# Bayesian methods for multi-modal posterior topologies

by

**Biljana Jonoska Stojkova**

M.Sc., Ss Cyril and Methodius University, 2011
B.Sc., Ss Cyril and Methodius University, 2003

Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of
Doctor of Philosophy

in the
Department of Statistics and Actuarial Science
Faculty of Science


**SIMON FRASER UNIVERSITY**
**Spring 2017**

# Approval

| | |
|---|---|
| **Name:** | **Biljana Jonoska Stojkova** |
| **Degree:** | **Doctor of Philosophy (Statistics)** |
| **Title:** | ***Bayesian methods for multi-modal posterior topologies*** |
| **Examining Committee:** | **Chair:** Dr. Tim Swartz<br>Professor |

**Dr. David A. Campbell**
Senior Supervisor
Associate Professor _____

**Dr. Liangliang Wang**
Supervisor
Assistant Professor _____

**Dr. Derek Bingham**
Internal Examiner
Associate Professor _____

**Dr. Russell Steele**
External Examiner
Associate Professor
Department of Mathematics and
Statistics
McGill University _____

**Date Defended:** 18 April 2017

# Abstract

The purpose of this thesis is to develop efficient Bayesian methods to address multi-modality in posterior topologies. In Chapter 2 we develop a new general Bayesian methodology that simultaneously estimates parameters of interest and probability of the model. The proposed methodology builds on the Simulated Tempering algorithm, which is a powerful sampling algorithm that handles multi-modal distributions, but it is difficult to use in practice due to the requirement to choose suitable prior for the temperature and temperature schedule. Our proposed algorithm removes this requirement, while preserving the sampling efficiency of the Simulated Tempering algorithm. We illustrate the applicability of the new algorithm to different examples involving mixture models of Gaussian distributions and ordinary differential equation models.

Chapter 3 proposes a general optimization strategy, which combines results from different optimization or parameter estimation methods to overcome shortcomings of a single method. Embedding the proposed optimization strategy in the Incremental Mixture Importance Sampling with Optimization algorithm (IMIS-Opt) significantly improves sampling efficiency and removes the dependence on the choice of the prior of the IMIS-Opt. We demonstrate that the resulting algorithm provides accurate parameter estimates, while the IMIS-Opt gets trapped in a local mode in the case of the ordinary differential equation (ODE) models. Finally, the resulting algorithm is implemented within the Approximate Bayesian Computation framework to draw likelihood-free inference.

Chapter 4 introduces a generalization of the Bayesian Information Criterion (BIC) that handles multi-modality in the posterior space. The BIC is a computationally efficient model selection tool, but it relies on the assumption that the posterior distribution is unimodal. When the posterior is multi-modal the BIC uses only one posterior mode, while discarding the information from the rest of the modes. We demonstrate that the BIC produces inaccurate estimates of the posterior probability of the bimodal model, which in some cases results in the BIC selecting the sub-optimal model. As a remedy, we propose a Multi-modal BIC (MBIC) that incorporates all relevant posterior modes, while preserving the computational efficiency of the BIC. The accuracy of the MBIC is demonstrated through bimodal models and mixture models of Gaussian distributions.

**Keywords:** Simulated tempering, importance sampling, optimization, model selection, thermodynamic integration, synthetic likelihood.

# Dedication

To my beloved husband Jane, my lovely family and friends

# Acknowledgements

I would like to thank all the people who made this thesis work possible. First and foremost, I have deepest gratitude and admiration for my supervisor, Dr. Dave Campbell, without whom this work would not have been possible. Thank you for being the greatest source of inspiration, for believing in me when I could not believe in myself, for teaching me to think broadly and deeply, and for being patient when my research was going slow. You are so much more than a supervisor, you are a friend. I will always be grateful for providing collaborative opportunities with academia and the local industry, for encouraging me to present our work at many conferences and seminars, and for guiding me through the process of writing.

I am thankful to the faculty members and the staff in the Department of Statistics and Actuarial Science for providing an excellent atmosphere to grow both professionally and personally. My deep appreciation goes to Dr. Richard Lockhart for sharing wisdom and knowledge during those stressful days before the comprehensive exam; to Dr. Tim Swartz and Dr. Rachel Altman for their helpful advices during difficult times; to Dr. Derek Bingham, for hosting me in the lab and for providing me with computational resources and amazing officemates. I gratefully acknowledge the support from Dr. Jinko Graham and Dr. Brad McNeney, to whom I owe the opportunity of doing the PhD in the Department. Many thanks to Kelly Jay, Charlene Bradbury and Sadika Jungic who were always happy to help with the administrative tasks, and to Sinisa Milosavljevic and Steve Obadia, who unhesitatingly helped at the times when my laptops crashed.

My gratitude extends to Dr. David Sivak for the opportunity to work on a x-ray crystallography problem, and for being supportive; to Dr. Jason Loepky and Dr. Shirin Golchi for the opportunity to collaborate with the Community Sift company. I am grateful to Dr. Luke Bornn for the constructive discussions on my papers, and to Dr. Jack Davis for providing feedback on my manuscripts. I would like to thank my committee members: Dr. Liangliang Wang, Dr. Derek Bingham and Dr. Russell Steele for the fruitful discussions on my thesis work, and to Dr. Tim Swartz for chairing my defense.

This thesis work was funded by the Department of Statistics and Actuarial Science, the Natural Sciences and Engineering Research Council (NSERC) of Canada, the Collaborative Research Grant (CRG) and the company FPIovations in Vancouver, the Mathematics of

Information Technology and Complex Systems (MITACS) Accelerate program and NSERC Engage involving the company Community Sift, and National Institutes of Health (NIH).

I would also like to thank the people who inspired me through earlier career days: Dr. Ljupcho Nastovski, Dr. Vesna Bucevska and Dr. Dragan Tevdovski who encouraged me to take my studies to a different level; and my wonderful colleagues and friends at the State Statistical Office in Macedonia.

The years spend in the PhD program were most enjoyable, for the most part thanks to the extremely friendly and pleasant atmosphere in the Department. To my fellow graduate students and friends Audrey, Shirin, Oksana, Ruth, Ofir, Mike, Huijing, Luyao, Maude, Ben, Charlie, Lassantha, Elena, Yi, Jack, Fei, Abdollah, Pulindu, Tianyu, Andrew, Charith, I want to say thank you for being there for me in difficult times; for cheering up my days through many dinners, concerts, hikes, trips and journal clubs. My appreciation goes to many more friends whose names should appear here, and to whom I would like to say thank you for being part of my life.

Finally, special thanks to my beloved husband who has always been by my side during the rainy and sunny days: there is a lot of you in this thesis. To the rest of my beloved family, my parents Jonche and Borka who instilled in me a love of math at an early age, Milena, Sasho, Angela, my in-law parents Todor and Ivanka, Aleksandra and Svetlana: thank you for your support and encouragement during my studies and for numerous skype conversations filled with love and positive energy.

# Contents

**Approval** ii

**Abstract** iii

**Dedication** v

**Acknowledgements** vi

**Table of Contents** viii

**List of Tables** xii

**List of Figures** xiii

**1 Introduction** 1

1.1 Overview . . . . . 1

1.1.1 Simulated Tempering Without Normalizing Constants . . . . . 1

1.1.2 Incremental Mixture Importance Sampling with Shotgun optimization 2

1.1.3 Multi-modal Bayesian Information Criterion . . . . . 3

1.2 Organization of the thesis . . . . . 4

**2 Parallel Tempering via Simulated Tempering Without Normalizing Constants** 5

2.1 Introduction . . . . . 5

2.2 Tempering Methods . . . . . 7

2.2.1 Parallel tempering . . . . . 7

2.2.2 Simulated tempering . . . . . 8

2.2.3 Outline of the Standard Simulated Tempering . . . . . 9

2.2.4 Related algorithms . . . . . 10

2.3 Simulated Tempering Without Normalizing Constants . . . . . 11

2.4 Parallel Tempering via Simulated Tempering with Normalizing Constants algorithm . . . . . 13

2.5 Estimation of Marginal Likelihoods via thermodynamic integration . . . . . 16

2.5.1 Computing the marginal likelihood via STWNC . . . . . 17
2.6 Example 1: Two-dimensional multi-modal mixture Gaussian model . . . . . 19
2.6.1 Results . . . . . 19
2.7 Example 2: Bimodal model . . . . . 19
2.7.1 Results . . . . . 20
2.7.2 Marginal likelihood estimation . . . . . 21
2.8 Example 3: Galaxy data . . . . . 23
2.8.1 Results . . . . . 24
2.8.2 Marginal likelihood estimation for the Galaxy data . . . . . 25
2.9 Example 4: Susceptible-Infected-Recovered (SIR) epidemiological model . . 26
2.9.1 Results . . . . . 28
2.10 Discussion . . . . . 28

**3 Incremental Mixture Importance Sampling with Shotgun optimization 32**
3.1 Introduction . . . . . 32
3.2 Shotgun optimization . . . . . 34
3.3 Incremental Mixture Importance Sampling with Optimization . . . . . 35
3.4 Incremental Mixture Importance Sampling with Shotgun optimization . . . 36
3.5 Ordinary differential equation models . . . . . 38
3.5.1 Motivating example – the FitzHugh-Nagumo ODE model . . . . . 39
3.6 Illustrative example – the FitzHugh-Nagumo model revisited . . . . . 41
3.6.1 Shotgun optimization strategy used to estimate the parameters in the FhN model . . . . . 41
3.6.2 Performance of the Shotgun optimization strategy in the FhN-ODE model. . . . . 44
3.6.3 Results . . . . . 45
3.7 Illustrative example – Susceptible-Infected-Removed (SIR) epidemiological model . . . . . 46
3.7.1 Shotgun optimization strategy used in SIR-ODE model . . . . . 47
3.7.2 Results . . . . . 47
3.8 Parameter estimation with IMIS-ShOpt using synthetic likelihood . . . . . 47
3.8.1 ABC framework . . . . . 48
3.8.2 IMIS-ShOpt with synthetic likelihood outline . . . . . 50
3.9 Illustration of the IMIS-ShOpt with synthetic likelihood performance through a chaotic stochastic model . . . . . 53
3.9.1 The Shotgun optimization . . . . . 54
3.9.2 Results . . . . . 54
3.10 Discussion . . . . . 55

**4 Bayesian Information Criterion (BIC) for Multimodal Distributions 58**

4.1 Introduction . . . 58
4.2 Bayesian approach to model selection and comparison . . . 60
4.3 Bayesian Information Criterion and Laplace Approximation . . . 60
4.3.1 Motivating example . . . 62
4.4 The Multi-modal BIC . . . 64
4.4.1 Illustration of the MBIC . . . 65
4.5 Model selection in mixture of Gaussian models . . . 65
4.5.1 The label switching problem . . . 66
4.5.2 Model selection performance of the MBIC, the LA and the BIC . . . 67
4.6 Discussion . . . 70

**5 Conclusion** **73**

**Bibliography** **75**

**Appendix A Analytical calculation of the marginal likelihood** **82**

**Appendix B Implementation, bimodal model** **83**

**Appendix C Implementation PT-STWNC, SIR model** **84**

**Appendix D Implementation of IMIS-Opt and IMIS-ShOpt, SIR model** **85**

**Appendix E Derivation of the MBIC** **88**

**Appendix F The unimodal model** **91**
F.0.1 The Laplace approximation (LA) in the unimodal model . . . 92
F.0.2 The MBIC in the unimodal model . . . 93

**Appendix G Scenario 1. Bimodal model with bimodal prior and unimodal likelihood** **94**
G.0.1 Analytical solution to the marginal likelihood in Scenario 1 . . . 95
G.0.2 The Laplace approximation (LA) in Scenario 1 . . . 96
G.0.3 The MBIC in Scenario 1 . . . 96
G.0.4 Bias in the MBIC, Scenario 1 . . . 97

**Appendix H Scenario 2. Bimodal model with bimodal likelihood and unimodal prior** **100**
H.0.1 Analytical solution to the marginal likelihood in Scenario 2 . . . 101
H.0.2 The Laplace approximation in Scenario 2 . . . 101
H.0.3 The MBIC in Scenario 2 . . . 101
H.0.4 Bias in the MBIC, Scenario 2 . . . 102

**Appendix I Discovering all the important modes** **104**

# List of Tables

Table 2.1 Parameter estimates v.s. theoretical values . . . . . 22
Table 2.2 Marginal log-likelihood of the bimodal model . . . . . 23
Table 2.3 Log marginal likelihood estimates of Galaxy data . . . . . 26
Table 2.4 log Bayes factors . . . . . 27

Table 3.1 The two FhN models – prior specifications . . . . . 41
Table 3.2 Computational time . . . . . 45
Table 3.3 Computational time . . . . . 47

Table 4.1 Number of modes discovered by ShOpt. . . . . 68

# List of Figures


Figure 2.1 Two-dimensional mixture of 20 Gaussian distributions model – contour plots of sampled $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$ obtained from the PT-STWNC 'tempered' chain (plot A), and 'target' chain (plot B). The plot C is generated by evaluating the model in (2.22) over the grid points from the plot B. The red dots denote the locations of the modes $\boldsymbol{\mu}$. 20

Figure 2.2 The bimodal model – sampled joint posterior distribution of $\mu$ and $\tau$ obtained from PT-STWNC 'tempered' chain: A perspective plot of the joint posterior distribution of $\mu$ and $\tau$, and B the corresponding contour plot. . . . . 21

Figure 2.3 The bimodal model – marginal posterior distributions of $\mu, \tau$ and $\sigma^2$. Distributions in gray, red, blue and green color correspond to samples obtained from the PT-STWNC 'tempered' chain, the PT-STWNC 'target' chain, the theoretical distribution and PAWL within ST, respectively. . . . . 22

Figure 2.4 Galaxy data, model with three components unequal variances – marginal posterior distributions of the sampled parameters $\mu_1, \mu_2, \sigma_1^2, \tau$ and bivariate plot of $(\mu_1, \mu_2)$. Gray and red color correspond to samples obtained from the 'tempered' and 'target' chain, respectively. The transparent violet color corresponds to the geometric temperature schedule used to obtain marginal likelihood estimate from TI via PT. . . . . 25

Figure 2.5 SIR model – marginal (diagonal) and bivariate joint (off-diagonal) posterior distributions of sampled parameters $\alpha, \beta$ and $I(0)$ obtained from the 'target' chain. . . . . 29

Figure 2.6 The SIR model – marginal posterior distributions of sampled parameters $\alpha, \beta, I(0)$ and $\tau$ obtained from the 'tempered' chain. . . . . 30

Figure 2.7 The SIR model – parameter space of the prior (gray color) versus the posterior space (red color) of the parameters in SIR model. The plot A shows contours of the joint prior distribution of the parameters $\alpha$ and $\beta$. The small red dots close to the origin correspond to the joint posterior distribution. Parameter space of the prior and posterior distribution of the parameter $I(0)$ are shown in the plot B. . . . . 30

Figure 3.1 The FhN-ODE model – impact of the disagreement between the log-likelihood and log posterior (plots A and B) and log prior (plot C) on the IMIS-Opt posterior estimate (plot D). The IMIS-Opt was run with D=3, B=1000 and J=10000. . . . . 40

Figure 3.2 The FhN-ODE model – re-sampled trajectories using IMIS-Opt on Model 1 (plot A), IMIS-ShOpt on Model 1 (plot B) and IMIS-ShOpt on Model 2 (plot C). The gray lines represent 10000 re-sampled trajectories, the solid thick blue and green thin lines correspond to the re-sampled trajectories at the posterior mean values for the state variables V and R, respectively. The red points represent the data, which were simulated from the vector of true parameters values $\boldsymbol{\theta} = (a = 0.2, b = 0.2, c = 3, V(0) = -1, R(0) = 1)'$. The IMIS-Opt was run with D=3, B=1000 and J=10000. The IMIS-ShOpt for both models, Model 1 and Model 2, was run with D=30, Q=3, B=1000 and J=10000. . . . . 45

Figure 3.3 The SIR-ODE model – marginal and bivariate joint posterior distributions of sampled parameters $\alpha, \beta$ and $I(0)$ obtained from the IMIS-Opt. The IMIS-Opt algorithm was run with $N_0 = 3000, D = 3, B = 1000, J = 10000, N = 1000$ (see Appendix D for implementation details). . . . . 48

Figure 3.4 The SIR-ODE model – marginal (diagonal) and bivariate joint (off-diagonal) posterior distributions of sampled parameters $\alpha, \beta$ and $I(0)$ obtained from the IMIS-ShOpt. The IMIS-ShOpt algorithm was run with $N_0 = 3000, Q = 10, D = 3, B = 1000, J = 10000, N = 1000$. . . 49

Figure 3.5 The theta-Ricker model – marginal posterior distributions of the parameters obtained from the final re-sampling stage. The vertical lines are drawn at the posterior mean (blue dashed) and the true value (red dotted). The gray distributions represent the priors. . . 55

Figure 3.6 The theta-Ricker model – weights of the particles in the importance sampling distribution before re-sampling. The vertical lines are drawn at the true parameter values. . . . . 56

Figure 4.1 Analytical marginal likelihood and the LA: A. Model in Scenario 1 – unimodal likelihood, bimodal prior; B. Model in Scenario 2 – bimodal likelihood and unimodal prior. . . . . 63
Figure 4.2 Analytical marginal likelihood and the MBIC: A. Model in Scenario 1 – unimodal likelihood, bi-modal prior; B. Model in Scenario 2 – bi-modal likelihood and unimodal prior. . . . . 66
Figure 4.3 Log marginal likelihood of the models according to the MBIC, the BIC at the global mode, the LA at the global mode, the PT-STWNC, the TI-PT-B and the TI-PT-NB. Studied models are given on the x-axis, log marginal likelihoods of the models according to the six studied model selection methods are given on the y axis. The bigger log marginal likelihood the better the model. . . . . 69
Figure 4.4 Galaxy data – Pairwise model comparisons using: A. the LA evaluated at each of the discovered modes; B. the BIC evaluated at each of the discovered modes. Each cell of the heat-maps correspond to the proportion of modes, $p_{M_i,M_j}$, (obtained as per (4.25)) at which the model $M_i$ from the y-axis was selected over the model $M_j$ from the x-axis, for $i, j \in \{1, .., 5\}$. . . . . 71

# Chapter 1

# Introduction

## 1.1 Overview

### 1.1.1 Simulated Tempering Without Normalizing Constants

When it comes to sampling from multi-modal distributions, single chain MCMC methods fail to efficiently explore the posterior space. Random walk variants, such as Simulated Tempering (ST) (Marinari and Parisi, 1992; Geyer and Thompson, 1995) and Parallel Tempering (PT) (Swendsen and Wang, 1986), have been proposed to efficiently sample from multi-modal target distributions. Sampling distribution of the tempering methods is defined by a sequence of distributions that bridge between the prior and the target distribution. The sequence of distributions is indexed at the inverse temperature, which is defined on the interval $[0, 1]$ and governs the influence of the likelihood to the posterior distribution. At the lower extreme of the distributions sequence, small values of the inverse temperature reduce the effect of the likelihood, and the resulting posterior distribution is close to the prior. This enables the sampler to move easily between the posterior modes. At the upper extreme of the distributions sequence, the resulting posteriors are approaching the target distribution, and the sampler has difficulties moving between the modes.

In standard ST, the inverse temperature is a discrete dynamic variable which is updated together with the parameters of interest. In order to ensure sampling of the temperature parameter, the standard ST requires from the user to perform preliminary runs to learn the prior for the inverse temperature and the inverse temperature schedule. This contributes in large to unpopularity of the standard ST in practice. We present a new Simulated Tempering algorithm that removes the requirement of the prior for the inverse temperature and the temperature schedule by introducing a continuous inverse temperature. We refer to the new algorithm as Simulated Tempering Without Normalizing Constants (STWNC).

The inverse temperature parameter is a nuisance parameter, and hence, its prior can be chosen for algorithmic convenience. We derive a formula for the prior of the inverse

temperature by imposing a constraint that profile distribution of the inverse temperature parameter over the parameters of interest is standard uniform.

Similar to the ST, the PT uses a sequence of distributions defined at inverse temperature. In PT, independent chains are run at different temperatures, while allowing the information between the chains to flow. Since the inverse temperature in PT and standard ST is discrete, the samples from the target distribution correspond to inverse temperature value of one. However, when the inverse temperature is continuous, obtaining samples from the target distribution is difficult because the event of continuous inverse temperature taking value of one has probability zero. To overcome this issue, we run PT with two chains, each having its own purpose. The first chain, which runs a STWNC chain, samples the parameters of interest at different temperatures, and the second chain draws samples from the target distribution using the Metropolis-Hastings. We refer to this hybrid algorithm as Parallel Tempering via Simulated Tempering Without Normalizing Constants (PT-STWNC).

Samples from the STWNC, which are updated at different inverse temperature values, can be used to obtain marginal likelihood estimates that are crucial for calculation of the Bayes factors and model selection thereof. Thermodynamic integration has been proposed (Friel and Pettitt, 2008; Calderhead and Girolami, 2009) to estimate marginal likelihood using samples from PT. Due to the discrete nature of the inverse temperature in PT, the thermodynamic integration requires temperature schedule to be known. We apply thermodynamic integration method to the samples from the STWNC with an aim to obtain marginal likelihood estimates. In STWNC the inverse temperature is continuous, and therefore, thermodynamic integration via STWNC removes the requirement for a temperature schedule, thus producing marginal likelihood estimates with negligible computational cost. Performance of the PT-STWNC is illustrated by mixture models of Gaussian distributions and an epidemiological ordinary differential equation model.

### 1.1.2 Incremental Mixture Importance Sampling with Shotgun optimization

Posterior topologies chmoniaracterized with many unimportant minor modes, ridges and ripples impose challenges to the sampling algorithms. Incremental Mixture of Importance Sampling with Optimization (IMIS-Opt) (Raftery and Bao, 2010) is designed to fully explore posterior spaces with multiple modes. However, the success of the IMIS-Opt depends on the choice of the initial importance distribution, and if the initial importance distribution disagrees with the posterior, the IMIS-Opt might miss important posterior modes. Often the prior serves as initial importance distribution, and if there are no initial samples from the prior in the unexplored regions, then the IMIS-Opt would have difficulties to cover these regions. To avoid this problem, one could choose a diffuse prior, but this implies that the prior is chosen for the algorithmic convenience rather than to represent the expert knowledge.

To remove the dependence of the IMIS-Opt on the choice of the prior, we modify the IMIS-Opt by replacing the optimization stage with a multiple-method optimization strategy. Consequently, we propose a general multiple-method optimization strategy that relies on the concept that no single method is the best in every problem (Wolpert and Macready, 1997). A single estimation method introduces a certain model relaxation which leads to exploring a certain region of the posterior topology. The proposed general optimization strategy combines results from different methods to overcome the shortcomings of a single method, and to discover all the important posterior modes thereof. We refer to the proposed general optimization strategy as a Shotgun optimization, and to the resulting algorithm as Incremental Mixture of Importance Sampling with Shotgun optimization (IMIS-ShOpt).

We apply the IMIS-ShOpt in variety of frameworks involving ordinary differential equation models (ODE) and chaotic stochastic difference equation models. Parameter estimation in ODE models is particularly challenging because their posterior spaces are characterized with many unimportant modes, ridges and ripples. We demonstrate through ODE model, that if the prior covers only one unimportant mode of the posterior, the IMIS-Opt gets trapped in the unimportant mode, while the IMIS-ShOpt successfully explores the full posterior space. The other framework in which the IMIS-ShOpt is applied using ideas from Approximate Bayesian Computation (ABC) (Tavaré et al., 1997; Pritchard et al., 1999). We demonstrate its parameter estimation performance as an likelihood-free approach.

### 1.1.3 Multi-modal Bayesian Information Criterion

The Bayesian Information Criterion (BIC) (Schwarz et al., 1978) is a computationally efficient model selection tool, because it approximates the posterior probability of the model while avoiding Monte Carlo simulations over parameter space. However, model selection from the BIC is based on the Laplace approximation (LA), which relies on the assumption that the posterior distribution is unimodal. Hence, when the posterior distribution is multi-modal the BIC inaccurately estimates the posterior probability of the model, which may result in selecting the incorrect model. We demonstrate through analytical calculations that the LA produces biased estimates of the posterior probability of the model in a simple bimodal case.

As a remedy we propose a Multi-modal BIC (MBIC) which extends the model selection abilities of the LA and the BIC to multi-modal posterior spaces. The proposed MBIC handles the multi-modality problem by taking into account all the important modes in the posterior parameter space. We use analytical derivations to show that the MBIC correctly estimates the posterior probability of the bimodal model with well separated modes. The MBIC has the advantage of exploiting all relevant posterior modes, which is a step closer to the fully exploring the posterior parameter space, while retaining the computational efficiency of the BIC.

Model selection performance of the MBIC, the LA and the BIC in the case of mixture model of Gaussian distributions with unknown number of components demonstrates that the BIC and the LA choose the correct model when evaluated at the global mode. However, when the LA and the BIC are evaluated at the local mode characterized with much smaller probability mass compared to the other modes, the LA and the BIC might fail to choose the correct model. The MBIC chooses the correct model. In order to obtain all the important posterior modes needed for calculation of the MBIC, we develop an algorithm which combines results from different runs of an optimizer started at different points. This strategy has the ability to explore different regions in the posterior topology which results in finding all the relevant posterior optima.

## 1.2 Organization of the thesis

The remainder of the thesis is organized as follows. In Chapter 2 we describe the proposed PT-STWNC algorithm after a brief overview of the tempering methods. Additionally, thermodynamic integration methodology for model selection is introduced and the PT-STWNC is applied to few examples. Mathematical derivations and implementation details are placed in Appendices A, B and C. In Chapter 3, before we propose the Shotgun optimization strategy, and the IMIS-ShOpt algorithm, we provide a brief introduction on the IMIS-Opt. The performance of the new compared to the existing methodology is illustrated through several examples. In addition, the IMIS-ShOpt with synthetic likelihood is described after a short introduction of the ABC framework, and the new algorithm is illustrated through a chaotic stochastic model. Chapter 4 is dedicated to introducing the MBIC, and studying its model selection abilities in comparison to the LA and the BIC. Mathematical derivations and the proposed algorithm for discovering all important modes are given in the Appendices E-I. The thesis is concluded in Chapter 5.

# Chapter 2

# Parallel Tempering via Simulated Tempering Without Normalizing Constants

## 2.1 Introduction

Basic random walk Markov Chain Monte Carlo (MCMC) methods are inefficient when faced with multi-modal posterior distributions, especially when modes are isolated by large gaps of low probability. Simulated Tempering (Marinari and Parisi, 1992; Geyer and Thompson, 1995) and Parallel Tempering (Swendsen and Wang, 1986; Geyer, 1991; Hukushima and Nemoto, 1996), are MCMC variants designed to ease the challenges of multi-modal distributions by introducing a temperature parameter to overcome the prohibitively low probability regions which otherwise trap samplers in local modes (Zhang and Ma, 2008). Sampling occurs at different temperatures, balancing short distance within-mode steps and longer distance between-mode steps.

Both Parallel Tempering (PT) and Simulated Tempering (ST) define a sequence of distributions for the vector of data $\boldsymbol{Y} \in \mathbb{R}^N$ and parameter $\boldsymbol{\theta} \in \mathbb{R}^d$, indexed by an inverse temperature $\tau \in [0, 1]$, along a path between the prior and the target distribution usually defined by

$$P(\boldsymbol{\theta} \mid \tau, \boldsymbol{Y}) \propto P(\boldsymbol{Y} \mid \boldsymbol{\theta}, \tau) P(\boldsymbol{\theta}) = P(\boldsymbol{Y} \mid \boldsymbol{\theta})^{\tau} P(\boldsymbol{\theta}).$$

At the extremes of the distribution sequence, $P(\boldsymbol{\theta} \mid \tau = 0, \boldsymbol{Y}) = P(\boldsymbol{\theta})$ is the prior, and $P(\boldsymbol{\theta} \mid \tau = 1, \boldsymbol{Y}) \propto P(\boldsymbol{Y} \mid \boldsymbol{\theta}) P(\boldsymbol{\theta})$ is the usual target distribution. Consequently, when $\tau = 0$ it is easy for the sampler to explore the parameter space, but at $\tau = 1$ the sampler explores the more challenging target distribution. Markov Chains in both PT and ST therefore avoid becoming trapped in a single mode. In PT, $T$ independent chains are run at different temperatures where information about $\boldsymbol{\theta}$ is allowed to flow between the $T$ chains.

ST samples the system at different temperatures using a single chain by augmenting the state space with the temperature parameter as a dynamic variable. In both standard ST and PT, samples of interest correspond to samples obtained at $\tau = 1$ i.e., $P(\boldsymbol{\theta} \mid \boldsymbol{Y}, \tau = 1)$.

Marginal likelihood estimates, which are crucial to Bayesian model comparison and selection, can be obtained using thermodynamic integration (TI) based on the samples obtained in all $T$ chains of PT (Friel and Pettitt, 2008; Calderhead and Girolami, 2009). TI approximates the log marginal likelihood, $\log P(\boldsymbol{Y})$ , by numerically solving the integral $\int_0^1 E_{\boldsymbol{\theta}|\tau,\boldsymbol{Y}}[log(P(\boldsymbol{Y} \mid \boldsymbol{\theta}))]d\tau$. As with any numerical integration scheme, accuracy of the integral depends on the discretization, here the number and location of $\tau$ values used in PT. Approximating the thermodynamic integral using PT approximations produces biased marginal likelihood estimates (Calderhead and Girolami, 2009). In this chapter, we propose a new ST algorithm that solves the Thermodynamic Integral without needing the user to select and tune the number and location of discrete $\tau$ values.

In order to update the inverse temperature parameter, standard ST requires the user to select a prior for $\tau$. Choice of $P(\tau)$ typically requires estimates of the normalizing constant, $z(\boldsymbol{Y} \mid \tau) = \int_{\Theta} P(\boldsymbol{Y} \mid \boldsymbol{\theta})^{\tau} P(\boldsymbol{\theta}) d\boldsymbol{\theta}$, which reduces to a finite dimensional problem if ST is performed over a fixed discretized sequence of $\tau$ values rather than treating $\tau$ as a continuous variable. Consequently, in standard ST, the temperature schedule and the normalizing constants must be estimated through preliminary runs.

Several other methods have been proposed for obtaining normalizing constants (Geyer and Thompson, 1995): iterative adjustment, Metropolis-Coupled Markov Chain Monte Carlo (MCMCMC) (Geyer, 1991), and stochastic approximation (Wasan, 1969; Wang and Landau, 2001). Stochastic approximation could be embedded within standard ST to adaptively estimate the normalizing constants algorithm, but this still requires preliminary runs to learn the temperature schedule (Atchade and Liu, 2004). Extending this further, the Parallel Adaptive Wang-Landau (PAWL) can be embedded in ST (Bornn et al., 2013; Bornn, 2014) to automatically learn the temperature schedule using an adaptive binning strategy. This approach is well suited to parameter estimation but does not resolve the discretization concerns when applied to TI.

Instead of performing preliminary runs to obtain the most suitable discretized temperature schedule and normalizing constants thereof, we propose a new ST algorithm with a continuous inverse temperature variable defined on [0,1]. We remove the requirement for calculating normalizing constants through imposition of the constraint that the distribution of the $\tau$ when profiling over $\boldsymbol{\theta}$ is Uniform. This constraint defines a formula for $P(\tau)$. In the rest of the chapter we refer to the new ST algorithm as 'Simulated Tempering Without Normalizing Constants' (STWNC). By sampling across a continuous temperature scale, samples from STWNC can be used for thermodynamic integral estimates of the marginal likelihood.

In PT and ST, samples from the target distribution are retained whenever $\tau = 1$ which is possible (though typically inefficient in ST) because of the discretization of the domain of $\tau$. However, in STWNC the temperature parameter is a continuous variable and therefore $P(\tau = 1) = 0$. To sample from the target distribution while maintaining a continuously valued $\tau$, we embed STWNC within the PT framework. The resulting PT and STWNC hybrid algorithm named 'Parallel Tempering - Simulated Tempering Without Normalizing Constants' (PT-STWNC), runs PT with $T = 2$ chains, each targeting a different goal. The first chain draws samples at continuous temperatures via STWNC. We refer to the first chain as a 'tempered' chain. The second chain draws samples from the target distribution where $\tau = 1$. We refer to the second chain as a 'target' chain in the rest of our presentation.

The remainder of the chapter is organized as follows. Section 2.2 gives an overview of tempering methods with emphasis on PT and standard ST as well as a review of modern variants of stochastic approximation. Section 2.3 proposes a new ST (STWNC) algorithm and provides detailed explanation of how STWNC removes the need for temperature dependent normalizing constants. It also introduces marginal likelihood estimation via STWNC. In Section 2.4 the hybrid PT-STWNC, which combines the two powerful sampling tempering algorithms Parallel Tempering and STWNC, is proposed. Section 2.5 presents an overview of marginal likelihood approximation via thermodynamic integration. Sections 2.6 -2.9 illustrate the PT-STWNC algorithm using several examples: a mixture model of two Gaussian distributions, a mixture model applied to Galaxy velocity data, a two-dimensional mixture model of twenty Gaussian distributions and a Susceptible-Infected-Recovered (SIR) epidemiological ordinary differential Equations (ODE) model. Section 2.10 provides discussion.

## 2.2 Tempering Methods

This section outlines Parallel and Simulated Tempering methods as well as their role in thermodynamic integration for marginal likelihood estimation.

### 2.2.1 Parallel tempering

Parallel Tempering (PT), also known as replica exchange or Population MCMC, is a sampling algorithm designed to improve the dynamic properties of the MCMC samplers especially when it comes to exploring the posterior surface with many isolated modes.

Given the likelihood $P(\mathbf{Y} \mid \boldsymbol{\theta})$, and prior distribution $P(\cdot)$, the inverse temperature sequence $0 \leq \tau_1 <, ..., < \tau_T = 1$ defines the $T$ approximations to the target posterior distribution:

$$P_t(\boldsymbol{\theta} \mid \mathbf{Y}) = \frac{P(\mathbf{Y} \mid \boldsymbol{\theta})^{\tau_t} P(\boldsymbol{\theta})}{z(\mathbf{Y} \mid \tau_t)}, \; t \in \{1, .., T\}, \tag{2.1}$$

where

$$z(\mathbf{Y} \mid \tau_t) = \int_{\boldsymbol{\Theta}} P(\mathbf{Y} \mid \boldsymbol{\theta})^{\tau_t} P(\boldsymbol{\theta}) d\boldsymbol{\theta} \tag{2.2}$$

is the temperature dependent normalizing constant of the $t^{th}$ approximation to the target posterior distribution $P_t(\boldsymbol{\theta} \mid \mathbf{Y})$. The parameter $\tau$ controls the contribution of the likelihood to the equation (2.1), thus enabling the sampler to move easily and explore the parameter space when $\tau$ is low-valued and to remain further within the basin of attraction of a local mode when $\tau$ is high-valued.

At each iteration PT performs one of two steps: a mutation step where, for example, Metropolis-Hastings (MH) is used to jitter $\boldsymbol{\theta}$ from each of the $T$ chains independently, and an exchange step where, with some probability two chains $t$ and $l$ are randomly chosen to exchange their parameters $\boldsymbol{\theta}_t$ and $\boldsymbol{\theta}_l$. The exchange is accepted with probability:

$$\begin{aligned}
& \min\left\{1, \frac{P_t(\boldsymbol{\theta}_l \mid \mathbf{Y})}{P_t(\boldsymbol{\theta}_t \mid \mathbf{Y})} \frac{P_l(\boldsymbol{\theta}_t \mid \mathbf{Y})}{P_l(\boldsymbol{\theta}_l \mid \mathbf{Y})}\right\} \\
= & \min\left\{1, \frac{P(\mathbf{Y} \mid \boldsymbol{\theta}_l)^{\tau_t} P(\boldsymbol{\theta}_l) P(\boldsymbol{Y} \mid \boldsymbol{\theta}_t)^{\tau_l} P(\boldsymbol{\theta}_t) z(\boldsymbol{Y} \mid \tau_l) z(\mathbf{Y} \mid \tau_t)}{z(\mathbf{Y} \mid \tau_t) z(\mathbf{Y} \mid \tau_l) P(\mathbf{Y} \mid \boldsymbol{\theta}_l)^{\tau_l} P(\boldsymbol{\theta}_l) P(\mathbf{Y} \mid \boldsymbol{\theta}_t)^{\tau_t} P(\boldsymbol{\theta}_t)}\right\} \\
= & \min\left\{1, \frac{P(\mathbf{Y} \mid \boldsymbol{\theta}_l)^{\tau_t} P(\mathbf{Y} \mid \boldsymbol{\theta}_t)^{\tau_l}}{P(\mathbf{Y} \mid \boldsymbol{\theta}_l)^{\tau_l} P(\mathbf{Y} \mid \boldsymbol{\theta}_t)^{\tau_t}}\right\}.
\end{aligned} \tag{2.3}$$

Note the cancellation of normalizing constants $z(\mathbf{Y} \mid \tau_l)$ and $z(\mathbf{Y} \mid \tau_t)$ when swapping parameters between two different temperature based chains. The cancellation of normalizing constants and the ease of movement between modes at low values of $\tau$ has made PT into a widely used algorithm. Samples from the $T - th$ chain correspond to samples from the target distribution.

### 2.2.2 Simulated tempering

Simulated tempering (ST), also known as serial tempering, is a single chain sampling method where the posterior parameter space is augmented by including the temperature, $\tau$, as a random variable. As with PT, $\tau$ controls the influence of likelihood on $P(\boldsymbol{\theta} \mid \mathbf{Y}, \tau)$. However, in ST different temperatures are explored in a random walk through the joint distribution of $\boldsymbol{\theta}$ and $\tau$ :

$$P(\boldsymbol{\theta}, \tau \mid \mathbf{Y}) = \frac{P(\mathbf{Y} \mid \boldsymbol{\theta})^{\tau} P(\boldsymbol{\theta}) P(\tau)}{P(\mathbf{Y})}, \tag{2.4}$$

where the normalizing constant for the joint density is:

$$P(\mathbf{Y}) = \int_0^1 \int_{\boldsymbol{\Theta}} P(\boldsymbol{Y} \mid \boldsymbol{\theta})^{\tau} P(\boldsymbol{\theta}) P(\tau) d\boldsymbol{\theta} d\tau.$$

Similar to PT, $\tau$ controls the influence of prior in the equation (2.4), and samples that arise at $\tau = 1$ are samples from target distribution, however, when $\tau$ is continuous, $P(\tau = 1) = 0$.

#### 2.2.2.1 Prior for $\tau$

In order for standard ST to mix well, a carefully chosen prior $P(\tau)$ needs to be defined. Following Geyer and Thompson (1995), the prior for the inverse temperature can be found by examining the marginal distribution of $\tau$:

$$\begin{aligned} P(\tau \mid \boldsymbol{Y}) = \int_{\boldsymbol{\Theta}} P(\tau, \boldsymbol{\theta} \mid \boldsymbol{Y}) d\boldsymbol{\theta} \quad &\propto P(\tau) \int_{\boldsymbol{\Theta}} P(\boldsymbol{Y} \mid \boldsymbol{\theta})^{\tau} P(\boldsymbol{\theta}) d\boldsymbol{\theta} \\ &\propto P(\tau) z(\boldsymbol{Y} \mid \tau) \end{aligned} \tag{2.5}$$

where $z(\boldsymbol{Y} \mid \tau) = \int_{\boldsymbol{\Theta}} P(\boldsymbol{Y} \mid \boldsymbol{\theta})^{\tau} P(\boldsymbol{\theta}) d\boldsymbol{\theta}$ is $\tau$ dependent normalizing constant of the conditional posterior distribution $P(\boldsymbol{\theta} \mid \tau, \boldsymbol{Y})$. If the prior for $\tau$ is chosen to be approximately proportional to the inverse normalizing constant, that is, if

$$P(\tau) \propto \frac{1}{z(\boldsymbol{Y} \mid \tau)}, \tag{2.6}$$

then the marginal distribution of $\tau$ is $\tau \mid \boldsymbol{Y} \sim Uniform(0, 1)$. Following Geyer and Thompson (1995), the standard ST algorithm starts with $n$ tempered distributions at $\tau_i = \frac{i}{T}$ where $i = 0, .., n \leq T$, and iterates between adjusting the prior of $\tau$ and adjusting the inverse temperature spacing until a desired rate for transitions between the interpolating distributions is met. Afterwards, new inverse temperatures are added, and a new iterative cycle of adjusting prior and inverse temperature spacing is started for the newly added temperature set.

### 2.2.3 Outline of the Standard Simulated Tempering

A Markov Chain using standard ST updates parameters through Gibbs steps, updating $\boldsymbol{\theta} \mid \tau, \boldsymbol{Y}$ and $\tau \mid \boldsymbol{\theta}, \boldsymbol{Y}$ in turn. In the first kind of transition, a fixed temperature mutation step is typically a Metropolis Hastings step identically to what would be used to sample within one of the fixed temperature chains $P_t(\boldsymbol{\theta} \mid \boldsymbol{Y})$ in PT.

Updating $\tau \mid \boldsymbol{\theta}^{(i+1)}, \boldsymbol{Y}$ through Metropolis Hastings occurs by proposing a value $\tau^*$ sampled from a symmetric distribution (for expositional simplicity). Using the prior in (2.6), the value $\tau^*$ is accepted (set $\tau^{(i+1)} = \tau^*$) with probability:

$$
\begin{aligned}
& \min\left\{1, \frac{P(\boldsymbol{\theta}^{(i+1)}, \tau^* \mid \boldsymbol{Y})}{P(\boldsymbol{\theta}^{(i+1)}, \tau^{(i)} \mid \boldsymbol{Y})}\right\} \\
= \quad & \min\left\{1, \frac{P(\boldsymbol{Y} \mid \boldsymbol{\theta}^{(i+1)})^{\tau^*} P(\boldsymbol{\theta}^{(i+1)}) P(\tau^*) P(\boldsymbol{Y})}{P(\boldsymbol{Y} \mid \boldsymbol{\theta}^{(i+1)})^{\tau^{(i)}} P(\boldsymbol{\theta}^{(i+1)}) P(\tau^{(i)}) P(\boldsymbol{Y})}\right\} \\
= \quad & \min\left\{1, \frac{P(\boldsymbol{Y} \mid \boldsymbol{\theta}^{(i+1)})^{\tau^*} z(\boldsymbol{Y} \mid \tau^{(i)})}{P(\boldsymbol{Y} \mid \boldsymbol{\theta}^{(i+1)})^{\tau^{(i)}} z(\boldsymbol{Y} \mid \tau^*)}\right\}. \qquad (2.7)
\end{aligned}
$$

Consequently, the acceptance probability relies on temperature dependent normalizing constant $z(\boldsymbol{Y} \mid \tau)$. Normalizing constants, also referred to as weights of the sequence of distributions (Neal, 1996), are generally unknown and finding suitable values requires pilot runs – defeating the purpose of using standard ST. Information from pilot runs also cannot be recycled when additional data becomes available because $P(\tau)$ depends on $\boldsymbol{Y}$. To make the normalizing constant problem simpler, $\tau$ is discretized instead of being continuously valued.

