABC}(\boldsymbol{\theta}|\boldsymbol{s}^{(i)})$ in the same manner as for those drawn from $\pi_{ABC}(\boldsymbol{\theta}|\boldsymbol{s}_{\text{obs}})$. It is possible to avoid the cost of performing a full ABC analysis by reusing the simulations from steps 1.1–1.3 of Algorithm 3, as is relatively common for ABC algorithms (Blum et al., 2013; Prangle et al., 2014). From the samples from $\pi_{ABC}(\boldsymbol{\theta}|\boldsymbol{s}^{(i)})$, the marginal distribution functions $\tilde{F}_{j,\boldsymbol{s}^{(i)}}(\cdot)$ can be constructed, for $j = 1, \ldots, d$, and the corresponding vector $\boldsymbol{p}^{(i)} = (p_1^{(i)}, \ldots, p_d^{(i)})^\top$ obtained via $p_j^{(i)} = \tilde{F}_{j,\boldsymbol{s}^{(i)}}(\theta_j^{(i)})$. Since $\boldsymbol{\theta}^{(i)}$ is an exact draw from the posterior distribution $\pi(\boldsymbol{\theta}|\boldsymbol{s}^{(i)})$, then Result 3 states that $\boldsymbol{p}^{(i)}$ is an exact draw from $G_{\boldsymbol{s}^{(i)}}(\boldsymbol{p})$.

If the ABC method produces the exact posterior so that $\pi_{ABC}(\boldsymbol{\theta}|\boldsymbol{s}^{(i)}) = \pi(\boldsymbol{\theta}|\boldsymbol{s}^{(i)})$, then Result 1 (see also Prangle et al., 2014) states that the resulting marginal distributions of $p_j^{(i)}$ would be $U(0,1)$. Of course, this is unlikely to be the case in practice, and so the marginal distributions $\tilde{F}_{j,\boldsymbol{s}^{(i)}}$ characterise the deviations away from uniformity, such

Algorithm 3 Recalibration of ABC output

Inputs:

- An observed dataset $\boldsymbol{y}_{\text{obs}}$.
- A prior $\pi(\boldsymbol{\theta})$ and intractable generative model $p(\boldsymbol{y}|\boldsymbol{\theta})$, with $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d)^\top$.
- An observed vector of summary statistics $\boldsymbol{s}_{\text{obs}} = S(\boldsymbol{y}_{\text{obs}})$.
- A smoothing kernel $K_h(u)$ with scale parameter $h > 0$.
- A positive integer N defining the number of ABC samples.

Data simulation and weighting:
For $i = 1, \ldots, N$:

1.1 Generate $\boldsymbol{\theta}^{(i)} \sim \pi(\boldsymbol{\theta})$ from the prior.
1.2 Generate $\boldsymbol{y}^{(i)} \sim p(\boldsymbol{y}|\boldsymbol{\theta}^{(i)})$ from the likelihood.
1.3 Compute the summary statistics $\boldsymbol{s}^{(i)} = S(\boldsymbol{y}^{(i)})$.
1.4 Compute the sample weight $w^{(i)} \propto K_h(||\boldsymbol{s}^{(i)} - \boldsymbol{s}_{\text{obs}}||)$.

Recalibration:

2.1 For $j = 1, \ldots, d$, construct $\tilde{F}_{j,\boldsymbol{s}_{\text{obs}}}(\cdot)$ based on the samples $\{(\boldsymbol{\theta}^{(i)}, w^{(i)})\}_{i=1}^N$.

For each i such that $w^{(i)} > 0$, and for $j = 1, \ldots, d$:

2.2 Construct $\tilde{F}_{j,\boldsymbol{s}^{(i)}}(\cdot)$ based on the samples $\{(\boldsymbol{\theta}^{(k)}, \boldsymbol{s}^{(k)})\}_{k=1,k\neq i}^N$ using the same procedure as in steps 1.4 and 2.1.
2.3 Set $p_j^{(i)} = \tilde{F}_{j,\boldsymbol{s}^{(i)}}(\theta_j^{(i)})$.
2.4 [Optional] Correct $p_j^{(i)}$ using a regression-adjustment (see Section 4.2.4).
2.5 Set $\hat{\theta}_j^{(i)} = \tilde{F}^{-1}_{j,\boldsymbol{s}_{\text{obs}}}(p_j^{(i)})$.

Outputs:

- Standard ABC output: a set of weighted samples $\{(\boldsymbol{\theta}^{(i)}, w^{(i)})\}_{i=1}^N$ from $\pi_{ABC}(\boldsymbol{\theta}|\boldsymbol{s}_{\text{obs}})$.
- A set of recalibrated weighted samples $\{(\hat{\boldsymbol{\theta}}^{(i)}, w^{(i)})\}_{i=1}^N$ from the recalibrated approximate posterior $\hat{\pi}_{ABC}(\boldsymbol{\theta}|\boldsymbol{s}_{\text{obs}})$.

as bias, or over-/under-estimation of variance. These deviations, contained within the marginal $p_j^{(i)}$, are then mapped onto the quantiles of the original ABC approximation of $\pi(\boldsymbol{\theta}|\boldsymbol{s}_{\text{obs}})$, producing the adjusted sample $\hat{\boldsymbol{\theta}}^{(i)} = (\hat{\theta}_1^{(i)}, \ldots, \hat{\theta}_d^{(i)})^\top$ where $\hat{\theta}_j^{(i)} = \tilde{F}^{-1}_{j,\boldsymbol{s}_{\text{obs}}}(p_j^{(i)})$ for $j = 1, \ldots, d$.

If $G_{\boldsymbol{s}^{(i)}}(\boldsymbol{p}) = G_{\boldsymbol{s}_{\text{obs}}}(\boldsymbol{p})$, then Result 2 states that the resulting $\hat{\boldsymbol{\theta}}^{(i)}$ would be a draw from $F_{\boldsymbol{s}_{\text{obs}}}(\boldsymbol{\theta})$. In practice, however, it must be assumed that $G_{\boldsymbol{s}^{(i)}}(\boldsymbol{p}) \approx G_{\boldsymbol{s}_{\text{obs}}}(\boldsymbol{p})$, and so the recalibrated draws $\hat{\boldsymbol{\theta}}^{(i)}$ will be draws from an approximation to $F_{\boldsymbol{s}_{\text{obs}}}(\boldsymbol{\theta})$. However, if similar biases and deviations away from the true posterior based on the approximation of $\pi(\boldsymbol{\theta}|\boldsymbol{s}^{(i)})$ are similar to those present in the approximation of $\pi(\boldsymbol{\theta}|\boldsymbol{s}_{\text{obs}})$, then the recalibration of an exact sample $\boldsymbol{\theta}^{(i)}$ from $\pi(\boldsymbol{\theta}|\boldsymbol{s}^{(i)})$ to $\hat{\boldsymbol{\theta}}^{(i)}$ approximately from $\pi(\boldsymbol{\theta}|\boldsymbol{s}_{\text{obs}})$ can be expected to be beneficial. We explore how well this works in practice in Section 4.3.

4.2.3 Recalibration with an auxiliary estimator

Algorithm 3 recalibrates the weighted samples $\{(\boldsymbol{\theta}^{(i)}, w^{(i)})\}_{i=1}^{N}$ from steps 1.1–1.4 by constructing a model to approximate the posterior distribution $\pi(\boldsymbol{\theta}|\boldsymbol{s})$ – namely $\pi_{ABC}(\boldsymbol{\theta}|\boldsymbol{s})$ – and construct the univariate marginals $\tilde{F}_{j,\boldsymbol{s}}(\cdot)$ required for the recalibration. However the ABC posterior $\pi_{ABC}(\boldsymbol{\theta}|\boldsymbol{s})$ is not the only model that can be used for this task.

Suppose that, more generally, we have an auxiliary model $g(\boldsymbol{y}|\boldsymbol{\theta})$ with an easily computable maximum likelihood estimator $\boldsymbol{s} = S(\boldsymbol{y})$, so that $g(\boldsymbol{y}|\boldsymbol{\theta}) = g(\boldsymbol{s}|\boldsymbol{\theta})$. Motivated by arguments in indirect inference (Gourieroux et al., 1993; Gleim and Pigorsch, 2013) and Bayesian indirect inference (Drovandi et al., 2017, 2015) the auxiliary model is commonly a close, but tractable surrogate of the intractable model $p(\boldsymbol{y}|\boldsymbol{\theta})$. Suppose also that given the prior distribution $\pi(\boldsymbol{\theta})$ it is computationally convenient to fit the associated posterior distribution $g(\boldsymbol{\theta}|\boldsymbol{s}) \propto g(\boldsymbol{s}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$ to $\boldsymbol{s}$. In this setting, the univariate marginal distributions of $g(\boldsymbol{\theta}|\boldsymbol{s}^{(i)})$ can be constructed as $\tilde{F}_{j,\boldsymbol{s}^{(i)}}(\cdot)$, and subsequently used for the recalibration of the weighted sample $(\boldsymbol{\theta}^{(i)}, w^{(i)})$ as before. With good choice of $g(\boldsymbol{\theta}|\boldsymbol{s})$ this procedure can be considerably faster and more efficient than using the ABC approximate posterior $\pi_{ABC}(\boldsymbol{\theta}|\boldsymbol{s})$ as the auxiliary estimator.

This use of the auxiliary model is different to some previous usages where the MAP or MLE of the auxiliary model defined summary statistics that were then used for a standard ABC analysis (e.g. Gleim and Pigorsch, 2013; Drovandi et al., 2015; Martin et al., 2017). Here, the whole auxiliary model is used to approximate the intractable posterior and produce univariate marginal distributions, rather than merely define a point estimate of the parameters.

Algorithm 4 lists the modifications to Algorithm 3 when using a more general auxiliary model. We explore the use of non-ABC auxiliary models in the simulation study in Section 4.3.1, and directly contrast ABC with non-ABC auxiliary models in the recalibration of an analysis of stereological extremes in Section 4.4.

Figure 4.1 provides a simple graphical illustration of the recalibration procedure. In this toy example, the assumption $G_{\boldsymbol{s}^{(i)}}(\boldsymbol{p}) \approx G_{\boldsymbol{s}_{\text{obs}}}(\boldsymbol{p})$ is numerically satisfied for all i in the acceptance region, so the recalibrated samples are exact. In general situations, it is difficult to anticipate whether this assumption will hold for a given choice of auxiliary model. Nevertheless, it is straightforward to retrospectively assess if $G_{\boldsymbol{s}^{(i)}}(\boldsymbol{p}) \approx G_{\boldsymbol{s}_{\text{obs}}}(\boldsymbol{p})$

Algorithm 4 Recalibration of an auxiliary estimator (Modifications to Algorithm 3)

Inputs:

- A tractable auxiliary model for the posterior $\pi(\boldsymbol{\theta}|\boldsymbol{y})$ with accessible maximum likelihood estimate (MLE) $\boldsymbol{s} = S(\boldsymbol{y})$ that admits auxiliary univariate marginal distribution functions $\tilde{F}_{j,\boldsymbol{s}}(\theta_j)$, $j = 1, \ldots, d$.

Data simulation and weighting:
For $i = 1, \ldots, N$:

1.3 Compute the MLE of the auxiliary model $\boldsymbol{s}^{(i)} = S(\boldsymbol{y}^{(i)})$.

Recalibration:

2.1 For $j = 1, \ldots, d$, construct $\tilde{F}_{j,\boldsymbol{s}_{\text{obs}}}(\cdot)$ based on the auxiliary MLE $\boldsymbol{s}_{\text{obs}}$.

For each i such that $w^{(i)} > 0$, and for $j = 1, \ldots, d$:

2.2 Construct $\tilde{F}_{j,\boldsymbol{s}^{(i)}}(\cdot)$ based on the auxiliary MLE $\boldsymbol{s}^{(i)}$.

Outputs:

- A set of recalibrated weighted samples $\{(\hat{\boldsymbol{\theta}}^{(i)}, w^{(i)})\}_{i=1}^{N}$ approximately from the posterior $\pi(\boldsymbol{\theta}|\boldsymbol{s}_{\text{obs}})$.

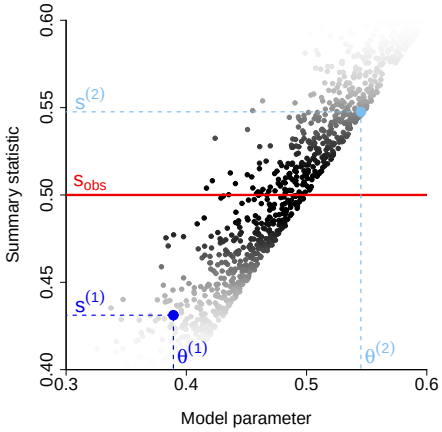
by checking if the pvalues histogram remains the same in different regions of the summary statistics space. In particular, one could test, formally or informally, if plot (b) in Figure 4.1 changes its pattern for different acceptance rates (that are less than the adopted value of 25%).

4.2.4 Regression-adjusted recalibration

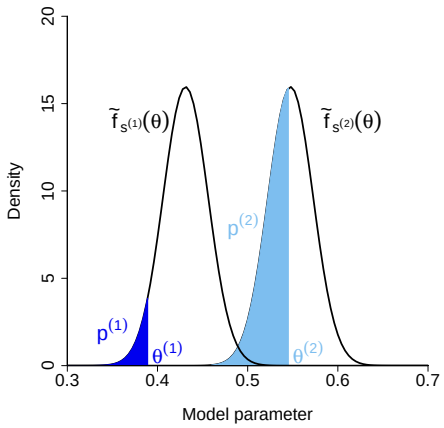
There are two natural ways in which regression-adjustment methods can be combined with recalibration in an ABC analysis. The most straightforward is where recalibration is employed to approximately correct for any biases incurred in a standard regression-adjustment ABC analysis (c.f. Marin et al., 2016; Frazier et al., 2017).

An alternative use of regression adjustment methods stems from the fact that the quality of a recalibrated posterior approximation rests on how well $G_{\boldsymbol{s}^{(i)}}(\boldsymbol{p})$ approximates $G_{\boldsymbol{s}_{\text{obs}}}(\boldsymbol{p})$. In the case where there are reasonable differences between $G_{\boldsymbol{s}^{(i)}}(\boldsymbol{p})$ and $G_{\boldsymbol{s}_{\text{obs}}}(\boldsymbol{p})$, one approach is to adjust the values of $\boldsymbol{p}^{(i)}$ given the predictors $\boldsymbol{s}^{(i)}$. In the case of a weighted local-linear regression (e.g. Beaumont et al., 2002) the model would be

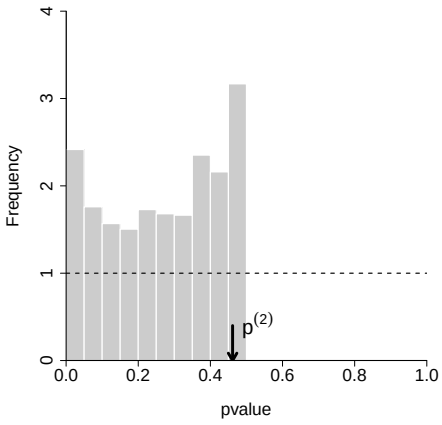
$$\eta(\boldsymbol{p}^{(i)}) = \boldsymbol{\alpha} + \boldsymbol{\beta}(\boldsymbol{s}^{(i)} - \boldsymbol{s}_{\text{obs}}) + \boldsymbol{\epsilon}^{(i)}$$



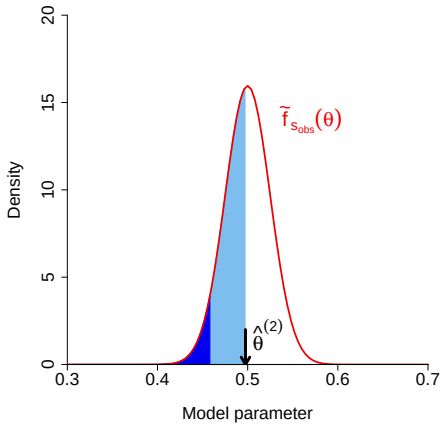
(a) Synthetic samples.



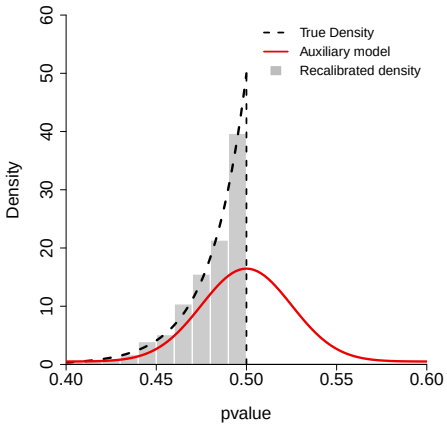
(b) Auxiliary posterior estimates.



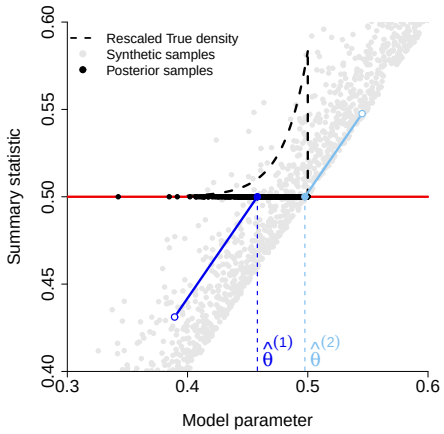
(c) Pvalues histogram.



(d) Recalibrating the initial estimate.



(e) Recalibration effect.



(f) Projection interpretation.

Figure 4.1: Let $(s|\theta) \sim \theta + \exp(50)$, $\theta \sim U(0,1)$, $\tilde{f}_{s^{(i)}}(\theta) = \mathrm{N}(s^{(i)}, \sigma = 0.025)$ and $s_{\mathrm{obs}} = 0.5$. Panel (a) shows the best 25% of 5000 synthetic samples, with points $i = 1, 2$ highlighted in blue. The gray intensity reflects the sample weight. Figure (b) presents the auxiliary posterior density estimates and the pvalues associated to the first two samples – see Step 2.3 in Algorithm 3. (c) and (d) show, respectively, the histogram of the observed pvalues and the recalibrated sample for $i = 2$ (Step 2.5 in Algorithm 3). Panel (e) illustrates the recalibration effect, while (f) provides a geometrical interpretation of the recalibration process as a projection technique that does not require the specification of a regression model.

for $i = 1, \ldots, N$, where $\boldsymbol{\alpha} \in \mathbb{R}^d$, $\boldsymbol{\beta}$ is a $d \times \dim(s^{(i)})$ matrix, $\boldsymbol{\epsilon}^{(i)} \sim N_d(0, \Sigma)$, $\eta(\cdot)$ is the logistic link function, and where the pair $(\boldsymbol{p}^{(i)}, \boldsymbol{s}^{(i)})$ is given the weight $K_h(\|\boldsymbol{s}^{(i)} - \boldsymbol{s}_{\text{obs}}\|)$. In this manner, the aim is to transform $\boldsymbol{p}^{(i)}$ so that it behaves as an approximate sample from $G_{\boldsymbol{s}_{\text{obs}}}(\boldsymbol{p})$ rather than an exact sample from $G_{s^{(i)}}(\boldsymbol{p})$. Of course for this adjustment to be beneficial it requires that the fitted regression model be highly accurate. If the model is poorly specified, as with standard regression-adjusted analyses, the final estimation error could easily increase compared to if it is not used. Both alternative uses of regression-adjustment with recalibration are examined in Section 4.3.2.

4.3 Simulation studies

We now examine the performance of the recalibration procedure of the previous Section on two simulated examples. The first makes use of a tractable Gaussian auxiliary model estimator for inference on a sum of lognormals distribution. The second examines the effect of recalibration on a "twisted normal" model under varied ABC inference configurations.

4.3.1 A sum of log-normals model

Consider a univariate random variable $Y = \sum_{\ell=1}^{L} X_\ell$, where $X_\ell \sim \text{LogNormal}(\mu, \sigma)$ are independent and identically distributed log-normal random variables with parameter $\boldsymbol{\theta} = (\mu, \sigma)^\top$. Log-normal distributions are commonly used to model heavy-tailed quantities, including stock prices and insurance claims. In these settings, Y can represent the complete value of a stock portfolio, or the total liability of claims for an insurance company (particularly if L is also random). Despite its structural simplicity, the associated likelihood function $p(\boldsymbol{y}|\boldsymbol{\theta})$ cannot be computed exactly, even numerically, for $L > 3$ (For $L = 2$ and possibly $L = 3$, the likelihood may viably be computed numerically through convolution integrals.) Several methods have been proposed to approximate this function (Fenton, 1960; Schwartz and Yeh, 1982; Jingxian et al., 2005), with the Fenton-Wilkinson approximation perhaps the most widely known (Fenton, 1960; Asmussen and Rojas-Nandayapa, 2008). Here, the intractable likelihood is approximated by another log-normal distribution with matching first and second moments. More precisely, it is assumed

that $p_Y(\boldsymbol{y}|\boldsymbol{\theta}) \approx p_Z(\boldsymbol{y}|\boldsymbol{\theta})$, where $Z \sim \text{LogNormal}(\alpha, \beta^2)$, with

$$\alpha = \mu + \log L + 0.5(\sigma^2 - \beta^2),$$
$$\beta^2 = \log[(\exp(\sigma^2) - 1)/L + 1].$$

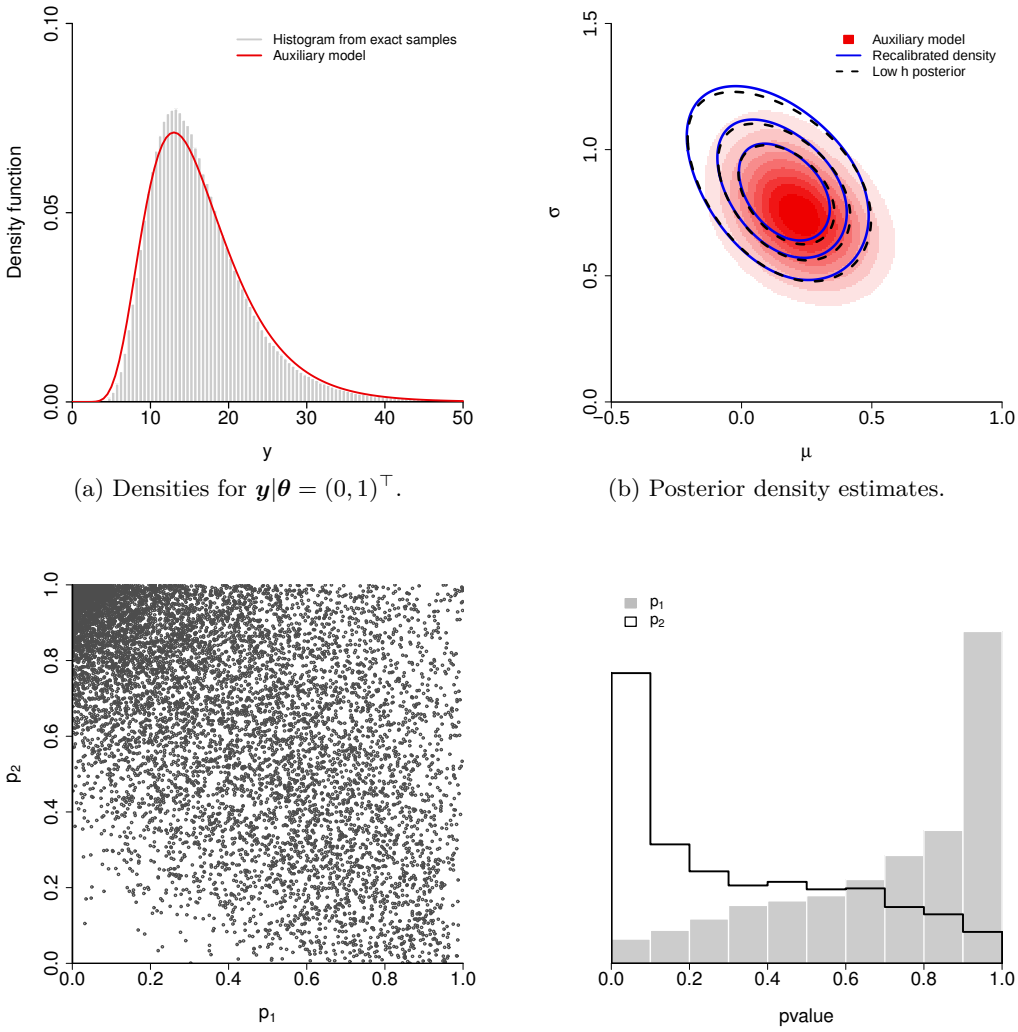
Suppose that we have n observations of Y, $\boldsymbol{y}_{\text{obs}} = (y_{\text{obs},1}, \ldots, y_{\text{obs},n})^\top$, and $\pi(\boldsymbol{\theta})$ is defined through the independent marginal prior distributions $\mu \sim N(0,1)$ and $\sigma^2 \sim \text{Gamma}(1,1)$, where $\boldsymbol{\theta} = (\mu, \sigma)^\top$. While the target posterior $\pi(\boldsymbol{\theta}|\boldsymbol{y}) = \pi_Y(\boldsymbol{\theta}|\boldsymbol{y}) \propto p_Y(\boldsymbol{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$ is intractable, the approximation $\pi_Z(\theta|\boldsymbol{y}) \propto p_Z(\boldsymbol{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$ is amenable to posterior simulation algorithms such as MCMC. In principle then, this lognormal approximation $\pi_Z(\boldsymbol{\theta}|\boldsymbol{y})$ could be used as the auxiliary posterior model $g(\boldsymbol{\theta}|\boldsymbol{s})$, where $\boldsymbol{s}$ is the MLE of $p_Z(\boldsymbol{y}|\boldsymbol{\theta})$. However, to do this would then require that a posterior simulation algorithm be implemented to draw samples from $\pi_Z(\boldsymbol{\theta}|\boldsymbol{y}^{(i)}) = g(\boldsymbol{\theta}|\boldsymbol{s}^{(i)})$, for each i for which $w^{(i)} = K_h(\|\boldsymbol{s}^{(i)} - \boldsymbol{s}_{\text{obs}}\|) > 0$, in order to construct the $\tilde{F}_{j,\boldsymbol{s}^{(i)}}(\cdot)$ marginal distributions. This would impose a large computational burden.

