

DEPARTMENT OF
STATISTICS

Monte Carlo Algorithms for Bayesian Inference

A thesis submitted in partial fulfilment of the requirements for the degree Doctorate of Philosophy in Statistics

December 2021

Emilia Pompe
St Peter's College

Department of Statistics
University of Oxford

Contents

1 Introduction and literature review **7**

1 Literature review on Adaptive MCMC 9
1 .1 The Metropolis–Hastings algorithm 9
1 .2 How to adapt the Metropolis algorithm? 11
1 .3 Theoretical properties of adaptive MCMC algorithms 13
1 .4 Theoretical contribution (Chapter 2) 16
2 Literature review on MCMC methods for targeting multimodal distributions 17
2 .1 Poor mixing of standard Markov chain Monte Carlo methods on multimodal target distributions 17
2 .2 Tempering-based methods 18
2 .3 Metropolis-Hastings algorithms with a special design of the proposal distribution 21
2 .4 Optimization-based methods 23
2 .5 Other methods 24
2 .6 Methodological contribution (Chapter 2) 25
3 Literature review on Bayesian inference under model misspecification 26
3 .1 Asymptotic results for standard Bayesian posteriors 26
3 .2 General methods of robust Bayesian inference 27
3 .3 Contribution (Chapter 3) 31
3 .4 Cutting feedback in Bayesian models 33
3 .5 Contribution (Chapter 4) 36

2 A Framework for Adaptive MCMC Targeting Multimodal Distributions **37**

1 Introduction 38
1 .1 Other approaches 39
1 .2 Contribution 40
2 Jumping Adaptive Multimodal Sampler (JAMS) 42
2 .1 Main algorithm 42
2 .2 Burn-in algorithm 46
2 .3 Further comments 48
3 Auxiliary Variable Adaptive MCMC 49
3 .1 Theoretical results for the class 50
3 .2 Adaptive Increasingly Rarely version of the class 54
4 Ergodicity of the Jumping Adaptive Multimodal Sampler 54
4 .1 Overview of the assumptions 54
4 .2 Theoretical results for mJAMS 55
5 Examples 60
5 .1 Mixture of Gaussians 60
5 .2 Mixture of t and banana-shaped distributions 62
5 .3 Sensor network localization 65
6 Summary and discussion 67

A Supplementary Material: theory . . . 68
A .1 Proofs for Section 3 . . . 69
A .2 Proofs for Section 4 . . . 81
A .3 Other algorithms in the Auxiliary Variable Adaptive MCMC class . . . 96
A .4 Spectral gap of a single JAMS kernel . . . 97
B Supplementary Material: methodology and numerical examples . . . 102
B .1 Updating $w_{\gamma,i}$ and $a_{\gamma,ik}$ in the main algorithm . . . 102
B .2 Burn-in algorithm . . . 103
B .3 Examples – further details . . . 106

3 Introducing prior information in Weighted Likelihood Bootstrap with applications to model misspecification 116
1 Introduction . . . 117
1 .1 Model misspecification and robust inference in Bayesian statistics . . . 117
1 .2 Properties of Weighted Likelihood Bootstrap . . . 118
1 .3 Edgeworth expansions . . . 121
2 Posterior Bootstrap via penalization on the weighted log-likelihood . . . 123
2 .1 Methodology for non-hierarchical models . . . 123
2 .2 Asymptotic results for Algorithm 5 . . . 124
2 .3 Choice of w_0 in Algorithm 5 . . . 129
3 Extending Posterior Bootstrap to hierarchical models . . . 132
3 .1 Fixed number of groups K . . . 132
3 .2 Number of groups K going to infinity . . . 135
4 Posterior Bootstrap via pseudo-samples from the prior predictive . . . 137
4 .1 Methodology . . . 137
4 .2 Choice of c in Algorithm 7 . . . 139
5 Illustrations . . . 140
5 .1 Toy examples . . . 140
5 .2 Overdispersed and underdispersed Poisson regression . . . 142
5 .3 Dirichlet allocation model for opioid crisis data . . . 144
6 Limitations of methods based on Weighted Likelihood Bootstrap . . . 146
A Proof of Theorem 3.2 . . . 148
B Proofs of theoretical results presented in Sections 2 and 4 . . . 150
B .1 Summary and Assumption 3.6 . . . 150
B .2 Notation . . . 151
B .3 Auxiliary results . . . 152
B .4 Proof of Theorem 3.10 . . . 155
B .5 Proof of Theorem 3.17 . . . 160
B .6 Edgeworth expansions for other methods based on Weighted Likelihood Bootstrap . . . 161
C Proofs of theoretical results presented in Section 3 . . . 163
C .1 Proof of Proposition 3.14 . . . 163
C .2 Additional information for Section 3 .2 and proof of Theorem 3.15 . . . 164
D Illustrations – further details . . . 175
D .1 Additional information for Section 5 .1 . . . 175
D .2 Toy example of a hierarchical model . . . 177
D .3 Additional information on the Dirichlet allocation model . . . 181

4 Asymptotics of cut distributions and robust modular inference using Posterior Bootstrap **183**
1 Introduction . . . 184
1 .1 Cutting feedback in Bayesian models . . . 184
1 .2 Computational challenges of Bayesian modular inference . . . 186
1 .3 Related work . . . 187
1 .4 Contribution . . . 188
2 Asymptotics of Bayesian versus point estimation in modular settings . . . 189
2 .1 Scenario 1 . . . 191
2 .2 Scenario 2 . . . 192
2 .3 Laplace approximation for the cut posterior . . . 193
3 Methodology and theoretical results . . . 194
3 .1 Algorithms and asymptotic results for modular inference . . . 194
3 .2 Prediction under Posterior Bootstrap for Modular Inference . . . 197
3 .3 Incorporating prior information . . . 199
4 Illustrations . . . 200
4 .1 Toy example . . . 200
4 .2 Biased data . . . 201
4 .3 Causal inference with propensity scores . . . 203
4 .4 Epidemiological study . . . 204
5 Discussion . . . 207
A Proofs of theoretical results . . . 208
A .1 Assumptions . . . 208
A .2 Proof of Proposition 4.3 . . . 210
A .3 Proofs of Theorems 4.4 and 4.5 . . . 211
A .4 Prediction under Posterior Bootstrap for Modular Inference versus Bayesian modular inference . . . 214
A .5 Extensions to semi-modular inference . . . 218

5 Conclusions and discussion **223**

Bibliography **227**

Abstract

This thesis addresses several issues appearing in Bayesian statistics. Firstly, computations for approximating Bayesian posteriors are often performed using Markov chain Monte Carlo (MCMC) methods. However, standard MCMC algorithms tend to perform poorly when the posterior distribution has multiple modes. Secondly, performance of MCMC methods heavily depends on the choice of hyperparameters. It is therefore beneficial to design adaptive methods which update the hyperparameters on the fly, as the algorithm runs. Another topic explored in this thesis is associated with potential misspecification of the considered Bayesian model. It has been well-understood that standard Bayesian inference, when applied to misspecified models, leads to misleading uncertainty quantification. This issue motivates departures from the standard Bayesian inference, and looking into more robust approaches.

In Chapter 1 we provide an introduction to the thesis and the literature review covering the above topics.

In Chapter 2, based on Pompe et al. (2020), we propose a new Monte Carlo method for sampling from multimodal distributions. The idea of this technique is based on splitting the task into two: finding the modes of a target distribution π and sampling, given the knowledge of the locations of the modes. The sampling algorithm relies on steps of two types: local ones, preserving the mode; and jumps to regions associated with different modes. Moreover, the method learns the optimal parameters of the algorithm as it runs, without requiring user intervention.

In order to design an adaptive scheme that facilitates both local and jump moves, we introduce an auxiliary variable representing each mode and we define a new target distribution $\tilde{\pi}$ on an augmented state space $\mathcal{X} \times \mathcal{I}$, where $\mathcal{X}$ is the original state space of π and $\mathcal{I}$ is the set of the modes. As the algorithm runs and updates its parameters, the target distribution $\tilde{\pi}$ also keeps being modified. This motivates a new class of algorithms, Auxiliary Variable Adaptive MCMC, generalizing Adaptive MCMC. We prove ergodic results for the class and apply them to our specific algorithm.

Chapter 3 is based on Pompe (2021). We propose Posterior Bootstrap, a set of algorithms extending Weighted Likelihood Bootstrap, to properly incorporate prior information and address the problem of model misspecification in Bayesian inference. We consider two approaches to incorporating prior knowledge: the first is based on penalization of the Weighted Likelihood Bootstrap objective function, and the second uses pseudo-samples from the prior predictive distribution. We also propose methodology for hierarchical models, which was not previously considered for methods based on Weighted Likelihood Bootstrap. We use Edgeworth expansions to gain new insight on properties of Weighted Likelihood Bootstrap and its extensions, as well as to guide the choice of hyperparameters in our algorithms.

In Chapter 4, based on Pompe and Jacob (2021), we focus on Bayesian models built of several components, called modules. Under standard Bayesian inference, misspecification of any module might propagate to other parts of the model, leading to unsatisfactory results. Cut distributions have been proposed as a remedy, where the information is prevented from flowing along certain directions. We consider cut distributions from an asymptotic perspective, find the equivalent of the Laplace approximation, and notice a lack of frequentist coverage for the associated credible regions. We propose algorithms based on the Posterior Bootstrap that deliver credible regions with the nominal frequentist asymptotic coverage.

In Chapter 5 we provide a discussion and indicate potential future research directions.

Acknowledgements

First of all, I would like to thank my primary supervisor, Chris Holmes. Over the course of my PhD I have become much more independent, and in many ways stronger. Certainly, Chris has heavily influenced the path that got me here. Moreover, I am really grateful that he pointed me to the literature on Weighted Likelihood Bootstrap, which opened for me a completely new research area.

I also want to express my gratitude to my co-supervisor, Judith Rousseau, for her patience in explaining the intricacies of asymptotic statistics. I was fortunate to have learnt so much from Judith.

Special thanks to Pierre Jacob, who is a great collaborator and a truly inspiring academic. I was extremely fortunate to have had an opportunity to visit Pierre's group at Harvard. It is fair to say that Pierre's enthusiasm for statistics revived my then slightly fading away passion for academic research. I also want to thank members of Pierre's lab, in particular Phyllis Ju, Niloy Biswas and Gonzalo Mena, who made me, an alien from Oxford, feel so welcome in the group.

My thanks are extended to my other collaborators, Luke Kelly and Krzysztof Łatuszyński, as well as the examiners of my PhD milestones, Geoff Nicholls, Arnaud Doucet and George Deligiannidis. I am grateful for everything I have learnt from them.

Looking back, I am indebted to those who inspired me to work on statistics. Special thanks to Jeffrey Collamore, whose course on Monte Carlo methods at the University of Copenhagen was my favourite academic course ever. Silly as it sounds, I fell in love with Monte Carlo methods... and that's how this journey started.

I was fortunate to have met during my PhD wonderful people, whom I now call friends. Many thanks to Petya Kindalova, who consumed with me more coffee than any human could handle, and rescued me countless times in the past few years, mainly from my own silliness. A huge thank you also to Chris Carmona, Habib Ganjgahi and Robert Pisarczyk. I already miss hanging out with them in Oxford.

I also thank the entire Oxford University Dancesport Team 2017/2018, and the Head Coach Bruce Richardson. The amazing dance classes were always making grey pre-pandemic Oxford days brighter, and in fact, were one of my favourite things about Oxford.

I am deeply indebted to my friends, Marek Stryjczyk a.k.a. Stryjczak, Iza Kanonowicz and Szymon Kanonowicz, who have been an enormous source of support throughout my PhD. If there exists a certificate of excellence in listening about PhD-related issues, they should all definitely get it. Also, kudos to Stryjczak, who managed to illustrate every single situation in my life with a meme or a gif.

I am also grateful to my family, in particular my Mom, Dad, and Kuba, for believing in me since I was a child. Special thanks to Kuba for subtly changing the topic whenever our parents were asking the most uncomfortable question: 'When are you submitting your thesis, Emilka?'.

Above all, I want to thank my partner, Romek Stasiński, for being the most wonderful source of support I could have imagined. Without Romek, I would have had neither the courage to start this journey, nor the determination to go through the most emotionally challenging moments. Also, it should be crystal clear that Romek's and Stryjczak's amazing sense of humour is what saved me in the most difficult times.

Finally, I am grateful to those who are not mentioned above, but who helped me develop the crucial skill of setting healthy boundaries.

I am leaving behind a few chapters of work, as well as an important, formative and in many ways unexpected chapter of life. And looking forward to the next one!

I gratefully acknowledge the financial support of the EPSRC and the Clarendon Fund.

Honesty and frankness make you vulnerable. Be honest and frank anyway.

Kent M. Keith

Each time a woman stands up for herself, without knowing it possibly, without claiming it, she stands up for all women.

Maya Angelou

Chapter 1

Introduction and literature review

Bayesian inference is gaining more and more popularity among applied statisticians in various fields. A parametric Bayesian model in its simplest form, with a parameter of interest θ, is given by

$$x_1, \ldots, x_n \sim f(x|\theta), \quad \theta \sim \pi(\theta), \quad \theta \in \Theta, \tag{1.1}$$

where f is the likelihood function and π denotes the prior on θ. We let $x_{1:n}$ denote the vector of observations $x_1, \ldots, x_n$. The density of the posterior distribution on θ given the observations is

$$\pi(\theta|x_{1:n}) = \frac{\pi(\theta)\prod_{i=1}^n f(x_i|\theta)}{\int \pi(\theta)\prod_{i=1}^n f(x_i|\theta)\mathrm{d}\theta}. \tag{1.2}$$

Bayesian inference could be seen as updating the belief about θ, from the prior to the posterior distribution, through an iterative use of Bayes' theorem, as more and more data become available. One of its main advantages is flexibility, which allows the user to fit complex realistic models to the data at hand. Furthermore, the Bayesian approach provides a natural way of incorporating into the analysis the user's prior beliefs about the parameter, or the expert knowledge, which is formally expressed via the prior distribution π. Bayesian inference has been successfully used for analysing datasets of various types, such as censored observations in survival analysis (Ibrahim et al., 2014), grouped data (using hierarchical models (Gelman et al., 2013)), time series (Barber et al., 2011), and many others. The use of Bayesian hierarchical modelling has been recently acclaimed as one of the eight most important statistical ideas of the past 50 years (Gelman and Vehtari, 2021). Moreover, the Bayesian approach provides a natural way of expressing uncertainty around parameters, even when the sample size is small, or the model has a complex structure.

The interest often lies in approximating integrals of the form $\int g(\theta)\pi(\theta|x_{1:n})\mathrm{d}\theta$, for some function g. If it is possible to draw samples directly from $\pi(\theta|x_{1:n})$, the estimation of this integral can be performed using the standard Monte Carlo approach. However, typically sampling directly from the posterior distribution is impossible, and therefore Markov chain Monte Carlo (MCMC) algorithms are employed to draw iteratively a sequence of samples, which are then used to approximate integrals of interest. In fact, the development of efficient MCMC methods has been one of the vital factors in popularizing the use of Bayesian methods observed in recent years.

Despite major developments in Bayesian theory, methodology and computation, many aspects of Bayesian statistics still remain a considerable challenge. This thesis addresses three problematic areas.

Firstly, standard MCMC methods tend to perform poorly when the posterior distribution has multiple modes. Moreover, their performance often depends on the hyperparameters of the MCMC method at hand. It is therefore beneficial to design adaptive methods which update the hyperparameters on the fly, as the algorithm runs. Secondly, the nature of MCMC methods is sequential, which hinders the benefit of having access to multiple cores and being able to run computations in parallel. Therefore, it is useful to design such algorithms for Bayesian inference, in which at least some of the computations can be performed in parallel.

Another problematic area arises from the potential misspecification of the considered model. It has been well-understood that standard Bayesian inference, when applied to misspecified models, leads to misleading uncertainty quantification (Kleijn and Van der Vaart, 2012). This explains the recent interest in methods that are more robust to misspecification while preserving the benefits of standard Bayesian inference.

In the remainder of this chapter, we review the literature on theoretical aspects of Adaptive MCMC, on MCMC methods for targeting multimodal distributions, and on robust Bayesian inference. We also summarize our main contributions, situating them in a wider literary context.

1 Literature review on Adaptive MCMC

1 .1 The Metropolis–Hastings algorithm

Markov chain Monte Carlo methods, dating back to Metropolis and Ulam (1949), rely on generating a Markov chain which converges to the target distribution of interest. In Sections 1 and 2 we will denote by π the target distribution, as well as its density (with a slight abuse of notation introduced in the previous section, where π was the density of the prior). By $P(x, \cdot)$ we will denote the transition kernel of the Markov chain at state $x \in \mathcal{X} \subseteq \mathbb{R}^d$, reversible with respect to π, that is

$$\pi(dx)P(x, dy) = \pi(dy)P(y, dx), \quad x, y \in \mathcal{X}.$$

Consider a ϕ-irreducible and aperiodic Markov chain on a state space $\mathcal{X}$, admitting a stationary distribution π (for definitions of ϕ-irreducibility and aperiodicity we refer the reader to Section 3.2 of Roberts and Rosenthal (2004)). We then have

$$\lim_{n\to\infty} \|P^n(x, \cdot) - \pi(\cdot)\|_{TV} = 0, \quad \pi\text{-a.e., } x \in \mathcal{X},$$

where $\|\cdot\|_{TV}$ denotes the total variation distance. Under the same conditions the following strong law of large numbers holds. Suppose that for some function g we have $\int g(x)\pi(x)dx < \infty$. Then

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} g(X_i) = \int g(x)\pi(x)dx,$$

with probability 1, where $X_1, X_2, \ldots$ are subsequent iterations of the Markov chain (see, for example, Theorem 4 and Fact 5 of Roberts and Rosenthal (2004)).

The Metropolis–Hastings (MH) algorithm (Metropolis et al., 1953; Hastings, 1970) is an important class of MCMC methods. Given a target distribution π and a family of proposal distributions $Q(x, \cdot)$ with their corresponding densities $q(x, \cdot)$, it generates a Markov chain as follows. If the current state of the chain is x, a proposal y for the next iteration is drawn from $Q(x, \cdot)$. This proposal is then accepted with probability

$$\alpha(x, y) = \min\left[1, \frac{\pi(y)q(y, x)}{\pi(x)q(x, y)}\right], \tag{1.3}$$

and rejected with probability $1 - \alpha(x, y)$. An important property of the MH algorithms is

that it is enough to be able to evaluate $\pi(x)$ up to the normalizing constant, as the constant cancels in the acceptance probability (1.3). This makes the MH algorithms particularly useful for targetting posteriors (1.2) in Bayesian models. It can be verified that such an algorithm is invariant with respect to π. Moreover, as long as $q(x, y)$ is positive as a function on $\mathcal{X} \times \mathcal{X}$, and the density $\pi(x)$ is finite everywhere, the resulting Markov chain is aperiodic and ϕ-irreducible, hence ergodic.

A well-studied special case of the Metropolis–Hastings methodology is the Random Walk Metropolis (Metropolis and Ulam, 1949), where the proposal distribution is symmetric, i.e. $q(x, y) = q(y, x)$. A typical choice for the proposal distribution at state x is $N(x, \Sigma)$ for some covariance matrix Σ. Other examples include random scan Gibbs samplers (Geman and Geman, 1984; Gelfand and Smith, 1990), Metropolis-adjusted Langevin algorithm (MALA) (Roberts and Tweedie, 1996) and Hamiltonian (Hybrid) Monte Carlo (HMC) (Duane et al., 1987; Neal, 2011).

We now define notions of uniform and geometric ergodicity, useful for analysing the rate of convergence to the limiting distribution. We say that a Markov chain is uniformly ergodic if

$$\|P^n(x, \cdot) - \pi(\cdot)\|_{TV} < M\rho^n \quad \pi\text{-a.e. } x \in \mathcal{X}, \quad n = 1, 2, \ldots,$$

for some $\rho \in (0, 1)$, $M < \infty$. In practice, there are few uniformly ergodic MCMC algorithms on non-compact state spaces. A weaker condition of geometric ergodicity is satisfied when

$$\|P^n(x, \cdot) - \pi(\cdot)\|_{TV} < M(x)\rho^n \quad \pi\text{-a.e. } x \in \mathcal{X}, \quad n = 1, 2, \ldots,$$

for some $\rho \in (0, 1)$, and a function $M(x) < \infty$ for π-a.e. $x \in \mathcal{X}$. We now define two conditions, closely linked with uniform and geometric ergodicity.

We say that a Markov chain satisfies a minorization condition if there exist a positive integer n_0, $\epsilon > 0$, a set $C \subseteq \mathcal{X}$, and a probability measure ν on $\mathcal{X}$ such that

$$P^{n_0}(x, \cdot) \geq \epsilon\nu(\cdot), \quad x \in C. \tag{1.4}$$

The set C is then called a small set for kernel P. If the whole state space $\mathcal{X}$ is small, then the chain is uniformly ergodic.

We say that a Markov chain satisfies a drift condition towards a small set C if for some

$0 < \lambda < 1$, $b < \infty$ and a function $V(x) : \mathcal{X} \to [1, \infty)$, we have

$$\int_{\mathcal{X}} P(x, dy)V(y) \leq \lambda V(x) + b\mathbb{I}_C(x), \quad x \in \mathcal{X}. \tag{1.5}$$

If a ϕ-irreducible, aperiodic Markov chain is invariant with respect to π, and satisfies conditions (1.4) and (1.5), it is geometrically ergodic (see for example Theorem 9 of Roberts and Rosenthal (2004), Chapter 15 of Meyn and Tweedie (2012)). We shall see in Sections 1 .3 and 1 .4 that the drift and minorization conditions are useful for establishing ergodicity of adaptive MCMC algorithms. Conditions required for geometric ergodicity of Metropolis algorithms were studied in detail by Jarner and Hansen (2000).

Even a geometrically ergodic chain may in practice mix very slowly. Its performance crucially depends on the design parameters of the given MCMC algorithm (see, for example Gelman et al. (1997); Gilks et al. (1998); Roberts and Rosenthal (2004)), such as the proposal covariance matrix in the case of Random Walk Metropolis. Intuitively, if too large variance is used, the proposed states will often lie within low probability regions, and therefore very few moves will be accepted. If, on the other hand, the steps are too small, we will observe a high acceptance rate, but the small increments will result in a slow exploration of the state space. Therefore, it is of interest to design adaptive algorithms that update their design parameter as they run, based on the history of the chain. This raises two vital questions, a methodological and a theoretical one.

1. How to adapt the parameters to improve the mixing of the chain?

2. Updating the parameters based on the entire history of the chain, rather than the most recent iteration, breaks the Markov property of the resulting process. Therefore, how to ensure that the resulting algorithm converges to the target of interest?

In Sections 1 .2 and 1 .3 we describe how these two questions have been addressed in the literature. In terms of methodology, we mainly focus on the Random Walk Metropolis algorithm.

1 .2 How to adapt the Metropolis algorithm?

The result on the optimal scaling of the Random Walk Metropolis algorithm was obtained by Gelman et al. (1997). In particular, the authors considered a target distribution $\pi_d(x)$ of a

product form $\prod_{j=1}^{d} f(x_j)$ and the proposal distribution $N(0, \sigma_d^2 I_d)$ for $\sigma_d > 0$. They showed that if Z^d is a speeded-up continuous-time version of the first coordinate of the Markov chain resulting from the Metropolis algorithm, then Z^d converges weakly to a certain diffusion process as $d \to \infty$. It turns out that the speed of this diffusion is maximized when the average acceptance rate is around 0.234. A consequence of this result is that the optimal covariance matrix in the Metropolis algorithm with Gaussian proposals is given by $2.38^2\Sigma/d$, where Σ is the covariance matrix of π.

Even though the constant 0.234 was obtained as an asymptotic result, Roberts and Rosenthal (2001) noticed that even for dimension $d = 5$ the optimal acceptance rate is very close to 0.234. As shown in Roberts and Rosenthal (2001), in dimension $d = 1$ the optimal acceptance rate is 0.44. Analogous asymptotic results about the optimal acceptance rate were also obtained for MALA (Roberts and Rosenthal, 2001) and HMC (Beskos et al., 2013).

The above results have served as a commonly used heuristic for tuning Metropolis algorithms, and gave rise to successful adaptive techniques. We present below two main ideas, noting that many modifications of these methods have been discussed in the literature. The adaptive scaling algorithm relies on increasing or decreasing the step size depending on the mean acceptance rate in the recent batch, consisting of a fixed number of iterations. Let α^* denote the optimal acceptance rate for a given dimension. Consider a Gaussian proposal distribution for batch n with a covariance matrix $\sigma_n^2 I_d$, and let α_n be the mean acceptance rate in batch n. The adapting scheme sets

$$\log \sigma_{n+1} = \log \sigma_n + \gamma_n(\alpha_n - \alpha^*),$$

where $(\gamma_i)_{i=1}^{\infty}$ is a deterministic sequence of positive numbers, such that $\sum_{i=1}^{\infty} \gamma_i = \infty$ and $\sum_{i=1}^{\infty} \gamma_i^{1+\lambda} < \infty$ for some $\lambda > 0$ Andrieu and Thoms (2008). Alternatively, one could set

$$\log \sigma_{n+1} = \log \sigma_n + \min\left[0.01, n^{-1/2}\right] \mathbb{I}_{\alpha_n > \alpha^*} - \min\left[0.01, n^{-1/2}\right] \mathbb{I}_{\alpha_n < \alpha^*},$$

as proposed by Roberts and Rosenthal (2009). This strategy could be used globally with α^* typically set to 0.234, or separately for each component with $\alpha^* = 0.44$ (Haario et al., 2005; Andrieu and Thoms, 2008; Roberts and Rosenthal, 2009).

Another idea (Haario et al., 1999, 2001; Roberts and Rosenthal, 2009) is to propose state

$n+1$ from a mixture of Gaussian distributions

$$(1-\beta)N\left(x, 2.38^2\Sigma_n/d\right) + \beta N\left(x, 0.01 I_d/d\right), \tag{1.6}$$

for a small constant $\beta \in (0,1)$, where x denotes the state at iteration n, and Σ_n is the empirical estimate of the covariance matrix of the target distribution based on the previous iterations. We use (1.6), rather than simply the Gaussian distribution $N\left(x, 2.38^2\Sigma_n/d\right)$, in order to bound away from 0 the variance of the proposal distribution. This condition helps ensure ergodicity of the adaptive algorithm. We discuss ergodicity of adaptive MCMC methods in more detail in Section 1 .3.

For a survey of other adaptation strategies for RWM algorithms, including state-dependent or region-dependent adaptation, we refer the reader to Andrieu and Thoms (2008); Roberts and Rosenthal (2009).

1 .3 Theoretical properties of adaptive MCMC algorithms

As mentioned earlier, one difficulty of designing adaptive MCMC algorithms is ensuring that they have the correct limiting distribution. There are cases where adaptation breaks the ergodicity of the chain, even though all the kernels are ergodic with the same limiting distribution, see Example 1 considered by Bai et al. (2011b). We present below various approaches used in the literature to establish ergodicity of adaptive MCMC methods.

The Adaptive Metropolis method considered by Haario et al. (1999) proposed a new state at iteration $n+1$ using a covariance matrix $\Sigma_{n+1} = \mathrm{cov}\left(X_n, \ldots, X_{n-H}\right)$ for a fixed memory parameter H. This trick allows for updating the transition kernel, while preserving the Markov property.

The next strategy that allows for updating the parameters on the fly, advocated by Gilks et al. (1998); Brockwell and Kadane (2005), is based on the following property of Markov chains. Suppose that an irreducible Markov chain satisfies the minorization condition (1.4). Then the Markov chain is regenerative (Nummelin, 1978; Mykland et al., 1995), that is, it has random times $T_1, T_2, \ldots$, such that the tours of the process between T_i and T_{i+1} for $i = 1, 2, \ldots$, are i.i.d. Hence, if updates of the parameter are made at regeneration times, the law of large numbers holds (for details see Gilks et al. (1998)). Even though there are ways of identifying regenerations for many MCMC algorithms (Mykland et al., 1995), this method

has several drawbacks. Firstly, regenerations rarely happen in large dimensions, which makes the adaptive scheme prohibitively inefficient. Additionally, identifying regeneration times using the method of Mykland et al. (1995) requires case-specific calculations, which precludes any generic implementation of an algorithm based on regenerations. Finally, the resulting identified regenerations are of orders of magnitude more infrequent than the true ones that are already rare.

This motivated looking into other ways of establishing ergodicity of adaptive algorithms. Haario et al. (2001) provided the proof of consistency of Monte Carlo averages for target distributions for their Adaptive Metropolis algorithm on bounded subsets of $\mathbb{R}^d$, based on demonstrating that a related process is a mixingale (McLeish, 1975). This was extended by Atchadé and Rosenthal (2005) to unbounded state spaces, while preserving the assumption that the parameter is bounded. The approach of Andrieu and Moulines (2006) was based on the martingale limit theory, and the existence and regularity of the solutions of Poisson's equation, not requiring the parameter to belong to a compact space. Other papers using martingale approximations include Fort et al. (2011); Atchadé and Fort (2010). In particular, the former weakened the assumption of geometric ergodicity of each single kernel, required by previous approaches.

Roberts and Rosenthal (2007) in their seminal work on theoretical properties of adaptive MCMC algorithms used the coupling technique to establish ergodicity. In particular, the authors considered a collection of Markov kernels P_γ for $\gamma \in \mathcal{Y}$, where each P_γ is ergodic to π. In what follows we will denote by $\sigma(A_1, \ldots, A_K)$ the σ-algebra generated by random variables $A_1, \ldots, A_K$. The adaptive algorithm could be viewed as a process $\{(X_n, \Gamma_n)\}_{n=0}^{\infty}$, $X_n \in \mathcal{X}, \Gamma_n \in \mathcal{Y}$. Let

$$\mathcal{F}_n = \sigma\left(X_0, \ldots, X_n, \Gamma_0, \ldots, \Gamma_n\right)$$

be the filtration generated by this process. Then for any measurable set A we have

$$\mathbb{P}\left(X_{n+1} \in A | X_n = x, \Gamma_n = \gamma, \mathcal{F}_{n-1}\right) = P_\gamma(x, A).$$

Theorem 2 of Roberts and Rosenthal (2007) states that an adaptive MCMC algorithm satisfying the two conditions defined below, Containment and Diminishing Adaptation, is ergodic.

To this end, define the $\varepsilon-$convergence time as

$$M_\varepsilon(x,\gamma) = \inf\{k \geq 1 : \|P_\gamma^k(x,\cdot) - \pi(\cdot)\|_{TV} \leq \varepsilon\}.$$

(a) (Containment). For all $\varepsilon > 0$ and all $\tilde{\delta} > 0$, there exists $N = N(\varepsilon, \tilde{\delta})$ such that

$$\mathbb{P}\left(M_\varepsilon(X_n, \Gamma_n) > N | X_0 = x, \Gamma_0 = \gamma\right) \leq \tilde{\delta} \quad \text{for all } n \in \mathbb{N}.$$

(b) (Diminishing Adaptation). The random variable

$$D_n = \sup_{x\in\mathcal{X}} \|P_{\Gamma_{n+1}}(x,\cdot) - P_{\Gamma_n}(x,\cdot)\|_{TV}$$

converges to 0 in probability.

Roberts and Rosenthal (2007) proved also the weak law of large numbers for all bounded functions for adaptive MCMC algorithms, under Diminishing Adaptation, and a stronger version of Containment, called Simultaneous Uniform Ergodicity (see Theorem 5 of Roberts and Rosenthal (2007)).

Diminishing Adaptation could informally be described as making the "amount" of adaptation at iteration n go to 0 as $n \to \infty$, so it can relatively easily be imposed on an algorithm. Analysing Containment, on the other hand, requires more subtle arguments. Theorem 3 of Roberts and Rosenthal (2007) shows that the simultaneous minorization and drift conditions imply that Containment holds. The former assumes that for some $C \subseteq \mathcal{X}$, and $\epsilon > 0$

$$P_\gamma(x,\cdot) \geq \epsilon\nu_\gamma(\cdot), \quad x \in C,$$

where each ν_γ is a probability measure on $\mathcal{X}$. The latter simultaneous condition assumes that

$$\int_{\mathcal{X}} P_\gamma(x,dy)V(y) \leq \lambda V(x) + b\mathbb{I}_C(x), \quad x \in \mathcal{X}, \tag{1.7}$$

for some $0 < \lambda < 1$ and $b < \infty$.