### 2.2.4 Related algorithms

Wang-Landau (WL) algorithm (Wang and Landau, 2001), a stochastic approximation algorithm, could be used within standard ST to automatically obtain $P(\tau)$ (Geyer and Thompson, 1995; Atchade and Liu, 2004). Standard WL algorithm requires a predefined temperature schedule $0 \leq \tau_1 <, .., < \tau_T = 1$ corresponding to partitioning the temperature state space $\mathbb{T}$ into $T$ different regions i.e., bins $E_1, .., E_T$. The goal is to construct a chain that could spend the same time in each $E_t$. The moves inside the $E_t$ are performed with a standard Metropolis-Hastings (MH) algorithm with target distribution $\pi$. The WL algorithm recursively re-weights $\pi$ in $E_t$ by a factor $\phi_i(t)$. Hence, given $\tau^{(i)}$ and some unnormalized weights $\phi_i$, the $\tau^{(i+1)}$ can be sampled using a MH algorithm with invariant density proportional to $\sum_{t=1}^{T} \frac{\pi(\tau)}{\phi_i(t)} \mathbb{I}_{E_t}(\tau)$, where $\mathbb{I}$ is the indicator function. The normalizing constants at the $i$-th iteration $\phi_i(t) = z(\mathbf{Y} \mid \tau_t^{(i)})$ are updated for each of the bins $t \in \{1, .., T\}$ until a predefined criterion is met. This criterion, also known as 'flat histogram', ensures that proportions of visits of the chain in each of the bins $E_1, .., E_T$ are approximately equal to $T^{-1}$. The updating rule for the normalizing constants yields $\phi_{i+1}(t) = \phi_i(t)(1 + \gamma_i \mathbb{I}_{E_t}(\tau^{(i+1)}))$, where $\gamma_i$ is a learning rate which decreases stochastically until the criterion of 'flat histogram' of visit frequency to the bins $E_1, .., E_T$ is met.

The Parallel Adaptive WL (PAWL) algorithm (Bornn et al., 2013) removes the need for preliminary runs of WL to learn the optimal partitioning of the state space by exploiting an adaptive binning strategy. The adaptive binning strategy requires initial bins and bin

range to be specified by the user. Since $\phi_i(t)$ represent normalizing constant for the $t-th$ bin, the adaptive binning strategy maintains uniformity within a bin to allow within-bin movement. This is achieved by determining presence of heavy tails in the distribution of the samples within each bin. If the distribution is skewed towards the left side, then the sampler will have difficulty moving to the neighboring bin on the left. Hence, the binning strategy divides the bin into two chains by the middle point of the bin, and then it measures discrepancy between the chains using a ratio of the number of points in any of the two chains and the number of points within the bin. If this ratio is close to 50% then the histogram of the within-bin distribution is close to uniform. Hence, if the ratio is below some threshold, for example 25%, two new bins are created using the middle point of the former bin, and otherwise, the bin remains unchanged. One can specify the threshold to be 50%, but then the number of newly created bins will be larger. PAWL checks if the bins have to be split until the 'flat histogram' criterion is met, and afterwards the bin splitting stops since the sampler can move easily between the bins. Embedding PAWL within standard ST has been explored with a purpose to automate the two input requirements for ST: choice of temperature schedule and calculation of $P(\tau)$ (Bornn, 2014).

The Equi-Energy (EE) sampler (Kou et al., 2006), which utilizes temperature-energy duality, also allows wide moves by performing jumps between the states with similar energy levels. The EE, which is a powerful sampling and estimation methodology that addresses multi-modality in high-dimensional target distributions, provides estimates of expectations under any fixed temperature. However, the discrete nature of the temperature in EE sampler requires careful tuning of the temperature schedule in order for the EE samples to be applicable to thermodynamic integration.

## 2.3 Simulated Tempering Without Normalizing Constants

The standard ST is not widely used in practice, because of the challenges that arise from finding suitable values of the unknown normalizing constants. The proposed Simulated Tempering Without Normalizing Constants (STWNC) algorithm removes the dependence on normalizing constants in the acceptance ratio in (2.7) by the way we define the prior for $\tau$. As with ST, STWNC is still moving through two kinds of transitions: updating $(\boldsymbol{\theta} \mid \tau, \boldsymbol{Y})$; and updating $(\tau \mid \boldsymbol{\theta}, \boldsymbol{Y})$. $\tau$ is a nuisance parameter and not of inferential interest, consequently its prior $P(\tau)$ can be selected for algorithmic convenience. Our algorithm therefore chooses $P(\tau)$ by imposing a constraint that the profile posterior distribution of $\tau$ while maintaining $\boldsymbol{\theta} \mid \tau, \boldsymbol{Y}$ at its maximum value results in a uniform distribution. Using this constraint we derive a formula for the prior of $\tau$ which is computationally inexpensive and does not require preliminary runs of the algorithm or discretization of $\tau$. In the remainder of this section we describe the derivation of formula for the prior of $\tau$.

We impose a constraint that the posterior distribution of $\tau$ while profiling over $\boldsymbol{\theta}$ is uniform:

$$\boldsymbol{\theta}_{max}(\tau) = \arg\max_{\boldsymbol{\theta}} P(\boldsymbol{\theta} \mid \tau, \boldsymbol{Y}), \tag{2.8}$$

$$P_{prof}(\tau, \boldsymbol{\theta} \mid \boldsymbol{Y}) = P(\tau, \boldsymbol{\theta} = \boldsymbol{\theta}_{max}(\tau) \mid \boldsymbol{Y}) = U(0, 1). \tag{2.9}$$

The equation (2.9) implies that $P_{prof}(\tau, \boldsymbol{\theta} \mid \boldsymbol{Y})$ has a ridge of constant maximum height from $\tau = 0$ to $\tau = 1$. In other words for any $\tau_i, \tau_j \in [0, 1]$,

$$P(\tau_i, \boldsymbol{\theta}_i = \boldsymbol{\theta}_{max}(\tau_i) \mid \boldsymbol{Y}) = P(\tau_j, \boldsymbol{\theta}_j = \boldsymbol{\theta}_{max}(\tau_j) \mid \boldsymbol{Y}). \tag{2.10}$$

To derive the prior of $\tau$, we expand the profile posterior distribution given by the equation (2.9):

$$\begin{aligned} P_{prof}(\tau, \boldsymbol{\theta} \mid \boldsymbol{Y}) \quad &= P(\tau, \boldsymbol{\theta} = \boldsymbol{\theta}_{max}(\tau) \mid \boldsymbol{Y}) \\ &= \frac{P(\boldsymbol{Y} \mid \boldsymbol{\theta} = \boldsymbol{\theta}_{max}(\tau))^{\tau} P(\boldsymbol{\theta} = \boldsymbol{\theta}_{max}(\tau)) P(\tau)}{P_{prof}(\boldsymbol{Y})} \end{aligned} \tag{2.11}$$

where $P(\boldsymbol{Y} \mid \boldsymbol{\theta} = \boldsymbol{\theta}_{max}(\tau))^{\tau}$ is tempered profile likelihood, $P_{prof}(\boldsymbol{Y}) = \int_0^1 P(\boldsymbol{Y} \mid \boldsymbol{\theta} = \boldsymbol{\theta}_{max}(\tau))^{\tau} P(\boldsymbol{\theta} = \boldsymbol{\theta}_{max}(\tau)) P(\tau) d\tau$ is the normalizing constant of the profile posterior distribution of $\tau$ over $\boldsymbol{\theta}$, $P(\tau)$ is the prior of the inverse temperature and $P(\boldsymbol{\theta} = \boldsymbol{\theta}_{max}(\tau))$ is prior of $\boldsymbol{\theta}$ evaluated at $\boldsymbol{\theta}_{max}(\tau)$. Expressing $P(\tau)$ from (2.11) gives

$$P(\tau) = \frac{P_{prof}(\tau, \boldsymbol{\theta} \mid \boldsymbol{Y}) P_{prof}(\boldsymbol{Y})}{P(\boldsymbol{Y} \mid \boldsymbol{\theta} = \boldsymbol{\theta}_{max}(\tau))^{\tau} P(\boldsymbol{\theta} = \boldsymbol{\theta}_{max}(\tau))}. \tag{2.12}$$

As a direct consequence of the constraint that profile posterior distribution of $\tau$ while maintaining $\boldsymbol{\theta} \mid \tau, \boldsymbol{Y}$ at its maximum value is uniform, i.e., (2.9) and (2.10), the numerator in the formula for prior of $\tau$ in (2.12) is constant with respect to $\tau$.

Using the equation (2.12) as prior for $\tau$, the STWNC acceptance ratio for a proposed $\tau^*$ alters (2.7) into:

$$\min\left\{1, \frac{P(\boldsymbol{Y} \mid \boldsymbol{\theta})^{\tau^*} P(\tau^*)}{P(\boldsymbol{Y} \mid \boldsymbol{\theta})^{\tau^{(i)}} P(\tau^{(i)})}\right\}$$

$$= \min\left\{1, \frac{P(\boldsymbol{Y} \mid \boldsymbol{\theta})^{\tau^*} P_{prof}(\tau^*, \boldsymbol{\theta} \mid \boldsymbol{Y}) P_{prof}(\boldsymbol{Y})}{P(\boldsymbol{Y} \mid \boldsymbol{\theta} = \boldsymbol{\theta}_{max}(\tau^*))^{\tau^*} P(\boldsymbol{\theta} = \boldsymbol{\theta}_{max}(\tau^*))} \times \right.$$
$$\left. \frac{P(\boldsymbol{Y} \mid \boldsymbol{\theta} = \boldsymbol{\theta}_{max}(\tau^{(i)}))^{\tau^{(i)}} P(\boldsymbol{\theta} = \boldsymbol{\theta}_{max}(\tau^{(i)}))}{P(\boldsymbol{Y} \mid \boldsymbol{\theta})^{\tau^{(i)}} P_{prof}(\tau^{(i)}, \boldsymbol{\theta} \mid \boldsymbol{Y}) P_{prof}(\boldsymbol{Y})}\right\}$$

$$= \min\left\{1, \frac{P(\boldsymbol{Y} \mid \boldsymbol{\theta})^{\tau^*} P(\boldsymbol{Y} \mid \boldsymbol{\theta} = \boldsymbol{\theta}_{max}(\tau^{(i)}))^{\tau^{(i)}} P(\boldsymbol{\theta} = \boldsymbol{\theta}_{max}(\tau^{(i)}))}{P(\boldsymbol{Y} \mid \boldsymbol{\theta})^{\tau^{(i)}} P(\boldsymbol{Y} \mid \boldsymbol{\theta} = \boldsymbol{\theta}_{max}(\tau^*))^{\tau^*} P(\boldsymbol{\theta} = \boldsymbol{\theta}_{max}(\tau^*))}\right\}. \tag{2.13}$$

The acceptance ratio in equation (2.13) does not depend on the temperature dependent normalizing constant $z(\boldsymbol{Y} \mid \tau)$ thus eliminating the need for discrete temperatures, normalizing constant estimates, and tuning of bin widths. The formula for the prior of $\tau$ imposes the property that the maximal contour of the joint posterior distribution is continuous for $\tau \in [0, 1]$. The pseudo-code of STWNC algorithm is given in Algorithm 1.

## 2.4 Parallel Tempering via Simulated Tempering with Normalizing Constants algorithm

In Parallel Tempering via Simulated Tempering with Normalizing Constants (PT-STWNC) inverse temperature is a continuous parameter and $P(\boldsymbol{\theta} \mid \boldsymbol{Y}) = \int_0^1 P(\boldsymbol{\theta}, \tau \mid \boldsymbol{Y}) d\tau$ is a marginal distribution, which does not coincide with target distribution which is the conditional distribution $P(\boldsymbol{\theta} \mid \boldsymbol{Y}, \tau = 1)$. In standard ST the inverse temperature is a discrete variable and the samples of interest correspond to draws from $P(\boldsymbol{\theta} \mid \boldsymbol{Y}, \tau = 1)$. To obtain samples from $P(\boldsymbol{\theta} \mid \boldsymbol{Y}, \tau = 1)$ we run STWNC within PT.

PT-STWNC runs STWNC within PT. We define the two PT-STWNC chains as follows:

- The first PT chain updates $\boldsymbol{\theta}$ and $\tau$ via STWNC using $P(\boldsymbol{\theta}, \tau \mid \boldsymbol{Y})$ and then evaluates untempered log likelihood producing samples from the tempered distributions, while keeping track of calculations involved in marginal likelihood estimate.

- The second PT chain updates $\boldsymbol{\theta}$ via standard Metropolis-Hastings or Gibbs sampling from the target distribution $P(\boldsymbol{\theta} \mid \boldsymbol{Y}, \tau = 1)$. The untempered log likelihood is also evaluated using the samples from the target chain.

**Algorithm 1** Simulated Tempering Without Normalizing Constants (STWNC)

**Goal:** Update $\boldsymbol{\theta}$ and $\tau$ from $P(\boldsymbol{\theta}, \tau \mid \boldsymbol{Y})$, where $\tau \in [0, 1]$ is continuous.
Initialize the algorithm with $i = 0$ and some values for $(\boldsymbol{\theta}^{(i)}, \tau^{(i)})$; define $N$ - the number of iterations.

**for** $i = 1 : N$ **do**

**Transition 1:** update $(\boldsymbol{\theta} \mid \tau^{(i)})$;

propose a $(\boldsymbol{\theta}^*)$;

calculate the MH ratio $\alpha_{\boldsymbol{\theta}}$ and accept or reject $\boldsymbol{\theta}^*$ :

$$\alpha_{\boldsymbol{\theta}} = \frac{P(\boldsymbol{\theta}^*, \tau^{(i)} \mid \boldsymbol{Y})}{P(\boldsymbol{\theta}^{(i)}, \tau^{(i)} \mid \boldsymbol{Y})} = \frac{P(\boldsymbol{\theta}^* \mid \tau^{(i)}, \boldsymbol{Y})}{P(\boldsymbol{\theta}^{(i)} \mid \tau^{(i)}, \boldsymbol{Y})};$$

sample a $u_{\boldsymbol{\theta}} \sim U(0, 1)$;

**if** $u_\theta < \alpha_{\boldsymbol{\theta}}$ **then** set $(\boldsymbol{\theta}^{(i+1)}, \tau^{(i)}) \leftarrow (\boldsymbol{\theta}^*, \tau^{(i)})$,
**else** retain $(\boldsymbol{\theta}^{(i+1)}, \tau^{(i)}) \leftarrow (\boldsymbol{\theta}^{(i)}, \tau^{(i)})$;
**end if**

**Transition 2:** update $(\tau \mid \boldsymbol{\theta}^{(i+1)})$;

propose a $\tau^*$;

calculate the MH ratio $(\alpha_\tau)$ and accept or reject $(\boldsymbol{\theta}^{(i+1)}, \tau^*)$:

$$\begin{aligned} \alpha_\tau &= \frac{P(\boldsymbol{\theta}^{(i+1)}, \tau^* \mid \boldsymbol{Y})}{P(\boldsymbol{\theta}^{(i+1)}, \tau^{(i)} \mid \boldsymbol{Y})} \\ &= \frac{P(\boldsymbol{Y} \mid \boldsymbol{\theta})^{\tau^*} P(\boldsymbol{Y} \mid \boldsymbol{\theta} = \boldsymbol{\theta}_{max}(\tau^{(i)}))^{\tau^{(i)}} P(\boldsymbol{\theta} = \boldsymbol{\theta}_{max}(\tau^{(i)}))}{P(\boldsymbol{Y} \mid \boldsymbol{\theta})^{\tau^{(i)}} P(\boldsymbol{Y} \mid \boldsymbol{\theta} = \boldsymbol{\theta}_{max}(\tau^*))^{\tau^*} P(\boldsymbol{\theta} = \boldsymbol{\theta}_{max}(\tau^*))} \end{aligned}$$

sample a $u_\tau \sim U(0, 1)$;

**if** $u_\tau < \alpha_\tau$ **then** set $(\boldsymbol{\theta}^{(i+1)}, \tau^{(i+1)}) \leftarrow (\boldsymbol{\theta}^{(i+1)}, \tau^*)$,
**else** retain $(\boldsymbol{\theta}^{(i+1)}, \tau^{(i+1)}) \leftarrow (\boldsymbol{\theta}^{(i+1)}, \tau^{(i)})$;
**end if**

**end for**

**Return**: a single chain of samples $\{\boldsymbol{\theta}, \tau\}$.

As in standard PT, the PT-STWNC algorithm goes through two transitions: exchange and mutation. In the exchange step, the exchange between the two chains is proposed. If the exchange was accepted, then the two chains swap the parameter values between each other, and otherwise the two chains go through mutation step. In the mutation step, the first chain updates the parameters of interest at different temperatures via STWNC; the second chain updates the parameters of interest at $\tau = 1$ via standard Metropolis-Hastings or Gibbs sampling. The pseudo-code of the PT-STWNC is given in Algorithm 2.

**Algorithm 2** Parallel Tempering using Simulated Tempering Without Normalizing Constants (PT-STWNC)

Initialize two parallel chains: 'tempered' chain – initialize the algorithm with some values for $(\boldsymbol{\theta}_1^{(i)}, \tau_1^{(i)})$ for $i = 0$ and 'target' chain – initialize the algorithm with values for $(\boldsymbol{\theta}_2^{(i)}, \tau_2 = 1)$ for $i = 0$; $N$ - the number of iterations.

**for** $i = 1 : N$ **do**
  **exchange:**
  propose an exchange between the two chains;
  **if** the exchange is accepted **then**
    calculate the exchange probability:

$$\alpha_\phi = \frac{P(\boldsymbol{Y} \mid \boldsymbol{\theta}_2^{(i)})^{\tau_1^{(i)}} P(\boldsymbol{Y} \mid \boldsymbol{\theta}_1^{(i)})^{\tau_2=1}}{P(\boldsymbol{Y} \mid \boldsymbol{\theta}_1^{(i)})^{\tau_1^{(i)}} P(\boldsymbol{Y} \mid \boldsymbol{\theta}_2^{(i)})^{\tau_2=1}} \tag{2.14}$$

    sample a $u \sim U(0, 1)$;
    **if** $u < \alpha_\phi$ **then**
      swap states of the two chains:
      $(\boldsymbol{\theta}_1^{(i+1)}, \boldsymbol{\theta}_2^{(i+1)}) \leftarrow (\boldsymbol{\theta}_2^{(i)}, \boldsymbol{\theta}_1^{(i)})$ ;
    **else**
      retain $(\boldsymbol{\theta_1}^{(i+1)}, \boldsymbol{\theta}_2^{(i+1)}) \leftarrow (\boldsymbol{\theta}_1^{(i)}, \boldsymbol{\theta}_2^{(i)})$;
    **end if**
  **else**
    **mutation:**
    Update 'tempered' chain, i.e., update $(\boldsymbol{\theta}_1^{(i+1)}, \tau_1)$ via STWNC algorithm (the pseudo-code given in Algorithm 1);
    Update 'target' chain, i.e., update $(\boldsymbol{\theta}_2^{(i+1)}, \tau_2 = 1)$ via standard Metropolis-Hastings or Gibbs sampler;
  **end if**
**end for**
**Return**: samples from the 'tempered' chain $\{\boldsymbol{\theta}_1, \tau_1\}$, and samples from the 'target' chain $\{\boldsymbol{\theta}_2, \tau_2 = 1\}$.

## 2.5 Estimation of Marginal Likelihoods via thermodynamic integration

Posterior model probabilities $P(M \mid \boldsymbol{Y})$ provide an intuitive framework for evaluating model $M$ within a model class. Expanding the model class requires rescaling all posterior model probabilities. Consequently, comparing models $M_1$ and $M_2$ is typically performed through the posterior odds,

$$\frac{P(M_1 \mid \boldsymbol{Y})}{P(M_2 \mid \boldsymbol{Y})} = \frac{P(\boldsymbol{Y} \mid M_1)}{P(\boldsymbol{Y} \mid M_2)} \frac{P(M_1)}{P(M_2)}. \tag{2.15}$$

The ratio of posterior and prior odds,

$$B_{12} = \frac{P(\boldsymbol{Y} \mid M_1)}{P(\boldsymbol{Y} \mid M_2)} \tag{2.16}$$

is Bayes Factor of $M_1$ against $M_2$ (Kass and Raftery, 1995).

When there are no prior preferences for models, $B_{12}$ is equal to the posterior odds. The marginal likelihoods for models $M_j, j = 1, 2$, in (2.16), are obtained by integrating over the parameter space,

$$P(\boldsymbol{Y} \mid M_j) = \int_{\boldsymbol{\Theta}_j} P(\boldsymbol{Y} \mid \boldsymbol{\theta}_j, M_j) P(\boldsymbol{\theta}_j \mid M_j) d\boldsymbol{\theta}_j, \tag{2.17}$$

where $\boldsymbol{\theta}_j$ are the parameters corresponding to the $j$-th model. For expositional simplicity we will assume no prior preference for models throughout this chapter. Computing meaningful Bayes Factors requires accurate estimates of the marginal likelihood in (2.17).

The Posterior Harmonic Mean estimator (PHM), uses importance sampling to integrate (2.17) (Newton and Raftery, 1994; Raftery et al., 2006) resulting in unbiased marginal likelihood estimates but potentially infinite variance. Steppingstone sampling (SS) (Xie et al., 2011), which uses importance sampling to estimate each ratio of normalizing constants of the sequence of interpolating distributions between prior and target distribution, provides reliable marginal likelihood estimates. Alternatively, thermodynamic integration (TI) (Friel and Pettitt, 2008) builds on ideas from path sampling (Gelman and Meng, 1998) to estimate the marginal likelihood via,

$$log(P(\boldsymbol{Y} \mid M_j)) = \int_0^1 \mathbb{E}_{\boldsymbol{\theta}|\boldsymbol{Y},\tau,M_j}[log(P(\boldsymbol{Y} \mid \boldsymbol{\theta}, M_j))] d\tau, \tag{2.18}$$

where the expectation in the integrand is with respect to the tempered posterior distribution in (2.1) (Friel and Pettitt, 2008; Calderhead and Girolami, 2009). A numerical approximation to the thermodynamic integral in (2.18) is possible through discretization of $\tau$. In Friel

and Pettitt (2008), samples from the discretized tempered posteriors were used from parallel chains in PT. At each discretized value of $\tau$, $\mathbb{E}_{\boldsymbol{\theta}|\boldsymbol{Y},\tau,M_j}[log(P(\boldsymbol{Y} \mid \boldsymbol{\theta}, M_j))]$ is evaluated and the marginal likelihood is approximated by applying a trapezoid rule to numerically integrate over $\tau$,

$$\begin{aligned} log(P(\boldsymbol{Y} \mid M_j)) = & \int_0^1 \mathbb{E}_{\boldsymbol{\theta}|\boldsymbol{Y},\tau,M_j}[log(P(\boldsymbol{Y} \mid \boldsymbol{\theta}, M_j))]d\tau \\ \approx & \quad \tfrac{1}{2}\sum_{t=2}^{T} \Delta\tau_t(\mathbb{E}_{t,M_j} + \mathbb{E}_{t-1,M_j}), \end{aligned} \tag{2.19}$$

where $\mathbb{E}_{t,M_j} = \mathbb{E}_{\boldsymbol{\theta}|\boldsymbol{Y},\tau_t,M_j}[log(P(\boldsymbol{Y} \mid \boldsymbol{\theta}, M_j))]$ and $\Delta\tau_t = \tau_t - \tau_{t-1}$. The discretized trapezoidal rule in (2.19) was improved by Calderhead and Girolami (2009) by correcting for integration bias in terms of Kullberg-Leibler divergence, $KL(p_{t-1,M_j} \| p_{t,M_j})$, of $p_{t,M_j}$ from $p_{t-1,M_j}$ for model $M_j$,

$$\begin{aligned} log(P(\boldsymbol{Y} \mid M_j)) \approx & \underbrace{\frac{1}{2}\sum_{t=2}^{T} \Delta\tau_t(\mathbb{E}_{t,M_j} + \mathbb{E}_{t-1,M_j})}_{\text{Approximation}} + \\ & \underbrace{\frac{1}{2}\sum_{t=2}^{T}[KL(p_{t-1,M_j} \| p_{t,M_j}) - KL(p_{t,M_j} \| p_{t-1,M_j})]}_{\text{Bias}}, \end{aligned} \tag{2.20}$$

where $p_{t,M_j}$ is the tempered posterior distribution for the model $M_j$ given by the equation (2.1).

The thermodynamic integration via PT applies a numerical integration approximation to Monte Carlo approximations. The approximation error should decrease with number of chains and number of samples in each chain. As with any numerical integration, discretization over $\tau$ determines the accuracy of the result. Determining the optimal temperature schedule requires preliminary experimentation which contributes to unpopularity of the thermodynamic integration in practice. Based on the idea from path sampling, (Gelman and Meng, 1998), Calderhead and Girolami (2009) proposed that the temperature schedule could be chosen such that the Monte Carlo variance of the marginal likelihood estimates is minimized. The authors' numerical simulations suggest that the optimal temperature schedule is $\tau_i = (\frac{i}{T})^5$, which puts more emphasis on values close to the prior.

### 2.5.1 Computing the marginal likelihood via STWNC

Following Friel and Pettitt (2008), in standard ST, samples $\{(\boldsymbol{\theta}^1, \tau_1), .., (\boldsymbol{\theta}^n, \tau_n)\}$ drawn from $P(\boldsymbol{\theta}, \tau \mid \boldsymbol{Y}, M_j)$ can be used to estimate the marginal likelihood by first obtaining Monte Carlo approximation $\mathbb{E}_{\boldsymbol{\theta}|\boldsymbol{Y},\tau,M_j}[logP(\boldsymbol{Y} \mid \boldsymbol{\theta}, M_j)]$, and then solving the thermodynamic integral in (2.19) via quadrature. This is based on the assumption that the prior of

$\tau$ is proportional to the temperature-dependent normalizing constant, $P(\tau) \propto z(\boldsymbol{Y} \mid \tau, M_j)$. According to Friel and Pettitt (2008), in single chain methods such as ST, the normalizing constant $z(\boldsymbol{Y} \mid \tau, M_j)$ varies by orders of magnitude with $\tau$ which leads to poor estimation of the log untempered likelihood $logP(\boldsymbol{Y} \mid \boldsymbol{\theta}, M_j)$. Thus in standard ST, small values of $\tau$ do not tend to be sampled with high frequencies.

However, in STWNC, where $\tau$ is continuous, the marginal distribution of $\tau$ tends to spend a lot of time at near zero values (as can be seen from the marginal distribution of $\tau$ in Figure 2.3). The shape of the marginal distribution of continuous $\tau$ in STWNC coincides with that of the recommended geometric temperature schedule (i.e., $\tau_i = (\frac{i}{T})^5$, Figure 2.3, purple dash dotted line) in thermodynamic integration via PT established by Calderhead and Girolami (2009).

Samples $\{(\boldsymbol{\theta}^1, \tau_1), .., (\boldsymbol{\theta}^n, \tau_n)\}$ from PT-STWNC can be used to solve the marginal likelihood integral in (2.18), by first ordering the samples with respect to $\tau$, and solving the integral numerically,

$$\begin{aligned} \log P(\boldsymbol{Y} \mid M_j) &= \int_0^1 \mathbb{E}_{\boldsymbol{\theta}|\boldsymbol{Y},\tau,M_j}[log(P(\boldsymbol{Y} \mid \boldsymbol{\theta}, M_j))]d\tau \\ &\approx \sum_{t=2}^{T} \Delta\tau_t \, \mathbb{E}_{t,M_j}, \end{aligned} \tag{2.21}$$

where $\mathbb{E}_{t,M_j} = \mathbb{E}_{\boldsymbol{\theta}|\boldsymbol{Y},\tau_t,M_j}[log(P(\boldsymbol{Y} \mid \boldsymbol{\theta}, M_j))]$ and $\Delta\tau_t = \tau_t - \tau_{t-1}$.

The TI via PT relies on two layers of approximation to produce marginal likelihood estimates. The first layer of approximation in TI via PT corresponds to the MCMC integration (i.e., obtaining PT samples), and the second layer occurs when $\tau$ is numerically integrated out from the thermodynamic integrals in (2.19) and (2.20). Similarly, in the TI via PT-STWNC, the first approximation layer corresponds to obtaining samples from PT-STWNC, and the second layer of approximation is a result of solving the marginal likelihood integral in (2.21) numerically. However, since the $\tau$ is continuous, the TI via PT-STWNC removes the need for optimal temperature discretization schedule. Consequently, PT-STWNC provides marginal likelihood estimate with negligible additional computational cost.

## 2.6 Example 1: Two-dimensional multi-modal mixture Gaussian model

The two-dimensional mixture of 20 Gaussians is characterized by,

$$f(\boldsymbol{Y}) = \sum_{k=1}^{20} \frac{p_k}{\sqrt{2\pi}\sigma_k^2} \exp\left\{-\frac{1}{2\sigma_k^2}(\boldsymbol{Y}-\mu_k)'(\boldsymbol{Y}-\mu_k)\right\}, \tag{2.22}$$

with equal variances $\sigma^2 = 0.1$, equal mixing probabilities $\boldsymbol{p} = (p_1, ..., p_{20})'$ (i.e., $p_i = 1/20$), and the modes $\boldsymbol{\mu} = \{\mu_{i,j}\}, i = 1, .., 20, j = 1, 2$, as given by Kou et al. (2006); Liang and Wong (2001). The goal is to sample bivariate data points $\boldsymbol{Y}_1, \boldsymbol{Y}_2$ from the likelihood in (2.22) under the Uniform joint prior for $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$ defined on the rectangle $[-1, 10] \times [-1, 10]$.

### 2.6.1 Results

The PT-STWNC was run for 35,000 iterations with the first 1000 samples being removed as burn-in. Figure 2.1 shows contour plots of the bivariate twenty-modal likelihood

$P(\boldsymbol{Y}_1, \boldsymbol{Y}_2 \mid \boldsymbol{\mu}, \sigma^2, \boldsymbol{p})$ obtained from the 'tempered' chain (the plot A) and the 'target' chain (the plot B). Both of the PT-STWNC chains visited and sampled well from all the 20 components. The target sampling distribution is included for comparison in plot C. These plots demonstrate that the PT-STWNC draws samples from all the 20 modes reasonably well.

## 2.7 Example 2: Bimodal model

The likelihood is bimodal with respect to $\mu$, but is unimodal with respect to $\boldsymbol{Y}$,

$$P(\boldsymbol{Y} \mid \mu, \sigma^2) = N(|\mu|, \sigma^2), \tag{2.23}$$

and the posterior distribution of $\mathrm{P}(\mu \mid \boldsymbol{Y}, \sigma^2)$ is bimodal. The data were simulated from (2.23) with n=25, $\mu = 1.5$ and $\sigma^2 = 1$.

The sampling distribution of the PT-STWNC is the tempered joint posterior distribution,

$$P(\mu, \sigma^2, \tau \mid \mathbf{Y}) \propto P(\mathbf{Y} \mid \mu, \sigma^2)^\tau P(\tau) P(\mu) P(\sigma^2).$$

Conjugate priors on $\mu$ and $\sigma^2$ were assigned, $P(\mu) \sim \text{Normal}(0, 1)$; $P(\sigma^2) \sim \text{InverseGamma}(1, 1)$. The prior of $\tau$ was evaluated using the formula (2.12).

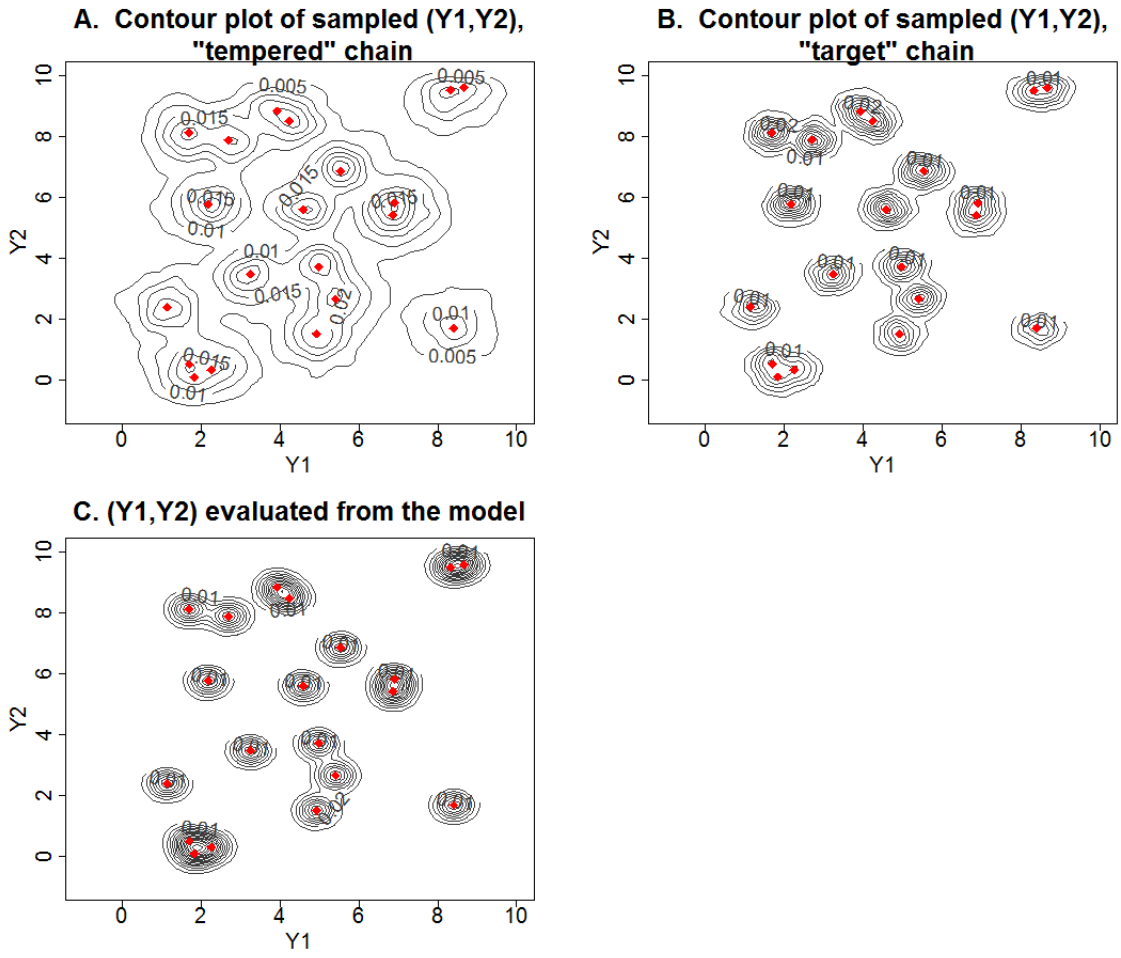


Figure 2.1: Two-dimensional mixture of 20 Gaussian distributions model – contour plots of sampled $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$ obtained from the PT-STWNC 'tempered' chain (plot A), and 'target' chain (plot B). The plot C is generated by evaluating the model in (2.22) over the grid points from the plot B. The red dots denote the locations of the modes $\boldsymbol{\mu}$.

### 2.7.1 Results

The PT-STWNC algorithm was run for 50,000 iterations with the first 15,000 samples discarded as burn-in. Figure 2.2 shows the sampled joint posterior distribution of $\mu$ and $\tau$ obtained from the PT-STWNC 'tempered' chain. The perspective and contour plots in Figure 2.2 illustrate that the two ridges of the posterior surface, which correspond to the maximum values of $\mu$, have approximately constant maximum height along the $\tau$ axis. The last observation is a direct consequence of the constraint that the profile posterior distribution of $\tau$ while profiling over $\mu$ is uniform on the interval [0,1].

The wide contours of Figure 2.2B at low values of $\tau$ demonstrate that the 'tempered' chain spends a lot of time sampling at low values of the inverse temperature, thus taking large steps to move between the two modes. Similarly, the marginal distribution of sampled $\tau$ demonstrates that low values of $\tau$ are sampled with higher frequencies (see Figure 2.3, gray color). The PT-STWNC 'target' chain updates the parameters of interest at $\tau = 1$, which results in drawing samples from the target posterior distribution (Figure 2.3, diagonal, blue color).

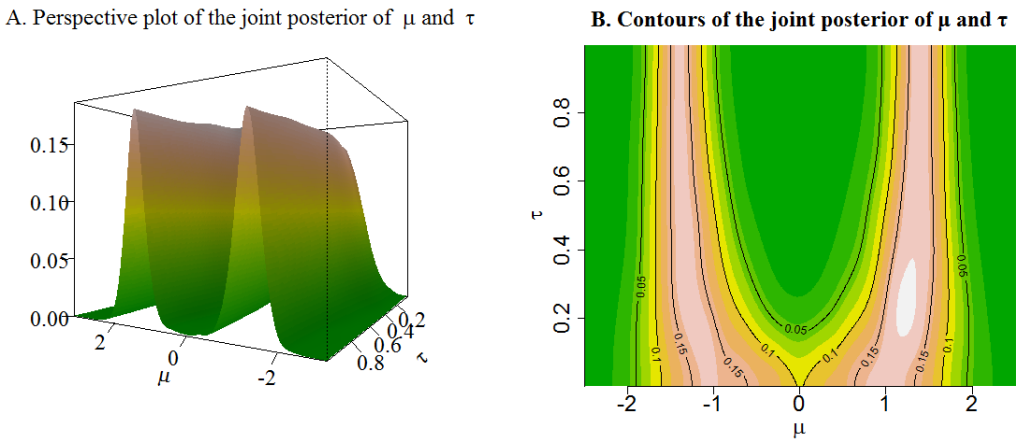


Figure 2.2: The bimodal model – sampled joint posterior distribution of $\mu$ and $\tau$ obtained from PT-STWNC 'tempered' chain: A perspective plot of the joint posterior distribution of $\mu$ and $\tau$, and B the corresponding contour plot.

PAWL within ST was also run on this toy example and the marginal distributions of $\mu$ and $\tau$ were compared to those from the PT-STWNC. The plot of the marginal distribution of $\mu$ is similar to the marginal distribution of $\mu$ obtained from the 'tempered' PT-STWNC chain (Figure 2.3, translucent green color). The marginal distribution of the discrete $\tau$ from the PAWL within ST has similar shape as that of the continuous $\tau$ from the PT-STWNC. Although both of the algorithms are powerful sampling and parameter estimation techniques, the PT-STWNC has the advantage of producing accurate marginal likelihood estimates with negligible added computational cost.

Table 2.1 illustrates that PT-STWNC produces accurate point estimates of the posterior means of $\mu$ and $\sigma^2$ compared to their theoretical values.

Implementation details are given in the Appendix B.

### 2.7.2 Marginal likelihood estimation

The PT-STWNC was used to estimate the marginal likelihood of the bimodal model introduced in the Section 2.7. The PT-STWNC marginal likelihood estimates were compared to the following three approaches: *i*). analytical solution to the marginal likelihood integral in (2.17), *ii*). the thermodynamic integration via PT with bias correction by Calderhead and

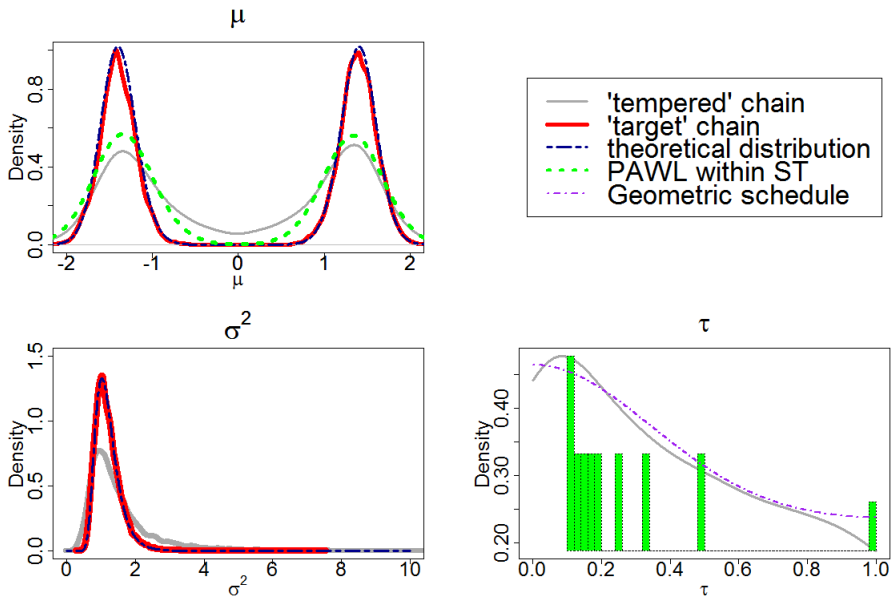


Figure 2.3: The bimodal model – marginal posterior distributions of $\mu, \tau$ and $\sigma^2$. Distributions in gray, red, blue and green color correspond to samples obtained from the PT-STWNC 'tempered' chain, the PT-STWNC 'target' chain, the theoretical distribution and PAWL within ST, respectively.

Table 2.1: Parameter estimates v.s. theoretical values

| | Parameter estimates | Theoretical results |
|---|---|---|
| $\mu$ | 1.398(0.0023); -1.404(0.0022) | 1.3995 ; -1.3995 |
| $\sigma^2$ | 1.21(0.017) | 1.29 |

The bimodal model – estimated posterior means (from the 'target' chain) and theoretical posterior means (from the target distribution) for each of the two modes of $\mu$ and for $\sigma^2$. Monte Carlo errors of the point estimates, obtained as per Craiu and Rosenthal (2014), are given in brackets.

Girolami (2009) (TI-PT-B) and *iii*). the thermodynamic integration via PT without bias correction by Friel and Pettitt (2008) (TI-PT-NB).

This is a toy example and one of the very few cases where a closed form of the marginal likelihood integral exists (see Appendix A). The PT-STWNC marginal likelihood estimates were obtained directly from the 'tempered' chain using the equation (2.21). TI via PT estimates were obtained by running PT with T=30 chains. The inverse temperature schedule was chosen to be a geometric schedule that tempers towards the prior i.e., $t_i = (\frac{i}{T})^5$.