Instead we approximate $\pi_Z(\boldsymbol{\theta}|\boldsymbol{y})$ by a bivariate normal density $N_2(\boldsymbol{\theta}_y^*, \Sigma_y)$, where $\boldsymbol{\theta}_y^* = \arg\max_{\boldsymbol{\theta}} p_Z(y|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$ and Σ_y is the inverse of the Hessian matrix of $-\log(p_Z(y|\boldsymbol{\theta})\pi(\boldsymbol{\theta}))$ (i.e. of the negative log of the tractable auxiliary posterior) evaluated at $\boldsymbol{\theta}_y^*$. In this manner, the auxiliary model $g(\boldsymbol{\theta}|\boldsymbol{s})$ is specified by this $N_2(\boldsymbol{\theta}_y^*, \Sigma_y)$ distribution, with $\boldsymbol{s} = (\boldsymbol{\theta}_y^*, \Sigma_y)^\top$, and the marginal distribution functions $\tilde{F}_{j,\boldsymbol{s}}(\cdot)$ are immediately available as univariate normal distribution functions. Calculation of $\boldsymbol{\theta}_y^*$ and Σ_y is very quick.

We simulate $n = 10$ observations from the true model $Y = \sum_{\ell=1}^{10} X_\ell$, where $X_\ell \sim \text{LogNormal}(0,1)$, to produce the observed dataset $\boldsymbol{y}_{\text{obs}}$. Algorithm 4 was then used to generate $N = 10,000$ approximate posterior samples. For simplicity, we specified $h = \infty$ so that the weights $w^{(i)} = 1/N$ were all equal. This provides a challenging scenario as we are then attempting to recalibrate all samples drawn from the prior to behave as approximate samples from $\pi(\boldsymbol{\theta}|\boldsymbol{s}_{\text{obs}})$.

Figure 4.2a compares the Fenton-Wilkinson lognormal density, $p_Z(\boldsymbol{y}|\boldsymbol{\theta})$, with the true density $p_Y(\boldsymbol{y}|\boldsymbol{\theta})$ at the true parameter values of $\boldsymbol{\theta} = (0,1)^\top$. The lognormal density is clearly a reasonable match for the true density in this case, although it is slightly more diffuse. However the resulting posterior estimate (shading) is inaccurate, as illustrated in Figure 4.2b, compared to that obtained under a highly computation intensive ABC

rejection sampler (dashed lines) with the vector $\boldsymbol{s} = S(\boldsymbol{y}) = \boldsymbol{\theta}^*_y$ as summary statistics and with the kernel scale parameter h reduced to a very low level. (The use of the MLE of a tractable approximation as summary statistics is a common approach.) In contrast, the resulting recalibrated posterior approximation (solid lines) appears visually very close to the low-h posterior.



(a) Densities for $\boldsymbol{y}|\boldsymbol{\theta} = (0,1)^\top$.

(b) Posterior density estimates.

(c) Realised samples from $G_{\boldsymbol{y}}(\boldsymbol{p})$.

(d) Marginal distributions of p_1 and p_2.

Figure 4.2: Panel (a) compares the true density (histogram), $p_Y(\boldsymbol{y}|\boldsymbol{\theta} = (0,1)^\top)$, with the corresponding Fenton-Wilkinson approximation $p_Z(\boldsymbol{y}|\boldsymbol{\theta} = (0,1)^\top)$ (solid line). Panel (b) compares kernel density estimates (KDE) of the approximate posterior resulting from: a low-h ABC sampler (dashed line), the Fenton-Wilkinson auxiliary model (shading) and the recalibrated posterior (solid lines). Panels (c) and (d) respectively present the joint and marginal $\boldsymbol{p} = (p_1, p_2)^\top$ values obtained during recalibration.

A bivariate scatterplot and univariate marginal histograms of the $\boldsymbol{p} = (p_1, p_2)^\top$ values produced in the recalibration are shown in Figures 4.2c and 4.2d. The non-uniformity of the marginal histograms suggests that the Fenton-Wilkinson method overestimates μ and

underestimates σ for this analysis, which is supported by the posterior density estimates in 4.2b. In this case the recalibration procedure corrects these errors successfully. In this analysis, the entire inference process took only a few seconds to complete on a desktop PC, with the computational cost dominated by the optimization process involved in computing $\boldsymbol{\theta}_y^*$. In comparison, the cost of recalibration was negligible, as it only involved calculating $\boldsymbol{p}$ and quantiles from univariate normal distributions.

4.3.2 A "twisted normal" model

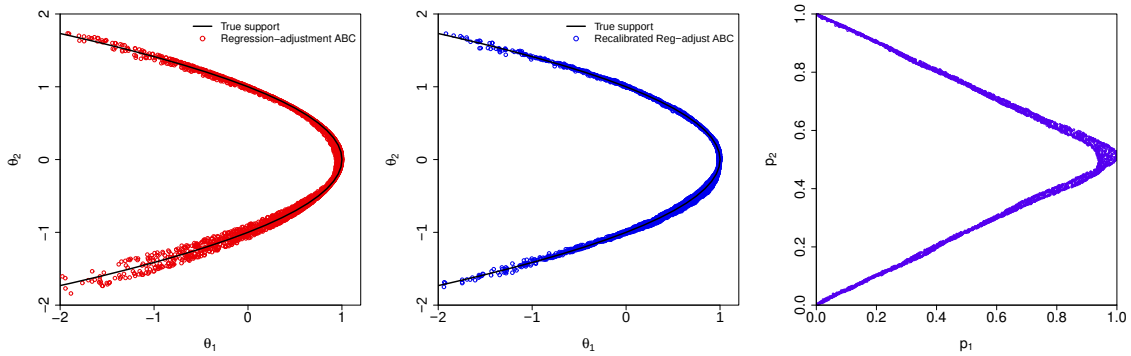
In this analysis, we investigate and quantify the effect of recalibration of standard ABC sampler output under various conditions. We consider the simple, deterministic data-generating model $Y = \theta_1 + \theta_2^2$, with $\boldsymbol{\theta} = (\theta_1, \theta_2)^\top$, and suppose that θ_1 and θ_2 have independent $N(0, 1)$ priors. For a single observed data point $\boldsymbol{y}_{\text{obs}} = y$, the resulting posterior mass is then concentrated on the set of points satisfying $\theta_1 = y - \theta_2^2$. For the below analysis we adopt $\boldsymbol{y}_{\text{obs}} = 1$.

We follow Algorithm 3, and draw $N = 10,000$ samples from the prior distribution, use the full dataset $\boldsymbol{y}$ (a single data point) as the summary statistic, and adopt the Epanechnikov kernel K_h, with h determined by giving the 3,000 samples $\boldsymbol{\theta}^{(i)}$ for which $\boldsymbol{s}^{(i)}$ is closest to $\boldsymbol{s}_{\text{obs}}$ non-zero weights $w^{(i)}$ (e.g. Biau et al., 2015). The 30% acceptance rate of the algorithm is approximately optimal for regression adjustment ABC in this analysis, in terms of producing the minimum mean square error (MSE) of a particular posterior functional (see below and Figure 4.4a).

Figure 4.3a illustrates the regression-adjusted ABC samples in comparison to the support of the true posterior, shown by the solid line. Figure 4.3b shows the same samples following recalibration, which includes the $\boldsymbol{p}$ value regression adjustment of Section 4.2.4. Standard regression-adjustment ABC is easily able to recover the twisted normal shape of the true posterior distribution, however the ABC approximation error is reflected by the extent of the samples lying far from the true posterior support (the solid line). The recalibrated samples, while still having some deviation away from the true posterior support, visibly produce an improved posterior approximation. This is particularly evident in the lower tail of the θ_2 margin.

Figure 4.3c shows the bivariate distribution of the realised $\boldsymbol{p} = (p_1, p_2)^\top$ values. Here, the univariate marginal distributions are almost uniform, indicating that the marginal

posterior distributions of the regression-adjusted ABC posterior approximation are close to the true posterior marginal distributions (c.f. Result 1 and Prangle et al., 2014), while the striking dependence structure is a direct result of the form of $\pi(\boldsymbol{\theta}|\boldsymbol{s}_{\text{obs}})$.



(a) Regression-adjusted ABC. (b) Recalibrated posterior samples. (c) Realised $\boldsymbol{p}$ values.

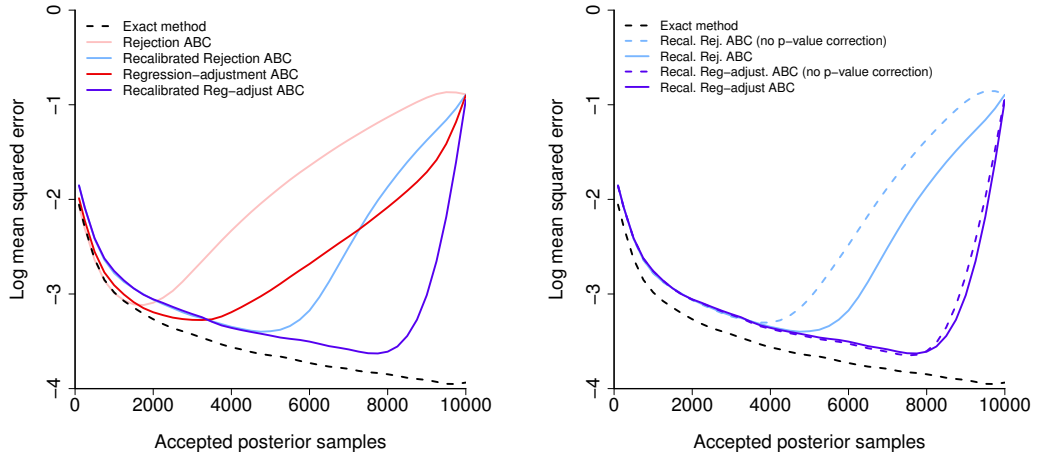
Figure 4.3: Panel (a) illustrates 3,000 samples from posterior distribution estimates using regression adjusted ABC and panel (b) the same samples following recalibration. The grey line indicates the support of the true posterior. Panel (c) presents the corresponding realised $\boldsymbol{p} = (p_1, p_2)^\top$ values.

More qualitatively, we consider estimation of the posterior expectation $E(\theta_1 - \theta_2|\theta_1 + \theta_2^2 = 1)$ under each of four ABC posterior approximation procedures: standard rejection sampling ABC both with and without regression adjustment, and each of these with a subsequent recalibration adjustment (including a regression adjustment on the $\boldsymbol{p}$ values). This computation was repeated 1,000 times and for a range of Epanechnikov kernel scale parameter values h, resulting in between 100 and all 10,000 samples with non-zero weight $w^{(i)} > 0$ being used for the computation. The log (base 10) mean squared error (MSE) over these 1,000 replicates was recorded. The conclusions of the below analysis were unchanged when other quantities of potential interest such as $P(\theta_1 > \theta_2|\theta_1 + \theta_2^2 = 1)$ were considered.

Figure 4.4a displays the log of the MSE for each method as a function of the number of posterior samples (out of 10,000). The same quantity based on samples drawn from the exact posterior is illustrated by the dashed line. Each of the ABC based log MSE curves behave in a similar way as the number of posterior samples increases (i.e. as the kernel scale parameter h increases). For small scale parameter values, the log MSE initially decreases as long as the quality of the posterior approximation for each method is high, with the decrease in log MSE achieved through an increase in the number of samples. That is, the high log MSE for low h is primarily driven by Monte Carlo error. At some

point, however, with increasing h the quality of the posterior approximation deteriorates too much, and the log MSE increases due to bias in the posterior approximation.

However, the relative performance of each ABC method differs in its performance for low h, and the point at which the bias in the posterior approximation begins to dominate the MSE. For low h values standard rejection ABC (light red line) performs as well as the exact posterior distribution until around 1,500 samples. For low h, implementing any post-processing method only increases the Monte Carlo error, as these require the estimation of regression parameters and/or marginal distribution functions $\tilde{F}_{j,\boldsymbol{s}}(\cdot)$, with more overheads required for recalibration than for regression adjustment. For larger h, however, there is a clear benefit to post-processing, with the quality of the regression adjusted posterior approximation (dark red line) meaning that it can reach a lower log MSE for an h equivalent to around 3,000 samples. The recalibrated posterior approximations perform even more efficiently, with the recalibrated regression-adjusted ABC posterior the most efficient of all, achieving their optimum log MSE values at around 5,000 and 8,000 samples. In fact, the minimum MSE obtained by recalibration (recalibrated regression-adjusted ABC) was 0.0002, which is a sizeable reduction from its uncalibrated counterpart of 0.0005 (regression-adjusted ABC) – especially taking into account the theoretical minimum, 0.0001, obtained by exact calculations.



(a) Effect of recalibration.

(b) Effect of correcting $\boldsymbol{p}$ values.

Figure 4.4: Log mean squared error of different ABC methods when estimating $E(\theta_1 - \theta_2 | \theta_1 + \theta_2^2 = 1)$, as a function of the number of posterior samples (out of $N = 10,000$). Panel (a) compares rejection (red lines) and recalibrated (blue lines) ABC estimators. Darker and lighter lines respectively denote rejection and regression adjustment ABC. The dashed black line depicts the case when the samples were drawn from the exact posterior. Panel (b) contrasts log MSE for recalibrated ABC methods both with and without regression adjusted $\boldsymbol{p}$ values.

Figure 4.4b presents the same information as Figure 4.4a but comparing the recalibration adjusted methods both with and without regression adjusted $\boldsymbol{p}$ values (Section 4.2.4). In this case, adjusting the $\boldsymbol{p}$ values clearly improves recalibrated rejection ABC, but recalibrated regression-adjustment ABC is only improved to a small extent. This primarily occurs as the linear regression model assumptions are not reasonable in this region.

In the above analysis, for ease of presentation, the same acceptance rate adopted in steps 1.4 and 2.1 of Algorithm 3 was used when computing the marginal estimates $\tilde{F}_{j,\boldsymbol{s}}(\cdot)$ in step 2.2. However, it could be computationally more efficient to use different rates for each step, such as using 30% of the synthetic samples to recalibrate a regression-adjustment ABC based on an acceptance rate of 10%.

4.4 Application: Estimation in Stereological extremes

During the production of a steel block, endogenous or exogenous chemical compounds are unavoidably embedded into the final product. Known as *inclusions*, these foreign substances affect the toughness, corrosion resistance and other features of the steel. The size of the largest inclusions, which cannot be directly observed, are particularly influential to the overall quality. Therefore, interest lies in an extreme value problem in which inference is required on the distribution of the largest inclusion sizes based on the inclusions observed in a two-dimensional planar slice through the block. Each observed cross-sectional inclusion size in $\boldsymbol{y}_{\text{obs}} = (y_{\text{obs},1}, \ldots, y_{\text{obs},n})^\top$ is related to an unknown inclusion size $V_i > y_{\text{obs},i}$ in 3-dimensional space. The number of inclusions in the sample is random, and, for any given i, the probability of observing $y_{\text{obs},i}$ depends on V_i – larger inclusions are more likely to intersect the planar slice.

To make inference in this stereological context, it is commonly assumed that the inclusion centres follow a homogeneous Poisson process with rate λ, and that inclusion sizes are mutually independent and independent of inclusion location. These assumptions are widely regarded as reasonable. When it comes to the shape of the inclusions, however, different formulations have been studied. Anderson and Coles (2002) assumed that inclusions were spherical, with "size" being characterized by the inclusion's diameter V. Subsequently Bortot et al. (2007) considered randomly oriented ellipsoidal shapes, where $y_{\text{obs},i}$ then refers to the largest principal diameter of the ith observed ellipse and V_i the

largest diameter of the corresponding ellipsoid. In both spherical and ellipsoidal constructions, a generalized Pareto distribution (GPD) is assigned to $V|V > v_0$, where v_0 is an appropriate threshold. The distribution function is given by

$$P(V \leq v|V > v_0) = 1 - \left[1 + \frac{\xi(v - v_0)}{\sigma}\right]_+^{-1/\xi},$$

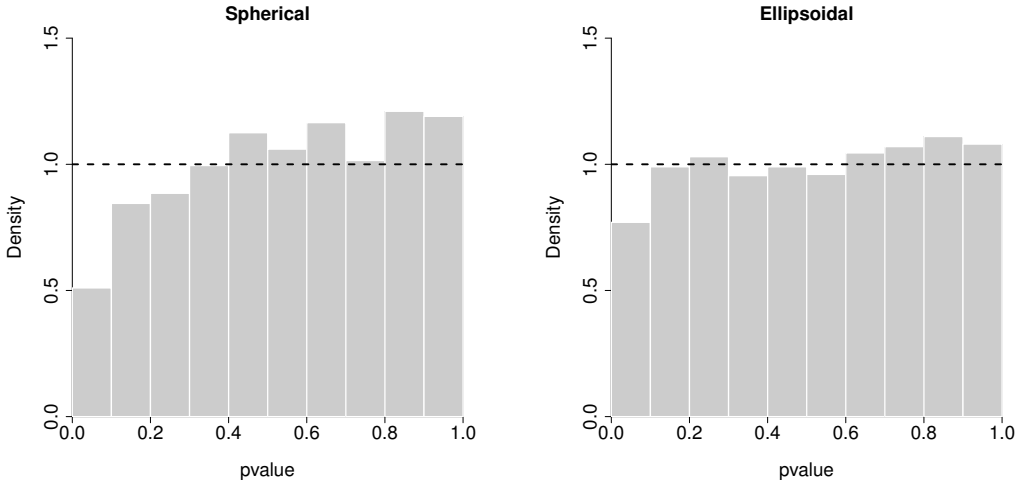
where $[a]_+ = \max\{0, a\}$, $v > v_0$, and $\sigma > 0$ and $-\infty < \xi < \infty$ are scale and shape parameters. To fully specify the model, Bortot et al. (2007) also assumed that the two non-leading principal diameters of a given ellipsoid are defined as $V_1 = U_1 V$ and $V_2 = U_2 V$, where U_1 and U_2 are independent standard uniform variables.

Anderson and Coles (2002) derived an exact MCMC sampler for the posterior distribution of their spherical model. However, the likelihood induced by the more plausible ellipsoidal model is computationally intractable, which motivated Bortot et al. (2007) to use ABC methods for inference on $\boldsymbol{\theta} = (\lambda, \sigma, \xi)^\top$. Erhardt and Sisson (2016) conducted a simulation study to investigate the performance of different ABC implementations in this context, demonstrating that regression-adjustment substantially improved the accuracy of rejection ABC. They adopted a uniform prior distribution for $\boldsymbol{\theta}$, restricted to a region that comfortably enveloped the effective support of the posterior distribution. In addition, they adopted the summary statistics

$$S(\boldsymbol{y}) = (n', q_{0.5}(\boldsymbol{y}), q_{0.7}(\boldsymbol{y}), q_{0.9}(\boldsymbol{y}), q_{0.95}(\boldsymbol{y}), q_{0.99}(\boldsymbol{y}), q_1(\boldsymbol{y}))^\top, \tag{4.2}$$

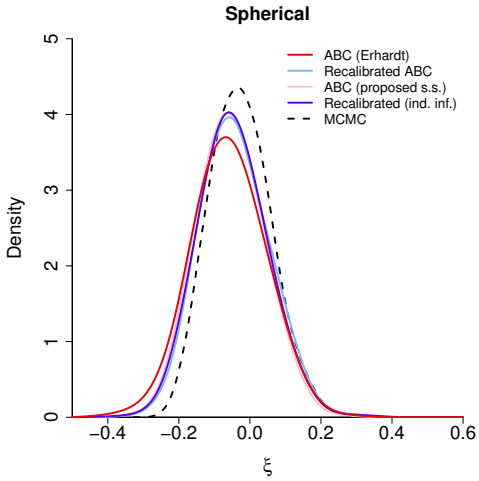
where $q_a(\boldsymbol{y})$ denotes the a-th quantile of $\boldsymbol{y}$, and n' is the (random) number of observations in $\boldsymbol{y}$. Their ABC analyses were performed using the best 2,000 out of $N = 2$ million generated samples $\{(\boldsymbol{\theta}^{(i)}, \boldsymbol{s}^{(i)})\}_{i=1}^N$.

With these same settings, we revisit the analysis in Erhardt and Sisson (2016), using Algorithm 3 to generate recalibrated samples from the regression-adjustment ABC posterior approximation. We focus our attention on the shape parameter ξ as it determines the tail behaviour of extreme value models. We also investigate recalibrating a computationally cheaper auxiliary method using Algorithm 4, similar to that implemented in Section 4.3.1. In the stereological context, intractability arises from the impossibility to measure the diameters V_i. We therefore use a tractable, but misspecified, auxiliary model which assumes that the observable diameters, $\boldsymbol{y}|\boldsymbol{y} > v_0$, follow a GPD with parameters σ' and

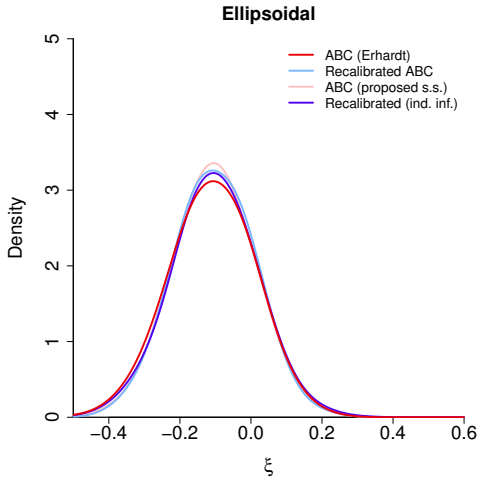


(a) Marginal distribution of p_ξ.