The Containment condition was studied in detail by Bai et al. (2011b), who proved a number of sufficient, and some necessary conditions for it to hold. They also showed that even though Diminishing Adaptation itself does not guarantee convergence to the correct limiting distribution, there are examples when the adaptive algorithm is ergodic even though

Containment does not hold. However, as shown by Łatuszyński and Rosenthal (2014), in such cases with positive probability, the adaptive algorithm will be asymptotically less efficient than any non-adaptive MCMC algorithm.

1 .4 Theoretical contribution (Chapter 2)

In Chapter 2 we introduce a new class of adaptive algorithms, called Auxiliary Variable Adaptive MCMC, as follows. For a given target distribution of interest π on $\mathcal{X}$, we define an augmented state space $\tilde{\mathcal{X}} = \mathcal{X} \times \Phi$, and consider a family of distributions $\tilde{\pi}_\gamma$ on $\tilde{\mathcal{X}}$. We assume that all distributions $\{\tilde{\pi}_\gamma\}_{\gamma \in \mathcal{Y}}$ have π as their marginal distribution with respect to $\mathcal{X}$, that is,

$$\tilde{\pi}_\gamma(B \times \Phi) = \pi(B), \quad B \in \mathcal{B}(\mathcal{X}), \gamma \in \mathcal{Y},$$

where $\mathcal{B}(\mathcal{X})$ denotes the Borel sets on $\mathcal{X}$. Let $\{\tilde{P}_\gamma\}_{\gamma \in \mathcal{Y}}$ be a collection of Markov chain transition kernels on $(\tilde{\mathcal{X}}, \mathcal{B}(\tilde{\mathcal{X}}))$, such that each $\tilde{P}_\gamma$ has $\tilde{\pi}_\gamma$ as its invariant distribution and is Harris ergodic. Examples of algorithms in this class are the Jumping Adaptive Multimodal Sampler introduced in Chapter 2, which motivated considering this class, as well as adaptive parallel tempering (Miasojedow et al., 2013).

Similarly to the adaptive MCMC setup defined in Section 1 .3, an algorithm in the Auxiliary Variable Adaptive MCMC class defines a process $\{(\tilde{X}_n, \Gamma_n)\}_{n=0}^{\infty}$. Let

$$\mathcal{G}_n = \sigma\{\tilde{X}_0, \ldots, \tilde{X}_n, \Gamma_0, \ldots, \Gamma_n\}$$

denote its filtration. The process satisfies

$$\mathbb{P}[\tilde{X}_{n+1} \in \tilde{B} | \tilde{X}_n = \tilde{x}, \Gamma_n = \gamma, \mathcal{G}_{n-1}] = \tilde{P}_\gamma(\tilde{x}, \tilde{B})$$

for $\tilde{x} \in \tilde{\mathcal{X}}, \gamma \in \mathcal{Y}, \tilde{B} \in \mathcal{B}(\tilde{\mathcal{X}})$.

We proved ergodicity and the weak law of large numbers under analogues of the Containment and Diminishing Adaptation conditions, defined below.

(a) (Containment). For all $\varepsilon > 0$ and all $\tilde{\delta} > 0$, there exists $N = N(\varepsilon, \tilde{\delta})$ such that

$$\mathbb{P}\left(M_\varepsilon(\tilde{X}_n, \Gamma_n) > N | \tilde{X}_0 = \tilde{x}, \Gamma_0 = \gamma\right) \leq \tilde{\delta} \quad \text{for all } n \in \mathbb{N},$$

where

$$M_\varepsilon(x,\gamma) = \inf\{k \geq 1 : \|\tilde{P}_\gamma^k(\tilde{x},\cdot) - \tilde{\pi}_\gamma(\cdot)\|_{TV} \leq \varepsilon\}.$$

(b) (Diminishing Adaptation). The random variable

$$D_n = \sup_{x\in\mathcal{X}} \|\tilde{P}_{\Gamma_{n+1}}(x,\cdot) - \tilde{P}_{\Gamma_n}(x,\cdot)\|_{TV}$$

converges to 0 in probability.

As opposed to Roberts and Rosenthal (2007), in our proof of ergodicity we provide an explicit construction of interpolating processes on the same probability space and show how to analyse them. Our rigorous coupling construction facilitated proving a more general form of the Weak Law of Large Numbers, requiring only the Containment condition, rather than Simultaneous Uniform Ergodicity.

We also show in Chapter 2 how to analyse Containment in our context. One of the difficulties in our case was the need to use drift functions depending on the target distribution $\tilde{\pi}_\gamma$. Instead of the simultaneous drift condition given by (1.7), in our setting it had the following form

$$\int_{\mathcal{X}} \tilde{P}_\gamma(\tilde{x}, d\tilde{y}) V_{\tilde{\pi}_\gamma}(\tilde{y}) \leq \lambda V_{\tilde{\pi}_\gamma}(\tilde{x}) + b, \quad \tilde{x} \in \tilde{\mathcal{X}},$$

for some $0 < \lambda < 1$ and $b < \infty$. Consequently, it was impossible to use the proof of Theorem 3 of Roberts and Rosenthal (2007) directly, and we developed our own framework for proving Containment. In particular, we had to add some additional assumptions, such as adapting on a compact set.

2 Literature review on MCMC methods for targeting multimodal distributions

2 .1 Poor mixing of standard Markov chain Monte Carlo methods on multimodal target distributions

Standard Markov chain Monte Carlo methods, such as RWM or MALA tend to perform poorly when the target distribution of interest, denoted by π, is multimodal, and this phenomenon can be observed even in low-dimensional state spaces. The reason for this is that the modes, defined loosely as high probability regions around surrounding local maxima of

π, are typically separated by low-probability barriers. Due to their design, standard methods struggle to accept moves to regions of low posterior probability, hence preventing good mixing between the modes. As proved by Mangoubi et al. (2018), even HMC, known to scale much better with dimension than Random Walk Metropolis on unimodal targets, turns out to perform as poorly as RWM when the target is multimodal. Diagnostic tools used in MCMC typically fail to detect these problems, as the MCMC sampler may not visit distant modes, while mixing well on the area it has explored.

Thus, an MCMC algorithm well-designed for targeting multimodal distributions needs to address the following issues.

1. Identify high probability regions, where the modes are located.

2. Move efficiently between the modes, which requires designing a mechanism of crossing low probability regions.

3. Ensure good local mixing by performing moves within each mode.

Additionally, it is well-understood that ideally such algorithms should be adaptive, and allow for performing at least some of the computations in parallel. Below we present various algorithms known from the literature that attempt to tackle the above issues in different ways, splitting these methods based on their main underlying idea.

2 .2 Tempering-based methods

The key idea behind tempering-based methods, such as parallel or simulated tempering, relies on observing that if π is a multimodal distribution of interest, then π^β for $0 < \beta < 1$ is flatter and more "spread out", and hence easier to sample from. Since eventually π remains the distribution of interest, some further steps are needed to obtain the samples from π, using samples from π^β as an intermediate step.

The parallel tempering (PT) method, proposed by Geyer (1991), is based on simulating a Markov chain defined on $\underbrace{\mathcal{X} \times \ldots \times \mathcal{X}}_{N+1}$ targeting

$$\bar{\pi}(x_0, x_1, \ldots, x_N) \propto \prod_{j=0}^{N} \pi(x_j)^{\beta_j} \tag{1.8}$$

for a sequence of parameters $\beta_0, \ldots, \beta_N$, often called "inverse temperatures" satisfying $0 \leq$

$\beta_0 < \ldots < \beta_N = 1$. Hence, we may view this as having $N+1$ chains on $\mathcal{X}$, each of them targeting the tempered distribution π^{β_j} for $j = 0, \ldots, N$. In particular, the last chain targets the distribution of interest. The algorithm alternates between the following two types of steps: update moves at each temperature levels, and temperature swaps. The former are Random Walk Metropolis steps performed separately for each temperature. To perform a swap, we first sample j from the set $\{0, \ldots, N-1\}$, and then propose a move from $(x_0, \ldots, x_j, x_{j+1}, \ldots, x_N)$ to $(x_0, \ldots, x_{j+1}, x_j, \ldots, x_N)$, subsequently accepted or rejected with probability such that the detailed balance holds for (1.8). The system is designed in such a way that the good mixing of chains associated with higher temperatures is transferred to chains at lower temperature levels. This helps address the first two issues outlined in Section 2 .1.

The design of the sequence of temperatures is of critical importance for ensuring good mixing. Atchadé et al. (2011) proved that the optimal choice of inverse temperatures for parallel tempering satisfies that the average acceptance rate of the temperature swaps, as well as update moves at each temperature level, is close to 0.234. This result was applied in practice by Miasojedow et al. (2013), who proposed an adaptive version of parallel tempering. The number of temperatures is chosen by the user and remains fixed, while parameters β_j and covariance matrices used for proposals at each level are updated on the fly, so that the acceptance rates get close to 0.234. Since parallel tempering is arguably the most popular MCMC method for multimodal distributions, in Chapter 2 we chose adaptive PT, as its optimized version, to be the benchmark for our own methodology.

Another improvement in parallel tempering was proposed by Tawn and Roberts (2019). The goal of their method, called QuanTA, is to make a temperature swap between two adjacent levels i and $i+1$ more independent of the current locations of x_i and x_{i+1}. To this end, they consider an appropriate transformation t and use it for proposing temperature swaps as moves from $(x_0, \ldots, x_i, x_{i+1}, \ldots, x_N)$ to $(x_0, \ldots, t(x_{i+1}, \beta_{i+1}, \beta_i), t(x_i, \beta_i, \beta_{i+1}), \ldots, x_N)$ Another novelty compared with the original PT algorithm is that QuanTA involves also approximating the posterior by a mixture of Gaussians and remembering the estimated centres of the modes.

An interesting strategy was proposed recently by Dharamshi et al. (2021), whose adaptive Divergence Minimization is designed for targeting irregularly-shaped target distributions. The main underlying idea is to minimize the information projection side of the Kullback–

Leibler divergence between the class of the proposal distributions and the target of interest. An extension of this technique, called Scout MCMC, is designed to be well-suited for multi-modal distributions. It relies on performing parallel tempering with just two chains (one of them targeting the original distribution π and the other one targeting its tempered version), which utilize Divergence Minimization steps to make update moves. This idea allows moving efficiently between the modes, as well as perform efficient local steps, even when the modes have irregular shapes.

The equi-energy sampler, developed by Kou et al. (2006) is based on a somewhat similar idea. Its additional advantage is providing estimates of quantities, such as entropy, not accessible via parallel tempering.

In this thesis we focus on reversible MCMC methods, however, recently irreversible versions of PT have been attracting interest, see for example Syed et al. (2019, 2021).

Simulated tempering, proposed by Marinari and Parisi (1992), relies on designing a target distribution $\tilde{\pi}$ on an augmented state space $\mathcal{X} \times \{0, \ldots, N\}$ defined as

$$\tilde{\pi}(x, j) \propto \pi(x)^{\beta_j}, \quad j = 0, \ldots, N, \quad 0 \leq \beta_0 < \ldots < \beta_N = 1.$$

Thus the conditional distribution $\tilde{\pi}(x, j)$ given $j = N$ is the distribution of interest. This algorithm also relies on two types of steps, update moves and temperature changes. In an update step we propose a move from (x, j) to (y, j) where y is drawn from a normal distribution centred at x. A temperature change proposes a move from (x, j) to either $(x, j-1)$ or $(x, j+1)$. A potential problem with this method may arise when $\tilde{\pi}$ has a very small probability of $j = N$. In such a case the number of samples that one can use to approximate π will also be small. An improvement to simulated tempering was more recently proposed by Tawn et al. (2020), who developed a so-called weight-preserving simulated tempering algorithm.

Theoretical properties of tempering-based methods have been well-studied in the literature, so we present below some of the main results. We first recall briefly the relation between mixing time and the spectral gap. Let P be a π-irreducible, aperiodic, non-negative definite and π-reversible Markov chain. Then P converges to π in total variation and the rate of this

convergence is characterized by

$$\|P^n(x,\cdot) - \pi(\cdot)\|_{TV} \leq M(x)\left\{1 - \mathbf{Gap}(P)\right\}^n$$

for a function M such that $M(x) < \infty$ for π-a.e. $x \in \mathcal{X}$. We say that an MCMC algorithm mixes torpidly when the spectral gap decreases exponentially as a function of the dimension of the state space, and we say it mixes rapidly when the gap decreases polynomially.

Mixing of parallel and simulated tempering on multimodal distributions was studied theoretically by Woodard et al. (2009a,b). In particular, they considered an example of a mixture of Gaussians given by

$$\pi(x) = \frac{1}{2}N\left(-\underbrace{(1,\ldots,1)}_{d}, \sigma_1^2 I_d\right) + \frac{1}{2}N\left(\underbrace{(1,\ldots,1)}_{d}, \sigma_2^2 I_d\right). \tag{1.9}$$

The analysis of the target distribution (1.9) motivated also one of our examples studied in Chapter 2. For $\sigma_1 = \sigma_2$, parallel and simulated tempering mix rapidly for certain choices of temperatures. If $\sigma_1 \neq \sigma_2$, parallel and simulated tempering turn out to mix torpidly on this example for any choice of the temperature ladder. A practical consequence of this fact is that if the target distribution contains both narrow and wide modes, tempering-based methods tend to get stuck in the wide ones, and this issue is particularly pronounced in high dimensions. Recent developments in this direction include the aforementioned article by Tawn et al. (2020). Their weight-preserving simulated tempering method provides an improvement to standard simulated tempering, which results in faster mixing between the modes on target distributions with wide and narrow modes, like the one considered in equation (1.9). The paper presents also theoretical results on the mixing time of the algorithm.

2 .3 Metropolis-Hastings algorithms with a special design of the proposal distribution

The methods presented below are characterized by non-standard designs of proposal distributions, which facilitate crossing the low-probability barriers, thus addressing the first two issues outlined in Section 2 .1. In the first two methods, these proposals are followed by Metropolis-Hastings acceptance/ rejection steps preserving invariance with respect to π. The last one operates on an extended state space following the auxiliary variable proposed

by Møller et al. (2006). In each case the algorithm alternates between the between-mode moves, based on non-standard transition kernels, and standard Random Walk Metropolis steps, used to move locally within the mode, thus addressing the third issue.

The method developed by Tjelmeland and Hegstad (2001) relies on two proposal distributions Q_0 and Q_1, used with probability 0.5 each. Let $\mu(z)$ denote the local maximum obtained via an optimization procedure started at z and let $\Sigma(z)$ be the local covariance matrix around $\mu(z)$. Let ψ be a vector of the same dimension as the current state x, drawn from a distribution with large variance. Then $Q_0(\cdot|x)$ is the normal distribution $N(\mu(x+\psi), \Sigma(x+\psi))$. In turn, $Q_1(\cdot|x)$ is given by $N(\mu(x-\psi), \Sigma(x-\psi))$. The above design allows the algorithm to propose moves to distant regions (via the use of ψ), while ensuring the proposed point lies in a high-probability region (via the use of maximization). The proposal from Q_i is subsequently accepted with probability

$$\alpha_i(y|x) = \min\left[1, \frac{\pi(y)q_{1-i}(x|y)}{\pi(x)q_i(y|x)}\right], \quad i = 0, 1,$$

where q_i denotes the density corresponding to Q_i, which ensures that the chain is invariant with respect to π. A significant drawback of this technique is the cost of each iteration associated with finding local optima.

The tempered transition method was developed by Neal (1996), who used the idea of tempering to design a proposal distribution for between-mode moves. To this end, we first define N pairs of transition kernels T_j and T_j' are constructed such that they satisfy

$$\pi(x)^{\beta_j} T_j(x'|x) = \pi(x')^{\beta_j} T_j'(x|x') \quad \text{for all } x, x' \in \mathcal{X} \text{ and } j = 0, \ldots, N-1,$$

where $0 \leq \beta_0 < \ldots < \beta_N = 1$. To propose a move from x to y, the algorithm passes twice through the temperature ladder in the following way. Setting $x_N = x$, we generate $x_j \sim T_j(\cdot|x_{j+1})$ for $j = N-1, \ldots, 0$. Then setting $x_0' := x_0$, for $j = 0, \ldots, N-1$ generate $x_{j+1}' \sim T_j'(\cdot|x_j')$, to finally obtain $y = x_N'$. The new state is then accepted with probability

$$\min\left[1, \left(\prod_{j=1}^{N} \pi(x_j)^{\beta_{j-1}-\beta_j}\right)\left(\prod_{j=1}^{N} \pi(x_j')^{\beta_j-\beta_{j-1}}\right)\right],$$

so that the Markov chain is invariant with respect to π.

Tak et al. (2017) developed another technique, based on the so-called down-up jumping

density q^{DU} used for proposing the next step. Given the current state x, first an intermediate downhill proposal x' is generated from density $q^D(x'|x)$, which is followed by an uphill proposal x^* generated from $q^U(x^*|x')$. the resulting proposal down-up jumping density is given by

$$q^{DU}(x^*|x) = \int q^D(x'|x)q^U(x^*|x')dx'.$$

The densities q^D and q^U are chosen such that they favour lower and higher values of the target distribution, respectively. Intuitively, x' will be located in the valley between modes and x^* will then more likely jump to a different mode than the one x belonged to. Since the acceptance probability for the proposal q^{DU} is intractable, the authors introduce an auxiliary variable, following the idea proposed by Møller et al. (2006). Note that unlike in the case of tempering-based methods, augmenting the state space is not used here to facilitate moving between modes, but purely to overcome the intractability issue.

One of the shortcomings of the methods described above is that they are not well-developed in the context of Adaptive MCMC.

2 .4 Optimization-based methods

Recall the three challenges that an MCMC sampler designed for targeting multimodal distributions needs to address, outlined in Section 2 .1. Methods presented so far tackle the first two of them simultaneously, that is, finding high-probability regions and moving between the modes happens at the same time. Optimization-based methods described below separate the two tasks. They rely on preliminary runs of optimization algorithms, which find the local maxima of π, hence addressing the first issue. A huge advantage of this approach is that the mode-finding stage can be performed fully in parallel. The knowledge of the location of the modes is then used to inform the appropriate design of the proposal distribution used for moves between the modes, which helps tackle the second issue. As we explain later, the methodology for Jumping Adaptive Multimodal Sampler presented in detail in Chapter 2, also belongs to this group.

Several optimization-based methods rely on the idea of Smart Darting Monte Carlo, introduced by Andricioaei et al. (2001). Suppose we know the locations of modes $\mu_1, \ldots, \mu_M$ of the target distribution. We then construct non-overlapping ϵ-spheres (called "jumping regions") centred at $\mu_1, \ldots, \mu_M$. We then perform steps of two types. Local moves are

Random Walk Metropolis steps. If the current state x is within jumping region i, for some $i = 1, \ldots, M$, jump moves are performed with probability p, and local steps are made otherwise. A jump move proposes state $y = x + \mu_j - \mu_i$, for j sampled uniformly from $1, \ldots, M$. The acceptance probability is such that the detailed balance condition holds. The intuition behind such a construction of jump moves is that the proposed state is the "corresponding" point in a different ϵ-sphere. The value of the target density is therefore expected to have the same order of magnitude for the proposed state and the current state, which ensures that the probability of acceptance is relatively high. Our JAMS methodology offers two ways of performing jumps, one of which, called "deterministic jumps", relies on a similar idea.

Sminchisescu and Welling (2011) generalized the Smart Darting Monte Carlo method, to allow for jumping regions to overlap, as well as to have arbitrary volume and shape. Regeneration Darting Monte Carlo (Ahn et al., 2013) extended this methodology further, by introducing updates of the jumping regions and parameters of the proposal distribution at regeneration times. In this algorithm the jump kernel is an independence sampler utilizing as a proposal distribution a mixture of Gaussian distributions truncated to the jumping regions.

Another optimization-based method, Wormhole Hamiltonian Monte Carlo (WHMC), was introduced by Lan et al. (2014) as an extension of Riemanian Manifold Hamiltonian Monte Carlo (Girolami and Calderhead, 2011). Given the locations of the modes, WHMC modifies the Riemannian geometric properties of the target distribution to create a network of so-called wormholes connecting the modes. The modification is designed in such a way that it shortens the distance between the modes, hence allowing for moving efficiently between high-probability regions. Updating the network is allowed only at regeneration times.

2 .5 Other methods

We have presented above three major strands of research, based on their underlying idea on how to tackle the problem of multimodality. The list of algorithms designed to deal with multimodal distributions is however much longer. The Wang–Landau algorithm (Wang and Landau, 2001a,b), as well as its adaptive version (Bornn et al., 2013), aims to explore various modes by pushing the sampler away from already known regions. Zhou (2011) proposed a method based on combining the optimization-based approach with the Wang–Landau method.

Other approaches include adaptive independence samplers, with a fixed number of components, as proposed by Giordani and Kohn (2010) or Luengo and Martino (2013), or with a possibility of adding new components on the fly (Maire et al., 2019). The pseudo-extended Hamiltonian Monte Carlo method Nemeth et al. (2019) creates a target defined on an augmented state space, in such a way that the modes are connected on that extended space. The pseudo-extended target is a special case of the target used in pseudo-marginal MCMC (Andrieu and Roberts, 2009). A post-hoc correction step is then required to convert the samples from the augmented target to samples from the original target distribution of interest.

2 .6 Methodological contribution (Chapter 2)

In Chapter 2 we propose the Jumping Adaptive Multimodal Sampler, which is an optimization-based framework for targeting multimodal distributions. It relies on a defining collection of target distributions $\tilde{\pi}_\gamma$, $\gamma \in \mathcal{Y}$, on the augmented state space $\mathcal{X} \times \mathcal{I}$, where $I = \{1, \ldots, N\}$ is the finite set of indices of the modes of π. Let $\mu_1, \ldots, \mu_N$ denote the modes found by initial optimization runs. By γ we denote the design parameter of the algorithm, adapted on the fly. For each γ the marginal distribution of $\tilde{\pi}_\gamma$ with respect to $\mathcal{X}$ is π.

We then design an MCMC sampler targeting $\tilde{\pi}_\gamma$, based on local moves and jumps. The former propose moves within the same mode, that is, from (x, i) to (y, i) for $x, y \in \mathcal{X}$, $i \in \mathcal{I}$. Any local sampler, such as RWM or MALA can be used for this purpose. To make a jump from mode i, we first sample mode $k \in \mathcal{I}$, and then propose a point y in the region around mode k. We study two ways of performing jump moves, either based on the independence sampler, or on proposing a "corresponding" point from another mode, which we call a deterministic jump.

Proposal distributions for both local moves and jumps are adjusted to the local geometry of the mode, which allows addressing efficiently the second and the third challenge from Section 2 .1. This is achieved through learning the parameters (typically, covariance matrices) of each mode based on points assigned to that mode. It is worth emphasizing that the augmented target distribution $\tilde{\pi}_\gamma$ is constructed in such a way that if a state (x, i) is accepted by the sampler, it is unlikely that x lies far from mode μ_i. Therefore, adaptation indeed relies on points from the region corresponding to the mode.

Since JAMS belongs to the Auxiliary Variable Adaptive MCMC class, we prove its ergodicity by verifying Containment and the Diminishing Adaptation, discussed in Section 1 .4.

In order to impose Diminishing Adaptation on the algorithm, we use the strategy of adapting increasingly rarely, as advocated by Chimisov et al. (2018). To prove the uniform drift condition (1.7), which is crucial for verifying Containment, we consider the drift function $V_{\tilde{\pi}_\gamma}(x,i) = \tilde{\pi}_\gamma(x,i)^{-1/2}$, and address separately local moves and jumps. We show that if the local moves are performed using Random Walk Metropolis, our algorithm satisfies Containment under mild conditions, similar to those required for geometric ergodicity of Metropolis algorithms (Jarner and Hansen, 2000).

3 Literature review on Bayesian inference under model misspecification

3 .1 Asymptotic results for standard Bayesian posteriors

We start by introducing a definition of model misspecification. We say that the model (1.1) is well-specified if the data $x_1, \ldots, x_n$ are generated from $f(\cdot|\theta_0)$ for some $\theta_0 \in \Theta$, and otherwise we say that it is misspecified. We let P^* denote the true underlying data-generating distribution, and by p^* we be its density. Note that we only consider misspecification with respect to the likelihood, not the prior.

Let us first consider the well-specified case. The Bernstein–von Mises theorem provides a link between Bayesian and frequentist inference when the number of observations goes to infinity, and has an important consequence for interpretation of Bayesian credible sets. Before we present it more formally, we define $\hat{\theta}$ as the maximum likelihood estimator in model (1.1), that is $\hat{\theta} = \arg\max_\theta \sum_{i=1}^n \log f(x_i|\theta)$. Note that even though we present below results for maximum likelihood estimators, we could more generally consider M-estimators (see Chapter 5 of Van der Vaart (2000)). We define also information matrices I and J as

$$I(\theta) = E_{p^*}\left[\{\nabla_\theta \log f(X|\theta)\}\{\nabla_\theta \log f(X|\theta)\}^T\right], \quad J(\theta) = -E_{p^*}\left[D^2_\theta \{\log f(X|\theta)\}\right]. \quad (1.10)$$

For well-specified regular models, the matrices $I(\theta_0)$ and $J(\theta_0)$ are equal (see, for example, Chapter 5 of Van der Vaart (2000)). The Bernstein-von Mises theorem states that under standard regularity conditions the asymptotic distribution of $\sqrt{n}\left(\theta - \hat{\theta}\right)$, for θ sampled from the Bayesian posterior, is $N\left(0, I(\theta_0)^{-1}\right)$, for I defined in (1.10). This implies that the credible sets under Bayesian inference are valid confidence sets in the frequentist sense. In

other words, central sets of probability $1-\alpha$ under the posterior distribution cover parameter θ_0 at the confidence level $1-\alpha$. This remarkable result on properties of Bayesian posteriors is named after R. von Mises and S. N. Bernstein, who proved results on asymptotic normality of the posterior distribution. The name of the theorem was however coined by L. Le Cam (LeCam, 1953), who extended their results to a more general setting.

In practice, users can rarely specify correctly the model they use, thus it is important to understand what happens when the model is misspecified. Suppose that the density of the true data-generating distribution, denoted by p^*, does not belong to the family $f(\cdot|\theta), \theta \in \Theta$. Then $\hat{\theta}$ asymptotically concentrates around the so-called pseudo-true parameter θ^* minimizing $D_{KL}(p^*||f)$, the Kullback–Leibler divergence between p^* and $f(\cdot|\theta)$, $\theta \in \Theta$. As shown by Berk (1966, 1970); Bunke and Milhaud (1998), an analogous result holds in Bayesian statistics, since under regularity conditions the posterior asymptotically also concentrates around θ^*. However, the asymptotic covariance matrix is typically different under the frequentist and the Bayesian scenario. In the frequentist case under misspecification the asymptotic covariance matrix is given by the "sandwich" formula Huber (1967), that is $\sqrt{n}\left(\hat{\theta} - \theta^*\right)$ converges in distribution to $N\left(0, J(\theta^*)^{-1} I(\theta^*) J(\theta^*)^{-1}\right)$. As shown by Kleijn and Van der Vaart (2012), for θ drawn from the Bayesian posterior for any measurable set A we have

$$\Pi\left\{n^{1/2}\left(\theta - \hat{\theta}\right) \in A | x_{1:n}\right\} \to \mathbb{P}\left(Z \in A\right), \quad Z \sim N\left\{0, J(\theta^*)^{-1}\right\}.$$

The above property implies that under misspecification credible sets are not valid asymptotic confidence sets for θ^* in the frequentist sense, and therefore using Bayesian inference may result in misleading quantification of uncertainty around parameters of interest. This caveat of the standard Bayesian approach has motivated a strand of research focused on performing inference using methods more robust to model misspecification. We present below some main approaches to performing Bayesian analysis under model misspecification.

3 .2 General methods of robust Bayesian inference

One of the methods that has gained a lot of popularity relies on using a power likelihood of the form $f(x|\theta)^\eta$ for a parameter $\eta > 0$ instead of the standard likelihood $f(x|\theta)$. Examples of the vast literature on this topic comprise Walker and Hjort (2001); Jiang and Tanner (2008); Grünwald (2012); Bissiri et al. (2016); Grünwald and van Ommen (2017). The resulting

power posterior is

$$\pi_\eta(\theta|x_{1:n}) \propto \pi(\theta) \prod_{i=1}^{n} f(x_i|\theta)^\eta = \pi(\theta) \exp\left\{\eta \sum_{i=1}^{n} \log f(x_i|\theta)\right\}. \tag{1.11}$$

More generally, one can consider directly a loss function $l(x, \theta)$ in place of the minus log likelihood, thus getting

$$\pi_\eta(\theta|x_{1:n}) \propto \pi(\theta) \exp\left\{ - \eta \sum_{i=1}^{n} l(x_i, \theta)\right\}. \tag{1.12}$$

Distributions π_η are termed as "general posteriors" or "Gibbs posteriors", and admit (1.11) as a special case. The work by Bissiri et al. (2016) formalized the above approach by showing that similarly to standard Bayesian inference, general posteriors provide a coherent update from a prior belief about parameter θ to a posterior belief expressed via $\pi_\eta(\theta|x_{1:n})$. General posteriors are closely linked with the PAC-Bayesian theory, which provides upper bounds on prediction risks associated with these distributions (Germain et al., 2016; Guedj, 2019; Alquier, 2020).

For general loss functions l, we set

$$\hat{\theta} = \arg\min_{\theta} \sum_{i=1}^{n} l(x_i, \theta)$$

and

$$\theta^* = \arg\min_{\theta} \mathbb{E}_{p^*} \left\{ l(X, \theta) \right\}.$$

Let us now state the asymptotic result for general posteriors (Chernozhukov and Hong, 2003; Müller, 2013). For θ drawn from a general posterior (1.12) the expression $\sqrt{n}\left(\theta - \hat{\theta}\right)$ is asymptotically normal with mean zero and covariance matrix $\eta^{-1} V(\theta^*)^{-1}$, where $V(\theta) = E_{p^*}\left\{ D^2_\theta l(X|\theta) \right\}$. In particular, if the loss function is the negative log-likelihood, the asymptotic covariance matrix is $\eta^{-1} J(\theta^*)^{-1}$. Therefore, this approach can only correct the scale, but not the shape of the covariance structure.

The interpretation of η is that it represents the learning rate, or the analyst's trust in the model. Indeed, for $\eta < 1$ the data have less influence than in standard Bayesian inference, while using $\eta > 1$ gives more prominence to the data. Since the choice of an appropriate power η has a crucial impact on the performance of the method, several articles attempted to address the issue of choosing the optimal value of η. The SafeBayes algorithm (Grünwald and

van Ommen, 2017) learns optimal η from data, relying on the prequential view on Bayesian inference (Dawid, 1984). Other methods (Holmes and Walker, 2017; Lyddon et al., 2019) propose different ways of optimizing η based on matching Fisher information, whereas Syring and Martin (2019) focus on setting η so that the corresponding credible region achieves the nominal frequentist coverage probability for θ^* (conservatively or exactly).

Power posteriors can also be seen as approximations to coarsened posteriors introduced by Miller and Dunson (2018). This is another method of robust Bayesian inference that relies on performing valid Bayesian inference conditioned on the event that the empirical distribution of the observed data is close to the empirical distribution of data generated by the assumed model, with respect to some discrepancy between probability distributions. The recent approach proposed by Loaiza-Maya et al. (2021) also draws its inspiration from the literature on power posteriors and is aimed at improving upon standard Bayesian prediction.

Another strand of research on robust Bayesian inference is focused around the BayesBag algorithm, which was first proposed in the field of phylogenetics (Waddell et al., 2002; Douady et al., 2003), although its name comes from Bühlmann (2014). The BayesBag algorithm relies on performing standard Bayesian inference on a certain number of bootstrapped datasets and combining all obtained samples.

The theory for BayesBag was developed by Huggins and Miller (2019). Suppose that original dataset has n observation, and that each bootstrapped dataset is obtained via sampling with replacement M observations from the original dataset. The authors proved that for $\bar{\theta}$ sampled from the BayesBag algorithm, under regularity conditions we have

$$\mathbb{P}\left\{n^{1/2}\left(\bar{\theta}-\hat{\theta}\right)\in A|x_{1:n}\right\}\to\mathbb{P}\left(Z\in A\right),\quad Z\sim N\left\{0,J(\theta^*)^{-1}/c+J(\theta^*)^{-1}I(\theta^*)J(\theta^*)^{-1}/c\right\}.$$

for any measurable set A, where $c=\lim_{n\to\infty}M/n$, $0<c<\infty$. Hence, the main hyperparameter determining the asymptotic covariance matrix is the size M of each bootstrapped dataset. The article provides also a discussion on the optimal choice of M. A related paper by Huggins and Miller (2020) focuses on applying the BayesBag method to model selection.