The TI-PT-NB and the TI-PT-B estimates in the Table 2.2, were obtained using the samples from a single run of the PT. The only difference between the TI-PT-NB and the

TI-PT-B is in the bias term as per equation (2.20). The results in Table 2.2 show that the thermodynamic integral bias term (as per the equation (2.20)) in the TI-PT-NB and the TI-PT-B has a small effect on the marginal likelihood estimate.

Table 2.2 demonstrates that increased number of PT chains (T=60 and T=100) do not yield substantially better TI-PT-NB and TI-PT-B estimates. This result complies with the finding by Ahlers and Engel (2008), who demonstrated that while thermodynamic integration via PT performs well in unimodal case, the method exhibits substantiative bias in a bimodal case with a tractable marginal likelihood. Moreover, according to the Ahlers and Engel (2008) the estimates do not improve with increased number of parallel chains. The authors trace back this problem to the incomplete equilibrium between the two modes, which leads to failure to reproduce the exact mixture probabilities. In addition, the bias in PT-STWNC does not reduce when the integral in (2.21) is obtained by combining all the samples from the 20 runs.

Table 2.2: Marginal log-likelihood of the bimodal model

| | Analytic solution | PT-STWNC | TI-PT-B | TI-PT-NB |
|---|---|---|---|---|
| T =30 | -38.946 | -37.767(0.02) | -38.254(0.006) | -38.261(0.006) |
| T =60 | - | - | -38.247(0.004) | -38.254(0.005) |
| T =100 | - | - | -38.247(0.004) | -38.254(0.003) |

The bimodal model – marginal log-likelihood estimates from: analytic solution, PT-STWNC, thermodynamic integration via PT with bias correction (Calderhead and Girolami, 2009) (TI-PT-B) and thermodynamic integration via PT without bias correction (Friel and Pettitt, 2008) (TI-PT-NB). The PT-STWNC, the TI-PT-B and the TI-PT-NB estimates are based on 20 independent runs. Standard deviations of the marginal likelihood estimates obtained from the 20 runs are given in brackets. The TI-PT-B and the TI-PT-NB estimates with T=60 and T=100 chains are also obtained from 20 independent runs. Details on the convergence of the PT chains are given in the Appendix B.

## 2.8 Example 3: Galaxy data

The Galaxy data comprises velocities of 82 galaxies that diverge from our galaxy studied by Postman et al. (1986); Carlin and Chib (1995); Neal (1999). The data are univariate identically and independently distributed samples from mixture of $K$ Gaussian components and denoted as $\mathbf{Y} = (y_1, y_2, .., y_n)^{'}$, with n=82. Parameters of the model are given as $\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\sigma^2}, \boldsymbol{p})^{'}$ where $\boldsymbol{\mu} = (\mu_1, .., \mu_K)^{'}$ is a vector of mixture component means, $\boldsymbol{\sigma^2} = (\sigma_1^2, .., \sigma_K^2)^{'}$ is a vector of mixture component variances and $\boldsymbol{p} = (p_1, .., p_K)^{'}$ is a vector of mixture probabilities. The $k$-th mixture component has distribution $N(\mathbf{Y} \mid \mu_k, \sigma_k^2)$. Then likelihood function is

$$P(\mathbf{Y} \mid \boldsymbol{\mu}, \boldsymbol{\sigma^2}, \boldsymbol{p}) = \prod_{i=1}^{n} \sum_{k=1}^{K} p_k N(y_i \mid \mu_k, \sigma_k^2). \tag{2.24}$$

Conjugate priors were assigned, $P(\mu_k) \sim \text{Normal}(20, 100)$, $P(\sigma_k^2) \sim \text{InverseGamma}(shape = 3, scale = \frac{1}{20})$ and $P(p_1, .., p_K) \sim \text{Dirichlet}(\alpha_1 = 1, .., \alpha_K = 1)$.

A latent variable $\mathbf{Z}$ such that $P(Z_{ik} = 1 \mid p_k) = p_k$ was introduced to help derive the sampling distribution for the PT-STWNC algorithm. $\mathbf{Z}$ is a $n \times K$ matrix of indicator variables in which $Z_{i,k} = 1$ indicates the data point $i$ belongs to the mixture component $k$. Then distribution of each of the data points $\{y_i\}$ conditional on $Z_{ik}$ is,

$$P(y_i \mid Z_{ik} = 1, \boldsymbol{\mu}, \boldsymbol{\sigma^2}, \boldsymbol{p}) = N(y_i \mid \mu_k, \sigma_k^2),$$

and the joint distribution of $\{y_i\}$ and $Z_{ik}$ is,

$$P(y_i, Z_{ik} = 1 \mid \boldsymbol{\mu}, \boldsymbol{\sigma^2}, \boldsymbol{p}) = p_k N(y_i \mid \mu_k, \sigma_k^2).$$

Each row of $\mathbf{Z}$ is a multinomial variable with probabilities,

$$P(Z_{ik} = 1 \mid y_i, \boldsymbol{\mu}, \boldsymbol{\sigma^2}, \boldsymbol{p}) = \frac{p_k N(y_i \mid \mu_k, \sigma_k^2)}{\sum_{j=1}^{K} p_j N(y_i \mid \mu_j, \sigma_j^2).} \tag{2.25}$$

Sampling distribution of the PT-STWNC is the joint posterior distribution of the parameters of interest $\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\sigma^2}, \boldsymbol{p})'$, latent variable $\mathbf{Z}$ and the inverse temperature parameter $\tau$,

$$\begin{aligned} P(\boldsymbol{\mu}, \boldsymbol{\sigma^2}, \boldsymbol{p}, \boldsymbol{Z}, \tau \mid \mathbf{Y}) \quad \propto \quad & \left( \prod_{i=1}^{n} \sum_{k=1}^{K} P(y_i \mid Z_{ik} = 1, \boldsymbol{\mu}, \boldsymbol{\sigma^2}, \boldsymbol{p}) \right)^{\tau} P_{\boldsymbol{p}}(\boldsymbol{p}) \times \\ & \left( \prod_{k=1}^{K} P_{\mu_k}(\mu_k) \right) \left( \prod_{k=1}^{K} P_{\sigma_k^2}(\sigma_k^2) \right) \left( \prod_{i=1}^{n} \sum_{k=1}^{K} P(Z_{ik} \mid p_k) \right) P(\tau). \end{aligned}$$

### 2.8.1 Results

The PT-STWNC algorithm was run for 35,000 iterations with the first 1000 generated samples being removed as burn-in. Multi-modality in the Galaxy data with three components is illustrated by marginal distributions of $\mu_1, \mu_2, \sigma_1^2$ and $\tau$ and bivariate joint posterior distribution of $(\mu_1, \mu_2)$ (Figure 2.4). Sampling from such complex posterior structures requires proper mixing of $\tau$ which is governed by proper proposal selection and tuning. Plot of the marginal distribution of $\tau$ shows that $\tau$ spends a lot of time at values close to zero, which is crucial for the sampler to move easily between the isolated modes and produce accurate estimates (Figure 2.4).

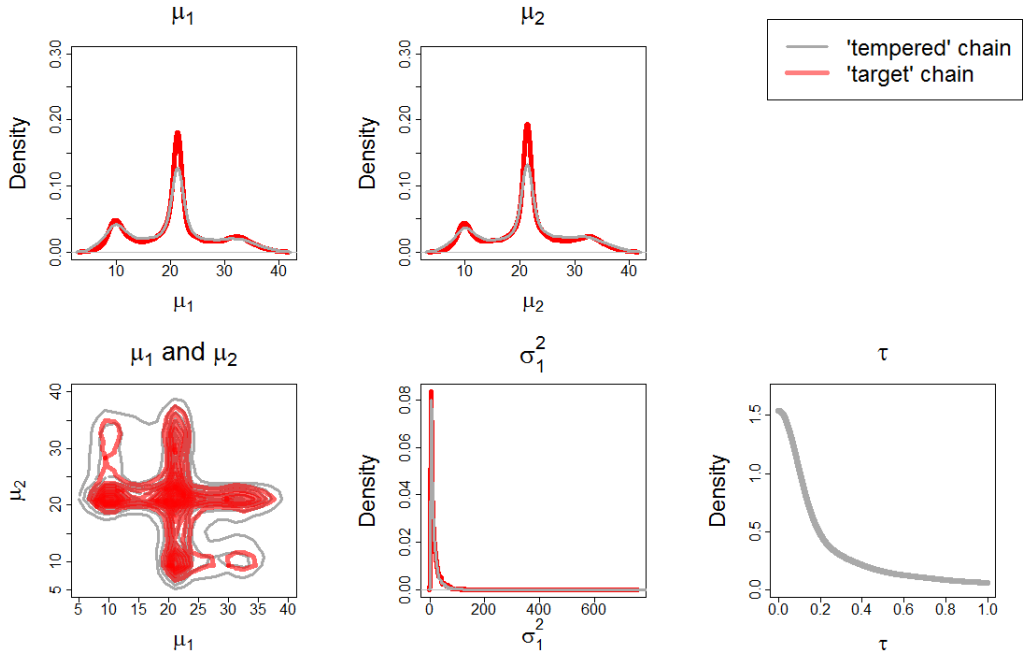


Figure 2.4: Galaxy data, model with three components unequal variances – marginal posterior distributions of the sampled parameters $\mu_1, \mu_2, \sigma_1^2, \tau$ and bivariate plot of $(\mu_1, \mu_2)$. Gray and red color correspond to samples obtained from the 'tempered' and 'target' chain, respectively. The transparent violet color corresponds to the geometric temperature schedule used to obtain marginal likelihood estimate from TI via PT.

### 2.8.2 Marginal likelihood estimation for the Galaxy data

The PT-STWNC was used to perform model selection on the Galaxy data set. The marginal likelihood was calculated for the Galaxy data to compare the following models: two components with equal variances, three components with unequal variances, three components with equal variances, four components with unequal variances and five components with unequal variances. A closed form expression of the marginal likelihood integral is not available.

The two thermodynamic integration approaches from the Section 2.5, were used to compare model selection abilities with PT-STWNC. Log marginal likelihood estimates and Bayes Factors demonstrate that all three estimation techniques agree that the model with five components is the best model (Tables 2.3 and 2.4). In addition, all the three estimation methods find support for the models with 3-5 components, while the worst model is the model with 2 components and equal variances. Findings from our model selection study comply with the results from previous studies. For instance, Steele and Raftery (2010)

argued that the number of components in the Galaxy data is not known, and concluded that the models with 3 and 6 components are reasonable fit to the data; Chib (1995) used Gibbs sampler to find that the model with 3 components is the best model; Liang and Wong (2001) used Evolutionary Monte Carlo (Liang and Wong, 2001) in combination with bridge sampling (Meng and Wong, 1996) to demonstrate that the model with 5 components is the best model and the model with 2 components equal variances is the worst model. In addition, Liang and Wong (2001) found support for the models with 3-5 components. Richardson and Green (1997) used Reversible Jump MCMC (RJMCMC) (Green, 1995) to conclude that the models with 5-7 components are good fit to the data.

The TI-PT-NB and the TI-PT-B estimates, in the Table 2.3, were obtained from a single run of the PT. Hence, the only difference between the TI-PT-NB and the TI-PT-B is in the bias term as per equation (2.20). The results in the Table 2.3 show that the estimates from the TI-PT-NB and the TI-PT-B are nearly the same, which suggests that the thermodynamic integral bias term has near zero effect in this example.

Table 2.3: Log marginal likelihood estimates of Galaxy data

| Model fitted | PT-STWNC | TI-PT-B | TI-PT-NB |
|---|---|---|---|
| 1. 2 components equal variances | -240.36(0.27) | -238.02(0.03) | -238.03(0.02) |
| 2. 3 components unequal variances | -226.96(0.48) | -224.26(0.03) | -224.28(0.03) |
| 3. 3 components equal variances | -232.68(0.54) | -224.20(0.06) | -224.23(0.05) |
| 4. 4 components unequal variances | -224.51(0.37) | -223.88(0.02) | -223.89(0.02) |
| 5. 5 components unequal variances | -222.91(0.23) | -223.85(0.02) | -223.85(0.02) |

Galaxy data – log marginal likelihood estimates obtained from the PT-STWNC, the TI-PT-B and the TI-PT-NB. Equations (2.21), (2.20) and (2.19) were used to obtain marginal likelihood estimates from the PT-STWNC, the TI-PT-B and the TI-PT-NB, respectively, for each of the five different models. The TI-PT-B and the TI-PT-NB estimates were obtained from T=30 PT chains using a geometric temperature schedule that tempers towards the prior $t_i = (\frac{i}{T})^5$. All the marginal likelihood estimates were obtained from 10 independent runs of each of the estimation techniques for each of the five models. Standard deviations of the marginal likelihood estimates from 10 runs are given in brackets.

## 2.9 Example 4: Susceptible-Infected-Recovered (SIR) epidemiological model

We illustrate the PT-STWNC on a Susceptible-Infected-Recovered (SIR) epidemiological model for number of daily deaths due to the black plague. The data were collected by the grave digger during the second black plague outbreak in the village of Eyam, UK, from June 19, 1666 to November 1, 1666 (Massad et al., 2004). The village had quarantined itself to avoid spreading the disease to the neighboring villages. Therefore, the population size is

Table 2.4: log Bayes factors

| | PT-STWNC | TI-PT-NB | TI-PT-B |
|---|---|---|---|
| $\log BF21$ | 13.4 | 13.74 | 13.75 |
| $\log BF31$ | 7.68 | 13.79 | 13.82 |
| $\log BF41$ | 15.85 | 14.14 | 14.13 |
| $\log BF51$ | 17.45 | 14.17 | 14.16 |
| $\log BF32$ | -5.72 | 0.05 | 0.06 |
| $\log BF42$ | 2.45 | 0.39 | 0.38 |
| $\log BF52$ | 4.05 | 0.43 | 0.41 |
| $\log BF43$ | 8.17 | 0.34 | 0.31 |
| $\log BF53$ | 9.77 | 0.37 | 0.34 |
| $\log BF54$ | 1.6 | 0.03 | 0.02 |

Bayes Factors obtained by applying the equation (2.16) to the log marginal likelihood estimates in the Table 2.3.

fixed to N=261, and the population is stratified into groups of susceptible S(t), infected I(t) and removed R(t) individuals, N=S(t)+I(t)+R(t). Since there is no recovery from the plague, the number of deaths correspond to the number of removed individuals up to time t, R(t) (Campbell and Lele, 2014; Golchi and Campbell, 2016)

The disease spread dynamics can be described by the following system of ordinary differential equations (ODE),

$$\frac{dS}{dt} = -\beta S(t)I(t),\ \frac{dI}{dt} = \beta S(t)I(t) - \alpha I(t),\ \frac{dR}{dt} = \alpha I(t) \tag{2.26}$$

where $\alpha$ describes the rate of death once the individual is infected and $\beta$ describes the plague transmission. Additional to the model parameters $\boldsymbol{\theta} = (\alpha, \beta)^{'}$, the ordinary differential equations model requires estimates of the initial states $(S(0), I(0), R(0))^{'}$. At the initial time the population consists of susceptible and infected individuals, and therefore $R(0) = 0$ and $S(0) = N - I(0)$. Consequently, the only initial state parameter is $I(0)$ so that the unknown parameters of the model are $\boldsymbol{\theta} = (\alpha, \beta, I(0))^{'}$. The data denoted as $\boldsymbol{Y} = (y_1, .., y_n)^{'}$ with $n = 136$, represent cumulative number of deaths up to times $t_1, .., t_n$. The data points $\boldsymbol{Y}$ were modeled by a binomial distribution with expected value equal to the solution to the ODE system in (2.26), $R_{(\alpha,\beta,I(0))}(t)$, where $t \in \{t_1, .., t_n\}$ .

The number of susceptible S(t) and infected I(t) are not observed. However, the number of infected at the end of the plague is 0, and the number of infected at time one before the end of the plague must therefore equal 1 (Campbell and Lele, 2014). Two additional data points on number of infected individuals $\boldsymbol{X} = (x_{n-1} = 1, x_n = 0)^{'}$ at times $(t_{n-1}, t_n)^{'}$ were modeled using binomial distribution with expected value equal to $I_{(\alpha,\beta,I(0))}(t)$ for

$t \in (t_{n-1}, t_n)'$,

$$P(\boldsymbol{Y} \mid \alpha, \beta, I(0)) = \prod_{i=1}^{n} \text{Binomial}\Big(y_i \mid N, \frac{R_{(\alpha,\beta,I(0))}(t_i)}{N}\Big) \times \prod_{i=n-1}^{n} \text{Binomial}\Big(x_i \mid N, \frac{I_{(\alpha,\beta,I(0))}(t_i)}{N}\Big) \tag{2.27}$$

Prior distributions for $\boldsymbol{\theta} = (\alpha, \beta, I(0))'$ were chosen to be: $\alpha, \beta \sim \text{Gamma}(1, 1)$, $I(0) \sim \text{Binomial}(N, \frac{5}{N})$.

Parameters $\alpha$ and $\beta$ are continuous and $I(0)$ is discrete. The discrete nature of the $I(0)$ induces multi-modality in the likelihood surface. This mixture of discrete and continuous parameters in the model imposes difficulties in sampling from the posterior distribution. Standard MCMC could get easily trapped in local modes of the posterior of the parameters of interest.

### 2.9.1 Results

The PT-STWNC was run on the SIR model for 35,000 iterations with 1000 burn-in samples. Multi-modality and topological challenges of the model are illustrated by the marginal distribution plots (Figure 2.5, diagonal) and by the bivariate joint posterior distributions plots (Figure 2.5, off-diagonal) of $\boldsymbol{\theta} = (\alpha, \beta, I(0))'$. The marginal distributions of $\alpha$ and $\beta$ exhibit structures with three isolated modes. In Figure 2.5 (off-diagonal), clouds in the joint posterior distribution of $\alpha$ and $\beta$ represent the modes which correspond to the discrete samples of $I(0) = \{6, 5, 4, 3\}$ from left to right.

Histograms of the marginal distributions of $\alpha, \beta, I(0)$ and $\tau$ obtained from the 'tempered' chain (Figure 2.6), illustrate the complexity and topological challenges of the model as well as the need for exploring the diffuse prior parameter space in order for PT-STWNC to draw samples from the target distribution. Figure 2.7 demonstrates that the prior parameter space (the grey contour lines) is much more diffused that that of the joint posterior distribution of $\alpha$ and $\beta$ (the red dots, which when zoomed-in assumes the shape of the target posterior distribution). Consequently, the algorithm spends much time sampling at near-zero values of $\tau$ thus exploring the prior parameter space.

For implementation details we refer the reader to Appendix C

## 2.10 Discussion

In this chapter we presented a hybrid Parallel Tempering Without Normalizing Constants (PT-STWNC) algorithm that tackles the problem of sampling from multi-modal distributions with isolated modes. The proposed PT-STWNC combines the sampling efficiency from the standard ST with the simplicity of its usage. Namely, the PT-STWNC removes the

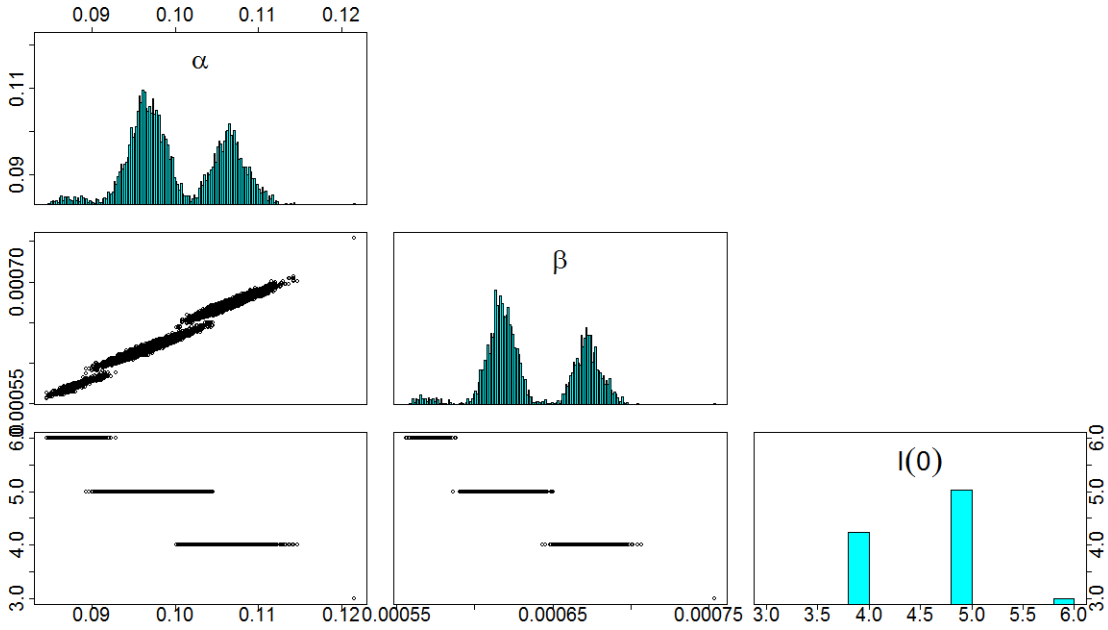

Figure 2.5: SIR model – marginal (diagonal) and bivariate joint (off-diagonal) posterior distributions of sampled parameters $\alpha, \beta$ and $I(0)$ obtained from the 'target' chain.

standard Simulated Tempering (ST) requirement for calculating normalizing constants and choosing a suitable temperature schedule. This is achieved by replacing the discrete temperature parameter in the standard ST with a continuous one, thus enabling the PT-STWNC to choose the prior of the inverse temperature parameter automatically. Moreover, employing continuous temperature enables PT-STWNC to produce marginal likelihood estimates and Bayes Factors thereof, at no additional computational cost. Our examples demonstrate that the PT-STWNC exhibits fast mixing and simultaneously produces accurate estimates of the parameters of interest and marginal likelihood.

In order for the PT-STWNC to produce reliable estimates, the inverse temperature has to mix well. Mixing of the inverse temperature parameter is directly affected by: *i*). optimizing the parameters of interest needed for calculation of its prior, $\mathrm{P}(\tau)$, and *ii*). choosing its proposal distribution.

Optimization of the posterior distribution with respect to the parameters, which is crucial for the selection of the prior for $\tau$, is computationally expensive in complex models. However, the optimization time decays to 0 with the number of MCMC iterations if the optimization at iteration $i$ of $P(\boldsymbol{\theta}^{(i)} \mid \tau^{(i)}, \boldsymbol{Y})$ is initialized with $\boldsymbol{\theta}^{(j)}$, where $(\boldsymbol{\theta}^{(j)}, \tau^{(j)})$ are chosen so that iteration $j = \arg\min_j |\tau^{(j)} - \tau^{(i)}|$.

When it comes to choosing the proposal distribution of the temperature parameter, proposing from independent standard uniform distribution works best when applied on

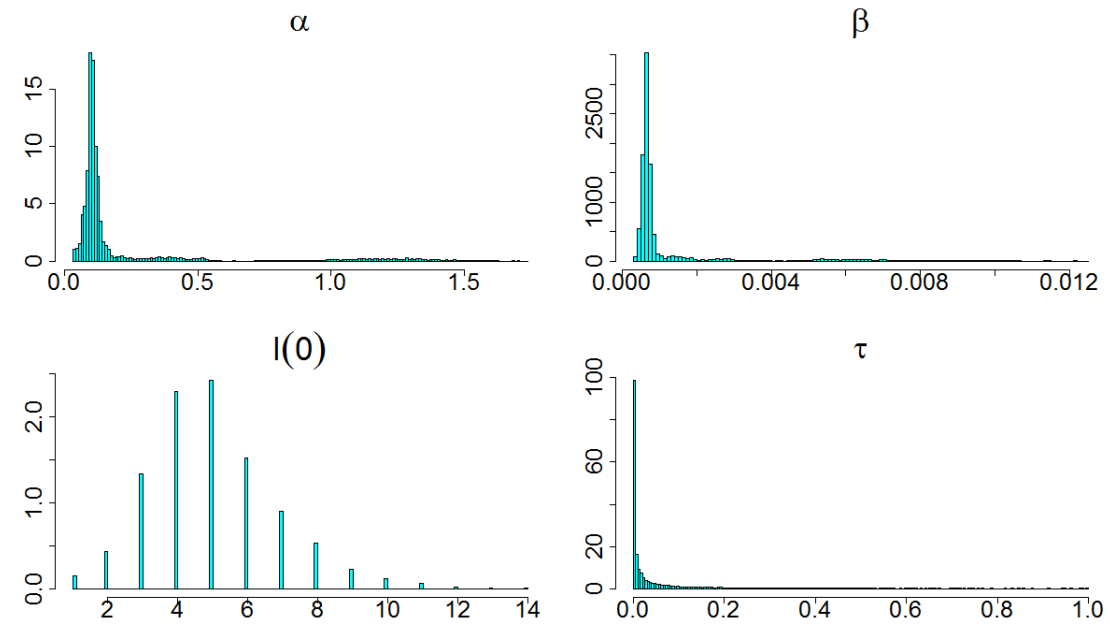


Figure 2.6: The SIR model – marginal posterior distributions of sampled parameters $\alpha, \beta, I(0)$ and $\tau$ obtained from the 'tempered' chain.

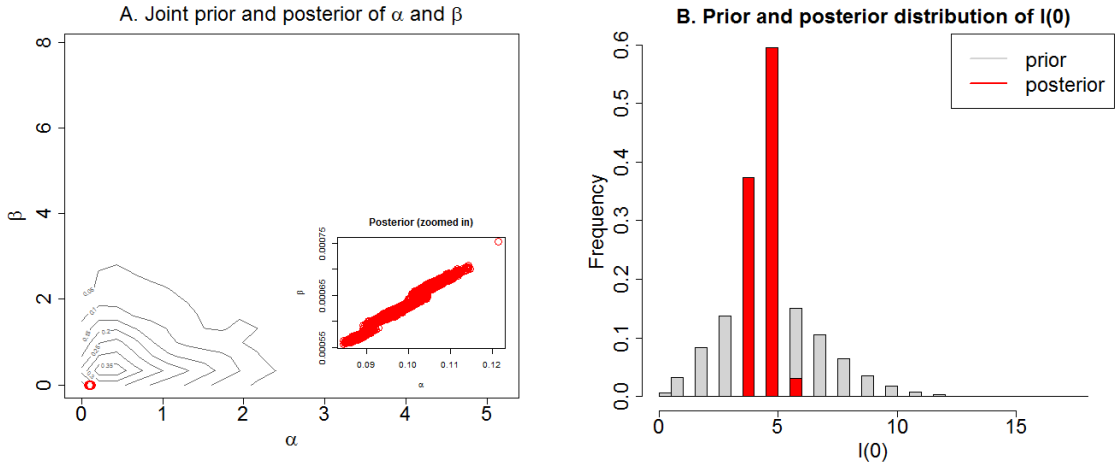


Figure 2.7: The SIR model – parameter space of the prior (gray color) versus the posterior space (red color) of the parameters in SIR model. The plot A shows contours of the joint prior distribution of the parameters $\alpha$ and $\beta$. The small red dots close to the origin correspond to the joint posterior distribution. Parameter space of the prior and posterior distribution of the parameter $I(0)$ are shown in the plot B.

simple examples as we have shown in the bimodal model and in the bivariate mixture of twenty Gaussian distributions. The mixture model of Gaussian distributions for the Galaxy data and the ODE model for the epidemiological model exhibit complex structures which

do not allow the temperature parameter to move freely in the temperature space. Carefully chosen proposal distribution coupled with tuning of the transition step solves the problem of bad mixing of $\tau$. $\tau$ needs to spend a lot of time at values close to zero in order for the PT-STWNC to capture the posterior structure especially when the likelihood is much more concentrated than the prior.

Bias in the the TI-PT-B and the TI-PT-NB estimates can be explained as follows. First, in order to obtain the TI-PT-B and the TI-PT-NB estimates, $\tau$ needs to be integrated out by numerical integration over the MCMC approximations of the conditional posterior means $\mathbb{E}_{\boldsymbol{\theta}|\boldsymbol{Y},\tau,M}[log(P(\boldsymbol{Y} \mid \boldsymbol{\theta}, M)]$ (equations (2.19) and (2.20)), which is equivalent to applying approximations recursively. Another possible contributing factor to the bias in the TI-PT-B and the TI-PT-NB estimates is the choice of the optimal temperature schedule. Namely, we did not test different temperature schedules to find the one that minimizes the bias and the Monte Carlo variance as suggested by Calderhead and Girolami (2009). Instead, we used the geometric temperature schedule which had been found by Calderhead and Girolami (2009) to be optimal for the linear regression model, but could be sub-optimal for our specific problem.

In complex problems, the PT-STWNC can be expanded to include more than two chains, by adding more STWNC chains. This strategy could aid in better mixing of inverse temperature from each of the STWNC chains.

# Chapter 3

# Incremental Mixture Importance Sampling with Shotgun optimization

## 3.1 Introduction

Sampling from a posterior density is challenging when the posterior modes are separated with deep valleys of low probability or when the posterior space is rife with many minor modes, ripples and ridges. Theoretically, standard Metropolis-Hastings or Gibbs algorithms converge to the target density if run infinitely long. Tempering methods such as Simulated Tempering (Marinari and Parisi, 1992; Geyer and Thompson, 1995; Zhang and Ma, 2008) and Parallel Tempering (Swendsen and Wang, 1986; Geyer, 1991; Hukushima and Nemoto, 1996), are random-walk variants designed to efficiently deal with sampling from multi-modal distributions.

However, Parallel Tempering could exacerbate topological challenges of the posterior if the prior is inconsistent with the likelihood, trapping the sampler in a local mode Campbell and Steele (2012). One could solve this problem using local optimal transition densities, which adapt to the local geometries of the posterior space (Atchadé et al., 2005).

Importance sampling algorithms such as Sampling Importance Re-sampling (SIR) (Rubin et al., 1988; Rubin, 1987; Poole and Raftery, 2000; Alkema et al., 2011) or Sequential Monte Carlo variants (SMC) (Del Moral et al., 2006) take advantage of computing the sampling weights in parallel. The difficulty with importance sampling methods is choosing the initial importance sampling density to cover the important modes of the target density. The prior is often chosen to be this initial importance density.

A frequentist alternative to MCMC methods would be to use optimization in order to find the modes, but in the presence of well isolated multiple modes, different starting points

for the optimizer result in multiple optima. Then the problem shifts to finding a way to combine these local optima.

Incremental Mixture Importance Sampling with Optimization (IMIS-Opt) (Raftery and Bao, 2010) is designed to discover all the important posterior modes by using the prior as a starting point for optimization, and then building a posterior through incrementally added optimized local posterior approximations. However, if the prior disagrees with the likelihood, i.e., if the prior covers the basin of attraction of local but not global likelihood modes, then the IMIS-Opt will miss the important modes. As a remedy, one can choose a diffuse prior, but this implies that the prior should be chosen for algorithmic convenience rather than to represent expert opinion.

In this chapter, we modify the IMIS-Opt algorithm by replacing the optimization step with a general optimization strategy, which is based on the idea that no single method outperforms other methods in every problem (Wolpert and Macready, 1997). The proposed multiple-method optimization strategy balances discovery of the global and the local modes by combining results from different regions of the posterior space, corresponding to local optima found by multiple parameter estimation methods. We refer to this strategy as Shotgun optimization (ShOpt), and the resulting algorithm as Incremental Mixture Importance Sampling with Shotgun optimization (IMIS-ShOpt). The IMIS-ShOpt relies on the Shotgun optimization, rather than on the prior choice. IMIS-ShOpt does not choose the prior for optimization convenience, but reaffirms its role of conveying expert opinion.

Shotgun optimization is a general methodology which is directly applicable to any model type including parameter estimation in differential equation models. The ordinary differential equation (ODE) models are particularly challenging because these models exhibit posterior topologies featuring multiple modes, ridges and ripples. Any of the existing methods for parameter estimation in ODEs might get trapped in a local mode for reasons specific to the method used. In this chapter we demonstrate that the IMIS-ShOpt produces accurate parameter estimates in ODEs by combining results from different methods. Furthermore, we showcase that the IMIS-ShOpt can be combined with the synthetic likelihood (Wood, 2010) to draw inference in models where the likelihood is intractable or costly to evaluate.

The rest of the chapter is organized as follows. Section 3.2 clarifies the need for multiple optimization techniques and discusses the differences between our Shotgun optimization strategy and the multi-objective optimization. Section 3.3 gives detailed overview of the IMIS-Opt algorithm, followed by a demonstration of the IMIS-Opt getting trapped in an unimportant mode in a simple ODE model. In Section 3.4 the proposed IMIS-ShOpt algorithm is presented. Sections 3.6 and 3.7 illustrate the performance of the IMIS-ShOpt algorithm through two examples involving ODE models. The IMIS-ShOpt via synthetic likelihood is proposed in the Section 3.8, and its parameter estimation performance is illustrated using a chaotic stochastic difference equation model. Section 3.10 follows with concluding remarks.

## 3.2 Shotgun optimization

For inference in complex problems there are generally multiple competing methods, none of which works uniformly best in every problem (Wolpert and Macready, 1997). Different competitive parameter estimation methods rely on different models (such as method of moments versus maximum likelihood estimators), or different optimization methods (such as gradient, simplex or simulated annealing). In practice, one has to decide between modifying the model specification or choosing an optimization strategy where each is tuned to the specific problem. Modifying the model leads to a variant of the desired answer, while choosing an optimization strategy requires validation if the answers are to be trusted. For example, for inference from ODE models, strict likelihood function optimization i.e., nonlinear least squares (NLS) based on the ODE solution (Bates and Watts, 1988; Seber, 1989), discover a local optima, whereas optimization of the profile likelihood using model based data smoothing instead of the ODE solution (Ramsay et al., 2007) will search widely for a global mode but results in higher variance estimates (Wu et al., 2014). Additionally, if there are multiple important modes the profile likelihood will not find them from different initializations, but NLS will find different modes with different initializations. Hence, different optimization strategies lead to different results. Then the Shotgun optimization strategy would be constructed as a combination of these two optimization methods in order to discover local and global optima (Berger et al., 1999; Walley and Moral, 1999).

Using Shotgun optimization introduces robustness to the shortcomings of a single method. Combining results from different optimization or parameter estimation methods ensures that posterior space has been more fully explored. The Shotgun optimization is analogous to the ensemble methods (Madigan and Raftery, 1994; Hoeting et al., 1999; Friedman et al., 2001; Mendes-Moreira et al., 2012; Montgomery et al., 2012) where relative importance of the predictions are determined using a combination of models. Ensemble methods rely on the notion that no particular model can fully capture the data features. Hence, some models better predict certain features of the data, while producing biased predictions in some areas. The ensemble methods overcome the induced bias by combing the models together. In the Shotgun optimization, certain methods provide better estimates of the parameters than others, and combining the results from different methods overcomes the problem of the introduced bias.

The way the Shotgun optimization combines results from different competing methods is substantially different from the multi-objective optimization. While the multi-objective strategy is a general optimization framework designed to optimize simultaneously several objectives, the proposed Shotgun optimization strategy is a single objective optimization that combines results from multiple optimization criteria. With multi-objective optimization strategy there is no single solution that is optimal with respect to all objectives. Instead of having one optimal solution, there is a set of alternative trade-offs, known as Pareto op-

timal solution set (Kuhn and Tucker, 1951; Miettinen, 2012), which balances among several objectives. Many multiple-objective methods have been proposed in the literature, such as weighted sum method (Cohon and Marks, 1973), goal programming (Levary and Avery, 1984), the minimax approach (Tseng and Lu, 1990)), evolutionary algorithms (EA) (Bäck et al., 1997; Zitzler, 1999; Deb, 2001; Cheng and Gen, 1997; Fonseca and Fleming, 1995; Valenzuela-Rendón et al., 1997) and Simulated Annealing for multi-objective optimization (Smith et al., 2004; Serafini, 1994; Alrefaei and Diabat, 2009).

## 3.3 Incremental Mixture Importance Sampling with Optimization

The main objective of Incremental Mixture Importance Sampling with Optimization (IMIS-Opt) (Raftery and Bao, 2010) is to iteratively construct an importance sampling distribution. The initial stage of the IMIS-Opt starts by drawing $N_0$ samples $\boldsymbol{\Theta}_0 = \{\boldsymbol{\theta}_1, .., \boldsymbol{\theta}_{N_0}\}$ from the prior and then calculating their weights based on the likelihood function. In the optimization stage, the D highest-weight points are selected to sequentially initialize the optimizer, which searches for the nearest mode in the target posterior space. Then B points, drawn from the multivariate Gaussian distribution centered at the modes found by the optimizer, are added to the current importance distribution. At each iteration of the importance stage, sampling weights are calculated, and B draws from the multivariate Gaussian distribution centered at the highest-weight point are added to the current importance sampling distribution. The weighting and sampling steps of the importance stage are iterated until the importance weights are reasonably uniform. After the stopping criterion is met, J inputs are re-sampled with replacement from $\{\boldsymbol{\theta}_1, .., \boldsymbol{\theta}_{N_K}\}$ with weights $(w_1, .., w_{N_K})'$ where $K$ is the total number of particles from the importance sampling distribution. The pseudo-code of the IMIS-Opt is given in Algorithm 3.

If optimization and importance sampling stages are excluded, then the algorithm becomes a Sampling Importance Re-sampling (SIR) algorithm (Rubin, 1987; Rubin et al., 1988; Poole and Raftery, 2000; Alkema et al., 2007). By excluding the optimization step, the algorithm becomes IMIS (Hesterberg, 1995; Steele et al., 2006; Raftery and Bao, 2010). The IMIS-Opt initializes the optimizer using the D highest-weight points which makes it a powerful method for exploring the posterior space. However, the successful mixing of the IMIS-Opt depends heavily on the consistency of the information in the prior and likelihood, and consequently, on whether or not samples from the prior cover all the important posterior modes. The implication is that the prior should be chosen for the optimization convenience rather than using the expert knowledge. For example, consider ordinary differential equation models, where the posterior topologies usually exhibit multi-modality, ripples and many unimportant local modes. If the prior covers the basin of attraction of unimportant mode, but not global mode, the IMIS-Opt gets easily trapped in the unim-

portant local mode covered by the prior, thus failing to explore the full posterior space. Consequently, using the estimates obtained from the IMIS-Opt will result in partial fit to the data.

## 3.4 Incremental Mixture Importance Sampling with Shotgun optimization

The IMIS-Opt success depends heavily on the consistency between the prior and the data. If the prior is inconsistent with the likelihood then the maximum height point needed to initialize the optimizer is in the basin of attraction of the local mode that is covered by the prior. Therefore, the sampler is prevented from fully exploring the posterior space. The Incremental Mixture Importance Sampling with Shotgun optimization (IMIS-ShOpt) builds on IMIS-Opt, by altering the optimization stage to incorporate the Shotgun optimization strategy, which consists of Q different competitive parameter estimation methods or optimization strategies. This implies using a variety of optimization methods in parallel or using a fixed optimizer on a variant of the function to optimize, such as likelihood or other objective function within the estimating framework. The Shotgun optimization strategy sequentially initializes Q different optimization methods (which could be run in parallel) for each of the D maximum weight points from the prior. Pseudo-code of the proposed Shotgun optimization strategy is given in Algorithm 4. Replacing the optimization step in the Algorithm 3 with the Shotgun optimization in the Algorithm 4 gives the pseudo code of the IMIS-ShOpt algorithm.

The IMIS-ShOpt explores true and alternative truth modes and merges the samples from different regions of the target posterior, $P(\boldsymbol{\theta} \mid \boldsymbol{Y})$, explored by the variety of criteria. Although the IMIS-ShOpt draws samples from the target posterior distribution $P(\boldsymbol{\theta} \mid \boldsymbol{Y})$, the optimization step uses different strategies of modifying the target posterior to improve the exploration of the parameter space. The modification of the posterior depends on the parameter estimation method used. For example, if a parameter of interest is a location parameter, the Multiple- method optimization in IMIS-ShOpt could be comprised of Q=2 methods: the Maximum Likelihood method and the Method of Moments. Therefore, the posterior modifications targeted by different optimization methods may give different results due to the differences in topology of the posterior space.

Sampling weights in the IMIS-ShOpt are obtained with respect to the target posterior distribution, ensuring that the unlikely points are not re-sampled in the final stage. Hence, keeping the unlikely points in the importance sampling distribution does not harm the algorithm, but it does improve the posterior exploration.

**Algorithm 3** IMIS-Opt

**Goal: Draw samples from the target distribution $P(\boldsymbol{\theta} \mid \boldsymbol{Y})$.**

**Input:** Data, model, likelihood function, prior distribution, $B$ - the number of incremental points, $D$ - the number of different initial points for the optimization, $N_0$ - the number of the initial samples from the prior and $J$ - the number of re-sampled points, $N$ - the number of iterations.

**Initial stage:** Draw $N_0$ samples $\boldsymbol{\Theta}_0 = \{\boldsymbol{\theta_1}, \boldsymbol{\theta_2}, ..., \boldsymbol{\theta_{N_0}}\}$ from the prior distribution $P(\boldsymbol{\theta})$.

**for** $k = 1 : N$ **do**

**if** k=1 **then**

For each $\{\boldsymbol{\theta_i}, i = 1, .., N_0\}$ calculate the sampling weights:

$$w_i^{(1)} = \frac{P\left(\boldsymbol{Y} \mid \boldsymbol{\theta_i}\right)}{\sum_{j=1}^{N_0} P(\boldsymbol{Y} \mid \boldsymbol{\theta_j})} \tag{3.1}$$

**Optimization stage:**

**for** $d = 1 : D$ **do**

Use $\boldsymbol{\theta}^{initial} = \underset{\boldsymbol{\theta}}{\text{argmax}}\ \boldsymbol{w}^{(1)}(\boldsymbol{\theta})$, $\boldsymbol{\theta} \in \boldsymbol{\Theta}_{d-1}$ to initialize the optimizer and get local posterior maxima $\boldsymbol{\theta}_d^{(Opt)} = \underset{\boldsymbol{\theta}}{\text{argmax}}\ P(\boldsymbol{\theta} \mid \boldsymbol{Y})$ along with the corresponding inverse negative Hessian $\boldsymbol{\Sigma}_d^{(Opt)}$.