(b) Marginal distributionsof p_ξ.



(c) Estimated posterior of ξ.



(d) Estimated posterior of ξ.

Figure 4.5: Panels (a) and (b) show for the spherical and elliptical cases, respectively, the realised $\boldsymbol{p}$ values, p_ξ, associated with the recalibration of regression-adjustment ABC using the summary statistics (4.2). Panels (c) and (d) compare the marginal posterior densities for ξ estimated by different methods and summary statistics.

ξ'. A new set of summary statistics may then be defined as

$$S'(\boldsymbol{y}) = (n', \tilde{\sigma}(\boldsymbol{y}), \tilde{\xi}(\boldsymbol{y}))^\top,$$

where $\tilde{\sigma}(\boldsymbol{y})$ and $\tilde{\xi}(\boldsymbol{y})$ are the MLEs of this auxiliary model. Although highly informative, $S'(\boldsymbol{y})$ is not itself an estimator for $\boldsymbol{\theta}$. So for each simulated dataset $\boldsymbol{s}'^{(i)}$, we follow Fearnhead and Prangle (2012) and estimate $\boldsymbol{\theta}^{(i)}$ by $\boldsymbol{\theta}^{+(i)}$, where $\boldsymbol{\theta}_j^{+(i)} = E(\boldsymbol{\theta}_j^{(i)} | \boldsymbol{s}_j'^{(i)})$, using univariate (splines) smoothers fitted to $\{(\boldsymbol{\theta}_j^{(i)}, \boldsymbol{s}_j'^{(i)})\}_{i=1}^N$ for $j = 1, \ldots, d$, using the default settings of the `smooth.spline` function in `R`. Finally, we define a Gaussian auxiliary marginal estimator as $\tilde{F}_{j,\boldsymbol{s}'^{(i)}}(\theta_j) = \Phi(\theta_j; \boldsymbol{\theta}_j^{+(i)}, \hat{\sigma}_j)$, where $\hat{\sigma}_j$ is the standard deviation of

the spline residuals for parameter j.

Figures 4.5a and 4.5b show the distribution of the marginal $\boldsymbol{p}$ values for ξ, p_ξ, obtained when recalibrating the best ABC estimator considered in Erhardt and Sisson (2016) – namely, regression-adjustment ABC with summary statistics given by (4.2). The left-skew of both plots indicates that this regression-adjustment ABC tends to underestimate ξ. A Kolmogorov-Smirnov test rejects the hypothesis that the p_ξ samples are from a $U(0,1)$ distribution (with p-values of 7×10^{-11} and 0.02 for the spherical and ellipsoidal cases respectively).

Figures 4.5c and 4.5d compare the marginal posterior density estimates for ξ using regression adjusted ABC (red line) and its recalibration (light blue line), with the posterior estimates using regression adjusted ABC using the summary statistics $S'(\boldsymbol{y})$ (pink line), and the recalibration of the Gaussian auxiliary estimator (dark blue line). Also shown for the spherical model (dashed line) is the exact posterior obtained from the MCMC sampler of Anderson and Coles (2002), although this is based on a partially-conjugate prior specification defined on a reparameterised space, and so this targets a different posterior to the ABC algorithms. Accordingly, a perfect correspondence between the exact posterior and the ABC methods should not be expected.

For the spherical model (Figure 4.5c) the underestimation of ξ reflected in the p_ξ values using the summary statistics (4.2) is visibly evident, and this is corrected under recalibration. For the ellipsoidal case, the initial bias in ξ was so mild that recalibration has barely affected the posterior estimate. For both spherical and ellipsoidal models, standard regression-adjusted ABC with the new summary statistics $S'(\boldsymbol{y})$ has performed as well as the recalibration of the Gaussian auxiliary estimator, with both densities appearing indistinguishable from the recalibrated standard ABC analysis. That these density estimates all lie in the same place strongly suggests that these are all good approximations to the true posterior in this case (with the uniform prior specification). It also suggests that the indirect inference-based summary statistics $S'(\boldsymbol{y})$ are highly informative for these models. Overall, either adoption of $S'(\boldsymbol{y})$ or any method of recalibration produces a more accurate posterior approximation than the analysis performed in Erhardt and Sisson (2016).

4.5 Discussion

This article introduces a recalibration procedure to post-process output from approximate Bayesian methods, in particular ABC techniques, based on the ideas in Prangle et al. (2014). Recalibration can improve the quality of an approximation of the posterior distribution by ensuring that the adjusted posterior estimate approximately satisfies the coverage property. This means that errors and biases induced by adopting various posterior approximations, such as the standard ABC posterior approximation or auxiliary model approximations, can be (approximately) corrected. Indeed, this may then be exploited so that the most computationally efficient approximate posterior can be adopted, which is not necessarily standard ABC, in the knowledge that a good adjustment is available to correct model mis-specification.

Accordingly, in Section 4.3.1 the error induced by the incorrect assumption that a sum of log-normal distributions follows a log-normal distribution was substantially reduced by recalibration. Section 4.3.2 illustrated that recalibration can serve as a non-parametric alternative to regression-adjustment ABC (when an appropriate regression model is not available), or as an additional layer of post-processing to correct the biases of the regression-adjustment itself. In the stereological extremes analysis in Section 4.4, using recalibration to correct a small bias in the results obtained by Erhardt and Sisson (2016), along with a more detailed investigation, provided a reassurance that more substantial errors have not been incurred in this analysis.

Recalibration does come with some computational cost, which may or may not be worthwhile, depending on a number of factors. An obvious practical requirement is that the auxiliary method used to construct the univariate marginal distributions $\tilde{F}_{j,\boldsymbol{s}}(\cdot)$ needs to be fast, or the computational overheads involved in recalibration will dominate those of the original analysis. Recalibration is also particularly appealing when simulation of datasets $\boldsymbol{y} \sim p(\boldsymbol{y}|\boldsymbol{\theta})$ under the model is computationally expensive. For instance, in the stereological extremes analysis of Section 4.4, the recalibration stage of Algorithm 4 required no more than 10% of the total computational time – a modest computational cost for this analysis.

As with standard ABC methods, the best choice of kernel scale parameter h is generally a non-trivial task. In principle, this choice is based on a balancing of Monte Carlo variation

and the intrinsic error arising from assuming that $G_{\boldsymbol{s}}(\boldsymbol{p})$ is nearly independent from $\boldsymbol{s}$ in the neighborhood of $\boldsymbol{s}_{\text{obs}}$, as visualised in Figure 4.4. Further, as observed by Prangle et al. (2014), marginal uniform distributions for the realised $\boldsymbol{p}$ values are possible from distributions other than the true posterior distribution. In particular, if ABC or the auxiliary method returns the prior distribution $\pi(\boldsymbol{\theta})$ as the approximate posterior (see also the noisy ABC of Fearnhead and Prangle (2012)), then as the prior automatically satisfies coverage (Prangle et al., 2014), recalibration post-processing will have no power to make a correction.

We have presented recalibration as a post-processing method for ABC and indirect inference based procedures. However, it may conceivably also be used for other methods for approximating posterior distributions, including variational methods and expectation propagation techniques. The `R` code used to perform the computations in this chapter is available online as supporting information to this article. In addition, an implementation of Algorithm 3 is available in the `abctools` `R` package.

Chapter 5

Likelihood-free approximate Gibbs sampling [1]

5.1 Introduction

Likelihood-free methods refer to a family of procedures that aim to perform likelihood-based statistical inference, but without direct evaluation of the likelihood function. This is attractive when the likelihood function is computationally prohibitive to evaluate due to dataset size or model complexity, or when the likelihood function is only known through a data generation process. Some classes of likelihood-free methods include pseudo-marginal methods (Beaumont, 2003; Andrieu and Roberts, 2009), indirect inference (Gourieroux et al., 1993) and approximate Bayesian computation (Beaumont, 2003; Sisson et al., 2017a).

In particular, approximate Bayesian computation (ABC) methods form an approximation to the computationally intractable posterior distribution by firstly sampling parameter vectors from the prior, and conditional on these, generating synthetic datasets under the model. The parameter vectors are then weighted by how well a vector of summary statistics of the synthetic datasets matches the same summary statistics of the observed data. ABC methods have seen extensive application and development over the past 15 years. See e.g. Sisson et al. (2017a) for an up to date overview of this area.

However, ABC methods have mostly been limited to analyses with moderate numbers of parameters (< 50) due to the inherent curse-of-dimensionality of matching larger num-

[1] In preparation for publication as: Rodrigues, G. S., Nott, D. J., Sisson, S. A., "Likelihood-free approximate Gibbs sampling."

bers of summary statistics, in what may be viewed as a high-dimensional kernel density estimation problem (Blum and François, 2010). For a fixed computational budget, the quality of the ABC posterior approximation deteriorates rapidly as the number of summary statistics (which is driven by the number of model parameters) increases (Nott et al., 2017).

A number of techniques for extending ABC methods to higher dimensional models have been developed. Post-processing techniques aim to reduce the approximation error by adjusting samples drawn from the ABC posterior approximation in a beneficial manner. These include regression-adjustments (Beaumont et al., 2002; Blum and François, 2010; Blum et al., 2013), marginal adjustment (Nott et al., 2014), and recalibration (Rodrigues et al., 2017b; Prangle et al., 2014). However, by their very nature post-processing techniques are a means to improve an existing analysis rather than an approach to extend ABC methods to higher dimensions. In addition, evidence is emerging that some of these procedures, in particular the regression-adjustments, perform less well than is generally believed (Marin et al., 2016; Frazier et al., 2017).

Alternative model-based approximations to the intractable posterior have been developed, including Gaussian copula models (Li et al., 2017), Gaussian mixture models (Bonassi et al., 2011), regression density estimation (Fan et al., 2013), Gaussian processes (Gutmann and Corander, 2016), Bayesian indirect inference (Drovandi et al., 2015, 2017), variational Bayes (Tran et al., 2017) and synthetic likelihoods (Wood, 2010; Ong et al., 2016). Each of these alternative models have appealing properties, although none of them fully address the high-dimensional ABC problem.

One technique that has some promise in helping extend ABC methods to higher dimensions is likelihood (or posterior) factorisation. When the likelihood can be factorised into lower dimensional components, lower dimensional comparisons of summary statistics can be made, thereby side-stepping the curse of dimensionality to some extend. This has been explored within hierarchical models by Bazin et al. (2010), within an expectation-propagation scheme by Barthelmé and Chopin (2014), for discretely observed Markov models by White et al. (2015), and within the copula-ABC approach of Li et al. (2017). However, such a factorisation is only available for particularly structured models (although see Li et al. (2017)). Other approaches include rephrasing the matching of summary statistics as a rare event problem (Prangle et al., 2016), and developing local Bayesian opti-

misation techniques for high-dimensional intractable models (Meeds and Welling, 2015; Gutmann and Corander, 2016).

In one particular take on posterior factorisation, Kousathanas et al. (2016) developed an ABC Markov chain Monte Carlo (MCMC) algorithm which only updates one parameter per iteration, so that the new candidate can be accepted or rejected based on a small subset of summary statistics. This approach can increase MCMC acceptance rates, although can be limited by the need to generate a synthetic dataset at each algorithm iteration, which may be computationally prohibitive if used for expensive simulators. It also requires the identification of conditionally sufficient statistics for each parameter.

In this article we introduce a likelihood-free approximate Gibbs sampler that targets the high-dimensional posterior indirectly by approximating its full conditional distributions. Low-dimensional regression-based models are constructed for each of these conditional distributions using standard ABC parameter and summary statistic samples, which then permit approximate Gibbs update steps. In contrast to Kousathanas et al. (2016), synthetic datasets are not generated during each sampler iteration, thereby providing efficiencies for expensive simulator models, and only require sufficient synthetic datasets to adequately construct the full conditional models (e.g. Fan et al. (2013)). Construction of the approximate conditional distributions can exploit known structures of the high-dimensional posterior, where available, to considerably reduce computational overheads. The models themselves can also be constructed in localised or global forms.

This manuscript is structured as follows. In Section 5.2 we introduce the method for constructing regression-based conditional distributions and for implementing the likelihood-free approximate Gibbs sampler, and discuss possible sampler variants. In Section 5.3, we explore the performance of the algorithm under various sampler and model settings, and provide a real data analysis of an *Airbnb* dataset using an intractable state space model with 13,140 parameters in Section 5.4. Section 5.5 concludes with a discussion.

5.2 Likelihood-free approximate Gibbs sampler

Suppose that $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_D)^\top$ is a D-dimensional parameter vector, with associated prior distribution $\pi(\boldsymbol{\theta})$, and a computationally intractable model for data $p(\boldsymbol{X}|\boldsymbol{\theta})$. Given the observed data, $\boldsymbol{X}_{\text{obs}}$, interest lies in the posterior distribution $\pi(\boldsymbol{\theta}|\boldsymbol{X}_{\text{obs}}) \propto p(\boldsymbol{X}_{\text{obs}}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$.

Algorithm 5 A simple importance sampling ABC algorithm

Inputs:

- An observed dataset $\boldsymbol{X}_{\text{obs}}$.
- A prior $\pi(\boldsymbol{\theta})$ and intractable generative model $p(\boldsymbol{X}|\boldsymbol{\theta})$.
- An observed vector of summary statistics $\boldsymbol{s}_{\text{obs}} = S(\boldsymbol{X}_{\text{obs}})$.
- A smoothing kernel $K_h(u)$ with scale parameter $h > 0$.
- A positive integer N defining the number of ABC samples.

Data simulation and weighting:
For $i = 1, \ldots, N$:

1.1 Generate $\boldsymbol{\theta}^{(i)} \sim \pi(\boldsymbol{\theta})$ from the prior.
1.2 Generate $\boldsymbol{X}^{(i)} \sim p(\boldsymbol{X}|\boldsymbol{\theta}^{(i)})$ from the model.
1.3 Compute the summary statistics $\boldsymbol{s}^{(i)} = S(\boldsymbol{X}^{(i)})$.
1.4 Compute the sample weight $w^{(i)} \propto K_h(||\boldsymbol{s}^{(i)} - \boldsymbol{s}_{\text{obs}}||)$.

Output:

- A set of weighted samples $\{(\boldsymbol{\theta}^{(i)}, w^{(i)})\}_{i=1}^N$ from $\pi_{ABC}(\boldsymbol{\theta}|\boldsymbol{s}_{\text{obs}})$.

The ABC approximation is given by

$$\pi_{\text{ABC}}(\boldsymbol{\theta}|\boldsymbol{s}_{\text{obs}}) \propto \pi(\boldsymbol{\theta}) \int K_h(\|S(\boldsymbol{X}) - \boldsymbol{s}_{\text{obs}}\|) p(\boldsymbol{X}|\boldsymbol{\theta}) d\boldsymbol{X}, \tag{5.1}$$

where $\boldsymbol{s} = S(\boldsymbol{X})$ is a vector of summary statistics, $\boldsymbol{s}_{\text{obs}} = S(\boldsymbol{X}_{\text{obs}})$ and $K_h(u) = K(u/h)/h$ is a smoothing kernel with bandwidth parameter $h > 0$. If the summary statistics $\boldsymbol{s}$ are sufficient then the approximation error can be made arbitrarily small by taking $h \to 0$ as in this case $\pi_{\text{ABC}}(\boldsymbol{\theta}|\boldsymbol{s}_{\text{obs}})$ will converge to the posterior distribution $\pi(\boldsymbol{\theta}|\boldsymbol{X}_{\text{obs}})$. Otherwise, for non-sufficient $\boldsymbol{s}$ and $h > 0$ the approximation is given as (5.1). See e.g. Sisson et al. (2017b) for further discussion on the construction of this approximation. A simple procedure to draw samples from $\pi_{\text{ABC}}(\boldsymbol{\theta}|\boldsymbol{s}_{\text{obs}})$ is given in Algorithm 5. More sophisticated algorithms are available.

Regression-adjustment post-processing methods (Beaumont et al., 2002; Blum and François, 2010; Blum et al., 2013) are commonly used to mitigate the effect of $h > 0$ in (5.1) by fitting regression models of the form $\theta_d|\boldsymbol{S} \sim f(\theta_d|\boldsymbol{\beta}_d^+, \boldsymbol{S})$, for $d = 1, \ldots, D$, based on the weighted samples $\{(\boldsymbol{\theta}^{(i)}, \boldsymbol{s}^{(i)}, w^{(i)})\}_{i=1}^N$, that are as close as possible to the corresponding intractable marginal distributions $\pi(\theta_d|\boldsymbol{S})$ in the region of $\boldsymbol{s}_{\text{obs}}$. For example, in the local linear approach of Beaumont et al. (2002) the fitted models are of the form

$$\theta_d^{(i)} = \alpha_d + \boldsymbol{\beta}_d^\top (\boldsymbol{s}^{(i)} - \boldsymbol{s}_{\text{obs}}) + \epsilon_d^{(i)},$$

for $i = 1, \ldots, N$ and $d = 1, \ldots, D$, where $\alpha_d \in \mathbb{R}$, $\boldsymbol{\beta}_d \in \mathbb{R}^q$, q is the length of the vector of summary statistics $\boldsymbol{s}$, and $\epsilon_d^{(i)} \sim N(0, \sigma_d^2)$. Here $\boldsymbol{\beta}_d^+ = (\alpha_d, \boldsymbol{\beta}_d, \sigma_d^2)^\top$ is the full vector of unknown regression parameters for model d. Regression-adjustment would then modify each $\theta_d^{(i)}$ to reduce the discrepancy between $\boldsymbol{s}^{(i)}$ and $\boldsymbol{s}_{\text{obs}}$ via $\theta_d^{*(i)} = \hat{\boldsymbol{\beta}}_d^\top \boldsymbol{s}_{\text{obs}} + (\theta_d^{(i)} - \hat{\boldsymbol{\beta}}_d^\top \boldsymbol{s}^{(i)})$ where $\hat{\boldsymbol{\beta}}_d$ denotes the least squares estimate of $\boldsymbol{\beta}_d$.

To construct the likelihood-free approximate Gibbs sampler we similarly build regression models, but in this case we construct regression models of the form $\theta_d | (\boldsymbol{S}, \boldsymbol{\theta}_{-d}) \sim f(\theta_d | \boldsymbol{\beta}_d^+, g_d(\boldsymbol{S}, \boldsymbol{\theta}_{-d}))$, where $\boldsymbol{\theta}_{-d}$ is the vector $\boldsymbol{\theta}$ but excluding θ_d, so that $f(\theta_d | \boldsymbol{\beta}_d^+, g_d(\boldsymbol{s}_{\text{obs}}, \boldsymbol{\theta}_{-d}))$ is as close as possible to the true conditional distribution $\pi(\theta_d | \boldsymbol{s}_{\text{obs}}, \boldsymbol{\theta}_{-d})$ of $\pi(\boldsymbol{\theta} | \boldsymbol{s}_{\text{obs}})$. The functions $g_d(\boldsymbol{S}, \boldsymbol{\theta}_{-d})$ indicate the combination of $\mathbf{S}$ and $\boldsymbol{\theta}_{-d}$ used in the regression model to determine the conditional distribution of θ_d, both as main effects, or interactions. Clearly the appropriate dependent variables will vary with d, but will typically be relatively low dimensional (see the analyses in Section 5.3 for a guide on how these may be selected). The approximate Gibbs sampler will then cycle through each of these conditional distributions in turn, drawing $\theta_d \sim f(\theta_d | \hat{\boldsymbol{\beta}}_d^+, \boldsymbol{s}_{\text{obs}}, \boldsymbol{\theta}_{-d})$ for $d = 1, \ldots, D$, conditioning on $\boldsymbol{s} = \boldsymbol{s}_{\text{obs}}$. If $f(\theta_d | \hat{\boldsymbol{\beta}}_d^+, \boldsymbol{s}_{\text{obs}}, \boldsymbol{\theta}_{-d}) = \pi(\theta_d | \boldsymbol{s}_{\text{obs}}, \boldsymbol{\theta}_{-d})$ then the resulting Gibbs sampler will exactly target $\pi(\boldsymbol{\theta} | \boldsymbol{s}_{\text{obs}})$. Otherwise, the resulting sampler will be an approximation (discussed further below). This procedure is outlined in Algorithm 6.

The algorithm begins similarly to many ABC algorithms, by drawing samples $\{(\boldsymbol{\theta}^{(i)}, \boldsymbol{s}^{(i)})\}_{i=1}^N$ from the predictive distribution $(\boldsymbol{\theta}^{(i)}, \boldsymbol{X}^{(i)}) \sim p(\boldsymbol{X} | \boldsymbol{\theta}) b(\boldsymbol{\theta})$ and computing $\boldsymbol{s}^{(i)} = S(\boldsymbol{X}^{(i)})$. In most standard ABC algorithms $b(\boldsymbol{\theta})$ is the prior distribution $\pi(\boldsymbol{\theta})$ or an importance sampling distribution. Then, a standard Gibbs sampler procedure is implemented by sampling each parameter in turn from an approximation to its full conditional distribution $\theta_d^{(m)} | (\boldsymbol{s}_{\text{obs}}, \boldsymbol{\theta}_{-d}) \sim f(\theta_d | \hat{\boldsymbol{\beta}}_d^+, g_d(\boldsymbol{s}_{\text{obs}}, \boldsymbol{\theta}_{-d}))$. These approximations are fitted using the pool of weighted samples $\{(\boldsymbol{\theta}^{(i)}, \boldsymbol{s}^{(i)}, w_d^{(i)})\}_{i=1}^N$, where the weights $w_d^{(i)} \propto K_h(\|(\boldsymbol{s}^{(i)}, \boldsymbol{\theta}_{-d}^{(i)}) - (\boldsymbol{s}_{\text{obs}}, \boldsymbol{\theta}_{-d}^\star)\|) \pi(\boldsymbol{\theta}) / b(\boldsymbol{\theta})$ ensure that higher importance is given to those samples which more closely match both the observed data $\boldsymbol{s}_{\text{obs}}$ and the conditioned values of the parameters $\boldsymbol{\theta}_{-d} = \boldsymbol{\theta}_{-d}^\star$.

Clearly it is important that consideration be given to appropriate scaling of summary statistics and parameter values within the distance measure $\|\cdot\|$ to avoid one or other dominating the comparison. Note also that it is only required that the full conditionals are estimated well in regions of high posterior density, rather than over the entirety of

the support of $\boldsymbol{\theta}$. In this manner, the importance density $b(\boldsymbol{\theta})$ can be chosen to place $\boldsymbol{\theta}^{(i)}$ samples in regions where the conditional distributions need to be well approximated, which may be a much smaller region than specified by the prior $\pi(\boldsymbol{\theta})$. One such strategy was successfully adopted by Fearnhead and Prangle (2012) who specified $b(\boldsymbol{\theta})$ as proportional to the prior $\pi(\boldsymbol{\theta})$ but restricted to a region of high posterior density as identified by a pilot simulation.