In the recent years we have been also observing a renewed interest in methods of robust inference based on Weighted Likelihood Bootstrap (WLB) (Newton, 1991; Newton and Raftery, 1994). This is largely due to developments and more widespread access to parallel computing, as WLB can be performed fully in parallel. Drawing a sample from WLB for

model (1.1) relies on maximizing the weighted log likelihood for weights $w_1, \ldots, w_n$ sampled i.i.d. from the Exp(1) distribution, that is

$$\tilde{\theta} = \arg\max_{\theta} \left\{ \sum_{i=1}^{n} w_i \log f(x_i|\theta) \right\}, \quad w_1, \ldots, w_n \sim \text{Exp}(1). \tag{1.13}$$

This algorithm was initially introduced as a method of sampling approximately from the Bayesian posterior. The asymptotic distribution of $\sqrt{n}(\tilde{\theta} - \hat{\theta})$ for $\tilde{\theta}$ drawn according (1.13) in a well-specified model (with a true parameter denoted by θ_0) was found by Newton (1991) and proved to be $N\left(0, I(\theta_0)^{-1}\right)$. Since this coincides with the asymptotics of the Bayesian posterior for well-specified models, this justifies why, for large sample sizes, the method was seen as an approximate alternative to Markov chain Monte Carlo methods. The authors suggested that the WLB procedure could alternatively be adjusted by the Sequential Importance Resampling algorithm (Gordon et al., 1993; Doucet et al., 2000), to correct for the fact that Weighted Likelihood Bootstrap is approximate.

The advantageous properties of Weighted Likelihood Bootstrap from the point of view of model misspecification were described by Fushiki (2010) and Lyddon et al. (2019). In particular, Theorem 1 of Lyddon et al. (2019) states that under standard regularity conditions for any measurable set A we have

$$\mathbb{P}\left\{ n^{1/2}\left(\tilde{\theta} - \hat{\theta}\right) \in A | x_{1:n} \right\} \to \mathbb{P}\left(Z \in A\right), \quad Z \sim N\left\{0, J(\theta^*)^{-1} I(\theta^*) J(\theta^*)^{-1}\right\}.$$

Thus, credible sets are valid confidence sets around θ^*. It is worth emphasizing that this property does not hold for any of the approaches presented above, that is general posteriors and BayesBag, unless matrices $I(\theta^*)$ and $J(\theta^*)$ are proportional.

Another advantage of Weighted Likelihood Bootstrap is that it outperforms Bayesian inference in terms of prediction. The work done by Fushiki (2005, 2010) shows that as long as $I(\theta^*) \neq J(\theta^*)$, WLB provides asymptotically better prediction than Bayesian inference (see also Theorem 1 of Lyddon et al. (2018)).

Even though the method appeared in the context of Bayesian models, it does not make use of any prior knowledge that the analyst may have about the parameter of interest. A similar idea was in fact being developed independently in the frequentist literature, where it was called “perturbing the minimand" (Jin et al., 2001). The analogy between WLB and ideas around perturbing the minimand was noticed by Parzen and Lipsitz (2007). Some of

these approaches consider adding a LASSO-type regularization term, see for example Minnier et al. (2011) and a follow-up paper by Das et al. (2019), which could from the Bayesian point of view be interpreted as incorporating a certain form of a prior distribution.

The question of incorporating prior information into Weighted Likelihood Bootstrap has been recently attracting attention in the Bayesian community. Approaches based on including in the dataset so-called pseudo-samples drawn from the posterior predictive or the prior predictive distribution were proposed by Lyddon et al. (2018) and Fong et al. (2019), whereas Newton et al. (2020) developed a method based on penalizing the weighted log-likelihood.

Several recent papers have looked into ways of performing model selection via a LASSO-type regression using Weighted Likelihood Bootstrap with shrinkage priors (Ng and Newton, 2020; Nie and Ročková, 2020). The algorithm developed by Ng and Newton (2020) should be interpreted as incorporating the Laplace prior into Weighted Likelihood Bootstrap, while Nie and Ročková (2020) focused on utilizing spike-and-slab priors (Mitchell and Beauchamp, 1988; George and McCulloch, 1993). As for theoretical properties of these methods, Ng and Newton (2020) proved the model selection consistency in the setting with a growing dimensionality. These results were then extended by Nie and Ročková (2020), who found the concentration rate in this setting.

The above review of methods aimed at tackling the issue of misspecification of Bayesian models is by no means comprehensive. There are other recent approaches, such as the reweighted probabilistic models, proposed by Wang et al. (2017), or the so-called robust Bayesian modelling, developed by Wang et al. (2018) and based on empirical Bayes estimation. We expect that given the growing complexity of the problems and their associated models, the topic of robust inference will keep attracting a lot of attention in the Bayesian literature.

3 .3 Contribution (Chapter 3)

The main goal of our work in Chapter 3 was to develop a methodology, which we further call Posterior Bootstrap, combining the advantages of Bayesian inference and Weighted Likelihood Bootstrap. The algorithms we propose are able to faithfully represent the user's prior beliefs, including when the prior has a hierarchical structure, while allowing for performing computations in parallel. The technique is also robust to model misspecification, in an analogous way to Weighted Likelihood Bootstrap.

We study two ways of incorporating prior information, in each case providing a rigorous analysis of the impact of incorporating the prior. Suppose the prior π on $\mathbb{R}^d$ factorizes, and let $\log \pi(\theta)$ denote a vector with coordinates $\log \pi_i(\theta_i)$ for $i = 1, \ldots, d$. In a non-hierarchical case, the first method relies on computing

$$\tilde{\theta} = \arg\max_{\theta} \left\{ \sum_{i=1}^{n} w_i \log f(x_i|\theta) + w_0^T \log \pi(\theta) \right\}, \quad w_1, \ldots, w_n \sim \text{Exp}(1),$$

for a hyperparameter $w_0 \in \mathbb{R}^d$. The second one relies on incorporating pseudosamples from the prior predictive, as proposed by Fong et al. (2019). To analyse the influence of the prior mathematically, we use second-order Edgeworth expansions, and study the prior-dependent terms in these expansions.

Furthermore, our analysis guides the choice of hyperparameters of the above methods. To our knowledge, this has been the first attempt to analyse theoretically the impact of the prior information incorporated into WLB, and provide mathematically principled ways of setting hyperparameters.

As mentioned earlier, Bayesian inference for hierarchical models is regarded as one of the most influential statistical ideas (Gelman and Vehtari, 2021). This was one of our motivations for extending WLB to hierarchical models, which again, to our knowledge, has not been attempted before. We consider the following type of hierarchical model

$$\begin{aligned} x_{k,1}, \ldots x_{k,n_k} &\sim f(x_{k,i}|\theta_k), \quad k = 1, \ldots, K, \\ \theta_k &\sim g(\theta_k|\lambda), \qquad k = 1, \ldots, K, \\ \lambda &\sim p(\lambda), \end{aligned}$$

where $x_{k,1}, \ldots, x_{k,n_k}$ for $k = 1, \ldots, K$ are observations and $\lambda, \theta_1, \ldots, \theta_K$ are parameters of interest. We study two scenarios: in the first one the number of groups K is fixed, in the second one it goes slowly to infinity, as $n_1, \ldots, n_K$ go to infinity. In Chapter 3 we propose principled algorithms for both cases. Apart from this methodological contribution, we also derive the asymptotic distribution of λ under these algorithms.

While the main focus of this work is on inference, we were also interested in comparing the predictive properties of general posteriors and BayesBag with those of WLB or Posterior Bootstrap under model (1.1). We prove in Chapter 3 that the answer depends on the

eigenvalues of the matrix $J(\theta^*)^{-1/2}I(\theta^*)J(\theta^*)^{-1/2}$. Consider a power posterior π_η of the form (1.11). If either all these eigenvalues are greater than 1, or all of them are smaller than 1, then η can be chosen in such a way that the corresponding general posterior π_η provides an asymptotically better prediction than WLB. Furthermore, under the same conditions, the hyperparameters of BayesBag can be chosen in such a way that it outperforms WLB in terms of prediction. Note that the above statement holds both for Posterior Bootstrap with penalization and with pseudosamples, for any choice of hyperparameters. Thus, this result could be of independent interest, and is separate from the issue of incorporating prior information into WLB.

3 .4 Cutting feedback in Bayesian models

The techniques described above can be used to deal with model misspecification, without assuming that the user suspects which parts of the model are likely to be misspecified. Here we present methods in which the user more actively addresses the problem, using their knowledge about the nature of misspecification of their model.

We consider a model built of different components, called modules. Formally, a model built of two modules is given by

$$\begin{aligned} &x_{1,1},\dots,x_{1,n_1} \sim f_1(x_1|\theta_1), \qquad \theta_1 \sim \pi_1(\theta_1), \quad \theta_1 \in \Theta_1 \subseteq \mathbb{R}^{d_1}, \\ &x_{2,1},\dots,x_{2,n_2} \sim f_2(x_2|\theta_1,\theta_2), \quad \theta_2 \sim \pi_2(\theta_2), \quad \theta_2 \in \Theta_2 \subseteq \mathbb{R}^{d_2}. \end{aligned} \tag{1.14}$$

We denote $x_{1,1:n_1} = (x_{1,1},\dots,x_{1,n_1})$ and $x_{2,1:n_2} = (x_{2,1},\dots,x_{2,n_2})$. If we suspect that some modules are misspecified, instead of using the full Bayesian model, we may want to cut the information feedback flowing from the modules we put less trust in to the ones we trust more. This allows us to prevent the misspecification from affecting the entire model. Cutting feedback, called also modular inference, is common practice, for example, in PK/PD studies in pharmacology. Pharmacokinetics (PK) is known to be more deeply understood by the scientific community, which gives analysts more trust in the PK part of the PK/PD models. Pharmacodynamics is believed to be less understood and consequently the PD modules are often suspected to be misspecified. This motivates the following two-stage procedure used, among others, by Bennett and Wakefield (2001); Zhang et al. (2003a,b); Lunn et al. (2009). The PK part of the model is fitted first, and then the obtained results are used as input for the PD module. The above procedure could be viewed as Bayesian multiple imputation

(Little, 1992).

Another area where the cut approach is commonly used is causal inference, as the use of propensity scores (Rosenbaum and Rubin, 1983) within a two-stage estimation procedure is a special case of cutting feedback. The first module is designed to estimate the propensity score (probability of being exposed to treatment) based on covariates. The second module is typically a regression model explaining the outcome variable based on the indicator of treatment and the propensity score. As argued by McCandless et al. (2010); Zigler et al. (2013); Zigler (2016), cutting feedback in these types of models is important for obtaining unbiased estimates of the treatment effects.

Modular inference in frequently used in applied statistics when the model combines multiple data sources, often corresponding to various areas of science, such as measuring the health effects of air pollution (Blangiardo et al., 2011). It has been successfully used in climate science (Shen et al., 2016), epidemiological models (Maucort-Boulch et al., 2008; Finucane et al., 2016), econometrics (Murphy and Topel, 2002) and computer models (Liu et al., 2009). Furthermore, performing Bayesian analysis on datasets with missingness using multiple imputation (rather than considering the full Bayesian model) is also associated with cutting feedback (Little, 1992; Rubin, 2004; Zhou and Reiter, 2010).

Let us now introduce the concept of cutting feedback more formally. Firstly, recall that the joint posterior distribution $\pi(\theta_1, \theta_2 | x_{1:n_1}, x_{2,1:n_2})$ associated with model (1.14) is given by

$$\pi(\theta_1, \theta_2 | x_{1,1:n_1}, x_{2,1:n_2}) \propto \pi_1(\theta_1) \prod_{j=1}^{n_1} f_1(x_{1,j}|\theta_1) \pi_2(\theta_2) \prod_{j=1}^{n_2} f_2(x_{2,j}|\theta_1, \theta_2). \tag{1.15}$$

Therefore, the marginal distribution of (1.15) with respect to θ_1 is given by

$$\pi(\theta_1 | x_{1,1:n_1}, x_{2,1:n_2}) \propto \pi(\theta_1 | x_{1,1:n_1}) \pi(x_{2,1:n_2} | \theta_1), \tag{1.16}$$

where $\pi(x_{2,1:n_2}|\theta_1) = \int \prod_{j=1}^{n_2} f_2(x_{2,j}|\theta_1, \theta_2) \pi_2(\theta_2) \mathrm{d}\theta_2$ can be interpreted as the feedback term of the second module on θ_1. The cut approach removes this feedback, thus leading to the following formula for the cut posterior distribution:

$$\pi_{\text{cut}}(\theta_1, \theta_2) \propto \frac{\pi(\theta_1, \theta_2 | x_{1,1:n_1}, x_{2,1:n_2})}{\pi(x_{2,1:n_2} | \theta_1)}. \tag{1.17}$$

Formula (1.17) combined with (1.16) implies that the marginal distribution of the cut posterior with respect to θ_1 is $\pi(\theta_1|x_{1:1:n_1})$. As proved by Yu et al. (2021), the cut posterior $\pi_{\text{cut}}(\theta_1, \theta_2)$ minimizes the Kullback–Leibler divergence between the families of densities $q(\theta_1, \theta_2)$ having $\pi(\theta_1|x_{1:1:n_1})$ as their marginal distribution with respect to θ_1, and the full posterior $\pi(\theta_1, \theta_2)$.

There are numerous interesting questions related to modular inference, regarding its methodological, computational and theoretical aspects. Firstly, an important question to answer is whether for a given model it is more beneficial to make inference based on the full Bayesian model, or to cut feedback. Ideas on how to choose between these two approaches have been described by Jacob et al. (2017) and Yu et al. (2021). An alternative methodology, called semi-modular inference, was designed by Carmona and Nicholls (2020). Instead of choosing between the cut or the joint model, the authors propose to regulate the amount of feedback flowing from the second module to the first one, which they do using ideas borrowed from power posteriors. The semi-modular distribution is given by

$$\begin{aligned}&\pi_{\text{SMI},\eta}(\theta_1, \theta_2|x_{1,1:n_1}, x_{2,1:n_2})\\ &\propto \frac{\pi_1(\theta_1|x_{1,1:n_1})\left\{\int \pi_2(\bar{\theta}_2)\prod_{j=1}^{n_2} f_2(x_{2,j}|\theta_1, \bar{\theta}_2)^\eta d\bar{\theta}_2\right\}\pi_2(\theta_2|\theta_1, x_{2,1:n_2})}{\pi(x_{2,1:n_2}|\theta_1)}.\end{aligned}$$

Another important aspect of modular inference is related to drawing samples from the cut posterior. As noticed by Plummer (2015), there is no known way of sampling from (1.17) using a single Markov chain. Thus, sampling from the cut distribution is usually done through a two-stage procedure. The first step involves running an MCMC algorithm targeting $\pi_1(\theta_1|x_{1,1:n_1})$. At the second stage for each sample θ_1 from the chain (typically thinned) the user runs a chain targeting

$$\pi_2(\theta_2|\theta_1, x_{2,1:n_2}) \propto \pi_2(\theta_2)\prod_{j=1}^{n_2} f_2(x_{2,j}|\theta_1, \theta_2).$$

In the case of semi-modular inference, the first sampling stage involves running an MCMC chain targeting

$$\pi_{1,\text{SMI}}(\theta_1, \bar{\theta}_2) \propto \pi_1(\theta_1|x_{1,1:n_1})\pi_2(\bar{\theta}_2)\prod_{j=1}^{n_2} f_2(x_{2,j}|\theta_1, \bar{\theta}_2)^\eta$$

The samples $\bar{\theta}_2$ are then discarded. The second sampling stage is analogous to the case of modular inference described above, that is, the user runs a chain targeting $\pi_2(\theta_2|\theta_1, x_{2,1:n_2})$ for samples θ_1 obtained in the first stage.

Since this procedure is not convenient from the point of view of implementation, finding efficient and scalable methods of sampling from cut distributions is an area of active research. Methods of sampling approximately from π_{cut} proposed so far rely on adaptive MCMC (Liu and Goudie, 2020) or variational inference (Yu et al., 2021). The approach proposed by Jacob et al. (2020) is based on the unbiased MCMC technique and allows for obtaining unbiased estimates of functionals of interest with respect to π_{cut} adaptive MCMC.

3 .5 Contribution (Chapter 4)

While the asymptotic theoretical properties of maximum likelihood estimators under two-stage estimators were studied by Murphy and Topel (2002), to our knowledge the asymptotics of π_{cut} have not been analysed in the literature. In an attempt to understand better the asymptotic properties of cut distributions, we find an equivalent of the Laplace approximation for π_{cut}. Based on this result, we notice that credible regions may not asymptotically be confidence regions in the frequentist sense, thus yielding misleading uncertainty estimations. In particular, we show that even when both modules are well-specified, the cut distribution may not have the nominal frequentist coverage.

In order to overcome this issue, we propose a new methodology for performing modular inference, by extending the Posterior Bootstrap algorithms presented in Chapter 4. We prove that our method asymptotically provides the nominal frequentist coverage, and demonstrate its performance on several toy and real data examples.

Chapter 2

A Framework for Adaptive MCMC Targeting Multimodal Distributions

Abstract

We propose a new Monte Carlo method for sampling from multimodal distributions. The idea of this technique is based on splitting the task into two: finding the modes of a target distribution π and sampling, given the knowledge of the locations of the modes. The sampling algorithm relies on steps of two types: local ones, preserving the mode; and jumps to regions associated with different modes. Besides, the method learns the optimal parameters of the algorithm while it runs, without requiring user intervention. Our technique should be considered as a flexible framework, in which the design of moves can follow various strategies known from the broad MCMC literature.

In order to design an adaptive scheme that facilitates both local and jump moves, we introduce an auxiliary variable representing each mode and we define a new target distribution $\tilde{\pi}$ on an augmented state space $\mathcal{X} \times \mathcal{I}$, where $\mathcal{X}$ is the original state space of π and $\mathcal{I}$ is the set of the modes. As the algorithm runs and updates its parameters, the target distribution $\tilde{\pi}$ also keeps being modified. This motivates a new class of algorithms, Auxiliary Variable Adaptive MCMC. We prove general ergodic results for the whole class before specializing to the case of our algorithm.

1 Introduction

Poor mixing of standard Markov Chain Monte Carlo (MCMC) methods on multimodal target distributions with isolated modes is a well-described problem in statistics. Due to their dynamics, these algorithms struggle with crossing low probability barriers separating the modes and thus take a long time before moving from one mode to another, even in low dimensions. Sequential Monte Carlo (SMC) has often empirically proven to outperform MCMC on this task, its robust behaviour, however, relies strongly on the good between-mode mixing of the Markov kernel used within the SMC algorithm (see Jasra et al. (2015)). Therefore, constructing an MCMC algorithm which enables fast exploration of the state space for complicated target functions is of great interest, especially as multimodal distributions are common in applications. The examples include, but are not limited to, problems in genetics (Craiu et al., 2009; Kou et al., 2006), astrophysics (Feroz et al., 2009, 2013; Tak et al., 2017) and sensor network localization (Ihler et al., 2005).

Moreover, multimodality is an inherent issue of Bayesian mixture models (e.g. Jasra et al. (2005)), where it may be caused by label-switching, or more generally, by model identifiability issues or model misspecification (see Drton and Richardson (2004)).

Designing MCMC algorithms for sampling from a multimodal target distribution π on a state space $\mathcal{X}$ needs to address three fundamental challenges:

(1) Identifying high probability regions where the modes are located;

(2) Moving between the modes by crossing low probability barriers;

(3) Sampling efficiently within the modes by accounting for inhomogeneity between them and their local geometry.

Existing MCMC methodology for multimodal distributions usually identifies these challenges separately, and a systematic way of addressing (1-3) is not available. In this paper we introduce a unifying framework for responding to challenges (1-3) simultaneously via a novel design of Auxiliary Variable Adaptive MCMC. The framework allows us to split the sampling task into mode finding, between-region jump moves and local moves. In addition, it incorporates parameter adaptations for optimization of the local and jump kernels, and identification of local regions.

1 .1 Other approaches

Numerous MCMC methods have been proposed to address the issue of multimodality. The most popular approach is based on tempering. The idea behind this type of method relies on an observation that raising a multimodal distribution π to the power $\beta \in (0, 1)$ makes the modes "flatter" and as a result, it is more likely to accept moves to the low probability regions. Hence, it is easier to explore the state space and find the regions where the modes of π are located, addressing challenge (1) above, and also to move between these regions, addressing challenge (2). The examples of such methods, which incorporate π^β by augmenting the state space, are parallel tempering proposed by Geyer (1991) and its adaptive version (Miasojedow et al., 2013), simulated tempering (Marinari and Parisi, 1992), tempered transitions (Neal, 1996) and the equi-energy sampler (Kou et al., 2006). Despite their popularity, tempering-based approaches, as noticed by Woodard et al. (2009a), tend to mix between modes exponentially slowly in dimension if the modes have different local covariance structures. Addressing this issue is an area of active research (Tawn et al., 2020).

Another strand of research is optimization-based methods, which address challenge (1) by running preliminary optimization searches in order to identify local maxima of the target distribution. They use this information in their between-mode proposal design to overcome challenge (2). A method called Smart Darting Monte Carlo, introduced in Andricioaei et al. (2001), relies on moves of two types: jumps between the modes, allowed only in non-overlapping ϵ-spheres around the local maxima identified earlier; and local moves (Random Walk Metropolis steps). This technique was generalized by Sminchisescu and Welling (2011), by allowing the jumping regions to overlap and have an arbitrary volume and shape. The authors of Ahn et al. (2013) went one step further by introducing updates of the jumping regions and parameters of the proposal distribution at regeneration times, hence the name of their method Regeneration Darting Monte Carlo (RDMC). This includes a possibility of adding new locations of the modes at regeneration times if they are detected by optimization searches running on separate cores. Another optimization-based method, Wormhole Hamiltonian Monte Carlo, was introduced by Lan et al. (2014) as an extension of Riemanian Manifold HMC. As before, updates of the parameters of this algorithm are allowed at regeneration times.

We will see later that the algorithm we propose also falls into the category of optimization-based methods.

The Wang-Landau algorithm (Wang and Landau, 2001b,a) or its adaptive version proposed by Bornn et al. (2013) belong to the exploratory strategies that aim to push the algorithm away from well-known regions and visit new ones, hence addressing challenge (1). The multi-domain sampling technique, proposed in Zhou (2011), combines the idea of the Wang-Landau algorithm with the optimization-based approach. This algorithm relies on partitioning the state space into domains of attraction of the modes. Local moves are Random Walk Metropolis steps proposed from a distribution depending on the domain of attraction of the current state. Jumps between the modes follow the independence sampler scheme, where the new states are proposed from a mixture of Gaussian distributions approximating π.

Other common approaches include the localized normal-symmetric Random Walk Metropolis algorithm (Andrieu and Thoms, 2008), MultiNest algorithms based on nested sampling (Feroz et al., 2009, 2013), and Metropolis-Hastings algorithms with a special design of the proposal distribution accounting for the necessity of moving between the modes (Tjelmeland and Hegstad, 2001; Tak et al., 2017).

1 .2 Contribution

The importance of addressing challenge (3) stems from the fact that when modes have distinct shapes, different local proposal distributions will work well in regions associated with different modes. However, the majority of the methods described above only employ a single transition kernel, regardless of the region.

In applied problems optimal parameters of the MCMC kernels are unknown, therefore recent approaches involve tuning them while the algorithm runs. In the case of unimodal target distributions, Adaptive MCMC techniques prove to be useful (Roberts and Rosenthal, 2009; Andrieu and Thoms, 2008; Green et al., 2015). The parameters of the involved transition kernels can be learned on the fly as the simulation progresses, based on the samples observed so far. The adaptive algorithms remain ergodic under suitable regularity conditions (Andrieu and Moulines, 2006; Roberts and Rosenthal, 2007; Fort et al., 2011; Atchadé and Fort, 2010; Chimisov et al., 2018).

In the case of multimodal distributions an analogous idea would be to apply these Adaptive MCMC methods separately to regions associated with different modes, to improve the within-mode mixing. Note that in order to sample from different proposal distributions in regions associated with different modes, one needs to control at each step of the algorithm

which region the current state belongs to. Besides, adapting parameters of the local proposal distributions on the fly must be based on samples that actually come from the corresponding region.

The multi-domain sampler (Zhou, 2011) discussed earlier has a mechanism of assigning samples to regions. However, in their setting keeping track of the regions requires running a gradient ascent procedure at each MCMC step, which imposes a high computational burden on the whole algorithm. Other optimization-based approaches known in the literature (e.g. Sminchisescu and Welling (2011) and Ahn et al. (2013)) tend to ignore the necessity of assigning samples to regions and the possibility of moving between the modes via local steps.

Another issue is that adaptive optimization-based methods presented above, such as those of Ahn et al. (2013) and Lan et al. (2014), allow for adaptations only at regeneration times. In high dimensions regenerations happen rarely, which in practice makes the adaptive scheme prohibitively inefficient. Besides, identifying regeneration times using the method of Mykland et al. (1995), as authors of both algorithms propose, requires case-specific calculations which precludes any generic implementation.

We aim to remedy these shortcomings by proposing a framework for designing an adaptive algorithm on an augmented state space $\mathcal{X} \times \mathcal{I}$, where $\mathcal{I} = \{1, \ldots, N\}$, and the auxiliary variable i of the resulting sample (x, i) encodes the corresponding region for x. Local MCMC kernels update x only, while jump kernels that move between the modes update x and i simultaneously. Furthermore, the design of the target distribution on the augmented state space prevents the algorithm from moving to a region associated with a different mode via local steps. We note that we do not define here formally regions associated with the modes, and it may happen that the chain samples both (x, i) and (x, j) for $i \neq j$. In practice this will happen when modes i and j are not well-separated.

In the sequel we make specific choices for the adaptive scheme, the local and jump kernels, as well as the burn-in routine used for setting up initial values of the parameters of the algorithm. However, the design is modular and different approaches can be incorporated in the framework. Besides, it allows for a multicore implementation of a large part of the algorithm.

This approach motivates introducing a new class of algorithms, Auxiliary Variable Adaptive MCMC, where not only transition kernels are allowed to be modified on the fly, but

also the augmented target distributions. It turns out that apart from our method, there is a wide range of algorithms that belong to this class, including adaptive parallel tempering or adaptive versions of pseudo-marginal MCMC. Thus our general ergodicity results, proved for the whole class under standard regularity conditions, can potentially be useful for analysing other methods.

The remainder of the paper is organized as follows. In Section 2 we present our algorithm, the Jumping Adaptive Multimodal Sampler (JAMS), and discuss its properties. In Section 3 we define the Auxiliary Variable Adaptive MCMC class and establish convergence in distribution and a Weak Law of Large Numbers for this class, under the uniform and the non-uniform scenario. We present theoretical results specialized to the case of our proposed algorithm in Section 4 . Ergodicity is derived here from the analogues of the Containment and Diminishing Adaptation conditions introduced in Roberts and Rosenthal (2007), as opposed to identifying regeneration times, which allows us to circumvent the issues described above. The proofs of all our theorems along with some additional comments about the theoretical results are gathered in Supplementary Material A . Section 5 demonstrates the performance of our method on two synthetic and one real data example. Additional details of our numerical experiments are available in Supplementary Material B . We conclude with a summary of our results and a discussion in Section 6 .

2 Jumping Adaptive Multimodal Sampler (JAMS)

2 .1 Main algorithm

Let π be the multimodal target distribution of interest defined on $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$. We introduce a collection of target distributions $\{\tilde{\pi}_\gamma\}_{\gamma\in\mathcal{Y}}$ on the augmented state space $\mathcal{X} \times \mathcal{I}$, where $\mathcal{I} := \{1, \ldots, N\}$ is a finite set of indices of the modes of π. We defer the discussion about finding the modes to Section 2 .2. Here γ denotes the design parameter of the algorithm that may be adapted on the fly. For a fixed $\gamma \in \mathcal{Y}$, $\tilde{\pi}_\gamma$ is defined as

$$\tilde{\pi}_\gamma(x, i) := \pi(x) \frac{w_{\gamma,i} Q_i(\mu_i, \Sigma_{\gamma,i})(x)}{\sum_{j\in\mathcal{I}} w_{\gamma,j} Q_j(\mu_j, \Sigma_{\gamma,j})(x)}, \tag{2.1}$$

where $Q_i(\mu_i, \Sigma_{\gamma,i})$ is an elliptical distribution (such as the normal or the multivariate t distribution) centred at μ_i with covariance matrix $\Sigma_{\gamma,i}$. We shall think of $\{\mu_i\}_{i\in\mathcal{I}}$ and $\{\Sigma_{\gamma,i}\}_{i\in\mathcal{I}}$ as locations and covariances of the modes of π, respectively. Firstly, notice that constructing a Markov chain targeting $\tilde{\pi}_\gamma$ provides a natural way of identifying the mode at each step by recording the auxiliary variable i. Besides, for each $B \in \mathcal{B}(\mathcal{X})$ and $\gamma \in \mathcal{Y}$ we have $\tilde{\pi}_\gamma(B \times \mathcal{I}) = \pi(B)$. Hence, π is the marginal distribution of $\tilde{\pi}_\gamma$ for each $\gamma \in \mathcal{Y}$, so sampling from $\tilde{\pi}_\gamma$ can be used to generate samples from π.

The sampling algorithm that we propose is summarized in Algorithm 1. The method relies on MCMC steps of two types, performed with probabilities $1-\epsilon$ and ϵ, respectively.

- **Local move**: Given the current state of the chain (x, i) and the current parameter γ, a local kernel $\tilde{P}_{\gamma,L,i}$ invariant with respect to $\tilde{\pi}_\gamma$ is used to update x, while i remains fixed, hence the mode is preserved.

- **Jump move**: Given the current state of the chain (x, i) and the current parameter γ, a new mode k is proposed with probability $a_{\gamma,ik}$. Then a new point y is proposed using a distribution $R_{\gamma,J,ik}(x, \cdot)$. The new pair is accepted or rejected using the standard Metropolis-Hastings formula such that the jump kernel is invariant with respect to $\tilde{\pi}_\gamma$.

Our choice for the local kernel is Random Walk Metropolis (RWM) with proposal $R_{\gamma,L,i}(x, \cdot)$ that follows either the normal or the t distribution. This allows us to employ well-developed adaptation strategies for RWM and build on its stability properties to establish ergodicity of JAMS in Section 4 . The acceptance probability formula for local moves is given by:

$$\alpha_{\gamma,L}\left((x,i) \to (y,i)\right) = \min\left[1, \frac{\pi(y)Q_i(\mu_i, \Sigma_{\gamma,i})(y)}{\pi(x)Q_i(\mu_i, \Sigma_{\gamma,i})(x)} \frac{\sum_{j\in\mathcal{I}} w_{\gamma,j} Q_j(\mu_j, \Sigma_{\gamma,j})(x)}{\sum_{j\in\mathcal{I}} w_{\gamma,j} Q_j(\mu_j, \Sigma_{\gamma,j})(y)}\right]. \tag{2.2}$$

As for the jump moves, we consider two different methods of proposing a new point y associated with mode k. The first one, which we call *independent proposal jumps*, is to draw y from an elliptical distribution centred at μ_k with covariance matrix $\Sigma_{\gamma,k}$, independently from the current point (x, i). Since there is no dependence on x and i, in the case of independent proposal jumps the proposal distribution to mode k will be denoted by $R_{\gamma,J,k}(\cdot)$. For independent proposal jumps, the acceptance probability is equal to

$$\alpha_{\gamma,J}\left((x,i) \to (y,k)\right) = \min\left[1, \frac{\tilde{\pi}_\gamma(y,k)}{\tilde{\pi}_\gamma(x,i)} \frac{a_{\gamma,ki} R_{\gamma,J,i}(x)}{a_{\gamma,ik} R_{\gamma,J,k}(y)}\right]. \tag{2.3}$$

Alternatively, given that the current state is (x, i), we can propose a "corresponding" point y in mode k, such that the Mahalanobis distance between the point and its corresponding mode is preserved, that is

$$(x - \mu_i)^T \Sigma_{\gamma,i}^{-1} (x - \mu_i) = (y - \mu_k)^T \Sigma_{\gamma,k}^{-1} (y - \mu_k).$$

The required equality is satisfied for

$$y := \mu_k + \Lambda_{\gamma,k} \Lambda_{\gamma,i}^{-1} (x - \mu_i), \tag{2.4}$$

where

$$\Sigma_{\gamma,i} = \Lambda_{\gamma,i} \Lambda_{\gamma,i}^T \quad \text{and} \quad \Sigma_{\gamma,k} = \Lambda_{\gamma,k} \Lambda_{\gamma,k}^T.$$

Herein this method will be called *deterministic jumps*. The acceptance probability is then given by

$$\alpha_{\gamma,J}\left((x, i) \to (y, k)\right) = \min\left[1, \frac{\tilde{\pi}_\gamma(y, k)}{\tilde{\pi}_\gamma(x, i)} \frac{a_{\gamma,ki} \sqrt{\det \Sigma_{\gamma,k}}}{a_{\gamma,ik} \sqrt{\det \Sigma_{\gamma,i}}}\right]. \tag{2.5}$$

In both cases the design of the jump moves takes into account the shapes of the two modes involved, which helps achieving high acceptance rates and consequently improves the between-mode mixing.