Update $\boldsymbol{\Theta}_d$ by excluding $\frac{N_0}{D}$ nearest neighbor points, $\boldsymbol{\theta}_k \in \boldsymbol{\Theta}_{d-1}$, that minimize the Mahalanobis distance,

$$(\boldsymbol{\theta}_k - \boldsymbol{\theta}_d^{(Opt)})'(\boldsymbol{\Sigma}_d^{(Opt)})^{-1}(\boldsymbol{\theta}_k - \boldsymbol{\theta}_d^{(Opt)}). \tag{3.2}$$

Draw B samples $\boldsymbol{\theta}_{1:B} \sim MVN(\boldsymbol{\theta}_d^{(Opt)}, \boldsymbol{\Sigma}_d^{(Opt)})$; add these samples to the importance sampling distribution and evaluate $H_k = MVN(\boldsymbol{\theta}_{1:B} \mid \boldsymbol{\theta}_d^{(Opt)}, \boldsymbol{\Sigma}_d^{(Opt)})$.

**end for**

**else**

**Importance sampling stage:**

For each $\{\boldsymbol{\theta_i}, i = 1, .., N_k\}$ calculate weights,

$$w_i^{(k)} = \frac{cP(\boldsymbol{Y} \mid \boldsymbol{\theta_i})P(\boldsymbol{\theta_i})}{\frac{N_0}{N_k}P(\boldsymbol{\theta_i}) + \frac{B}{N_k}\sum_{s=1}^{k} H_s(\boldsymbol{\theta_i})}, \tag{3.3}$$

where $N_k = N_0 + B(D + k)$ and $c = 1/\sum_{i=1}^{N_k} w_i^{(k)}$ is the normalizing constant.

Choose the maximum weight input $\boldsymbol{\theta}_k$ and estimate $\boldsymbol{\Sigma}_k$ as the weighted covariance of B inputs with smallest Mahalanobis distance,

$$w_p(\boldsymbol{\theta})\left(\boldsymbol{\theta} - \boldsymbol{\theta}_k\right)'\left(\boldsymbol{\Sigma_\pi}\right)^{-1}\left(\boldsymbol{\theta} - \boldsymbol{\theta}_k\right),$$

where the weights $w_p(\boldsymbol{\theta})$ are proportional to the average of the importance weights and the uniform weights $\frac{1}{N_k}$, $\boldsymbol{\Sigma}_\pi$ is the covariance of the initial importance distribution.

**Algorithm 3** IMIS-Opt - continued

Draw B samples $\boldsymbol{\theta}_{1:B} \sim MVN(\boldsymbol{\theta}_k, \boldsymbol{\Sigma}_k)$; add these points to the importance sampling distribution and evaluate $H_k = MVN(\boldsymbol{\theta}_{1:B} \mid \boldsymbol{\theta}_k, \boldsymbol{\Sigma}_k)$.

**end if**

**if** $\sum_{1}^{N_k}(1-(1-w^{(k)})^J) \geq J(1-\exp(-1))$ i.e., importance sampling weights are approximately uniform **then** exit for loop

**end if**

**end for**

**Re-sampling stage:**

Re-sample J points with replacement from $\{\boldsymbol{\theta}_1, .., \boldsymbol{\theta}_{N_k}\}$ and weights $(w_1, .., w_{N_k})'$.

**Algorithm 4** The Shotgun optimization

**Optimization stage:**

**for** $d = 1 : D$ **do**

Find the d-th maximum weight point $\boldsymbol{\theta}_d^{(initial)} = \underset{\boldsymbol{\theta}}{\text{argmax}}\ \boldsymbol{w}^{(k)}(\boldsymbol{\theta})$, $\boldsymbol{\theta} \in \boldsymbol{\Theta}_{d-1}$ to initialize $Q$ optimizers.

**for** $q = 1 : Q$ **do**

Use q-th optimization method initialized at $\boldsymbol{\theta}_d^{initial}$ to obtain local maxima $\boldsymbol{\theta}_{d,q}^{(Opt)}$ along with the corresponding inverse negative Hessian $\boldsymbol{\Sigma}_{d,q}^{(Opt)}$ (this step can be parallelized).

Update $\boldsymbol{\Theta}_d$ by excluding $\frac{N_0}{QD}$ nearest neighbor points, $\boldsymbol{\theta}_k \in \boldsymbol{\Theta}_{d-1}$, that minimize the Mahalanobis distance,

$$(\boldsymbol{\theta}_k - \boldsymbol{\theta}_{d,q}^{(Opt)})'(\boldsymbol{\Sigma}_{d,q}^{(Opt)})^{-1}(\boldsymbol{\theta}_k - \boldsymbol{\theta}_{d,q}^{(Opt)}). \tag{3.4}$$

Draw B samples $\boldsymbol{\theta}_{1:B} \sim MVN(\boldsymbol{\theta}_{d,q}^{(Opt)}, \boldsymbol{\Sigma}_{d,q}^{(Opt)})$; add these points to the importance sampling distribution and evaluate $H_k = MVN(\boldsymbol{\theta}_{1:B} \mid \boldsymbol{\theta}_{d,q}^{(Opt)}, \boldsymbol{\Sigma}_{d,q}^{(Opt)})$.

**end for**

**end for**

## 3.5 Ordinary differential equation models

Ordinary differential equation (ODE) models are mechanistic models which describe the rate of change of system states $\boldsymbol{X}(\boldsymbol{\theta}, t)$ which are realizations of a S-dimensional process $\boldsymbol{X}$ at time t with parameters $\boldsymbol{\theta} \in \boldsymbol{\Theta}^P$,

$$\frac{d\boldsymbol{X}(\boldsymbol{\theta}, t)}{dt} = f(\boldsymbol{X}(\boldsymbol{\theta}, t), \boldsymbol{\theta}). \tag{3.5}$$

The s-th system state,

$$\frac{dX_s(\boldsymbol{\theta}, t)}{dt} = f_s(\boldsymbol{X}(\boldsymbol{\theta}, t), \boldsymbol{\theta}), \tag{3.6}$$

relies on a known function $f_s$ that depends on the entire set of S system states. The ODE systems are designed to capture complex phenomena using few parameters while preserving interpretability. The goal is to estimate the parameters $\boldsymbol{\theta}$, given the noisy observations $\boldsymbol{Y} = \{y_{sj}\}$ at times $\boldsymbol{t} = \{t_{sj}\}$, for $s = 1, .., S, j = 1, ..n_s$. Usually the analytical solution to (3.5) does not exist, and hence, a numerical solver must be used with initial state $\boldsymbol{X}(0) = \boldsymbol{X}(\boldsymbol{\theta}, 0)$ to obtain the solution $\boldsymbol{X}(\boldsymbol{\theta}, t)$. In practice, the initial state vector is not known, and has to be estimated together with the unknown parameters $\boldsymbol{\theta}$.

Using a Gaussian error structure centered at the solution to the ODE model in (3.5), $\boldsymbol{X}(\boldsymbol{\theta}, t)$, the likelihood for observation vector $\boldsymbol{y}_s = (y_{s1}, .., y_{sn_s})'$ from states is:

$$P(y_{sj} \mid \boldsymbol{X}(\boldsymbol{\theta}, t_{sj}), \boldsymbol{\theta}) = N\left(\boldsymbol{X}(\boldsymbol{\theta}, t_{sj}), \boldsymbol{\sigma}^2_{\boldsymbol{y}_s}\right). \tag{3.7}$$

Small changes in parameters can lead to big changes in the dynamics of the model. Consequently, multi-modality, ridges and deep valleys of low-probability areas are common characteristics of the likelihoods in ODE models (Campbell and Steele, 2012). Standard random walk MCMC algorithms could easily get trapped in a local mode. Model relaxation methods that use model based smoothing, rather than numerically solving the ODE system in (3.5) have been designed to overcome the topological challenges (Brunel et al., 2008; Liang and Wu, 2008; Ramsay et al., 2007). These methods will be discussed in the Section 3.6.1.

### 3.5.1 Motivating example – the FitzHugh-Nagumo ODE model

The FitzHugh-Nagumo model (FitzHugh, 1961; Nagumo et al., 1962) captures the behavior of spike potentials in the giant axon of squid neurons. The FitzHugh-Nagumo model is described by a system of two non-linear differential equations, corresponding to the two state variables: voltage across the membrane, $V$, and outward currents (recovery), $R$, with a vector of parameters of interest $\boldsymbol{\theta} = (a, b, c)'$,

$$\frac{dV}{dt} = c\left(V(t) - V(t)^3/3 + R(t)\right) \text{ and } \frac{dR}{dt} = -\frac{1}{c}\left(V(t) - a + bR(t)\right). \tag{3.8}$$

The analytic solution of the ODE system (3.8) does not exist and therefore the numerical solution to the system can be used with initial states values $\{V(0), R(0)\} = \{V(\boldsymbol{\theta}, 0), R(\boldsymbol{\theta}, 0)\}$. The likelihood follows the measurement error model in (3.7), centered about the solution of (3.8), $V(\boldsymbol{\theta}, t)$ and $R(\boldsymbol{\theta}, t)$,

$$\boldsymbol{Y}_V(t) \mid \boldsymbol{\theta} \sim N\left(V(\boldsymbol{\theta}, t), \sigma^2_V\right) \text{ and } \boldsymbol{Y}_R(t) \mid \boldsymbol{\theta} \sim N\left(R(\boldsymbol{\theta}, t), \sigma^2_R\right). \tag{3.9}$$

The vector of parameters of interest in the model including the initial points is $\boldsymbol{\theta} = (a, b, c, \sigma^2_V, \sigma^2_R, V(0), R(0))'$. For expositional simplicity, we consider a one parameter model while holding the rest of the parameters fixed to the values,

$(a = 0.2, b = 0.2, \sigma_V^2 = 0.05^2, \sigma_R^2 = 0.05^2, V(0) = -1, R(0) = 1)'$, with $\boldsymbol{\theta} = c$ being the only parameter to estimate.

In order to determine the prior for this system, one could solve the ODE system numerically over a coarse grid of values of $\boldsymbol{\theta}$ and place 95% of the prior mass over those values that produce oscillations (Campbell and Steele, 2012).

As an illustrative example, placing a prior which assumes that oscillations occur an integer multiples of the true frequency of the oscillation, induces inconsistency between the prior and the data. For example, the prior,

$$P(c) = N(14, 2), \tag{3.10}$$

suggests that there is only one full oscillation in the system (Figure 3.2 A), while for the true value $c = 3$, the data exhibit two full oscillations (Figures 3.2 B and C).

Figure 3.1 A and B show that the likelihood and the target distribution exhibit multiple modes separated with deep valleys of near-zero probability regions measuring several thousands on the log scale. Standard random walk algorithms could get easily trapped in an unimportant local mode around $c = 12.05$. The prior given by the equation (3.10) covers only one of the local modes in the likelihood (Figures 3.1 A and C), and does not cover the basin of attraction of the global mode. Consequently, IMIS-Opt is trapped in the local mode that is covered by the prior (Figure 3.1D).

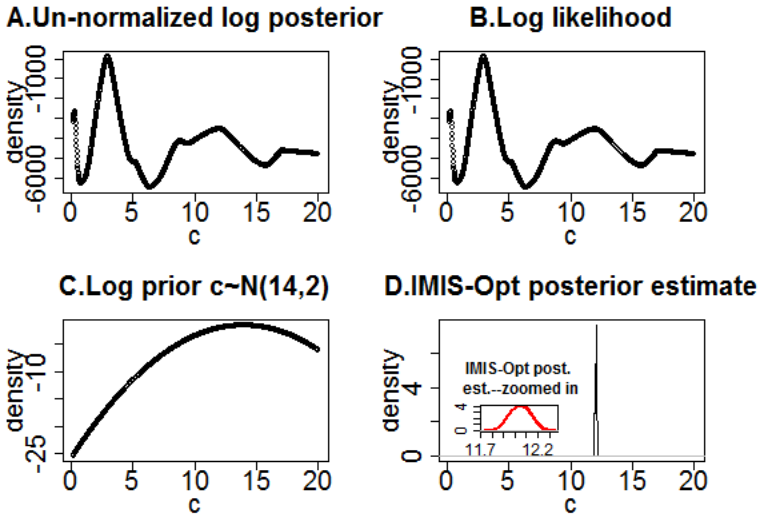


Figure 3.1: The FhN-ODE model – impact of the disagreement between the log-likelihood and log posterior (plots A and B) and log prior (plot C) on the IMIS-Opt posterior estimate (plot D). The IMIS-Opt was run with D=3, B=1000 and J=10000.

## 3.6 Illustrative example – the FitzHugh-Nagumo model revisited

We illustrate the performance of the IMIS-ShOpt using the following two models: the FhN-ODE model from Section 3.5.1 (Model 1), where only one parameter is estimated while the rest of the parameters are held fixed to their true values, and the full FhN-ODE model (Model 2) with parameters of interest $\boldsymbol{\theta} = \left(a, b, c, \sigma_V^2, \sigma_R^2, V(0), R(0)\right)'$. For comparison, the results from the performance of the IMIS-Opt on the Model 1 are also presented and discussed. Table 3.1 presents prior specifications of the two models.

Table 3.1: The two FhN models – prior specifications

| | a | b | c | $\sigma_V^2$ | $\sigma_R^2$ | $V(0)$ | $R(0)$ |
|---|---|---|---|---|---|---|---|
| Model 1 | 0.2 | 0.2 | $N(14, 2)$ | 0.05 | 0.05 | -1 | 1 |
| Model 2 | $N(0, .4)$ | $N(0, .4)$ | $N(14, 2)$ | $IGamma(3, 3)$ | $IGamma(3, 3)$ | $N(-1, .5)$ | $N(1, .5)$ |

The two FhN models – in the Model 1, prior has been assigned only for the parameter $c$, while the rest of the parameters are fixed to their true values. In Model 2 prior distributions have been assigned for all parameters.

### 3.6.1 Shotgun optimization strategy used to estimate the parameters in the FhN model

The Shotgun optimization strategy used to estimate the parameters of the FhN model comprises three different parameter estimation methods in ODE models: *i*). Non-linear Least Squares (NLS) (Bates and Watts, 1988; Seber, 1989), *ii*). Two Stage estimator (Brunel et al., 2008; Liang and Wu, 2008; Varah, 1982) and *iii*). Generalized Profiling (GP) (Ramsay et al., 2007). All three are described bellow.

#### The NLS method

Following Bates and Watts (1988), the maximum likelihood estimate $\hat{\boldsymbol{\theta}}$ is obtained by minimizing the negative log-likelihood, which in Gaussian distribution (as per (3.9)) becomes a sum of squared difference between observations and the numerical solution solution to the ODE model in (3.5),

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \sum_{s=1}^{S} \sum_{j=1}^{n_s} \left[y_{sj} - \boldsymbol{X}(\boldsymbol{\theta}, t_{sj})\right]^2. \tag{3.11}$$

The NLS method has several drawbacks. First, in order to minimize the sum of squared error in (3.11), NLS requires numerically solving the ODE system in (3.5) at each evaluation of the optimization criteria, which, in turn, requires the initial system states. NLS estimates depend on the initial guesses of the parameters of interest especially in the cases

when the sum of square error function in (3.11) exhibits multiple modes. As a result, the starting points determine whether the parameter estimate will converge to a local or global mode. Consequently, the NLS performs well in the cases when the neighborhood of the true parameters values are used as initial optimization guesses.

**The Two-Stage method**

The Two-Stage method first smooths the data as an estimate $\hat{\boldsymbol{X}}(\boldsymbol{\theta}, t)$ and then differentiates that smooth to approximate $\frac{d\boldsymbol{X}(\boldsymbol{\theta},t)}{dt}$ (Liang and Wu, 2008; Brunel et al., 2008; Varah, 1982). Parameter estimates are obtained by maximizing fidelity to the ODE model in (3.5) using the estimates from the smoothing step.

The local polynomial procedure (Fan and Gijbels, 1996) approximates the $s$-th state $\boldsymbol{X}_s(\boldsymbol{\theta}, t_{sj})$ by a $\nu$-th order polynomial, in a neighborhood of the time point $t_{s0}$, with $a_i(\boldsymbol{\theta}, t_{s0}) = \boldsymbol{X}_s^{(i)}(\boldsymbol{\theta}, t_{s0})$ for $i = 0, .., \nu$,

$$\begin{aligned} \boldsymbol{X}_s(\boldsymbol{\theta}, t_{sj}) &\approx \boldsymbol{X}_s(\boldsymbol{\theta}, t_{s0}) + (t_{sj} - t_{s0})\boldsymbol{X}_s^{(1)}(\boldsymbol{\theta}, t_{s0}) + .. + (t_{sj} - t_{s0})^s \boldsymbol{X}_s^{(\nu)}(\boldsymbol{\theta}, t_{s0})/\nu! \\ &= \sum_{i=0}^{\nu} a_i(\boldsymbol{\theta}, t_{s0})(t_{sj} - t_{s0})^i, \text{ for } s = 1, .., S, j = 1, .., n_s. \end{aligned} \tag{3.12}$$

Following Fan and Gijbels (1996), the estimators $\widehat{\boldsymbol{X}_s^{(i)}}(\boldsymbol{\theta}, t), i = 0, 1$, are obtained by minimizing the locally weighted least-square criterion,

$$\sum_{j=1}^{n_s} \left[ y_{sj} - \sum_{i=0}^{\nu} a_i (t_{sj} - t_{s0})^i \right]^2 K_h(t_{sj} - t_{s0}), \tag{3.13}$$

where h controls the size of the neighborhood around $t_{s0}$, $K_h(.) = K_h/h$ controls the weights, and $K(.)$ is a Kernel weight function.

In the second stage, the estimate $\hat{\boldsymbol{\theta}}$ is obtained by minimizing the sum of squared difference between the derivative estimate and the derivative from the ODE model,

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \sum_{s=1}^{S} \sum_{j=1}^{n_s} \left[ \widehat{\boldsymbol{X}_s^{(1)}}(\boldsymbol{\theta}, t_{sj}) - f_s(\hat{\boldsymbol{X}}(\boldsymbol{\theta}, t_{sj}), \boldsymbol{\theta}) \right]^2. \tag{3.14}$$

Although the objective function (3.14) resembles the least squares, the error term is not independently distributed. Hence, the estimator $\hat{\boldsymbol{\theta}}$ is called pseudo-least squares (PsLS) estimator. Alternatively, the SIMEX (Carroll et al., 2006) algorithm can be used to deal with measurement error in covariates for nonlinear regression models.

The Two-Stage method is computationally more efficient than the NLS, since it avoids employing the numerical solver at each evaluation of the objective function. However, this gain of computational efficiency comes at the cost of accuracy. Namely, in the first stage the data are smoothed without using the ODE model information. The ODE model is only

used in the second stage to obtain $\hat{\boldsymbol{\theta}}$ based on the first stage smoothing results. Separating the estimation procedure in two stages results in a reduced estimation accuracy of the ODE parameters (Ding and Wu, 2014). Combining the Two-Stage and the NLS method can improve parameter estimates by first obtaining the neighborhood of the estimates from the Two-Stage method and then using them as initial points for the NLS (Wu et al., 2014).

**The Generalized Profiling method**

Avoiding the numerical solution to the ODE system, the GP method uses collocation to smooth out the data which is governed by the ODE model through penalizing the deviation at the level of the derivative.

Collocation methods (Varah, 1982) approximate $\boldsymbol{X}_s(\boldsymbol{\theta}, t)$, with a linear combination of $L_s$ basis functions, $\boldsymbol{\Phi}_s(t) = (\phi_{s1}(t), .., \phi_{sL_s}(t))'$,

$$\hat{\boldsymbol{X}}_s(\boldsymbol{\theta}, t) = \boldsymbol{\Phi}_s(t)\boldsymbol{c}_s = \sum_{l=1}^{L_s} c_{sl}\phi_{sl}(t), \tag{3.15}$$

where $\boldsymbol{c}_s = (c_{s1}, .., c_{sL_s})'$ is a corresponding vector of basis coefficients, and the $\boldsymbol{c}$ is the composite vector of length $L = \sum_{s=1}^{S} L_s$, obtained by concatenating $\boldsymbol{c}_s$. The $\boldsymbol{\Phi}_s$ is a $n_s \times L_s$ matrix with values $\phi_l(t_{sj})$ for $j = 1, .., n_s$, and the $\boldsymbol{\Phi}$ is a block-diagonal matrix of dimension $n = \sum_{s=1}^{S} n_s \times L$, with matrices $\boldsymbol{\Phi}_s$ on the diagonal and otherwise zero.

Following Wang et al. (2014), the B-splines were chosen as basis functions. The GP is a cascade optimization procedure which first profiles out the basis coefficients of the ODE model based data smooth, and then estimates the ODE parameters using the profile likelihood.

The model based data smoothing is performed to obtain the basis functions coefficients. Being nuisance parameters, the basis coefficients are obtained by keeping $\boldsymbol{\theta}$ fixed, while optimizing the inner optimization criterion,

$$G\left(\boldsymbol{C} \mid \boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{Y}\right) = \sum_{s=1}^{S_0}\sum_{j=1}^{n_s} \omega_{sj}\left[y_{sj} - \boldsymbol{\Phi}_s(t_{sj})\boldsymbol{c_s}\right]^2 + \sum_{s=1}^{S} \lambda_s \int_{\boldsymbol{T}} \left[\frac{d\boldsymbol{\Phi}_s(t)\boldsymbol{c_s}}{dt} - f_s(\boldsymbol{\Phi}(t)\boldsymbol{c}, \boldsymbol{\theta})\right]^2 dt, \tag{3.16}$$

where $t$ is integrated out over the interval of maximum range of observation times over all observed states, $\boldsymbol{T} = [\min_s(\boldsymbol{t}_s), \max_s(\boldsymbol{t}_s)]$, with $\boldsymbol{t}_s = (t_{s1}, .., t_{sn_s})'$, and the $S_0$ is the dimension of the observed system states such that $S_0 \leq S$. The first term of G represents a weighted sum of squares which is a measure of how well the observed states are approximated by the basis functions, while the second term of G measures the fidelity of the basis functions to the ODE model. The smoothing parameter $\lambda$ controls the trade-off between fit to the data and fidelity to the ODE model. For notational simplicity, the dependence of $\boldsymbol{c}_s$ on $\boldsymbol{\theta}$

in (3.16) is omitted. Hence, having $\boldsymbol{c}_s(\boldsymbol{\theta})$ in (3.16) implicates that for any set of $\boldsymbol{\theta}$ the inner optimization criteria is optimized with respect to the basis functions coefficients $\boldsymbol{c}_s$.

The outer optimization criterion,

$$J\left(\boldsymbol{\theta} \mid \boldsymbol{C}, \boldsymbol{Y}\right) = \sum_{s=1}^{S_0} \sum_{j=1}^{n_s} \omega_{sj} \left[y_{sj} - \boldsymbol{\Phi}_s(t_{sj}) \boldsymbol{c}_s(\boldsymbol{\theta})\right]^2, \tag{3.17}$$

produces $\hat{\boldsymbol{\theta}}$ estimates using the basis functions coefficients estimates obtained from the inner optimization.

### 3.6.2 Performance of the Shotgun optimization strategy in the FhN-ODE model.

Rather than optimizing the posterior target distribution to find the important modes as per the IMIS-Opt optimization step, the Shotgun optimization strategy in the IMIS-ShOpt employs various modifications of the target posterior depending on the optimization method used, thus improving the exploration of the posterior space. For example, the NLS method uses the sum of squared error in (3.11) as optimization criteria, the Two-Stage method optimizes the fidelity of the data smooth to the model in (3.14), and the GP uses cascade optimization where the basis coefficients are first profiled out by the inner optimization criterion in (3.16), and then the ODE parameters are estimated using the outer optimization criterion in (3.17). In the Shotgun optimization, the parameter estimates $\hat{\boldsymbol{\theta}}$ were obtained by combining the results from different optimization criteria, while the Hessian matrices evaluated at $\hat{\boldsymbol{\theta}}$ were obtained numerically using the target posterior distribution of the FhN-ODE model.

The three methods (NLS, Two-Stage and GP) combined together discovered global and local optima of the parameters of the FhN-ODE model while widely exploring the posterior space. Namely, the prior of the parameter $c$ covers only the unimportant local mode of the target posterior centered around $c$=12.05, and therefore, the initial particles in the IMIS-ShOpt are in the basin of attraction of that local mode, thus missing the global mode. The results from the NLS were highly affected by the initial points, and consequently, the optima from the NLS were in the basin of attraction of the local mode at c=12.05. The GP method was occasionally discovering both the local and the global mode. The two-stage method proved to be the least sensitive to the initial points and hence, it was the only method among the three that discovered the global mode with any starting point. The exploration of global and local maxima obtained from the Shotgun optimization is the goal of IMIS-ShOpt.

Shotgun optimization strategy is computationally efficient due to its ability to run in parallel its constituting methods (here NLS, Two-Stage and GP). Table 3.2 shows the computational time in seconds needed to run the IMIS-ShOpt for the Model 1 and Model 2 and

the IMIS-Opt for the Model 1. The IMIS-ShOpt in Model 1 is faster than IMIS-Opt for the Model 1, because the Shotgun optimization explores the posterior space efficiently thus enabling the sampler to converge in just 2 iterations. By contrast, the optimization stage in the IMIS-Opt is less efficient and it takes 150 iterations for the algorithm to converge.

Table 3.2: Computational time

| IMIS-Opt, Model 1 | IMIS-ShOpt, Model 1 | IMIS-ShOpt, Model 2 |
|---|---|---|
| 697.325 | 521.531 | 3784.042 |

Wall-clock time in seconds of the runs from IMIS-ShOpt and IMIS-Opt on Model 1 and Model 2.

### 3.6.3 Results

Figure 3.2, B and C demonstrate that although the prior for the parameter $c$ does not cover the global mode, the IMIS-ShOpt recovers the two and a half oscillations of the true trajectories in Model 1 and Model 2. By contrast, the re-sampled trajectories obtained from the IMIS-Opt (Figure 3.2 A), recover only one oscillation of the true trajectories, while missing the other one-and-a-half oscillation. If IMIS-Opt used a Stochastic global optimizer or an evolutionary optimizer instead of gradient descent, the global maximum could have been found.

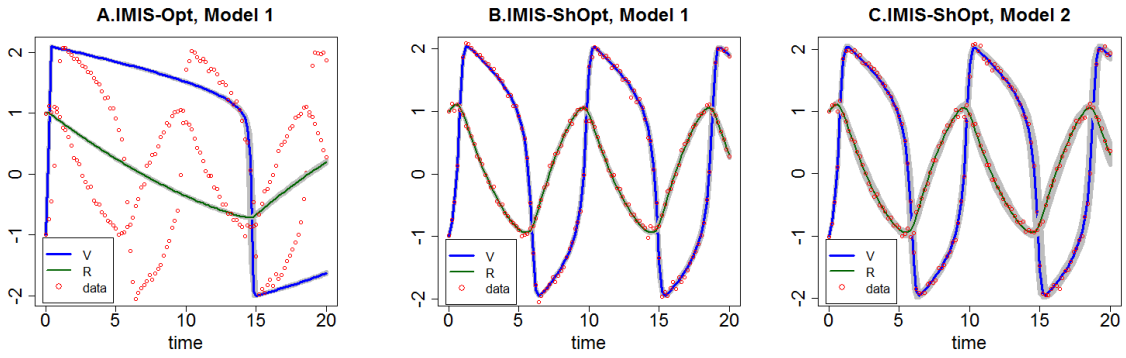


Figure 3.2: The FhN-ODE model – re-sampled trajectories using IMIS-Opt on Model 1 (plot A), IMIS-ShOpt on Model 1 (plot B) and IMIS-ShOpt on Model 2 (plot C). The gray lines represent 10000 re-sampled trajectories, the solid thick blue and green thin lines correspond to the re-sampled trajectories at the posterior mean values for the state variables V and R, respectively. The red points represent the data, which were simulated from the vector of true parameters values $\boldsymbol{\theta} = (a = 0.2, b = 0.2, c = 3, V(0) = -1, R(0) = 1)'$. The IMIS-Opt was run with D=3, B=1000 and J=10000. The IMIS-ShOpt for both models, Model 1 and Model 2, was run with D=30, Q=3, B=1000 and J=10000.

## 3.7 Illustrative example – Susceptible-Infected-Removed (SIR) epidemiological model

In this section we consider a Susceptible-Infected-Removed (SIR) epidemiological model using the data from the second black plague outbreak in the village of Eyam UK, from June 19, 1666 to November 1, 1666 (Massad et al., 2004). Since the village had been quarantined, the population size is fixed to N=261 and is stratified into states of susceptible S(t), infected I(t) and removed R(t) individuals, N=S(t)+I(t)+R(t). $R(t)$ corresponds to the number of deaths up to time t, because there is no recovery from the plague (Campbell and Lele, 2014; Golchi and Campbell, 2016). The following system of ordinary differential equations (ODE) models the disease spread dynamics:

$$\frac{dS}{dt} = -\beta S(t)I(t),\ \frac{dI}{dt} = \beta S(t)I(t) - \alpha I(t),\ \frac{dR}{dt} = \alpha I(t) \tag{3.18}$$

where $\alpha$ describes the rate of death once the individual is infected and $\beta$ describes the plague transmission. In order for the ODE system in (3.18) to be numerically solved, the initial states $S(0), I(0)$ and $R(0)$ are required. Since the number of removed at the initial time is 0, $R(0) = 0$, it follows that $S(0) = N - I(0)$, the initial states of the system reduce to $I(0)$. Hence, parameters of the model are $\boldsymbol{\theta} = (\alpha, \beta, I(0))'$. The data $\boldsymbol{Y} = (y_1, .., y_n)'$ comprise of the cumulative number of deaths up to times $(t_1, .., t_n), n = 136$. The likelihood of the data followed a binomial distribution with expected value equal to the solution $R_{(\alpha,\beta,I(0))}(t)$ to the system in (3.18).

The states S(t) and I(t) are not observed, however, the number of infected at the end of the plague is 0, and the number of infected at time one before the end of the plague must therefore equal 1 (Campbell and Lele, 2014). Two additional data points on number of infected individuals $\boldsymbol{X} = (x_{n-1} = 1, x_n = 0)'$ at times $(t_{n-1}, t_n)'$ were modeled using binomial distribution with expected value equal to the solution $I_{(\alpha,\beta,I(0))}(t)$ to the system in (3.18) at $t \in (t_{n-1}, t_n)'$ time points,

$$\begin{aligned} P(\boldsymbol{Y} \mid \alpha, \beta, I(0)) \ &= \ \prod_{i=1}^{n} \text{Binomial}\Big(y_i \mid N, \frac{R_{(\alpha,\beta,I(0))}(t_i)}{N}\Big) \times \\ &\quad \prod_{i=n-1}^{n} \text{Binomial}\Big(x_i \mid N, \frac{I_{(\alpha,\beta,I(0))}(t_i)}{N}\Big). \end{aligned} \tag{3.19}$$

Prior distributions for $\boldsymbol{\theta} = (\alpha, \beta, I(0))'$ were chosen to be:

$$\alpha, \beta \sim \text{Gamma}(1, 1), I(0) \sim \text{Binomial}(N, \frac{5}{N}). \tag{3.20}$$

### 3.7.1 Shotgun optimization strategy used in SIR-ODE model

Shotgun optimization strategy applied to the SIR-ODE model employs the NLS method described in the Section 3.6.1 using the D=3 highest weights points to initialize the optimizer, and Q=10 different optimization criteria that correspond to optimizing $\alpha$ and $\beta$ at fixed discrete values of $I(0) \in \{1, 2, 3, .., 10\}$. Optimization function was defined as the sum of squared errors of the observed data and data simulated from the model (3.11), with the optimization criteria varying with $I(0) \in \{1, 2, 3, .., 10\}$, i.e., the parameters of interest $\hat{\boldsymbol{\theta}}$ were optimized conditional on $I(0) \in \{1, 2, 3, .., 10\}$, $\hat{\boldsymbol{\theta}} \mid I(0) = \left(\hat{\alpha}, \hat{\beta}\right)'$. The Hessian matrix evaluated at $\hat{\boldsymbol{\theta}} \mid I(0) = \left(\hat{\alpha}, \hat{\beta}\right)'$ was numerically calculated using the target conditional posterior distribution $P(\alpha, \beta \mid I(0), \boldsymbol{Y})$ based on the likelihood defined in (3.19).

Implementation details are given in the Appendix D.

Table 3.3 shows the computational time in seconds needed to run the IMIS-ShOpt in comparison to that of the IMIS-Opt for the SIR model.

Table 3.3: Computational time

| IMIS-Opt | IMIS-ShOpt |
|---|---|
| 169.244 | 319.739 |

Wall-clock time in seconds of the runs from IMIS-ShOpt and IMIS-Opt on the SIR model.

### 3.7.2 Results

The mixture of continuous and discrete parameters in the SIR-ODE model induces multi-modality in the posterior space. Figure 3.4 illustrates multi-modality and topological challenges of the posterior space of the SIR model. Marginal distributions of the two continuous parameters $\alpha$ and $\beta$ exhibit three isolated modes. Clouds in the bivariate plot of $\alpha$ and $\beta$ depicts the four modes corresponding to the discrete values of $I(0) = \{6, 5, 4, 3\}$ from left to right. While the IMIS-ShOpt captures all the multiple modes in the posterior space, the IMIS-Opt gets trapped in the mode around the initial state $I(0) = 5$ (Figure 3.3). The IMIS-Opt uses gradient optimization to optimize the target posterior distribution. Regardless of the many different starting points, the gradient optimization discovers only the global mode.

## 3.8 Parameter estimation with IMIS-ShOpt using synthetic likelihood

In this section we introduce the IMIS-ShOpt with synthetic likelihood (Wood, 2010) which borrows ideas from the Approximate Bayesian Computation (ABC) framework. ABC methods (Tavaré et al., 1997; Pritchard et al., 1999) provide a framework for inference in cases

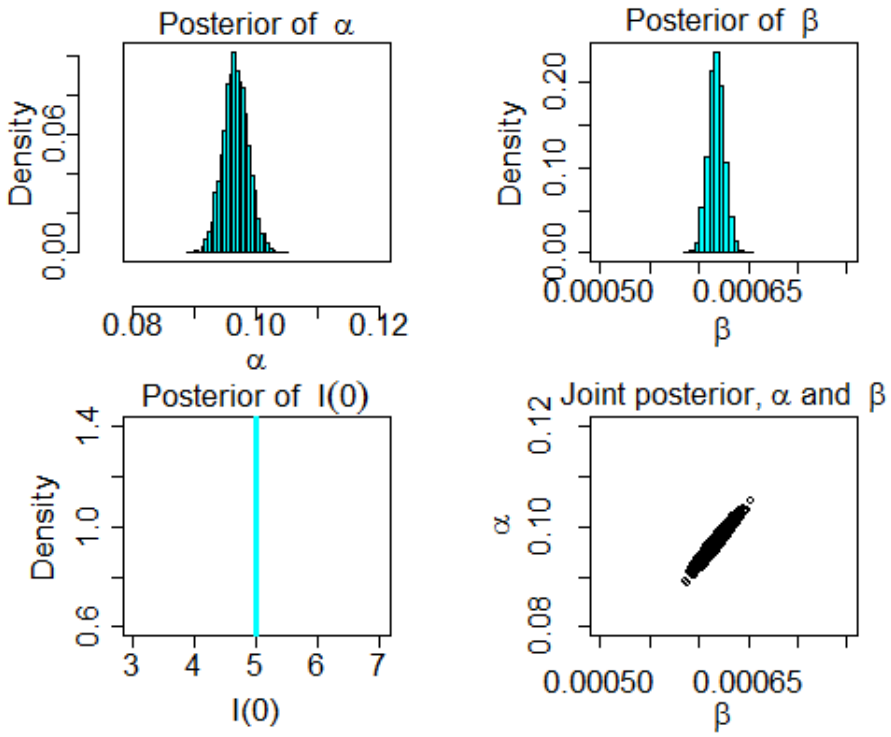


Figure 3.3: The SIR-ODE model – marginal and bivariate joint posterior distributions of sampled parameters $\alpha, \beta$ and $I(0)$ obtained from the IMIS-Opt. The IMIS-Opt algorithm was run with $N_0 = 3000, D = 3, B = 1000, J = 10000, N = 1000$ (see Appendix D for implementation details).

where the likelihood is intractable or very costly to evaluate, but simulating data from the model is relatively easy.

### 3.8.1 ABC framework

ABC posterior distribution is constructed by first simulating replicates $\boldsymbol{Z}$ for a given set of parameters $\boldsymbol{\theta}$, and then replacing the likelihood function with an indicator function $\mathbb{I}_{\mathcal{A}}(\boldsymbol{Z})$ over a matching set $\mathcal{A}$:

$$P_{\mathcal{A}}(\boldsymbol{\theta}, \boldsymbol{Z} \mid \boldsymbol{Y}) = \frac{P(\boldsymbol{\theta})P(\boldsymbol{Z} \mid \boldsymbol{Y})\mathbb{I}_{\mathcal{A}}(\boldsymbol{Z})}{\int_{\mathcal{A}} P(\boldsymbol{\theta})P(\boldsymbol{Z} \mid \boldsymbol{Y})d\boldsymbol{Z}}. \tag{3.21}$$

The matching set $\mathcal{A}$ compares the replicate data $\boldsymbol{Z}$ generated from the model $P(\boldsymbol{Z} \mid \boldsymbol{\theta})$ with the observed data $\boldsymbol{Y}$. In particular, the matching set $\mathcal{A}$ measures similarity between summary statistics $\boldsymbol{s}(.)$ of the simulated data and observed data. For discrete data, the matching can be exact, $\mathcal{A} = \{\boldsymbol{Z} \mid s(\boldsymbol{Z}) = s(\boldsymbol{Y})\}$. The approximation error of the ABC depends on the matching criteria $\mathcal{A}$, and it goes to zero if the summary statistics $\boldsymbol{s}(.)$ is sufficient. In the continuous data case, since $P(\boldsymbol{Z} \in \mathcal{A}) = 0$, a distance metric $\rho(.)$ and a tolerance $\epsilon > 0$ are defined in order to control the approximation error, and thus, the

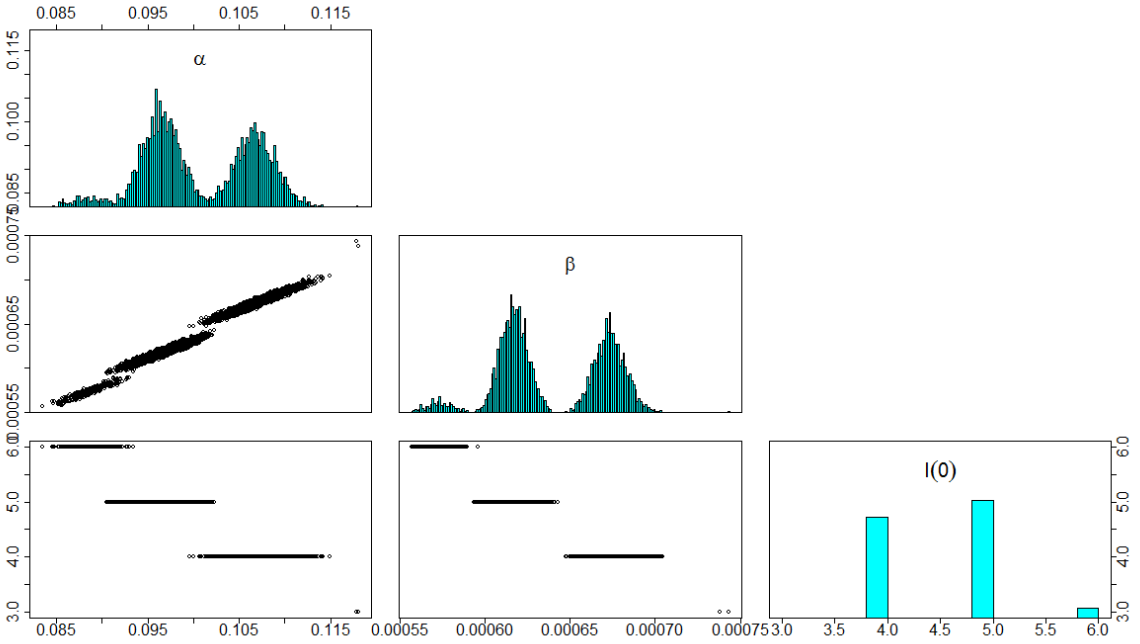


Figure 3.4: The SIR-ODE model – marginal (diagonal) and bivariate joint (off-diagonal) posterior distributions of sampled parameters $\alpha, \beta$ and $I(0)$ obtained from the IMIS-ShOpt. The IMIS-ShOpt algorithm was run with $N_0 = 3000, Q = 10, D = 3, B = 1000, J = 10000, N = 1000$.

matching set $\mathcal{A}$ becomes:

$$\mathcal{A} = \{\boldsymbol{Z} \mid \rho(s(\boldsymbol{Z}), s(\boldsymbol{Y})) < \epsilon\}. \tag{3.22}$$

The choice of the summary statistics $\boldsymbol{s}(.)$ is crucial to the accuracy of ABC. In practice, sufficient statistics are rarely known, and instead approximately sufficient summary statistics are used. Choice of the most informative summary statistics has been studied in the literature. Joyce and Marjoram (2008) studied effectiveness of sequentially added summary statistics using a likelihood ratio test. Prangle et al. (2014) proposed a method for semi-automatic selection of good summary statistics for a given model choice based on fitting regressions on simulated data. Fearnhead and Prangle (2010) developed another semi-automated approach, where the posterior means estimates, which are obtained from the simulated parameters and data sets, are used as summary statistics.

The basic ABC algorithm starts with parameters draws, $\boldsymbol{\theta}^*$, from the prior distribution. Then, a replicate data set $\boldsymbol{Z}$ is simulated for each draw from the prior, and if $\boldsymbol{Z} \in \mathcal{A}$ then the $\boldsymbol{\theta}^*$ is accepted as a draw from the posterior, and it is otherwise rejected.

When available, a set of summary statistics $\boldsymbol{s}(.) = (\boldsymbol{s_1}(.), .., \boldsymbol{s_T}(.))^T$ could be used in order to reduce the approximation error. The matching set of target posterior comprises the entire collection of summary statistics. Approximations to the target posterior that

correspond to the subsets

$$\mathcal{A}_t = \{\boldsymbol{Z} \mid \rho(s_1(\boldsymbol{Z}), s_1(\boldsymbol{Y})) < \epsilon_1, .., \rho(s_t(\boldsymbol{Z}), s_t(\boldsymbol{Y})) < \epsilon_t\}, t = 1, .., T, \tag{3.23}$$

are more dispersed and easier to sample from. For example, in order to increase the strictness of a model constraint, the Sequentially Constrained Markov Chain with ABC algorithm (Golchi and Campbell, 2016), employed a sequence of approximate posterior distributions $\{\pi_{\mathcal{A}_t}\}_{t=1}^{T}$ defined on a decreasing sequence of matching sets $\mathcal{A}_1 \supseteq \mathcal{A}_2 \supseteq ... \supseteq \mathcal{A}_T$, where $\mathcal{A}_t$ is given in (3.23). This idea originates from the SMC ABC algorithm by Del Moral et al. (2012); Peters et al. (2012) where in order to reduce the approximation error, a sequence of approximations to the target posterior is defined using a decreasing sequence of tolerance levels $\epsilon_1 > .. > \epsilon_T$, with $\epsilon_T$ corresponding to the target posterior distribution. Consequently, introducing a set of summary statistics has an advantage of moving the $\boldsymbol{s}(.)$ closer to the sufficient summary statistics. The proposed IMIS-ShOpt employs the idea of using different subsets of the summary statistics collection to explore different parts of the posterior space within a synthetic likelihood framework.