Any appropriate regression technique can be used to construct the models $f(\theta_d|\boldsymbol{\beta}_d^+, g_d(\boldsymbol{s}_{\text{obs}}, \boldsymbol{\theta}_{-d}))$ such as non-parametric models, GLMs, neural networks, semi-parametric models, lasso etc. There are two possible ways to draw samples from each conditional regression model, as in step 2.2.4 in Algorithm 6. The first is when a parametric error distribution has been assumed, in which case a new sample may be drawn directly from the fitted distribution. For example, if the regression model is specified such that $\theta_d \sim N(\hat{\mu}, \hat{\sigma}^2)$ for specified $\hat{\mu}$ and $\hat{\sigma}^2$, then a new value of θ_d may be drawn directly from $N(\hat{\mu}, \hat{\sigma}^2)$. Alternatively, when a parametric error distribution is not assumed, the (weighted) distribution of empirical residuals $r_d^{(i)} = \theta_d^{(i)} - \hat{\mu}$ can be constructed as $R_d^N(r) = \sum_{i=1}^N w_d^{*(i)} \delta_{r_d^{(i)}}(r)$ where $w_d^{*(i)} = w_d^{(i)} / \sum_{j=1}^N w_d^{(j)}$, and $\delta_Z(z)$ is the Dirac measure, defined as $\delta_Z(z) = 1$ if $z \in Z$ and $\delta_Z(z) = 0$ otherwise. A new value of θ_d is then given by $\theta_d = \hat{\mu} + r$ where $r \sim R_d^N(r)$.

The computational overheads in Algorithm 6 are in the initial data simulation stage (steps 1.1–1.3) which is standard in many ABC algorithms, and in the fitting of a separate regression model for each parameter θ_d in each stage of the Gibbs sampler (steps 2.2.2–2.2.3). In the case of the latter, while it can be computationally cheap to fit any one regression model, repeating this ND times during sampler implementation can clearly raise the computational burden. There are two approaches that can reduce these costs, which can be implemented either separately or simultaneously.

In certain cases, the model $p(\boldsymbol{\theta}|\boldsymbol{s}_{\text{obs}})$ will have a structure such that several of the model parameters will have exactly the same form of full conditional distribution $\pi(\theta_d|\boldsymbol{s}_{\text{obs}}, \boldsymbol{\theta}_{-d})$. One such example is a hierarchical model (see section 5.3.2) where $x_{dj} \sim p(x|\theta_d)$ for $j = 1, \ldots, n_d$, and $\theta_1, \ldots, \theta_{D-1} \sim N(\theta_D, \sigma^2)$. Here the form of $\pi(\theta_d|\boldsymbol{s}_{\text{obs}}, \boldsymbol{\theta}_{-d})$ is identical for $d = 1, \ldots, D-1$. Accordingly the regression model $f(\theta_d|\boldsymbol{\beta}_d^+, g_d(\boldsymbol{S}, \boldsymbol{\theta}_{-d}))$ can be fitted by pooling the weighted samples $\{(\boldsymbol{\theta}^{(i)}, \boldsymbol{s}^{(i)}, w_d^{(i)})\}_{i=1}^N$ for $d = 1, \ldots, D-1$ (each using different sub-elements of the vectors), thereby allowing computational savings in allowing the value of N to be reduced. Further, in the case where the conditional independence

graph structure of the posterior is known (again, consider the hierarchical model), then the choice of which elements of $\boldsymbol{\theta}_{-d}$ should be included within the regression function $g_d(\boldsymbol{S}, \boldsymbol{\theta}_{-d})$ is immediately specified as the neighbours of θ_d on the conditional independence graph, and this does not then require independent elicitation.

A second approach is to choose the regression model $f(\theta_d|\boldsymbol{\beta}_d^+, g_d(\boldsymbol{S}, \boldsymbol{\theta}_{-d}))$ sufficiently flexibly so that not only is it a good approximation of $\pi(\theta_d|\boldsymbol{s}_{\text{obs}}, \boldsymbol{\theta}_{-d})$ when $\boldsymbol{\theta}_{-d}$ is fixed at a particular value, $\boldsymbol{\theta}_{-d}^\star$, within the Gibbs sampler, but that the regression model holds globally for any $\boldsymbol{\theta}_{-d}$. Within Algorithm 6, the approximation of $\pi(\theta_d|\boldsymbol{s}_{\text{obs}}, \boldsymbol{\theta}_{-d}^\star)$ with $\boldsymbol{\theta}_{-d} = \boldsymbol{\theta}_{-d}^\star$ is achieved by weighting the $\{(\boldsymbol{\theta}^{(i)}, \boldsymbol{s}^{(i)})\}_{i=1}^N$ samples in the region of $\boldsymbol{\theta}_{-d}^\star$ according to step 2.2.2. If the regression model $f(\theta_d|\boldsymbol{\beta}_d^+, g_d(\boldsymbol{S}, \boldsymbol{\theta}_{-d}))$ was a good approximation of $\pi(\theta_d|\boldsymbol{s}_{\text{obs}}, \boldsymbol{\theta}_{-d})$ for any value of $\boldsymbol{\theta}_{-d}$ (in the region of high posterior density), then the $\boldsymbol{\theta}_{-d}^\star$ specific weighting of step 2.2.2 can be removed, all samples weighted as $w^{(i)} \propto K_h(|\boldsymbol{s}^{(i)} - \boldsymbol{s}_{\text{obs}}\|)\pi(\boldsymbol{\theta})/b(\boldsymbol{\theta})$, thereby localising on summary statistics only, and the regression models fitted prior to implementing the Gibbs sampler. This global model likelihood-free approximate Gibbs sampler is described in Algorithm 7. Clearly the computational overheads of Algorithm 7 are substantially lower than for the localised model version. However, the localised version can be expected to be more accurate in practice, precisely due to the localised approximation of the full conditional distributions, and the difficulty in deriving sufficiently accurate global regression models.

In certain circumstances it can be seen that the likelihood-free approximate Gibbs sampler will exactly target the true partial posterior $\pi(\boldsymbol{\theta}|\boldsymbol{s}_{\text{obs}})$. In the case where the true conditional distributions $\pi(\theta_d|\boldsymbol{s}_{\text{obs}}, \boldsymbol{\theta}_{-d})$ are nested within the family of distributions described by $f(\theta_d|\boldsymbol{\beta}_d^+, g_d(\boldsymbol{S}, \boldsymbol{\theta}_{-d}))$, then as $N \to \infty$, which in turn allows $h \to 0$, then

$$f(\theta_d|\hat{\boldsymbol{\beta}}_d^+, g_d(\boldsymbol{S}, \boldsymbol{\theta}_{-d})) \to \pi(\theta_d|\boldsymbol{s}_{\text{obs}}, \boldsymbol{\theta}_{-d})$$

due to the law of large numbers ($N \to \infty$) and $h \to 0$ eliminating the usual local ABC approximation error. In this case, then Algorithms 6 and 7 will be exact. In any other cases, $f(\theta_d|\hat{\boldsymbol{\beta}}_d^+, g_d(\boldsymbol{S}, \boldsymbol{\theta}_{-d}))$ will be an approximation of $\pi(\theta_d|\boldsymbol{s}_{\text{obs}}, \boldsymbol{\theta}_{-d})$. This can be either a strong approximation, whereby $f(\theta_d|\boldsymbol{\beta}_d^+, g_d(\boldsymbol{S}, \boldsymbol{\theta}_{-d}))$ can exactly describe $\pi(\theta_d|\boldsymbol{s}_{\text{obs}}, \boldsymbol{\theta}_{-d})$ but where $\hat{\boldsymbol{\beta}}_d^+$ has not converged to $\boldsymbol{\beta}_d^+$ (i.e. finite N), or a weak approximation, whereby $\pi(\theta_d|\boldsymbol{s}_{\text{obs}}, \boldsymbol{\theta}_{-d})$ is not nested within the family $f(\theta_d|\boldsymbol{\beta}_d^+, g_d(\boldsymbol{S}, \boldsymbol{\theta}_{-d}))$, and

so $f(\theta_d|\hat{\boldsymbol{\beta}}_d^+, g_d(\boldsymbol{S}, \boldsymbol{\theta}_{-d}))$ represents the closest approximation to $\pi(\theta_d|\boldsymbol{s}_{\text{obs}}, \boldsymbol{\theta}_{-d})$ available within the regression model's functional constraints. This latter (weak) approximation can be arbitrarily good or poor.

When the fitted regression models only approximate the true posterior conditionals, then these may be *incompatible* in the sense that the set of approximate conditional distributions may not imply a joint distribution that is unique or even exists. This is equally a criticism of the ABC-MCMC sampler of Kousathanas et al. (2016) as it is of the likelihood-free approximate Gibbs sampler, unless for the former it can be guaranteed that the subset of summary statistics used to update θ_d in an ABC Metropolis-Hastings update step is sufficient for the full conditional distribution. See e.g. Arnold et al. (1999) for a comprehensive treatment of conditional specification of statistical models.

Incompatible conditional distributions are commonly encountered in the area of multivariate imputation by chained equations (MICE) also known as fully conditional specification (FCS), which is specifically designed for incomplete data problems (van Buuren and Groothuis-Oudshoorn, 2011). In the simplified case of multivariate conditional distributions within exponential families, Arnold et al. (1999) found that determining appropriate constraints on the model parameters to ensure a valid joint density was "daunting" and often "unattainable." Other authors have expressed uncertainty on the effects of incompatibility, although simulation studies have suggested that the problem may not be serious in practice (van Buuren and Groothuis-Oudshoorn, 2011; van Buuren et al., 2006; Drechsler and Rassler, 2008). Chen and Ip (2015) have investigated the behaviour of the Gibbs sampler when the conditional distributions are potentially incompatible.

However, in a more general study of parameterisation within Bayesian modelling, Gelman (2004) proposes the use of "inconsistent conditional distributions" as a new class of models, motivated by computational and analytical convenience in order to bypass the limitations of joint models.

5.3 Simulation studies

We examine the performance of the likelihood-free approximate Gibbs sampler in two simulation studies: a Gaussian mixture model using global regression models, and in a simple hierarchical model with both local and global regression models.

5.3.1 A Gaussian mixture model

We consider the D-dimensional Gaussian mixture model of Nott et al. (2014) where

$$p(\boldsymbol{s}|\boldsymbol{\theta}) = \sum_{b_1=0}^{1} \dots \sum_{b_D=0}^{1} \left[\prod_{i=1}^{D} \omega^{1-b_i}(1-\omega)^{1-b_i} \right] \phi_D(\boldsymbol{s}|\boldsymbol{\mu}(\boldsymbol{b},\boldsymbol{\theta}), \boldsymbol{\Sigma}),$$

where $\phi_D(\boldsymbol{x}|\boldsymbol{a},\boldsymbol{B})$ denotes the multivariate Gaussian density with mean $\boldsymbol{a}$ and covariance $\boldsymbol{B}$ evaluated at $\boldsymbol{x}$, $\omega \in [0,1]$ is a mixture weight, $\boldsymbol{\mu}(\boldsymbol{b},\boldsymbol{\theta}) = ((1-2b_1)\theta_1, \dots, (1-2b_D)\theta_D)^\top$, $\boldsymbol{b} = (b_1, \dots, b_D)^\top$ with $b_i \in \{0,1\}$, and $\boldsymbol{\Sigma} = [\Sigma_{ij}]$ is such that $\Sigma_{ii} = 1$ and $\Sigma_{ij} = \rho$ for $i \neq j$. For illustration we consider the $D = 2$ dimensional case, with $\boldsymbol{s}_{\text{obs}} = (5/2, 5/2)^\top$, fix $\omega = 0.3$ and $\rho = 0.7$ as known constants and specify $\pi(\theta_d)$ as $U(-20, 40)$ for $d = 1, 2$.

In this setting, the full conditional distributions for θ_1 and b_1 are given by

$$\begin{aligned}
\theta_1|(\theta_2,\boldsymbol{b},\boldsymbol{s}) &\sim N(\mu_{\theta_1}, \sqrt{1-\rho^2})I(-20 < \theta_1 < 40), \\
\mu_{\theta_1} &= s_1 - \rho s_2 + \rho\theta_2 - 2s_1b_1 + 2\rho s_2 b_1 - 2\rho b_1\theta_2 - 2\rho\theta_2 b_2 + 4\rho b_1 b_2 \theta_2 \qquad (5.2) \\
b_1|(\boldsymbol{\theta},b_2,\boldsymbol{s}) &\sim \text{Bernoulli}(L(p_{b_1})), \\
p_{b_1} &= \ln\left(\frac{1-\omega}{\omega}\right) + \frac{-2}{1-\rho^2}s_1\theta_1 + \frac{2\rho}{1-\rho^2}s_2\theta_1 + \frac{-2\rho}{1-\rho^2}\theta_1\theta_2 + \frac{4\rho}{1-\rho^2}b_2\theta_1\theta_2,
\end{aligned}$$

where $L(x) = 1/(1+\exp(-x))$ denotes the logistic function, with the full conditional distributions for θ_2 and b_2 obtained by switching the indices in the above. For this simple model we construct global regression models (Algorithm 7). We generate $N = 1,000,000$ samples from the prior predictive distribution (i.e. with $b(\boldsymbol{\theta}) = \pi(\boldsymbol{\theta})$) and specify $K_h(u)$ as the uniform kernel ($h = \infty$).

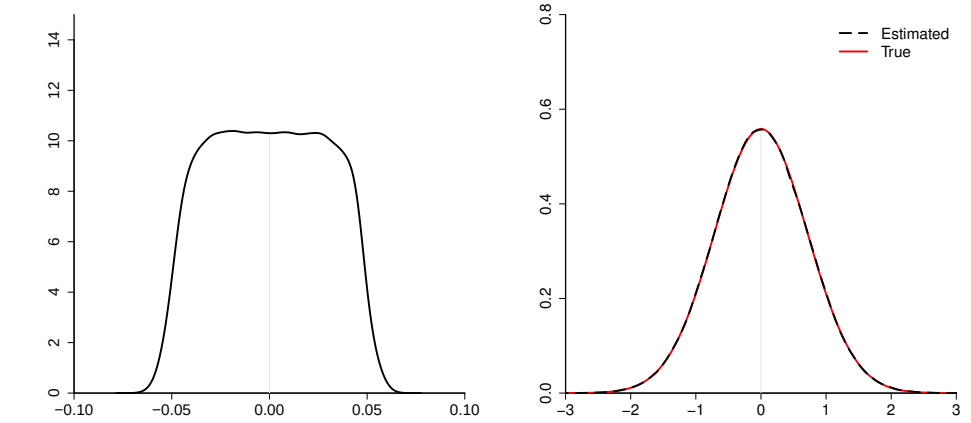
As an illustration, we first naively attempt to approximate the full conditional distribution of θ_1 by a main-effects only (excluding $\boldsymbol{b}$) Gaussian regression model $\theta_1|(\theta_2,\boldsymbol{b},\boldsymbol{s}) \sim N(\beta_0 + \beta_1 s_1 + \beta_2 s_2 + \beta_3\theta_2, \sigma^2)$. The resulting MLEs were $\hat{\boldsymbol{\beta}} = (8.76, -0.31, 0, 0)^\top$ (s.e. $= (0.019, 0.001, 0.001, 0.001)$) and $\hat{\sigma} = 16.16$, which suggests that θ_1 is conditionally independent of s_2 and θ_2. This can clearly be seen to be incorrect based on a simple graphical exploration of the synthetic samples. This is a clear warning of the need to consider sufficiently flexible regression models, with interaction effects (as discussed in Nott et al. (2014) and as is evident in the form of μ_{θ_1}). Instead, we specify the regression mean with all main effects and interactions and, because the number of samples N is large, the resulting MLEs of $\boldsymbol{\beta}$ (and σ^2) matched the true values in (5.2) up to at least one decimal

place.

Figure 5.1a shows a kernel density estimate (KDE) of the differences between the fitted and true conditional mean values $(\hat{\mu}_{\theta_1}^{(i)} - \mu_{\theta_1}^{(i)})$ for each of the N data points used in the regression. In most cases, the absolute difference was less than 0.05. Figure 5.1b shows a KDE of the empirical residuals and the true $N(0, \sqrt{1-\rho^2})$ error density. The similarity suggests that in sampling from the regression model, randomly choosing a residual is essentially equivalent to sampling from the true Gaussian error distribution. Given that we are fitting a regression model in the same family as the true conditional distribution, we have a strong approximation of $\theta_1|(\theta_2, \boldsymbol{b}, \boldsymbol{s})$ (as defined in section 5.2) in this case.

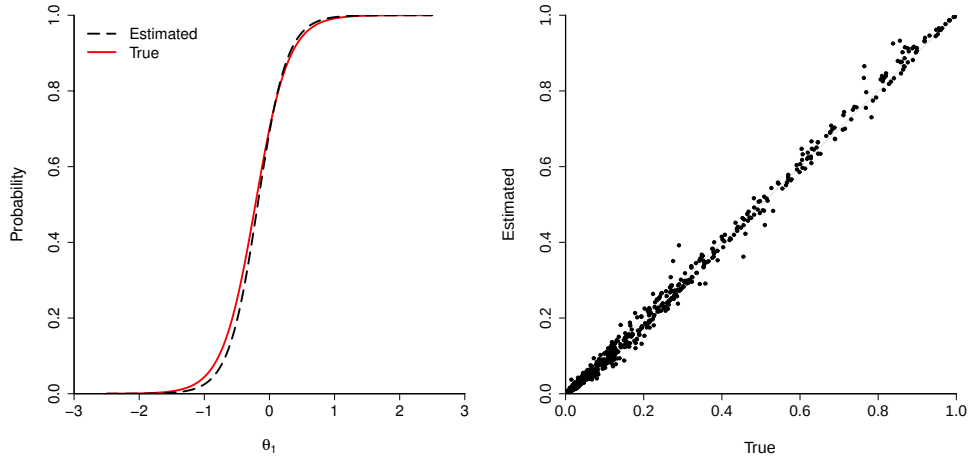
In a similar manner, we naturally model the conditional distribution of $b_1|(\boldsymbol{\theta}, b_2, \boldsymbol{s})$ as a Bernoulli GLM with logistic link function, and all possible conditional main effects and interactions. Figure 5.1c examines the quality of this approximation by presenting the cdf's of the fitted and the true probabilities of $p(b_1 = 1|\theta_1, s_1 = s_2 = 2.5, b_2 = 0, \theta_2 = -2.5)$. The distributions are very similar, though still distinguishable. An explanation for this is that for most of the N samples, the conditional probability of b_1 is either (numerically) 0 or 1. In other words, only the samples such that $\boldsymbol{\theta}$ is close to the origin are informative for the regression parameters. This regression model is again a strong approximation to the true conditional distribution.

Figure 5.2 illustrates the output of $M = 20,000$ iterations of the resulting likelihood-free approximate Gibbs sampler, when initialised at $(\theta_1, \theta_2, b_1, b_2) = (0, -10, 1, 0)$. The sampler moves around the parameter space well, and visually appears to target the true posterior distribution. During sampler implementation the true and estimated probabilities of switching the value of b_1 were recorded, and are illustrated in Figure 5.1d. Only a small proportion of the M probabilities are larger than 0.2 (due to the form of the posterior), but on the whole the estimated probabilities are generally accurate, with a few exceptions. For this example, the estimated conditional distributions (and associated switching probabilities of b_1) will approximate their true counterparts arbitrarily well as N gets large, essentially due to the simple form of the true posterior distribution. A better mixing approximate Gibbs sampler could also have been constructed for this posterior distribution, using a 4-level multinomial regression for the full conditional of $\boldsymbol{b}|(\boldsymbol{\theta}, \boldsymbol{s})$ and a bivariate Gaussian regression model for $\boldsymbol{\theta}|(\boldsymbol{b}, \boldsymbol{s})$.



(a) KDE of diff. between fitted and true means

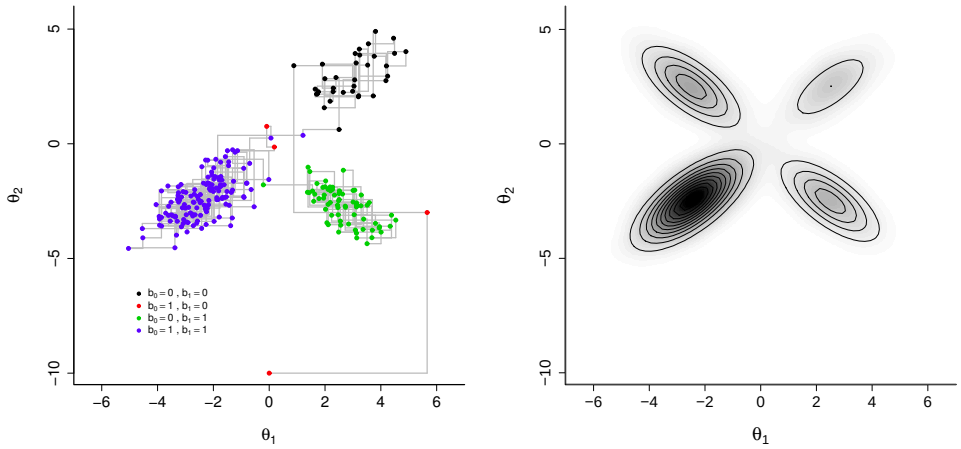
(b) Residual error distribution



(c) $p(b_1 = 1|\theta_1, s_1 = s_2 = 2.5, b_2 = 0, \theta_2 = -2.5)$

(d) Probability of changing the state of b_1

Figure 5.1: Assessing the quality of the regression approximation. (a) A kernel density estimate (KDE) of the differences between the fitted and true conditional mean values $\hat{\mu}_{\theta_1}^{(i)} - \mu_{\theta_1}^{(i)}$. (b) The true $N(0, \sqrt{1-\rho^2})$ error density and the KDE of the fitted regression residuals. (c) The fitted and the true conditional distribution $p(b_1 = 1|\theta_1, s_1 = s_2 = 2.5, b_2 = 0, \theta_2 = -2.5)$. (d) True versus estimated probability of changing the the state of the cluster indicator variable b_1.

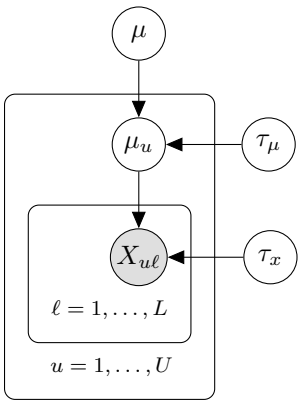


(a) First 150 Gibbs iterations (b) Posterior distribution

Figure 5.2: Likelihood-free approximate Gibbs sampler output. (a) Sample path of the first 250 iterations of (θ_1, θ_2), with the values of (b_1, b_2) indicated by coloured points. (b) Posterior density estimates (shading) based on 20,000 sampler iterations, and true posterior density contours.

5.3.2 A simple hierarchical model

We now contrast the performance of the approximate Gibbs sampler when using local and global regression models for a simple Gaussian hierarchical model. Hierarchical methods have been previously considered in the likelihood-free framework by Bazin et al. (2010) and Rodrigues et al. (2016). The simple hierarchical model is defined as



$$
\begin{aligned}
X_{u\ell} &\sim N(\mu_u, \tau_x^{-1}) \qquad (5.3)\\
\mu_u &\sim N(\mu, \tau_\mu^{-1})\\
\tau_x &\sim \text{Gamma}(\alpha_x, \nu_x)\\
\tau_\mu &\sim \text{Gamma}(\alpha_\mu, \nu_\mu)\\
\mu &\sim N(0, 1),
\end{aligned}
$$

where $X_{u\ell}$ denotes the ℓ-th observation in group u, for $\ell = 1, \ldots, L$ and $u = 1, \ldots, U$, and ν_μ and ν_y are rate parameters. This model is fully tractable and allows a direct comparison between the exact and approximate posterior distributions.