As presented in Algorithm 1, the method involves learning the parameters on the fly. We design an adaptation scheme of three lists of parameters: covariance matrices (used both for adapting the target distribution $\tilde{\pi}_\gamma$ and the proposal distributions), weights $w_{\gamma,i}$ and probabilities $a_{\gamma,ik}$ of proposing mode k in a jump from mode i. Hence, formally $\mathcal{Y}$ refers to the product space of $\Sigma_{\gamma,i}$, $w_{\gamma,i}$ and $a_{\gamma,ik}$ for $i, k \in \{1, \ldots, N\}$ restricted by $\sum_{j \in \mathcal{I}} w_{\gamma,j} = 1$ and $\sum_{k \in \mathcal{I}} a_{\gamma,ik} = 1$ for each $\gamma \in \mathcal{Y}$ and each $i \in \mathcal{I}$. In Algorithm 1 we let $m_1, \ldots, m_N$ and $S_1, \ldots, S_N$ denote the empirical means and empirical covariance matrices, respectively, based on samples collected so far in the chain for modes $1, \ldots, N$.

An adaptive scheme for $w_{\gamma,i}$ and $a_{\gamma,ik}$ that follows an intuitive heuristic is discussed briefly in Supplementary Material B .

Our method of adapting the covariance matrices $\Sigma_{\gamma,i}$ is presented in Algorithm 2. For every $i \in \mathcal{I}$ the matrix $\Sigma_{\gamma,i}$ is based on the empirical covariance matrix of the samples from the region associated with mode i obtained so far. This is possible in our framework by keeping track of the auxiliary variable i. Updates are performed every certain number of

Algorithm 1 JAMS: main algorithm, iteration $n+1$

Input: current point (x_n, i_n), list of modes $\{\mu_1, \dots \mu_N\}$, constant $\epsilon \in (0,1)$, parameter $\gamma_n = \{\Sigma_i, w_i, a_{ik}\}_{i,k \in \{1,\dots,N\}}$, empirical means $m_1, \dots, m_N$ and empirical covariance matrices $S_1, \dots, S_N$.
Generate $u \sim U[0,1]$.
if $u > \epsilon$ **then**
    Local move:
    Propose a new value $y \sim R_{\gamma_n, L, i_n}(x_n, \cdot)$.
    Accept y with probability $\alpha_{\gamma_n, L}\left((x_n, i_n) \to (y, i_n)\right)$.
    if y accepted **then**
        $(x_{n+1}, i_{n+1}) := (y, i_n)$.
    else
        $(x_{n+1}, i_{n+1}) := (x_n, i_n)$.
    end if
else
    Jump move:
    Propose a new mode $k \sim (a_{i1}, \dots, a_{iN})$.
    Propose a new value $y \sim R_{\gamma_n, J, ik}(x_n, \cdot)$.
    Accept (y, k) with probability $\alpha_{\gamma_n, J}\left((x_n, i_n) \to (y, k)\right)$.
    if (y, k) accepted **then**
        $(x_{n+1}, i_{n+1}) := (y, k)$.
    else
        $(x_{n+1}, i_{n+1}) := (x_n, i_n)$.
    end if
end if
Update the empirical mean $m_{i_{n+1}}$ and the empirical covariance matrix $S_{i_{n+1}}$ by including x_{n+1}.
Update the parameter γ_n to γ_{n+1} according to Algorithm 2.
return New sample (x_{n+1}, i_{n+1}), parameter γ_{n+1}, $m_{i_{n+1}}$ and $S_{i_{n+1}}$.

iterations (denoted by AC_2 in Algorithm 2). This method follows the classical Adaptive Metropolis methodology (cf. Haario et al. (2001); Roberts and Rosenthal (2009)) applied separately to the covariance structure associated with each mode. For the local proposal distributions the covariance matrices are additionally scaled by the factor $2.38^2/d$, which is commonly used as optimal for Adaptive Metropolis algorithms (Gelman et al., 1997; Roberts and Rosenthal, 2001). Since representing a covariance matrix in high dimensions reliably typically requires a large number of samples, we do not apply this method straight away. Instead, we perform adaptive scaling, aiming to achieve the optimal acceptance rate (typically fixed at 0.234; see Gelman et al. (1997); Roberts and Rosenthal (2001)) for local moves, until the number of samples observed in a given mode exceeds a pre-specified constant (denoted by AC_1 in Algorithm 2).

It is worth outlining that this special construction of the target distribution $\tilde{\pi}_\gamma$ makes it unlikely for the algorithm to escape via local steps from the mode it is assigned to and settle

Algorithm 2 Updating the parameters, iteration $n+1$

Input: (in addition to the parameters of iteration $n+1$ of the main algorithm) number of samples observed so far in each mode $n_1, \ldots, n_N$, auxiliary matrices $\tilde{\Sigma}_1, \ldots, \tilde{\Sigma}_N$, positive integers AC_1 and AC_2, constants $\alpha, \alpha_{\text{opt}} \in (0,1)$, $\beta > 0$.
if $n_{i_{n+1}} < AC_1$ **then**
  if Local move then
    $\tilde{\Sigma}_{i_{n+1}} := \exp\left(n_{i_{n+1}}^{-\alpha}\left(\alpha_{\gamma_n,L} - \alpha_{\text{opt}}\right)\right)\tilde{\Sigma}_{i_{n+1}}$.
    $\Sigma_{\gamma_{n+1},i_{n+1}} := \tilde{\Sigma}_{i_{n+1}} + \beta I_d$.
  end if
else
  if $n_{i_{n+1}}$ is divisible by AC_2 **then**
    $\Sigma_{\gamma_{n+1},i_{n+1}} := S_{i_{n+1}} + \beta I_d$.
    Update $w_{\gamma_n,i}$ and $a_{\gamma_n,ik}$ to $w_{\gamma_{n+1},i}$ and $a_{\gamma_{n+1},ik}$ for $i,k = 1,\ldots,N$.
  end if
end if

in another mode included in the augmented target distribution. Indeed, if a proposed point y is very distant from the current mode μ_i and close to another mode μ_k, the acceptance probability becomes very small due to the expression $Q_i(\mu_i, \Sigma_{\gamma,i})(y)$ in the numerator of (2.2) and $Q_k(\mu_k, \Sigma_{\gamma,k})(y)$ in the denominator, as $Q_i(\mu_i, \Sigma_{\gamma,i})(y)$ will typically be tiny in such case and $Q_k(\mu_k, \Sigma_{\gamma,k})(y)$ will be large. We note that the above property only refers to known modes, included in the formula for $\tilde{\pi}$. It may happen that some modes are missed by the initial optimization runs – in such cases JAMS performs on regions around the unnoticed modes like a standard local MCMC sampler.

2 .2 Burn-in algorithm

Algorithm 1 takes mode locations $\{\mu_1, \ldots, \mu_N\}$ and initial values of the matrices

$$\{\Sigma_{\gamma_0,1}, \ldots, \Sigma_{\gamma_0,N}\}$$

as input. Since further improvements in the estimation of $\Sigma_{\gamma,i}$ are possible only after some samples in mode i have been observed, matrices

$$\{\Sigma_{\gamma_0,1}, \ldots, \Sigma_{\gamma_0,N}\}$$

need to represent well the shapes of the corresponding modes so that jumps to all the modes are accepted reasonably quickly. We address the issue of setting up these values by introducing a burn-in algorithm, which runs before the main MCMC sampler (Algorithm 1)

is initiated. Alternatively, one may prefer a version of this method in which the burn-in routine continues running in parallel to the main sampler on multiple cores, e.g. to include modes possibly missed in the initial run. We sketch different stages of the burn-in routine below, additional details are given in Supplementary Material B .

Starting points for the optimization procedure

We sample the starting points for optimization searches uniformly on a compact set which is a product of intervals $[L_1, U_1] \times \ldots \times [L_d, U_d]$, provided by the user. Note that if the domain of attraction of each mode overlaps with $[L_1, U_1] \times \ldots \times [L_d, U_d]$, then asymptotically all modes will be found, as we will have at least one starting point in each domain.

When dealing with Bayesian models, one can alternatively sample the starting points from the prior distribution.

Mode finding via an optimization procedure

The BFGS optimization algorithm (Nocedal and Wright, 2006) is initiated from every starting point. The BFGS method method provides the optimum point and the Hessian matrix at this point which is particularly useful in the next step of mode merging.

For numerical reasons, instead of working directly with π, we typically use the BFGS algorithm to find the local minima of $-\log(\pi)$.

Mode merging

Starting the optimization procedure from different points belonging to the same basin of attraction will take us to points which are close to the true local maxima, but numerically different, an issue that seems to be ignored in optimization-based MCMC literature.

We deal with this by classifying two vectors m_i and m_j as corresponding to the same mode if the squared Mahalanobis distance between them is smaller than some pre-specified value. If we let H_i and H_j denote the Hessian matrices of $-\log(\pi)$ at m_i and m_j, respectively, the above Mahalanobis distance is calculated for H_i^{-1} and H_j^{-1} (for symmetry, we average over these two values). This method is scale invariant as the Hessian captures the local shape and scale.

An alternative approach, which we however do not explore further in this paper, would be to merge modes if the local sampler can move efficiently between them. In this case,

jumps would only be performed between well-separated regions of the state space.

Initial covariance matrix estimation

In order to find initial covariance matrix estimates $\Sigma_{\gamma_0,1}, ..., \Sigma_{\gamma_0,N}$ that accurately reflect the geometry of different modes, we employ the augmented target machinery of Algorithm 1 in the following way. We run Algorithm 1 without jumps, i.e. with $\epsilon = 0$, in parallel, starting from each of the modes $\mu_1, \ldots, \mu_N$. This implies that we run N chains and each of them adapts only the matrix Σ_i corresponding to the mode μ_i which was its starting point. We make a number of rounds of this procedure, and after each round we update the target distribution $\tilde{\pi}$ by exchanging the knowledge about the adapted covariance matrices between cores. The final covariance matrices passed to the main MCMC sampler are calculated based on the samples collected in all rounds.

The reason why we exchange information between rounds, despite the additional cost of communication between cores, is that we want to keep each chain close to its corresponding mode thanks to the acceptance probability in (2.2). The algorithm is then less likely to visit regions corresponding to other modes and contaminate the estimate. Essentially, the initial covariance estimation revisits the problem of collecting samples only from the corresponding regions, discussed earlier.

The initial value of the matrix corresponding to mode i is the inverse of the Hessian evaluated at μ_i. The values of $w_{\gamma,i}$ and $a_{\gamma,ik}$ are set to $1/N$ and not updated during those runs. In Supplementary Material B we propose a method of choosing the number of rounds automatically based on monitoring the inhomogeneity factor (see Roberts and Rosenthal (2009) and Rosenthal (2011)).

2 .3 Further comments

It is important to point out that the auxiliary variable approach presented above should be thought of as a flexible framework rather than one specific method. The BFGS algorithm used for mode finding could be replaced with another optimization procedure. Similarly, instead of the Random Walk Metropolis algorithm, local moves could be performed with a different MCMC sampler, such as HMC or MALA . One could also consider another scheme for updating the parameters, for example, combining adaptive scaling with covariance matrix estimation (see Vihola (2011)).

Moreover, we note that the way of handling different stages of the burn-in algorithm, described in the previous section, was done from the point of view of the user who does not have any prior knowledge about the locations of the modes. However, in applied examples, a thorough understanding of the studied problem often offers a good guidance as for where in the state space one should expect to have large probability masses. This knowledge should be taken into account in the burn-in algorithm, e.g. in the choice of the starting points for optimization searches.

A notable example here is mixture models, which often lead to multimodal posteriors. The symmetry of the posterior in mixture models will typically help find the modes of the target. That is, if a mode has been found, the user can add other modes by permuting certain coordinates. It should be noted that in such cases we advise incorporating all modes, including those obtained via permutation, in the augmented target distribution, rather than working with restricted state spaces.

3 Auxiliary Variable Adaptive MCMC

We introduce a general class of Auxiliary Variable Adaptive MCMC algorithms, as follows.

Recall that $\pi(\cdot)$ is a fixed target probability density on $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$. For an auxiliary pair $(\Phi, \mathcal{B}(\Phi))$, define $\tilde{\mathcal{X}} := \mathcal{X} \times \Phi$, and for an index set $\mathcal{Y}$, consider a family of probability measures $\{\tilde{\pi}_\gamma(\cdot)\}_{\gamma\in\mathcal{Y}}$ on $(\tilde{\mathcal{X}}, \mathcal{B}(\tilde{\mathcal{X}}))$. Moreover, let $\{\tilde{P}_\gamma\}_{\gamma\in\mathcal{Y}}$ be a collection of Markov chain transition kernels on $(\tilde{\mathcal{X}}, \mathcal{B}(\tilde{\mathcal{X}}))$. We say that an algorithm belongs to the Auxiliary Variable Adaptive MCMC if the following two conditions hold.

1. The family of distributions $\{\tilde{\pi}_\gamma(\cdot)\}_{\gamma\in\mathcal{Y}}$ satisfies

$$\tilde{\pi}_\gamma(B \times \Phi) = \pi(B) \quad \text{for every } B \in \mathcal{B}(\mathcal{X}) \text{ and } \gamma \in \mathcal{Y}. \tag{2.6}$$

2. Each $\tilde{P}_\gamma$ has $\tilde{\pi}_\gamma$ as its invariant distribution and is Harris ergodic.

It can be easily checked that JAMS belongs to the class defined above. There exist other algorithms falling into this category, therefore the results presented later in this paper, in particular Theorems 2.2, 2.3 and 2.4, may be useful for analysing their ergodicity. Examples of other algorithms in this class include adaptive parallel tempering (Miasojedow et al., 2013) and adaptive versions of pseudo-marginal algorithms (Andrieu and Roberts, 2009; Andrieu

and Vihola, 2015). A more detailed discussion on this may be found in Supplementary Material A .

In order to define the dynamics of the Auxiliary Variable Adaptive MCMC sequence $\{(\tilde{X}_n, \Gamma_n)\}_{n=0}^{\infty}$, where Γ represents a random variable taking values in $(\mathcal{Y}, \mathcal{B}(\mathcal{Y}))$, denote its filtration as

$$\mathcal{G}_n := \sigma\{\tilde{X}_0, \ldots, \tilde{X}_n, \Gamma_0, \ldots, \Gamma_n\}.$$

The conditional distribution of Γ_{n+1} given $\mathcal{G}_n$ will be specified by the adaptive algorithm being used, such as Algorithm 1, while the dynamics of the $\tilde{X}$ coordinate follows

$$\mathbb{P}[\tilde{X}_{n+1} \in \tilde{B} | \tilde{X}_n = \tilde{x}, \Gamma_n = \gamma, \mathcal{G}_{n-1}] = \tilde{P}_\gamma(\tilde{x}, \tilde{B}) \tag{2.7}$$

for $\tilde{x} \in \tilde{\mathcal{X}}, \gamma \in \mathcal{Y}, \tilde{B} \in \mathcal{B}(\tilde{\mathcal{X}})$. Note that depending on the adaptive update rule for Γ_n, the sequence $\{(\tilde{X}_n, \Gamma_n)\}_{n=0}^{\infty}$, defined above is not necessarily a Markov chain. By $\tilde{A}_n(\cdot)$ denote the distribution of the $\tilde{\mathcal{X}}$-marginal of $\{(\tilde{X}_n, \Gamma_n)\}_{n=0}^{\infty}$ at time n, conditionally on the starting points, i.e.

$$\tilde{A}_n(\tilde{B}) := \mathbb{P}[\tilde{X}_n \in \tilde{B} | \tilde{X}_0 = \tilde{x}, \Gamma_0 = \gamma] \quad \text{for } \tilde{B} \in \mathcal{B}(\tilde{\mathcal{X}}).$$

By $A_n(\cdot)$ denote the further marginalization of $\tilde{A}_n(\cdot)$ onto the space of interest $\mathcal{X}$, namely

$$A_n(B) := \tilde{A}_n(B \times \Phi) \quad \text{for } B \in \mathcal{B}(\mathcal{X}).$$

Finally, in order to define ergodicity of the Auxiliary Variable Adaptive MCMC, let

$$T_n(\tilde{x}, \gamma) := \|A_n(\cdot) - \pi(\cdot)\|_{TV} = \sup_{B \in \mathcal{B}(\mathcal{X})} |A_n(B) - \pi(B)|.$$

Definition 2.1. *We say that the Auxiliary Variable Adaptive MCMC algorithm generating* $\{(\tilde{X}_n, \Gamma_n)\}_{n=0}^{\infty}$, *is* ergodic, *if*

$$\lim_{n \to \infty} T_n(\tilde{x}, \gamma) = 0 \quad \textit{for all} \quad \tilde{x} \in \tilde{\mathcal{X}}, \gamma \in \mathcal{Y}.$$

3 .1 Theoretical results for the class

The two main approaches to verifying ergodicity of Adaptive MCMC are based on martingale approximations (Andrieu and Moulines, 2006; Fort et al., 2011; Atchadé and Fort, 2010) or

coupling (Roberts and Rosenthal, 2007). Here we extend the latter to the Auxiliary Variable Adaptive MCMC class by constructing explicit couplings. In particular, ergodicity of this class of algorithms will be verified for the uniform and the non-uniform case, providing results analogous to Theorems 1 and 2 of Roberts and Rosenthal (2007).

For the uniform case analogues of the usual conditions of Simultaneous Uniform Ergodicity and Diminishing Adaptation will be required.

Theorem 2.2 (Ergodicity – uniform case)**.** *Consider an Auxiliary Variable Adaptive MCMC algorithm on a state space $\tilde{\mathcal{X}} = \mathcal{X} \times \Phi$, following dynamics (2.7) with a family of transition kernels $\{\tilde{P}_\gamma\}_{\gamma\in\mathcal{Y}}$ satisfying (2.6) and such that $\tilde{P}_\gamma$ is $\tilde{\pi}_\gamma$-ergodic for each $\gamma \in \mathcal{Y}$. If conditions (a) and (b) below are satisfied, then the algorithm is ergodic in the sense of Definition 2.1.*

(a) (Simultaneous Uniform Ergodicity). For all $\varepsilon > 0$, there exists $N = N(\varepsilon) \in \mathbb{N}$ such that

$$\|\tilde{P}_\gamma^N(\tilde{x}, \cdot) - \tilde{\pi}_\gamma(\cdot)\|_{TV} \leq \varepsilon, \qquad \text{for all} \quad \tilde{x} \in \tilde{\mathcal{X}} \quad \text{and} \quad \gamma \in \mathcal{Y}.$$

(b) (Diminishing Adaptation). The random variable

$$D_n := \sup_{\tilde{x}\in\tilde{\mathcal{X}}} \|\tilde{P}_{\Gamma_{n+1}}(\tilde{x}, \cdot) - \tilde{P}_{\Gamma_n}(\tilde{x}, \cdot)\|_{TV} \tag{2.8}$$

converges to 0 in probability.

In fact assumption (a) of Theorem 2.2 can be relaxed. To this end, we define the $\varepsilon-$convergence time as

$$M_\varepsilon(\tilde{x}, \gamma) := \inf\{k \geq 1 : \|\tilde{P}_\gamma^k(\tilde{x}, \cdot) - \tilde{\pi}_\gamma(\cdot)\|_{TV} \leq \varepsilon\}. \tag{2.9}$$

It is enough that the random variable $M_\varepsilon(\tilde{X}_n, \Gamma_n)$ is bounded in probability. Precisely, the following ergodicity result holds for the non-uniform case.

Theorem 2.3 (Ergodicity – non-uniform case)**.** *Consider an Auxiliary Variable Adaptive MCMC algorithm, under the assumptions of Theorem 2.2 and replace condition (a) with the following:*

(a) (Containment). For all $\varepsilon > 0$ and all $\tilde{\delta} > 0$, there exists $N = N(\varepsilon, \tilde{\delta})$ such that

$$\mathbb{P}\left(M_\varepsilon(\tilde{X}_n, \Gamma_n) > N | \tilde{X}_0 = \tilde{x}, \Gamma_0 = \gamma\right) \leq \tilde{\delta} \quad \text{for all } n \in \mathbb{N}. \tag{2.10}$$

Then the algorithm is ergodic in the sense of Definition 2.1.

In our proof of Theorems 2.2 and 2.3 we introduce two auxiliary processes $\{(\tilde{X}_n^m, \Gamma_n^m)\}_{n=0}^{\infty}$ and $\{(\tilde{X}_n^i, \Gamma_n^i)\}_{n=0}^{\infty}$ on $\mathcal{X} \times \Phi$, to be thought of as "Markovian" and "intermediate". The latter interpolates between the Markovian one and the original adaptive process. Crucial for our proof is the coupling construction, which shows that the total variation distance between the two pairs of processes goes to 0. We then observe that the total variation distance between processes on $\mathcal{X} \times \Phi$ bounds from above the distance between their marginalizations onto $\mathcal{X}$. Finally, we use the triangle inequality to conclude ergodicity of the adaptive process. All the processes are constructed explicitly on the same probability space, which makes our proofs adaptable to other cases.

We use the same coupling construction to establish the Weak Law of Large Numbers for the class of Auxiliary Variable Adaptive MCMC algorithms for both the uniform and the non-uniform case (corresponding to the scenario of Theorem 2.2 and 2.3, respectively). By letting Φ be a singleton, our result applies to the standard Adaptive MCMC setting and extends the result of Roberts and Rosenthal (2007) where the WLLN was provided for the uniform case only.

Theorem 2.4 (WLLN)**.** *Consider an Auxiliary Variable Adaptive MCMC algorithm, as in Theorem 2.3, together with assumptions a) and b) of this theorem. Let $g : \mathcal{X} \to \mathbb{R}$ be a bounded measurable function. Then*

$$\frac{\sum_{i=1}^{n} g(X_i)}{n} \to \pi(g) \quad \textit{in probability as } n \to \infty.$$

While Containment is a weaker condition than Simultaneous Uniform Ergodicity, it is less tractable and in the standard Adaptive MCMC setting drift conditions are typically used to verify it (Roberts and Rosenthal, 2007; Bai et al., 2011b). Lemma 2.5 helps verify Containment via geometric drift conditions in the Auxiliary Variable framework.

Lemma 2.5. *Assume that the following conditions are satisfied.*

a) For each $\gamma \in \mathcal{Y}$ $\|\tilde{P}_\gamma^k(\tilde{x}, \cdot) - \tilde{\pi}_\gamma(\cdot)\|_{TV} \to 0$ as $k \to \infty$.

b) There exists $\lambda < 1$, $b < \infty$ and a collection of functions $V_{\tilde{\pi}_\gamma} : \tilde{\mathcal{X}} \to [1, \infty)$ for $\gamma \in \mathcal{Y}$, such that the following simultaneous drift condition is satisfied:

$$\tilde{P}_\gamma V_{\tilde{\pi}_\gamma}(\tilde{x}) \leq \lambda V_{\tilde{\pi}_\gamma}(\tilde{x}) + b \quad \textit{for all } \tilde{x} \in \tilde{\mathcal{X}} \textit{ and } \gamma \in \mathcal{Y}, \tag{2.11}$$

where for $\tilde{x} \in \tilde{\mathcal{X}}$

$$\tilde{P}_\gamma V_{\tilde{\pi}_\gamma}(\tilde{x}) := \mathbb{E}\left(V_{\tilde{\pi}_\gamma}(\tilde{X}_{n+1}) | \tilde{X}_n = \tilde{x}, \Gamma_n = \gamma\right).$$

Moreover, $V_{\tilde{\pi}_\gamma}(\tilde{x})$ *is bounded on compact sets as a function of* $(\tilde{x}, \gamma)$.

c) *There exist* $\delta > 0$, $v > 2n_0 b/(1 - \lambda^{n_0})$ *and a positive integer* n_0, *such that the following minorization condition holds: for each* $\gamma \in \mathcal{Y}$ *we can find a probability measure* ν_γ *on* $\tilde{\mathcal{X}}$ *satisfying*

$$\tilde{P}_\gamma^{n_0}(\tilde{x}, \cdot) \geq \delta \nu_\gamma(\cdot) \quad \textit{for all } \tilde{x} \textit{ with } V_{\tilde{\pi}_\gamma}(\tilde{x}) \leq v. \tag{2.12}$$

d) $\mathcal{Y}$ *is compact in some topology.*

e) *There exists a compact set* A *such that if* $X_n \notin A$, *then* $\Gamma_{n+1} = \Gamma_n$.

f) $\mathbb{E} V_{\tilde{\pi}_{\Gamma_0}}(\tilde{X}_0) < \infty$.

Then the Containment condition (2.10) *holds.*

Lemma 2.5 bears resemblance to Theorem 3 of Roberts and Rosenthal (2007). However, in our setting drift functions need to depend on the target distribution $\tilde{\pi}_\gamma$, which modifies the form of the standard drift condition and as a consequence, we cannot use the proof of this theorem directly to obtain our result. In particular, our lemma additionally assumes that the adaptation happens on a compact set only. Intuitively, condition e) is imposed to facilitate bounding from above the following expression

$$\frac{E\left(V_{\tilde{\pi}_{\Gamma_{n+1}}}(\tilde{X}_{n+1}) | \tilde{X}_n, \Gamma_n\right)}{V_{\tilde{\pi}_{\Gamma_n}}(\tilde{X}_n)}. \tag{2.13}$$

We deal with (2.13) separately for $\tilde{X}_n \in A$ and for $\tilde{X}_n \notin A$. In the first case the upper bound can be obtained by the assumption that $V_{\tilde{\pi}_\gamma}$ is bounded on compact sets as a function of $(\tilde{x}, \gamma)$. In the second case, in order to bound (2.13) we can use directly the drift condition (2.11) as we have $\Gamma_{n+1} = \Gamma_n$.

3 .2 Adaptive Increasingly Rarely version of the class

Adaptive Increasingly Rarely (AIR) MCMC algorithms were introduced in Chimisov et al. (2018) as an alternative to classical Adaptive MCMC methods. While they share the same self-tuning properties, their ergodic properties are mathematically easier to analyse and their computational cost of adaptation is smaller.

The key idea behind the AIR algorithms is to allow for updates of parameters only at pre-specified times N_j with an increasing sequence of lags n_k between them. For the sequence $\{n_k\}_{k>1}$ Chimisov et al. (2018) proposed using any scheme that satisfies $c_2 k^\kappa \geq n_k \geq c_1 k^\kappa$ for some positive c_1, c_2 and κ. In order to ensure that the random variable D_n defined by (2.8) converges to 0 in probability, i.e. Diminishing Adaptation holds, the following modification is introduced. The updates happen at times N_j^*, where

$$N_j^* = \sum_{k=1}^{j} n_k^* \quad \text{with } N_0^* = 0$$

$$\text{and} \quad n_k^* = n_k + \text{Uniform}[0, \lfloor k^{\kappa^*} \rfloor] \quad \text{for some } \kappa^* \in (0, \kappa) \text{ and } n_0^* = 0.$$

Observe that D_n is only positive if $n+1 \in \{N_j^*\}_{j\geq 1}$. Besides, if $n+1 > N_k$ then $\mathbb{P}(D_n > 0) \leq \frac{1}{\lfloor k^{\kappa^*} \rfloor}$, so in particular D_n goes to 0 as n tends to infinity.

We apply the same idea to Auxiliary Variable Adaptive MCMC algorithms, by adapting the parameters of the transition kernels and the target distributions only at times N_j^*, as described above, so that Diminishing Adaptation is automatically satisfied for these algorithms. In Section 4 we study in detail an AIR version of JAMS.

4 Ergodicity of the Jumping Adaptive Multimodal Sampler

The main results of this section are stated in Theorems 2.6 and 2.7, which establish convergence of our algorithm to the correct limiting distribution under the uniform and the non-uniform scenario, respectively. To prove these theorems, we will use our results from Section 3 .

4 .1 Overview of the assumptions

In order to prove ergodic results for JAMS, we consider a slightly modified version of Algorithm 1, called mJAMS herein. While being easier to analyse mathematically, it inherits the

main properties of Algorithm 1. The modifications are two-fold:

1. We update the parameters only if the most recent sample (x_n, i_n) is such that x_n belongs to some fixed compact set A_{i_n}. Note that these sets can be adapted but only as long as they are contained within some larger pre-specified compact set.

2. We adapt the parameters "increasingly rarely" (see Section 3 .2).

In Supplementary Material A we present the pseudocode of the resulting algorithm and make a remark on how to choose the compact sets $A_1, \ldots, A_N$. If jumps are proposed deterministically, we additionally assume that they are allowed only on "jumping regions" $JR_{\gamma,i}$ defined as

$$JR_{\gamma,i} = \{x \in \mathcal{X} : (x - \mu_i)^T \Sigma_{\gamma,i}^{-1} (x - \mu_i) \leq R\} \tag{2.14}$$

for $i \in \mathcal{I}$ and some $R > 0$. Note that equation (2.4) ensures that if x belongs to $JR_{\gamma,i}$ and we propose a deterministic jump from (x, i) to (y, k), then y must be in $JR_{\gamma,k}$. Thus the detailed balance condition is satisfied. The reasons for these modifications will become clearer when we present the proofs of the ergodic theorems.

Observe that each matrix $\Sigma_{\gamma,i}$ is based on samples belonging to a compact set A_i, so these matrices are bounded from above. Since we keep adding βI_d to the covariance matrix at each step, they are also bounded from below. Consequently, there exist positive constants m and M for which

$$mI_d \preceq \Sigma_{\gamma,i}, \preceq MI_d \quad \text{for all } \gamma \in \mathcal{Y} \text{ and } i \in \mathcal{I}. \tag{2.15}$$

As for the adaptive scheme for $w_{\gamma,i}$ and $a_{\gamma,ik}$, we only require that these values be bounded away from 0, i.e. there exist ϵ_a and ϵ_w such that

$$w_{\gamma,i} > \epsilon_w \quad \text{and} \quad a_{\gamma,ik} > \epsilon_a \quad \text{for all } \gamma \in \mathcal{Y} \text{ and } i, k \in \mathcal{I}. \tag{2.16}$$

Therefore, the parameter space $\mathcal{Y}$ may be considered as compact.

4 .2 Theoretical results for mJAMS

We begin with the case when the jump moves are proposed independently from distributions $R_{\gamma,J,i}$ with heavier tails than the tails of the target distribution π for all $i \in \mathcal{I}$ and $\gamma \in \mathcal{Y}$,

i.e.

$$\sup_{x\in\mathcal{X}}\sup_{\gamma\in\mathcal{Y}}\frac{\pi(x)}{R_{\gamma,J,i}(x)}<\infty \quad \text{for each } i\in\mathcal{I}. \tag{2.17}$$

We prove that under this assumption Simultaneous Uniform Ergodicity is satisfied for mJAMS and consequently, by Theorem 2.2, the algorithm is ergodic.

Theorem 2.6. *Consider the mJAMS algorithm and assume that the relationship between the target distribution π and the proposal distributions $R_{\gamma,J,i}$ satisfies (2.17). Then mJAMS is ergodic.*

When the tails of the distribution π are heavier then the tails of the proposal distributions $R_{\gamma,J,i}$, or when the jumps follow the deterministic scheme, Simultaneous Uniform Ergodicity does not hold. However, it turns out that under some additional regularity conditions mJAMS is still ergodic, as it satisfies the assumptions of Lemma 2.5.