### 3.8.2 IMIS-ShOpt with synthetic likelihood outline

The likelihood defined as an indicator function $\mathbb{I}_{\mathcal{A}}(\boldsymbol{Z})$ as per (3.21), with matching set defined over the entire set of available summary statistics, could be used to calculate the importance sampling weights. However, the matching set $\mathcal{A}$ defined in (3.22) relies on vector of tolerance levels, $\boldsymbol{\epsilon}$, and a distance measure $\rho(.)$, which need to be tuned to ensure that there are at least few initial particles with non-zero weights so that the Shotgun optimization can be initialized. Finding a trade-off between small tolerance levels and initial particles with non-zero weights is not an easy task and harms the efficiency of the IMIS-ShOpt algorithm.

Namely, small tolerance levels yield only few initial particles with non-zero weights, which in turn contributes to loss of efficiency. Having only few initial non-zero weight particles leads to adding only small amount of new incremental samples to the importance sampling distribution during the optimization stage. Moreover, due to the small tolerance levels, there would be only small number of non-zero particles after the importance sampling distribution is repopulated with new samples from both the optimization stage and the importance sampling stage. Consequently, many iterations in the importance sampling stage would be needed for the IMIS-ShOpt algorithm to converge.

In chaotic systems, likelihood-based inference breaks down because small changes in parameters induce big changes in the system states. To avoid the requirement of the tolerance levels and the distance measure, and to gain the efficiency from the Shotgun optimization thereof, we approximate the likelihood function with a synthetic likelihood (Wood, 2010). As an alternative to the likelihood approach, the synthetic likelihood captures important

dynamics in the data using the summary statistics of the replicate data rather than the noisy observations. Although synthetic likelihood approach employs ideas from the ABC framework, the log synthetic likelihood behaves like a conventional log likelihood in the limit, when the number of replicated data sets approaches infinity, but acts with reduced efficiency because of the lack of sufficient statistics.

Following Wood (2010), synthetic likelihood of any given value of $\boldsymbol{\theta}$, given the summary statistics of the replicate data sets, can be constructed as follows. For a given set of parameters $\boldsymbol{\theta}$, $N_{\boldsymbol{Z}}$ replicate data sets $\boldsymbol{Z} = \{\boldsymbol{Z}_1, .., \boldsymbol{Z}_{N_Z}\}$ are simulated from the model $P(\boldsymbol{Z} \mid \boldsymbol{\theta})$, and the vector of summary statistics $\boldsymbol{S} = \{\boldsymbol{s}(\boldsymbol{Z}_1), .., \boldsymbol{s}(\boldsymbol{Z}_{N_{\boldsymbol{Z}}})\}$ is calculated for each replicate data set. The mean of the $N_{\boldsymbol{Z}}$ summary statistics can be calculated as $\hat{\boldsymbol{\mu}}_\theta = \frac{\sum_{i=1}^{N_{\boldsymbol{Z}}} \boldsymbol{s}(\boldsymbol{Z}_i)}{N_{\boldsymbol{Z}}}$, $\boldsymbol{S}^* = \{\boldsymbol{s}(\boldsymbol{Z}_1) - \hat{\boldsymbol{\mu}}_\theta, .., \boldsymbol{s}(\boldsymbol{Z}_{N_{\boldsymbol{Z}}}) - \hat{\boldsymbol{\mu}}_\theta\}$, and the variance-covariance matrix can be calculated as $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}} = \frac{\boldsymbol{S}^* \boldsymbol{S}^{*'}}{N_{\boldsymbol{Z}}-1}$. Then the log-synthetic likelihood is given by a multivariate normal distribution, $MVN(\boldsymbol{S} \mid \hat{\boldsymbol{\mu}}_\theta, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}})$, i.e.,

$$\mathcal{L}_s(\boldsymbol{\theta} \mid \boldsymbol{S}) = -\frac{1}{2}(\boldsymbol{S} - \hat{\boldsymbol{\mu}}_\theta)' \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}}^{-1} (\boldsymbol{S} - \hat{\boldsymbol{\mu}}_\theta) - \frac{1}{2} log|\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}}|. \tag{3.24}$$

When a set of candidate summary statistics is available, the target likelihood is defined over the entire set of available summary statistics.

The proposed IMIS-ShOpt algorithm draws samples from the posterior distribution of the parameters of interest in models where likelihood function is computationally very costly to evaluate. In the initial stage, $N_0$ samples $\{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, ..., \boldsymbol{\theta}_{N_0}\}$ are drawn from the prior distribution $P(\boldsymbol{\theta})$. Sampling weights are calculated using the target synthetic likelihood $\mathcal{L}_s(\boldsymbol{\theta} \mid \boldsymbol{S})$, defined over the entire set of available summary statistics.

The objective function in the Shotgun optimization step uses different approximations to the synthetic likelihood $\mathcal{L}_s(\boldsymbol{\theta} \mid \boldsymbol{S})$ defined over subsets of the entire set of summary statistics. These approximations to the target synthetic likelihood might explore different regions of the posterior space. The strategy of defining different approximations to the synthetic likelihood was used to construct several different optimization criteria in the Shotgun optimization stage of the IMIS-ShOpt, one for each random subset of summary statistics. The Hessian matrix is calculated using the target synthetic likelihood which operates on the entire set of the available summary statistics.

The pseudo-code of the IMIS-ShOpt algorithm with synthetic likelihood is given in the Algorithm 5.

**Algorithm 5** The IMIS-ShOpt with synthetic likelihood

**Goal: Parameter estimation**

**Input:** Data, likelihood function, synthetic likelihood function, prior distribution and the model.

Initialize $N$ – the number of iterations, $B$ –the number of incremental points, $D$ – the number of different initial points for the optimization, $Q$ – the number of different optimization criteria, $N_0$ – the number of initial samples from the prior and $J$ – the number of re-sampled points.

**Initial stage:** Draw $N_0$ samples $\boldsymbol{\Theta}_0 = \{\boldsymbol{\theta_1}, \boldsymbol{\theta_2}, ..., \boldsymbol{\theta_{N_0}}\}$ from the prior distribution $P(\boldsymbol{\theta})$.

**for** $k = 1 : N$ **do**

**if** $k = 1$ **then**

For each $\boldsymbol{\theta}_i$, $i = 1, .., N_0$, simulate $N_{\boldsymbol{Z}}$ vectors of replicate data $\boldsymbol{Z}_i = \{\boldsymbol{Z}_1, .., \boldsymbol{Z}_{N_{\boldsymbol{Z}}}\}$ from the model, $P(\boldsymbol{Z} \mid \boldsymbol{\theta}_i)$.

For each $\boldsymbol{\theta}_i$, $i = 1, .., N_0$, calculate the vector of entire set of available summary statistics, $\boldsymbol{S} = \{\boldsymbol{s}(\boldsymbol{Z}_1), .., \boldsymbol{s}(\boldsymbol{Z}_{N_{\boldsymbol{Z}}})\}$ and construct the synthetic likelihood using (3.24).

For each $\boldsymbol{\theta}_i, i = 1, .., N_0$ calculate the sampling weights,

$$w_i^{(k)} = \frac{\mathcal{L}_s(\boldsymbol{\theta}_i \mid \boldsymbol{S})}{\sum_{j=1}^{N_0} \mathcal{L}_s(\boldsymbol{\theta}_j \mid \boldsymbol{S})} \tag{3.25}$$

**Optimization stage:**

**for** $d = 1 : D$ **do**

Find the d-th maximum weight point $\boldsymbol{\theta}_d^{initial} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, \boldsymbol{w}^{(k)}(\boldsymbol{\theta})$, $\boldsymbol{\theta} \in \boldsymbol{\Theta}_{d-1}$ to initialize Q optimizers.

**for** $q = 1 : Q$ **do**

Use q-th optimization method to optimize $\boldsymbol{\theta}$, the objective function is $\mathcal{L}_s(\boldsymbol{\theta} \mid \boldsymbol{S})$ based on a subset of summary statistics, i.e., obtain local maxima $\boldsymbol{\theta}_{d,q}^{(Opt)}$. Obtain the corresponding inverse Hessian $\boldsymbol{\Sigma}_{d,q}^{(Opt)}$ using the target synthetic likelihood.

Update $\boldsymbol{\Theta}_d$ by excluding $\frac{N_0}{DQ}$ nearest neighbor points, $\boldsymbol{\theta}_k \in \boldsymbol{\Theta}_{d-1}$, that minimize the Mahalanobis distance,

$$(\boldsymbol{\theta}_k - \boldsymbol{\theta}_{d,q}^{(Opt)})'(\boldsymbol{\Sigma}_{d,q}^{(Opt)})^{-1}(\boldsymbol{\theta}_k - \boldsymbol{\theta}_{d,q}^{(Opt)}). \tag{3.26}$$

Draw $B$ samples $\boldsymbol{\theta}_{1:B} \sim MVN(\boldsymbol{\theta}_{d,q}^{(Opt)}, \boldsymbol{\Sigma}_{d,q}^{(Opt)})$; add these points to the importance sampling distribution $P(\boldsymbol{\theta} \mid \boldsymbol{Y})$ and evaluate $H_k = MVN(\boldsymbol{\theta}_{1:B} \mid \boldsymbol{\theta}_{d,q}^{(Opt)}, \boldsymbol{\Sigma}_{d,q}^{(Opt)})$.

**end for**

**end for**

**else**

**Importance sampling stage:**

For each $\boldsymbol{\theta}_i, i = 1, .., N_k$ calculate weights:

$$w_i^{(k)} = \frac{cP(\boldsymbol{\theta}_i)\mathcal{L}_s(\boldsymbol{\theta}_i \mid \boldsymbol{S})}{\frac{N_0}{N_k}P(\boldsymbol{\theta}_i) + \frac{B}{N_k}\sum_{s=1}^{k} H_s(\boldsymbol{\theta}_i)}, \tag{3.27}$$

**Algorithm 5** The IMIS-ShOpt with synthetic likelihood - continued

where $N_k = N_0 + B(QD + k)$ and $c = 1/\sum_{i=1}^{N_k} w_i^{(k)}$.

Choose a maximum weight input, $\boldsymbol{\theta}_k$, and estimate $\boldsymbol{\Sigma}_k$ as the weighted covariance of B inputs with smallest Mahalanobis distance,

$$w_p(\boldsymbol{\theta})\left(\boldsymbol{\theta}-\boldsymbol{\theta}_k\right)'\left(\boldsymbol{\Sigma}_{\boldsymbol{\pi}}\right)^{-1}\left(\boldsymbol{\theta}-\boldsymbol{\theta}_k\right),$$

where the weights are $w_p(\boldsymbol{\theta}) = c_1(\boldsymbol{w}^{(k)} + 1/N_k)$, $\Sigma_\pi$ is the covariance of the initial importance distribution and $c_1 = 1/w_p(\boldsymbol{\theta})$.

Draw $B$ samples $\boldsymbol{\theta}_{1:B} \sim MVN(\boldsymbol{\theta}_k, \boldsymbol{\Sigma}_k)$; add these points to the importance sampling distribution and evaluate $H_k = MVN(\boldsymbol{\theta}_{1:B} \mid \boldsymbol{\theta}_{m,k}, \boldsymbol{\Sigma}_k)$.

**end if**

**if** $\sum_{1}^{N_k}(1-(1-w^{(k)})^J) \geq J(1-\exp(-1))$ i.e., importance sampling weights are approximately uniform **then** exit for loop

**end if**

**end for**

**Re-sampling stage:**

Re-sample $J$ points with replacement from $\{\boldsymbol{\theta_1}, .., \boldsymbol{\theta_{N_k}}\}$ and weights $w^{(k)}$.

## 3.9 Illustration of the IMIS-ShOpt with synthetic likelihood performance through a chaotic stochastic model

Consider a chaotic stochastic difference model, where full likelihood-based inference fails. The model exhibits intractable or expensive-to-evaluate likelihoods, but it is relatively easy to simulate data from the model.

Following Gilpin and Ayala (1973), the ecological theta-Ricker model, states that the abundance of the population in the next time point, $N_{t+1}$, is equal to the abundance at the current time point $N_t$, multiplied by the exponent of the growth rate, $\exp\left(r(1-\frac{N_t}{K})^{\tilde{\theta}} + \epsilon_t\right)$, over the time step $t$. The process noise, also known as environmental noise is modeled as $\epsilon_t \sim N(0, \sigma_p^2)$ and $K$ quantifies carrying capacity. The theta-Ricker model can be written as follows,

$$N_{t+1} = N_t \exp\left(r\left(1-\left(\frac{N_t}{K}\right)^{\tilde{\theta}}\right)+\epsilon_t\right), \tag{3.28}$$

.

The theta-Ricker model is defined with parameters $\boldsymbol{\theta} = (r, \phi, \sigma_p^2, \tilde{\theta})'$. The data are outcomes of the Poisson distribution with mean $\phi N_t$, where $\phi$ is a scaling parameter,

$$y_t \sim Poisson(\phi N_t).$$

The IMIS-ShOpt algorithm was used to estimate the parameters of the theta-Ricker model. The data were simulated from $\boldsymbol{\theta} = (\log r = 0.5, \phi = 4, \sigma^2 = 0.01, \log\tilde{\theta} = 1)'$ at

$T$=50 time steps with initial population $N_0 = 3$ and $K = 100$. Prior distributions were defined independently, $\log r \sim N(0.5, 1), \phi \sim \chi^2(df = 4), \sigma_p^2 \sim IGamma(shape = 2, scale = 0.05), \log \tilde{\theta} = N(1, 1)$. The IMIS-ShOpt was initialized with $B = 1000, J = 3000, N_0 = 10000, N = 500, D = 10, Q = 3, N_Z = 30$.

The set of summary statistics used in IMIS-ShOpt is a modification of the set from Golchi and Campbell (2016),

$$\begin{aligned}\boldsymbol{S} = \{&median(\boldsymbol{Y}), \sum_{i=1}^{n} \frac{y_i}{n}, \frac{\sum_{i=1}^{n} y\mathbb{I}_{(1,\infty)}(y_i)}{\sum_{i=1}^{n} \mathbb{I}_{(1,\infty)}(y_i)}, \sum_{i=1}^{n} y\mathbb{I}_{(10,\infty)}(y_i), \sum_{i=1}^{n} \mathbb{I}_0(y_i), \\ &Quantile_{0.75}(\boldsymbol{Y}), max(\boldsymbol{Y}), \sum_{i=1}^{n} \mathbb{I}_{(100,\infty)}(y_i), \sum_{i=1}^{n} \mathbb{I}_{(300,\infty)}(y_i), \\ &\sum_{i=1}^{n} \mathbb{I}_{(500,\infty)}(y_i), \sum_{i=1}^{n} y\mathbb{I}_{(800,\infty)}(y_i)\}. \end{aligned} \tag{3.29}$$

### 3.9.1 The Shotgun optimization

The target optimization function is the synthetic likelihood in (3.24) defined over the entire set of summary statistics $\boldsymbol{S}$ in (3.29). The $q$-th optimization method in the Shotgun optimization strategy initialized at fixed $d$, corresponds to an approximation to the target synthetic likelihood, $\mathcal{L}_s(\boldsymbol{\theta} \mid \tilde{\boldsymbol{S}})$, defined over a random subset of seven unique summary statistics from the entire set of summary statistics, $\tilde{\boldsymbol{S}} = \{s_i, s_j, s_k, s_l, s_m, s_o, s_p \mid i, j, k, l, m, o, p = 1, .., 11\} \subseteq \boldsymbol{S}$. Different approximations to the target synthetic likelihood explore different parts of the posterior space, and therefore, combining the results should lead to discovering all important posterior modes. Hence, multiple optimization criteria in the Shotgun optimization were defined to correspond to distinct approximations to the synthetic likelihood.

Although the locations of the posterior modes were discovered using different approximations to the target synthetic likelihood, the Hessian matrices of the posterior modes were obtained numerically using the target synthetic likelihood, $\mathcal{L}_s(\boldsymbol{\theta} \mid \boldsymbol{S})$.

### 3.9.2 Results

The IMIS-ShOpt with synthetic likelihood produces reasonable parameter estimates. The results, presented as kernel density estimates of the approximate marginal posteriors, are given in Figure 3.5. Figure 3.6 shows that the weights of all the particles in the importance sampling distribution before the final re-sampling stage are non-zero in the neighborhood of the true parameter values. In addition, Figure 3.6 demonstrates that before the final re-sampling stage the importance sampling distribution of the process noise variance, $\sigma_p^2$, contains particles with negative values. These points are added to the importance sampling

distribution during the optimization and importance sampling stage, but do not survive the final re-sampling stage because they have zero weights as shown in Figures 3.5 and 3.6. Rather than harming the importance sampling distribution, the negative-valued points help in better exploration of the posterior surface.

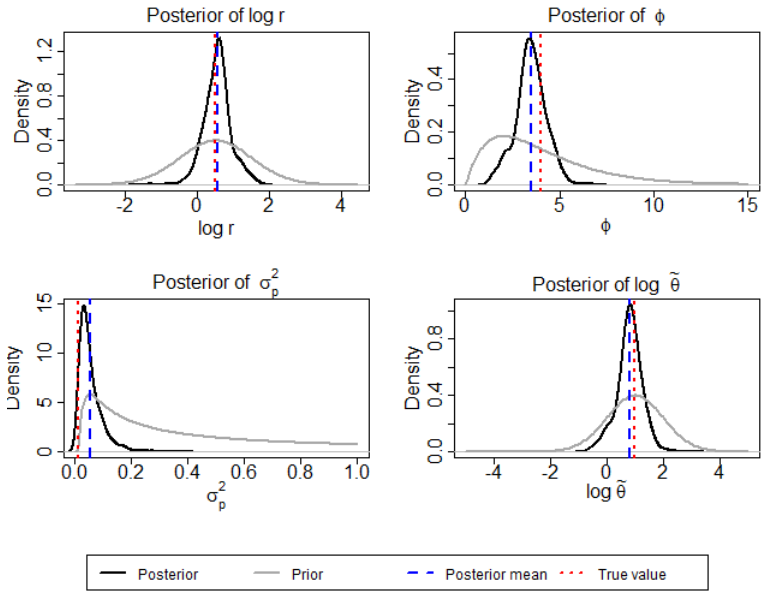


Figure 3.5: The theta-Ricker model – marginal posterior distributions of the parameters obtained from the final re-sampling stage. The vertical lines are drawn at the posterior mean (blue dashed) and the true value (red dotted). The gray distributions represent the priors.

The Shotgun optimization helps exploring the parameter space through the approximations to the target synthetic likelihood. Namely, the target synthetic likelihood, which employs the entire set of the summary statistics, exhibits narrow spiky modes which leads to optimization difficulties. Approximations to the target synthetic likelihood constructed by randomly chosen subsets of seven summary statistics, are more diffuse then the target synthetic likelihood, and hence, easier to optimize. Shotgun optimization combines results from different approximations to the target synthetic likelihood, thus resulting in more fully exploration of the parameter space.

## 3.10 Discussion

This chapter proposes a general optimization framework, the Shotgun optimization, which relies on the idea that no single method outperforms other methods in every situation. Different methods employ different model variations which leads to exploring different regions of the posterior space. Combining the results from different methods balances discovery of global and local modes, which results in more fully explored posterior space. Some meth-

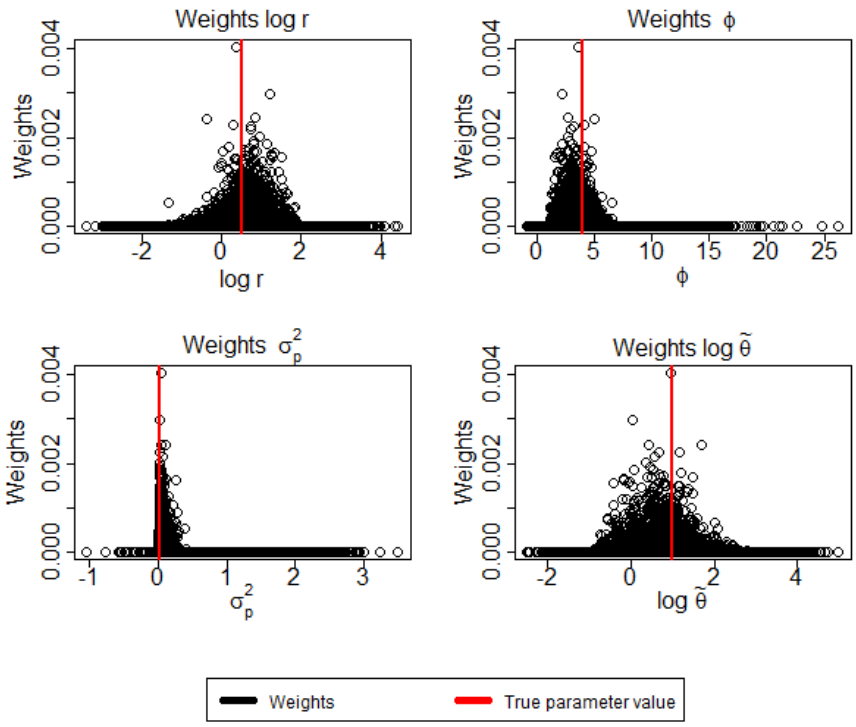


Figure 3.6: The theta-Ricker model – weights of the particles in the importance sampling distribution before re-sampling. The vertical lines are drawn at the true parameter values.

ods produce better estimates of the parameters, while others introduce bias. Merging the results from different methods together can overcome the introduced bias.

Throughout this chapter, we employed the Shotgun optimization in various scenarios in the framework of the Incremental Mixture Importance Sampling with Optimization (IMIS-Opt) algorithm, and demonstrated that the performance of the Shotgun optimization is superior to that of a single optimization method. The Incremental Mixture Importance Sampling with Shotgun optimization (IMIS-ShOpt), is useful in cases where posterior topologies are complex and exhibit multiple isolated modes separated with deep valleys of low probabilities. In Section 4.3.1 we demonstrated that the gradient based optimization in the IMIS-Opt performs poorly when the prior caries information that is inconsistent with the likelihood, which implies that the prior is chosen for optimization purposes rather than for representing the expert opinion. The IMIS-ShOpt removes this dependence on the prior by combining results from several different optimization methods, and its success depends on Shotgun optimization step discovering all the relevant optima. Consequently, the IMIS-ShOpt reestablishes the role of the prior to convey the expert knowledge rather than being chosen for optimization convenience. In addition, in Section 3.8 we demonstrated that the IMIS-ShOpt combined with synthetic likelihood (SL) can be used to estimate parameters of interest in models where likelihood is extremely costly to evaluate.

The Shotgun optimization strategy is a general framework which can be applied in any model type. Given a model type, competing parameter estimation methods deal with the posterior topologies in different ways, which leads to exploring diverse and potentially informative locations of the parameter space. The Shotgun optimization incorporates results from different competing methods or from different optimization criteria, and ensures that the parameter space is more fully explored. In addition, the Shotgun optimization is computationally efficient, since it can be easily parallelized. For instance, in the FhN-ODE model, the Shotgun optimization method runs in parallel the following three parameter estimation techniques for ODE models: the Non-linear Least Squares, the Two-Stage and the Generalized Profiling. Each of the methods discovers either a local mode or the global mode, but combined together the three methods find all the important modes. Similarly, in the SIR-ODE model, the Shotgun optimization consists of fitting the Non-linear Least Squares locally at several possible locations of the posterior modes corresponding to different values of the initial infection state. The Shotgun optimization strategy in the IMIS-ShOpt with synthetic likelihood merges results from different optimization criteria defined by different approximations to the target synthetic likelihood. Each approximation to the target synthetic likelihood corresponds to a randomly chosen subset of summary statistics from the entire collection of seven summary statistics thus exploring different locations of the posterior space.

# Chapter 4

# Bayesian Information Criterion (BIC) for Multimodal Distributions

## 4.1 Introduction

Many methods have been proposed for statistically selecting a model from a set of candidate models, such as likelihood ratio tests, Bayes Factors, and information criteria. For a general overview on model selection techniques, including both frequentist and Bayesian approaches, see (Kadane and Lazar, 2004; Chipman et al., 2001). Classical frequentist methods are based on comparing extra residual sum of squares from nested models with and without a subset of variables, for example the well known $C_p$ statistic (Kennard, 1971) and likelihood ratio test (Neyman and Pearson, 1992; Wilks, 1938). Some modern variations of frequentist techniques are the Risk Inflation Criterion (Foster and George, 1994), defined as the maximum possible increase in risk due to including the variables in the model, and the Covariance Inflation Criterion (Tibshirani and Knight, 1999) used in prediction problems. Bayes factors compare the posterior probabilities of models conditional on the data, while integrating over uncertainty in the model parameters. Information criterion strategies are approximations to these frequentists or Bayesian approaches (Konishi and Kitagawa, 2008). The Bayesian Information Criterion (BIC) (Schwarz et al., 1978) approximates the posterior probability of the model conditional on the data by applying Laplace's method to integrate over model parameters. The BIC relies on the Laplace approximation (LA) to estimate the posterior probability of the modes, and hence it works well under the assumption that the posterior distribution over model parameters is highly peaked around one mode and the sample size is large.

However, even when the posterior distribution for model parameters is multi-modal, the BIC takes into account only one mode. Evaluating the BIC at a local mode may result in

selecting the incorrect model. When multiple modes exist evaluating the BIC within one mode will give a poor representation of the relative quality of model fit.

This chapter addresses the problem of multi-modality in the posterior space of the model parameters, while preserving the computational advantages of the BIC. For example, in X-ray crystallography one crystal protein molecule is built of thousands of atoms and the goal is to learn the three-dimensional coordinates and atomic displacement for each of the atoms within this molecule. The BIC is used for choosing isotropic or anisotropic atomic displacement parameters, and to select the model with the best number of conformers (Woldeyes et al., 2014). The BIC is a useful model selection tool in problems such as crystallography, because it approximates the posterior probability of the model and avoids high dimensional Monte Carlo integration over the model parameters. However, the BIC assumptions aren't valid because of the inherent multi-modality.

The proposed Multimodal Bayesian Information Criterion (MBIC) combines information from multiple modes by approximating the posterior distribution by a mixture of locally weighted unimodal distributions (Gelman et al., 2014). Laplace approximations are applied locally to each of the weighted unimodal distributions to approximate the marginal likelihood over the model parameters.

The MBIC combines information for all discovered modes, and hence, it is a step closer to full consideration of the posterior space compared to the BIC. To discover all the relevant posterior modes, we propose an optimization strategy that uses the idea of combining the results from different runs of the optimizer started at different points. Different initial points lead the optimizer to explore different parts of the posterior topology thus resulting in discovering all the posterior modes. We refer to this optimization strategy as Shotgun optimization (ShOpt).

The rest of the chapter is organized as follows. In the Section 4.2 the Bayesian approach to model selection and comparison is described; the underlying LA of the BIC is introduced along with demonstration of its failure to approximate the analytical marginal likelihood in a simple bimodal model; the Multi-modal BIC is proposed, and its performance is demonstrated in bimodal case. In the Section 4.5, results from the model selection abilities of the BIC, the LA and the Multi-modal BIC in a mixture model of Gaussian distributions are presented and discussed. The concluding remarks follow in the Section 4.6, and mathematical derivations together with the description of the ShOpt algorithm are given in the Appendices E-I.

## 4.2 Bayesian approach to model selection and comparison

The Bayesian approach for model uncertainty given a set of candidate models $\mathbb{M} = \{M_1, .., M_L\}$ for data $\boldsymbol{Y} = (y_1, .., y_n)^{'}$, relies on the posterior probability of the model,

$$P(M_l \mid \boldsymbol{Y}) = \frac{P(\boldsymbol{Y} \mid M_l)P(M_l)}{\sum\limits_{i=1}^{L} P(\boldsymbol{Y} \mid M_i)P(M_i)}, \tag{4.1}$$

where the marginal likelihood, $P(\boldsymbol{Y} \mid M_l)$, is found by integrating over the $q_l$ dimensional parameter vector $\boldsymbol{\theta}_l$,

$$P(\boldsymbol{Y} \mid M_l) = \int P(\boldsymbol{Y} \mid \boldsymbol{\theta}_l, M_l)P(\boldsymbol{\theta}_l \mid M_l)d\boldsymbol{\theta}_l = \int P(\boldsymbol{\theta}_l, \boldsymbol{Y} \mid M_l)d\boldsymbol{\theta}_l. \tag{4.2}$$

Pairwise comparison between $M_1$ and $M_2$ is performed via posterior odds,

$$\underbrace{\frac{P(M_1 \mid \boldsymbol{Y})}{P(M_2 \mid \boldsymbol{Y})}}_{\text{Posterior odds}} = \underbrace{\frac{P(\boldsymbol{Y} \mid M_1)}{P(\boldsymbol{Y} \mid M_2)}}_{\text{Bayes factor}} \times \underbrace{\frac{P(M_1)}{P(M_2)}}_{\text{Prior odds}} . \tag{4.3}$$

The Bayes Factor indicates how the data update the prior odds to yield the posterior odds. Under the uniform prior for the models, i.e., $P(M_1) = P(M_2)$, the Bayes Factor equals the posterior odds. The posterior probability $P(M_l \mid \boldsymbol{Y})$ is a measure of 'trueness' to the data, consequently, the model selection strategy chooses the model with the largest posterior odds Chipman et al. (2001). Hence, obtaining accurate estimates of the marginal likelihood in (4.2) is crucial to Bayesian model selection.

## 4.3 Bayesian Information Criterion and Laplace Approximation

Following Tierney and Kadane (1986); Konishi and Kitagawa (2008), the integral in (4.2), can be approximated by using Laplace's method for integrals,

$$P(\boldsymbol{Y} \mid M_l) = \int \exp\left(\log\left(P(\boldsymbol{Y} \mid \boldsymbol{\theta}_l, M_l)\right)\right)P(\boldsymbol{\theta}_l \mid M_l)d\boldsymbol{\theta}_l. \tag{4.4}$$

Let $l(\boldsymbol{\theta}_l) = \log\left(P(\boldsymbol{Y} \mid \boldsymbol{\theta}_l, M_l)\right))$. The Laplace's method works well under under the asymptotic argument that as the sample size, $n$, increases the integrand in (4.4) becomes a Gaussian distribution with peak at the maximum likelihood estimate $\hat{\boldsymbol{\theta}}_l$, and hence, the approximation will depend only on the behavior of the likelihood function $l(\boldsymbol{\theta}_l)$ around its mode $\hat{\boldsymbol{\theta}}_l$.

In the case of a unimodal model, the Laplace approximation is applied as follows. The log-likelihood $l(\boldsymbol{\theta}_l)$ in (4.4) is replaced with its second order Taylor expansion at $\hat{\boldsymbol{\theta}}_l$,

$$l(\boldsymbol{\theta}_l) = l(\hat{\boldsymbol{\theta}}_l) - \frac{n}{2}(\theta_l - \hat{\boldsymbol{\theta}}_l)' H(\hat{\boldsymbol{\theta}}_l)(\theta_l - \hat{\boldsymbol{\theta}}_l) + .., \tag{4.5}$$

where

$$H(\hat{\boldsymbol{\theta}}_l) = -\frac{1}{n}\frac{\partial^2 l(\boldsymbol{\theta}_l)}{\partial \boldsymbol{\theta}_l \partial \boldsymbol{\theta}_l'}\Big|_{\boldsymbol{\theta}_l = \hat{\boldsymbol{\theta}}_l} = -\frac{1}{n}\frac{\partial^2 \log\left(P(\boldsymbol{Y} \mid \boldsymbol{\theta}_l, M_l)\right))}{\partial \boldsymbol{\theta}_l \partial \boldsymbol{\theta}_l'}\Big|_{\boldsymbol{\theta}_l = \hat{\boldsymbol{\theta}}_l} \tag{4.6}$$

is the Hessian matrix evaluated at the mode $\hat{\boldsymbol{\theta}}_l$.

Similarly, the prior is replaced with its Taylor expansion around the mode $\hat{\boldsymbol{\theta}}_l$,

$$P(\boldsymbol{\theta}_l \mid M_l) = P(\hat{\boldsymbol{\theta}}_l \mid M_l) + (\boldsymbol{\theta}_l - \hat{\boldsymbol{\theta}}_l)' \frac{\partial P(\boldsymbol{\theta}_l \mid M_l)}{\partial \boldsymbol{\theta}_l}\Big|_{\boldsymbol{\theta}_l = \hat{\boldsymbol{\theta}}_l} + .. \tag{4.7}$$

The marginal likelihood integral is then,

$$\begin{aligned} P(\boldsymbol{Y} \mid M_l) &= \int \exp\left(l(\hat{\boldsymbol{\theta}}_l) - \frac{n}{2}(\theta_l - \hat{\boldsymbol{\theta}}_l)' H(\hat{\boldsymbol{\theta}}_l)(\theta_l - \hat{\boldsymbol{\theta}}_l) + ..\right) \\ &\times \left(P(\hat{\boldsymbol{\theta}}_l \mid M_l) + (\boldsymbol{\theta}_l - \hat{\boldsymbol{\theta}}_l)' \frac{\partial P(\boldsymbol{\theta}_l \mid M_l)}{\partial \boldsymbol{\theta}_l}\Big|_{\boldsymbol{\theta}_l = \hat{\boldsymbol{\theta}}_l} + ..\right) d\boldsymbol{\theta}_l \end{aligned} \tag{4.8}$$

$$P(\boldsymbol{Y} \mid M_l) = \exp\Big(l(\hat{\boldsymbol{\theta}}_l)\Big) P(\hat{\boldsymbol{\theta}}_l \mid M_l) \int \exp\left(-\frac{n}{2}(\theta_l - \hat{\boldsymbol{\theta}}_l)' H(\hat{\boldsymbol{\theta}}_l)(\theta_l - \hat{\boldsymbol{\theta}}_l)\right) d\boldsymbol{\theta}_l. \tag{4.9}$$

Equation (4.9) yields therefore due to the fact that the MLE, $\hat{\boldsymbol{\theta}}_l$, converges in probability to $\boldsymbol{\theta}_l$ with order $\hat{\boldsymbol{\theta}}_l - \boldsymbol{\theta}_l = O(n^{-\frac{1}{2}})$, and therefore the following equation holds:

$$\int (\hat{\boldsymbol{\theta}}_l - \boldsymbol{\theta}_l) \exp\left(-\frac{n}{2}(\hat{\boldsymbol{\theta}}_l - \boldsymbol{\theta}_l)' H(\hat{\boldsymbol{\theta}}_l)(\hat{\boldsymbol{\theta}}_l - \boldsymbol{\theta}_l)\right) d\boldsymbol{\theta}_l = 0. \tag{4.10}$$

The integrand in (4.9), which originates from the second order term of the Taylor expansion, is a kernel of $q_l$ - dimensional Gaussian distribution with mean $\hat{\boldsymbol{\theta}}_l$ and covariance matrix $\frac{H^{-1}(\hat{\boldsymbol{\theta}}_l)}{n}$. Consequently, the integral in (4.9) can be solved analytically,

$$\int \exp\left(-\frac{n}{2}(\theta_l - \hat{\boldsymbol{\theta}}_l)' H(\hat{\boldsymbol{\theta}}_l)(\theta_l - \hat{\boldsymbol{\theta}}_l)\right) d\boldsymbol{\theta}_l = (2\pi)^{\frac{q_l}{2}} n^{-\frac{q_l}{2}} \left|H(\hat{\boldsymbol{\theta}}_l)\right|^{-\frac{1}{2}}, \tag{4.11}$$

and the solution to the marginal likelihood is,

$$P(\boldsymbol{Y} \mid M_l) \approx P(\boldsymbol{Y} \mid \hat{\boldsymbol{\theta}}_l, M_l) P(\hat{\boldsymbol{\theta}}_l \mid M_l) (2\pi)^{\frac{q_l}{2}} n^{-\frac{q_l}{2}} \left|H(\hat{\boldsymbol{\theta}}_l)\right|^{-\frac{1}{2}}. \tag{4.12}$$

To coincide with likelihood ratio tests which make model comparisons based on differences at the scale of $-2log(P(\boldsymbol{Y} \mid M_l))$, we define the analogue based on the Laplace Approximation, LA, as

$$LA \approx -2\log\left((2\pi)^{\frac{q_l}{2}} P(\boldsymbol{Y} \mid \hat{\boldsymbol{\theta}}_l, M_l) P(\hat{\boldsymbol{\theta}}_l \mid M_l) n^{-\frac{q_l}{2}} \left|H(\hat{\boldsymbol{\theta}}_l)\right|^{-\frac{1}{2}}\right), \tag{4.13}$$

such that

$$\exp(-\frac{1}{2}LA) \approx P(\boldsymbol{Y} \mid M_l). \tag{4.14}$$

The Laplace's method in (4.11) relies on asymptotic arguments but also produces a good approximation when the likelihood is approximately Gaussian with a unique mode around $\hat{\boldsymbol{\theta}}_l$.

By ignoring the terms of order $O(1)$ with respect to the sample size $n$, such as $(2\pi)^{\frac{q_l}{2}}$, $P(\hat{\boldsymbol{\theta}}_l \mid M_l)$, and $\left|H(\hat{\boldsymbol{\theta}}_l)\right|^{-\frac{1}{2}}$, the equation (4.13) reduces to the BIC (Schwarz et al., 1978),

$$BIC \approx -2\log P(\boldsymbol{Y} \mid \hat{\boldsymbol{\theta}}_l, M_l) + q_l \log n. \tag{4.15}$$

In the next section we show analytically that LA fails to approximate twice the log marginal likelihood correctly in a bimodal model.

### 4.3.1 Motivating example

When the joint density $P(\boldsymbol{\theta}_l, \boldsymbol{Y} \mid M_l)$ is unimodal and the sample size is large, Laplace approximation produces a good approximation to $-2\log P(\boldsymbol{Y} \mid M_l)$ (see Appendix F). For expositional simplicity we drop the notation of the model $M_l$ in the rest of the section, while focusing on approximating (4.2) within a single model. To motivate the need for the MBIC, we consider a bimodal distribution with well separated modes, where an analytical solution to the integral in (4.2) is available. We consider two scenarios, where bi-modality is induced by a bimodal prior or by a bimodal likelihood.

**Scenario 1.**

The likelihood is Gaussian, $P(\boldsymbol{Y} \mid \boldsymbol{\theta}) = N(\boldsymbol{Y} \mid \boldsymbol{\theta}, s^2_{\boldsymbol{Y}})$, and the prior is bimodal,

$$P(\boldsymbol{\theta}) = \frac{1}{2}N(\boldsymbol{\theta} \mid \lambda_1, s^2_{\boldsymbol{\theta}}) + \frac{1}{2}N(\boldsymbol{\theta} \mid \lambda_2, s^2_{\boldsymbol{\theta}}), \tag{4.16}$$

induces a bimodal posterior $P(\boldsymbol{\theta} \mid \boldsymbol{Y}) \propto P(\boldsymbol{Y} \mid \boldsymbol{\theta})P(\boldsymbol{\theta}) = P(\boldsymbol{\theta}, \boldsymbol{Y})$ (see Appendix G).

Since each mode is Gaussian, the Laplace approximation gives exactly the contribution of a single mode to the marginal likelihood in (4.2). However, the assumption in LA of a unimodal posterior space eliminates the entire contribution of the second mode. Hence,

rather than the result in (4.14), LA evaluated at one mode is,

$$\exp(-\frac{1}{2}LA) = \delta P(\boldsymbol{Y}), \tag{4.17}$$

where $\delta$ depends on the relative importance of the posterior mode used (see Appendix G).

Figure 4.1A visualizes the relationship between the analytical solution to the marginal likelihood $P(\boldsymbol{Y})$ and $\exp(-\frac{1}{2}LA)$ in (4.17) for a special case where the data were simulated such that $\frac{\sum_{i=1}^{n} y_i}{n} = \frac{\lambda_1+\lambda_2}{2}$, and $\lambda_1 = 8, \lambda_2 = -8$, $n = 50$, $s_{\boldsymbol{Y}} = 3.5$ and $s^2_{\boldsymbol{\theta}} = 1$ were chosen such that the posterior $P(\boldsymbol{\theta} \mid \boldsymbol{Y}) \propto P(\boldsymbol{\theta}, \boldsymbol{Y})$ satisfies the 'multi-modality condition' in Appendix G, and moreover, the posterior modes are well separated. LA values were calculated over the grid for the first prior mean $\lambda_1$, centered around the mean $\bar{y}$, and $\lambda_2$ was constrained to maintain $\frac{\sum_{i=1}^{n} y_i}{n} = \frac{\lambda_1+\lambda_2}{2}$. The orange region corresponds to the grid points where the posterior modes are well separated. In these regions of bi-modality, $\exp(-LA/2) = \frac{1}{2}P(\boldsymbol{Y})$ half of the analytical marginal likelihood (see the zoomed-in plot). In the cases where the 'multi-modality condition' in the Appendix G is not satisfied, the BIC correctly estimates the marginal likelihood.

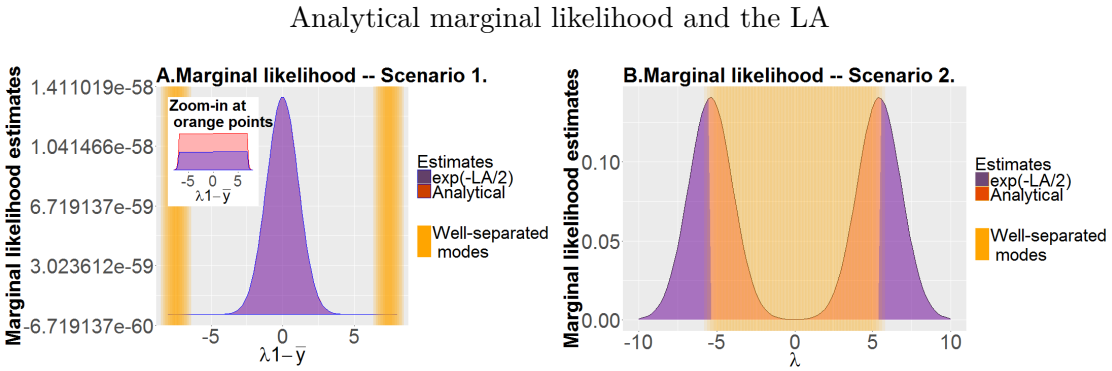


Figure 4.1: Analytical marginal likelihood and the LA: A. Model in Scenario 1 – unimodal likelihood, bimodal prior; B. Model in Scenario 2 – bimodal likelihood and unimodal prior.