To estimate the $U + 3$ parameters $\boldsymbol{\theta} = (\mu, \mu_1, \ldots, \mu_U, \tau_\mu, \tau_x)^\top$ of this model, we adopt the summary statistics $\boldsymbol{S} = (\overline{X}_1, \ldots, \overline{X}_U, \hat{\sigma}_1^{-1}, \ldots, \hat{\sigma}_U^{-1}, \overline{\overline{X}}, \bar{\sigma}^{-1})^\top$, where $\overline{X}_u$ and $\hat{\sigma}_u$ are the sample mean and variance of the data in group u, respectively, $\overline{\overline{X}}$ is the overall sample mean

and $\bar{\sigma}$ denotes the mean of the sample variances $\hat{\sigma}_1, \ldots, \hat{\sigma}_U$. To implement the approximate Gibbs sampler we specify simple and non-exact regression model approximations of the full conditional distributions.

The true full conditional distribution for μ is given by

$$\pi(\mu|\ldots) \sim N\left(\frac{U\tau_\mu \overline{\mu}}{1+U\tau_\mu}, (1+U\tau_\mu)^{-1}\right),$$

and so the conditional mean $E(\mu|\ldots)$ is a nonlinear function of τ_μ and $\overline{\mu} = (1/U)\sum_{u=1}^{U}\mu_u$, and the conditional variance is not constant throughout the covariate space. From the models' DAG, it is immediate that given $(\mu_1, \ldots, \mu_U, \tau_\mu)$, μ is independent of all other nodes in the graph, including the observed data. This permits the immediate discarding of several uninformative nodes within $g_\mu(\cdot)$. Where available, exploring the model structure to reduce the complexity of the regression models can be very valuable. Note that it is only necessary to approximate the full conditional distribution of parameters that are conditionally dependent on intractable quantities. Here we approximate $\pi(\mu|\ldots)$ by assuming a simple linear and homoscedastic regression model of the form

$$\mu|(\boldsymbol{S}, \boldsymbol{\theta}_{-\mu}) \sim N\left((1, \tau_\mu, \overline{\mu})^\top \boldsymbol{\beta}_\mu, V_\mu\right).$$

The distribution of each unit mean $\mu_1, \ldots, \mu_U$ depends on the data exclusively through the corresponding unit-specific summary statistics. For $u = 1, \ldots, U$,

$$\pi(\mu_u|\ldots) \sim N\left(\frac{\mu\tau_\mu + L\tau_x\overline{X}_u}{\tau_\mu + L\tau_x},\ (\tau_\mu + L\tau_x)^{-1}\right)$$

is the conditional distribution of μ_u, which we approximate by the single regression model

$$\mu_u|(\boldsymbol{S}, \boldsymbol{\theta}_{-\mu_u}) \sim N\left((1, \mu, \tau_\mu, \tau_x, \overline{X}_u, \hat{\sigma}_u^{-1})^\top \boldsymbol{\beta}_{\mu_u}, V_{\mu_u}\right).$$

As the units are exchangeable one can use all synthetic observations (irrespective the unit label) when fitting this single regression model for all groups u. This is because a synthetic sample $(\mu, \tau_\mu, \tau_x, \overline{X}_u, \hat{\sigma}_u^{-1})$ is as informative for $\hat{\boldsymbol{\beta}}_{\mu_u}$ as it is to the regression parameters of any other group u'. This strategy takes advantage of this symmetric nature of the DAG to increase the accuracy of the regression parameter estimates.

The true conditional distribution of τ_μ is given by

$$\pi(\tau_\mu|\ldots) \sim \text{Gamma}\left(\alpha_\mu + \frac{U}{2},\ \nu_\mu + \frac{\sum_{u=1}^{U}(\mu_u - \mu)^2}{2}\right).$$

To maintain the support of τ_μ on $\mathbb{R}$, we specify the regression model approximation as

$$\log(\tau_\mu)|(\boldsymbol{S}, \boldsymbol{\theta}_{-\tau_\mu}) \sim N\left((1, \mu, \overline{\mu}, \hat{\sigma}_\mu^{-1})^\top \boldsymbol{\beta}_{\tau_\mu}, V_{\tau_\mu}\right),$$

where $\hat{\sigma}_\mu = \frac{1}{U-1}\sum_{u=1}^{U}(\mu_u - \overline{\mu})^2$. This distribution is again conditionally independent of the observed data. We deliberately chose to work with the normal model as opposed to a more appropriate gamma GLM, to further compromise the validity of the model assumptions and better assess the approximate Gibbs sampler's robustness to misspecification.

Finally, we similarly model the true distribution of the precision parameter τ_x,

$$\pi(\tau_x|\ldots) \sim \text{Gamma}\left(\alpha_x + \frac{UL}{2},\ \nu_x + \frac{\sum_{u=1}^{U}\sum_{\ell=1}^{L}(X_{u\ell} - \mu_u)^2}{2}\right),$$

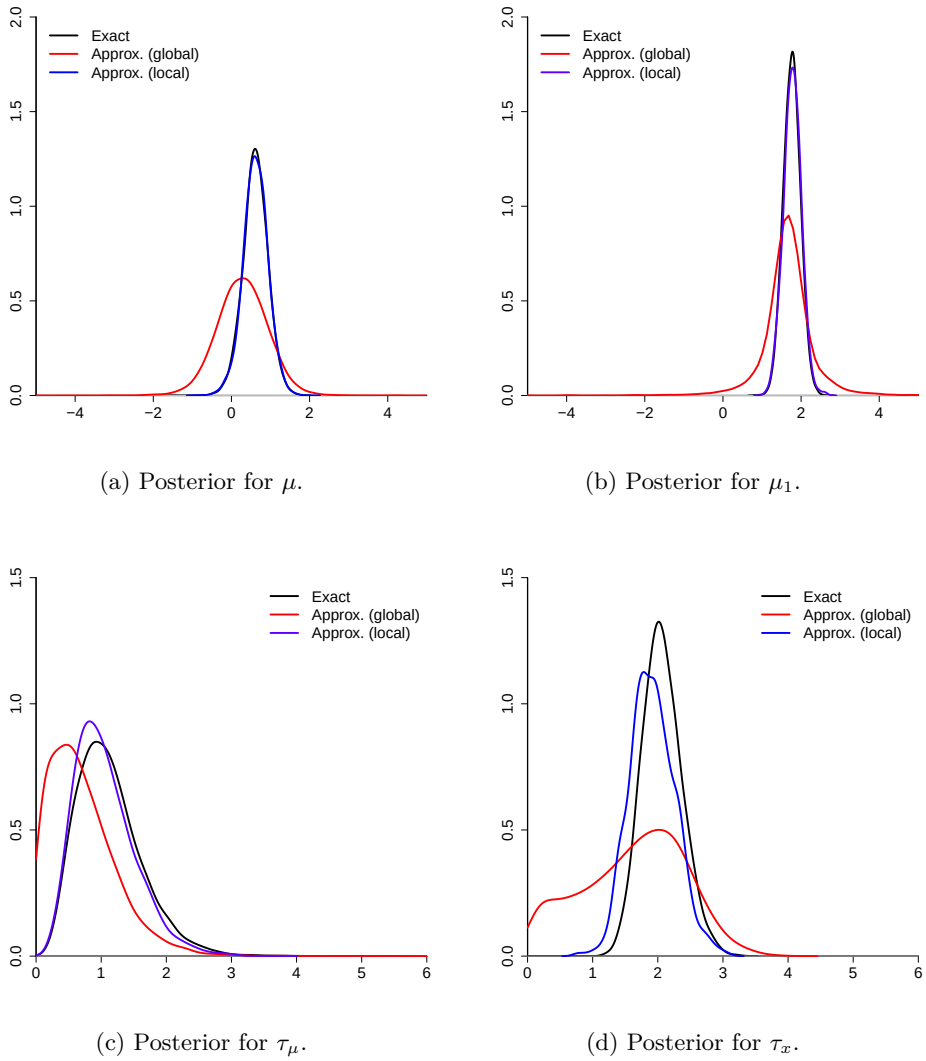
by the approximation

$$\log(\tau_x)|(\boldsymbol{S}, \boldsymbol{\theta}_{-\tau_x}) \sim N\left((1, \overline{\overline{X}}, \bar{\sigma}^{-1})^\top \boldsymbol{\beta}_{\tau_x}, V_{\tau_x}\right).$$

For simulated data generated with $U = L = 10$ and $\alpha_\mu = \nu_\mu = \alpha_x = \nu_x = 1$, we draw $N = 100,000$ samples from the prior predictive distribution and a subsequent $M = 5,000$ samples from the approximate Gibbs sampler, with the chain initialised at the sample estimate of each parameter (i.e., the corresponding entries of the observed summary statistics). We implement both Algorithm 6 in which the regression models are localised to best approximate the true conditional distributions at each stage of the Gibbs sampler, and Algorithm 7 in which the models are localised on $\boldsymbol{s} = \boldsymbol{s}_{\text{obs}}$ only at the start of the algorithm, and then are assumed to hold globally. K_h was chosen to be uniform, with h determined to select the best $1,000$ samples for each localisation.

Figure 5.3 presents the marginal posterior densities for μ, μ_1, τ_μ and τ_x obtained for the exact and approximate Gibbs samplers, with both global and local models. In each case, while the global regression model sampler (red lines) can broadly estimate the right location, it is clearly a poor approximation of the true marginal posterior density (black

lines). This is hardly surprising in light of the large differences between the exact conditional distributions and the chosen regression models. However, localising the regression models (blue lines) at each stage of the Gibbs sampler produces a major improvement in the quality of the approximation.



(a) Posterior for μ.

(b) Posterior for μ_1.

(c) Posterior for τ_μ.

(d) Posterior for τ_x.

Figure 5.3: Exact and approximate posteriors for (a) the overall mean μ; (b) the mean of group 1: μ_1; (c) the precision of the means τ_μ; (d) the precision of the observations within groups τ_x. Black line depicts the exact Gibbs posterior. Red and blue lines represent the approximate posterior densities obtained using global and local regression models, respectively.

Note that unlike the mixture example (section 5.3.1), the true conditional distributions are not contained within the family of the chosen regression models. Accordingly, the approximation errors will not decrease as the number of synthetic samples N gets large (and h gets small). However, the regression models could be improved by including interaction terms and transforming parameters to better suit the regression model assumptions.

In terms of computation, on a HP device with an Intel Core i7-4790 CPU (3.6GHz) with 16 GB of RAM, sampling the $N = 100,000$ synthetic samples took roughly 51 seconds to run in `R`. The approximate Gibbs sampling stage took 40 seconds and 22 minutes, respectively, when using the global and local regression models. This is a substantial computational increase when considering that the exact Gibbs sampler takes less than 0.5 seconds to execute.

5.4 A state space model of *Airbnb data*

We analyze a time series dataset containing Airbnb rental prices in the city of Seattle, WA, USA. The dataset, compiled on 4 January 2016, was made available at *kaggle.com*, and can also be accessed through *insideairbnb.com*. It consists of 928151 entries, each of which corresponding to an available listed space (property, room, etc) at a given date (spanning the year of 2016). A fundamental feature of this dataset is its non-adherence to the normal distribution – even after taking the logarithm, the prices fails to be well approximated by any symmetric distribution.

The so-called g-and-k distribution (Haynes, 1998) appears as an attractive, more suitable alternative. The g-and-k is characterized by its cumulative distribution function (CDF), defined by

$$Q(q|A,B,g,k) = A + B\left[1 + c\frac{1-\exp\{-gz(q)\}}{1+\exp\{-gz(q)\}}\right](1+z(q)^2)^k z(q),$$

for $B > 0$ and $k > -0.5$, where $z(q)$ denotes the q-th quantile of the standard Gaussian distribution. We follow Rayner and MacGillivray (2002) and set $c = 0.8$. Due to its ability to accommodate different combinations of location, scale, skewness and kurtosis, this 4-parameters distribution has recently gained popularity, particularly in the ABC literature (Drovandi and Pettitt, 2011; Fearnhead and Prangle, 2012). Nevertheless, the g-and-k has no closed form density, and evaluating the likelihood is an expensive computational exercise. Peters et al. (2016) introduced a fast L-moments estimator for the parameters of the g-and-k, enabling these authors to study an insurance dataset similar in structure to the one considered here. Their approach, however, involved fitting an independent g-and-k for each time step, ignoring potential longitudinal correlations.

We illustrate how to apply our approximate Gibbs method to fit an intractable non-

linear state space model that, for each time step, the observations are assumed to follow a g-and-k distribution. The general form of such model is defined as follows.

$$\text{Observation distribution:} \quad \boldsymbol{y}_t \sim gk(\boldsymbol{\beta}_t), \quad t = 1, \ldots, T \tag{5.4a}$$

$$\text{Link function:} \quad g(\boldsymbol{\beta}_t) = \boldsymbol{\lambda}_t = \boldsymbol{F}_t'\boldsymbol{\theta}_t \tag{5.4b}$$

$$\text{System equation:} \quad (\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}) = \boldsymbol{G}_t\boldsymbol{\theta}_{t-1} + \boldsymbol{w}_t, \quad \boldsymbol{w}_t \sim N(\boldsymbol{0}, \boldsymbol{W}_t) \tag{5.4c}$$

$$\text{Prior distribution:} \quad \boldsymbol{\theta}_0 \sim N(\boldsymbol{m}_0, \boldsymbol{C}_0), \tag{5.4d}$$

where $\boldsymbol{y}_t$ denotes the vector of (log) prices observed at time t, $\boldsymbol{F}_t$ is a known $p \times 4$ design matrix that maps the state vector $\boldsymbol{\theta}_t$ to the linear predictor $\boldsymbol{\lambda}_t$, $\boldsymbol{G}_t$ is a known $p \times p$ evolution matrix that dictates the system's dynamics, $\boldsymbol{W}_t$ is a possibly unknown covariance matrix, and $\boldsymbol{\beta}_t = (\lambda_{1,t}, \exp(\lambda_{2,t}), \lambda_{3,t}, \exp(\lambda_{4,t}) - 0.5) = (A, B, g, k)_t$ represents the parameters of the g-and-k. The link function $g(\cdot)$ ensures that $\boldsymbol{\beta}_t$ respects the constraints imposed by the observation distribution. We assume that given $\boldsymbol{\theta}_t$, the observations $\boldsymbol{y}_t$ are independent and identically distributed. The sequence of errors $\boldsymbol{w}_t$ are also assumed to be independent.

State space models provide a flexible and well-structured framework to probabilistically describe an extensive array of applied problems. For a comprehensive treatment of Bayesian state space models, also known as Dynamic models(see West and Harrison, 1997). Petris (2010) developed the useful *dlm* R package that makes available several functions that facilitate the practical use of Gaussian linear state space models, otherwise known as Dynamic linear models. West et al. (1985), in turn, introduced the so-called Dynamic generalized linear models, which relaxes the linearity and Gaussian assumptions, allowing the observations to follow other members of the exponential family. Other works have focused on specific observation distributions, such as the Beta Da-Silva et al. (2011) and the Dirichlet Da-Silva and Rodrigues (2013).

Computational hurdles have limited the use of intractable dynamic models such as the one considered here, but an increasing effort to tackle this issue has been recently observed (Jasra et al., 2012; Dean et al., 2014; Martin et al., 2014; Calvet and Czellar, 2012; Yildirim et al., 2013; Picchini and Samson, 2016; Martin et al., 2016).

In this study, the prior distribution for $\boldsymbol{\theta}_0$ is set as $\boldsymbol{m}_0 = \boldsymbol{0}$ and $\boldsymbol{C}_0 = 10^7\mathbf{I}$, where $\boldsymbol{0}$ is a vector of zeros and $\mathbf{I}$ denotes the identity matrix. It is further assumed that $\boldsymbol{W}_t = \boldsymbol{W} = \text{diag}(1/\tau_1, \ldots, 1/\tau_p)$, with $\tau_i \sim \text{gamma}(\alpha = 10^{-10}, \nu = 10^{-10})$, for $i = 1, \ldots, p$.

That is, $\boldsymbol{W}_t$ is a static covariance matrix with the inverse of the diagonal elements (the precisions) following independent and identical gamma priors.

5.4.1 Inferential procedure

The joint distribution factorizes as

$$p(\boldsymbol{\Theta}, \boldsymbol{W}, \boldsymbol{y}_1, \ldots, \boldsymbol{y}_T) = p(\boldsymbol{\theta}_0)p(\boldsymbol{W}) \prod_{t=1}^{T} \left[p(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}, \boldsymbol{W})p(\boldsymbol{y}_t|\boldsymbol{\lambda}_t)\right], \tag{5.5}$$

where $\boldsymbol{\Theta} = (\boldsymbol{\theta}_0, \ldots, \boldsymbol{\theta}_T)$. The data only depends on the system state through $\boldsymbol{\lambda}_t$, so the full conditional distribution for $\boldsymbol{\theta}_t$ can be conveniently factorized as

$$\begin{aligned} p(\boldsymbol{\theta}_t|\cdot) &= p(\boldsymbol{\theta}_t|\boldsymbol{\lambda}_t, \cdot)p(\boldsymbol{\lambda}_t|\cdot) \\ &= p(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}, \boldsymbol{\theta}_{t+1}, \boldsymbol{W}, \boldsymbol{\lambda}_t)p(\boldsymbol{\lambda}_t|\boldsymbol{\theta}_{t-1}, \boldsymbol{\theta}_{t+1}, \boldsymbol{W}, \boldsymbol{y}_t). \end{aligned}$$

Therefore, one can sample from this distribution in two stages – first $\boldsymbol{\lambda}_t^* \sim p(\boldsymbol{\lambda}_t|\cdot)$ and then $\boldsymbol{\theta}_t^* \sim p(\boldsymbol{\theta}_t|\boldsymbol{\lambda}_t^*, \cdot)$. By design, the intractability only occurs on the rightmost term of Eq. 5.5, and with the exception of $p(\boldsymbol{\lambda}_t|\cdot)$, all FCDs allow efficient, exact sampling. While often a tedious task, deriving those distributions only requires basic matrix algebra operations. The FCDs are fully identified hereafter.

System's initial state $\boldsymbol{\theta}_0$

$$p(\boldsymbol{\theta}_0|\cdot) \sim N(\boldsymbol{a}_0, \boldsymbol{\Sigma}_0),$$

where $\boldsymbol{\Sigma}_0 = (\boldsymbol{G}_1'\boldsymbol{W}^{-1}\boldsymbol{G}_1 + \boldsymbol{C}_0^{-1})^{-1}$ and $\boldsymbol{a}_0 = \boldsymbol{\Sigma}_0(\boldsymbol{C}_0^{-1}\boldsymbol{m}_0 + \boldsymbol{G}_1'\boldsymbol{W}^{-1}\boldsymbol{\theta}_1)$.

System's future state $\boldsymbol{\theta}_{T+1}$

To facilitate sampling $\boldsymbol{\theta}_T$, we augmented the parameter space to keep track of the parameter $\boldsymbol{\theta}_{T+1}$, with FCD given by

$$p(\boldsymbol{\theta}_{T+1}|\cdot) \sim N(\boldsymbol{G}_{T+1}\boldsymbol{\theta}_T, \boldsymbol{W}).$$

Evolution error's precisions τ_i

$$p(\tau_i|\cdot) \sim \text{Gamma}\left(\alpha + \frac{T+1}{2}, \nu + \frac{\sum_{t=1}^{T+1} \boldsymbol{w}_{ti}^2}{2}\right),$$

where $\boldsymbol{w}_t = \boldsymbol{\theta}_t - \boldsymbol{G}_t\boldsymbol{\theta}_{t-1}$ represents the system innovation at time t.

System state $\boldsymbol{\theta}_t$

The model equations imply that

$$\begin{pmatrix} \boldsymbol{\theta}_t \\ \boldsymbol{\lambda}_t \end{pmatrix} \Bigg| \; \boldsymbol{\theta}_{t-1}, \boldsymbol{\theta}_{t+1}, \boldsymbol{W} \sim \text{N}\left[\begin{pmatrix} \boldsymbol{a}_t \\ \boldsymbol{f}_t \end{pmatrix}, \quad \begin{pmatrix} \boldsymbol{R}_t & \boldsymbol{R}_t F_t \\ F_t'\boldsymbol{R}_t & \boldsymbol{q}_t \end{pmatrix}\right],$$

where

$$\boldsymbol{f}_t = \boldsymbol{F}_t'\boldsymbol{a},$$

$$\boldsymbol{q}_t = F_t'\boldsymbol{R}_t F_t,$$

$$\boldsymbol{a}_t = \boldsymbol{R}_t(\boldsymbol{W}^{-1}\boldsymbol{G}_t\boldsymbol{\theta}_{t-1} + \boldsymbol{G}_{t+1}'\boldsymbol{W}^{-1}\boldsymbol{\theta}_{t+1}),$$

$$\boldsymbol{R}_t = (\boldsymbol{G}_{t+1}'\boldsymbol{W}^{-1}\boldsymbol{G}_{t+1} + \boldsymbol{W}^{-1})^{-1}.$$

Therefore, it follows from the conditional properties of the multivariate normal distribution that

$$p(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}, \boldsymbol{\theta}_{t+1}, \boldsymbol{W}, \boldsymbol{\lambda}_t) = N(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t),$$

where

$$\boldsymbol{\mu}_t = \boldsymbol{a}_t + \boldsymbol{R}_t\boldsymbol{F}_t\boldsymbol{q}_t^{-1}(\boldsymbol{\lambda}_t - \boldsymbol{f}_t),$$

$$\boldsymbol{\Sigma}_t = \boldsymbol{R}_t - \boldsymbol{R}_t\boldsymbol{F}_t\boldsymbol{q}_t^{-1}\boldsymbol{F}_t'\boldsymbol{R}_t.$$

Linear predictor $\boldsymbol{\lambda}_t$

The linear predictor's conditional distribution, $p(\boldsymbol{\lambda}_t|\boldsymbol{\theta}_{t-1}, \boldsymbol{\theta}_{t+1}, \boldsymbol{W}, \boldsymbol{y}_t)$, is intractable and needs to be approximated. To do so, the dimension of the conditioning set is drastically reduced by replacing the observed data $\boldsymbol{y}_t$ by the summary statistic $\boldsymbol{s}_t = g(\hat{\boldsymbol{\beta}}_t)$, where $\hat{\boldsymbol{\beta}}_t$ represents the L-moments estimator of $\boldsymbol{\beta}_t$ given $\boldsymbol{y}_t$ and $g(\cdot)$ is the link function previously defined. Although we cannot guarantee the summary statistics are sufficient, we believe it retain nearly all the relevant information available in the full dataset. In addition, as

stated in Peters et al. (2016), the L-moment estimators are nearly unbiased for all sample sizes and all distributions.

It is useful to recognize that $p(\boldsymbol{\lambda}_t|\boldsymbol{\theta}_{t-1}, \boldsymbol{\theta}_{t+1}, \boldsymbol{W}, \boldsymbol{s}_t) = p(\boldsymbol{\lambda}_t|\boldsymbol{\phi}_t, \boldsymbol{s}_t)$, where $\boldsymbol{\phi}_t = (\boldsymbol{f}_t, \boldsymbol{q}_t, n_t)$ and n_t is the sample size at time t. Because this structure is valid throughout the evolution period, the time label can be effectively dropped, which reduces the problem to approximating the distribution of a 4-dimensional vector conditional on 13 variables ($\boldsymbol{q}_t$ is a diagonal matrix).