Theorem 2.7. *Consider the mJAMS algorithm and assume that the following conditions are satisfied.*

a) For each $i\in\mathcal{I},\gamma\in\mathcal{Y}$ the proposal distribution for local moves $R_{\gamma,L,i}$ follows an elliptical distribution parametrized by $\Sigma_{\gamma,i}$. Furthermore, the family of distributions $R_{\gamma,L,i}(\boldsymbol{0},\cdot)$, $\gamma\in\mathcal{Y}$, has uniformly bounded probability density functions, and for any compact set $C\subset\mathcal{X}$ we have

$$\inf_{x,y\in C}\inf_{\gamma\in\mathcal{Y}}R_{\gamma,L,i}(x,y)>0 \quad \text{for each } i\in\mathcal{I}. \tag{2.18}$$

b) Let $r_{\gamma,i}(x)$ be the rejection set for local moves, i.e. $r_{\gamma,i}(x):=\{y\in\mathcal{X}:\tilde{\pi}_\gamma(y,i)<\tilde{\pi}_\gamma(x,i)\}$. We assume that

$$\limsup_{|x|\to\infty}\sup_{\gamma\in\mathcal{Y}}\int_{r_{\gamma,i}(x)}R_{\gamma,L,i}(x,y)dy<1 \quad \text{for each } i\in\mathcal{I}. \tag{2.19}$$

c) The target distribution π is super-exponential, i.e. it is positive with continuous first derivatives and satisfies

$$\lim_{|x|\to\infty}\frac{x}{|x|}\cdot\nabla\log\pi(x)=-\infty. \tag{2.20}$$

d) Every Q_i, $i\in\mathcal{I}$, is an elliptical distribution parametrized by $\Sigma_{\gamma,i}$ positive on $\mathcal{X}$ and additionally, the following condition is satisfied:

$$\sup_{x\in\mathcal{X}}\frac{Q_i(\mu_i,\Sigma_{\gamma_1,i})(x)}{Q_k(\mu_k,\Sigma_{\gamma_2,k})(x)}<\infty \quad \text{for all } i,k\in\mathcal{I} \text{ and } \gamma_1,\gamma_2\in\mathcal{Y}. \tag{2.21}$$

Additionally, one of the following two conditions for jump moves holds.

e1) Jump moves follow the procedure for deterministic jumps, as described in Section 2 .1.

e2) Jump moves follow the independent proposal procedure, as described in Section 2 .1. The proposal distributions for jumps have uniformly bounded probability density functions and satisfy

$$\inf_{x \in B(\mu_i, r)} \inf_{\gamma \in \mathcal{Y}} R_{\gamma,J,i}(x) > 0 \quad \text{for each } i \in \mathcal{I} \text{ and some } r > 0, \tag{2.22}$$

where $B(\mu_i, r)$ is a ball of radius r and centre μ_i. Moreover, the relationship between the target distribution $R_{\gamma,J,i}$ is given by

$$\sup_{x \in \mathcal{X}} \sup_{\gamma \in \mathcal{Y}} \frac{R_{\gamma,J,i}(x)}{\pi(x)^{s_J}} < \infty \quad \text{for each } i \in \mathcal{I} \text{ and some } s_J \in (0, 1]. \tag{2.23}$$

Then mJAMS is ergodic.

The most challenging part of the proof is verifying drift condition (2.11), which we do for $V_{\tilde{\pi}_\gamma}((x,i)) = c\tilde{\pi}_\gamma(x,i)^{-s}, c > 0, s \in (0,1)$, separately for local and jump kernels. When proving the result for local moves, we show that $\limsup_{|x| \to \infty} \sup_{\gamma \in \mathcal{Y}} \frac{\tilde{P}_{\gamma,L,i} V_{\tilde{\pi}_\gamma}((x,i))}{V_{\tilde{\pi}_\gamma}((x,i))} < 1$, following the proof of Theorem 4.1 of Jarner and Hansen (2000). That is, we consider analogues of the regularity conditions of this theorem (assumptions b) and c) above) and we split the state space $\mathcal{X}$ into three disjoint sets, like in Jarner and Hansen (2000), to show that for large x the value of the integral of $R_{\gamma,L,i}(x,y) \min\left[1, \frac{\tilde{\pi}_\gamma(y,i)}{\tilde{\pi}_\gamma(x,i)}\right] \frac{V_{\tilde{\pi}_\gamma}((y,i))}{V_{\tilde{\pi}_\gamma}((x,i))}$ with respect to y is arbitrarily small on each set. The main difference between our approach and that of Jarner and Hansen (2000) of is that we work with adaptive kernels and target distributions, which necessitates bounding the suprema of the involved expressions over $\gamma \in \mathcal{Y}$. Besides, assumption d) is introduced to ensure that $\tilde{\pi}_\gamma$ does not "deviate too much" from π. This condition can be easily verified if every Q_i, $i \in \mathcal{I}$, follows the t distribution with the same number of degrees of freedom. Note that for non-compact state spaces $\mathcal{X}$ the condition in assumption d) does not hold when Q_i, $i \in \mathcal{I}$, follow the normal distributions.

For the jump kernels the idea is to show that $\sup_{\gamma \in \mathcal{Y}} V_{\tilde{\pi}_\gamma}((y,k))$ is "not too large" when the algorithm jumps to (y,k). For deterministic jumps performed on jumping regions it follows from the boundedness of $V_{\tilde{\pi}_\gamma}((x,i))$ on compact sets. For independent proposal jumps we show that under (2.23) $\sup_{\gamma \in \mathcal{Y}} \int_{\mathcal{X}} R_{\gamma,J,k}(y) V_{\tilde{\pi}_\gamma}((y,k))\, dy < \infty$.

Note that condition a) holds automatically for our algorithm if we assume that the

proposal distributions for local moves follow either the normal or the t distribution (see Section 2 .1) and when (2.15) holds. Condition (2.22) is satisfied if the proposal distributions for jumps follow, for example, the normal distribution. The following lemmas are useful in verifying assumption b) of Theorem 2.7.

Lemma 2.8. *Let* $r(x) := \{y \in \mathcal{X} : \pi(y) < \pi(x)\}$ *and* $a(x) := \{y \in \mathcal{X} : \pi(y) \geq \pi(x)\}$. *Consider mJAMS together with conditions a), c) and d) of Theorem 2.7. Assume additionally that for some* $\gamma^* \in \mathcal{Y}$

$$\limsup_{|x|\to\infty} \int_{r(x)} R_{\gamma^*,L,i}(x,y)dy < 1 \quad \text{for each } i \in \mathcal{I}. \tag{2.24}$$

Then condition (2.19) *holds.*

Lemma 2.9. *Consider mJAMS together with conditions a), c) and d) of Theorem 2.7. Assume additionally that the target distribution* π *satisfies*

$$\limsup_{|x|\to\infty} \frac{x}{|x|} \cdot \frac{\nabla\pi(x)}{|\nabla\pi(x)|} < 0. \tag{2.25}$$

Then condition (2.19) *holds.*

The following corollary shows that mJAMS in a standard setting is successful at targeting mixtures of normal distributions.

Corollary 2.10. *Let the target distribution* π *be given by*

$$\pi(x) \propto w_1 \exp\left(-p_1(x)\right) + \ldots + w_n \exp\left(-p_n(x)\right),$$

where $w_i > 0$ *and* p_i *is a polynomial of order* ≥ 2 *for each* $i = 1, \ldots, n$. *If additionally* Q_i *for* $i \in \mathcal{I}$ *follows the multivariate t distribution with the same number of degrees of freedom, and* $R_{\gamma,L,i}(\boldsymbol{0}, \cdot)$ *follows the normal distribution, the assumptions of Lemma 2.9 are satisfied.*

The result stated below is a corollary from Theorem 2.4 and establishes the Weak Law of Large Numbers for our algorithm.

Theorem 2.11. *Consider mJAMS and assume that conditions of either Theorem 2.6 or Theorem 2.7 are satisfied. Then the Weak Law of Large Numbers holds for all bounded and measurable functions.*

Remark 2.12. Note that Theorems 2.6, 2.7 and 2.11 are based on an assumption that the list of modes is fixed. Let us now consider mJAMS in the version with mode finding running in parallel to the main MCMC sampler. Assume additionally that $\mathbb{P}(\tau < t) \to 1 \quad \text{as} \quad t \to \infty$, where τ is the time of adding the last mode. In this case Theorems 2.6, 2.7 and 2.11 still hold. Indeed, as the parallel burn-in algorithm runs independently of mJAMS, we can rephrase all the probabilistic limiting statements in the proofs on the set $C_t := \{\tau < t\}$ and then let $t \to \infty$.

An important question to ask is how the performance of the algorithm changes with the dimension of the state space. In the case of (non-adaptive) MCMC methods this is typically measured by the rate at which the spectral gap of a transition kernel decreases. In Supplementary Material A we analyse the spectral gap of an individual JAMS kernel. To obtain a formal result, we consider the following target density:

$$\pi(x) = \frac{1}{2} N\left(-\underbrace{(1,\ldots,1)}_{d}, \sigma_1^2 I_d\right) + \frac{1}{2} N\left(\underbrace{(1,\ldots,1)}_{d}, \sigma_2^2 I_d\right). \tag{2.26}$$

Parallel tempering is known to mix torpidly on this example (see Woodard et al. (2009a)) for any choice of temperatures when $\sigma_1 \neq \sigma_2$. We show that given the locations of the modes and a certain rate of approximation of the proposal covariance matrices to the true ones, an individual JAMS kernel with independent proposal jumps mixes rapidly on this target density. The proof of this result relies on observing that the jump kernel following the independent proposal scheme is in fact an independence sampler kernel, thus we can use a result from Gåsemyr (2006) to obtain its spectral gap.

We also argue, without proving a formal result, why the number of iterations of covariance matrix estimation in the burn-in algorithm, required to obtain the above rate of approximation is polynomial with respect to the dimension. Hence, given the locations of the modes, this would give an overall polynomial time of an algorithm composed of the covariance matrix estimation in the burn-in algorithm, followed by a non-adaptive version of JAMS targeting (2.26).

5 Examples

In this section we present empirical results for our method (Algorithm 1 preceded by the burn-in algorithm). We test its performance on three examples – the first one is a mixture of two Gaussians motivated by Woodard et al. (2009a); the second one is a mixture of fifteen multivariate t distributions and five banana-shaped ones; the third one is a Bayesian model for sensor network localization. Our implementation admits three versions, varying in the way the jumps between modes are performed. In particular, we consider here the deterministic jump and two independent proposal jumps, with Gaussian and t-distributed proposals.

Additionally, we compare the performance of our algorithm against adaptive parallel tempering (Miasojedow et al., 2013), which was chosen here as it is the refined version of the most commonly used MCMC method for multimodal distributions (parallel tempering). What is more, this algorithm has a generic implementation, where the user only needs to provide the target density function. In order to make a comparison between the efficiency of these algorithms, among other things, we analyse the Root Mean Square Error (RMSE) divided by the square root of the dimension of the state space, given a computational budget. We measure the computational cost by the number of evaluations of the target distribution (and its gradient, if applicable), as this is typically the dominating factor in real data examples. Herein we define RMSE as the Euclidean distance between the true d-dimensional expected value (if known) and its empirical estimate based on MCMC samples.

In order to depict the variability in the results delivered by both methods, each simulation was repeated 20 times. For exact settings of the experiments, as well as some additional results, we refer the reader to Supplementary Material B .

5 .1 Mixture of Gaussians

We looked at the results for the target distribution (2.26) in several different dimensions d ranging between 10 and 200, for $\sigma_1^2 = 0.5\sqrt{d/100}$ and $\sigma_2^2 = \sqrt{d/100}$. The results for our method shown below are based on 500,000 iterations of the main algorithm, preceded by the burn-in algorithm including 1500 BFGS runs, started from points sampled uniformly on $[-2, 2]^d$.

The length of the covariance matrix estimation was chosen automatically using the rule

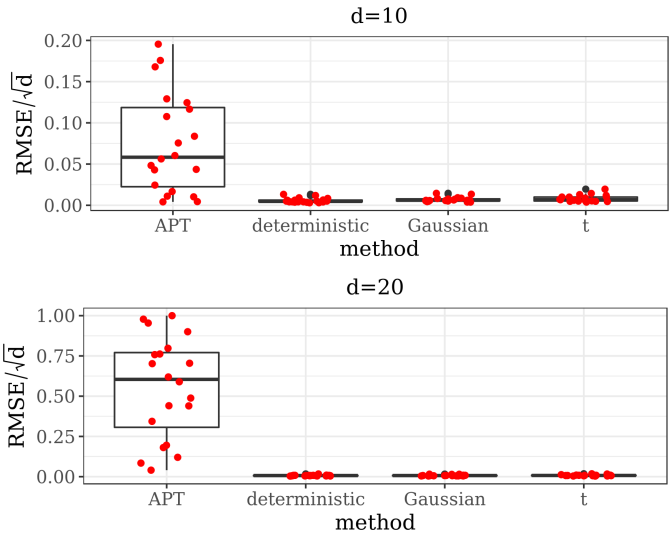


Figure 2.1: Boxplots of the values of RMSE/$\sqrt{d}$ for the mixture of Gaussians across 20 runs of the experiment, dimensions 10 and 20. We compare the results of APT with the three JAMS versions: deterministic, Gaussian and t-distributed jumps. Note different scales on the y-axis.

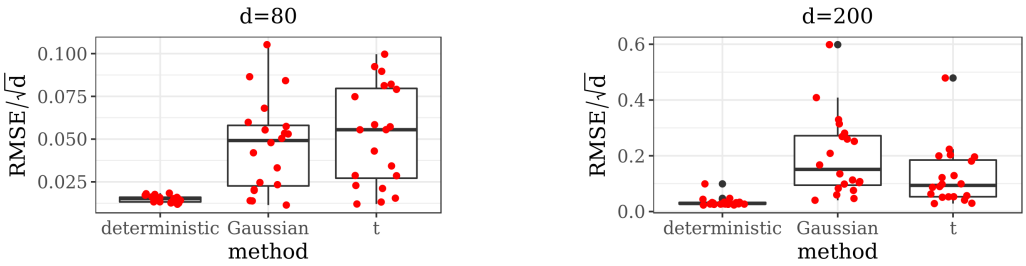


Figure 2.2: Boxplots of the values of RMSE/$\sqrt{d}$ for the mixture of Gaussians across 20 runs of the experiment, dimensions 80 and 200. Note different scales on the y-axis.

described in Supplementary Material B and varied between 3000 (for $d = 10$) to 1,023,000 iterations (for $d = 200$) per mode. For dimensions $d = 10$ and $d = 20$ we ran also the adaptive parallel tempering (APT) algorithm, with 700,000 iterations and 5 temperatures. Overall this requires 3,500,000 evaluations of the target density that cannot be performed in parallel, despite the name of the method, as the communication between chains running at different temperatures is needed after every iteration. In the light of the tendency of the parallel tempering algorithm to stay in wider modes, each time the APT algorithm was started in $-\underbrace{(1, \ldots, 1)}_{d} \in \mathbb{R}^d$. In order to base our analysis on the same sample size of 500,000 for the two methods, in the case of adaptive parallel tempering we applied an initial burn-in period of 200,000 steps.

The results presented in the boxplots of Figure 2.1, as well as the upper panel of density plots (Figure 2.3) show that our method outperforms adaptive parallel tempering on this example, even when the latter method is given a much larger computational budget. The

	deterministic		Gaussian		t-distributed	
	Lowest	Highest	Lowest	Highest	Lowest	Highest
d=10	0.98	0.99	0.85	0.87	0.71	0.73
d=20	0.98	0.99	0.79	0.83	0.66	0.68
d=80	0.91	0.98	0.23	0.41	0.24	0.39
d=130	0.72	0.98	0.04	0.13	0.06	0.15
d=160	0.79	0.97	0.01	0.07	0.03	0.07
d=200	0.64	0.97	0.01	0.05	0.02	0.06

Table 2.1: The lowest and the highest value (across 20 runs of the experiment) of the acceptance rates of jump moves between the two modes for the mixture of Gaussians for different jump methods and dimensions.

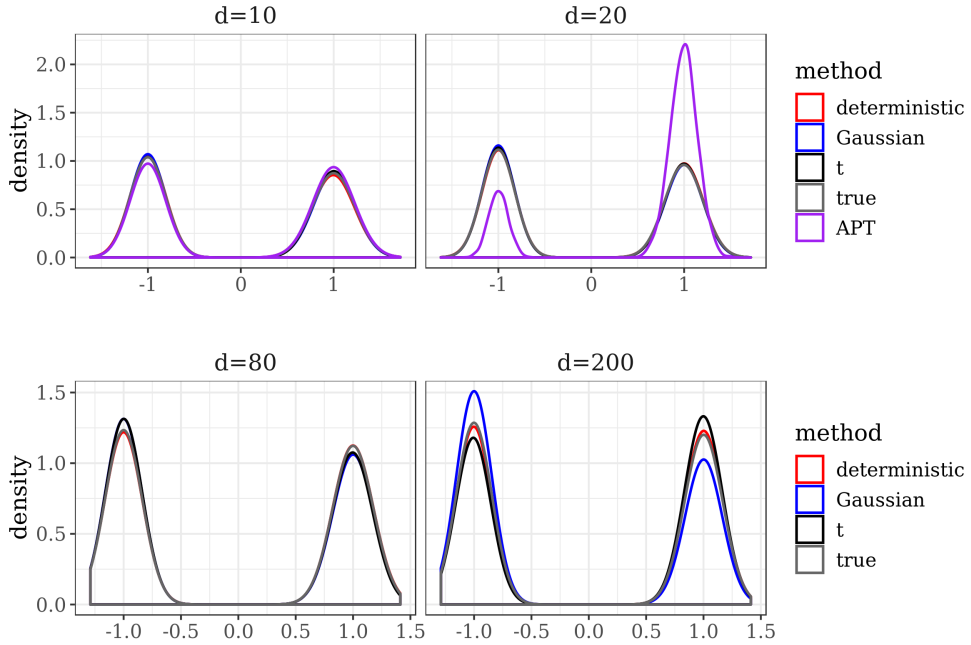


Figure 2.3: Density plots for the mixture of Gaussians in dimensions 10, 20, 80 and 200. The upper panel shows a comparison between APT and the three JAMS versions. The simulations chosen for the analysis correspond to the median value of RMSE across 20 experiments (the tenth largest value of RMSE).

summary of the acceptance rates of the jump moves presented in Table 2.1 demonstrates that the algorithm preserves good mixing between the modes in all its jump versions up to dimension 80. It is remarkable that the deterministic jump ensures excellent mixing even in much higher dimensions, outperforming the remaining two methods (see Figure 2.2 and the lower panel of Figure 2.3), with the acceptance rate between 0.64 and 0.97 in dimension 200.

5 .2 Mixture of t and banana-shaped distributions

In order to test our method on a multimodal target with highly non-elliptical modes, we used a modified and more challenging version of a classic example introduced in Kou et al. (2006) (a mixture of 20 bivariate Gaussian distributions), studied also later by Miasojedow

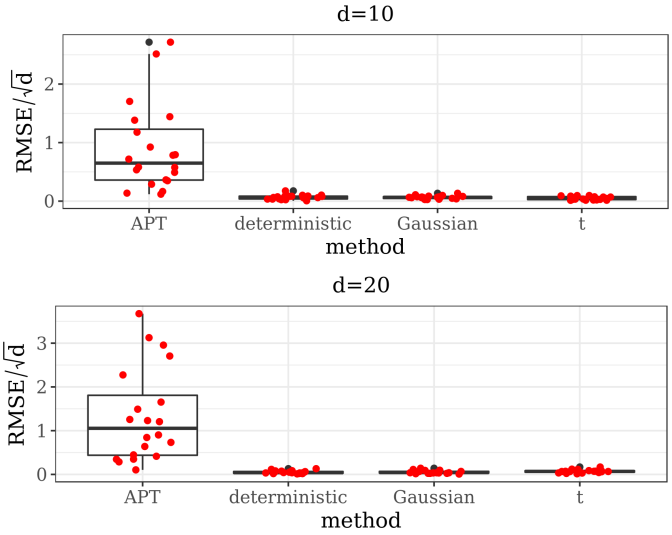


Figure 2.4: Boxplots of the values of RMSE/$\sqrt{d}$ for the mixture of banana-shaped and t-distributions across 20 runs of the experiment, dimensions 10 and 20. We compare the results of APT with the three JAMS versions: deterministic, Gaussian and t-distributed jumps. Note different scales on the y-axis.

et al. (2013) and Tak et al. (2017).

In our case, instead of the Gaussian distribution, modes 1-5 follow the banana-shaped distribution with t tails, where the shape of the banana for mode k is present in the projection onto coordinates 1 and $2k$. The remaining ones follow the multivariate t distribution with 7 degrees of freedom and the covariance matrices $0.01\sqrt{d}I_d$, where d is the dimension. The weights are assumed to be equal to 0.05. We consider dimensions $d = 10$ and $d = 20$ by repeating the original coordinates of the centres of the modes five and ten times, respectively.

The results below are based on 500,000 iterations, preceded by 40,000 BFGS runs. The starting points for these runs were sampled uniformly on $[-2, 12]^d$. The number of iterations of the covariance matrix estimation varied between 7,000 and 15,000 steps per mode for dimension $d = 10$ and between 15,000 and 63,000 steps per mode for dimension $d = 20$. For adaptive parallel tempering we used 2,100,000 iterations and 5 temperatures. We applied an initial burn-in period of 600,000 steps and we thinned the chain keeping every third sample.

In Supplementary Material B we present results for the same example obtained using JAMS in dimensions $d = 50$ and $d = 80$ assuming that the modes of the target distribution are known, since mode finding (in particular, getting to each basin of attraction) is the main bottleneck for this example.

For dimensions $d = 10$ and $d = 20$ all modes were found by the BFGS runs in each of the 20 simulations. Figure 2.4 illustrates that the empirical means based on JAMS samples approximate well the true expected value of the target distribution, consistently across all

	deterministic		Gaussian		t-distributed	
	Lowest	Highest	Lowest	Highest	Lowest	Highest
d=10	0.27	0.77	0.12	0.52	0.20	0.65
d=20	0.20	0.75	0.09	0.35	0.11	0.48

Table 2.2: The lowest and the highest value (across 20 runs of the experiment) of the acceptance rate of jumps from a given mode, in dimensions 10 and 20, for the mixture of banana-shaped and t-distributions.

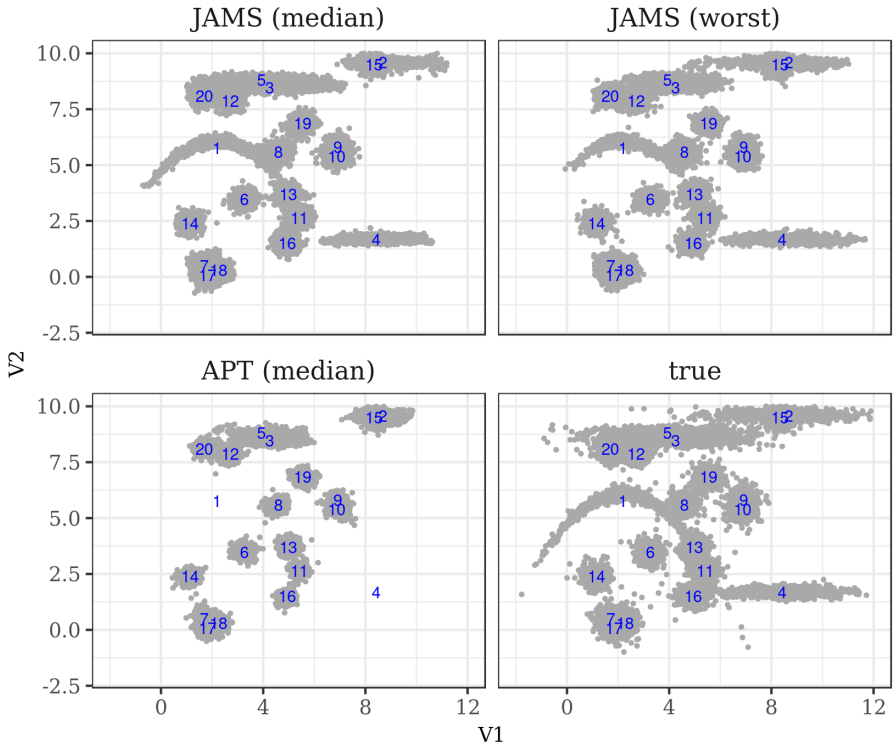


Figure 2.5: Scatterplots of the first and second coordinate of the mixture of t and banana-shaped distributions in dimension 20. Recall that modes 1-5 follow the banana-shaped distribution with t tails, where the shape of the banana for mode k is present in the projection onto coordinates 1 and $2k$. The blue numbers denote the centres of the modes. We compare the results obtained using JAMS (with t-distributed jumps) and APT against samples generated from the true distribution. The simulations chosen for the analysis correspond to the median or maximum ("worst") value of RMSE across 20 experiments.

experiments, and that our method significantly outperforms APT with a smaller computational cost. Figure 2.5 shows that APT struggles with identifying some of the banana-shaped modes, whereas JAMS mixes well between all the modes and reflects well the shape of the true distribution. Due to Gaussian local proposal both methods exhibit slightly lighter tails than the true distribution. Good between-mode mixing properties of JAMS are confirmed by the results gathered in Table 2.2.

5 .3 Sensor network localization

We consider here an example from Ihler et al. (2005), analysed later by Ahn et al. (2013), Lan et al. (2014) and, in a modified version by Tak et al. (2017). There are 11 sensors with locations $x_1, \ldots, x_{11}$ scattered on a space $[0,1]^2$. The locations of sensors $x_1, \ldots, x_8$ are unknown, the remaining three locations are known. For any two sensors i and j we observe the distance y_{ij} between them with probability $\exp\left(-\frac{\|x_i - x_j\|^2}{2\times 0.3^2}\right)$. Once observed, the distance y_{ij} follows the normal distribution given by $y_{ij} \sim N\left(\|x_i - x_j\|, 0.02^2\right)$. Let w_{ij} be equal to 1 when y_{ij} is observed and 0 otherwise, and denote $y := \{y_{ij}\}$ and $w := \{w_{ij}\}$. The goal of the study is to make inference about the unknown locations $x_i = (z_{i1}, z_{i2})$ for $i = 1, \ldots 8$ given y and w. Following Ahn et al. (2013) and Lan et al. (2014) we put a uniform prior on $[0,1]$ for each of the coordinates z_{i1} and z_{i2} for $i = 1, \ldots 8$. The resulting posterior distribution is given by

$$\pi(x_1, \ldots x_8 | y, w) \propto \prod_{\substack{j=2,\ldots,11 \\ i=1,\ldots,8 \\ i<j}} f_{ij}\left(x_i, x_j | y_{ij}, w_{ij}\right) \mathbb{I}_{x_i, x_j \in [0,1]}, \quad \text{where}$$

$$f_{ij}\left(x_i, x_j | y_{ij}, w_{ij}\right) = \begin{cases} \exp\left(-\frac{\|x_i - x_j\|^2}{2\times 0.3^2}\right) \exp\left(-\frac{\left(y_{ij} - \|x_i - x_j\|\right)^2}{2\times 0.02^2}\right) & \text{if } w_{ij} = 1, \\ 1 - \exp\left(-\frac{\|x_i - x_j\|^2}{2\times 0.3^2}\right) & \text{otherwise.} \end{cases}$$

Note that the posterior distribution is defined on a compact space, therefore the assumptions of Theorem 2.6 will be satisfied for normally distributed or t-distributed proposals.

Since there are few observed distances with known locations (see: top left panel of Figure 2.6), the model is non-identifiable which results in multimodality of the posterior distribution. We ran JAMS on this example for 500,000 iterations of the main algorithm. This was preceded by 10,000 BFGS runs (with starting points for optimization sampled from the prior) and covariance matrix estimation (between 7000 and 15,0000 iterations per mode). For parallel tempering we used 700,000 iterations (with a burn-in period of 200,000) and 4 temperatures. If JAMS is implemented on 8 cores, this means that running an APT simulation is about twice as costly as running a JAMS one (see Supplementary Material B for details).

Despite the fact that for all 20 APT experiments the acceptance rates at all temperature levels, as well as for between-temperature swaps, converged to the optimal acceptance rate

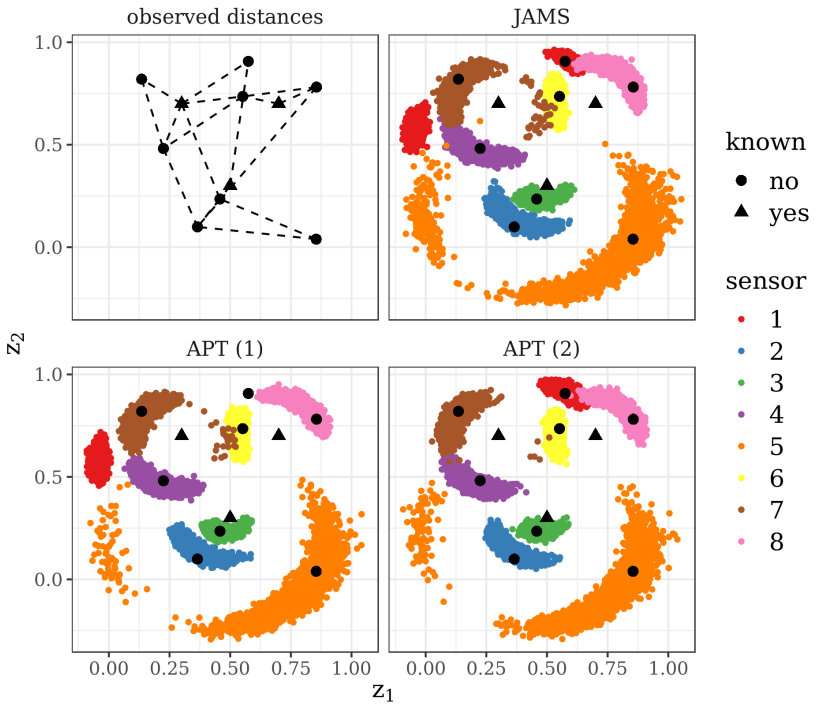


Figure 2.6: Black triangles and dots denote true locations of the sensors with known and unknown locations, respectively. Top left panel: dashed lines represent observed distances between sensors. Top right panel: posterior samples obtained using JAMS (with Gaussian jumps) for locations $x_1, \dots x_8$. Bottom panels: posterior samples using APT for two different starting points.

0.234 (see Atchadé et al. (2011)), the behaviour of this algorithm was unstable. As shown in Figure 2.6, in the case of APT the estimation of the location of sensor 1 depends on the starting point. In the case of JAMS, both modes for x_1 (in red) are represented. Figure 2.7 illustrates stability of JAMS across all experiments and jump methods. In Supplementary Material B we assign an even higher computational budget to adaptive parallel tempering allowing for 5 temperatures and observe a substantial improvement in mixing and stability, but the results are still worse than those of JAMS.

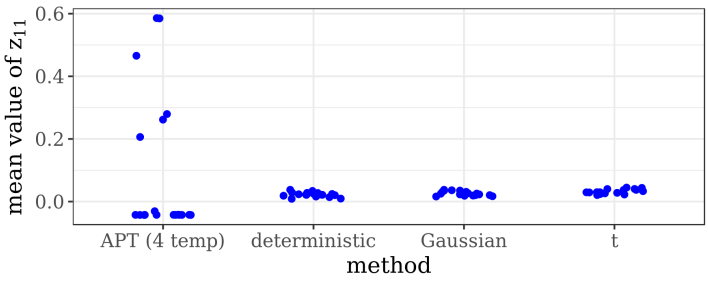


Figure 2.7: A boxplot of the mean value of the first coordinate of sensor 1 for 20 runs of the experiment for APT (with 4 temperature levels) and three versions of JAMS.

6 Summary and discussion

The framework we proposed here is based on three fundamental ideas. Firstly, we split the task into mode finding and sampling from the target distribution. This allows the user to choose freely the best optimization algorithm available. Secondly, we base our approach on local moves responsible for mixing within the same mode and jumps that move between the modes. Finally, we account for inhomogeneity between the modes by using different proposal distributions at each mode and adapting their parameters separately.

To develop a methodological approach and prove ergodic results for our algorithm, we introduced the Auxiliary Variable Adaptive MCMC class. There are other adaptive algorithms falling in this category, so our theoretical results may potentially be useful beyond the scope of JAMS. We have shown that the Auxiliary Variable Adaptive MCMC methods enjoy robust ergodicity properties analogous to those of Adaptive MCMC.

We believe that our approach overcomes to a large extent deficiencies of tempering-based methods. Its important advantage is an inbuilt mechanism of remembering the locations of the modes identified so far, by including them in the augmented target distributions. Hence, in contrast to the tempering-based methods, it does not need to "re-discover" a mode each time in order to move there, which is of particular importance when some basins of attraction are much smaller than the others. By analysing the spectral gap of a single JAMS kernel (with independent proposal jumps) and commenting on the rate of convergence of the estimation of covariance matrices during the burn-in algorithm, we have argued why for a mixture of Gaussians with known locations of the modes, the required number of iterations should depend polynomially on the dimension. A very interesting research avenue, suggested by the Reviewers, would be establishing formal results on the performance of our adaptive algorithm as a function of dimension, possibly also for a larger class of target distributions and the deterministic jump scheme.