**Scenario 2.**

The multi-modality in this scenario is induced by a likelihood, defined as a mixture of $K = 2$ Gaussian densities,

$$P(\boldsymbol{Y} \mid \boldsymbol{\theta}) = \sum_{k=1}^{K} p_k N(\boldsymbol{Y} \mid \boldsymbol{\theta}_k, s^2_{\boldsymbol{Y}_k}), \tag{4.18}$$

while the prior is a unimodal Gaussian, $P(\boldsymbol{\theta}) = N(\boldsymbol{\theta} \mid \lambda, s^2_{\boldsymbol{\theta}})$.

The relationship between the LA and the analytical marginal likelihood follows (4.17) (see Appendix H).

In general, the marginal likelihood for this model is intractable. Namely, in order to solve the marginal likelihood integral in (4.2), one needs to introduce a latent variable that indicates which one of the mixture components the data point $y_i$ is assigned to. Integrating out this latent variable is prohibitively expensive, since there exist $K^n$ possible ways to assign $n$ data points to $K$ mixture components.

Figure 4.1B visualizes the relationship in (4.17) for Scenario 2, over a grid for the prior mean $\lambda$ for $n = 1$ so that analytical solution to the marginal likelihood integral is available. The LA values were calculated over a grid for the prior mean $\lambda$ using only $n = 1$ data point simulated from the bimodal likelihood according to (4.18) with $K$=2, $p_1 = p_2 = \frac{1}{2}$, $\theta_2 = -\theta_1$, $\theta_1 = 6$, and $s^2_{\boldsymbol{Y}_1} = s^2_{\boldsymbol{Y}_2} = 1, s^2_{\boldsymbol{\theta}} = 1$. The one data point was simulated such that the posterior modes are well separated, and hence, the $\delta$ in the equation (4.17) is exactly $\frac{1}{2}$. The golden region represents the cases where the modes are well separated and where the $\exp(-\frac{1}{2}LA)$ fails to approximate the marginal likelihood, $P(\boldsymbol{Y})$, correctly.

## 4.4 The Multi-modal BIC

The Multi-modal BIC (MBIC) modifies the unimodal LA, and hence the BIC, with aim to tackle the multi-modality in the joint density $P(\boldsymbol{\theta}, \boldsymbol{Y})$. Following Gelman et al. (2014), the multi-modal joint density function, $P(\boldsymbol{\theta}, \boldsymbol{Y})$, with fairly widely separated modes, can be approximated with a mixture of $K$ locally weighted unimodal densities, $g(\hat{\boldsymbol{\theta}}_k, \boldsymbol{Y}) = p_k f_k(\boldsymbol{\theta}, \boldsymbol{Y})$, each centered at the $k^{th}$ mode, $\hat{\boldsymbol{\theta}}_k$, $P(\boldsymbol{\theta}, \boldsymbol{Y}) \approx \sum_{k=1}^{K} g(\hat{\boldsymbol{\theta}}_k, \boldsymbol{Y})$.

The MBIC approximates negative twice the log marginal likelihood using the unimodal densities,

$$\begin{aligned} -2\log P(\boldsymbol{Y}) &= -2\log\left(\int P(\boldsymbol{\theta}, \boldsymbol{Y})d\boldsymbol{\theta}\right) \\ &\approx -2\log\left(\sum_{k=1}^{K}\int g(\hat{\boldsymbol{\theta}}_k, \boldsymbol{Y})d\boldsymbol{\theta}\right) \\ &= -2\log\left(\sum_{k=1}^{K}\int exp(log(p_k f_k(\boldsymbol{\theta}, \boldsymbol{Y})))d\boldsymbol{\theta}\right). \end{aligned} \tag{4.19}$$

Laplace approximations described Section 4.3 are applied to each mixture component in (4.19) to produce $K$ Gaussian integrals. The $k-th$ mixture component $\log(p_k f_k(\boldsymbol{\theta}, \boldsymbol{Y}))$ is replaced with its second order Taylor expansion at $\hat{\boldsymbol{\theta}}_k$,

$$\log(p_k f_k(\boldsymbol{\theta}, \boldsymbol{Y})) = \log\left(p_k f_k(\hat{\boldsymbol{\theta}}_k, \boldsymbol{Y})\right) - \frac{n}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_k)' H(\hat{\boldsymbol{\theta}}_k)(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_k) + \cdots, \tag{4.20}$$

where

$$H(\hat{\boldsymbol{\theta}}_k) \equiv -\frac{1}{n}\frac{\partial^2 \log\left(p_k f_k(\boldsymbol{\theta}, \boldsymbol{Y})\right)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_k} \tag{4.21}$$

is the Hessian matrix at the mode around $\hat{\boldsymbol{\theta}}_k$. The second term of the Taylor expansion in (4.20) represents a $q$-dimensional Normal distribution with mean $\hat{\boldsymbol{\theta}}_k$ and variance-covariance matrix $\frac{H^{(-1)}(\hat{\boldsymbol{\theta}}_k)}{n}$.

Following the derivation of Laplace approximation described Section 4.3, Taylor expansions in (4.20) produce $K$ Gaussian integrals. Solving these Gaussian integrals results in the MBIC,

$$MBIC = -2\log\left((2\pi)^{\frac{q}{2}}\sum_{k=1}^{K} g(\hat{\boldsymbol{\theta}}_k, \boldsymbol{Y})n^{-\frac{q}{2}}\left|H(\hat{\boldsymbol{\theta}}_k)\right|^{-\frac{1}{2}}\right). \tag{4.22}$$

Since the MBIC approximates the multi-modal density with weighted unimodal distributions in (4.4), it relies on the assumption that the multiple modes are well separated. In addition, the Laplace's method approximates the integrands in (4.19) with Gaussian densities, and hence, the MBIC works well under the assumption that each locally weighted mixture component is not grossly non-Gaussian, which is the case with large sample size. When the joint density $P(\boldsymbol{\theta}, \boldsymbol{Y})$ is unimodal, i.e, $K = 1$, the MBIC reduces to the unimodal LA, and BIC (see Appendices F and G for details of derivations).

### 4.4.1 Illustration of the MBIC

Consider the two bimodal models introduced earlier. In both scenarios, the models are Gaussian with widely separated modes, and the MBIC provides an exact solution to the integral in (4.2),

$$\exp\{-\frac{1}{2}MBIC\} = P(\boldsymbol{Y}). \tag{4.23}$$

(see Appendices G and H).

Figures 4.2A and 4.2B graphically illustrate the relationship in (4.23) for models in Scenario 1 and Scenario 2, respectively. Analytical marginal likelihood and the MBIC were calculated under same special conditions as the plots in Figure 4.1.

## 4.5 Model selection in mixture of Gaussian models

We consider the Galaxy data set, which provides information on velocities of 82 galaxies that diverge from our galaxy studied by Postman et al. (1986); Carlin and Chib (1995); Neal (1999). The data, denoted as $\mathbf{Y} = (y_1, y_2, .., y_n)'$ for n=82, are univariate identically

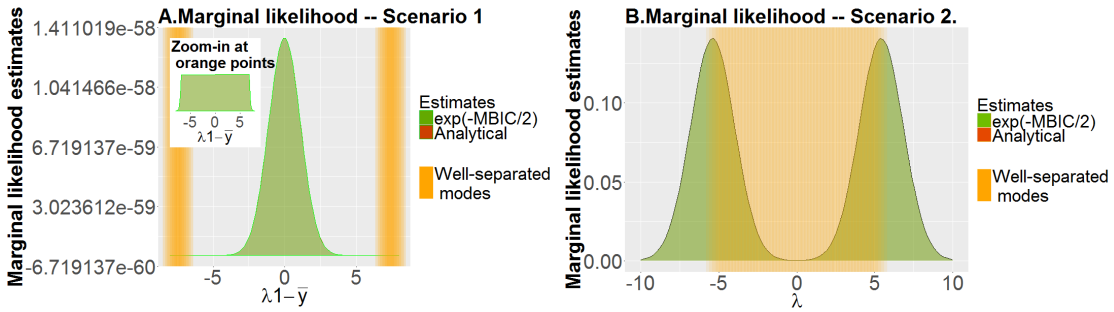


Figure 4.2: Analytical marginal likelihood and the MBIC: A. Model in Scenario 1 – unimodal likelihood, bi-modal prior; B. Model in Scenario 2 – bi-modal likelihood and unimodal prior.

and independently distributed samples from a mixture of $K$ Gaussian components. The data are modeled as,

$$P(\mathbf{Y} \mid \boldsymbol{\mu}, \boldsymbol{\sigma^2}, \boldsymbol{p}) = \prod_{i=1}^{n} \sum_{k=1}^{K} p_k N(y_i \mid \mu_k, \sigma_k^2), \tag{4.24}$$

with parameters $\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\sigma^2}, \boldsymbol{p})'$ where $\boldsymbol{\mu} = (\mu_1, .., \mu_K)'$ is a vector of mixture component means, $\boldsymbol{\sigma^2} = (\sigma_1^2, .., \sigma_K^2)'$ is a vector of mixture component variances and $\boldsymbol{p} = (p_1, .., p_K)'$ is a vector of mixture probabilities. Conjugate priors were assigned, $P(\mu_k) \sim \text{Normal}(20, 100)$, $P(\sigma_k^2) \sim \text{InverseGamma}(3, scale = \frac{1}{20})$ and $P(p_1, .., p_K) \sim \text{Dirichlet}(\alpha_1 = 1, .., \alpha_K = 1)$.

The number of components $K$ is unknown, and so we perform model selection using following models:

- $2Ecomp$ – the model with 2 components and equal variances,
- $3Ecomp$ – the model with 3 components and equal variances,
- $3comp$ – the model with 3 components and unequal variances,
- $4comp$ – the model with 4 components and unequal variances and
- $5comp$ – the model with 5 components and unequal variances.

### 4.5.1 The label switching problem

Parameter estimation and model selection in mixture models induces a label switching problem, which arises due to the invariance of the likelihood when relabeling the mixture components (Redner and Walker, 1984; Richardson and Green, 1997; Stephens, 2000; Jasra et al., 2005; Diebolt and Robert, 1994; Stephens, 1997). This invariance implies that the

likelihood in (4.24) is the same for all permutations of $\boldsymbol{\theta}$, thus producing highly multi-modal and symmetric likelihoods. Hence, for each mode in the mixture Gaussian model there are $K!$ corresponding symmetric or label switching modes.

### 4.5.2 Model selection performance of the MBIC, the LA and the BIC

In this section we present and discuss results from the model selection performance of the following methods: the BIC, the LA and the MBIC using the five models introduced earlier. The model with the lowest BIC, LA or MBIC was considered as a selected model. For comparison, the estimates obtained from the following thermodynamic integration methods: *i*). thermodynamic integration via Parallel Tempering and Simulated Tempering Without Normalizing Constants (PT-STWNC) introduced in Chapter 2, *ii*). thermodynamic integration via Parallel Tempering with bias correction (TI-PT-B) by Calderhead and Girolami (2009) and *iii*). thermodynamic integration via Parallel Tempering without bias correction (TI-PT-NB) by Friel and Pettitt (2008), are included.

To discover all the important posterior modes, we developed an optimization strategy which we refer to as Shotgun optimization (ShOpt). The ShOpt algorithm, which is tailored for this specific problem, discovers all the important modes while taking into account the $K!$ number of label switching modes for each of the studied models.

The ShOpt algorithm starts with drawing $N_0$ samples from the prior distribution, $P(\boldsymbol{\theta})$, followed by calculating the weights for each sample using the likelihood function. The highest weights points are used to initialize the optimizer and find the local optima. For each newly discovered mode, the corresponding $K!$ label switching modes are added to the set of newly discovered modes. At each iteration, after a mode is found, samples from the prior that are within the basin of attraction of the newly discovered mode are discarded, thus enabling the algorithm to move to the unexplored regions of the parameter space. Mahalanobis distance between the newly discovered modes and the previously discovered modes is used to make a decision whether or not to accept the newly discovered mode. The algorithm is described in details in the Appendix I.

Table 4.1 demonstrates that in the Galaxy data example, each of the studied models exhibits a highly multi-modal and symmetric posterior space as a result of the label switching problem. The difference in number of modes with respect to the Maximum A Posteriori (MAP) and Maximum Likelihood Estimate (MLE) estimates observed in the Table 4.1 can be explained as follows. The ShOpt algorithm employs a conservative measure (Mahalanobis distance) to decide whether or not to accept a newly discovered candidate mode as distinct from the others. In addition, the BIC is derived from LA under the assumption that the sample size is large, in which case the likelihood overwhelms the prior and the MAP estimate is almost equal to the MLE. Hence, insufficiently large data set can lead to the differences in number of modes with respect to the MLE and MAP in Table 4.1.

Table 4.1: Number of modes discovered by ShOpt.

| Model fitted | $K = 2$ components, equal variances | $K = 3$ components, equal variances | $K = 3$ components | $K = 4$ components | $K = 5$ components |
|---|---|---|---|---|---|
| MAP | $8 \times 2!$ | $5 \times 3!$ | $9 \times 3!$ | $20 \times 4!$ | $18 \times 5!$ |
| MLE | $37 \times 2!$ | $68 \times 3!$ | $338 \times 3!$ | $741 \times 4!$ | $1008 \times 5!$ |

Galaxy data – number of modes for each of the five models, discovered by the ShOpt algorithm. The first and the second line correspond to the number of modes according to the MAP and MLE, respectively. The MLE estimates were used to obtain the BIC estimates, while the MAP estimates were used to obtain the MBIC and the LA estimates. The ShOpt algorithm was run with $N_0 = 10000$ samples from the prior.

We compare the models using log marginal likelihood of a given model for a given estimate (such as MBIC, LA, BIC, PT-STWNC, TI-PT-B and TI-PT-NB). The estimates from the PT-STWNC, the TI-PT-B and TI-PT-NB are given in Table 2.3. Following the Bayesian interpretation of the MBIC, LA and BIC as minus twice the log marginal likelihood by Tierney and Kadane (1986) (equation in (4.19)), the MBIC, LA and BIC estimates were transformed to log marginal likelihood estimates.

According to the MBIC, the best model is the one with 5 components, and the worst is the model with 2 components equal variances (Figure 4.3). Log marginal likelihood estimates of the rest of the models with respect to the MBIC are approximately the same, which indicates that any of these models could be a good fit to the data. These findings are consistent with studies by Liang and Wong (2001); Chib (1995); Neal (1999). Liang and Wong (2001) used bridge sampling introduced by Meng and Wong (1996) on Evolutionary Monte Carlo Liang and Wong (2001) outputs, to obtain probabilities of the models with number of components 2-5. The authors findings suggest that the model with 5 components is the best, while the model with 2 components is the worst. Richardson and Green (1997) found support for the models with number of components between 5 and 7, using full Bayesian approach via Reversible Jump MCMC (RJMCMC) Green (1995). Steele and Raftery (2010) argued that the correct model for the Galaxy data is unknown, however, they concluded that the models with 3 and 6 components are reasonable fit to the data.

The LA and the BIC evaluated at the global modes behave similarly to the MBIC: they find that the models with 5 and 3 components are the best, respectively, while the model with 2 components is the worst (Figure 4.3).

Figure 4.3 demonstrates that all of the thermodynamic estimators, the PT-STWNC, the TI-PT-B and the TI-PT-NB, agree that the best model is the one with 5 components and the worst model is the model with 2 components. In addition, all the three estimators find

support for the models with 3-5 components, and the log marginal likelihoods of models 3-5 indicate these models are well separated from the worst model (Figure 4.3).

Galaxy data: model selection

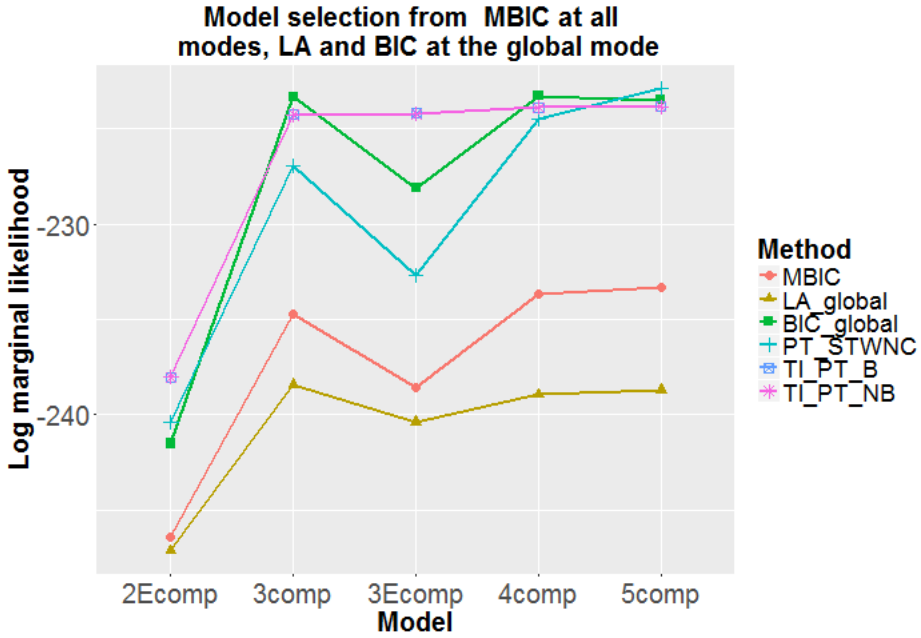


Figure 4.3: Log marginal likelihood of the models according to the MBIC, the BIC at the global mode, the LA at the global mode, the PT-STWNC, the TI-PT-B and the TI-PT-NB. Studied models are given on the x-axis, log marginal likelihoods of the models according to the six studied model selection methods are given on the y axis. The bigger log marginal likelihood the better the model.

In practice, model selection via the BIC is performed by evaluating the BIC at only one discovered mode. The most commonly used optimizers are searching for the local modes, and hence, without a full exploration of the posterior space, there is no guarantee that the global mode has been found. Consequently, depending on the relative importance of the discovered mode, the BIC might select the wrong model.

In the Galaxy data example, the BIC and the LA were evaluated at each of the discovered modes as per Table 4.1. Pairwise comparisons of all the studied models were performed using all possible pairs of modes per pairs of models. The results were summarized by calculating proportions of the modes at which the model $M_i$ was chosen over the model $M_j$, $p_{M_i,M_j}$, for $i, j \in \{1, .., 5\}$. The proportion of modes at which the LA chooses the model $M_i$ over the model $M_j$ is,

$$p_{M_i,M_j} = \sum_{k=1}^{N_{M_j}} \frac{1}{N_{M_j}} \frac{\sum_{m=1}^{N_{M_i}} \mathbb{I}_{LA(M_i(m))<LA(M_j(k))}}{N_{M_i}}, \tag{4.25}$$

where $N_{M_i}$ and $N_{M_j}$ are the total number of modes in the models $M_i$ and $M_j$, respectively. Hence, to obtain the proportions at which the LA chooses the model $M_i$ over the model $M_j$, the modes were averaged under the assumption that each mode is equally likely to be discovered. $\sum_{m=1}^{N_{M_i}} \mathbb{I}_{LA(M_i(m))<LA(M_j(k))}$ is the number of modes in the model $M_i$, where the LA of the model $M_i$ is lower than that of the model $M_2$ at mode $k$. Analogously, the proportion of modes at which the BIC selects the model $M_i$ over the model $M_j$ were obtained as per (4.25) by replacing the LA with the BIC.

Pairwise comparisons between each of the studied models suggest that the LA at any mode selects the model with 3 components (Figure 4.4, plot A), whereas the BIC at any mode chooses the model with 5 components (Figure 4.4, plot B). Although the choices of the best models are in-line with the previous studies, the proportions of modes at which the LA and the BIC at any local mode select a sub-optimal model model are large. For instance, the proportions of modes at which the LA selects the model with 2 components over the model with 3 components equal variances is 0.7778. Similarly, the BIC chooses the model with 2 components equal variances over the model with 3 components equal variances in 35.7% of the modes.

The model selection study in this section reveals that the LA and the BIC model selection performance is reasonably well when the information on the global mode is used. However, the LA and the BIC fail to select the correct model when evaluated at local mode which has significantly smaller probability mass then that of the other posterior modes. These findings demonstrate clearly that there is a need for a multi-modal approach that incorporates information about all the relevant posterior modes.

## 4.6 Discussion

In this chapter, we proposed a Multi-modal Bayesian Information Criterion (MBIC), which is a generalization of the Bayesian Information Criterion (BIC) that addresses multi-modality in the posterior parameter space. While BIC approximates the posterior probability of the model using only one mode, the approximation from the MBIC employs information from all relevant posterior modes. The MBIC first approximates the multi-modal un-normalized posterior density with a mixture of weighted unimodal densities, and then applies the Laplace's method locally to each unimodal density to integrate over model parameters. Consequently, the MBIC correctly approximates the posterior probability of the model under the following two assumptions: *i*). the posterior modes are fairly widely separated, and *ii*). the posterior modes are not grossly non-Gaussian, which is the case with large sample size.

The analytical derivations in examples in Scenario 1 and Scenario 2, indicate that the unimodal Laplace approximation (LA) underestimates, and the MBIC correctly estimates

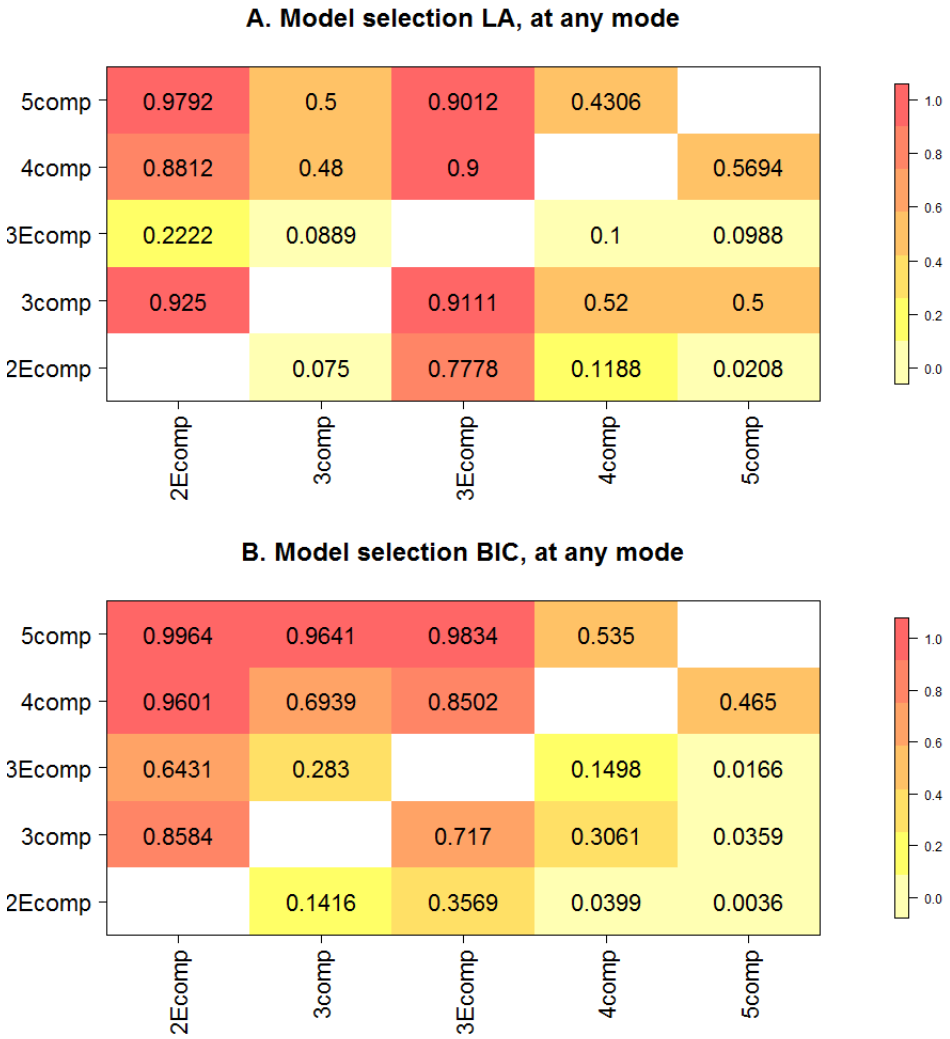


Figure 4.4: Galaxy data – Pairwise model comparisons using: A. the LA evaluated at each of the discovered modes; B. the BIC evaluated at each of the discovered modes. Each cell of the heat-maps correspond to the proportion of modes, $p_{M_i,M_j}$, (obtained as per (4.25)) at which the model $M_i$ from the y-axis was selected over the model $M_j$ from the x-axis, for $i, j \in \{1, .., 5\}$.

the posterior probability of the model in the cases where the two posterior modes are well separated.

The model selection study in Section 4.5 demonstrates that the MBIC correctly determines the best and the worst model. Log marginal likelihoods for the rest of the models with respect to the MBIC are approximately the same, which indicates that any of these models could be a good fit to the data. This finding conforms with the results from the previous studies. While the LA and the BIC evaluated at the global mode correctly choose

the best model, the problem of choosing the wrong model arises when the LA and the BIC are evaluated at local modes. The last result is a direct consequence of the BIC and the LA relying on only one mode, instead of employing information from all relevant posterior modes.

The MBIC combines the computational efficiency of the BIC with the advantage of combining all the information from fully explored posterior space. Hence, the MBIC is a useful model selection tool in high-dimensional problems that exhibit inherent multi-modality, and where Monte Carlo integration over the parameter space is prohibitively expensive.

In practice, where posterior topologies are high-dimensional and calculation of the Hessian is prohibitively computationally expensive, the MBIC could be applied with some modifications. The proposed algorithm for discovering all important modes, the Shotgun optimization (ShOpt), could be modified to select the most important posterior modes based on their relative importance heights. The relative importance height $h$ of the mode, $\hat{\boldsymbol{\theta}}_k$, can be obtained as a proportion of the probability mass of the mode with respect to the total probability mass of the all the posterior modes, i.e., $h_{\hat{\boldsymbol{\theta}}_k} = \frac{g(\hat{\boldsymbol{\theta}}_k|\boldsymbol{Y})}{\sum_{j=1}^{K} g(\hat{\boldsymbol{\theta}}_j|\boldsymbol{Y})}$. Modifications of the MBIC and the ShOpt to accommodate high-dimensional problems where Hessian matrix is not available are left for future research.

# Chapter 5

# Conclusion

In this thesis we have developed three efficient and effective computational methods for parameter estimation and model selection in complex posterior topologies: Parallel Tempering via Simulated Tempering Without Normalizing Constants (PT-STWNC), Incremental Mixture Importance Sampling with Shotgun optimization (IMIS-ShOpt) and Multi-modal Bayesian Information Criterion (MBIC). The Simulated Tempering Without Normalizing Constants (STWNC) builds on the standard Simulated Tempering (ST), which is an efficient tempering algorithm designed to handle the multi-modality in the posterior space, but it is unpopular in practice because of the requirement for the normalizing constants and temperature schedule. The PT-STWNC removes this requirement by introducing a continuous temperature parameter, which not only eliminates the requirement for normalizing constants and temperature schedule, but also enables calculation of the probability of the model at a negligible additional computational cost. Chapter 2 demonstrates that the PT-STWNC is an efficient and user-friendly computational tool that concurrently estimates the parameters of interest and the probability of the model.

The Shotgun optimization is a general optimization methodology that combines results from different optimization methods to ensure that all the relevant posterior modes have been discovered. Replacing the optimization stage in the Incremental Mixture of Importance sampling with our proposed Shotgun optimization strategy results in an efficient algorithm, named Incremental Mixture Importance Sampling with Shotgun optimization. The IMIS-ShOpt algorithm addresses the problem of sampling from posterior distribution characterized with rife surfaces, many unimportant modes, riffles and ridges. Furthermore, while the IMIS-Opt requires the prior to agree with the data, the success of the IMIS-ShOpt depends on the Shotgun optimization finding all the important posterior modes, thus allowing the prior to reflect the expert knowledge, rather than being chosen for algorithmic convenience. The performance of the IMIS-ShOpt is demonstrated by two ordinary differential equation examples. Comparisons between performance of the IMIS-ShOpt and the IMIS-Opt in complex topologies show that the IMIS-Opt gets trapped in a single mode,

while the IMIS-ShOpt explores the full posterior surface. The IMIS-ShOpt with synthetic likelihood via Approximate Bayesian Computation framework (IMIS-ShOpt) is illustrated as an efficient and effective likelihood-free approach for sampling from posterior distribution.

We have also proposed a generalization of the Bayesian Information Criterion (BIC), the Multi-modal BIC (MBIC), which handles model selection in multi-modal posterior spaces. Although the BIC is a computationally efficient model selection tool, it employs information from one posterior mode only, while discarding the information from the rest of the modes. Our analytical derivations and model selection study demonstrated that the BIC produces biased estimates of the probability of the model, which might lead to choosing the wrong model. The MBIC improves the estimates from the BIC by taking into account all the important posterior modes. Moreover, we developed an algorithm that helps discovering all the important modes needed for calculating MBIC. Through the analytical calculations and model selection study we demonstrated that the MBIC handles the multi-modality in the posterior spaces, while preserving the computational efficiency of the BIC.

# Bibliography


Ahlers, H. and Engel, A. (2008). Prior-predictive value from fast growth simulations. *The European Physical Journal B*, 62(3):357–364.

Alkema, L., Raftery, A. E., and Clark, S. J. (2007). Probabilistic projections of hiv prevalence using bayesian melding. *The Annals of Applied Statistics*, pages 229–248.

Alkema, L., Raftery, A. E., Gerland, P., Clark, S. J., Pelletier, F., Buettner, T., and Heilig, G. K. (2011). Probabilistic projections of the total fertility rate for all countries. *Demography*, 48(3):815–839.

Alrefaei, M. H. and Diabat, A. H. (2009). A simulated annealing technique for multi-objective simulation optimization. *Applied mathematics and computation*, 215(8):3029–3035.

Atchade, Y. F. and Liu, J. S. (2004). The wang-landau algorithm for monte carlo computation in general state spaces. *Statistica Sinica*, 20:209–33.

Atchadé, Y. F., Rosenthal, J. S., et al. (2005). On adaptive markov chain monte carlo algorithms. *Bernoulli*, 11(5):815–828.

Bäck, T., Hammel, U., and Schwefel, H.-P. (1997). Evolutionary computation: Comments on the history and current state. *Evolutionary computation, IEEE Transactions on*, 1(1):3–17.

Bates, D. M. and Watts, D. G. (1988). *Nonlinear regression analysis and its applications*. Wiley.

Berger, J. O., Liseo, B., Wolpert, R. L., et al. (1999). Integrated likelihood methods for eliminating nuisance parameters. *Statistical Science*, 14(1):1–28.

Bornn, L. (2014). Pawl-forced simulated tempering. In *The Contribution of Young Researchers to Bayesian Statistics*, pages 61–65. Springer.

Bornn, L., Jacob, P. E., Del Moral, P., and Doucet, A. (2013). An adaptive interacting wang–landau algorithm for automatic density exploration. *Journal of Computational and Graphical Statistics*, 22(3):749–773.

Brunel, N. J. et al. (2008). Parameter estimation of ode's via nonparametric estimators. *Electronic Journal of Statistics*, 2:1242–1267.

Calderhead, B. and Girolami, M. (2009). Estimating bayes factors via thermodynamic integration and population mcmc. *Computational Statistics & Data Analysis*, 53(12):4028–4045.

Campbell, D. and Lele, S. (2014). An anova test for parameter estimability using data cloning with application to statistical inference for dynamic systems. *Computational Statistics & Data Analysis*, 70:257–267.

Campbell, D. and Steele, R. (2012). Smooth functional tempering for nonlinear differential equation models. *Statistics and Computing*, 22(2):429–443.

Carlin, B. P. and Chib, S. (1995). Bayesian model choice via markov chain monte carlo methods. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(3):pp. 473–484.

Carroll, R. J., Ruppert, D., Stefanski, L. A., and Crainiceanu, C. M. (2006). *Measurement error in nonlinear models: a modern perspective*. CRC press.

Cheng, R. and Gen, M. (1997). Genetic algorithms and engineering design. *New York*.

Chib, S. (1995). Marginal likelihood from the gibbs output. *Journal of the American Statistical Association*, 90(432):1313–1321.

Chipman, H., George, E. I., McCulloch, R. E., Clyde, M., Foster, D. P., and Stine, R. A. (2001). The practical implementation of bayesian model selection. *Lecture Notes-Monograph Series*, pages 65–134.

Cohon, J. L. and Marks, D. H. (1973). Multiobjective screening models and water resource investment. *Water Resources Research*, 9(4):826–836.

Craiu, R. V. and Rosenthal, J. S. (2014). Bayesian computation via markov chain monte carlo. *Annual Review of Statistics and Its Application*, 1(1):179–201.

Deb, K. (2001). *Multi-objective optimization using evolutionary algorithms*, volume 16. John Wiley & Sons.

Del Moral, P., Doucet, A., and Jasra, A. (2006). Sequential monte carlo samplers. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(3):411–436.

Del Moral, P., Doucet, A., and Jasra, A. (2012). An adaptive sequential monte carlo method for approximate bayesian computation. *Statistics and Computing*, 22(5):1009–1020.

Diebolt, J. and Robert, C. P. (1994). Estimation of finite mixture distributions through bayesian sampling. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 363–375.

Ding, A. A. and Wu, H. (2014). Estimation of ordinary differential equation parameters using constrained local polynomial regression. *Statistica Sinica*, 24(4):1613.

Fan, J. and Gijbels, I. (1996). *Local polynomial modelling and its applications: monographs on statistics and applied probability 66*, volume 66. CRC Press.

Fearnhead, P. and Prangle, D. (2010). Constructing summary statistics for approximate bayesian computation: Semi-automatic abc. *arXiv preprint arXiv:1004.1112*.

FitzHugh, R. (1961). Impulses and physiological states in theoretical models of nerve membrane. *Biophysical journal*, 1(6):445.

Fonseca, C. M. and Fleming, P. J. (1995). Multiobjective genetic algorithms made easy: selection sharing and mating restriction. In *Genetic Algorithms in Engineering Systems: Innovations and Applications, 1995. GALESIA. First International Conference on (Conf. Publ. No. 414)*, pages 45–52. IET.

Foster, D. P. and George, E. I. (1994). The risk inflation criterion for multiple regression. *The Annals of Statistics*, pages 1947–1975.

Friedman, J., Hastie, T., and Tibshirani, R. (2001). *The elements of statistical learning*, volume 1. Springer series in statistics Springer, Berlin.

Friel, N. and Pettitt, A. N. (2008). Marginal likelihood estimation via power posteriors. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(3):589–607.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2014). *Bayesian data analysis, Chapter 13*, volume 2. Taylor & Francis.

Gelman, A. and Meng, X.-L. (1998). Simulating normalizing constants: from importance sampling to bridge sampling to path sampling. *Statistical Science*, 13(2):163–185.

Gelman, A., Roberts, G., and Gilks, W. (1996). Efficient metropolis jumping rules. *Bayesian statistics*, 5(599-608):42.

Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical science*, pages 457–472.

Geyer, C.J. (1991). Markov chain monte carlo maximum likelihood. *Computing Science and Statistics*, page 156–163.

Geyer, C. J. and Thompson, E. A. (1995). Annealing markov chain monte carlo with applications to ancestral inference. *Journal of the American Statistical Association*, 90(431):909–920.

Gilpin, M. E. and Ayala, F. J. (1973). Global models of growth and competition. *Proceedings of the National Academy of Sciences*, 70(12):3590–3593.

Golchi, S. and Campbell, D. A. (2016). Sequentially constrained monte carlo. *Computational Statistics & Data Analysis*, 97:98–113.

Green, P. J. (1995). Reversible jump markov chain monte carlo computation and bayesian model determination. *Biometrika*, 82(4):711–732.

Hesterberg, T. (1995). Weighted average importance sampling and defensive mixture distributions. *Technometrics*, 37(2):185–194.

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999). Bayesian model averaging: a tutorial. *Statistical science*, pages 382–401.

Hukushima, K. and Nemoto, K. (1996). Exchange monte carlo method and application to spin glass simulations. *Journal of the Physical Society of Japan*, 65(6):1604–1608.

Jasra, A., Holmes, C., and Stephens, D. (2005). Markov chain monte carlo methods and the label switching problem in bayesian mixture modeling. *Statistical Science*, pages 50–67.

Joyce, P. and Marjoram, P. (2008). Approximately sufficient statistics and bayesian computation. *Statistical applications in genetics and molecular biology*, 7(1).

Kadane, J. B. and Lazar, N. A. (2004). Methods and criteria for model selection. *Journal of the American statistical Association*, 99(465):279–290.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the american statistical association*, 90(430):773–795.

Kennard, R. W. (1971). A note on the cp statistic. *Technometrics*, 13(4):899–900.

Konishi, S. and Kitagawa, G. (2008). *Information criteria and statistical modeling, Chapter 9*. Springer Science & Business Media.

Kou, S. C., Zhou, Q., and Wong, W. H. (2006). Discussion paper equi-energy sampler with applications in statistical inference and statistical mechanics. *The Annals of Statistics*, 34(4):pp. 1581–1619.

Kuhn, H. and Tucker, A. (1951). Proceedings of 2nd berkeley symposium.

Levary, R. and Avery, M. (1984). On the practical application of weighting equities in a portfolio via goal programming. *Opsearch*, 21:246–261.

Liang, F. and Wong, W. H. (2001). Real-parameter evolutionary monte carlo with applications to bayesian mixture models. *Journal of the American Statistical Association*, 96(454):653–666.

Liang, H. and Wu, H. (2008). Parameter estimation for differential equation models using a framework of measurement error in regression models. *Journal of the American Statistical Association*, 103(484):1570–1583.

Madigan, D. and Raftery, A. E. (1994). Model selection and accounting for model uncertainty in graphical models using occam's window. *Journal of the American Statistical Association*, 89(428):1535–1546.

Marinari, E. and Parisi, G. (1992). Simulated tempering: A new monte carlo scheme.

Massad, E., Coutinho, F., Burattini, M., and Lopez, L. (2004). The eyam plague revisited: did the village isolation change transmission from fleas to pulmonary? *Medical hypotheses*, 63(5):911–915.

Mendes-Moreira, J., Soares, C., Jorge, A. M., and Sousa, J. F. D. (2012). Ensemble approaches for regression: A survey. *ACM Computing Surveys (CSUR)*, 45(1):10.

Meng, X.-L. and Wong, W. H. (1996). Simulating ratios of normalizing constants via a simple identity: a theoretical exploration. *Statistica Sinica*, pages 831–860.

Miettinen, K. (2012). *Nonlinear multiobjective optimization*, volume 12. Springer Science & Business Media.

Montgomery, J. M., Hollenbach, F. M., and Ward, M. D. (2012). Improving predictions using ensemble bayesian model averaging. *Political Analysis*, 20(3):271–291.

Nagumo, J., Arimoto, S., and Yoshizawa, S. (1962). An active pulse transmission line simulating nerve axon. *Proceedings of the IRE*, 50(10):2061–2070.

Neal, R. M. (1996). Sampling from multimodal distributions using tempered transitions. *Statistics and computing*, 6(4):353–366.

Neal, R. M. (1999). Erroneous results in marginal likelihood from the gibbs output. *minmeo, University of Toronto*.

Newton, M. A. and Raftery, A. E. (1994). Approximate bayesian inference with the weighted likelihood bootstrap. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 3–48.

Neyman, J. and Pearson, E. S. (1992). On the problem of the most efficient tests of statistical hypotheses. In *Breakthroughs in Statistics*, pages 73–108. Springer.

Peters, G. W., Fan, Y., and Sisson, S. A. (2012). On sequential monte carlo, partial rejection control and approximate bayesian computation. *Statistics and Computing*, 22(6):1209–1222.

Poole, D. and Raftery, A. E. (2000). Inference for deterministic simulation models: the bayesian melding approach. *Journal of the American Statistical Association*, 95(452):1244–1255.

Postman, M., Huchra, J., and Geller, M. (1986). Probes of large-scale structure in the corona borealis region. *The Astronomical Journal*, 92:1238–1247.

Prangle, D., Fearnhead, P., Cox, M. P., Biggs, P. J., and French, N. P. (2014). Semi-automatic selection of summary statistics for abc model choice. *Statistical applications in genetics and molecular biology*, 13(1):67–82.

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

Raftery, A. E. and Bao, L. (2010). Estimating and projecting trends in hiv/aids generalized epidemics using incremental mixture importance sampling. *Biometrics*, 66(4):1162–1173.

Raftery, A. E., Newton, M. A., Satagopan, J. M., and Krivitsky, P. N. (2006). Estimating the integrated likelihood via posterior simulation using the harmonic mean identity.

Ramsay, J. O., Hooker, G., Campbell, D., and Cao, J. (2007). Parameter estimation for differential equations: a generalized smoothing approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(5):741–796.

Redner, R. A. and Walker, H. F. (1984). Mixture densities, maximum likelihood and the em algorithm. *SIAM review*, 26(2):195–239.

Reschenhofer, E. (2001). The bimodality principle. *J Stat Educ*, 9(1):1–16.

Richardson, S. and Green, P. J. (1997). On bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society: series B (statistical methodology)*, 59(4):731–792.

Rubin, D. B. (1987). The calculation of posterior distributions by data augmentation: Comment: A noniterative sampling/importance resampling alternative to the data augmentation algorithm for creating a few imputations when fractions of missing information are modest: The sir algorithm. *Journal of the American Statistical Association*, 82(398):543–546.

Rubin, D. B. et al. (1988). Using the sir algorithm to simulate posterior distributions. *Bayesian statistics*, 3(1):395–402.

Schwarz, G. et al. (1978). Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464.

Seber, G. (1989). F, and wild, cj (1989), nonlinear regression.

Serafini, P. (1994). Simulated annealing for multi objective optimization problems. In *Multiple criteria decision making*, pages 283–292. Springer.

Smith, K. I., Everson, R. M., and Fieldsend, J. E. (2004). Dominance measures for multi-objective simulated annealing. In *Evolutionary Computation, 2004. CEC2004. Congress on*, volume 1, pages 23–30. IEEE.