Without loss of generality, we write

$$(\boldsymbol{\lambda}|\boldsymbol{\phi}, \boldsymbol{s}) = \boldsymbol{\mu}_\lambda + \boldsymbol{\Sigma}_\lambda^{1/2}\boldsymbol{\epsilon}_\lambda, \tag{5.6}$$

where $\boldsymbol{\mu}_\lambda$ and $\boldsymbol{\Sigma}_\lambda^{1/2}$, which are functions of $\boldsymbol{\phi}$ and $\boldsymbol{s}$, denote the mean and the (Cholesky) squared root of the covariance matrix of $(\boldsymbol{\lambda}|\boldsymbol{\phi}, \boldsymbol{s})$, respectively. $\boldsymbol{\epsilon}_\lambda$ follows an unknown standardized distribution (that may also depend on the conditioning list). Even without knowledge of the distribution of $\boldsymbol{\epsilon}_\lambda$, given the moments of the joint vector,

$$\left(\begin{array}{c|c} \boldsymbol{\lambda} & \\ \boldsymbol{s} & \boldsymbol{\phi} \end{array}\right) \sim \left[\begin{pmatrix} \boldsymbol{f} \\ \boldsymbol{f} \end{pmatrix}, \quad \boldsymbol{\Omega}_{\boldsymbol{\phi}} = \begin{pmatrix} \boldsymbol{\Omega}_{11} & \boldsymbol{\Omega}_{12} \\ \boldsymbol{\Omega}_{21} & \boldsymbol{\Omega}_{22} \end{pmatrix}\right],$$

Linear Bayes (Hartigan, 1969; Goldstein, 1976) can be employed to define the following estimators:

$$\hat{\boldsymbol{\mu}}_\lambda = \boldsymbol{f} + \boldsymbol{\Omega}_{12}\boldsymbol{\Omega}_{22}^{-1}(\boldsymbol{s} - \boldsymbol{f}) \tag{5.7a}$$

$$\hat{\boldsymbol{\Sigma}}_\lambda = \boldsymbol{\Omega}_{11} - \boldsymbol{\Omega}_{12}\boldsymbol{\Omega}_{22}^{-1}\boldsymbol{\Omega}_{21} \tag{5.7b}$$

For each time step in each Gibbs iteration, the ABC machinery allows the execution of a sequence of tasks that ultimately generate an approximate sample from $p(\boldsymbol{\lambda}|\boldsymbol{\phi}, \boldsymbol{y})$. These are: estimate the covariance matrix $\boldsymbol{\Omega}_{\boldsymbol{\phi}}$; compute the conditional moments in Eq. 5.7; draw an approximate sample from $\boldsymbol{\epsilon}_\lambda$; and plug-in the obtained values into Eq. 5.6.

We proceed as follows. Only once, before running the Gibbs sampler, we generated $M = 5000$ synthetic samples of $\boldsymbol{\phi}$ uniformly on a hypercube that roughly covers the region that might be visited during the Gibbs run. Precisely, the synthetic means $\boldsymbol{f}$ have the same range as observed in $\boldsymbol{s}_{\text{obs}}$, n spans the spectrum of actual sample sizes and the

diagonal elements of $\boldsymbol{q}$ are in the interval $(0, 1^{-5})$. Next, for each $\boldsymbol{\phi}^m = (\boldsymbol{f}^m, \boldsymbol{q}^m, n^m)$, $m = 1, \ldots, M$, we sampled $(\boldsymbol{\lambda}, \boldsymbol{s})^m \sim (\boldsymbol{\lambda}, \boldsymbol{s}|\boldsymbol{\phi}^m) = p(\boldsymbol{s}|\boldsymbol{\lambda}, n^m)N(\boldsymbol{\lambda}|\boldsymbol{f}^m, \boldsymbol{q}^m)$. Notice that, in contrast to the previous sections, we do not generate synthetic samples from the full generative process. Instead, these are drawn exclusively in a small portion of the model.

Throughout the approximate Gibbs sampler, the unknown matrix

$$\boldsymbol{\Omega}_{\boldsymbol{\phi}^*} = \mathrm{V}\left(\begin{array}{c|c} \boldsymbol{\lambda} & \\ \boldsymbol{s} & \boldsymbol{\phi}^* \end{array} \right) \approx \int \mathrm{V}\left(\begin{array}{c|c} \boldsymbol{\lambda} - \boldsymbol{f} & \\ \boldsymbol{s} - \boldsymbol{f} & \boldsymbol{q}, n \end{array} \right) K_h(\|\boldsymbol{\phi} - \boldsymbol{\phi}^*\|)p(\boldsymbol{\phi})d\boldsymbol{\phi}$$

is estimated for different states of $\boldsymbol{\phi}^*$. To do so, we suggest computing the weighted covariance matrix calculated over the centered synthetic samples $(\boldsymbol{\lambda}^m - \boldsymbol{f}^m, \boldsymbol{s}^m - \boldsymbol{f}^m)$, $m = 1, \ldots, M$. Weighting and centering the samples reduce the estimation error and allow us to have a unique set of synthetic samples (as opposed to generating a new set for each $\boldsymbol{\phi}^*$, what would be computationally prohibitive). For setting the weights, which are refreshed for each $\boldsymbol{\phi}^*$, we adopted the Epanechnikov kernel, with bandwidth chosen such that the best 2000 samples were accepted.

For $m = 1, \ldots, M$, we calculate $\boldsymbol{\Omega}_{\boldsymbol{\phi}^m}$ and compute $\hat{\boldsymbol{\mu}}_\lambda^m$ and $\hat{\boldsymbol{\Sigma}}_\lambda^m$ from Eq. 5.7. The empirical residual are then given by $\boldsymbol{\epsilon}_\lambda^m = (\hat{\boldsymbol{\Sigma}}_\lambda^m)^{-1/2}(\boldsymbol{\lambda}^m - \hat{\boldsymbol{\mu}}_\lambda^m)$. Finally, an approximate sample from the full conditional distribution for $\boldsymbol{\lambda}_t$, $p(\boldsymbol{\lambda}_t|\boldsymbol{\phi}_t, \boldsymbol{s}_t)$, is obtained by

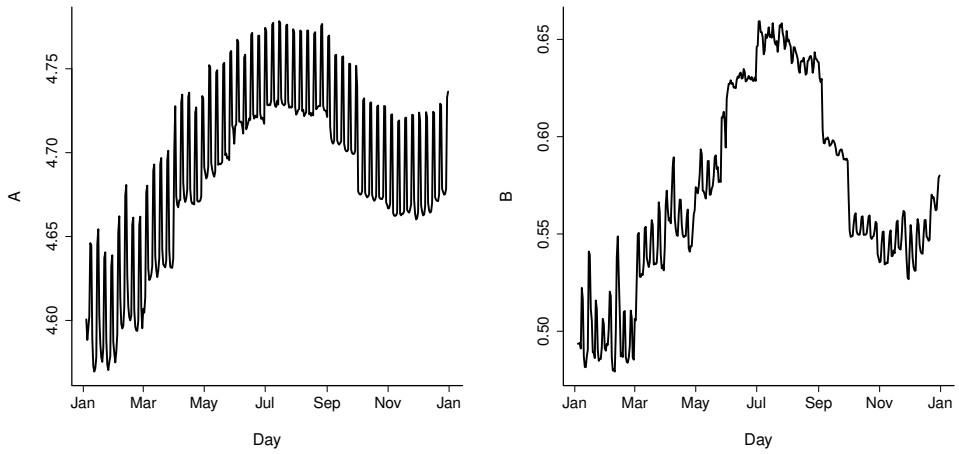
$$\boldsymbol{\lambda}_t^* = \hat{\boldsymbol{\mu}}_\lambda^* + (\hat{\boldsymbol{\Sigma}}_\lambda^*)^{1/2}\boldsymbol{\epsilon}_\lambda^k \sim \int p(\boldsymbol{\lambda}_t|\boldsymbol{\phi}, \boldsymbol{s}_t)K_h(\|\boldsymbol{\phi} - \boldsymbol{\phi}_t\|)p(\boldsymbol{\phi})d\boldsymbol{\phi},$$

where the index k is sampled from $(1, \ldots, M)$ with probability proportional to $K_h(\|\boldsymbol{\phi}^m - \boldsymbol{\phi}_t\|)$, and $\hat{\boldsymbol{\mu}}_\lambda^*$ and $(\hat{\boldsymbol{\Sigma}}_\lambda^*)^{1/2}$ are implicitly defined by $\boldsymbol{\phi}_t$.

5.4.2 Complete model specification and results

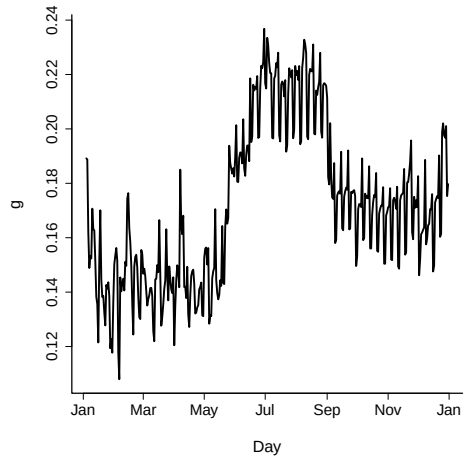
Figure 5.4 shows the L-moments estimates $\hat{\boldsymbol{\beta}}_t$. Each individual series seems to be dominated by 4 components – a dynamic level; a sudden shift induced by the start and end of the extended summer season (from 1 April to 31 September); a weekly seasonal effect; an error term that accounts for other minor non-considered factors. Also, the series are clearly interdependent. In particular, a strong negative correlation between the parameters of variability (Figure 5.4d) and kurtosis (Figure 5.4b) is present, as one would expect in advance.

To accommodate the referred factors, each individual time series $i = 1, \ldots, 4$, is asso-

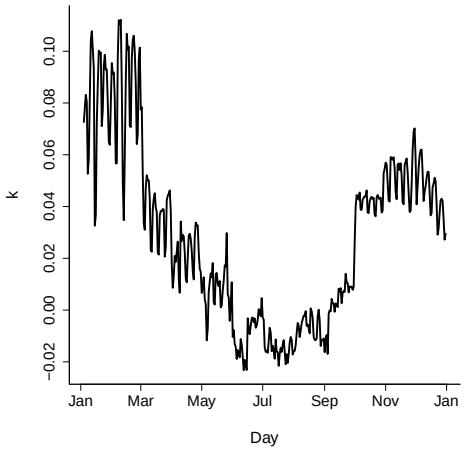


(a) Location

(b) Scale



(c) Skewness



(d) Kurtosis

Figure 5.4: L-moments estimates, $\hat{\boldsymbol{\beta}}_t = g^{-1}(\boldsymbol{s}_{\text{obs}})$, of the g-and-k parameters – obtained by fitting the distributions independently, one for each day.

ciated to its own system parameters, $\boldsymbol{\theta}_t^{(i)}$, and the following matrices:

$$\boldsymbol{F}_t^{(i)} = (\boldsymbol{E}_2, \boldsymbol{E}_6, \delta(t))$$

and

$$\boldsymbol{G}_t^{(i)} = \boldsymbol{G}^{(i)} = \begin{pmatrix} \boldsymbol{J}_2 & \mathbf{0}_{2\times 6} & \mathbf{0}_{2\times 1} \\ \mathbf{0}_{6\times 2} & \boldsymbol{P}_6 & \mathbf{0}_{6\times 1} \\ \mathbf{0}_{1\times 2} & \mathbf{0}_{1\times 6} & 1 \end{pmatrix}, \quad \text{where } \boldsymbol{J}_2 = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, \boldsymbol{P}_6 = \begin{pmatrix} -\mathbf{1}_{1\times 5} & -1 \\ \mathbf{I}_5 & \mathbf{0}_{5\times 1} \end{pmatrix},$$

$\boldsymbol{E}_n = (1, 0, \ldots, 0)$ is an n-dimensional vector, $\delta(t)$ is an indicator function that takes value 1 if t is in the summer season and 0 otherwise, $\mathbf{1}$ denotes a matrix of ones. $\boldsymbol{J}_2$, which is a Jordan block, implies a local-linear trend for the latent level $\theta_{1,t}^{(i)}$. $\boldsymbol{P}_6$ is a permutation matrix that relates to the weekly seasonal effect, which impact the series though $\theta_{3,t}^{(i)}$. The summer-effect, in turn, is encapsulated by $\theta_{9,t}^{(i)}$.

The multivariate model in Eq. 5.4 becomes fully specified by setting

$$\boldsymbol{F}_t = \boldsymbol{F}_t^{(i)} \otimes \mathbf{I_4},\ \boldsymbol{G} = \boldsymbol{G}^{(i)} \otimes \mathbf{I_4} \text{ and } \boldsymbol{\theta}_t = (\boldsymbol{\theta}_t^{(1)}, \ldots, \boldsymbol{\theta}_t^{(4)}),$$

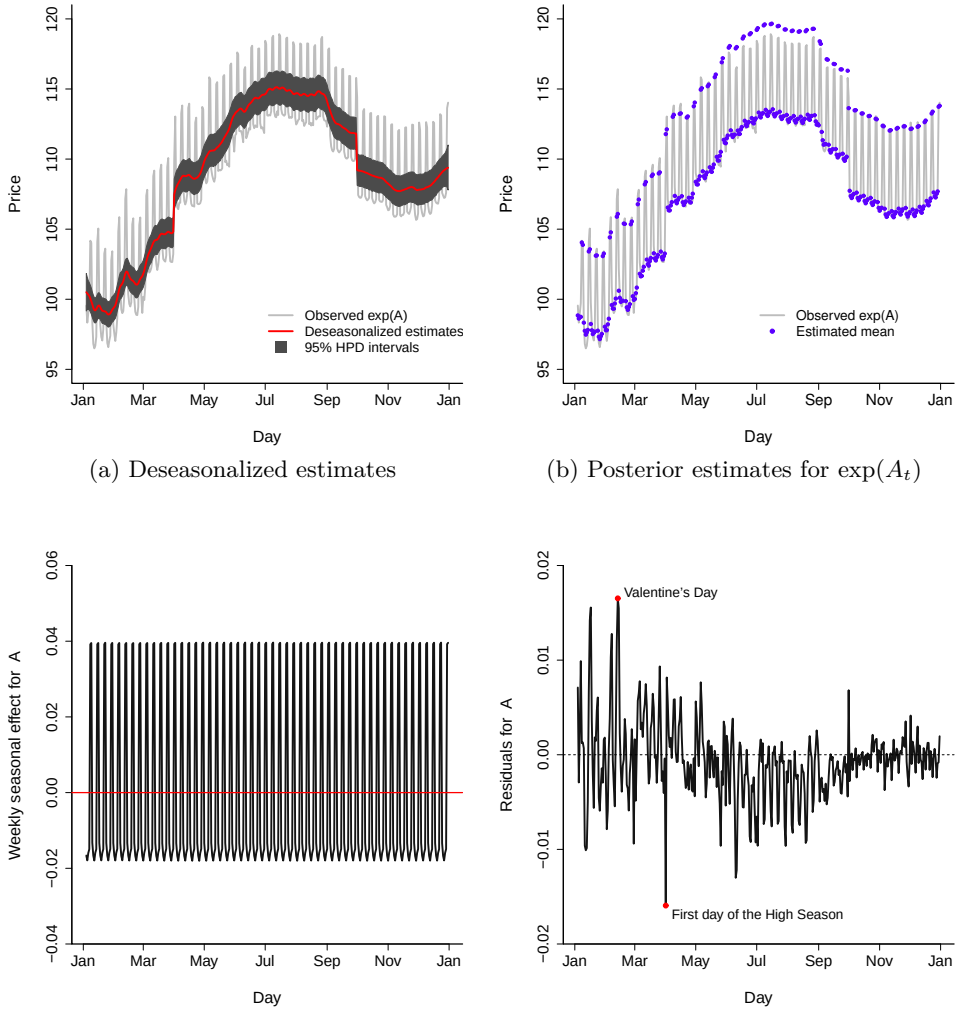
where $\otimes$ represents the Kronecker product defined as follows. Consider an $a \times b$ matrix $\boldsymbol{M}$ and a $c \times d$ matrix $\boldsymbol{N}$, then the Kronecker product of $\boldsymbol{M}$ and $\boldsymbol{N}$ is the $ac \times bd$ matrix

$$\boldsymbol{M} \otimes \boldsymbol{N} = \begin{pmatrix} m_{11}\boldsymbol{N} & \ldots & m_{1b}\boldsymbol{N} \\ \vdots & \ddots & \vdots \\ m_{a1}\boldsymbol{N} & \ldots & m_{ab}\boldsymbol{N} \end{pmatrix}.$$

The chosen specification superimpose the features driving this data, but could be easily adapted to other settings. For more details on how to specify the matrix of a dynamic model (see Petris et al., 2009).

Figure 5.5 shows some of the estimated model parameters associated to $\boldsymbol{\beta}_{1,t}$. In Figure 5.5a, the deseasonalized posterior estimates are plot over the raw series. For easier interpretation, the original scale of rental prices is adopted. The estimated series level changed substantially throughout the year, going from $\exp(\hat{\theta}_{1,1}^{(1)}) = 100.5$ to $\exp(\hat{\theta}_{1,T}^{(1)}) = 109.4$, a 8.8% increase. Caution must be exercised when interpreting this result. As the data was compiled when the series started, we are not comparing the prices actually paid for spaces

at given day, but the listing prices at the date of compilation. Another interesting feature is the sharp increase observed in the high season. The parameter $\boldsymbol{\theta}_{9,t}^{(i)}$ is estimated to be nearly-static along the year, with $\hat{\boldsymbol{\theta}}_{9,t}^{(1)} \approx 0.024$ for all t. Therefore, prices are expected to increase by about $\exp(0.024) - 1 = 2.4\%$ when the high season starts (and reduce by the same amount when it finished). If more years of data were available, it would be straightforward to include in the model an yearly seasonal effect.



(a) Deseasonalized estimates

(b) Posterior estimates for $\exp(A_t)$

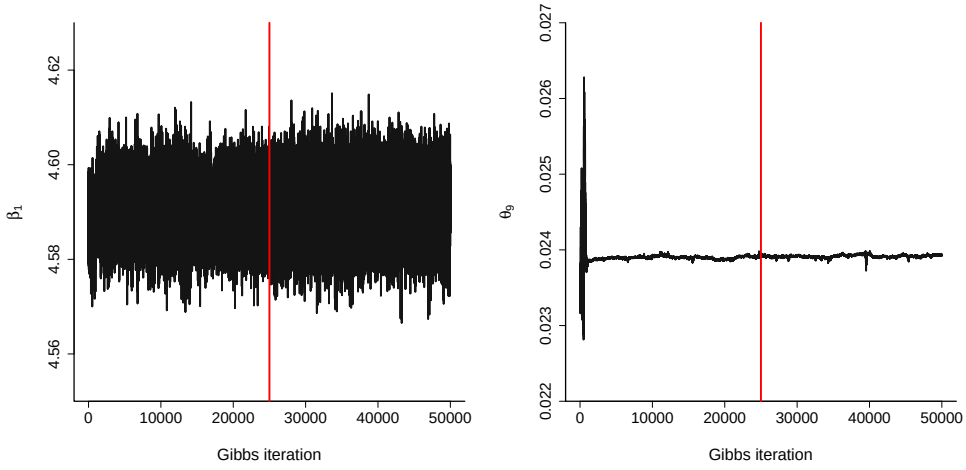
(c) Estimated seasonal effect

(d) Residual plot

Figure 5.5: All plots refer to the g-and-k location parameters A_t. (a) Light-gray lines are the observed L-moments estimates. The red line depicts the posterior mean of the deseasonalized parameter $\exp(\theta_{1,t}^{(1)} + \theta_{3,t}^{(1)}\delta(t))$, while the gray shade shows the 95% HPD credibility intervals; (b) the posterior estimates of $\exp(A_t)$ are given in blue; (c) The estimated seasonal effect on the linear predictor $\lambda_{1,t}$, given by the posterior mean of $\theta_{3,t}^{(1)}$; (d) The residual plot for A_t, showing the differences $(\boldsymbol{s}_{\text{obs}1,t} - \hat{\boldsymbol{\lambda}}_{1,t})$.

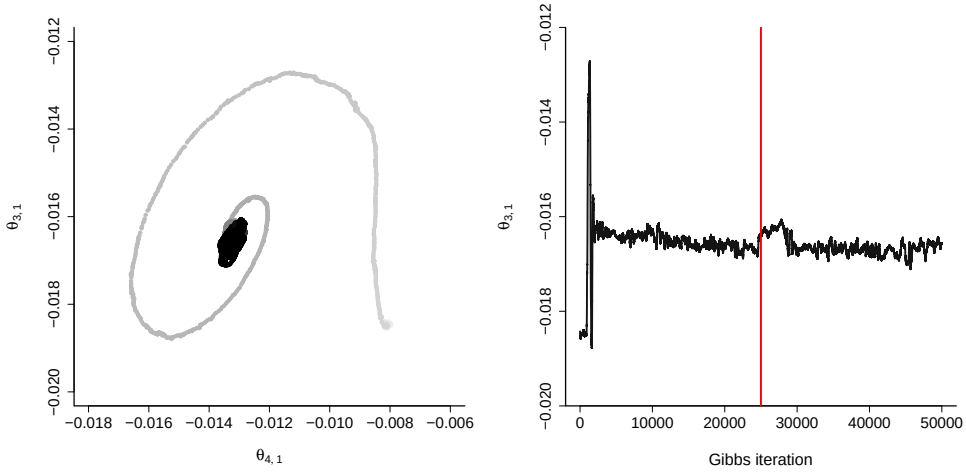
The blue points in Figure 5.5b are the estimated $\exp(\boldsymbol{\beta}_{1,t})$, obtained by adding the

seasonality effect to the red line in the previous plot. The seasonality effect is summarized in Figure 5.5c. The density location is estimated to increase by 5.6% from Thursdays to Fridays, for example. The residual plot in Figure 5.5d exhibits a slight lack-of-fit around the middle of the year. The highest residual was observed on the Valentine's weekend. Surprising at the first sight, this phenomenal might be related to an increase in demand driven by local couples, rather than visitors from other cities. This however would have to be confirmed by further analysis. The lowest residual was on the first day of the High season – Friday, 1 April.



(a) Parameter A at time $t = 1$.

(b) Average Summer-effect



(c) Bivariate trace of seasonality parameters

(d) Seasonality parameter

Figure 5.6: Trace plots for a selection of model parameters. The vertical red lines delimit the burn-in period. (a) Location parameter A at time $t = 1$; (b) Average (over the time period) of the nearly-static summer-effect parameter; (c) Bivariate trace of the seasonality parameters $\theta^{(1)}_{3,1}$ and $\theta^{(1)}_{4,1}$; (d) Univariate trace of $\theta^{(1)}_{3,1}$.

In this study, we generated 1 million approximate Gibbs samples, from which we keep

only 1 every 20 (the chain thinning), and then discarded the first 25000 as burn-in. Some of the MCMC trace plots are displayed in Figure 5.6. There are 13140 unknown parameters in the model, and assessing the convergence of the chain is not a trivial task. However, our exploration indicates that chains are seemingly stable (except, perhaps, if precise estimates of the posterior quantiles are required). The bivariate plot in Figure 5.6c shows a curious spiral trajectory towards the region of high probability density (this behavior might be explained by the rotation of the seasonality parameters induced by the permutation matrix $\boldsymbol{P}_6$). The MCMC was started from estimates provided by fitting, via Kalman smoothing, an auxiliary, simple state space model (that assumes that each series in Figure 5.4 follows an independent DLM, with pre-specified matrices $\boldsymbol{W}^{(i)}$).