The methodology proposed in this paper has its limitations. Firstly, it is designed under the assumption that the target distribution has a finite number of local maxima. The construction of JAMS facilitates moving between the regions around these maxima. However, some posterior distributions have a very large number of local maxima, or even infinitely many maxima. We believe that JAMS is much better suited for posteriors with several distinct modes than for targets with highly jagged surfaces.

Currently, the main bottleneck of the method is mode finding, and in particular, sampling starting points for optimization runs in such a way that there is at least one point in the basin of attraction of each mode. As well as other optimization-based algorithms, in cases when a mode is missed by optimization runs, JAMS will inherit the properties of the standard Random Walk Metropolis algorithm on the region associated with this mode. If the probability mass of the regions corresponding to the missed mode is large, the improvement of JAMS relative to the local sampler may become negligible. A potential remedy to this issue could be replacing with certain probability standard RWM with Gaussian or t-distributed proposals, with local samplers being Metropolis-Hastings algorithms using a special design of the proposal distribution, which facilitates crossing low probability barriers, e.g. Tak et al. (2017). Improving the way of sampling the starting points for optimization runs, as well as investigating different strategies of exploring well the state space even when some modes have been missed, is an area of future work.

Acknowledgements

We thank Louis Aslett, Ewan Cameron, Arnaud Doucet, Geoff Nicholls, Jeffrey Rosenthal, Radek Adamczak, George Deligiannidis, two anonymous Referees and the Associate Editor for helpful comments, and Shiwei Lan for pointing us to the data for the sensor network example. We also thank Radu Craiu for providing the data set for the LOH example studied in Supplementary Material B .

A Supplementary Material: theory

Supplementary Material A consists of Sections A .1, A .2, A .3 and A .4. In Section A .1 we present the proofs of our theoretical results stated in Section 3 . In Section A .2 we prove the results stated in Section 4 . In Section A .3 we give some comments about other algorithms in the Auxiliary Variable Adaptive MCMC class. In Section A .4 we comment on the dependence of the performance of the algorithm on the dimension, by analysing the spectral gap of an individual JAMS kernel.

A .1 Proofs for Section 3

The proofs of our results presented in Section 3 will rely on a coupling construction analogous to Roberts and Rosenthal (2007) (see also Roberts and Rosenthal (2013) for a more rigorous presentation).

To prove Theorems 2.2, 2.3 and 2.4 we define two auxiliary processes $\{(\tilde{X}_n^m, \Gamma_n^m)\}_{n=0}^{\infty}$ and $\{(\tilde{X}_n^i, \Gamma_n^i)\}_{n=0}^{\infty}$ on the space $\mathcal{X} \times \Phi$, to be thought of as "Markovian" and "intermediate". The latter is designed in such a way that it interpolates between the Markovian one and the original adaptive process. The key part of the proof is the explicit coupling construction which shows that the total variation distance between the two pairs of processes goes to 0. The remainder of this section is organized as follows. We first define formally the two processes and introduce necessary notation. We next prove some auxiliary lemmas, from which we finally conclude ergodicity (Theorems 2.2 and 2.3) and the Weak Law of Large Numbers (Theorem 2.4).

Recall that $\{(\tilde{X}_n, \Gamma_n)\}_{n=0}^{\infty}$ is the adaptive process on $\tilde{X} := \mathcal{X} \times \Phi$ defined in Section 3 with dynamics governed by equation (2.7). On the same probability space we define two additional sequences, namely $\{(\tilde{X}_n^{m(t)}, \Gamma_n^{m(t)})\}_{n=0}^{\infty}$ and $\{(\tilde{X}_n^{i(t,\kappa)}, \Gamma_n^{i(t,\kappa)})\}_{n=0}^{\infty}$ (the meaning of the parameter κ will be explained later), which are identical to $\{(\tilde{X}_n, \Gamma_n)\}_{n=0}^{\infty}$ before pre-specified time t, i.e.

$$(\tilde{X}_n^{m(t)}, \Gamma_n^{m(t)}) = (\tilde{X}_n^{i(t,\kappa)}, \Gamma_n^{i(t,\kappa)}) = (\tilde{X}_n, \Gamma_n) \qquad \text{for } n \leq t.$$

After time t, the adaptive parameter $\Gamma^{m(t)}$ of $\{(\tilde{X}_n^{m(t)}, \Gamma_n^{m(t)})\}_{n=0}^{\infty}$ freezes and $\tilde{X}_n^{m(t)}$ becomes a Markov chain with the marginal dynamics defined for $n+1 > t$ as:

$$\Gamma_{n+1}^{m(t)} := \Gamma_n^{m(t)} \quad (= \Gamma_t^{m(t)}), \tag{2.27}$$

$$\mathbb{P}\left[\tilde{X}_{n+1}^{m(t)} \in \tilde{B} | \tilde{X}_n^{m(t)} = \tilde{x}, \Gamma_t^{m(t)} = \gamma\right] = \tilde{P}_\gamma(\tilde{x}, \tilde{B}), \quad \tilde{B} \in \mathcal{B}(\tilde{\mathcal{X}}). \tag{2.28}$$

The second sequence $\{(\tilde{X}_n^{i(t,\kappa)}, \Gamma_n^{i(t,\kappa)})\}_{n=0}^{\infty}$ interpolates between

$$\{(\tilde{X}_n, \Gamma_n)\}_{n=0}^{\infty} \quad \text{and} \quad \{(\tilde{X}_n^{m(t)}, \Gamma_n^{m(t)})\}_{n=0}^{\infty}.$$

We define the dynamics of $\Gamma^{i(t,\kappa)}$ for $n+1 > t$, as:

$$\Gamma_{n+1}^{i(t,\kappa)} := \begin{cases} \Gamma_{n+1} & \text{if } \sup_{\tilde{x}\in\tilde{\mathcal{X}}} \|\tilde{P}_{\Gamma_{n+1}}(\tilde{x},\cdot) - \tilde{P}_{\Gamma_n^{i(t,\kappa)}}(\tilde{x},\cdot)\|_{TV} \leq \kappa, \\ \Gamma_n^{i(t,\kappa)} & \text{otherwise;} \end{cases} \tag{2.29}$$

Let $\tau_{i(t,\kappa)}$ denote an auxiliary stopping time that records decoupling of Γ_n and $\Gamma_n^{i(t,\kappa)}$, that is

$$\tau_{i(t,\kappa)} := \min\{n : \Gamma_n^{i(t,\kappa)} \neq \Gamma_n\},$$

with the convention $\min\emptyset = \infty$. Now, let us define the dynamics of $\tilde{X}_n^{i(t,\kappa)}$ as:

$$\tilde{X}_n^{i(t,\kappa)} := \tilde{X}_n \quad \text{for } n \leq \tau_{i(t,\kappa)}, \tag{2.30}$$

$$\text{and} \quad \mathbb{P}\left[\tilde{X}_{n+1}^{i(t,\kappa)} \in \tilde{B} \mid \tilde{X}_n^{i(t,\kappa)} = \tilde{x}, \Gamma_n^{i(t,\kappa)} = \gamma\right] = \tilde{P}_\gamma(\tilde{x}, \tilde{B}) \tag{2.31}$$

for $n+1 > \tau_{i(t,\kappa)}$ and $\tilde{B} \in \mathcal{B}(\tilde{\mathcal{X}})$. Informally speaking, we will later show that even though processes $\tilde{X}_n$, $\tilde{X}_n^{m(t)}$ $\tilde{X}_n^{i(t,\kappa)}$ split, they remain "close to each other".

Let $\tilde{A}_n^{\mathcal{G}_t}(\cdot)$ denote the distribution of the $\tilde{\mathcal{X}}$-marginal of $\{(\tilde{X}_n, \Gamma_n)\}_{n=0}^\infty$ at time n, conditionally on the history up to time t, i.e.

$$\tilde{A}_n^{\mathcal{G}_t}(\tilde{B}) := \mathbb{P}[\tilde{X}_n \in \tilde{B} | \tilde{X}_0 = \tilde{x}, \tilde{X}_1 = \tilde{x}_1, \ldots, \tilde{X}_t = \tilde{x}_t, \Gamma_0 = \gamma, \Gamma_1 = \gamma_1, \ldots, \Gamma_t = \gamma_t]$$

for $\tilde{B} \in \mathcal{B}(\tilde{\mathcal{X}})$. Recall that for $t = 0$ we have

$$\tilde{A}_n(\tilde{B}) := \tilde{A}_n^{\mathcal{G}_0}(\tilde{B}) = \mathbb{P}[\tilde{X}_n \in \tilde{B} | \tilde{X}_0 = \tilde{x}, \Gamma_0 = \gamma] \quad \text{for } \tilde{B} \in \mathcal{B}(\tilde{\mathcal{X}}).$$

Let $A_n^{\mathcal{G}_t}(\cdot)$ denote the marginalization of $\tilde{A}_n^{\mathcal{G}_t}(\cdot)$ onto $\mathcal{X}$, that is

$$A_n^{\mathcal{G}_t}(B) := \tilde{A}_n^{\mathcal{G}_t}(B \times \Phi) \quad \text{for } B \in \mathcal{B}(\mathcal{X}).$$

Define also the filtration $\{\mathcal{G}_n^*\}_{n=0}^\infty$ as an extension of $\{\mathcal{G}_n\}_{n=0}^\infty$ by:

$$\mathcal{G}_n^* := \sigma\left\{\{(\tilde{X}_k, \Gamma_k)\}_{k=0}^n,\ \{(\tilde{X}_k^{m(t)}, \Gamma_k^{m(t)})\}_{k=0}^n,\ \{(\tilde{X}_k^{i(t,\kappa)}, \Gamma_k^{i(t,\kappa)})\}_{k=0}^n\right\}. \tag{2.32}$$

Let the distributions

$$\tilde{A}_n^{m(t)}(\cdot), \quad \tilde{A}_n^{m(t),\mathcal{G}_t^*}(\cdot), \quad A_n^{m(t)}(\cdot), \quad A_n^{m(t),\mathcal{G}_t^*}(\cdot),$$

and

$$\tilde{A}_n^{i(t,\kappa)}(\cdot), \quad \tilde{A}_n^{i(t,\kappa),\mathcal{G}_t^*}(\cdot), \quad A_n^{i(t,\kappa)}(\cdot), \quad A_n^{i(t,\kappa),\mathcal{G}_t^*}(\cdot),$$

be analogues of $\tilde{A}_n(\cdot)$, $\tilde{A}_n^{\mathcal{G}_t}(\cdot)$, $A_n(\cdot)$, $A_n^{\mathcal{G}_t}(\cdot)$, where in the definitions of the above terms, instead of $\{(\tilde{X}_n, \Gamma_n)\}_{n=0}^{\infty}$, we use the sequences

$$\{(\tilde{X}_n^{m(t)}, \Gamma_n^{m(t)})\}_{n=0}^{\infty} \quad \text{and} \quad \{(\tilde{X}_n^{i(t,\kappa)}, \Gamma_n^{i(t,\kappa)})\}_{n=0}^{\infty}$$

respectively, and condition on the extended σ-algebras defined in (2.32).

Lemma 2.13. *Let $\tilde{\nu}_1$ and $\tilde{\nu}_2$ be probability measures on $\tilde{\mathcal{X}} = \mathcal{X} \times \Phi$ and let ν_1 and ν_2 be their marginals on $\mathcal{X}$. Then*

$$\|\nu_1 - \nu_2\|_{TV} \leq \|\tilde{\nu}_1 - \tilde{\nu}_2\|_{TV}.$$

Proof. Total variation distances on $\tilde{\mathcal{X}}$ involve suprema over larger classes of sets than those on $\mathcal{X}$, in particular $|\nu_1(B) - \nu_2(B)| = |\tilde{\nu}_1(B \times \Phi) - \tilde{\nu}_2(B \times \Phi)|$. □

Lemma 2.14. *Let Simultaneous Uniform Ergodicity, i.e. condition (a) of Theorem 2.2, hold. Then for all $\varepsilon > 0$, there exists $N_0 = N_0(\varepsilon)$ such that for all $N \geq N_0$*

$$\|A_{t+N}^{m(t)}(\cdot) - \pi(\cdot)\|_{TV} \leq \varepsilon \quad \textit{for all} \quad t \in \mathbb{N}, \tilde{x} \in \tilde{\mathcal{X}} \textit{ and } \gamma \in \mathcal{Y}.$$

Proof. Note that for any $B \in \mathcal{B}(\mathcal{X})$ the object $A_{t+N}^{m(t),\mathcal{G}_t^*}(B)$ is a $\mathcal{G}_t^*$-measurable random variable. We then Jensen's inequality to obtain (2.33) below. Next, we use Lemma 2.13 to obtain (2.34). To get (2.35) recall that $\{\tilde{X}_k^{m(t)}\}_{k=t}^{\infty}$ is a Markov chain started from $\tilde{X}_t$ with dynamics $\tilde{P}_{\Gamma_t}$. Finally, we use monotonicity of the total variation for Markov chains, and

pick $N_0 = N_0(\varepsilon)$ via assumption (a) of Theorem 2.2 to conclude (2.36).

$$\begin{aligned}
\|A_{t+N}^{m(t)}(\cdot) - \pi(\cdot)\|_{TV} &\leq \mathbb{E}\left[\|A_{t+N}^{m(t),\mathcal{G}_t^*}(\cdot) - \pi(\cdot)\|_{TV}\right] && (2.33)\\
&\leq \mathbb{E}\left[\|\tilde{A}_{t+N}^{m(t),\mathcal{G}_t^*}(\cdot) - \tilde{\pi}_{\Gamma_t}(\cdot)\|_{TV}\right] && (2.34)\\
&= \mathbb{E}\left[\|\tilde{P}_{\Gamma_t}^{N}(\tilde{X}_t,\cdot) - \tilde{\pi}_{\Gamma_t}(\cdot)\|_{TV}\right]\\
&\leq \mathbb{E}\left[\|\tilde{P}_{\Gamma_t}^{N_0}(\tilde{X}_t,\cdot) - \tilde{\pi}_{\Gamma_t}(\cdot)\|_{TV}\right] && (2.35)\\
&\leq \mathbb{E}[\varepsilon] = \varepsilon. && (2.36)
\end{aligned}$$

□

Lemma 2.15. *Let Containment, i.e. condition (a) of Theorem 2.3, hold. Then for all $\tilde{x} \in \tilde{\mathcal{X}}$, $\gamma \in \mathcal{Y}$ and $\varepsilon > 0$, there exists $N_0 = N_0(\varepsilon, \tilde{x}, \gamma)$, such that for all $t \in \mathbb{N}$ and all $N \geq N_0$,*

$$\|A_{t+N}^{m(t)}(\cdot) - \pi(\cdot)\|_{TV} \leq \mathbb{E}\left[\|A_{t+N}^{m(t),\mathcal{G}_t^*}(\cdot) - \pi(\cdot)\|_{TV}\right] \leq \varepsilon. \tag{2.37}$$

Proof. First, we use the Containment condition, that is, we choose $N_0 = N_0(\varepsilon, \tilde{x}, \gamma)$ such that

$$\mathbb{P}\left(M_{\varepsilon/2}(\tilde{X}_n, \Gamma_n) > N_0 | \tilde{X}_0 = \tilde{x}, \Gamma_0 = \gamma\right) \leq \varepsilon/2 \quad \text{for all } n \in \mathbb{N}.$$

Let $G := \{M_{\varepsilon/2}(\tilde{X}_t, \Gamma_t) \leq N_0\}$, we then have $\mathbb{P}(G') \leq \varepsilon/2$ and

$$\|\tilde{P}_{\Gamma_t}^{N_0}(\tilde{X}_t,\cdot) - \tilde{\pi}_{\Gamma_t}(\cdot)\|_{TV} \leq \varepsilon/2 \quad \text{on } G.$$

Now for $N > N_0$ we can reiterate the arguments used in the proof of Lemma 2.14 until equation (2.35). This in particular yields the first part of (2.37). To obtain the second part, note that

$$\begin{aligned}
\mathbb{E}\left[\|\tilde{P}_{\Gamma_t}^{N_0}(\tilde{X}_t,\cdot) - \tilde{\pi}_{\Gamma_t}(\cdot)\|_{TV}\right] &= \mathbb{E}\left[\|\tilde{P}_{\Gamma_t}^{N_0}(\tilde{X}_t,\cdot) - \tilde{\pi}_{\Gamma_t}(\cdot)\|_{TV}\mathbb{I}_G\right]\\
&\quad + \mathbb{E}\left[\|\tilde{P}_{\Gamma_t}^{N_0}(\tilde{X}_t,\cdot) - \tilde{\pi}_{\Gamma_t}(\cdot)\|_{TV}\mathbb{I}_{G'}\right]\\
&\leq (\varepsilon/2)\cdot\mathbb{P}(G) + 1\cdot\mathbb{P}(G') \;\leq\; \varepsilon,
\end{aligned}$$

as required. □

Lemma 2.16. *Let Diminishing Adaptation, i.e. condition (b) of Theorem 2.2, hold. Then*

for all $\varepsilon > 0$, $\kappa > 0$ *and* $N_0 \in \mathbb{N}$ *there exists* $t_0 = t_0(\varepsilon, \kappa, N_0)$ *such that for every* $t \geq t_0$ *and every* $N \leq N_0$,

$$\|\tilde{A}_{t+N}(\cdot) - \tilde{A}^{i(t,\kappa)}_{t+N}(\cdot)\|_{TV} \leq \varepsilon.$$

Proof. Recall D_n defined in condition (b) of Theorem 2.2 and let

$$H_n := \{D_n \geq \kappa\}.$$

Note that by Diminishing Adaptation for every $n \geq t_0 = t_0(\varepsilon, \kappa, N_0)$ we have $\mathbb{P}(H_n) \leq \varepsilon/N_0$. For $t \geq t_0$, we define

$$E := \bigcap_{n=t}^{t+N_0-1} H_n^c, \qquad \text{satisfying} \qquad \mathbb{P}(E) \geq 1 - \varepsilon.$$

We will consider the process $\{(\tilde{X}_n^{i(t,\kappa)}, \Gamma_n^{i(t,\kappa)})\}_{n=0}^{\infty}$ with $t \geq t_0$. Note that on E we have $\tilde{X}_n = \tilde{X}_n^{i(t,\kappa)}$, for $n = 0, 1, \ldots, t + N_0$, and therefore the coupling inequality (see e.g. Section 4.1 of Roberts and Rosenthal (2004)) for every $N \leq N_0$ yields as claimed

$$\|\tilde{A}_{t+N}(\cdot) - \tilde{A}^{i(t,\kappa)}_{t+N}(\cdot)\|_{TV} \leq \mathbb{P}(E^c) \leq \varepsilon. \tag{2.38}$$

□

Lemma 2.17. *For every* $\kappa > 0$, $t \in \mathbb{N}$ *and* $N \in \mathbb{N}$, *the distributions of*

$$\{(\tilde{X}_n^{m(t)}, \Gamma_n^{m(t)})\}_{n=0}^{\infty} \quad \textit{and} \quad \{(\tilde{X}_n^{i(t,\kappa)}, \Gamma_n^{i(t,\kappa)})\}_{n=0}^{\infty}$$

satisfy the following:

$$\begin{aligned} \|\tilde{A}^{i(t,\kappa)}_{t+N}(\cdot) - \tilde{A}^{m(t)}_{t+N}(\cdot)\|_{TV} &\leq \mathbb{E}\Big[\|\tilde{A}^{i(t,\kappa),\mathcal{G}_t^*}_{t+N}(\cdot) - \tilde{P}^N_{\Gamma_t}(\tilde{X}_t, \cdot)\|_{TV}\Big] \\ &\leq \kappa N^2. \end{aligned} \tag{2.39}$$

Proof. We begin with applying Jensen's inequality to get

$$\|\tilde{A}^{i(t,\kappa)}_{t+N}(\cdot) - \tilde{A}^{m(t)}_{t+N}(\cdot)\|_{TV} \leq \mathbb{E}\Big[\|\tilde{A}^{i(t,\kappa),\mathcal{G}_t^*}_{t+N}(\cdot) - \tilde{A}^{m(t),\mathcal{G}_t^*}_{t+N}(\cdot)\|_{TV}\Big]. \tag{2.40}$$

Now recall equations (2.27) and (2.28) to note that

$$\tilde{A}_{t+N}^{m(t),\mathcal{G}_t^*}(\cdot) = \tilde{P}_{\Gamma_t}^N(\tilde{X}_t^{m(t)}, \cdot) = \tilde{P}_{\Gamma_t}^N(\tilde{X}_t, \cdot), \tag{2.41}$$

that is $\{\tilde{X}_n^{m(t)}\}_{n=t}^{t+N}$ is a Markov chain started from $\tilde{X}_t$ with dynamics $\tilde{P}_{\Gamma_t}$. Combining (2.40) with (2.41) yields (2.39).

Now recall (2.29), (2.30) and (2.31), i.e. the dynamics of

$$\{(\tilde{X}_n^{i(t,\kappa)}, \Gamma_n^{i(t,\kappa)})\}_{n=K-N}^{K},$$

and observe that (2.29) yields

$$\sup_{\tilde{x}\in\tilde{\mathcal{X}}} \|\tilde{P}_{\Gamma_t}(\tilde{x}, \cdot) - \tilde{P}_{\Gamma_n^{i(t,\kappa)}}(\tilde{x}, \cdot)\|_{TV} \leq N\kappa \quad \text{for } n = t, \ldots, t+N-1. \tag{2.42}$$

We now recall a fact stated in Proposition 3(g) of Roberts and Rosenthal (2004). Given probability measures μ_1 and μ_2, there exist jointly defined random variables $Y_1 \sim \mu_1(\cdot)$ and $Y_2 \sim \mu_2(\cdot)$, such that $\mathbb{P}(Y_1 \neq Y_2) = \|\mu_1(\cdot) - \mu_2(\cdot)\|_{TV}$. Hence, in our case, for every $n = t, \ldots, t+N-1$, if $\tilde{X}_n^{m(t)} = \tilde{X}_n^{i(t,\kappa)}$, then by (2.42) the fact stated above, there exists a coupling of $\tilde{X}_{n+1}^{m(t)}$ and $\tilde{X}_{n+1}^{i(t,\kappa)}$, such that

$$\mathbb{P}[\tilde{X}_{n+1}^{m(t)} = \tilde{X}_{n+1}^{i(t,\kappa)}] \geq 1 - N\kappa.$$

The intuition behind this step is that $\tilde{X}_n^{m(t)} = \tilde{X}_n^{i(t,\kappa)}$ implies that these processes are equal at the next iteration with high probability.

Reiterating the above construction N times from $n = t$ to $n = t+N-1$ implies that there exists a coupling such that

$$\mathbb{P}[\tilde{X}_{t+N}^{m(t)} = \tilde{X}_{t+N}^{i(t,\kappa)} \mid \tilde{X}_t^{m(t)} = \tilde{X}_t^{i(t,\kappa)}] \geq 1 - N^2\kappa.$$

Hence by the coupling inequality

$$\|\tilde{A}_{t+N}^{i(t,\kappa),\mathcal{G}_t^*}(\cdot) - \tilde{P}_{\Gamma_t}^N(\tilde{X}_t, \cdot)\|_{TV} \leq \kappa N^2,$$

as required. □

Proof of Theorem 2.2

Proof. By the triangle inequality, for any n, t, and κ, we have

$$\begin{aligned} T_n(\tilde{x}, \gamma) = \|A_n(\cdot) - \pi(\cdot)\|_{TV} &\leq \|A_n(\cdot) - A_n^{i(t,\kappa)}(\cdot)\|_{TV} \\ &\quad + \|A_n^{i(t,\kappa)}(\cdot) - A_n^{m(t)}(\cdot)\|_{TV} \\ &\quad + \|A_n^{m(t)}(\cdot) - \pi(\cdot)\|_{TV} \\ &\leq \|\tilde{A}_n(\cdot) - \tilde{A}_n^{i(t,\kappa)}(\cdot)\|_{TV} \\ &\quad + \|\tilde{A}_n^{i(t,\kappa)}(\cdot) - \tilde{A}_n^{m(t)}(\cdot)\|_{TV} \\ &\quad + \|A_n^{m(t)}(\cdot) - \pi(\cdot)\|_{TV} \\ &=: \diamondsuit_n^{(1)} + \diamondsuit_n^{(2)} + \diamondsuit_n^{(3)}, \end{aligned}$$

where in the second inequality, for the first two terms, we have used Lemma 2.13.

Now fix $\delta > 0$. To prove the claim, it is enough to construct a target time $K_0 = K_0(\delta, \tilde{x}, \gamma)$, s.t.

$$T_K(\tilde{x}, \gamma) \leq \delta \quad \text{for all} \quad K > K_0.$$

We shall find such a target time of the form $K_0 = t_0 + N_0$. To this end let $\varepsilon = \delta/3$.

First, we use Lemma 2.14 to fix $N_0 = N_0(\varepsilon)$ such that

$$\diamondsuit_{t+N_0}^{(3)} \leq \varepsilon \quad \text{for all} \quad t \in \mathbb{N}, \tilde{x} \in \tilde{\mathcal{X}} \text{ and } \gamma \in \mathcal{Y}. \tag{2.43}$$

Next, we set $\kappa := \varepsilon/N_0^2$ and we use Lemma 2.17 to conclude that

$$\diamondsuit_{t+N_0}^{(2)} \leq \varepsilon \quad \text{for all} \quad t \in \mathbb{N}, \tilde{x} \in \tilde{\mathcal{X}} \text{ and } \gamma \in \mathcal{Y}. \tag{2.44}$$

Finally, use Lemma 2.16 to find $t_0 = t_0(\varepsilon, \kappa, N_0)$ such that

$$\diamondsuit_{t+N_0}^{(1)} \leq \varepsilon \quad \text{for all} \quad t \geq t_0, \tilde{x} \in \tilde{\mathcal{X}} \text{ and } \gamma \in \mathcal{Y}. \tag{2.45}$$

Letting $K_0 := t_0 + N_0$ allows us to decompose every $K \geq K_0$ into $K = t + N_0$, so that (2.43), (2.44), (2.45) are satisfied, which yields the claim. □

Proof of Theorem 2.3

Proof. The proof is identical, except that we use Lemma 2.15 instead of Lemma 2.14 to find $N_0 = N_0(\varepsilon, \tilde{x}, \gamma)$ in (2.43). □

Proof of Theorem 2.4

Proof. To prove that the Weak Law of Large Numbers holds for the Auxiliary Variable Adaptive MCMC class, recall again the sequences

$$\{(\tilde{X}_n^{m(t)}, \Gamma_n^{m(t)})\}_{n=0}^{\infty} \quad \text{and} \quad \{(\tilde{X}_n^{i(t,\kappa)}, \Gamma_n^{i(t,\kappa)})\}_{n=0}^{\infty}$$

defined above. Without loss of generality we will assume that $\pi(g) = 0$ and that $|g(x)| < a$.

By Markov's inequality

$$\mathbb{P}\left(\frac{1}{T}\Big|\sum_{i=1}^{T} g\left(X_i\right)\Big| \geq \delta^{1/2}\right) \leq \frac{1}{\delta^{1/2}}\mathbb{E}\left(\frac{1}{T}\Big|\sum_{i=1}^{T} g\left(X_i\right)\Big|\right),$$

hence to obtain the WLLN it is enough to show that for every $\delta > 0$ there exists such $T_0 = T_0(\delta) = T_0(\delta, \tilde{x}, \gamma)$, where $(\tilde{x}, \gamma)$ are the starting points of $(\tilde{X}_n, \Gamma_n)$, that for all $T > T_0$

$$\mathbb{E}\left(\frac{1}{T}\Big|\sum_{i=1}^{T} g\left(X_i\right)\Big|\right) \leq \delta. \tag{2.46}$$

To prove (2.46) we will need to consider second moments which makes it necessary to deal with mixed terms of the form $\mathbb{E}g(X_i)g(X_j)$. Let $\varepsilon > 0$ be fixed (a specific value will be fixed later). Firstly, for $i < j$ consider the following calculation.

$$\begin{aligned}
\left|\mathbb{E}g(X_i)g(X_j^{m(i)})\right| &\leq \left|\mathbb{E}\left(\mathbb{E}\left[g(X_i)g(X_j^{m(i)})|\mathcal{G}_i^*\right]\right)\right| \\
&\leq \left|\mathbb{E}\left(g(X_i)\mathbb{E}\left[g(X_j^{m(i)})|\mathcal{G}_i^*\right]\right)\right| \\
&\leq \mathbb{E}\left(|g(X_i)|\left|\mathbb{E}\left[g(X_j^{m(i)}) - \pi(g)|\mathcal{G}_i^*\right]\right|\right) \\
&\leq a^2\mathbb{E}\left[\|A_j^{m(i),\mathcal{G}_i^*}(\cdot) - \pi(\cdot)\|_{TV}\right] \\
&\leq a^2\varepsilon, \quad \text{for all } i \in \mathbb{N} \text{ and } j - i \geq N_0 = N_0(\varepsilon, \tilde{x}, \gamma),
\end{aligned} \tag{2.47}$$

where $N_0(\varepsilon, \tilde{x}, \gamma)$ has been obtained from Lemma 2.14, if assuming Simultaneous Uniform Ergodicity, or from Lemma 2.15, if assuming Containment.

Secondly, given N_0 in (2.47), fix $N_1 \geq N_0$ such that $1/N_1 < \varepsilon$ and consider pairs i, j satisfying $N_1 \leq j - i \leq N_1^2$. Set $\kappa := \varepsilon/N_1^4$, and compute

$$\begin{aligned}
\left|\mathbb{E}g(X_i)(g(X_j^{i(i,\kappa)}) - g(X_j^{m(i)}))\right| &\leq E|g(X_i)||g(X_j^{i(i,\kappa)}) - g(X_j^{m(i)})| \\
&\leq a^2 \|A_j^{i(i,\kappa)}(\cdot) - A_j^{m(i)}(\cdot)\|_{TV} \\
\text{(by Lemmas 2.13 and 2.17)} \quad &\leq a^2 \kappa (j-i)^2 \quad \text{(for all } \kappa, i, j) \\
&\leq a^2 \varepsilon
\end{aligned} \tag{2.48}$$

The last inequality in the expression above follows since $N_1 \leq j - i \leq N_1^2$ and $\kappa = \varepsilon/N_1^4$.

Finally, use Lemma 2.16 to find $t_0 = t_0(\varepsilon, \kappa, N_1^2)$ and conclude that

$$\begin{aligned}
\left|\mathbb{E}g(X_i)(g(X_j) - g(X_j^{i(i,\kappa)}))\right| &\leq \mathbb{E}|g(X_i)||g(X_j) - g(X_j^{i(i,\kappa)})| \\
&\leq a^2 \|A_j(\cdot) - A_j^{i(i,\kappa)}(\cdot)\|_{TV} \\
&\leq a^2 \varepsilon
\end{aligned} \tag{2.49}$$

for all $i > t_0 = t_0(\varepsilon, \kappa, N_1^2)$ and $j - i \leq N_1^2$.