Steele, R. J. and Raftery, A. E. (2010). Performance of bayesian model selection criteria for gaussian mixture models. *Frontiers of statistical decision making and bayesian analysis*, pages 113–130.

Steele, R. J., Raftery, A. E., and Emond, M. J. (2006). Computing normalizing constants for finite mixture models via incremental mixture importance sampling (imis). *Journal of Computational and Graphical Statistics*, 15(3):712–734.

Stephens, M. (1997). Bayesian methods for mixtures of normal distributions.

Stephens, M. (2000). Dealing with label switching in mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(4):795–809.

Swendsen, R. H. and Wang, J.-S. (1986). Replica monte carlo simulation of spin-glasses. *Physical Review Letters*, 57(21):2607.

Tavaré, S., Balding, D. J., Griffiths, R. C., and Donnelly, P. (1997). Inferring coalescence times from dna sequence data. *Genetics*, 145(2):505–518.

Tibshirani, R. and Knight, K. (1999). The covariance inflation criterion for adaptive model selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3):529–546.

Tierney, L. and Kadane, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the american statistical association*, 81(393):82–86.

Tseng, C. and Lu, T. (1990). Minimax multiobjective optimization in structural design. *International Journal for Numerical Methods in Engineering*, 30(6):1213–1228.

Valenzuela-Rendón, M., Uresti-Charre, E., and Monterrey, I. (1997). A non-generational genetic algorithm for multiobjective optimization. In *in Proceedings of the Seventh International Conference on Genetic Algorithms*. Citeseer.

Varah, J. (1982). A spline least squares method for numerical parameter estimation in differential equations. *SIAM Journal on Scientific and Statistical Computing*, 3(1):28–46.

Walley, P. and Moral, S. (1999). Upper probabilities based only on the likelihood function. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, pages 831–847.

Wang, F. and Landau, D. (2001). Efficient, multiple-range random walk algorithm to calculate the density of states. *Physical review letters*, 86(10):2050.

Wang, L., Cao, J., Ramsay, J., Burger, D., Laporte, C., and Rockstroh, J. (2014). Estimating mixed-effects differential equation models. *Statistics and Computing*, 24(1):111–121.

Wasan, M. (1969). Stochastic approximation. *Cambridge tracts in mathematics and mathematical physics (*, (58).

Wilks, S. S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics*, 9(1):60–62.

Woldeyes, R. A., Sivak, D. A., and Fraser, J. S. (2014). E pluribus unum, no more: from one crystal, many conformations. *Current Opinion in Structural Biology*, 28:56 – 62.

Wolpert, D. H. and Macready, W. G. (1997). No free lunch theorems for optimization. *Evolutionary Computation, IEEE Transactions on*, 1(1):67–82.

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

Wu, S., Liu, Z.-P., Qiu, X., and Wu, H. (2014). Modeling genome-wide dynamic regulatory network in mouse lungs with influenza infection using high-dimensional ordinary differential equations. *PloS one*, 9(5):e95276.

Zhang, C. and Ma, J. (2008). Comparison of sampling efficiency between simulated tempering and replica exchange. *The Journal of chemical physics*, 129(13):134112.

Zitzler, E. (1999). *Evolutionary algorithms for multiobjective optimization: Methods and applications*, volume 63. Citeseer.

Xie, Wangang and Lewis, Paul O and Fan, Yu and Kuo, Lynn and Chen, Ming-Hui (2011). Improving marginal likelihood estimation for Bayesian phylogenetic model selection. *Oxford University Press*, 60(2):150-160.

# Appendix A

# Analytical calculation of the marginal likelihood

In this section we provide the details on analytic calculation of the marginal likelihood in (2.17) for the bimodal model example given in the Section (2.7).

The posterior distribution of the unknown parameter $\mu$ is,

$$\begin{aligned} P(\mu \mid \sigma^2, \boldsymbol{Y}) &\propto P(\boldsymbol{Y} \mid |\mu|, \sigma^2) P(\mu) = \prod_{i=1}^{n} N(y_i \mid |\mu|, \sigma^2) P(\mu) \\ &= \left( \prod_{i=1}^{n} N(y_i \mid \mu, \sigma^2) \mathbb{I}(\mu > 0) + \prod_{i=1}^{n} N(-y_i \mid \mu, \sigma^2) \mathbb{I}(\mu < 0) \right) P(\mu) \end{aligned} \tag{A.1}$$

where the variance was fixed at $\sigma^2 = 1$. The prior of $\mu$ was Gaussian: $P(\mu) \sim N(\lambda = 0, \beta = 1)$.

The Thermodynamic Integral is:

$$\begin{aligned} \int_{\mu} P(\mu \mid \sigma^2, \boldsymbol{Y}) d\mu &= (2\pi)^{-\frac{n}{2}} (\sigma^2)^{-\frac{n}{2}} \beta^{-1} \frac{1}{2} \sqrt{a^{-1}} \\ &\times \left[ \exp\left( \frac{1}{2} \frac{b^2}{a} - \frac{1}{2} c \right) + \exp\left( \frac{1}{2} \frac{b_1^2}{a} - \frac{1}{2} c \right) \right], \end{aligned} \tag{A.2}$$

where $a = \frac{n}{\sigma^2} + \frac{1}{\beta}$, $b = \frac{\sum_i y_i}{\sigma^2} + \frac{\lambda}{\beta}$, $b_1 = \frac{-\sum_i y_i}{\sigma^2} + \frac{\lambda}{\beta}$ and $c = \frac{\sum_i y_i^2}{\sigma^2} + \frac{\lambda^2}{\beta}$.

Plug in the values of $\{\boldsymbol{Y}, \lambda, \beta, \sigma^2, n\}$ in the solution of the integral given by the equation (A.2) and take a log to obtain the analytical marginal likelihood reported in Table 2.2.

# Appendix B

# Implementation, bimodal model

The transition kernel of $\mu$ was updated using the optimal symmetric jumping kernel for Gaussian target distributions by Gelman et al. (1996). The proposal distribution for of $\mu$ at the the $i$-th iteration is:

$$\mu^{(i+1)} \sim N\left(\mu^{(i)}, [2.4/\sqrt{d}]^2 \mathrm{Var}_{P(\mu|\sigma^2,\mathbf{Y},\tau)}(\mu)\right) \tag{B.1}$$

where $\mathrm{Var}_{P(\mu|\sigma^2,\mathbf{Y},\tau)}(\mu)$ is the target variance of $\mu$ with respect to the target posterior distribution $P(\mu \mid \sigma^2, \boldsymbol{Y}, \tau)$, $[2.4/\sqrt{d}]^2$ is the optimal scale factor of the target variance found by Gelman et al. (1996) with d being the dimension of the parameters updated in the MCMC step. The variance parameter $\sigma^2$ was sampled from a log normal proposal distribution. The transition step was tuned so that the acceptance rate is 44 %. The inverse temperature parameter was updated by drawing independent samples from the standard uniform proposal distribution.

In order to evaluate the prior $P(\tau)$ in (2.12), $\mu$ and $\sigma^2$ were optimized using closed forms of the conditional posterior mean of $P(\mu \mid \boldsymbol{Y}, \sigma^2, \tau)$ and the conditional posterior mode of $P(\sigma^2 \mid \boldsymbol{Y}, \mu, \tau)$, respectively. In particular, $\mu$ and $\sigma^2$ were maximized in a conditional iterative manner by optimizing each of them conditional on the last optimized value of the other. Iterations were repeated until the optimized values of the both parameters stopped changing within a tolerance level of $10^{-3}$. Using explicit formulae of posterior means (or modes) avoids numerical issues that are usually associated with optimization routines.

**Marginal likelihood estimation using thermodynamic integration via Parallel Tempering – bimodal model**

Convergence of each of the PT chains used to obtain the TI-PT-NB and the TI-PT-B estimates was assessed using the Potential Scale Reduction Factor (PSRF) or $\hat{R}$ statistics by Gelman and Rubin (1992), which compares the in-chain and between-chain variances of the chains for each of the 20 runs. The observed $\hat{R} < 1.1$ in our runs indicates that the chains have converged.

# Appendix C

# Implementation PT-STWNC, SIR model

The proposal distribution of $\tau$ was a truncated standard normal at [0,1]. The proposal distributions of $\alpha$ and $\beta$ were log-normal. The proposal distribution of the $I(0)$ was binomial, set up as follows: the proposed at the $i-th$ iteration is: $I(0)^{(i)} = binomial(N, \frac{I(0)^{(i-1)}}{N})$, where N=261 is the total population.

Optimization of $\boldsymbol{\theta} = (\alpha, \beta, I(0))^{'}$, which is needed to evaluate the prior of the inverse temperature $\tau$, was carried out by first optimizing the continuous $(\alpha, \beta)^{'}$ conditional on fixed discrete values of $I(0) = \{1, .., 8\}$ using the Nelder-Mead optimization routine. Then, out of the eight optimized values $(\alpha_{max}, \beta_{max}, I(0) \mid I(0) \in \{1, 2, .., 8\})^{'}$, the one that maximizes the posterior distribution $P(\alpha, \beta, I(0) \mid \boldsymbol{Y}, \tau)$ was chosen as a maximum.

# Appendix D

# Implementation of IMIS-Opt and IMIS-ShOpt, SIR model

The IMIS-ShOpt for the SIR model draws samples from the target posterior $P(\alpha, \beta, I(0) \mid \boldsymbol{Y})$ by sampling the two continuous parameters $\alpha, \beta$ conditionally on the $I(0)$ while updating $I(0)$ uniformly over $\{1, 2.., 10\}$. The algorithm starts with initial particles $\{\alpha, \beta, I(0)\}$ from the prior given in (3.20), and then calculates initial weights using the likelihood in (3.19). The Shotgun optimization, optimizes the sum of squared error function in (3.11), by finding local maxima of $\alpha, \beta$ conditional on $I(0) \in \{1, 2, .., 10\}$, i.e., $(\alpha, \beta \mid I(0) \in \{1, 2, .., 10\})$. For each newly discovered local mode, B samples $(\alpha, \beta)$ are drawn from the multivariate Gaussian, while $I(0)$ is updated with the corresponding value from $\{1, 2, .., 10\}$. Similarly, in the importance sampling stage, the maximum weight point is selected and the weighted covariance is calculated using the $\alpha, \beta \mid I(0) \in \{1, 2, .., 10\}$. The new B samples $(\alpha, \beta)$ are drawn from the multivariate Gaussian, while fixing the B samples from $I(0)$ to the value of $I(0)$ from the currently selected maximum weight point. Pseudo code of the implementation of the IMIS-ShOpt on the SIR model is given in the Algorithm 6.

Similar to the IMIS-ShOpt, the IMIS-Opt on the SIR model, updates the two continuous parameters $\alpha, \beta$ conditionally on the $I(0)$ while updating $I(0)$ uniformly over $\{1, 2.., 10\}$. The optimization stage is implemented as follows. Optimization of $\boldsymbol{\theta} = (\alpha, \beta, I(0))^{'}$, was carried out by first optimizing the conditional posterior distribution $P(\alpha, \beta \mid I(0), \boldsymbol{Y}, \tau)$ for each $I(0) = \{1, .., 10\}$. Then, out of the ten optimized values $(\alpha_{max}, \beta_{max}, I(0) \mid I(0) \in \{1, 2, .., 10\})^{'}$, the one that maximizes the posterior distribution $P(\alpha, \beta, I(0) \mid \boldsymbol{Y}, \tau)$ was chosen as a maximum. Hence, instead of keeping all the 10 optima and using them to repopulate the importance sampling distribution as in IMIS-ShOpt, the IMIS-Opt uses only one optima to repopulate the importance sampling distribution. The Hessian matrix was obtained using the conditional posterior $P(\alpha, \beta \mid I(0), \boldsymbol{Y}, \tau)$ for the corresponding $I(0) = \{1, .., 10\}$. The importance sampling stage follows the importance sampling stage in the IMIS-ShOpt.

**Algorithm 6** IMIS-ShOpt for the SIR model

**Goal: Draw samples from the target distribution $P(\boldsymbol{\theta} \mid \boldsymbol{Y})$ of the SIR model, where $\boldsymbol{\theta} = (\alpha, \beta, I(0))'$.**

**Input:** Data, model, likelihood function, prior distribution, $B$ - the number of incremental points, $D$ - the number of different initial points for the optimization, $N_0$ - the number of the initial samples from the prior and $J$ - the number of re-sampled points, $N$ - the number of iterations.

**Initial stage:** Draw $N_0$ samples $\boldsymbol{\Theta}_0 = \{\boldsymbol{\theta_1}, \boldsymbol{\theta_2}, ..., \boldsymbol{\theta_{N_0}}\}$ from the prior distribution $P(\boldsymbol{\theta})$ as per (3.20).

**for** $k = 1 : N$ **do**

**if** k=1 **then**

For each $\{\boldsymbol{\theta_i}, i = 1, .., N_0\}$ calculate the sampling weights:

$$w_i^{(1)} = \frac{P(\boldsymbol{Y} \mid \boldsymbol{\theta_i})}{\sum_{j=1}^{N_0} P(\boldsymbol{Y} \mid \boldsymbol{\theta_j})}, \tag{D.1}$$

using the likelihood function in (3.19)

**Optimization stage:**

**for** $d = 1 : D$ **do**

Find the d-th maximum weight point $\boldsymbol{\theta}_d^{(initial)} = \underset{\boldsymbol{\theta}}{\text{argmax}}\ \boldsymbol{w}^{(k)}(\boldsymbol{\theta})$, $\boldsymbol{\theta} \in \boldsymbol{\Theta}_{d-1}$ to initialize $Q$ optimizers.

**for** $q = 1 : 10$ **do**

Let $\check{\boldsymbol{\theta}} = (\boldsymbol{\theta} \mid I(0) = q)$ denote a vector of parameters of interest conditional on $I(0) = q$.

Use NLS method as per (3.11) initialized at $\boldsymbol{\theta}_d^{initial}$ to obtain local maxima

$$\check{\boldsymbol{\theta}}_{d,q}^{(Opt)} = \arg\min_{\check{\boldsymbol{\theta}}} \sum_{s=1}^{S} \sum_{j=1}^{n_s} \left[y_{sj} - \boldsymbol{X}(\boldsymbol{\theta}, t_{sj})\right]^2, \tag{D.2}$$

and obtain the corresponding inverse negative Hessian, $\boldsymbol{\Sigma}_{d,q}^{(Opt)}$, using the conditional target posterior $P(\alpha, \beta \mid I(0) = q, \boldsymbol{Y})$.

Update $\boldsymbol{\Theta}_d$ by excluding $\frac{N_0}{QD}$ nearest neighbor points, $\boldsymbol{\theta}_k \in \boldsymbol{\Theta}_{d-1}$, that minimize the Mahalanobis distance,

$$(\check{\theta}_k - \check{\boldsymbol{\theta}}_{d,q}^{(Opt)})' (\boldsymbol{\Sigma}_{d,q}^{(Opt)})^{-1} (\check{\theta}_k - \check{\boldsymbol{\theta}}_{d,q}^{(Opt)}). \tag{D.3}$$

Draw B samples $\check{\boldsymbol{\theta}}_{1:B} \sim MVN(\check{\boldsymbol{\theta}}_{d,q}^{(Opt)}, \boldsymbol{\Sigma}_{d,q}^{(Opt)})$ and repopulate B samples with $I(0) = q$; add these points to the importance sampling distribution and evaluate $H_k = MVN(\check{\boldsymbol{\theta}}_{1:B} \mid \check{\boldsymbol{\theta}}_{d,q}^{(Opt)}, \boldsymbol{\Sigma}_{d,q}^{(Opt)})$.

**end for**

**end for**

**else**

**Importance sampling stage:**

For each $\{\boldsymbol{\theta_i}, i = 1, .., N_k\}$ calculate weights,

**Algorithm 6** IMIS-ShOpt for the SIR model - continued

$$w_i^{(k)} = \frac{cP(\boldsymbol{Y} \mid \boldsymbol{\theta_i})P(\check{\boldsymbol{\theta}}_i)}{\frac{N_0}{N_k}P(\check{\boldsymbol{\theta}}_i) + \frac{B}{N_k}\sum_{s=1}^{k} H_s(\check{\boldsymbol{\theta}}_i)}, \tag{D.4}$$

where $N_k = N_0 + B(D+k)$ and $c = 1/\sum_{i=1}^{N_k} w_i^{(k)}$ is the normalizing constant.

Choose the maximum weight input $\boldsymbol{\theta}_k$ and extract $s = I(0)$ for this point; then estimate $\boldsymbol{\Sigma}_k$ as the weighted covariance of B inputs with smallest Mahalanobis distance,

$$w_p(\boldsymbol{\theta})\left(\check{\boldsymbol{\theta}} - \check{\boldsymbol{\theta}}_k\right)' (\boldsymbol{\Sigma_\pi})^{-1}\left(\check{\boldsymbol{\theta}} - \check{\boldsymbol{\theta}}_k\right),$$

where $\check{\boldsymbol{\theta}} = (\boldsymbol{\theta}_k \mid I(0) = s)$ corresponds to the vector of parameters of interest conditional on the current value of $I(0) = s$, the weights $w_p(\boldsymbol{\theta})$ are proportional to the average of the importance weights and the uniform weights $\frac{1}{N_k}$, $\boldsymbol{\Sigma}_\pi$ is the covariance of the initial importance distribution.

Draw B samples $\check{\boldsymbol{\theta}}_{1:B} \sim MVN(\check{\boldsymbol{\theta}}_k, \boldsymbol{\Sigma}_k)$; add these points to the importance sampling distribution and re-populate the new B samples for $I(0) = s$; then evaluate $H_k = MVN(\check{\boldsymbol{\theta}}_{1:B} \mid \check{\boldsymbol{\theta}}_k, \boldsymbol{\Sigma}_k)$.

**end if**

**if** $\sum_{1}^{N_k}(1-(1-w^{(k)})^J) \geq J(1-\exp(-1))$ i.e., importance sampling weights are approximately uniform **then** exit for loop

**end if**

**end for**

**Re-sampling stage:**

Re-sample J points with replacement from $\{\boldsymbol{\theta}_1, .., \boldsymbol{\theta}_{N_k}\}$ and weights $(w_1, .., w_{N_k})'$.

Both algorithms, the IMIS-Opt and the IMIS-ShOpt, used diffuse prior densities for the SIR-ODE model parameters. As a result, a big proportion of the initial importance samples drawn from the prior distribution fall outside the domain of the ODE model where the solution does not exist. For algorithmic convenience, the log-likelihood for the points outside the domain of the ODE system was set to take very small values (e.g., - 999999) so that the weights of these points were effectively zero. Hence, both algorithms were initialized with only few non-zero weight samples from the prior. The IMIS-ShOpt employed Q=10 optimization methods initialized at the highest weight point to discover 10 different modes, whereas the IMIS-Opt used only one optimization routine initialized at the highest weight point to find only one mode. The rest of the non-zero weight initial points were within the basin of attraction of the previously discovered modes, and hence, they were excluded from the set of candidates initial optimization points.

Optimization step in both algorithms continued initializing the optimizers with zero-weight points, which in turn did not contribute in discovering new modes. These 'bad' points could have been either physically removed from the importance distribution or kept with their likelihood set to an extremely small value (e.g, -9999999). Keeping the 'bad' points in the importance sampling distribution did not harm the convergence of the both algorithms, because ultimately the highest-weight points were re-sampled in the final stage.

# Appendix E

# Derivation of the MBIC

The MBIC approximately solves the log of the marginal likelihood integral,

$$-2\log P(\boldsymbol{Y} \mid M_l) = -2\log \int P(\boldsymbol{Y} \mid \boldsymbol{\theta}_l, M_l) P(\boldsymbol{\theta}_l \mid M_l) d\boldsymbol{\theta}_l, \tag{E.1}$$

where $\boldsymbol{\theta}_l$ is a $q_l$ dimensional vector of parameters of the model $M_l$. Following Gelman et al. (2014), the multi-modal conditional posterior distribution of $\boldsymbol{\theta}_l$ with fairly widely separated modes, given the data $\boldsymbol{Y} = (y_1, .., y_n)'$,

$$P(\boldsymbol{\theta}_l \mid \boldsymbol{Y}, M_l) \propto \prod_{i=1}^{n} P(y_i \mid \boldsymbol{\theta}_l, M_l) P(\boldsymbol{\theta}_l \mid M_l) = P(\boldsymbol{\theta}_l, \boldsymbol{Y} \mid M_l), \tag{E.2}$$

can be approximated with a mixture of $K$ locally weighted unimodal densities $g(\hat{\boldsymbol{\theta}}_{l,k}, \boldsymbol{Y} \mid M_l) = p_k f_k(\boldsymbol{\theta}_l, \boldsymbol{Y} \mid M_l)$ about the $k$-th mode, $\hat{\boldsymbol{\theta}}_{l,k}$,

$$P(\boldsymbol{\theta}_l, \boldsymbol{Y} \mid M_l) \approx \sum_{k=1}^{K} g(\hat{\boldsymbol{\theta}}_{l,k}, \boldsymbol{Y} \mid M_l). \tag{E.3}$$

The $\boldsymbol{\theta}_l$ increases its dimension to a $K \times q_l$ matrix, where $K$ is the number of mixture components. Since MBIC requires the maximum value of the $p_k f_k$, the second derivative of $g(\hat{\boldsymbol{\theta}}_{l,k}, \boldsymbol{Y} \mid M_l) = p_k f_k(\boldsymbol{\theta}_l, \boldsymbol{Y} \mid M_l)\Big|_{\boldsymbol{\theta}_l = \hat{\boldsymbol{\theta}}_{l,k}}$ around the mode $\hat{\boldsymbol{\theta}}_{l,k}$ must exist, and thus the $p_k f_k(\boldsymbol{\theta}_l, \boldsymbol{Y} \mid M_l)$ is maximized when $\frac{\partial P(\boldsymbol{\theta}_l, \boldsymbol{Y} \mid M_l)}{\partial \boldsymbol{\theta}_l}\Big|_{\boldsymbol{\theta}_l = \hat{\boldsymbol{\theta}}_{l,k}} = \frac{\partial p_k f_k(\boldsymbol{\theta}_l, \boldsymbol{Y} \mid M_l)}{\partial \boldsymbol{\theta}_l}\Big|_{\boldsymbol{\theta}_l = \hat{\boldsymbol{\theta}}_{l,k}} = 0$ and

$\left.\frac{\partial^2 P(\boldsymbol{\theta}_l, \boldsymbol{Y}|M_l)}{\partial \boldsymbol{\theta}_l \partial \boldsymbol{\theta}_l^{'}}\right|_{\boldsymbol{\theta}_l=\hat{\boldsymbol{\theta}}_{l,k}} < 0$. Then, the marginal likelihood is,

$$\begin{aligned} P(\boldsymbol{Y} \mid M_l) = \int P(\boldsymbol{\theta}_l, \boldsymbol{Y} \mid M_l) d\boldsymbol{\theta}_l &\approx \int \sum_{k=1}^{K} p_k f_k(\boldsymbol{\theta}_l, \boldsymbol{Y} \mid M_l) d\boldsymbol{\theta}_l \\ &= \sum_{k=1}^{K} \int p_k f_k(\boldsymbol{\theta}_l, \boldsymbol{Y} \mid M_l) d\boldsymbol{\theta}_l \\ &= \sum_{k=1}^{K} \int \exp\left(\log\left(p_k f_k(\boldsymbol{\theta}_l, \boldsymbol{Y} \mid M_l)\right)\right) d\boldsymbol{\theta}_l. \quad \text{(E.4)} \end{aligned}$$

Laplace's method, which is applied locally to each of the $K$ integrals in (E.4), relies on the second order Taylor expansion of $p_k f_k(\boldsymbol{\theta}_l, \boldsymbol{Y} \mid M_l)$ around the $k$-th mode $\hat{\boldsymbol{\theta}}_{l,k}$,

$$\log p_k f_k(\boldsymbol{\theta}_l, \boldsymbol{Y} \mid M_l) = \log p_k f_k(\hat{\boldsymbol{\theta}}_{l,k}, \boldsymbol{Y} \mid M_l) - \frac{n}{2}(\boldsymbol{\theta}_l - \hat{\boldsymbol{\theta}}_{l,k})^{'} H(\hat{\boldsymbol{\theta}}_{l,k})(\boldsymbol{\theta}_l - \hat{\boldsymbol{\theta}}_{l,k}) + \cdots, \quad \text{(E.5)}$$

where

$$H(\hat{\boldsymbol{\theta}}_{l,k}) \equiv -\frac{1}{n} \left.\frac{\partial^2 \log p_k f_k(\boldsymbol{\theta}_l, \boldsymbol{Y} \mid M_l)}{\partial \boldsymbol{\theta}_l \partial \boldsymbol{\theta}_l^{'}}\right|_{\boldsymbol{\theta}_l=\hat{\boldsymbol{\theta}}_{l,k}} \quad \text{(E.6)}$$

is the Hessian matrix at the mode around $\hat{\boldsymbol{\theta}}_{l,k}$. The second term of the Taylor expansion in (E.5) represents a $q_l$-dimensional Normal distribution with mean $\hat{\boldsymbol{\theta}}_{l,k}$ and variance-covariance matrix $\frac{H^{(-1)}(\hat{\boldsymbol{\theta}}_{l,k})}{n}$.

Next, each of the integrands in (E.4) is replaced with their Taylor expansions to obtain Gaussian integrals. Solving these integrals gives,

$$\begin{aligned} & P(\boldsymbol{Y} \mid M_l) \approx \\ \approx & \sum_{k=1}^{K} \int \exp\left(\log\left(p_k f_k(\hat{\boldsymbol{\theta}}_{l,k}, \boldsymbol{Y} \mid M_l)\right)\right) \exp\{-\frac{n}{2}(\boldsymbol{\theta}_l - \hat{\boldsymbol{\theta}}_{l,k})^{'} H(\hat{\boldsymbol{\theta}}_{l,k})(\boldsymbol{\theta}_l - \hat{\boldsymbol{\theta}}_{l,k})\} d\boldsymbol{\theta}_l \\ = & (2\pi)^{\frac{q_l}{2}} \sum_{k=1}^{K} p_k f_k(\hat{\boldsymbol{\theta}}_{l,k}, \boldsymbol{Y} \mid M_l) n^{-\frac{q_l}{2}} \left|H(\hat{\boldsymbol{\theta}}_{l,k})\right|^{-\frac{1}{2}} \\ = & (2\pi)^{\frac{q_l}{2}} \sum_{k=1}^{K} P(\hat{\boldsymbol{\theta}}_{k,l}, \boldsymbol{Y} \mid M_l) n^{-\frac{q_l}{2}} \left|H(\hat{\boldsymbol{\theta}}_{l,k})\right|^{-\frac{1}{2}}. \quad \text{(E.7)} \end{aligned}$$

Taking a logarithm of both sides of (E.7) and multiplying by $-2$ gives the MBIC,

$$MBIC = -2 \log\left((2\pi)^{\frac{q_l}{2}} \sum_{k=1}^{K} g(\hat{\boldsymbol{\theta}}_{l,k}, \boldsymbol{Y} \mid M_l) n^{-\frac{q_l}{2}} \left|H(\hat{\boldsymbol{\theta}}_{l,k})\right|^{-\frac{1}{2}}\right). \quad \text{(E.8)}$$

When $n \to \infty$, the influence of the prior in the equation (E.2) diminishes, and therefore $\sum_{k=1}^{K} p_k f_k(\boldsymbol{\theta}_l, \boldsymbol{Y} \mid M_l) = P(\boldsymbol{\theta}_l, \boldsymbol{Y} \mid M_l) \to \prod_{i=1}^{n} P(\boldsymbol{y_i} \mid \boldsymbol{\theta}_l, M_l)$, the variance term approaches zero, $|H(\hat{\boldsymbol{\theta}}_{l,k})| \to 0$ and the $p_k f_k(\boldsymbol{\theta}_l, \boldsymbol{Y} \mid M_l)$ becomes a Dirac delta function with point

mass around $\hat{\boldsymbol{\theta}}_{l,k}$, i.e., $\left\{ \begin{array}{ll} \infty, & \boldsymbol{\theta}_l = \hat{\boldsymbol{\theta}}_{l,k} \\ 0, & \boldsymbol{\theta}_l \neq \hat{\boldsymbol{\theta}}_{l,k} \end{array} \right\}$. Potential numerical issues induced by using the non-log form of the joint density, $P(\hat{\boldsymbol{\theta}}_{l,k}, \boldsymbol{Y} \mid M_l)$, in (E.8) can be avoided by evaluating the log form, $\log P(\hat{\boldsymbol{\theta}}_{l,k}, \boldsymbol{Y} \mid M_l)$, before exponentiating it. When the exponentiated value is zero, a constant can be added to the $\log P(\hat{\boldsymbol{\theta}}_{l,k}, \boldsymbol{Y} \mid M_l)$ and subtracted after the MBIC is calculated.

When $P(\boldsymbol{\theta}_l, \boldsymbol{Y} \mid M_l)$ is unimodal, the MBIC reduces to the unimodal Laplace's approximation method by Tierney and Kadane (1986), i.e.,

$$MBIC = LA = -2 \log \left( (2\pi)^{\frac{q_l}{2}} P(\boldsymbol{Y} \mid \hat{\boldsymbol{\theta}}_{l,k}, M_l) P(\hat{\boldsymbol{\theta}}_{l,k} \mid M_l) n^{-\frac{q_l}{2}} \left| H(\hat{\boldsymbol{\theta}}_{l,k}) \right|^{-\frac{1}{2}} \right), \tag{E.9}$$

which for large sample size reduces to the BIC (Schwarz et al., 1978),

$$BIC = -2 \log P(\boldsymbol{Y} \mid \hat{\boldsymbol{\theta}}_{l,k}, M_l) + q_l \log n. \tag{E.10}$$

The BIC is a special case of the MBIC, and hence, the model selection strategy chooses the model with lowest MBIC.

# Appendix F

# The unimodal model

Consider univariate data $\boldsymbol{Y} = (y_1, .., y_n)^{'}$, and a Gaussian likelihood characterized by parameter $\boldsymbol{\theta}$. The likelihood of the data given $\boldsymbol{\theta}$ is Gaussian, $\boldsymbol{Y} \mid \boldsymbol{\theta} \sim N(\boldsymbol{\theta}, s^2_{\boldsymbol{Y}})$, and the prior distribution of the unknown parameter $\boldsymbol{\theta}$ is also Gaussian, $\boldsymbol{\theta} \sim N(\lambda, s^2_{\boldsymbol{\theta}})$. The model notation $M_l$ is omitted for notational simplicity, since this section assesses the quality of the approximation to the marginal likelihood from the BIC and the MBIC within a single model. The unimodal joint density of $\boldsymbol{\theta}$ and $\boldsymbol{Y}$ is given by,

$$\begin{aligned} P_{uni}(\boldsymbol{\theta}, \boldsymbol{Y}) &= P(\boldsymbol{Y} \mid \boldsymbol{\theta})P(\boldsymbol{\theta}) = \prod_{i=1}^{n} P(y_i \mid \boldsymbol{\theta})P(\boldsymbol{\theta}) \qquad \text{(F.1)} \\ &= \left(\frac{1}{\sqrt{2\pi}s_{\boldsymbol{Y}}}\right)^n \exp\{-\frac{1}{2s^2_{\boldsymbol{Y}}}\sum_{i=1}^{n}(y_i - \boldsymbol{\theta})^2\}\frac{1}{\sqrt{2\pi}s_{\boldsymbol{\theta}}}\exp\{-\frac{1}{2s^2_{\boldsymbol{\theta}}}(\boldsymbol{\theta} - \lambda)^2\} \\ &= \frac{1}{(2\pi)^{\frac{n+1}{2}} s^n_{\boldsymbol{Y}} s_{\boldsymbol{\theta}}} \exp\{-\frac{1}{2g_{s_{\boldsymbol{\theta}}}(s^2_{\boldsymbol{Y}}, s^2_{\boldsymbol{\theta}})}(\boldsymbol{\theta} - g_{\boldsymbol{\theta}}(\boldsymbol{Y}, \lambda, s^2_{\boldsymbol{Y}}, s^2_{\boldsymbol{\theta}}))^2\} \\ &\times \exp\{-\frac{1}{2}(g(\boldsymbol{Y}, \lambda, s^2_{\boldsymbol{Y}}, s^2_{\boldsymbol{\theta}}))\}, \qquad \text{(F.2)} \end{aligned}$$

where

$$g_{\boldsymbol{\theta}}(\boldsymbol{Y}, \lambda, s^2_{\boldsymbol{Y}}, s^2_{\boldsymbol{\theta}}) = \frac{\frac{\sum_{i=1}^{n} y_i}{s^2_{\boldsymbol{Y}}} + \frac{\lambda}{s^2_{\boldsymbol{\theta}}}}{\frac{n}{s^2_{\boldsymbol{Y}}} + \frac{1}{s^2_{\boldsymbol{\theta}}}}, \qquad \text{(F.3)}$$

$$g_{s_{\boldsymbol{\theta}}}(s^2_{\boldsymbol{Y}}, s^2_{\boldsymbol{\theta}}) = \frac{1}{\frac{n}{s^2_{\boldsymbol{Y}}} + \frac{1}{s^2_{\boldsymbol{\theta}}}}, \qquad \text{(F.4)}$$

$$g(\boldsymbol{Y}, \lambda, s^2_{\boldsymbol{Y}}, s^2_{\boldsymbol{\theta}}) = \frac{\lambda^2}{s^2_{\boldsymbol{\theta}}} + \frac{\sum_{i=1}^{n} y_i^2}{s^2_{\boldsymbol{Y}}} - \frac{\left(\frac{\sum_{i=1}^{n} y_i}{s^2_{\boldsymbol{Y}}} + \frac{\lambda}{s^2_{\boldsymbol{\theta}}}\right)^2}{\frac{n}{s^2_{\boldsymbol{Y}}} + \frac{1}{s^2_{\boldsymbol{\theta}}}}. \qquad \text{(F.5)}$$

**The analytical solution of the marginal likelihood in the unimodal model**

$$\begin{aligned} P_{ana\_uni}(\boldsymbol{Y}) &= \int P_{uni}(\boldsymbol{\theta}, \boldsymbol{Y}) d\boldsymbol{\theta} \\ &= \frac{1}{(2\pi)^{\frac{n+1}{2}} s_{\boldsymbol{Y}}^n s_{\boldsymbol{\theta}}} \exp\{-\frac{1}{2}(g(\boldsymbol{Y}, \lambda, s_{\boldsymbol{Y}}^2, s_{\boldsymbol{\theta}}^2))\} \\ &\times \int \exp\{-\frac{1}{2g_{s_{\boldsymbol{\theta}}}(s_{\boldsymbol{Y}}^2, s_{\boldsymbol{\theta}}^2)}(\boldsymbol{\theta} - g_{\boldsymbol{\theta}}(\boldsymbol{Y}, \lambda, s_{\boldsymbol{Y}}^2, s_{\boldsymbol{\theta}}^2))^2\} d\boldsymbol{\theta}. \end{aligned} \tag{F.6}$$

The integrand in (F.6) is a kernel of a Gaussian distribution,

$$N(g_{\boldsymbol{\theta}}(\boldsymbol{Y}, \lambda, s_{\boldsymbol{Y}}^2, s_{\boldsymbol{\theta}}^2), g_{s_{\boldsymbol{\theta}}}(s_{\boldsymbol{Y}}^2, s_{\boldsymbol{\theta}}^2)), \tag{F.7}$$

with probability density function that integrates to one, and therefore, the solution to the integral in (F.6) is the normalizing constant of (F.7), $\sqrt{2\pi g_{s_{\boldsymbol{\theta}}}(s_{\boldsymbol{Y}}^2, s_{\boldsymbol{\theta}}^2)}$. Consequently, the analytical solution to (4.2) is,

$$P_{ana\_uni}(\boldsymbol{Y}) = \frac{\sqrt{g_{s_{\boldsymbol{\theta}}}(s_{\boldsymbol{Y}}^2, s_{\boldsymbol{\theta}}^2)}}{(2\pi)^{\frac{n}{2}} s_{\boldsymbol{Y}}^n s_{\boldsymbol{\theta}}} \exp\{-\frac{1}{2} g(\boldsymbol{Y}, \lambda, s_{\boldsymbol{Y}}^2, s_{\boldsymbol{\theta}}^2)\}.$$

### F.0.1 The Laplace approximation (LA) in the unimodal model

The LA evaluated at the posterior means or Maximum A Posterior estimate, $\hat{\boldsymbol{\theta}} = g_{\boldsymbol{\theta}}(\boldsymbol{Y}, \lambda, s_{\boldsymbol{Y}}^2, s_{\boldsymbol{\theta}}^2)$, is,

$$\begin{aligned} \exp\{-\frac{1}{2} LA(\hat{\boldsymbol{\theta}})\} &= P_{uni}(\hat{\boldsymbol{\theta}}, \boldsymbol{Y}) n^{-\frac{q}{2}} |H(\hat{\boldsymbol{\theta}})|^{-\frac{1}{2}} \sqrt{2\pi} \\ &= \frac{1}{(2\pi)^{\frac{n+1}{2}} s_{\boldsymbol{Y}}^n s_{\boldsymbol{\theta}}} \exp\{-\frac{1}{2} g(\boldsymbol{Y}, \lambda, s_{\boldsymbol{Y}}^2, s_{\boldsymbol{\theta}}^2)\} \sqrt{2\pi g_{s_{\boldsymbol{\theta}}}(s_{\boldsymbol{Y}}^2, s_{\boldsymbol{\theta}}^2)}. \end{aligned}$$

Since the model is Gaussian, the approximation obtained from the LA is exactly equal to the solution to (4.2),

$$\exp\{-\frac{1}{2} LA(\hat{\boldsymbol{\theta}})\} = P_{ana\_uni}(\boldsymbol{Y}).$$

### F.0.2 The MBIC in the unimodal model

The MBIC for the unimodal case is exactly the same as the LA and the analytical solution,

$$\begin{aligned}
\exp\{-\frac{1}{2}MBIC(\hat{\boldsymbol{\theta}})\} &= P_{uni}(\hat{\boldsymbol{\theta}}, \boldsymbol{Y}) n^{-\frac{q}{2}} |H(\hat{\boldsymbol{\theta}})|^{-\frac{1}{2}} \sqrt{2\pi} \\
&= \frac{\sqrt{2\pi g_{s_{\boldsymbol{\theta}}}(s^2_{\boldsymbol{Y}}, s^2_{\boldsymbol{\theta}})}}{(2\pi)^{\frac{n+1}{2}} s^n_{\boldsymbol{Y}} s_{\boldsymbol{\theta}}} \exp\{-\frac{1}{2} g(\boldsymbol{Y}, \lambda, s^2_{\boldsymbol{Y}}, s^2_{\boldsymbol{\theta}})\} \\
&= P_{ana_uni}(\boldsymbol{Y}).
\end{aligned}$$

# Appendix G

# Scenario 1. Bimodal model with bimodal prior and unimodal likelihood

The likelihood is Gaussian $P(\boldsymbol{Y} \mid \boldsymbol{\theta}) = N(\boldsymbol{Y} \mid \boldsymbol{\theta}, s^2_{\boldsymbol{Y}})$, and the prior is bimodal,

$$P(\boldsymbol{\theta}) = \frac{1}{2} N(\boldsymbol{\theta} \mid \lambda_1, s^2_{\boldsymbol{\theta}}) + \frac{1}{2} N(\boldsymbol{\theta} \mid \lambda_2, s^2_{\boldsymbol{\theta}}). \tag{G.1}$$

The joint density of the $\boldsymbol{\theta}$ and $\boldsymbol{Y}$ is given by,

$$\begin{aligned}
P_{bi}(\boldsymbol{\theta}, \boldsymbol{Y}) &= \frac{1}{2} \frac{1}{(2\pi)^{\frac{n+1}{2}} s^n_{\boldsymbol{Y}} s_{\boldsymbol{\theta}}} \exp\{-\frac{1}{2s^2_{\boldsymbol{Y}}} \sum_{i=1}^{n} (y_i - \boldsymbol{\theta})^2\} \exp\{-\frac{1}{2s^2_{\boldsymbol{\theta}}} (\boldsymbol{\theta} - \lambda_1)^2\} \\
&+ \frac{1}{2} \frac{1}{(2\pi)^{\frac{n+1}{2}} s^n_{\boldsymbol{Y}} s_{\boldsymbol{\theta}}} \exp\{-\frac{1}{2s^2_{\boldsymbol{Y}}} \sum_{i=1}^{n} (y_i - \boldsymbol{\theta})^2\} \exp\{-\frac{1}{2s^2_{\boldsymbol{\theta}}} (\boldsymbol{\theta} - \lambda_2)^2\} \\
&= \frac{1}{2} \frac{1}{(2\pi)^{\frac{n+1}{2}} s^n_{\boldsymbol{Y}} s_{\boldsymbol{\theta}}} \exp\{-\frac{1}{2g_{s_{\boldsymbol{\theta}}}(s^2_{\boldsymbol{Y}}, s^2_{\boldsymbol{\theta}})} (\boldsymbol{\theta} - g_{\boldsymbol{\theta}}(\boldsymbol{Y}, \lambda_1, s^2_{\boldsymbol{Y}}, s^2_{\boldsymbol{\theta}}))^2\} \\
&\times \exp\{-\frac{1}{2}(g(\boldsymbol{Y}, \lambda_1, s^2_{\boldsymbol{Y}}, s^2_{\boldsymbol{\theta}}))\} \\
&+ \frac{1}{2} \frac{1}{(2\pi)^{\frac{n+1}{2}} s^n_{\boldsymbol{Y}} s_{\boldsymbol{\theta}}} \exp\{-\frac{1}{2g_{s_{\boldsymbol{\theta}}}(s^2_{\boldsymbol{Y}}, s^2_{\boldsymbol{\theta}})} (\boldsymbol{\theta} - g_{\boldsymbol{\theta}}(\boldsymbol{Y}, \lambda_2, s^2_{\boldsymbol{Y}}, s^2_{\boldsymbol{\theta}}))^2\} \\
&\times \exp\{-\frac{1}{2}(g(\boldsymbol{Y}, \lambda_2, s^2_{\boldsymbol{Y}}, s^2_{\boldsymbol{\theta}}))\}.
\end{aligned} \tag{G.2}$$

#### G.0.0.0.1 Multi-modality condition.

Following Reschenhofer (2001), depending on the distance between the mixture component means $\lambda_1$ and $\lambda_2$, the prior density given by the equation (G.1) will exhibit either a maximum at $\frac{1}{2}(\lambda_1 + \lambda_2)$ (unimodal case), or a local minimum at $\frac{1}{2}(\lambda_1 + \lambda_2)$ (bimodal case). Hence, $\frac{\partial P(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}=\frac{1}{2}(\lambda_1+\lambda_2)} = 0$, and the 'multi-modality condition' can be derived by setting the

second derivative $\left.\frac{\partial^2 P(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right|_{\boldsymbol{\theta}=\frac{1}{2}(\lambda_1+\lambda_2)} > 0$, which results in $|\lambda_1 - \lambda_2| > 2s^2_{\boldsymbol{\theta}}$. Analogously to the prior distribution, the 'multi-modality condition' for the unnormalized conditional posterior density, $P(\boldsymbol{\theta} \mid \boldsymbol{Y})$, states that the absolute distance between the posterior means $\hat{\boldsymbol{\theta}}_1 = g_{\boldsymbol{\theta}}(\boldsymbol{Y}, \lambda_1, s^2_{\boldsymbol{Y}}, s^2_{\boldsymbol{\theta}})$ and $\hat{\boldsymbol{\theta}}_2 = g_{\boldsymbol{\theta}}(\boldsymbol{Y}, \lambda_2, s^2_{\boldsymbol{Y}}, s^2_{\boldsymbol{\theta}})$ is larger than twice the posterior standard deviation $\sqrt{g_{s_{\boldsymbol{\theta}}}(s^2_{\boldsymbol{Y}}, s^2_{\boldsymbol{\theta}})}$, i.e.,

$$\left| \frac{\frac{\sum_{i=1}^{n} y_i}{s^2_{\boldsymbol{Y}}} + \frac{\lambda_1}{s^2_{\boldsymbol{\theta}}}}{\frac{n}{s^2_{\boldsymbol{Y}}} + \frac{1}{s^2_{\boldsymbol{\theta}}}} - \frac{\frac{\sum_{i=1}^{n} y_i}{s^2_{\boldsymbol{Y}}} + \frac{\lambda_2}{s^2_{\boldsymbol{\theta}}}}{\frac{n}{s^2_{\boldsymbol{Y}}} + \frac{1}{s^2_{\boldsymbol{\theta}}}} \right| > \frac{2}{\sqrt{\frac{n}{s^2_{\boldsymbol{Y}}} + \frac{1}{s^2_{\boldsymbol{\theta}}}}}, \tag{G.3}$$

which reduces to $|\lambda_1 - \lambda_2| > 2s^2_{\boldsymbol{\theta}}\sqrt{\frac{n}{s^2_{\boldsymbol{Y}}} + \frac{1}{s^2_{\boldsymbol{\theta}}}}$.