The above example highlights the capacity of our technique to handle complex model structures. The inferential procedure applied here slightly differs from the steps described in Algorithm 1, even though they share the same underling principles. The use of the Linear Bayes allowed us to model the vector $\boldsymbol{\lambda}$ at once, instead of fitting one regression for each of its elements. An advantage of this approach is to account for the full correlation structure and to naturally handle heteroscedasticity. With univariate regressions, one would have to accommodate possible interaction terms and explicitly model the variance.

In regards to the computational cost, this was an expensive endeavor – although most of it was accounted for by the steps that would be executed in a hypothetical exact Gibbs sampler. It took nearly 10 days to generate the 1 million MCMC samples, running R on a personal computer. In addition, the chain mixes rather slowly.

The overall ABC error is successfully reduced by splitting the problem into small pieces. Unfortunately, this prompts the parameters to be highly correlated. In resume, there is a fundamental trade-off between the ABC and the Monte Carlo errors. But even so, for some problems, combining these tools are potentially the only feasible option.

5.5 Discussion

Modeling the posterior indirectly, through the full conditional distributions, may be appealing for a number of reasons. The proposed method naturally captures the covariance structure of the posterior distribution (at least when appropriate regressions models are employed), which represents a useful advantage over most ABC algorithms, as discussed

in the introduction section.

When compared to the regression-adjustment ABC, the approximate Gibbs sampler can massively simplify the regression models structures. Handling a set of low-dimensional models is frequently easier than deriving an appropriate multivariate regression that describes $f(\boldsymbol{\theta}|\boldsymbol{S})$ at once. That becomes particularly beneficial when generating synthetic samples is computationally expensive (e.g. in genetic and epidemiological studies).

The strategies described in Subsection 5.3.2 explore the model's architecture to further reduce the regressions' complexity, leading ultimately to lower approximation errors. These solutions make our approach suitable for inference in some well-structured problems, such as the one considered in Section 5.4. However, for complex and non-structured models where little information about the form of the (exact) full conditional distributions is available, the regressions may not provide fully reasonable conditional model representations for all parameters, and accordingly should be elicited with care. In this regard, a single poorly-approximated conditional distribution could in principle severely impact the estimation of the remaining parameters, as it would also happen with the standard Gibbs sampler if one of the full conditional distributions was wrongly derived.

As discussed in Section 5.2, under certain theoretical conditions, the likelihood-free approximate Gibbs sampler will exactly target the true partial posterior $\pi(\boldsymbol{\theta}|\boldsymbol{s}_{\text{obs}})$. In more realistic cases, where the computational resources are limited, the results are approximate, and might be sensitive to the tuning parameters and the validity of the regression assumptions. We therefore recommend the practitioner to conduct a detailed diagnostic analysis to ensure that the algorithm's inputs have been properly specified.

Localizing the regressions is also beneficial in terms of making the method more robust, even though that is not always the case. While it is generally adequate to overcome lack of normality, linearity or homoscedasticity, it is not as effective against missing variables and/or interactions terms.

Inheriting some of the limitations of the Gibbs sampler is another considerable drawback of this approach. Potentially poor mixing, convergence issues and the impact of the approximations in the behavior of the chain may impose important practical constraints. On that note, other decompositions of the posterior can be derived to avoid those concerns.

Specifically, considering the following decomposition:

$$f(\boldsymbol{\theta}|\boldsymbol{S}) = f(\theta_1|\boldsymbol{S})f(\theta_2|\theta_1, \boldsymbol{S}) \dots f(\theta_D|\theta_1, \dots, \theta_{D-1}, \boldsymbol{S}),$$

one can still resort on synthetic samples to model each term in the right-hand side. It is then possible to draw an approximate posterior sample $\boldsymbol{\theta}^*$ by sequentially sampling from $\theta_1^* \sim \hat{f}(\theta_1|\boldsymbol{S})$ to $\theta_D^* \sim \hat{f}(\theta_D|\boldsymbol{\theta}_{-D}^*, \boldsymbol{S})$.

Algorithm 6 Likelihood-free approximate Gibbs sampling (localised models)

Inputs:

- An observed dataset $\boldsymbol{X}_{\text{obs}}$.
- A prior $\pi(\boldsymbol{\theta})$ and intractable generative model $p(\boldsymbol{X}|\boldsymbol{\theta})$.
- A sampling distribution $b(\boldsymbol{\theta})$ describing a region of high posterior density.
- An observed vector of summary statistics $\boldsymbol{s}_{\text{obs}} = S(\boldsymbol{X}_{\text{obs}})$.
- A smoothing kernel $K_h(u)$ with scale parameter $h > 0$.
- A positive integer N defining the number of ABC samples.
- A positive integer M defining the number of Gibbs sampler iterations.
- A collection of regression models $f(\theta_d|\boldsymbol{\beta}_d^+, g_d(\boldsymbol{S}, \boldsymbol{\theta}_{-d}))$ to approximate each full conditional distribution $\pi(\theta_d|\boldsymbol{s}_{\text{obs}}, \boldsymbol{\theta}_{-d})$ for $d = 1, \ldots, D$.

Data simulation:
For $i = 1, \ldots, N$:

1.1 Generate $\boldsymbol{\theta}^{(i)} \sim b(\boldsymbol{\theta})$ from some suitable distribution $b(\boldsymbol{\theta})$.
1.2 Generate $\boldsymbol{X}^{(i)} \sim p(\boldsymbol{X}|\boldsymbol{\theta}^{(i)})$ from the model.
1.3 Compute the summary statistics $\boldsymbol{s}^{(i)} = S(\boldsymbol{X}^{(i)})$.

Approximate Gibbs sampling:

2.1 Initialise $\tilde{\boldsymbol{\theta}}^{(0)} = (\tilde{\theta}_1^{(0)}, \ldots, \tilde{\theta}_D^{(0)})^\top$.

2.2 For $m = 1, \ldots, M$:
For $d = 1, \ldots, D$:

2.2.1 Denote by $\boldsymbol{\theta}^\star_{-d} = (\tilde{\theta}_1^{(m)}, \ldots, \tilde{\theta}_{d-1}^{(m)}, \tilde{\theta}_{d+1}^{(m-1)}, \ldots, \tilde{\theta}_D^{(m-1)})^\top$ the vector containing the most recently updated values of $\tilde{\theta}_j^{(\cdot)}$, $j \neq d$.

2.2.2 Set the regression weights $w_d^{(i)} = K_h(\|(\boldsymbol{s}^{(i)}, \boldsymbol{\theta}_{-d}^{(i)}) - (\boldsymbol{s}_{\text{obs}}, \boldsymbol{\theta}^\star_{-d})\|)\pi(\boldsymbol{\theta})/b(\boldsymbol{\theta})$ for $i = 1, \ldots, N$.

2.2.3 Fit a suitable regression model $\theta_d|(\boldsymbol{S}, \boldsymbol{\theta}_{-d}) \sim f(\theta_d|\boldsymbol{\beta}_d^+, g_d(\boldsymbol{S}, \boldsymbol{\theta}_{-d}))$ using the weighted samples $\{(\boldsymbol{\theta}^{(i)}, \boldsymbol{s}^{(i)}, w_d^{(i)})\}_{i=1}^N$, so that $f(\theta_d|\hat{\boldsymbol{\beta}}_d^+, g_d(\boldsymbol{s}_{\text{obs}}, \boldsymbol{\theta}^\star_{-d}))$ locally approximates the full conditional distribution $\pi(\theta_d|\boldsymbol{s}_{\text{obs}}, \boldsymbol{\theta}^\star_{-d})$.

2.2.4 Gibbs update: sample $\tilde{\theta}_d^{(m)}|(\boldsymbol{s}_{\text{obs}}, \boldsymbol{\theta}^\star_{-d}) \sim f(\theta_d|\hat{\boldsymbol{\beta}}_d^+, g_d(\boldsymbol{s}_{\text{obs}}, \boldsymbol{\theta}^\star_{-d}))$.

Output:

- Realised Gibbs sampler output $\tilde{\boldsymbol{\theta}}^{(0)}, \ldots, \tilde{\boldsymbol{\theta}}^{(M)}$ with target distribution $\approx \pi(\boldsymbol{\theta}|\boldsymbol{s}_{\text{obs}})$.

Algorithm 7 Likelihood-free approximate Gibbs sampling (global models) [Changes from Algorithm 6.]

Approximate Gibbs sampling:

2.1 Initialise $\tilde{\boldsymbol{\theta}}^{(0)} = (\tilde{\theta}_1^{(0)}, \ldots, \tilde{\theta}_D^{(0)})^\top$.

2.2 Compute the sample weights $w^{(i)} \propto K_h(\|\boldsymbol{s}^{(i)} - \boldsymbol{s}_{\text{obs}}\|)\pi(\boldsymbol{\theta})/b(\boldsymbol{\theta})$, for $i = 1, \ldots N$.

2.3 For $d = 1, \ldots, D$:
Fit a suitable regression model $\theta_d|(\boldsymbol{S}, \boldsymbol{\theta}_{-d}) \sim f(\theta_d|\boldsymbol{\beta}_d^+, g_d(\boldsymbol{S}, \boldsymbol{\theta}_{-d}))$ using the weighted samples $\{(\boldsymbol{\theta}^{(i)}, \boldsymbol{s}^{(i)}, w^{(i)})\}_{i=1}^N$, so that $f(\theta_d|\hat{\boldsymbol{\beta}}_d^+, g_d(\boldsymbol{s}_{\text{obs}}, \boldsymbol{\theta}_{-d}))$ locally approximates the full conditional distribution $\pi(\theta_d|\boldsymbol{s}_{\text{obs}}, \boldsymbol{\theta}_{-d})$.

2.4 For $m = 1, \ldots, M$:
For $d = 1, \ldots, D$:

2.4.1 Denote by $\boldsymbol{\theta}_{-d}^\star = (\tilde{\theta}_1^{(m)}, \ldots, \tilde{\theta}_{d-1}^{(m)}, \tilde{\theta}_{d+1}^{(m-1)}, \ldots, \tilde{\theta}_D^{(m-1)})^\top$ the vector containing the most recently updated values of $\tilde{\theta}_j^{(\cdot)}$, $j \neq d$.

2.4.2 Gibbs update: sample $\tilde{\theta}_d^{(m)}|(\boldsymbol{s}_{\text{obs}}, \boldsymbol{\theta}_{-d}^\star) \sim f(\theta_d|\hat{\boldsymbol{\beta}}_d^+, g_d(\boldsymbol{s}_{\text{obs}}, \boldsymbol{\theta}_{-d}^\star))$.

Discussion

Throughout this thesis, we have introduced new methods and algorithms that have meaningfully contributed to the advancement of the Approximate Bayesian Computation literature. Together, the proposals foster the capacity of ABC to handle very high dimensional models, alleviate the intrinsic ABC approximation error and enhance the estimation of posterior dependences. We have also extended ABC into the non-parametric framework, demonstrating its potential to estimate infinite-dimensional parameters.

In Chapter 2, we combined some of the state-of-the-art ABC techniques to make inference possible in a challenging epidemiological setting. The fitted parameters lead to the biologically important conclusion that the vast majority of cases of multidrug resistant tuberculosis come from transmission of strains that are already multidrug resistant. This suggests that, although it is imperative to support treatment adherence (to avoid treatment failures that could lead to multidrug resistance), public health policies that reduce the incidence of new cases may be effective in controlling drug resistance. In addition, the study concludes that there is a positive probability of conversion directly from doubly drug sensitive to doubly resistant infections. Lastly, resistance to isoniazid occurs much more frequently than resistance against rifampicin, which is consistent with previously documented estimates based on *in vitro* experiments (David, 1970; Ford et al., 2013).

Computationally, estimating this model was a highly demanding endeavor. If the fitting process was faster, we could have increased the model's complexity in a number of ways, including assigning prior distributions to some of the parameters assumed to be known (based on values previously reported in the epidemiological literature). In hindsight, the lazy ABC approach introduced by Prangle (2016) could potentially have accelerated the computations, given the high cost of generating each synthetic dataset ($\approx$ 40 sec). Other extensions to this Chapter could involve accounting for the incubation period of the pathogen or performing model selection in a more formal way – instead of

basing this on the posterior predictive distribution.

Chapter 3 was devoted to the problem of non-parametrically estimating a set of associated density functions. We proposed specifying the prior distribution using Gaussian processes in a hierarchical architecture. Exact inference using MCMC would have been infeasible. However, we showed that by extending the principles of regression adjustment to the case of functional objects, one could achieve accurate posterior inference with practically viable computational costs. The method proved to be competitive against other well-established non-parametric approaches. An important advantage of the ABC construction is that the inferential procedure is somewhat detached from the actual model specification, in the sense that changes to the base density, for example, do not require the tedious exercise of re-deriving the sampling equations. It also liberates us from the need to adopt conjugate distributions. In this work, the "closeness" of two samples was defined in terms of the KL-divergence between the respective Kernel Density Estimators. This sort of summarization may be useful in other ABC constructions, not necessarily involving functional parameters.

Despite our encouraging results, there is still room for improvement. Specifically, we adopted an arbitrarily chosen regular grid to discretize the continuous function. A better alternative would be to, as in Tokdar (2007), infer an appropriate irregular grid from the observed data. However, while it is obvious that regions where the densities vary more rapidly should be allocated more grid points, it is not so obvious how to actually implement this concept in practice.

Deriving an appropriate functional regression model that reasonably mirrors the underlying probabilistic relationship was anything but trivial. That is hardly surprising, and provides yet another illustrative example of the nature of regression adjustment techniques: the gain (or loss) obtained from performing the adjustment is determined by the modeler's ability (or resources) to derive a good regression model.

As an alternative to regression adjustment, we introduced in Chapter 4 an algorithm that allows one to recalibrate (in the coverage property sense) a posterior estimate. It would be fair to state that the core idea is to use ABC to improve an ABC estimator – or, equivalently, it is about generating synthetic ABC estimates, from which we can learn how the ABC estimator behaves in a neighborhood of the observed data. It is important to clarify that the diagnostic part of this approach was previously proposed by Prangle

et al. (2014). Our contribution is on how to enhance the initial estimator in response to what was learned from the diagnostic analysis.

We showed through a simulation study that this approach serves both as an alternative and as a supplement to regression-adjustment – our approach outperformed regression-adjustment alone by a comfortable margin. The method may also be used to correct other cheaper, non-ABC estimators, such as the one presented in Subsection 4.3.1. This indirect inference approach mitigates the computational burden, which is often substantial, given that one needs to estimate the posterior distribution numerous times. Operationally, the method only requires the tuning of the acceptance fraction. Still, we were unable to establish a reliable recommendation in regards to how this parameter should be chosen in a general setting – future contributions along these lines would certainly strengthen the results available at this stage.

Finally, in Chapter 5, we introduced a likelihood-free approximate Gibbs sampler. The leading motivation for creating this algorithm is the realization that approximating univariate conditional distributions is considerably easier than approximating a high-dimensional posterior distribution. The Gibbs sampler mechanics provided a straightforward way to implement this idea. We therefore proposed the use of ABC to draw approximate samples from these full conditional distributions that are intractable, within the Gibbs sampler. As with the recalibration algorithm, the covariance structure of the posterior distribution is properly captured here.

The algorithm successfully handled the remarkably difficult problem of estimating an intractable multivariate high-dimension state-space model (where more than 13000 interrelated parameters were present). This demonstrates the methods aptitude to address well-structured models for time series, spatial or hierarchical data. Although we didn't focused on these sorts of applications, we anticipate that the method will also be attractive in problems where it is costly to generate synthetic samples.

In terms of limitations, the sampler inherits the extensively-researched limitations of the exact Gibbs sampler (e.g. poor mixing). Developing a stronger theoretical foundation would also be useful to identify situations where the algorithm may perform poorly. Similarly to the work by van Buuren and Groothuis-Oudshoorn (2011) our approach is also sensitive to problems encountered where the conditional distributions (or their approximations) are not compatible with the existence of a unique joint distribution. Issues

involving potential incompatibility, and its consequences, therefore need further investigation. Exploiting ABC to approximate other decompositions of the posterior is one also possibility, which could eliminate some of the raised concerns.

List of Figures

List of Tables

List of Algorithms

References

Aandahl, R. Z., Reyes, J. F., Sisson, S. A., and Tanaka, M. M. (2012), "A model-based Bayesian estimation of the rate of evolution of VNTR loci in *Mycobacterium tuberculosis*." *PLoS Computational Biology*, 8, e1002573.

Adams, R. P., Murray, I., and MacKay, D. J. C. (2009), "Nonparametric Bayesian Density Modeling with Gaussian processes," *arXiv preprint arXiv:0912.4896*.

Aeschbacher, S., Beaumont, M. A., and Futschik, A. (2012), "A Novel Approach for Choosing Summary Statistics in Approximate Bayesian Computation," *Genetics*, 192, 1027–1047.

Anderson, C. and Coles, S. (2002), "The Largest Inclusion in a Piece Of Steel," *Extremes*, 5, 237–252.

Anderson, L. F., Tamne, S., Brown, T., Watson, J. P., Mullarkey, C., Zenner, D., and Abubakar, I. (2014), "Transmission of multidrug-resistant tuberculosis in the UK: a cross-sectional molecular and epidemiological study of clustering and contact tracing," *The Lancet Infectious Diseases*, 14, 406–415.

Andrieu, C. and Roberts, G. O. (2009), "The pseudo-marginal approach for efficient Monte Carlo computations," *Annals of Statistics*, 37, 697–725.

Arnold, B. C., Castillo, E., and Sarabia, J. M. (1999), *Conditional Specification of Statistical Models*, Spinger Series in Statistics, New York, NY: Springer New York.

Asmussen, S. and Rojas-Nandayapa, L. (2008), "Asymptotics of sums of lognormal random variabiles with Gaussian copula," *Statistics and Probability Letters*, 78, 2709–2714.

Barthelmé, S. and Chopin, N. (2014), "Expectation propagation for likelihood-free

inference," *Journal of the American Statistical Association*, 109, 315–333.

Bazin, E., Dawson, K. J., and Beaumont, M. A. (2010), "Likelihood-free inference of population structure and local adaptation in a Bayesian hierarchical model," *Genetics*, 185, 587–602.

Beaumont, M. (2010), "Approximate Bayesian computation in evolution and ecology," *Annual Review of Ecology, Evolution, and Systematics*, 41, 379–406.

Beaumont, M. A. (2003), "Estimation of population growth or decline in genetically monitored populations," *Genetics*, 164, 1139–1160.

Beaumont, M. A., Zhang, W., and Balding, D. J. (2002), "Approximate Bayesian computation in population genetics," *Genetics*, 162, 2025–2035.

Beran, R. (1987), "Prepivoting to reduce level error of confidence sets," *Biometrika*, 74, 457–468.

Bertorelle, G., Benazzo, A., and Mona, S. (2010), "ABC as a flexible framework to estimate demography over space and time: Some cons, many pros," *Molecular Ecology*, 19, 2609–2625.

Biau, G., Céreou, F., and Guyader, A. (2015), "New insights into approximate Bayesian computation," *Ann. Inst. H. Poincaré Probab. Statist.*, 51, 376–403.

Blower, S. M. and Chou, T. (2004), "Modeling the emergence of the 'hot zones': tuberculosis and the amplification dynamics of drug resistance." *Nature Medicine*, 10, 1111–1116.

Blower, S. M., McLean, A. R., Porco, T. C., Small, P. M., Hopewell, P. C., Sanchez, M. A., and Moss, A. R. (1995), "The intrinsic transmission dynamics of tuberculosis epidemics." *Nature Medicine*, 1, 815–821.

Blum, M. G. B. (2010), "Approximate Bayesian Computation: A Nonparametric Perspective," *Journal of the American Statistical Association*, 105, 491–1178.

Blum, M. G. B. and François, O. (2010), "Non-linear regression models for approximate Bayesian computation," *Statistics and Computing*, 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, 189–208.

Bonassi, F. V., You, L., and West, M. (2011), "Bayesian learning from marginal data in bionetwork models," *Statistical Applications in Genetics and Molecular Biology*, 10, Article 49.

Bortot, P., Coles, S. G., and Sisson, S. a. (2007), "Inference for Stereological Extremes," *Journal of the American Statistical Association*, 102, 84–92.

Calvet, L. E. and Czellar, V. (2012), "Accurate Methods for Approximate Bayesian Computation Filtering," *Journal of Financial Econometrics*, 13, 798–838.

Casali, N., Nikolayevskyy, V., Balabanova, Y., Harris, S. R., Ignatyeva, O., Kontsevaya, I., Corander, J., Bryant, J., Parkhill, J., Nejentsev, S., Horstmann, R. D., Brown, T., and Drobniewski, F. (2014), "Evolution and transmission of drug resistant tuberculosis in a Russian population," in *Nature genetics*.

Chen, S.-H. and Ip, E. H. (2015), "Behaviour of the Gibbs sampler when conditional distributions are potentially incompatible," *Journal of Statistical Computation and Simulation*, 85.

Chen, S. X. (1999), "Beta Kernels Estimators for Density Functions," *Computational Statistics & Data Analysis*, 31, 131–145.

Cohen, T. and Murray, M. (2004), "Modeling epidemics of multidrug-resistant *M. tuberculosis* of heterogeneous fitness." *Nature Medicine*, 10, 1117–1121.

Colijn, C., Cohen, T., Ganesh, A., and Murray, M. (2011), "Spontaneous emergence of multiple drug resistance in tuberculosis before and during therapy." *PLoS One*, 6, e18327.

Cook, S. R., Gelman, A., and Rubin, D. B. (2006), "Validation of Software for Bayesian Models Using Posterior Quantiles," *Journal of Computational and Graphical Statistics*, 15, 675–692.

Csillery, K., Blum, M. G. B., Gaggiotti, O. E., and François, O. (2010), "Approximate Bayesian computation in practice," *Trends in Ecology and Evolution*, 25, 410–418.

Csilléry, K., François, O., and Blum, M. G. B. (2012), "abc: an R package for approximate Bayesian computation (ABC)," *Methods in Ecology and Evolution*, 3, 475–479.

Da-Silva, C. Q., Migon, H. S., and Correia, L. T. (2011), "Dynamic Bayesian beta models," *Computational Statistics and Data Analysis*, 55, 2074–2089.

Da-Silva, C. Q. and Rodrigues, G. S. (2013), "Bayesian Dynamic Dirichlet Models," *Communications in Statistics - Simulation and Computation*, 44, 787–818.

Dai, J. and Sperlich, S. (2010), "Simple and Effective Boundary Correction for Kernel Densities and Regression With An Application to the World Income and Engel Curve Estimation," *Computational Statistics & Data Analysis*, 54, 2487–2497.

David, H. L. (1970), "Probability distribution of drug-resistant mutants in unselected populations of *Mycobacterium tuberculosis*." *Applied Microbiology*, 20, 810–814.

Davidson, R. and MacKinnon, J. G. (2002), "Fast Bootstrap Tests of Nonnested Linear Regression Models," *Econometric Reviews*, 21, 419–429.

— (2007), "Improving the reliability of bootstrap tests with the fast double bootstrap," *Computational Statistics and Data Analysis*, 51, 3259 – 3281.

De Iorio, M., Muller, P., Rosner, G. L., and MacEachern, S. N. (2004), "An ANOVA model for dependent random measures," *Journal of the American Statistical Association*, 99, 205–215.

Dean, T. A., Singh, S. S., Jasra, A., and Peters, G. W. (2014), "Parameter estimation for hidden markov models with intractable likelihoods," *Scandinavian Journal of Statistics*, 41, 970–987.

Drechsler, J. and Rassler, S. (2008), "Does convergence really matter?" in *Recent Advances in Linear Models and Related Areas – Essays in Honour of Helge Toutenburg*, eds. Shalabh and Heumann, C., Springer-Verlag, Berlin.