We will now address the mixed term $\mathbb{E}g(X_i)g(X_j)$. Since $|g| < a$, trivially for any i, j

$$|\mathbb{E}g(X_i)g(X_j)| \leq a^2. \tag{2.50}$$

Moreover, for $\varepsilon > 0$, $N_1 \geq N_0(\varepsilon, \tilde{x}, \gamma)$, $\kappa = \varepsilon/N_1^4$, pairs i, j such that $N_1 \leq j - i \leq N_1^2$, and $i > t_0(\varepsilon, \kappa, N_1^2)$ equations (2.47), (2.48) and (2.49) yield

$$\begin{aligned}
|\mathbb{E}g(X_i)g(X_j)| \leq &\left|\mathbb{E}g(X_i)(g(X_j) - g(X_j^{i(i,\kappa)}))\right| \\
&+ \left|\mathbb{E}g(X_i)(g(X_j^{i(i,\kappa)}) - g(X_j^{m(i)}))\right| \\
&+ \left|\mathbb{E}g(X_i)g(X_j^{m(i)})\right| \\
\leq &\, 3a^2\varepsilon.
\end{aligned} \tag{2.51}$$

Consequently, for any $N_1 > \max\{1/\varepsilon, N_0\}$ and $t_1 > t_0$, chosen as above, we can compute

$$\begin{aligned}\mathbb{E}\left(\frac{1}{N_1^2}\sum_{k=t_1+1}^{t_1+N_1^2} g(X_k)\right)^2 &\leq \frac{1}{N_1^4}\Bigg(\sum_{i,j=t_1+1}^{t_1+N_1^2} \mathbb{I}_{\{|j-i|\geq N_1\}}|\mathbb{E}g(X_i)g(X_j)| \\ &\qquad + \sum_{i,j=t_1+1}^{t_1+N_1^2} \mathbb{I}_{\{|j-i|<N_1\}}|\mathbb{E}g(X_i)g(X_j)|\Bigg) \\ &\leq \frac{2}{N_1^4}\left(N_1^4 3a^2\varepsilon + N_1^3 a^2\right) \leq 8a^2\varepsilon, \end{aligned} \tag{2.52}$$

where we have used (2.51) and (2.50) to bound the first and second sum, respectively. The Cauchy-Schwartz inequality combined with (2.52) implies

$$\mathbb{E}\left|\frac{1}{N_1^2}\sum_{k=t_1+1}^{t_1+N_1^2} g(X_k)\right| \leq 2\sqrt{2}a\varepsilon^{1/2} \tag{2.53}$$

for $N_1 > \max\{1/\varepsilon, N_0(\varepsilon, \tilde{x}, \gamma)\}$ and $t_1 > t_0(\varepsilon, \kappa, N_1^2)$. Following the proof of Theorem 5 of Roberts and Rosenthal (2007), fix T so large that

$$\max\left[\frac{at_0}{T}, \frac{aN_1^2}{T}\right] \leq \varepsilon^{1/2}. \tag{2.54}$$

Use (2.53) and (2.54) to observe that

$$\begin{aligned}\mathbb{E}\left(\frac{1}{T}\Big|\sum_{i=1}^{T} g\left(X_i\right)\Big|\right) &\leq \mathbb{E}\left(\frac{1}{T}\Big|\sum_{i=1}^{t_0} g\left(X_i\right)\Big|\right) \\ &\qquad + \mathbb{E}\left(\frac{N_1^2}{T}\sum_{j=1}^{\lfloor\frac{T-t_0}{N_1^2}\rfloor}\frac{1}{N_1^2}\Big|\sum_{k=1}^{N_1^2} g\left(X_{t_0+(j-1)N_1^2+k}\right)\Big|\right) \\ &\qquad + \mathbb{E}\left(\frac{1}{T}\Big|\sum_{i=t_0+\lfloor\frac{T-t_0}{N_1^2}\rfloor N_1^2+1}^{T} g\left(X_i\right)\Big|\right) \\ &\leq \frac{at_0}{T} + 2\sqrt{2}a\varepsilon^{1/2} + \frac{aN_1^2}{T} \\ &\leq 2(\sqrt{2}a+1)\varepsilon^{1/2}.\end{aligned}$$

Setting $\varepsilon := (\delta/2(\sqrt{2}a+1))^2$ in the above argument yields (2.46) as desired. □

Proof of Lemma 2.5

Proof. We will begin the proof by showing that assumption (2.11) implies that an analogous drift condition is satisfied for $\tilde{P}_\gamma^{n_0}$, n_0 defined in (2.12), perhaps with different constants λ and b, which we define below. For any $k \in \{1, \ldots, n_0\}$ we have

$$\tilde{P}_\gamma^k(\tilde{x}) := \mathbb{E}\left(V_{\tilde{\pi}_\gamma}(\tilde{X}_{n+k}) | \tilde{X}_n = \tilde{x}, \Gamma_n = \gamma, \Gamma_{n+1} = \gamma, \ldots, \Gamma_{n+k-1} = \gamma\right).$$

For $k = 2$ we have

$$\begin{aligned}
\tilde{P}_\gamma^2(\tilde{x}) &= \mathbb{E}\left(V_{\tilde{\pi}_\gamma}(\tilde{X}_{n+2}) | \tilde{X}_n = \tilde{x}, \Gamma_n = \gamma, \Gamma_{n+1} = \gamma\right) \\
&= \mathbb{E}\left(\mathbb{E}\left(V_{\tilde{\pi}_\gamma}(\tilde{X}_{n+2}) | \tilde{X}_{n+1}, \tilde{X}_n = \tilde{x}, \Gamma_n = \gamma, \Gamma_{n+1} = \gamma\right) | \tilde{X}_n = \tilde{x}, \Gamma_n = \gamma, \Gamma_{n+1} = \gamma\right) \\
&\leq \mathbb{E}\left(\lambda V_{\tilde{\pi}_\gamma}(\tilde{X}_{n+1}) + b | \tilde{X}_n = \tilde{x}, \Gamma_n = \gamma, \Gamma_{n+1} = \gamma\right) \\
&= b + \lambda \mathbb{E}\left(V_{\tilde{\pi}_\gamma}(\tilde{X}_{n+1}) | \tilde{X}_n = \tilde{x}, \Gamma_n = \gamma, \Gamma_{n+1} = \gamma\right) \\
&\leq b + \lambda\left(\lambda V_{\tilde{\pi}_\gamma}(\tilde{x}) + b\right) \leq \lambda^2 V_{\tilde{\pi}_\gamma}(\tilde{x}) + 2b.
\end{aligned}$$

By similar calculations and induction we obtain

$$\tilde{P}_\gamma^{n_0}(\tilde{x}) \leq \lambda^{n_0} V_{\tilde{\pi}_\gamma}(\tilde{x}) + n_0 b, \tag{2.55}$$

as required.

By Theorem 12 of Rosenthal (1995) and conditions (2.12) and (2.55), there exists $K < \infty$ and $\rho < 1$, depending only on λ, b, v, n_0 and δ, such that for each $\gamma \in \mathcal{Y}$ and for any $k \in \mathbb{N}$ we have

$$\|\tilde{P}_\gamma^{n_0 \cdot k}(\tilde{x}, \cdot) - \tilde{\pi}_\gamma(\cdot)\|_{TV} \leq \left(K + V_{\tilde{\pi}_\gamma}(\tilde{x})\right) \rho^k.$$

We now use the monotonicity of $\|\tilde{P}_\gamma^n(\tilde{x}, \cdot) - \tilde{\pi}_\gamma(\cdot)\|_{TV}$ in n (see Proposition 3b) of Roberts and Rosenthal (2004)) to argue that

$$\begin{aligned}
\|\tilde{P}_\gamma^{n_0 \cdot k + n_0 - 1}(\tilde{x}, \cdot) - \tilde{\pi}_\gamma(\cdot)\|_{TV} \leq \ldots &\leq \|\tilde{P}_\gamma^{n_0 \cdot k + 1}(\tilde{x}, \cdot) - \tilde{\pi}_\gamma(\cdot)\|_{TV} \\
&\leq \|\tilde{P}_\gamma^{n_0 \cdot k}(\tilde{x}, \cdot) - \tilde{\pi}_\gamma(\cdot)\|_{TV} \\
&\leq \left(K + V_{\tilde{\pi}_\gamma}(\tilde{x})\right) \rho^k \\
&= \left(K/\rho + V_{\tilde{\pi}_\gamma}(\tilde{x})/\rho\right) \rho^{k+1}.
\end{aligned}$$

Let $\tilde{\rho} := \rho^{\frac{1}{n_0}}$. It follows that for every $m \in \mathbb{N}$

$$\|\tilde{P}_\gamma^m(\tilde{x}, \cdot) - \tilde{\pi}_\gamma(\cdot)\|_{TV} \leq \left(K/\rho + V_{\tilde{\pi}_\gamma}(\tilde{x})/\rho\right) \tilde{\rho}^m. \tag{2.56}$$

The next step of the proof will be to show that the sequence $V_{\tilde{\pi}_{\Gamma_n}}(\tilde{X}_n)$ is bounded in probability. By Lemma 3 in Roberts and Rosenthal (2007), it suffices to show that $\sup_n \mathbb{E} V_{\tilde{\pi}_{\Gamma_n}}(\tilde{X}_n) < \infty$. Firstly, let us show that $\tilde{P} V_{\tilde{\pi}_\gamma}(\tilde{x})$ is bounded for $\gamma \in \mathcal{Y}$ and $\tilde{x} \in A$. Note that

$$\begin{aligned}
\sup_{\gamma \in \mathcal{Y}} \sup_{\tilde{x} \in A} \tilde{P}_\gamma V_{\tilde{\pi}_\gamma}(\tilde{x}) &= \sup_{\gamma \in \mathcal{Y}} \sup_{\tilde{x} \in A} \left(\frac{\tilde{P}_\gamma V_{\tilde{\pi}_\gamma}(\tilde{x})}{V_{\tilde{\pi}_\gamma}(\tilde{x})} V_{\tilde{\pi}_\gamma}(\tilde{x}) \right) \\
&\leq \sup_{\gamma \in \mathcal{Y}} \sup_{\tilde{x} \in \mathcal{X}} \frac{\tilde{P}_\gamma V_{\tilde{\pi}_\gamma}(\tilde{x})}{V_{\tilde{\pi}_\gamma}(\tilde{x})} \sup_{\gamma \in \mathcal{Y}} \sup_{\tilde{x} \in A} V_{\tilde{\pi}_\gamma}(\tilde{x}).
\end{aligned}$$

Since A and $\mathcal{Y}$ were assumed to be compact, $\sup_{\gamma \in \mathcal{Y}} \sup_{\tilde{x} \in A} V_{\tilde{\pi}_\gamma}(\tilde{x}) < \infty$. Additionally, the drift condition (2.11) yields

$$\sup_{\gamma \in \mathcal{Y}} \sup_{\tilde{x} \in \mathcal{X}} \frac{\tilde{P}_\gamma V_{\tilde{\pi}_\gamma}(\tilde{x})}{V_{\tilde{\pi}_\gamma}(\tilde{x})} \leq \sup_{\gamma \in \mathcal{Y}} \sup_{\tilde{x} \in \mathcal{X}} \frac{\lambda V_{\tilde{\pi}_\gamma}(\tilde{x}) + b}{V_{\tilde{\pi}_\gamma}(\tilde{x})} \leq \lambda + b.$$

Therefore we can define $M := \sup_{\gamma \in \mathcal{Y}} \sup_{\tilde{x} \in A} \tilde{P}_\gamma V_{\tilde{\pi}_\gamma}(\tilde{x}) < \infty$. It follows that

$$\begin{aligned}
E\left(V_{\tilde{\pi}_{\Gamma_{n+1}}}(\tilde{X}_{n+1}) | \tilde{X}_n, \Gamma_n\right) &= \mathbb{E}\left(V_{\tilde{\pi}_{\Gamma_{n+1}}}(\tilde{X}_{n+1}) | \tilde{X}_n, \Gamma_n\right) \mathbb{I}_{\tilde{X}_n \in A} \\
&\quad + \mathbb{E}\left(V_{\tilde{\pi}_{\Gamma_{n+1}}}(\tilde{X}_{n+1}) | \tilde{X}_n, \Gamma_n\right) \mathbb{I}_{\tilde{X}_n \notin A} \\
&= \mathbb{E}\left(V_{\tilde{\pi}_{\Gamma_{n+1}}}(\tilde{X}_{n+1}) | \tilde{X}_n, \Gamma_n\right) \mathbb{I}_{\tilde{X}_n \in A} \\
&\quad + \mathbb{E}\left(V_{\tilde{\pi}_{\Gamma_n}}(\tilde{X}_{n+1}) | \tilde{X}_n, \Gamma_n\right) \mathbb{I}_{\tilde{X}_n \notin A} \\
&\leq \sup_{\tilde{x} \in A} \tilde{P}_{\Gamma_{n+1}} V_{\tilde{\pi}_{\Gamma_{n+1}}}(\tilde{x}) \\
&\quad + \mathbb{E}\left(V_{\tilde{\pi}_{\Gamma_n}}(\tilde{X}_{n+1}) | \tilde{X}_n, \Gamma_n\right) \\
&\leq \sup_{\gamma \in \mathcal{Y}} \sup_{\tilde{x} \in A} \tilde{P}_\gamma V_{\tilde{\pi}_\gamma}(\tilde{x}) + \lambda V_{\tilde{\pi}_{\Gamma_n}}(\tilde{X}_n) + b \\
&\leq M + \lambda V_{\tilde{\pi}_{\Gamma_n}}(\tilde{X}_n) + b.
\end{aligned} \tag{2.57}$$

By the law of total expectation,

$$\mathbb{E} V_{\tilde{\pi}_{\Gamma_{n+1}}}(\tilde{X}_{n+1}) = \mathbb{E}\mathbb{E}\left(V_{\tilde{\pi}_{\Gamma_{n+1}}}(\tilde{X}_{n+1}) | \tilde{X}_n, \Gamma_n\right),$$

which combined with (2.57) gives

$$\mathbb{E}V_{\tilde{\pi}_{\Gamma_{n+1}}}(\tilde{X}_{n+1}) \leq \lambda \mathbb{E}V_{\tilde{\pi}_{\Gamma_n}}(\tilde{X}_n) + M + b.$$

This implies, using Lemma 2 in Roberts and Rosenthal (2007) that

$$\sup_n \mathbb{E}V_{\tilde{\pi}_{\Gamma_n}}(\tilde{X}_n) \leq \max\left[\mathbb{E}V_{\tilde{\pi}_{\Gamma_0}}(\tilde{X}_0), \frac{M+b}{1-\lambda}\right].$$

Lemma 2.5 will now follow from combining the fact that the sequence $V_{\tilde{\pi}_{\Gamma_n}}(\tilde{X}_n)$ is bounded in probability with (2.56). Note that for any fixed ε and $\tilde{\delta}$, there exists N such that

$$\begin{aligned}
\mathbb{P}\left(M_\varepsilon(\tilde{X}_n, \Gamma_n) \leq N\right) &= \mathbb{P}\left(\|\tilde{P}^N_{\Gamma_n}(\tilde{X}_n, \cdot) - \tilde{\pi}_\gamma(\cdot)\|_{TV} \leq \varepsilon\right) \\
&\geq \mathbb{P}\left(\left(K/\rho + V_{\tilde{\pi}_{\Gamma_n}}(\tilde{X}_n)/\rho\right)\tilde{\rho}^N \leq \varepsilon\right) \\
&= \mathbb{P}\left(V_{\tilde{\pi}_{\Gamma_n}}(\tilde{X}_n) \leq \varepsilon\tilde{\rho}^{-N}\rho - K\right) \\
&\geq 1 - \tilde{\delta}
\end{aligned}$$

for all $n \in \mathbb{N}$. The last inequality holds since $\varepsilon\tilde{\rho}^{-N}\rho - K \to \infty$ as $N \to \infty$ and $V_{\Gamma_n}(\tilde{X}_n)$ is bounded in probability. □

A .2 Proofs for Section 4

Algorithm 3 is the modified version of JAMS, considered in Section 4 . Remark (2.18) shows one possible method of constructing the compact sets $A_1, \ldots A_N$.

Remark 2.18. Even though the theory presented in Section 4 works for any choice of the compact sets $A_1, \ldots, A_N$, we propose to define these sets in the following way. Recall that the burn-in algorithm provides the list of mode locations $\{\mu_1, ..., \mu_N\}$ and initial estimates of covariance matrices $\{\Sigma_{\gamma_0,1}, ..., \Sigma_{\gamma_0,N}\}$. By λ_i denote the maximum eigenvalue of $\Sigma_{\gamma_0,i}$ and let $\lambda_M = \max\{\lambda_1, ..., \lambda_N\}$. Let C be the convex hull of $\{\mu_1, ..., \mu_N\}$ and D_C its diameter. Define

$$A_i := \{x \in \mathcal{X} : \|x - \mu_i\| \leq 2D_C + 100(d\lambda_M)^{1/2}\},$$

where d is the dimension of $\mathcal{X}$.

In our proofs $\tilde{P}_{\gamma,L,i}$ will denote the kernel associated with the local move around mode i

and analogously $\tilde{P}_{\gamma,J,i}$ will be the kernel of the jump to mode i. The full transition kernel $\tilde{P}_\gamma$ is thus defined as

$$\tilde{P}_\gamma\left((x,i),(dy,k)\right) := (1-\epsilon)\tilde{P}_{\gamma,L,i}\left((x,i),(dy,k)\right)\delta_{i=k} + \epsilon a_{\gamma,ik}\tilde{P}_{\gamma,J,k}\left((x,i),(dy,k)\right).$$

Algorithm 3 mJAMS (modified JAMS): main algorithm, iteration $n+1$

Input: current point (x_n, i_n), list of modes $\{\mu_1, \dots \mu_N\}$, constant $\epsilon \in (0,1)$, parameter $\gamma_n = \{\Sigma, w_i, a_{ik}\}_{i,k\in\{1,\dots,N\}}$, empirical means $m_1, \dots, m_N$ and covariance matrices $S_1, \dots, S_N$, integer $N_j^* \geq n+1$, next element of the lag sequence n_{j+1}.
Generate $u \sim U[0,1]$.
if $u > \epsilon$ **then**
    Local move:
    Propose a new value $y \sim R_{\gamma_n,L,i_n}(x_n, \cdot)$.
    Accept y with probability $\alpha_{\gamma_n,L}\left((x_n,i_n) \to (y,i_n)\right)$.
    if y accepted **then**
        $(x_n, i_n) := (y, i_n)$.
    else
        $(x_n, i_n) := (x_n, i_n)$.
    end if
else
    if $x_n \notin JR_{\gamma_n,i_n}$ **then go to** 4 ▷ **(only for deterministic jumps)**
    end if
    Jump move:
    Propose a new mode $k \sim (a_{i1}, \dots, a_{iN})$.
    Propose a new value $y \sim R_{\gamma_n,J,ik}(x_n, \cdot)$.
    Accept (y,k) with probability $\alpha_{\gamma_n,J}\left((x_n,i_n) \to (y,k)\right)$.
    if (y,k) accepted **then**
        $(x_{n+1}, i_{n+1}) := (y,k)$.
    else
        $(x_{n+1}, i_{n+1}) := (x_n, i_n)$.
    end if
end if
if $x_{n+1} \in A_{i_{n+1}}$ **then**
    Update the empirical mean $m_{i_{n+1}}$ and covariance matrix $S_{i_{n+1}}$ by including x_{n+1}.
    if $n+1$ is equal to N_j^* **then**
        Update the parameter γ_n to γ_{n+1} according to Algorithm 2.
        Sample $N_{j+1}^* = N_j^* + n_{j+1} + \text{Uniform}\left[0, \lfloor (j+1)^{\kappa^*} \rfloor\right]$.
    end if
end if
return New sample (x_{n+1}, i_{n+1}), parameter γ_{n+1}, $m_{i_{n+1}}$ and $S_{i_{n+1}}$.

Proof of Theorem 2.6

Proof. The aim of the proof is to verify the assumptions of Theorem 2.2 and conclude. Diminishing Adaptation has been addressed in Section 3 .2, so it is enough to prove that Simultaneous Uniform Ergodicity holds. Note that assumption (2.17) implies that for some positive constant c_1

$$\frac{R_{\gamma,J,k}(y)}{\tilde{\pi}_\gamma(y,k)} = \frac{R_{\gamma,J,k}(y)}{\pi(y)} \frac{\sum_{j\in\mathcal{I}} w_{\gamma,j} Q_j(\mu_j, \Sigma_{\gamma,j})(y)}{w_{\gamma,k} Q_k(\mu_k, \Sigma_{\gamma,k})(y)} > \frac{R_{\gamma,J,k}(y)}{\pi(y)} > c_1$$

for each $k \in \mathcal{I}$, $y \in \mathcal{X}$ and $\gamma \in \mathcal{Y}$. For any $(x,i) \in \mathcal{X}\times\mathcal{I}$, any set $\hat{C} \subset \mathcal{X}$ and any $k \in \mathcal{I}$ we can compute

$$\begin{aligned}
\int_{\hat{C}} \tilde{P}_\gamma\left((x,i),(dy,k)\right) &\geq \epsilon a_{\gamma,ik} \int_{\hat{C}} \tilde{P}_{\gamma,J,k}\left((x,i),(dy,k)\right) \\
&= \epsilon a_{\gamma,ik} \int_{\hat{C}} R_{\gamma,J,k}(y) \min\left[1, \frac{\tilde{\pi}_\gamma(y,k) a_{\gamma,ki} R_{\gamma,J,i}(x)}{\tilde{\pi}_\gamma(x,i) a_{\gamma,ik} R_{\gamma,J,k}(y)}\right] dy \\
&\geq \epsilon \int_{\hat{C}} \min\left[a_{\gamma,ik} R_{\gamma,J,k}(y), \tilde{\pi}_\gamma(y,k) a_{\gamma,ki} \frac{R_{\gamma,J,i}(x)}{\tilde{\pi}_\gamma(x,i)}\right] dy \\
&\geq \epsilon\epsilon_a \int_{\hat{C}} \min\left[c_1 \tilde{\pi}_\gamma(y,k), c_1 \tilde{\pi}_\gamma(y,k)\right] dy \\
&= \epsilon\epsilon_a c_1 \int_{\hat{C}} \tilde{\pi}_\gamma(y,k) dy,
\end{aligned}$$

where ϵ_a is as in equation (2.16).

Furthermore, any set $C \subset \mathcal{X}\times\mathcal{I}$ may be decomposed as $C = \bigcup_{k\in\mathcal{I}} \hat{C}_k \times \{k\}$, therefore

$$\sum_{k\in\mathcal{I}} \int_{\hat{C}_k} \tilde{P}_\gamma\left((x,i),(dy,k)\right) \geq \sum_{k\in\mathcal{I}} \epsilon\epsilon_a c_1 \int_{\hat{C}_k} \tilde{\pi}_\gamma(y,k) dy = \epsilon\epsilon_a c_1 \tilde{\pi}_\gamma(C). \tag{2.58}$$

Since $\tilde{\pi}_\gamma$ is a probability measure on $\mathcal{X}\times I$ for each $\gamma \in \mathcal{Y}$ and (2.58) holds for all $(x,i) \in \mathcal{X}\times I$, by Theorem 8 of Roberts and Rosenthal (2004) we have

$$\|\tilde{P}_\gamma^n((x,i),\cdot) - \tilde{\pi}_\gamma(\cdot)\|_{TV} \leq (1-\epsilon\epsilon_a c_1)^n \quad \text{for all } (x,i) \in \mathcal{X}\times\mathcal{I} \text{ and } \gamma \in \mathcal{Y},$$

which completes the proof. □

Proof of Theorem 2.7

We will show that the assumptions of Theorem 2.3 are satisfied. Since Diminishing Adaptation was discussed in Section 3 .2, it suffices to prove that the Containment condition holds, which we will do using Lemma 2.5. Assumptions a) and d) were discussed in Section 4 .1. Assumption e) follows directly from the construction of the algorithm for $A := \bigcup_{i\in\mathcal{I}} A_i \times \{i\}$. Assumption f) holds trivially, since $\tilde{X}_0$ and Γ_0 are deterministic (chosen by the user of the algorithm). The remaining part of the proof is organized as follows.

We show that the drift condition expressed in assumption b) of Lemma 2.5 is satisfied under the assumptions of Theorem 2.7. To this end, we consider a drift function of the form

$$V_{\tilde{\pi}_\gamma}(\tilde{x}) := c\tilde{\pi}_\gamma(\tilde{x})^{-s} = c\tilde{\pi}_\gamma((x,i))^{-s} \tag{2.59}$$

for some $s \in (0,1)$ and c such that $c\pi(x)^{-s} \geq 1$ (thus enforcing $V_{\tilde{\pi}_\gamma}(\tilde{x}) > 1$). We first focus on obtaining the appropriate result for the local kernels and subsequently we combine it with the result for jumps. Finally, we prove that assumption c) of Lemma 2.5 is satisfied for $n_0 = 3$.

Assumption b) of Lemma 2.5 (local kernels)

Proof. The drift function $V_{\tilde{\pi}_\gamma}$ defined as above is a jointly continuous function of (x,γ) so it is bounded on compact sets in $\mathcal{X}\times\mathcal{Y}$ for each $i \in \mathcal{I}$, as required by assumption b) of Lemma 2.5. Therefore, it is also bounded on compact sets in $\tilde{\mathcal{X}}\times\mathcal{Y}$. The proof will be continued for $s = \frac{1}{2}$ but analogous reasoning would be valid for any $s \in (0,1)$.

We will prove that there exists $\lambda_L < 1$ such that for the local move kernels we have

$$\limsup_{|x|\to\infty} \sup_{\gamma\in\mathcal{Y}} \frac{\tilde{P}_{\gamma,L,i} V_{\tilde{\pi}_\gamma}((x,i))}{V_{\tilde{\pi}_\gamma}((x,i))} \leq \lambda_L \tag{2.60}$$

for all $i \in \mathcal{I}$. We will refer multiple times to the proof of Theorem 4.1 of Jarner and Hansen (2000). Following the notation used there, let $C_{\pi(x)}(\delta)$ denote the radial δ-zone around $C_{\pi(x)}$, where $C_{\pi(x)}$ is the contour manifold corresponding to $\pi(x)$. Firstly, there exists R_0 such that for $|x| > R_0$ the contour manifold $C_{\pi(x)}$ is parametrized by S^{d-1} and encloses the acceptance set for π defined as $a(x) := \{y \in \mathcal{X} : \pi(y) > \pi(x)\}$ (we refer to Section 4 of Jarner and Hansen (2000) for the details of this argument). In our proof we will only consider $|x| > R_0$.

Define also

$$\lambda_{L,i} := \limsup_{|x|\to\infty} \sup_{\gamma\in\mathcal{Y}} \int_{r_{\gamma,i}(x)} R_{\gamma,L,i}(x,y)dy. \tag{2.61}$$

By assumption (2.19) $\lambda_{L,i} < 1$.

Fix $i \in \mathcal{I}$ and $\epsilon > 0$. We will show that for sufficiently large x

$$\frac{\tilde{P}_{\gamma,L,i} V_{\tilde{\pi}_\gamma}((x,i))}{V_{\tilde{\pi}_\gamma}((x,i))} \leq \lambda_{i,L} + 3\epsilon + \epsilon^{1/2}. \tag{2.62}$$

The idea of this proof is to split $\mathcal{X}$ into disjoint sets $\mathcal{X} \setminus B(x,K)$, $B(x,K) \cap C_{\pi(x)}(\delta)$ and $B(x,K) \setminus C_{\pi(x)}(\delta)$ and show that for any x with a sufficiently large norm the integral representing acceptance, that is, of the function $R_{\gamma,L,i}(x,y) \min\left[1, \frac{\tilde{\pi}_\gamma(y,i)}{\tilde{\pi}_\gamma(x,i)}\right] \frac{V_{\tilde{\pi}_\gamma}((y,i))}{V_{\tilde{\pi}_\gamma}((x,i))}$ on those sets is bounded from above by ϵ, ϵ and $\epsilon^{1/2}$, respectively. We fix the values of K and δ below. As for the rejection part, we use (2.61) to show that the corresponding integral is bounded by $\lambda_{L,i} + \epsilon$, for all x at a sufficient distance from $\mathbf{0}$. Putting all these upper bounds together, we obtain the required $\lambda_{i,L} + 3\epsilon + \epsilon^{1/2}$.

Firstly, observe that by assumption a), condition (2.15), and the fact that the covariance matrices for the local proposal distributions are scaled by a fixed factor $2.38^2/d$, the family of distributions $R_{\gamma,L,i}(x,\cdot)$ is tight. Thus, there exists K such that

$$\sup_{\gamma\in\mathcal{Y}} \int_{\mathcal{X}\setminus B(x,K)} R_{\gamma,L,i}(x,y)dy \leq \epsilon. \tag{2.63}$$

Furthermore, as shown the proof of Theorem 4.1 of Jarner and Hansen (2000), under assumption c) of Theorem 2.7 for any positive δ and K

$$\mu^{\text{Leb}}\left(B(x,K) \cap C_{\pi(x)}(\delta)\right) \leq \delta \left(\frac{|x|+K}{|x|-K}\right)^{d-1} \frac{\mu^{\text{Leb}}\left(B(x,3K)\right)}{K}. \tag{2.64}$$

Fix K satisfying (2.63). Since $\lim_{x\to\infty} \left(\frac{|x|+K}{|x|-K}\right)^{d-1} = 1$, there exists $R_1 > 0$ such that for $|x| > \max[R_0, R_1]$

$$\left(\frac{|x|+K}{|x|-K}\right)^{d-1} < 1 + \epsilon.$$

Recall that by assumption a) for any $x \in \mathcal{X}$ we have

$$\sup_{\gamma\in\mathcal{Y}} \sup_{y\in\mathcal{X}} R_{\gamma,L,i}(x,y) > 0.$$

Now let us choose δ such that for $|x| > R_1$

$$\mu^{\text{Leb}}\left(C_{\pi(x)}(\delta) \cap B(x,K)\right) \leq \frac{\epsilon}{\sup_{\gamma\in\mathcal{Y}} \sup_{y\in\mathcal{X}} R_{\gamma,L,i}(x,y)},$$

therefore getting

$$\sup_{\gamma\in\mathcal{Y}} \int_{C_{\pi(x)}(\delta)\cap B(x,K)} R_{\gamma,L,i}(x,y)dy \leq \epsilon. \tag{2.65}$$

Let $r(x) = \{y \in \mathcal{X} : \pi(y) < \pi(x)\}$ and $a(x) = \{y \in \mathcal{X} : \pi(y) \geq \pi(x)\}$. We now split $B(x,K) \setminus C_{\pi(x)}(\delta)$ into $(r(x) \cap B(x,K)) \setminus C_{\pi(x)}(\delta)$ and $(a(x) \cap B(x,K)) \setminus C_{\pi(x)}(\delta)$ and we estimate the value of $\min\left[1, \frac{\tilde{\pi}_\gamma(y,i)}{\tilde{\pi}_\gamma(x,i)}\right] \frac{V_{\tilde{\pi}_\gamma}((y,i))}{V_{\tilde{\pi}_\gamma}((x,i))}$ on each of those sets separately. Fix $\tilde{K}$ such that

$$\frac{\sum_{j\in\mathcal{I}} w_{\gamma,j} Q_j(\mu_j, \Sigma_{\gamma,j})(x)}{w_{\gamma,i} Q_i(\mu_i, \Sigma_{\gamma,i})(x)} \leq \tilde{K} \quad \text{for all } x \in \mathcal{X} \text{ and } \gamma \in \mathcal{Y}. \tag{2.66}$$

This is possible by assumption d) combined with conditions (2.15) and (2.16). Since π is super-exponential, there exists R_2 so large that for $|x| > \max[R_0, R_2]$:

1) If $y \in (r(x) \cap B(x,K)) \setminus C_{\pi(x)}(\delta)$, then $\frac{\pi(y)}{\pi(x)} \leq \frac{\epsilon}{\tilde{K}}$.

2) If $y \in (a(x) \cap B(x,K)) \setminus C_{\pi(x)}(\delta)$, then $\frac{\pi(x)}{\pi(y)} \leq \frac{\epsilon}{\tilde{K}}$.