### G.0.1 Analytical solution to the marginal likelihood in Scenario 1

$$\begin{aligned}
P_{ana_bi}(\boldsymbol{Y}) &= \int P_{bi}(\boldsymbol{\theta}, \boldsymbol{Y}) d\boldsymbol{\theta} \\
&= \frac{1}{2} \frac{1}{(2\pi)^{\frac{n+1}{2}} s^n_{\boldsymbol{Y}} s_{\boldsymbol{\theta}}} \exp\{-\frac{1}{2}(g(\boldsymbol{Y}, \lambda_1, s^2_{\boldsymbol{Y}}, s^2_{\boldsymbol{\theta}}))\} \\
&\times \int \exp\{-\frac{1}{2g_{s_{\boldsymbol{\theta}}}(s^2_{\boldsymbol{Y}}, s^2_{\boldsymbol{\theta}})}(\boldsymbol{\theta} - g_{\boldsymbol{\theta}}(\boldsymbol{Y}, \lambda_1, s^2_{\boldsymbol{Y}}, s^2_{\boldsymbol{\theta}}))^2\} d\boldsymbol{\theta} \\
&+ \frac{1}{2} \frac{1}{(2\pi)^{\frac{n+1}{2}} s^n_{\boldsymbol{Y}} s_{\boldsymbol{\theta}}} \exp\{-\frac{1}{2}(g(\boldsymbol{Y}, \lambda_2, s^2_{\boldsymbol{Y}}, s^2_{\boldsymbol{\theta}}))\} \\
&\times \int \exp\{-\frac{1}{2g_{s_{\boldsymbol{\theta}}}(s^2_{\boldsymbol{Y}}, s^2_{\boldsymbol{\theta}})}(\boldsymbol{\theta} - g_{\boldsymbol{\theta}}(\boldsymbol{Y}, \lambda_2, s^2_{\boldsymbol{Y}}, s^2_{\boldsymbol{\theta}}))^2\} d\boldsymbol{\theta}.
\end{aligned} \tag{G.4}$$

The analytical solution of the marginal likelihood in bimodal case is,

$$\begin{aligned}
P_{ana_bi}(\boldsymbol{Y}) &= \frac{1}{2} \frac{\sqrt{g_{s_{\boldsymbol{\theta}}}(s^2_{\boldsymbol{Y}}, s^2_{\boldsymbol{\theta}})}}{(2\pi)^{\frac{n}{2}} s^n_{\boldsymbol{Y}} s_{\boldsymbol{\theta}}} \exp\{-\frac{1}{2}(g(\boldsymbol{Y}, \lambda_1, s^2_{\boldsymbol{Y}}, s^2_{\boldsymbol{\theta}}))\} \\
&+ \frac{1}{2} \frac{\sqrt{g_{s_{\boldsymbol{\theta}}}(s^2_{\boldsymbol{Y}}, s^2_{\boldsymbol{\theta}})}}{(2\pi)^{\frac{n}{2}} s^n_{\boldsymbol{Y}} s_{\boldsymbol{\theta}}} \exp\{-\frac{1}{2}(g(\boldsymbol{Y}, \lambda_2, s^2_{\boldsymbol{Y}}, s^2_{\boldsymbol{\theta}}))\}.
\end{aligned} \tag{G.5}$$

### G.0.2 The Laplace approximation (LA) in Scenario 1

The LA in the bimodal case, uses one of the two posterior means, $\hat{\boldsymbol{\theta}}_1 = g_{\boldsymbol{\theta}}(\boldsymbol{Y}, \lambda_1, s^2_{\boldsymbol{Y}}, s^2_{\boldsymbol{\theta}})$ and $\hat{\boldsymbol{\theta}}_2 = g_{\boldsymbol{\theta}}(\boldsymbol{Y}, \lambda_2, s^2_{\boldsymbol{Y}}, s^2_{\boldsymbol{\theta}})$. The LA at the mode $\hat{\boldsymbol{\theta}}_1$ is,

$$\begin{aligned}
\exp\{-\frac{1}{2}LA(\hat{\boldsymbol{\theta}}_1)\} &= P_{bi}(\hat{\boldsymbol{\theta}}_1, \boldsymbol{Y}) n^{-\frac{q}{2}} |H(\hat{\boldsymbol{\theta}})|^{-\frac{1}{2}} \sqrt{2\pi} \qquad \text{(G.6)} \\
&= \frac{1}{2} \frac{1}{(2\pi)^{\frac{n+1}{2}} s^n_{\boldsymbol{Y}} s_{\boldsymbol{\theta}}} \sqrt{2\pi g_{s_{\boldsymbol{\theta}}}(s^2_{\boldsymbol{Y}}, s^2_{\boldsymbol{\theta}})} \exp\{-\frac{1}{2} g(\boldsymbol{Y}, \lambda_1, s^2_{\boldsymbol{Y}}, s^2_{\boldsymbol{\theta}})\} \\
&\times \underbrace{\exp\{-\frac{1}{2} \frac{(g_{\boldsymbol{\theta}}(\boldsymbol{Y}, \lambda_1, s^2_{\boldsymbol{Y}}, s^2_{\boldsymbol{\theta}}) - g_{\boldsymbol{\theta}}(\boldsymbol{Y}, \lambda_1, s^2_{\boldsymbol{Y}}, s^2_{\boldsymbol{\theta}}))^2}{g_{s_{\boldsymbol{\theta}}(s^2_{\boldsymbol{Y}}, s^2_{\boldsymbol{\theta}})}}\}}_{1} \\
&+ \frac{1}{2} \frac{1}{(2\pi)^{\frac{n+1}{2}} s^n_{\boldsymbol{Y}} s_{\boldsymbol{\theta}}} \sqrt{2\pi g_{s_{\boldsymbol{\theta}}}(s^2_{\boldsymbol{Y}}, s^2_{\boldsymbol{\theta}})} \exp\{-\frac{1}{2} g(\boldsymbol{Y}, \lambda_2, s^2_{\boldsymbol{Y}}, s^2_{\boldsymbol{\theta}})\} \\
&\times \underbrace{\exp\{-\frac{1}{2} \frac{(g_{\boldsymbol{\theta}}(\boldsymbol{Y}, \lambda_1, s^2_{\boldsymbol{Y}}, s^2_{\boldsymbol{\theta}}) - g_{\boldsymbol{\theta}}(\boldsymbol{Y}, \lambda_2, s^2_{\boldsymbol{Y}}, s^2_{\boldsymbol{\theta}}))^2}{g_{s_{\boldsymbol{\theta}}(s^2_{\boldsymbol{Y}}, s^2_{\boldsymbol{\theta}})}}\}}_{\approx 0, \text{ when } |\hat{\boldsymbol{\theta}}_1 - \hat{\boldsymbol{\theta}}_2| >> 6\sqrt{g_{s_{\boldsymbol{\theta}}(s^2_{\boldsymbol{Y}}, s^2_{\boldsymbol{\theta}})}} \text{ and } \frac{\sum_{i=1}^{2} y_i}{n} = \frac{\lambda_1+\lambda_2}{2}} . \qquad \text{(G.7)}
\end{aligned}$$

When the posterior is multi-modal, the last term in the equation $(G.7)$, is approaching zero. Then the LA evaluated at the mode $\hat{\boldsymbol{\theta}}_1$, satisfies $\exp\{-\frac{LA}{2}(\hat{\boldsymbol{\theta}}_1)\} = \delta P(\boldsymbol{Y})$ where $\delta$ depends on the importance of the posterior mode used. At the special case where $\frac{\sum_{i=1}^{n} y_i}{n} = \frac{\lambda_1+\lambda_2}{2}$ and the modes are well separated, i.e., $|\hat{\boldsymbol{\theta}}_1 - \hat{\boldsymbol{\theta}}_2| >> 6\sqrt{g_{s_{\boldsymbol{\theta}}(s^2_{\boldsymbol{Y}}, s^2_{\boldsymbol{\theta}})}}$, the LA is one half of the analytical solution of the marginal likelihood of the bimodal density,

$$\begin{aligned}
\exp\{-\frac{1}{2}LA(\hat{\boldsymbol{\theta}}_1)\} &= \frac{1}{2} \frac{\sqrt{g_{s_{\boldsymbol{\theta}}}(s^2_{\boldsymbol{Y}}, s^2_{\boldsymbol{\theta}})}}{(2\pi)^{\frac{n}{2}} s^n_{\boldsymbol{Y}} s_{\boldsymbol{\theta}}} \exp\{-\frac{1}{2} g(\boldsymbol{Y}, \lambda_1, s^2_{\boldsymbol{Y}}, s^2_{\boldsymbol{\theta}})\} \\
&= \frac{1}{2} P_{ana_bi}(\boldsymbol{Y}). \qquad \text{(G.8)}
\end{aligned}$$

The condition for well separated modes, which states that the distance between the two posterior modes are bigger than six standard deviations is chosen such that the second term in (G.7) has near-zero value. At this special case, the LA evaluated at the mode $\hat{\boldsymbol{\theta}}_2$ also satisfies (G.8), since the two modes around $\hat{\boldsymbol{\theta}}_1$ and $\hat{\boldsymbol{\theta}}_2$ have equal standard deviations and weights.

### G.0.3 The MBIC in Scenario 1

The MBIC takes into account the two posterior modes $\hat{\boldsymbol{\theta}}_1 = g_{\boldsymbol{\theta}}(\boldsymbol{Y}, \lambda_1, s^2_{\boldsymbol{Y}}, s^2_{\boldsymbol{\theta}})$ and $\hat{\boldsymbol{\theta}}_2 = g_{\boldsymbol{\theta}}(\boldsymbol{Y}, \lambda_2, s^2_{\boldsymbol{Y}}, s^2_{\boldsymbol{\theta}})$. In the special case when the modes are Gaussian and fairly widely separated, the local Laplace approximations applied to each of the modes give exact solutions

to the local marginal likelihood integrals,

$$\begin{aligned}
& \exp\{-\frac{1}{2}MBIC(\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2)\} \\
= & \sum_{k=1}^{2} g(\hat{\boldsymbol{\theta}}_k, \boldsymbol{Y}) n^{-\frac{q}{2}} |H(\hat{\boldsymbol{\theta}}_k)|^{-\frac{1}{2}} \sqrt{2\pi} = \sum_{k=1}^{2} g(\hat{\boldsymbol{\theta}}_k, \boldsymbol{Y}) \sqrt{g_{s_{\boldsymbol{\theta}}}(s^2_{\boldsymbol{Y}}, s^2_{\boldsymbol{\theta}})} \sqrt{2\pi} \\
= & \frac{1}{2} \frac{\sqrt{g_{s_{\boldsymbol{\theta}}}(s^2_{\boldsymbol{Y}}, s^2_{\boldsymbol{\theta}})}}{(2\pi)^{\frac{n}{2}} s^n_{\boldsymbol{Y}} s_{\boldsymbol{\theta}}} \exp\{-\frac{1}{2} g(\boldsymbol{Y}, \lambda_1, s^2_{\boldsymbol{Y}}, s^2_{\boldsymbol{\theta}})\} \\
+ & \frac{1}{2} \frac{\sqrt{g_{s_{\boldsymbol{\theta}}}(s^2_{\boldsymbol{Y}}, s^2_{\boldsymbol{\theta}})}}{(2\pi)^{\frac{n}{2}} s^n_{\boldsymbol{Y}} s_{\boldsymbol{\theta}}} \exp\{-\frac{1}{2} g(\boldsymbol{Y}, \lambda_2, s^2_{\boldsymbol{Y}}, s^2_{\boldsymbol{\theta}})\} \\
= & P_{ana_bi}(\boldsymbol{Y}).
\end{aligned} \tag{G.9}$$

Hence, the MBIC exactly matches the analytical solution $P_{ana\_bi}(\boldsymbol{Y})$.

### G.0.4 Bias in the MBIC, Scenario 1

The bias in MBIC arises when the modes are not well separated,

$$
\begin{aligned}
\exp\{-\frac{1}{2}MBIC(\hat{\boldsymbol{\theta}}_1,\hat{\boldsymbol{\theta}}_2)\} &= \sum_{k=1}^{2} g(\hat{\boldsymbol{\theta}}_k,\boldsymbol{Y}) n^{-\frac{q}{2}} |H(\hat{\boldsymbol{\theta}})|^{-\frac{1}{2}} \sqrt{2\pi} \\
&= \frac{1}{2}\frac{\sqrt{2\pi g_{s_{\boldsymbol{\theta}}}(s^2_{\boldsymbol{Y}},s^2_{\boldsymbol{\theta}})}}{(2\pi)^{\frac{n+1}{2}} s^n_{\boldsymbol{Y}} s_{\boldsymbol{\theta}}} \exp\{-\frac{1}{2} g(\boldsymbol{Y},\lambda_1,s^2_{\boldsymbol{Y}},s^2_{\boldsymbol{\theta}})\} \\
&\times \underbrace{\exp\{-\frac{1}{2}\frac{(\hat{\boldsymbol{\theta}}_1-\hat{\boldsymbol{\theta}}_1)^2}{g_{s_{\boldsymbol{\theta}}(s^2_{\boldsymbol{Y}},s^2_{\boldsymbol{\theta}})}}\}}_{1} \\
&+ \frac{1}{2}\frac{\sqrt{2\pi g_{s_{\boldsymbol{\theta}}}(s^2_{\boldsymbol{Y}},s^2_{\boldsymbol{\theta}})}}{(2\pi)^{\frac{n+1}{2}} s^n_{\boldsymbol{Y}} s_{\boldsymbol{\theta}}} \exp\{-\frac{1}{2} g(\boldsymbol{Y},\lambda_2,s^2_{\boldsymbol{Y}},s^2_{\boldsymbol{\theta}})\} \\
&\times \underbrace{\exp\{-\frac{1}{2}\frac{(\hat{\boldsymbol{\theta}}_1-\hat{\boldsymbol{\theta}}_2)^2}{g_{s_{\boldsymbol{\theta}}(s^2_{\boldsymbol{Y}},s^2_{\boldsymbol{\theta}})}}\}}_{\approx 0,\ \text{when}\ |\hat{\boldsymbol{\theta}}_1-\hat{\boldsymbol{\theta}}_2| >> 6\sqrt{g_{s_{\boldsymbol{\theta}}(s^2_{\boldsymbol{Y}},s^2_{\boldsymbol{\theta}})}}\ \text{and}\ \frac{\sum_{i=1}^{2} y_i}{n} = \frac{\lambda_1+\lambda_2}{2}} \\
&+ \frac{1}{2}\frac{\sqrt{2\pi g_{s_{\boldsymbol{\theta}}}(s^2_{\boldsymbol{Y}},s^2_{\boldsymbol{\theta}})}}{(2\pi)^{\frac{n+1}{2}} s^n_{\boldsymbol{Y}} s_{\boldsymbol{\theta}}} \exp\{-\frac{1}{2} g(\boldsymbol{Y},\lambda_1,s^2_{\boldsymbol{Y}},s^2_{\boldsymbol{\theta}})\} \\
&\times \underbrace{\exp\{-\frac{1}{2}\frac{(\hat{\boldsymbol{\theta}}_2-\hat{\boldsymbol{\theta}}_1)^2}{g_{s_{\boldsymbol{\theta}}(s^2_{\boldsymbol{Y}},s^2_{\boldsymbol{\theta}})}}\}}_{\approx 0,\ \text{when}\ |\hat{\boldsymbol{\theta}}_1-\hat{\boldsymbol{\theta}}_2| >> 6\sqrt{g_{s_{\boldsymbol{\theta}}(s^2_{\boldsymbol{Y}},s^2_{\boldsymbol{\theta}})}}\ \text{and}\ \frac{\sum_{i=1}^{2} y_i}{n} = \frac{\lambda_1+\lambda_2}{2}} \\
&+ \frac{1}{2}\frac{\sqrt{2\pi g_{s_{\boldsymbol{\theta}}}(s^2_{\boldsymbol{Y}},s^2_{\boldsymbol{\theta}})}}{(2\pi)^{\frac{n+1}{2}} s^n_{\boldsymbol{Y}} s_{\boldsymbol{\theta}}} \exp\{-\frac{1}{2} g(\boldsymbol{Y},\lambda_2,s^2_{\boldsymbol{Y}},s^2_{\boldsymbol{\theta}})\} \\
&\times \underbrace{\exp\{-\frac{1}{2}\frac{(\hat{\boldsymbol{\theta}}_2-\hat{\boldsymbol{\theta}}_2)^2}{g_{s_{\boldsymbol{\theta}}(s^2_{\boldsymbol{Y}},s^2_{\boldsymbol{\theta}})}}\}}_{1}. \qquad \text{(G.10)}
\end{aligned}
$$

If the two modes are not well separated, i.e., $|\hat{\boldsymbol{\theta}}_1 - \hat{\boldsymbol{\theta}}_2| > 2\sqrt{g_{s_{\boldsymbol{\theta}}(s^2_{\boldsymbol{Y}},s^2_{\boldsymbol{\theta}})}}$ and $|\hat{\boldsymbol{\theta}}_1 - \hat{\boldsymbol{\theta}}_2| < 6\sqrt{g_{s_{\boldsymbol{\theta}}(s^2_{\boldsymbol{Y}},s^2_{\boldsymbol{\theta}})}}$,

$$
\begin{aligned}
\exp\{-\frac{1}{2}MBIC(\hat{\boldsymbol{\theta}}_1,\hat{\boldsymbol{\theta}}_2)\} &> P_{ana\_bi}(Y) + \exp(-2) P_{ana\_bi}(\boldsymbol{Y}) \\
&= (1+\exp(-2)) P_{ana\_bi}(\boldsymbol{Y}), \qquad \text{(G.11)}
\end{aligned}
$$

where $P_{ana\_bi}(\boldsymbol{Y})$ is given in (G.5).

The condition for well separated modes, which states that the distance between the two posterior modes are bigger than six standard deviations is chosen such that the second and third terms in (G.10) has near-zero value.

# Appendix H

# Scenario 2. Bimodal model with bimodal likelihood and unimodal prior

The likelihood is a mixture of two Gaussian distributions with a single observed data point $\boldsymbol{Y} = y$,

$$P(\boldsymbol{Y} \mid \boldsymbol{\theta}) = \frac{1}{2} N(y \mid \boldsymbol{\theta}, s^2_{\boldsymbol{Y}}) + \frac{1}{2} N(y \mid -\boldsymbol{\theta}, s^2_{\boldsymbol{Y}}), \tag{H.1}$$

and the prior is Gaussian, $P(\boldsymbol{\theta}) = N(\boldsymbol{\theta} \mid \lambda, s^2_{\boldsymbol{\theta}})$. The joint density $P_{bi\_lik}(\boldsymbol{\theta}, \boldsymbol{Y})$ at a single observed data point $\boldsymbol{Y} = y$ is given by,

$$\begin{aligned}
P_{bi_lik}(\boldsymbol{\theta}, \boldsymbol{Y}) &= \frac{1}{2} \frac{1}{2\pi s_{\boldsymbol{Y}} s_{\boldsymbol{\theta}}} \exp\{-\frac{1}{2s^2_{\boldsymbol{Y}}}(y - \boldsymbol{\theta})^2\} \exp\{-\frac{1}{2s^2_{\boldsymbol{\theta}}}(\boldsymbol{\theta} - \lambda)^2\} \\
&+ \frac{1}{2} \frac{1}{2\pi s_{\boldsymbol{Y}} s_{\boldsymbol{\theta}}} \exp\{-\frac{1}{2s^2_{\boldsymbol{Y}}}(y + \boldsymbol{\theta})^2\} \exp\{-\frac{1}{2s^2_{\boldsymbol{\theta}}}(\boldsymbol{\theta} - \lambda)^2\} \\
&= \frac{1}{2} \frac{1}{2\pi s_{\boldsymbol{Y}} s_{\boldsymbol{\theta}}} \exp\{-\frac{1}{2g_{s_{\boldsymbol{\theta}}(s^2_{\boldsymbol{Y}}, s^2_{\boldsymbol{\theta}})}}(\boldsymbol{\theta} - g_{\boldsymbol{\theta}}(y, \lambda, s^2_{\boldsymbol{Y}}, s^2_{\boldsymbol{\theta}}))^2\} \\
&\times \exp\{-\frac{1}{2} g(y, \lambda, s^2_{\boldsymbol{Y}}, s^2_{\boldsymbol{\theta}})\} \\
&+ \frac{1}{2} \frac{1}{2\pi s_{\boldsymbol{Y}} s_{\boldsymbol{\theta}}} \exp\{-\frac{1}{2g_{s_{\boldsymbol{\theta}}(s^2_{\boldsymbol{Y}}, s^2_{\boldsymbol{\theta}})}}(\boldsymbol{\theta} - g_{\boldsymbol{\theta}}(-y, \lambda, s^2_{\boldsymbol{Y}}, s^2_{\boldsymbol{\theta}}))^2\} \\
&\times \exp\{-\frac{1}{2} g(-y, \lambda, s^2_{\boldsymbol{Y}}, s^2_{\boldsymbol{\theta}})\}.
\end{aligned} \tag{H.2}$$

### H.0.1 Analytical solution to the marginal likelihood in Scenario 2

$$\begin{aligned} P_{ana\_bi\_lik}(\boldsymbol{Y}) &= \int P_{bi\_lik}(\boldsymbol{\theta}, \boldsymbol{Y}) d\boldsymbol{\theta} \\ &= \frac{1}{2}\frac{1}{2\pi s_{\boldsymbol{Y}} s_{\boldsymbol{\theta}}} \exp\{-\frac{1}{2} g(y, \lambda, s^2_{\boldsymbol{Y}}, s^2_{\boldsymbol{\theta}})\}\sqrt{2\pi g_{s_{\boldsymbol{\theta}}}(s^2_{\boldsymbol{Y}}, s^2_{\boldsymbol{\theta}})} \\ &+ \frac{1}{2}\frac{1}{2\pi s_{\boldsymbol{Y}} s_{\boldsymbol{\theta}}} \exp\{-\frac{1}{2} g(-y, \lambda, s^2_{\boldsymbol{Y}}, s^2_{\boldsymbol{\theta}})\}\sqrt{2\pi g_{s_{\boldsymbol{\theta}}}(s^2_{\boldsymbol{Y}}, s^2_{\boldsymbol{\theta}})}. \end{aligned} \tag{H.3}$$

#### H.0.1.0.1 The multi-modality condition.

The 'multi-modality condition', $|\hat{\boldsymbol{\theta}}_1 - \hat{\boldsymbol{\theta}}_2| > 2\sqrt{g_{s_{\boldsymbol{\theta}}}(s^2_{\boldsymbol{Y}}, s^2_{\boldsymbol{\theta}})}$, becomes,

$$\left| \frac{\frac{y}{s^2_{\boldsymbol{Y}}} + \frac{\lambda}{s^2_{\boldsymbol{\theta}}}}{\frac{1}{s^2_{\boldsymbol{Y}}} + \frac{1}{s^2_{\boldsymbol{\theta}}}} - \frac{\frac{-y}{s^2_{\boldsymbol{Y}}} + \frac{\lambda}{s^2_{\boldsymbol{\theta}}}}{\frac{1}{s^2_{\boldsymbol{Y}}} + \frac{1}{s^2_{\boldsymbol{\theta}}}} \right| > \frac{2}{\sqrt{\frac{1}{s^2_{\boldsymbol{Y}}} + \frac{1}{s^2_{\boldsymbol{\theta}}}}}, \tag{H.4}$$

### H.0.2 The Laplace approximation in Scenario 2

The LA in bimodal case, uses only one of the two posterior means,

$$\hat{\boldsymbol{\theta}}_1 = g_{\boldsymbol{\theta}}(y, \lambda, s^2_{\boldsymbol{Y}}, s^2_{\boldsymbol{\theta}}); \hat{\boldsymbol{\theta}}_2 = g_{\boldsymbol{\theta}}(-y, \lambda, s^2_{\boldsymbol{Y}}, s^2_{\boldsymbol{\theta}}). \tag{H.5}$$

At the special case where the two modes are well separated i.e., $|\hat{\boldsymbol{\theta}}_1 - \hat{\boldsymbol{\theta}}_2| >> 6\sqrt{g_{s_{\boldsymbol{\theta}}}(s^2_{\boldsymbol{Y}}, s^2_{\boldsymbol{\theta}})}$, the LA evaluated at the first mode is exactly half of the analytical solution to the marginal likelihood integral in (4.2),

$$\begin{aligned} \exp\{-\frac{1}{2} LA(\hat{\boldsymbol{\theta}}_1)\} &= \frac{1}{2}\frac{\sqrt{g_{s_{\boldsymbol{\theta}}}(s^2_{\boldsymbol{Y}}, s^2_{\boldsymbol{\theta}})}}{2\pi s_{\boldsymbol{Y}} s_{\boldsymbol{\theta}}} \exp\{-\frac{1}{2} g(y, \lambda, s^2_{\boldsymbol{Y}}, s^2_{\boldsymbol{\theta}})\} \\ &= \frac{1}{2} P_{ana\_bi\_lik}(\boldsymbol{Y}). \end{aligned} \tag{H.6}$$

Analogously, the LA evaluated at $\hat{\boldsymbol{\theta}}_2$ is one half of $P_{ana\_bi\_lik}(\boldsymbol{Y})$.

### H.0.3 The MBIC in Scenario 2

The MBIC takes into account the two posterior modes given in the equation (H.5), and provided that the two modes are fairly widely separated i.e., $|\hat{\boldsymbol{\theta}}_1 - \hat{\boldsymbol{\theta}}_2| >> 6\sqrt{g_{s_{\boldsymbol{\theta}}}(s^2_{\boldsymbol{Y}}, s^2_{\boldsymbol{\theta}})}$, it

produces an exact solution to the marginal likelihood integral in (4.2),

$$\begin{aligned}
\exp\{-\frac{1}{2}MBIC(\hat{\boldsymbol{\theta}}_1,\hat{\boldsymbol{\theta}}_2)\} &= \frac{1}{2}\frac{\sqrt{g_{s_{\boldsymbol{\theta}}}(s^2_{\boldsymbol{Y}},s^2_{\boldsymbol{\theta}})}}{2\pi s_{\boldsymbol{Y}} s_{\boldsymbol{\theta}}}\exp\{-\frac{1}{2}g(y,\lambda,s^2_{\boldsymbol{Y}},s^2_{\boldsymbol{\theta}})\} \\
&+ \frac{1}{2}\frac{\sqrt{g_{s_{\boldsymbol{\theta}}}(s^2_{\boldsymbol{Y}},s^2_{\boldsymbol{\theta}})}}{2\pi s_{\boldsymbol{Y}} s_{\boldsymbol{\theta}}}\exp\{-\frac{1}{2}g(-y,\lambda,s^2_{\boldsymbol{Y}},s^2_{\boldsymbol{\theta}})\} \\
&= P_{ana_bi_lik}(\boldsymbol{Y}). \qquad \text{(H.7)}
\end{aligned}$$

### H.0.4 Bias in the MBIC, Scenario 2

The bias in MBIC arises when the modes are not well separated,

$$\begin{aligned}
\exp\{-\frac{1}{2}MBIC(\hat{\boldsymbol{\theta}}_1,\hat{\boldsymbol{\theta}}_2)\} &= \sum_{k=1}^{2} g(\hat{\boldsymbol{\theta}}_k,\boldsymbol{Y})n^{-\frac{q}{2}}|H(\hat{\boldsymbol{\theta}})|^{-\frac{1}{2}}\sqrt{2\pi} \\
&= \frac{1}{2}\frac{\sqrt{g_{s_{\boldsymbol{\theta}}}(s^2_{\boldsymbol{Y}},s^2_{\boldsymbol{\theta}})}}{2\pi s_{\boldsymbol{Y}} s_{\boldsymbol{\theta}}}\exp\{-\frac{1}{2}g(y,\lambda,s^2_{\boldsymbol{Y}},s^2_{\boldsymbol{\theta}})\} \\
&\times \underbrace{\exp\{-\frac{1}{2}\frac{(\hat{\boldsymbol{\theta}}_1-\hat{\boldsymbol{\theta}}_1)^2}{g_{s_{\boldsymbol{\theta}}}(s^2_{\boldsymbol{Y}},s^2_{\boldsymbol{\theta}})}\}}_{1} \\
&+ \frac{1}{2}\frac{\sqrt{g_{s_{\boldsymbol{\theta}}}(s^2_{\boldsymbol{Y}},s^2_{\boldsymbol{\theta}})}}{2\pi s_{\boldsymbol{Y}} s_{\boldsymbol{\theta}}}\exp\{-\frac{1}{2}g(-y,\lambda,s^2_{\boldsymbol{Y}},s^2_{\boldsymbol{\theta}})\} \\
&\times \underbrace{\exp\{-\frac{1}{2}\frac{(\hat{\boldsymbol{\theta}}_1-\hat{\boldsymbol{\theta}}_2)^2}{g_{s_{\boldsymbol{\theta}}}(s^2_{\boldsymbol{Y}},s^2_{\boldsymbol{\theta}})}\}}_{\approx 0,\ \text{when}\ |\hat{\boldsymbol{\theta}}_1-\hat{\boldsymbol{\theta}}_2|>>6\sqrt{g_{s_{\boldsymbol{\theta}}}(s^2_{\boldsymbol{Y}},s^2_{\boldsymbol{\theta}})}} \\
&+ \frac{1}{2}\frac{\sqrt{g_{s_{\boldsymbol{\theta}}}(s^2_{\boldsymbol{Y}},s^2_{\boldsymbol{\theta}})}}{2\pi s_{\boldsymbol{Y}} s_{\boldsymbol{\theta}}}\exp\{-\frac{1}{2}g(y,\lambda,s^2_{\boldsymbol{Y}},s^2_{\boldsymbol{\theta}})\} \\
&\times \underbrace{\exp\{-\frac{1}{2}\frac{(\hat{\boldsymbol{\theta}}_2-\hat{\boldsymbol{\theta}}_1)^2}{g_{s_{\boldsymbol{\theta}}}(s^2_{\boldsymbol{Y}},s^2_{\boldsymbol{\theta}})}\}}_{\approx 0,\ \text{when}\ |\hat{\boldsymbol{\theta}}_1-\hat{\boldsymbol{\theta}}_2|>>6\sqrt{g_{s_{\boldsymbol{\theta}}}(s^2_{\boldsymbol{Y}},s^2_{\boldsymbol{\theta}})}} \\
&+ \frac{1}{2}\frac{\sqrt{g_{s_{\boldsymbol{\theta}}}(s^2_{\boldsymbol{Y}},s^2_{\boldsymbol{\theta}})}}{2\pi s_{\boldsymbol{Y}} s_{\boldsymbol{\theta}}}\exp\{-\frac{1}{2}g(-y,\lambda,s^2_{\boldsymbol{Y}},s^2_{\boldsymbol{\theta}})\} \\
&\times \underbrace{\exp\{-\frac{1}{2}\frac{(\hat{\boldsymbol{\theta}}_2-\hat{\boldsymbol{\theta}}_2)^2}{g_{s_{\boldsymbol{\theta}}}(s^2_{\boldsymbol{Y}},s^2_{\boldsymbol{\theta}})}\}}_{1}. \qquad \text{(H.8)}
\end{aligned}$$

If the two modes are not well separated, i.e., $|\hat{\boldsymbol{\theta}}_1-\hat{\boldsymbol{\theta}}_2| > 2\sqrt{g_{s_{\boldsymbol{\theta}}}(s^2_{\boldsymbol{Y}},s^2_{\boldsymbol{\theta}})}$, and $|\hat{\boldsymbol{\theta}}_1-\hat{\boldsymbol{\theta}}_2| < 6\sqrt{g_{s_{\boldsymbol{\theta}}}(s^2_{\boldsymbol{Y}},s^2_{\boldsymbol{\theta}})}$,

$$\begin{aligned} \exp\{-\frac{1}{2}MBIC(\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2)\} &> P_{ana\_bi\_lik}(Y) + \exp(-2)P_{ana\_bi\_lik}(\boldsymbol{Y}) \\ &= (1 + \exp(-2))P_{ana\_bi\_lik}(\boldsymbol{Y}), \end{aligned} \tag{H.9}$$

where $P_{ana\_bi\_lik}(\boldsymbol{Y})$ is given in (H.3).

The condition for well separated modes, which states that the distance between the two posterior modes are bigger than six standard deviations is chosen such that the second and third terms in (H.8) has near-zero value.

# Appendix I

# Discovering all the important modes

The efficiency of the MBIC relies on the information from all the relevant posterior modes, while avoiding the full Bayesian approach to the posterior exploration. To solve the problem of discovering all the important posterior modes, we propose an optimization strategy which combines results from the optimization initialized at different points. The idea relies on the notion that different initial points lead the optimizer to exploring different regions of the posterior distribution, thus discovering different posterior modes. The proposed strategy is closely related to the Incremental Mixture Importance Sampling with Optimization (IMIS-Opt) Raftery and Bao (2010) algorithm, which combines importance sampling with optimization to iteratively build the target importance distribution, and to fully explore the posterior space thereof. Here, the goal is to build an optimization strategy for discovering all the relevant posterior modes, rather than sampling from the posterior distribution. We refer to the proposed optimization strategy as Shotgun optimization (ShOpt).

The initial stage of the ShOpt starts with drawing $N_0$ samples from the prior distribution, $P(\boldsymbol{\theta})$, followed by calculating the weights for each sample using the likelihood function. Then, in the optimization stage, the ShOpt sequentially chooses highest weight points to initialize the optimizer, which leads to exploring different regions of the posterior space. At each iteration, after a mode is found, samples from the prior that are within the basin of attraction of the newly discovered mode are discarded, thus enabling the algorithm to discover a new mode.

In order for the ShOpt to discover all the label switching modes, at each iteration after a new mode is discovered, the corresponding $K!$ modes are found by permuting the newly discovered mode. Then, samples from the prior that are within the basin of attraction of the $K!$ newly discovered modes are discarded based on the Mahalanobis distance, and the algorithm continues to explore the remainder of the target posterior space.

Although the samples from within the basin of attraction of the identified modes are excluded, the optimizer initialized at the points from the prior located close to some of excluded modes, might lead to finding modes that have already been discovered. To solve the problem of duplicate modes, the ShOpt takes additional step to decide whether the newly discovered mode and its corresponding label switching modes are duplicates. The decision

rule is constructed using pairwise Mahalanobis distances between the $K!$ label switching newly discovered modes and the previously discovered modes. Since the Mahalanobis distance is not a symmetric measure, the distances in both directions between the new modes and the previously discovered modes, are calculated. In the first direction, the Mahalanobis distances, $MA_1$, are calculated using the inverse negative Hessian matrices of the new modes, while in the other direction, the Mahalanobis distances, $MA_2$, are calculated using the inverse negative Hessian matrices of the previously discovered modes. The $K!$ newly discovered modes are added to the set of all discovered modes if all element-wise minimum distances of $MA_1$ and $MA_2$ are larger than two standard deviations. The pseudo code of the ShOpt is presented in Algorithm 7.

**Algorithm 7** The Shotgun optimization (ShOpt)

**Goal: Discover all the important posterior modes.**

**Input:** Data, the likelihood $P(\boldsymbol{Y} \mid \boldsymbol{\theta})$ and the prior density $P(\boldsymbol{\theta})$. $N_0$ - the number of samples from the prior and $K$ - the number of components in the mixture model.

**Initial stage:** Draw $N_0$ samples $\boldsymbol{\Theta}_0 = \{\boldsymbol{\theta_1}, \boldsymbol{\theta_2}, ..., \boldsymbol{\theta_{N_0}}\}$ from the prior distribution $P(\boldsymbol{\theta})$.

For each $\boldsymbol{\theta}_i \in \boldsymbol{\Theta}_0$ calculate the sampling weights:

$$w_i^{(1)} = \frac{P\left(\boldsymbol{Y} \mid \boldsymbol{\theta_i}\right)}{\sum\limits_{j=1}^{N_0} P(\boldsymbol{Y} \mid \boldsymbol{\theta_j})}. \tag{I.1}$$

**Optimization stage:**

Set a counter for the number of discovered modes, $N_m = 0$.

Set the counter for the while loop d=1.

**do**

Use $\boldsymbol{\theta}^{initial} = \underset{\boldsymbol{\theta}}{\text{argmax}}\ \boldsymbol{w}^{(1)}(\boldsymbol{\theta})$, $\boldsymbol{\theta} \in \boldsymbol{\Theta}_{d-1}$, to initialize the optimizer and get local posterior maxima $\boldsymbol{\theta}_d^{(Opt)} = \underset{\boldsymbol{\theta}}{\text{argmax}}\ P(\boldsymbol{\theta} \mid \boldsymbol{Y}), \boldsymbol{\theta} \in \boldsymbol{\Theta}_{d-1}$ along with the corresponding inverse negative Hessian, $\boldsymbol{\Sigma}_d^{(Opt)}$.

Find all $K!$ permutations of the newly discovered mode, $\boldsymbol{\theta}_{d,k}^{(Opt)}$, and obtain their respective inverse negative Hessian matrices, $\boldsymbol{\Sigma}_{d,k}^{(Opt)}$.

**for** $k = 1 : K!$ **do**

Update $\boldsymbol{\Theta}_d$ by excluding nearest neighbor points, $\boldsymbol{\theta}_l \in \boldsymbol{\Theta}_{d-1}$, with Mahalanobis distance less than 2 standard deviations from the newly discovered mode,

$$(\boldsymbol{\theta}_l - \boldsymbol{\theta}_{d,k}^{(Opt)})'(\boldsymbol{\Sigma}_{d,k}^{(Opt)})^{-1}(\boldsymbol{\theta}_l - \boldsymbol{\theta}_{d,k}^{(Opt)}) < 4. \tag{I.2}$$

**end for**

**Decide whether to accept the newly discovered $K!$ modes.**

**if** d=1 **then**

Add the $K!$ newly discovered modes the and their corresponding inverse negative Hessian matrices to the set of discovered modes $\mathcal{D} = \{\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\Sigma}}\}$.

Set $N_m = N_m + K!$.

**else**

**for** $k = 1 : K!$ **do**

**for** $m = 1 : N_m$ **do**

Calculate Mahalanobis distance, $\boldsymbol{MA_1}$, between the $k$-th permutation of the newly discovered mode and the $m$-th previously discovered mode,

$$\boldsymbol{MA_1}_{k,m} = (\hat{\boldsymbol{\theta}}_m - \boldsymbol{\theta}_{d,k}^{(Opt)})'(\boldsymbol{\Sigma}_{d,k}^{(Opt)})^{-1}(\hat{\boldsymbol{\theta}}_m - \boldsymbol{\theta}_{d,k}^{(Opt)}). \tag{I.3}$$

Calculate Mahalanobis distance, $\boldsymbol{MA_2}$, between the $m$-th previously discovered mode and the $k$-th permutation of the newly discovered mode,

$$\boldsymbol{MA_2}_{k,m} = (\boldsymbol{\theta}_{d,k}^{(Opt)} - \hat{\boldsymbol{\theta}}_m)'(\hat{\boldsymbol{\Sigma}}_m)^{-1}(\boldsymbol{\theta}_{d,k}^{(Opt)} - \hat{\boldsymbol{\theta}}_m). \tag{I.4}$$

**Algorithm 7** The Shotgun optimization (ShOpt) - continued

Find the minimum Mahalanobis distance,

$$\boldsymbol{MA}_{k,m} = \min\{\boldsymbol{MA}_{\boldsymbol{1}k,m}, \boldsymbol{MA}_{\boldsymbol{2}k,m}\}. \tag{I.5}$$

**end for**
**end for**
**if** Each element in $\boldsymbol{MA}$ is larger than 4 **then**
Add the $K!$ newly discovered modes and their corresponding inverse negative Hessian matrices to the set of all discovered modes $\mathcal{D} = \{\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\Sigma}}\}$.
Set $N_m = N_m + K!$.
**else** Discard the $K!$ newly discovered modes.
**end if**
**end if**
d=d+1;
**while** $\boldsymbol{\Theta}_{d-1} \neq \emptyset$
**Output:** The set of all discovered modes and their corresponding inverse Hessian matrices, $\mathcal{D} = \{\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\Sigma}}\}$.