Drovandi, C. C., Mengersen, K. L., and Robert, C. P. (2017), "Approximating the likelihood in Approximate Bayesian Computation," in *Handbook of Approximate Bayesian Computation*, eds. Sisson, S. A., Fan, Y., and Beaumont, M. A., Chapman & Hall/CRC Press, p. in press.

Drovandi, C. C. and Pettitt, A. N. (2011), "Likelihood-free Bayesian estimation of multivariate quantile distributions," *Computational Statistics and Data Analysis*, 55, 2541–2556.

Drovandi, C. C., Pettitt, A. N., and Lee, A. (2015), "Bayesian Indirect Inference

Using a Parametric Auxiliary Model," *Statistical Science*, 30, 72–95.

Dye, C. and Espinal, M. A. (2001), "Will tuberculosis become resistant to all antibiotics?" *Proceedings of the Royal Society of London B: Biological Sciences*, 268, 45–52.

Efron, B. (1987), "Better Bootstrap Confidence Intervals," *Journal of the American Statistical Association*, 82, 171–185.

Erhardt, R. and Sisson, S. A. (2016), "Modelling extremes using approximate Bayesian computation," 281-306.

Escobar, M. D. and West, M. (1995), "Bayesian Density Estimation and Inference Using Mixtures," *Journal of the American Statistical Association*, 90, 577–588.

Ewens, W. J. (1972), "The sampling theory of selectively neutral alleles." *Theoretical Population Biology*, 3, 87–112.

Fan, J. and Zhang, W. (1999), "Statistical Estimation in Varying Coefficient Models," *The Annals of Statistics*, 27, 1491–1518.

Fan, Y., Dortet-Bernadet, J.-L., and Sisson, S. A. (2010), "On Bayesian Curve Fitting via Auxiliary Variables," *Journal of Computational and Graphical Statistics*, 19, 626–644.

Fan, Y., Nott, D. J., and Sisson, S. A. (2013), "Approximate Bayesian computation via regression density estimation," *Stat*, 2, 34–48.

Fearnhead, P. and Prangle, D. (2012), "Constructing summary statistics for approximate Bayesian computation: Semi-automatic approximate Bayesian computation," *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 74, 419–474.

Fenton, L. F. (1960), "The Sum of Log-Normal Probability Distributions in Scatter Transmission Systems," *IRE Transactions on Communications Systems*, 8, 57–67.

Ferguson, T. S. (1973), "A Bayesian analysis of some nonparametric problems," *The Annals of Statistics*, 1, 209–230.

Flaxman, S., Gelman, A., Neill, D., Smola, A., and Vehtari, A. (2016), "Fast hierarchical Gaussian processes," *Manuscript in preparation*.

Ford, C. B., Shah, R. R., Maeda, M. K., Gagneux, S., Murray, M. B., Cohen, T., Johnston, J. C., Gardy, J., Lipsitch, M., and Fortune, S. M. (2013), "*Mycobacterium tuberculosis* mutation rate estimates from different lineages predict substantial differences in the emergence of drug-resistant tuberculosis." *Nature Genetics*, 45, 784–790.

Frazier, D. T., Robert, C. P., and Rousseau, J. (2017), "Model misspecification in ABC: Consequences and diagnostics," *In preparation*.

Gandhi, N. R., Moll, A., Sturm, A. W., Pawinski, R., Govender, T., Lalloo, U., Zeller, K., Andrews, J., and Friedland, G. (2006), "Extensively drug-resistant tuberculosis as a cause of death in patients co-infected with tuberculosis and HIV in a rural area of South Africa," *The Lancet*, 368, 1575–1580.

Geenens, G. (2014), "Probit Transformation for Kernel Density Estimation on the Unit Interval," *Journal of the American Statistical Association*, 109, 346–358.

Gelman, A. (2004), "Parameterisation and Bayesian modelling," *Journal of the American Statistical Association*, 99, 537–545.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004), *Bayesian Data Analysis*, Chapman & Hall/CRC, 2nd ed.

Gillespie, S. H. (2002), "Evolution of drug resistance in *Mycobacterium tuberculosis*: clinical and molecular perspective." *Antimicrobial Agents and Chemotherapy*, 46, 267–274.

Gleim, A. and Pigorsch, C. (2013), "Approximate Bayesian computation with indirect summary statistics," Tech. rep., http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.665.5503.

Goldstein, M. (1976), "Bayesian analysis of regression problems," *Biometrika*, 63, 51–58.

Gourieroux, C., Monfort, A., and Renault, E. (1993), "Indirect inference," *Journal of Applied Econometrics*, 8, S85–S118.

Gutmann, M. U. and Corander, J. (2016), "Bayesian optimisation for likelihood-free inference of simulator-based statistical models," *Journal of Machine Learning Research*, 17, 1–47.

Gutmann, M. U., Dutta, R., Kaski, S., and Corander, J. (2017), "Likelihood-free inference via classification," *Statistics and Computing*.

Hall, P. (1986), "On the Bootstrap of Confidence Intervals," 14.

Hall, P. and Martin, M. A. (1988), "On bootstrap resampling and iteration," *Biometrika*, 75, 661–671.

Hartigan, J. (1969), "Linear Bayesian Methods," *Journal of the Royal Statistical Society. Series B (Methodological)*, 31, 446–454.

Haynes, M. A. (1998), "Flexible distributions and statistical models in ranking and selection procedures with applications," Ph.D. thesis, Queensland University of Technology.

Hubbell, S. P. (2001), *The unified neutral theory of biodiversity and biogeography (MPB-32)*, vol. 32, Princeton University Press.

Jara, A., Hanson, T., Quintana, F., Müller, P., and Rosner, G. (2011), "DPpackage: Bayesian Semi- and Nonparametric Modeling in R," *Journal of Statistical Software*, 40, 1–30.

Jasra, A., Singh, S. S., Martin, J. S., and McCoy, E. (2012), "Filtering via approximate Bayesian computation," *Statistics and Computing*, 22, 1223–1237.

Jingxian, W., Mehta, N. B., and Jin, Z. (2005), "A flexible lognormal sum approximation method," *GLOBECOM - IEEE Global Telecommunications Conference*, 6, 3413–3417.

Jones, M. C. (1993), "Simple boundary correction in kernel density estimation," *Statistics and Computing*, 3, 135–146.

Jones, M. C. and Henderson, D. A. (2007), "Kernel-Type Density Estimation on the Unit Interval," *Biometrika*, 94, 977–984.

Joyce, P. and Marjoram, P. (2008), "Approximately sufficient statistics and bayesian computation." *Statistical applications in genetics and molecular biology*, 7, Article26.

Kousathanas, A., Leuenberger, C., Helfer, J., Quinodoz, M., Foll, M., and Wegmann, D. (2016), "Likelihood-free inference in high-dimensional models," *Genetics*, 203, 893–904.

Lee, S. M. S. and Young, G. A. (1995), "Asymptotic Iterated Bootstrap Confidence Intervals," *Ann. Statist.*, 23, 1301–1330.

— (2003), "Prepivoting by weighted bootstrap iteration," *Biometrika*, 90, 393–410.

Lenk, P. J. (1988), "The Logistic Normal Distribution for Bayesian, Nonparametric, Predictive Densities," *Journal of the American Statistical Association*, 83, 509–516.

— (1991), "Towards a practicable Bayesian nonparametric density estimator," *Biometrika*, 78, 531.

Leonard, T. (1978), "Density Estimation, Stochastic Processes and Prior Information," *Journal of the Royal Statistical Society. Series B*, 40, 113–146.

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.

Lindsay, B. G. (1988), "Composite Likelihood Methods," .

Lintusaari, J., Gutmann, M. U., Dutta, R., Kaski, S., and Corander, J. (2017), "Fundamentals and Recent Developments in Approximate Bayesian Computation," *Syst. Biol*, 00, 1–17.

Lo, A. Y. (1984), "On a class of Bayesian nonparametric estimates: I, Density estimates," *Annals of Statistics*, 12, 351–357.

Lopes, J. S. and Beaumont, M. A. (2010), "ABC: A useful Bayesian tool for the analysis of population data," *Infection, Genetics and Evolution*, 10, 825–832.

Luciani, F., Francis, A. R., and Tanaka, M. M. (2008), "Interpreting genotype cluster sizes of *Mycobacterium tuberculosis* isolates typed with *IS*6110 and spoligotyping." *Infection, Genetics and Evolution*, 8, 182–190.

Luciani, F., Sisson, S. A., Jiang, H., Francis, A. R., and Tanaka, M. M. (2009), "The epidemiological fitness cost of drug resistance in *Mycobacterium tuberculosis*." *Proceedings of the National Academy of Sciences USA*, 106, 14711–14715.

Marin, J.-M., Pillai, N., Robert, C. P., and Rousseau, J. (2014), "Relevant statistics for Bayesian model choice," *Journal of the Royal Statistical Society, Series B*, 76, 833–859.

Marin, J.-M., Pudlo, P., and Robert, C. P. (2017), "Likelihood-free Model Choice," in *Handbook of Approximate Bayesian Computation*, eds. Sisson, S. A., Fan, Y., and Beaumont, M. A., CRC Press Taylor & Francis Group.

Marin, J.-M., Raynal, L., Pudlo, P., Ribatet, M., and Robert, C. P. (2016), "ABC random forests for Bayesian parameter inference," *https://arxiv.org/abs/1605.05537*.

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, 15324–8.

Martin, G. M., McCabe, B. P., Frazier, D. T., Maneesoonthorn, W., and Robert, C. P. (2017), "Auxiliary Likelihood-Based Approximate Bayesian Computation in State Space Models," *arXiv preprint arXiv:1604.07949*.

Martin, G. M., McCabe, B. P. M., Maneesoonthorn, W., and Robert, C. P. (2016), "Approximate Bayesian Computation in State Space Models," *arXiv:1409.8363*, 1–38.

Martin, J. S., Jasra, A., Singh, S. S., Whiteley, N., Del Moral, P., and McCoy, E. (2014), "Approximate Bayesian Computation for Smoothing," *Stochastic Analysis and Applications*, 32, 397–420.

Martin, M. A. (1990), "On Bootstrap Iteration for Coverage Correction in Confidence Intervals," *Journal of the American Statistical Association*, 85, 1105–1118.

Mazars, E., Lesjean, S., Banuls, A. L., Gilbert, M., Vincent, V., Gicquel, B., Tibayrenc, M., Locht, C., and Supply, P. (2001), "High-resolution minisatellite-based typing as a portable approach to global analysis of *Mycobacterium tuberculosis* molecular epidemiology." *Proceedings of the National Academy of Sciences of the United States of America*, 98, 1901–1906.

Meeds, T. and Welling, M. (2015), "Optimization Monte Carlo: Efficient and embarrassingly parallel likelihood-free inference," in *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, vol. 28, p. paper 5881.

Menéndez, P., Fan, Y., Garthwaite, P. H., and Sisson, S. A. (2014), "Simultaneous adjustment of bias and coverage probabilities for confidence intervals," *Computational Statistics and Data Analysis*, 70, 35–44.

Mitchison, D. A. (1951), "The segregation of streptomycin-resistant variants of *Mycobacterium tuberculosis* into groups with characteristic levels of resistance." *Microbiology*, 5, 596–604.

Monteserin, J., Camacho, M., Barrera, L., Palomino, J. C., Ritacco, V., and Martin, A. (2013), "Genotypes of *Mycobacterium tuberculosis* in patients at risk of drug resistance in Bolivia," *Infection, Genetics and Evolution*, 17, 195–201.

Muller, P., Quintana, F., and Rosner, G. (2004), "A method for combining inference across related nonparametric Bayesian models," *J. Roy. Statist. Soc., Series B*, 66, 735–749.

Nachega, J. B. and Chaisson, R. E. (2003), "Tuberculosis drug resistance: a global threat," *Clinical Infectious Diseases*, 36, S24–S30.

Nott, D. J., Fan, Y., Marshall, L., and Sisson, S. A. (2014), "Approximate Bayesian computation and Bayes linear analysis: Towards high-dimensional ABC," *Journal of Computational and Graphical Statistics*, 23, 65–86.

Nott, D. J., Ong, V. J.-H., Fan, Y., and Sisson, S. A. (2017), "High-dimensional ABC," in *Handbook of Approximate Bayesian Computation*, eds. Sisson, S. A., Fan, Y., and Beaumont, M. A., Chapman and Hall/CRC Press, p. in press.

Nunes, M. and Balding, D. J. (2010), "On optimal selection of summary statistics for approximate Bayesian computation." *Statistical applications in genetics and molecular biology*, 9, Article34.

OECD (2013), "Pisa in Focus, n° 28: What makes urban schools different?" URL: http://www.oecd.org/pisa/pisaproducts/pisainfocus/pisa(Last accessed on 17/10/2014).

Ong, V. J.-H., Nott, D. J., Tran, M.-N., Sisson, S. A., and Drovandi, C. C. (2016), "Variational Bayes with synthetic likelihood," *https://arxiv.org/abs/1608.03069*.

Peters, G. W., Chen, W. Y., and Gerlach, R. H. (2016), "Estimating Quantile Families of Loss Distributions for Non-Life Insurance Modelling via L-moments," *risks*, 42.

Petris, G. (2010), "An R Package for Dynamic Linear Models," *Journal of Statistical Software*, 36, 1–16.

Petris, G., Petrone, S., and Campagnoli, P. (2009), *Dynamic Linear Models with R*, useR!, Springer-Verlag, New York.

Picchini, U. and Samson, A. (2016), "Coupling stochastic EM and Approximate Bayesian Computation for parameter inference in state-space models," 46, 1–22.

Prangle, D. (2016), "Lazy abc," *Statistics and Computing*, 26, 171–185.

Prangle, D., Blum, M. G. B., Popovic, G., and Sisson, S. A. (2014), "Diagnostic tools for approximate Bayesian computation using the coverage property," *Australia and New Zealand Journal of Statistics*, 56, 309–329.

Prangle, D., Everitt, R. G., and Kypraios, T. (2016), "A rare event approach to high dimensional approximate Bayesian computation," *https://arxiv.org/abs/1611.02492*.

Pym, A. S., Saint-Joanis, B., and Cole, S. T. (2002), "Effect of katG Mutations on the Virulence of Mycobacterium tuberculosis and the Implication for Transmission in Humans," *Infection and Immunity*, 70, 4955–4960.

Ragheb, M. N., Ford, C. B., Chase, M. R., Lin, P. L., Flynn, J. L., and Fortune, S. M. (2013), "The mutation rate of mycobacterial repetitive unit loci in strains of *M. tuberculosis* from cynomolgus macaque infection." *BMC Genomics*, 14, 145.

Ramaswamy, S. and Musser, J. M. (1998), "Molecular genetic basis of antimicrobial agent resistance in *Mycobacterium tuberculosis*: 1998 update." *Tubercle and Lung Disease*, 79, 3–29.

Ramsay, J. and Silverman, B. W. (2005), *Functional Data Analysis*, Springer, New York, 2nd ed.

Ramsay, J. O., Wickham, H., Graves, S., and Hooker, G. (2013), *fda: Functional Data Analysis*, r package version 2.3.8.

Rasmussen, C. E. and Williams, C. K. I. (2006), *Gaussian Processes for Machine Learning*, MIT Press.

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, 10576–10581.

Rayner, G. D. and MacGillivray, H. L. (2002), "Numerical maximum likelihood estimation for the g-and-k and generalized g-and-h distributions," *Statistics and Computing*, 12, 57–75.

Reyes, J. F. and Tanaka, M. M. (2010), "Mutation rates of spoligotypes and variable numbers of tandem repeat loci in *Mycobacterium tuberculosis*." *Infection, Genetics and Evolution*, 10, 1046–1051.

Robert, C. P., Corunet, J.-M., Marin, J.-M., and Pillai, N. (2011), "Lack of confidence in approximate Bayesian computational (ABC) model choice," *Proceedings of the National Academy of Sciences of the USA*, 108, 15112–15117.

Rodrigues, G. S., Francis, A. R., Sisson, S. A., and Tanaka, M. M. (2017a), "Inferences on the acquisition of multidrug resistance in Mycobacterium turbeculosis using molecular epidemiolgical data," in *Handbook of Approximate Bayesian Computation*, eds. Sisson, S. A., Fan, Y., and Beaumont, M. A., CRC Press Taylor & Francis Group, p. in press.

Rodrigues, G. S., Nott, D. J., and Sisson, S. A. (2016), "Functional regression approximate Bayesian computation for Gaussian process density estimation," *Computational Statistics and Data Analysis*, 103, 229–241.

Rodrigues, G. S., Prangle, D., and Sisson, S. A. (2017b), "Recalibration: A post-processing method for approximate Bayesian computation," *Submitted for publication*.

Schwartz, S. C. and Yeh, Y. S. (1982), "On the Distribution Function and Moments of Power Sums With Log-Normal Components," *Bell System Technical Journal*, 61, 1441–1462.

Scott, D. W. (1992), *Multivariate Density Estimation: Theory, Practice, and Visualization*, John Wiley, New York.

Shi, J. Q. and Choi, T. (2011), *Gaussian Process Regression Analysis for Functional Data*, CRC Press.

Sisson, S. A. and Fan, Y. (2011), "Likelihood-free Markov chain Monte Carlo," in *Handbook of Markov chain Monte Carlo*, eds. Brooks, S. P., Gelman, A., Jones, G., and Meng, X. L., Boca Raton, FL: CRC Press, pp. 319–341.

Sisson, S. A., Fan, Y., and Beaumont, M. A. (eds.) (2017a), *Handbook of Approximate Bayesian Computation*, Chapman and Hall/CRC Press.

Sisson, S. A., Fan, Y., and Beaumont, M. A. (2017b), "Overview of approximate Bayesian computation," in *Handbook of Approximate Bayesian Computation*, eds. Sisson, S. A., Fan, Y., and Beaumont, M. A., Chapman & Hall/CRC, p. in press.

Sisson, S. A., Fan, Y., and Tanaka, M. M. (2007), "Sequential Monte Carlo without likelihoods," *Proceedings of the National Academy of Sciences of the United States of America*, 104, 1760–1765. Errata (2009), *106*, 16889.

Supply, P., Allix, C., Lesjean, S., Cardoso-Oelemann, M., Rüsch-Gerdes, S., Willery, E., Savine, E., de Haas, P., van Deutekom, H., Roring, S., Bifani, P., Kurepina, N., Kreiswirth, B., Sola, C., Rastogi, N., Vatin, V., Gutierrez, M. C., Fauville, M., Niemann, S., Skuce, R., Kremer, K., Locht, C., and van Soolingen, D. (2006), "Proposal for standardization of optimized mycobacterial interspersed repetitive unit-variable-number tandem repeat typing of *Mycobacterium tuberculosis*," *Journal of Clinical Microbiology*, 44, 4498–4510.

Supply, P., Niemann, S., and Wirth, T. (2011), "On the mutation rates of spoligotypes and variable numbers of tandem repeat loci of *Mycobacterium tuberculosis*." *Infection, Genetics and Evolution*, 11, 251–252.

Tanaka, M. M., Francis, A. R., Luciani, F., and Sisson, S. A. (2006), "Using approximate Bayesian computation to estimate tuberculosis transmission parameters from genotype data." *Genetics*, 173, 1511–1520.

Teh, Y. W., Jordan, M. I., Beale, M. J., and Blei, D. M. (2006), "Hierarchical dirichlet processes," *Journal of the American Statistical Association*, 101, 1566–1581.

Thorburn, D. (1986), "A Bayesian Approach to Density Estimation," *Biometrika*, 73, 65–75.

Thornton, K. and Andolfatto, P. (2006), "Approximate Bayesian inference reveals evidence for a recent, severe bottleneck in a Netherlands population of *Drosophila melanogaster*," *Genetics*, 172, 1607–1619.

Tokdar, S. T. (2007), "Towards a faster implementation of density estimation with logistic Gaussian process priors," *Journal of Computational and Graphical Statistics*, 16, 633–655.

Tokdar, S. T. and Ghosh, J. K. (2007), "Posterior consistency of logistic Gaussian process priors in density estimation," *Journal of Statistical Planning and Inference*, 137.

Tomlinson, G. and Escobar, M. (1999), "Analysis of densities," Tech. rep., University of Toronto.

Tran, M.-N., Nott, D. J., and Kohn, R. (2017), "Variational Bayes with intractable likelihood," *Journal of Computational and Graphical Statistics*, in press.

Turner, B. M. and Van Zandt, T. (2012), "A tutorial on approximate Bayesian computation," *Journal of Mathematical Psychology*, 56, 69–85.

van Buuren, S., Brand, J. P. L., Groothius-Oudshoorn, C. G. M., and Rubin, D. B. (2006), "Fully conditional specification in multivariate imputation," *Journal of Computational and Graphical Statistics*, 76, 1049–1064.

van Buuren, S. and Groothuis-Oudshoorn, J. (2011), "Mice: multivariate imputation by chained equations in R," *Journal of Statistical Software*, 45.

Wand, M. P., Marron, J. S., and Ruppert, D. (1991), "Transformations in Density Estimation," *Journal of the American Statistical Association*, 86, 343–353.

Wegmann, D., Leuenberger, C., and Excoffier, L. (2009), "Efficient Approximate Bayesian Computation Coupled With Markov Chain Monte Carlo Without Likelihood," *Genetics*, 182, 1207–1218.

Wegmann, D., Leuenberger, C., Neuenschwander, S., and Excoffier, L. (2010), "ABCtoolbox: a versatile toolkit for approximate Bayesian computations," *BMC bioinformatics*, 11, 116.

West, M. and Harrison, J. (1997), *Bayesian Forecasting and Dynamic Models*, Springer Series in Statistics, New York: Springer-Verlag, 2nd ed.

West, M., Harrison, P. J., and Migon, H. S. (1985), "Dynamic Generalized Linear Models and Bayesian Forecasting," *Journal of the American Statistical Association*, 80, 73–83.

West, M., Muller, P., and Escobar, M. D. (1994), "Hierarchical priors and mixture models, with applications in regression and density estimation," *In: Aspects of Uncertainty: A Tribute to D.V. Lindley*, 363–386.

White, S., Kypraios, T., and Preston, S. (2015), "Piecewise approximate Bayesian computation: Fast inference for discretely observed Markov models using a factorised posterior distribution," *Statistics and Computing*, 25, 289–301.

WHO (2015), "Global tuberculosis report 2015," Tech. rep., World Health Organization.

Wirth, T., Hildebrand, F., Allix-Béguec, C., Wölbeling, F., Kubica, T., Kremer, K., van Soolingen, D., Rüsch-Gerdes, S., Locht, C., Brisse, S., Meyer, A., Supply, P., and Niemann, S. (2008), "Origin, spread and demography of the *Mycobacterium tuberculosis* complex." *PLoS Pathogens*, 4, e1000160.

Wood, S. N. (2010), "Statistical inference for noisy nonlinear ecological dynamic systems," in *Nature*, eds. Dey, D. and Yan, J., Chapman and Hall/CRC Press, vol. 466, pp. 1102–1104.

Yildirim, S., Dean, T., and Jasra, A. (2013), "Parameter Estimation in Hidden Markov Models with Intractable Likelihoods Using Sequential Monte Carlo," *Journal of Computational and Graphical Statistics*, 8600, 1–22.

Zhao, Y., Xu, S., Wang, L., Chin, D. P., Wang, S., Jiang, G., Xia, H., Zhou, Y., Li, Q., Ou, X., et al. (2012), "National survey of drug-resistant tuberculosis in China," *New England Journal of Medicine*, 366, 2161–2170.