In the first case we have (using (2.66)):

$$\begin{aligned}\frac{\tilde{\pi}_\gamma(y,i)}{\tilde{\pi}_\gamma(x,i)} &= \frac{\pi(y)}{\pi(x)} \frac{\sum_{j\in\mathcal{I}} w_{\gamma,j} Q_j(\mu_j, \Sigma_{\gamma,j})(x)}{w_{\gamma,i} Q_i(\mu_i, \Sigma_{\gamma,i})(x)} \frac{w_{\gamma,i} Q_i(\mu_i, \Sigma_{\gamma,i})(y)}{\sum_{j\in\mathcal{I}} w_{\gamma,j} Q_j(\mu_j, \Sigma_{\gamma,j})(y)} \\ &\leq \frac{\pi(y)}{\pi(x)} \frac{\sum_{j\in\mathcal{I}} w_{\gamma,j} Q_j(\mu_j, \Sigma_{\gamma,j})(x)}{w_{\gamma,i} Q_i(\mu_i, \Sigma_{\gamma,i})(x)} \leq \tilde{K} \frac{\pi(y)}{\pi(x)} \leq \epsilon.\end{aligned}$$

Similarly for $y \in (a(x) \cap B(x,K)) \setminus C_{\pi(x)}(\delta)$ we get $\frac{\tilde{\pi}_\gamma(x,i)}{\tilde{\pi}_\gamma(y,i)} \leq \epsilon$. Hence, on $B(x,K) \setminus C_{\pi(x)}(\delta)$ we have

$$\min\left[1, \frac{\tilde{\pi}_\gamma(y,i)}{\tilde{\pi}_\gamma(x,i)}\right] \frac{V_{\tilde{\pi}_\gamma}((y,i))}{V_{\tilde{\pi}_\gamma}((x,i))} = \min\left[\frac{\tilde{\pi}_\gamma(x,i)^{1/2}}{\tilde{\pi}_\gamma(y,i)^{1/2}}, \frac{\tilde{\pi}_\gamma(y,i)^{1/2}}{\tilde{\pi}_\gamma(x,i)^{1/2}}\right] \leq \epsilon^{1/2}. \tag{2.67}$$

Furthermore, by assumption (2.61) we can choose R_3 such that for $|x| > R_3$

$$\sup_{\gamma\in\mathcal{Y}} \int_{r_{\gamma,i}(x)} R_{\gamma,L,i}(x,y)dy \leq \lambda_{L,i} + \epsilon. \tag{2.68}$$

Finally, for $|x| > \max[R_0, R_1, R_2, R_3]$ we obtain

$$\begin{aligned}
\frac{\tilde{P}_{\gamma,L,i}V_{\tilde{\pi}_\gamma}((x,i))}{V_{\tilde{\pi}_\gamma}((x,i))} &= \int_{\mathcal{X}} R_{\gamma,L,i}(x,y)\min\left[1,\frac{\tilde{\pi}_\gamma(y,i)}{\tilde{\pi}_\gamma(x,i)}\right]\frac{V_{\tilde{\pi}_\gamma}((y,i))}{V_{\tilde{\pi}_\gamma}((x,i))}dy \\
&\quad + \int_{\mathcal{X}} R_{\gamma,L,i}(x,y)\left(1-\min\left[1,\frac{\tilde{\pi}_\gamma(y,i)}{\tilde{\pi}_\gamma(x,i)}\right]\right)dy \\
&= \int_{\mathcal{X}} R_{\gamma,L,i}(x,y)\min\left[\frac{\tilde{\pi}_\gamma(x,i)^{1/2}}{\tilde{\pi}_\gamma(y,i)^{1/2}},\frac{\tilde{\pi}_\gamma(y,i)^{1/2}}{\tilde{\pi}_\gamma(x,i)^{1/2}}\right]dy \\
&\quad + \int_{r_{\gamma,i}(x)} R_{\gamma,L,i}(x,y)\left(1-\min\left[1,\frac{\tilde{\pi}_\gamma(y,i)}{\tilde{\pi}_\gamma(x,i)}\right]\right)dy \\
&=: \blacklozenge_1 + \blacklozenge_2 + \blacklozenge_3 + \blacklozenge_4,
\end{aligned}$$

where

$$\begin{aligned}
\blacklozenge_1 &:= \int_{\mathcal{X}\setminus B(x,K)} R_{\gamma,L,i}(x,y)\min\left[\frac{\tilde{\pi}_\gamma(x,i)^{1/2}}{\tilde{\pi}_\gamma(y,i)^{1/2}},\frac{\tilde{\pi}_\gamma(y,i)^{1/2}}{\tilde{\pi}_\gamma(x,i)^{1/2}}\right]dy \\
\blacklozenge_2 &:= \int_{B(x,K)\cap C_{\pi(x)}(\delta)} R_{\gamma,L,i}(x,y)\min\left[\frac{\tilde{\pi}_\gamma(x,i)^{1/2}}{\tilde{\pi}_\gamma(y,i)^{1/2}},\frac{\tilde{\pi}_\gamma(y,i)^{1/2}}{\tilde{\pi}_\gamma(x,i)^{1/2}}\right]dy \\
\blacklozenge_3 &:= \int_{B(x,K)\setminus C_{\pi(x)}(\delta)} R_{\gamma,L,i}(x,y)\min\left[\frac{\tilde{\pi}_\gamma(x,i)^{1/2}}{\tilde{\pi}_\gamma(y,i)^{1/2}},\frac{\tilde{\pi}_\gamma(y,i)^{1/2}}{\tilde{\pi}_\gamma(x,i)^{1/2}}\right]dy \\
\blacklozenge_4 &:= \int_{r_{\gamma,i}(x)} R_{\gamma,L,i}(x,y)\left(1-\frac{\tilde{\pi}_\gamma(y,i)}{\tilde{\pi}_\gamma(x,i)}\right)dy.
\end{aligned}$$

We will now bound from above each of the four terms $\blacklozenge_1$, $\blacklozenge_2$, $\blacklozenge_3$ and $\blacklozenge_4$. We have:

$$\begin{aligned}
\blacklozenge_1 &\le \int_{\mathcal{X}\setminus B(x,K)} R_{\gamma,L,i}(x,y)dy && \le \epsilon \quad \text{(by (2.63))}, \\
\blacklozenge_2 &\le \int_{B(x,K)\cap C_{\pi(x)}(\delta)} R_{\gamma,L,i}(x,y)dy && \le \epsilon \quad \text{(by (2.65))}, \\
\blacklozenge_3 &\le \int_{B(x,K)\setminus C_{\pi(x)}(\delta)} R_{\gamma,L,i}(x,y)\epsilon^{1/2}dy && \le \epsilon^{1/2} \quad \text{(by (2.67))} \\
\blacklozenge_4 &\le \int_{r_{\gamma,i}(x)} R_{\gamma,L,i}(x,y)dy && \le \lambda_{L,i}+\epsilon \quad \text{(by (2.68))}.
\end{aligned}$$

We therefore get

$$\frac{\tilde{P}_{\gamma,L,i}V_{\tilde{\pi}_\gamma}((x,i))}{V_{\tilde{\pi}_\gamma}((x,i))} = \blacklozenge_1 + \blacklozenge_2 + \blacklozenge_3 + \blacklozenge_4 \le \lambda_{L,i} + 3\epsilon + \epsilon^{1/2},$$

which completes the proof of (2.62). Consequently, by setting λ_L such that

$$\max_{i\in\mathcal{I}} \lambda_{i,L} < \lambda_L < 1,$$

we obtain (2.60). Observe that there exists $R_L > 0$ such that if $|x| > R_L$, then

$$\tilde{P}_{\gamma,L,i} V_{\tilde{\pi}_\gamma}((x,i)) \leq \lambda_L V_{\tilde{\pi}_\gamma}((x,i)).$$

For $|x| \leq R_L$ we have

$$\sup_{|x|<R_L} \sup_{\gamma\in\mathcal{Y}} \tilde{P}_{\gamma,L,i} V_{\tilde{\pi}_\gamma}((x,i)) \leq \sup_{|x|<R_L} \sup_{\gamma\in\mathcal{Y}} \frac{\tilde{P}_{\gamma,L,i} V_{\tilde{\pi}_\gamma}((x,i))}{V_{\tilde{\pi}_\gamma}((x,i))} \sup_{|x|<R_L} \sup_{\gamma\in\mathcal{Y}} V_{\tilde{\pi}_\gamma}((x,i)).$$

Now analogously to $r_{\gamma,i}(x)$, let us define the acceptance region for $\tilde{\pi}_\gamma$ as

$$a_{\gamma,i}(x) = \{y \in \mathcal{X} : \tilde{\pi}_\gamma(y,i) \geq \tilde{\pi}_\gamma(x,i)\}. \tag{2.69}$$

Note that

$$\begin{aligned}
\frac{\tilde{P}_{\gamma,L,i} V_{\tilde{\pi}_\gamma}((x,i))}{V_{\tilde{\pi}_\gamma}((x,i))} &= \int_{a_{\gamma,i}(x)} R_{\gamma,i}(x,y) \frac{V_{\tilde{\pi}_\gamma}((y,i))}{V_{\tilde{\pi}_\gamma}((x,i))} dy \\
&\quad + \int_{r_{\gamma,i}(x)} R_{\gamma,L,i}(x,y) \frac{\tilde{\pi}_\gamma(y,i)}{\tilde{\pi}_\gamma(x,i)} \frac{V_{\tilde{\pi}_\gamma}((y,i))}{V_{\tilde{\pi}_\gamma}((x,i))} dy \\
&\quad + \int_{r_{\gamma,i}(x)} R_{\gamma,L,i}(x,y) \left(1 - \frac{\tilde{\pi}_\gamma(y,i)}{\tilde{\pi}_\gamma(x,i)}\right) dy \\
&= \int_{a_{\gamma,i}(x)} R_{\gamma,L,i}(x,y) \frac{\tilde{\pi}_\gamma(x,i)^{1/2}}{\tilde{\pi}_\gamma(y,i)^{1/2}} dy \\
&\quad + \int_{r_{\gamma,i}(x)} R_{\gamma,L,i}(x,y) \left(1 - \frac{\tilde{\pi}_\gamma(y,i)}{\tilde{\pi}_\gamma(x,i)} + \frac{\tilde{\pi}_\gamma(y,i)^{1/2}}{\tilde{\pi}_\gamma(x,i)^{1/2}}\right) dy \\
&\leq 2 \int_{\mathcal{X}} R_{\gamma,L,i}(x,y) dy = 2.
\end{aligned}$$

Besides

$$\sup_{|x|<R_L} \sup_{\gamma\in\mathcal{Y}} V_{\tilde{\pi}_\gamma}((x,i)) < \infty$$

as for each i the function $V_{\tilde{\pi}_\gamma}((x,i))$ is jointly continuous with respect to x and γ. By setting

$$b_L := 2 \max_{i\in\mathcal{I}} \sup_{|x|<R_L} \sup_{\gamma\in\mathcal{Y}} V_{\tilde{\pi}_\gamma}((x,i))$$

we obtain

$$\tilde{P}_{\gamma,L,i}V_{\tilde{\pi}_\gamma}((x,i)) \leq \lambda_L V_{\tilde{\pi}_\gamma}((x,i)) + b_L \tag{2.70}$$

for all $(x,i) \in \mathcal{X} \times \mathcal{I}$. □

Assumption b) of Lemma 2.5 (jump kernels)

Proof. Firstly recall that under assumption e1) of Theorem 2.7 we have:

$$\begin{aligned}\tilde{P}_\gamma V_{\tilde{\pi}_\gamma}((x,i)) &= \sum_{k\in\mathcal{I}} \int_{\mathcal{X}} \tilde{P}_\gamma((x,i),(dy,k)) V_{\tilde{\pi}_\gamma}((y,k)) \\ &= (1-\epsilon)\int_{\mathcal{X}} \tilde{P}_{\gamma,L,i}((x,i),(dy,i)) V_{\tilde{\pi}_\gamma}((y,i)) \\ &\quad + \epsilon \sum_{k\in\mathcal{I}} a_{\gamma,ik} \int_{\mathcal{X}} \tilde{P}_{\gamma,J,k}((x,i),(dy,k)) V_{\tilde{\pi}_\gamma}((y,k)).\end{aligned} \tag{2.71}$$

if x belongs to the jumping region $JR_{\gamma,i}$, and

$$\tilde{P}_\gamma V_{\tilde{\pi}_\gamma}((x,i)) = \int_{\mathcal{X}} \tilde{P}_{\gamma,L,i}((x,i),(dy,i)) V_{\tilde{\pi}_\gamma}((y,i))$$

otherwise. Recall as well that all the jumping regions $JR_{\gamma,i}$ for $\gamma \in \mathcal{Y}$, $i \in \mathcal{I}$ are contained within a compact set D and consequently any point (y,k) proposed in a deterministic jump satisfies $(y,k) \in D \times \{k\}$. Let us now define

$$b_J := \sup_{\gamma\in\mathcal{Y}} \max_{k\in\mathcal{I}} \sup_{y\in D} V_{\tilde{\pi}_\gamma}((y,k)) < \infty.$$

Observe that

$$\sup_{\gamma\in\mathcal{Y}} \max_{k\in\mathcal{I}} \sup_{x\in D} \int_{\mathcal{X}} \tilde{P}_{\gamma,J,k}((x,i),(dy,k)) V_{\tilde{\pi}_\gamma}((y,k)) \leq b_J$$

and so for all (x,i)

$$\begin{aligned}\tilde{P}_\gamma V_{\tilde{\pi}_\gamma}((x,i)) \leq & \int_{\mathcal{X}} \tilde{P}_{\gamma,L,i}((x,i),(dy,i)) V_{\tilde{\pi}_\gamma}((y,i)) \\ &+ \sum_{k\in\mathcal{I}} a_{\gamma,ik} \int_{\mathcal{X}} \tilde{P}_{\gamma,J,k}((x,i),(dy,k)) V_{\tilde{\pi}_\gamma}((y,k)) \\ &\leq \lambda_L V_{\tilde{\pi}_\gamma}((x,i)) + b_L + b_J.\end{aligned}$$

Finally, setting $\lambda := \lambda_L$ and $b := b_L + b_J$ yields (2.11) under assumption e1).

Let us now consider assumption e2). Recall that for any $s \in (0,1)$ if $V_{\tilde{\pi}_\gamma}((x,i)) = c\tilde{\pi}_\gamma(x,i)^{-s}$, then (2.70) holds for some λ_L, b_L and R_L. Furthermore,

$$\begin{aligned}
&\int_{\mathcal{X}} \tilde{P}_{\gamma,J,k}((x,i),(dy,k))\, V_{\tilde{\pi}_\gamma}((y,k)) \\
&= \int_{\mathcal{X}} R_{\gamma,J,k}(y) \min\left[1, \frac{\tilde{\pi}_\gamma(y,k)}{\tilde{\pi}_\gamma(x,i)} \frac{a_{\gamma,ki} R_{\gamma,J,i}(x)}{a_{\gamma,ik} R_{\gamma,J,k}(y)}\right] V_{\tilde{\pi}_\gamma}((y,k))\, dy \\
&\quad + \left(1 - \int_{\mathcal{X}} R_{\gamma,J,k}(y) \min\left[1, \frac{\tilde{\pi}_\gamma(y,k)}{\tilde{\pi}_\gamma(x,i)} \frac{a_{\gamma,ki} R_{\gamma,J,i}(x)}{a_{\gamma,ik} R_{\gamma,J,k}(y)}\right] dy\right) V_{\tilde{\pi}_\gamma}((x,i)) \\
&\leq \int_{\mathcal{X}} R_{\gamma,J,k}(y) V_{\tilde{\pi}_\gamma}((y,k))\, dy + V_{\tilde{\pi}_\gamma}((x,i)).
\end{aligned} \tag{2.72}$$

By assumption (2.23) there exists a constant c_2 such that $\frac{R_{\gamma,J,i}(x)}{\pi(x)^{s_J}} < c_2$ for each $x \in \mathcal{X}$, $i \in \mathcal{I}$ and $\gamma \in \mathcal{Y}$ and as a consequence,

$$\frac{R_{\gamma,J,i}(x)}{\tilde{\pi}_\gamma(x,i)^{s_J}} = \frac{R_{\gamma,J,i}(x)}{\pi(x)^{s_J}} \left(\frac{\sum_{j\in\mathcal{I}} w_{\gamma,j} Q_j(\mu_j, \Sigma_{\gamma,j})(x)}{w_{\gamma,i} Q_i(\mu_i, \Sigma_{\gamma,i})(x)}\right)^{s_J} \leq c_2 \tilde{K}^{s_J},$$

where the last inequality follows from (2.66). Fix $s < s_J$ and observe that

$$\begin{aligned}
b_J &:= \sup_{\gamma\in\mathcal{Y}} \max_{k\in\mathcal{I}} \int_{\mathcal{X}} R_{\gamma,J,k}(y) V_{\tilde{\pi}_\gamma}((y,k))\, dy \\
&= \sup_{\gamma\in\mathcal{Y}} \max_{k\in\mathcal{I}} \int_{\mathcal{X}} R_{\gamma,J,k}(y) \tilde{\pi}_\gamma(y,k)^{-s} dy \\
&= \sup_{\gamma\in\mathcal{Y}} \max_{k\in\mathcal{I}} \int_{\mathcal{X}} \frac{R_{\gamma,J,k}(y)}{\tilde{\pi}_\gamma(y,k)^{s_J}} \tilde{\pi}_\gamma(y,k)^{s_J - s} dy \\
&\leq \sup_{\gamma\in\mathcal{Y}} \max_{k\in\mathcal{I}} c_2 \tilde{K}^{s_J} \int_{\mathcal{X}} \tilde{\pi}_\gamma(y,k)^{s_J - s} dy \\
&\leq c_2 \tilde{K}^{s_J} \int_{\mathcal{X}} \pi(y)^{s_J - s} dy < \infty,
\end{aligned} \tag{2.73}$$

where the last inequality follows from π being super-exponential and $s_J - s$ positive.

Recall additionally that under e2) (2.71) holds for all $(x,i) \in \mathcal{X} \times \mathcal{I}$. Putting together (2.70), (2.72), (2.73) and (2.71) yields

$$\begin{aligned}
\tilde{P}_\gamma V_{\tilde{\pi}_\gamma}((x,i)) &\leq (1-\epsilon)\lambda_L V_{\tilde{\pi}_\gamma}((x,i)) + (1-\epsilon) b_L + \epsilon \sum_{k\in\mathcal{I}} a_{\gamma,ik} b_J + \epsilon V_{\tilde{\pi}_\gamma}((x,i)) \\
&= \left((1-\epsilon)\lambda_L + \epsilon\right) V_{\tilde{\pi}_\gamma}((x,i)) + (1-\epsilon) b_L + \epsilon b_J.
\end{aligned}$$

By setting $\lambda := (1-\epsilon)\lambda_L + \epsilon$ and $b := (1-\epsilon)b_L + \epsilon b_J$, we obtain the drift condition as given by (2.11). □

Assumption c) of Lemma 2.5

Proof. Proving the minorization condition (2.12) amounts to specifying n_0, δ, ν_γ and v, and verifying that $\tilde{P}_\gamma^{n_0}(\tilde{x}, B) \geq \delta\nu_\gamma(B)$ for all measurable sets B and all $\tilde{x}$ satisfying $V_{\tilde{\pi}_\gamma}(\tilde{x}) \leq v$. Let C_v be defined as $C_v := \{x \in \mathcal{X} : c\pi(x)^{-s} \leq v\}$, where c is defined in (2.59). We specify the value of v below, separately for assumptions e1) and e2).

Note that if $x \in C_v$, then $V_{\tilde{\pi}_\gamma}((x,i)) \leq v$ for each $i \in \mathcal{I}$ and each $\gamma \in \mathcal{Y}$. Observe also that C_v is a compact set. Let ν_γ be the uniform distribution on $C_v \times \mathcal{I}$ (and 0 everywhere else) i.e. for $A \subseteq C_v$ we have $\nu_\gamma(A \times \{i\}) = \frac{1}{N}\frac{\mu^{\mathrm{Leb}}(A)}{\mu^{\mathrm{Leb}}(C_v)}$. To prove the claim, it is enough to show that

$$\tilde{P}_\gamma^{n_0}(\tilde{x}, \hat{B}) \geq \delta\nu_\gamma\left(\hat{B}\right)$$

for $\hat{B}$ of the form $B \times \{k\}$, for any $B \subseteq C_v$ and any $k \in \mathcal{I}$.

Firstly, note that for any $i, k \in \mathcal{I}$

$$\begin{aligned} &\inf_{x,y \in C_v} \inf_{\gamma \in \mathcal{Y}} \min\left[1, \frac{\tilde{\pi}_\gamma(y,k)}{\tilde{\pi}_\gamma(x,i)}\right] \\ \geq &\inf_{x,y \in C_v} \inf_{\gamma \in \mathcal{Y}} \min\left[1, \frac{\pi(y)}{\pi(x)} \frac{w_{\gamma,k} Q_k(\mu_k, \Sigma_{\gamma,k})(y)}{\sum_{j \in \mathcal{I}} w_{\gamma,j} Q_j(\mu_j, \Sigma_{\gamma,j})(y)}\right] \geq 0, \end{aligned} \tag{2.74}$$

where the last inequality is satisfied by assumption c) and equation (2.66).

We will first focus on verifying the minorization condition under assumption e1). Recall that D is a compact set in $\mathcal{X}$ such that for each $\gamma \in \mathcal{Y}$ and $i \in \mathcal{I}$ we have $JR_{\gamma,i} \subseteq D$. Recall also that by the construction of the jumping regions there exists r_1 such that for each $\gamma \in \mathcal{Y}$ and $i \in \mathcal{I}$ the ball $B(\mu_i, r_1) \subseteq JR_{\gamma,i}$. Let us now pick v so large that $D \subseteq C_v$ and $v > 2n_0 b/(1 - \lambda^{n_0})$ for $n_0 = 3$.

The minorization condition will be proved for $n_0 = 3$. Indeed, three steps of the algorithm are enough to get from a point (x,i) to a set $B \times \{k\}$ (a local step within mode i to reach its jumping region, a jump to mode k and a local move within mode k to set B).

Fix $(x,i) \in C_v \times \mathcal{I}$ and a set $\hat{B} = B \times \{k\}$ for $B \subseteq C_v$ (we allow for the case $k = i$). Note

that since $JR_{\gamma,i} \subset C_v$ for all $i \in \mathcal{I}$ and $\gamma \in \mathcal{Y}$ we have

$$\begin{aligned}
&\inf_{x \in C_v} \inf_{\gamma \in \mathcal{Y}} \tilde{P}_{\gamma,L,i}\left((x,i), JR_{\gamma,i} \times \{i\}\right) \\
&\geq \inf_{x \in C_v} \inf_{\gamma \in \mathcal{Y}} \int_{JR_{\gamma,i}} R_{\gamma,L,i}(x,y) \min\left[1, \frac{\tilde{\pi}_\gamma(y,i)}{\tilde{\pi}_\gamma(x,i)}\right] dy \\
&\geq \inf_{x,y \in C_v} \inf_{\gamma \in \mathcal{Y}} R_{\gamma,L,i}(x,y) \inf_{x,y \in C_v} \inf_{\gamma \in \mathcal{Y}} \min\left[1, \frac{\tilde{\pi}_\gamma(y,i)}{\tilde{\pi}_\gamma(x,i)}\right] \mu^{\mathrm{Leb}}(JR_{\gamma,i}) \\
&\geq \inf_{x,y \in C_v} \inf_{\gamma \in \mathcal{Y}} R_{\gamma,L,i}(x,y) \inf_{x,y \in C_v} \inf_{\gamma \in \mathcal{Y}} \min\left[1, \frac{\tilde{\pi}_\gamma(y,i)}{\tilde{\pi}_\gamma(x,i)}\right] \mu^{\mathrm{Leb}}\left(B(\mu_i, r_1)\right) \\
&=: p_{1,i}.
\end{aligned} \tag{2.75}$$

By equations (2.18) and (2.74) we get that $p_{1,i}$ defined above is strictly positive for $i \in \mathcal{I}$. Considering the probability of accepting a deterministic jump from mode i to mode k, we obtain

$$\begin{aligned}
&\inf_{x \in JR_{\gamma,i}} \inf_{\gamma \in \mathcal{Y}} \tilde{P}_{\gamma,J,k}\left((x,k), B \times \{k\}\right) \\
&\geq \inf_{x \in JR_{\gamma,i}} \inf_{y \in JR_{\gamma,k}} \inf_{\gamma \in \mathcal{Y}} \min\left[1, \frac{\tilde{\pi}(y,k)}{\tilde{\pi}(x,i)} \frac{a_{\gamma,ki}\sqrt{\det \Sigma_{\gamma,k}}}{a_{\gamma,ik}\sqrt{\det \Sigma_{\gamma,i}}}\right] \\
&\geq \inf_{x,y \in C_v} \inf_{\gamma \in \mathcal{Y}} \min\left[1, \frac{\tilde{\pi}(y,k)}{\tilde{\pi}(x,i)} \frac{a_{\gamma,ki}\sqrt{\det \Sigma_{\gamma,k}}}{a_{\gamma,ik}\sqrt{\det \Sigma_{\gamma,i}}}\right] \\
&=: p_{2,ik}.
\end{aligned} \tag{2.76}$$

It follows from equations (2.74), (2.16) and (2.15) that $p_{2,ik} > 0$ for $i, k \in \mathcal{I}$. Analogous arguments show that

$$\begin{aligned}
&\inf_{x \in JR_{\gamma,k}} \inf_{\gamma \in \mathcal{Y}} \tilde{P}_{\gamma,L,k}\left((x,k), B \times \{k\}\right) \\
&\geq \inf_{x \in C_v} \inf_{\gamma \in \mathcal{Y}} \int_B R_{\gamma,L,k}(x,y) \min\left[1, \frac{\tilde{\pi}_\gamma(y,k)}{\tilde{\pi}_\gamma(x,k)}\right] dy \\
&\geq \underbrace{\inf_{x,y \in C_v} \inf_{\gamma \in \mathcal{Y}} R_{\gamma,L,k}(x,y) \inf_{x,y \in C_v} \inf_{\gamma \in \mathcal{Y}} \min\left[1, \frac{\tilde{\pi}_\gamma(y,k)}{\tilde{\pi}_\gamma(x,k)}\right]}_{=:p_{3,k}} \mu^{\mathrm{Leb}}(B).
\end{aligned} \tag{2.77}$$

Combining (2.75), (2.76) and (2.77) yields

$$\tilde{P}^3_\gamma((x,i), B \times \{k\}) \geq (1-\epsilon)^2 \epsilon \epsilon_a p_{1,i} p_{2,ik} p_{3,k} \mu^{\mathrm{Leb}}(B).$$

Setting $\delta := (1-\epsilon)^2 \epsilon \epsilon_a \min_{i,k \in \mathcal{I}} p_{1,i} p_{2,ik} p_{3,k} N \mu^{\mathrm{Leb}}(C_v)$ completes the proof.

We will now verify the minorization condition under assumption e2). Let v be so large that $B(\mu_i, r) \subseteq C_v$ (see assumption (2.22)) for $i \in \mathcal{I}$ and $v > 2n_0 b/(1-\lambda^{n_0})$, for $n_0 = 3$. We will prove that indeed the minorization condition holds for $n_0 = 3$. Note that if we want to move from (x, i) to a set $B \times \{k\}$, it is enough to make a local step to $B(\mu_i, r)$, and then a jump to $B(\mu_k, r)$ followed by a local step to B.

As before, fix $(x, i) \in C_v \times \mathcal{I}$ and a set $\hat{B} = B \times \{k\}$ for $B \subseteq C_v$. Again, we include the case $k = i$. Analogous calculations to (2.75) show that

$$\begin{aligned} &\inf_{x \in C_v} \inf_{\gamma \in \mathcal{Y}} \tilde{P}_{\gamma, L, i}\left((x, i), B(\mu_i, r) \times \{i\}\right) \\ &\geq \inf_{x, y \in C_v} \inf_{\gamma \in \mathcal{Y}} R_{\gamma, L, i}(x, y) \inf_{x, y \in C_v} \inf_{\gamma \in \mathcal{Y}} \min\left[1, \frac{\tilde{\pi}_\gamma(y, i)}{\tilde{\pi}_\gamma(x, i)}\right] \mu^{\mathrm{Leb}}\left(B(\mu_i, r)\right) \\ &=: p_{4,i} > 0. \end{aligned} \tag{2.78}$$

For the jump kernel involved we obtain

$$\begin{aligned} &\inf_{x \in B(\mu_i, r)} \inf_{\gamma \in \mathcal{Y}} \tilde{P}_{\gamma, J, k}\left((x, i), B(\mu_k, r) \times \{k\}\right) \\ &\geq \inf_{x \in B(\mu_i, r)} \inf_{\gamma \in \mathcal{Y}} \int_{B(\mu_k, r)} R_{\gamma, J, k}(y) \min\left[1, \frac{\tilde{\pi}_\gamma(y, k)}{\tilde{\pi}_\gamma(x, i)} \frac{a_{\gamma, ki} R_{\gamma, J, i}(x)}{a_{\gamma, ik} R_{\gamma, J, k}(y)}\right] dy \\ &\geq \inf_{x \in B(\mu_i, r)} \inf_{y \in B(\mu_k, r)} \inf_{\gamma \in \mathcal{Y}} R_{\gamma, J, k}(y) \min\left[1, \frac{\tilde{\pi}_\gamma(y, k)}{\tilde{\pi}_\gamma(x, i)} \frac{a_{\gamma, ki} R_{\gamma, J, i}(x)}{a_{\gamma, ik} R_{\gamma, J, k}(y)}\right] \mu^{\mathrm{Leb}}\left(B(\mu_k, r)\right) \\ &=: p_{5, ik}. \end{aligned} \tag{2.79}$$

Note that $p_{5,ik}$ is positive by equations (2.74) and (2.16), and assumption e2). Finally, similar calculations to (2.77) yield

$$\begin{aligned} &\inf_{x \in B(\mu_k, r)} \inf_{\gamma \in \mathcal{Y}} \tilde{P}_{\gamma, L, k}\left((x, k), B \times \{k\}\right) \\ &\geq \inf_{x \in C_v} \inf_{\gamma \in \mathcal{Y}} \int_B R_{\gamma, L, k}(x, y) \min\left[1, \frac{\tilde{\pi}_\gamma(y, k)}{\tilde{\pi}_\gamma(x, k)}\right] dy \\ &\geq \underbrace{\inf_{x, y \in C_v} \inf_{\gamma \in \mathcal{Y}} R_{\gamma, L, k}(x, y) \inf_{x, y \in C_v} \inf_{\gamma \in \mathcal{Y}} \min\left[1, \frac{\tilde{\pi}_\gamma(y, k)}{\tilde{\pi}_\gamma(x, k)}\right]}_{= p_{3,k}} \mu^{\mathrm{Leb}}(B). \end{aligned} \tag{2.80}$$

for $p_{3,k}$ defined in the previous part of the proof. We now combine (2.78), (2.79) and (2.80) to get

$$\tilde{P}_\gamma^3((x, i), B \times \{k\}) \geq (1-\epsilon)^2 \epsilon \epsilon_a p_{4,i} p_{5,ik} p_{3,k} \mu^{\mathrm{Leb}}(B).$$

Setting $\delta := (1-\epsilon)^2\epsilon\epsilon_a \min_{i,k\in\mathcal{I}} p_{4,i}p_{5,ik}p_{3,k}N\mu^{\mathrm{Leb}}(C_v)$ completes the proof. □

Proof of Theorem 2.11

Proof. This theorem is a direct corollary from Theorem 2.4. The assumptions of this theorem were verified in the proofs of Theorems 2.6 or 2.7, under the uniform and the non-uniform scenario, respectively. □

Proof of Lemma 2.8

Proof. Fix $i \in \mathcal{I}$ and let ϵ_L be such that for $|x|$ larger than some R_0

$$\int_{a(x)} R_{\gamma^*,L,i}(x,y) \geq \epsilon_L,$$

(such ϵ_L can be found due to assumption (2.24)). Hence, for K sufficiently large

$$\int_{a(x)\cap B(x,K)} R_{\gamma^*,L,i}(x,y) \geq \frac{\epsilon_L}{2},$$

which implies that for any $|x| > R_0$

$$\begin{aligned}\mu^{\mathrm{Leb}}\left(a(x)\cap B(x,K)\right) &\geq \frac{\epsilon_L}{2\sup_{y\in B(x,K)} R_{\gamma^*,L,i}(x,y)} \\ &= \frac{\epsilon_L}{2\sup_{y\in B(\mathbf{0},K)} R_{\gamma^*,L,i}(\mathbf{0},y)}\end{aligned}$$

and consequently

$$\begin{aligned}&\inf_{\gamma\in\mathcal{Y}} \int_{a(x)\cap B(x,K)} R_{\gamma,L,i}(x,y)dy \\ &\geq \mu^{\mathrm{Leb}}\left(a(x)\cap B(x,K)\right) \inf_{\gamma\in\mathcal{Y}} \inf_{y\in B(x,K)} R_{\gamma,L,i}(x,y) \\ &\geq \frac{\epsilon_L \inf_{\gamma\in\mathcal{Y}} \inf_{y\in B(\mathbf{0},K)} R_{\gamma,L,i}(\mathbf{0},y)}{2\sup_{y\in B(\mathbf{0},K)} R_{\gamma^*,L,i}(\mathbf{0},y)} =: \tilde{\epsilon}_L > 0.\end{aligned} \tag{2.81}$$

By assumption a) of Theorem 2.7 $\tilde{\epsilon}_L$ is indeed positive.

Let the acceptance region $a_{\gamma,i}(x)$ be given by (2.69). We will show that for $|x|$ sufficiently large and for each $\gamma \in \mathcal{Y}$

$$\int_{a_{\gamma,i}(x)} R_{\gamma,L,i}(x,y)dy \geq \frac{\tilde{\epsilon}_L}{2},$$

which will prove the claim. We shall now repeat similar arguments to those used in the proof

of Theorem 2.7, in the part for the local kernels. Firstly, we use formula (2.64) to conclude that for $|x|$ larger than some R_1 (which may depend on K) and for sufficiently small δ (which may depend on K, R_1 and $\tilde{\epsilon}_L$), we have

$$\sup_{\gamma\in\mathcal{Y}} \int_{C_{\pi(x)}(\delta)\cap B(x,K)} R_{\gamma,L,i}(x,y)dy \leq \frac{\tilde{\epsilon}_L}{2}. \tag{2.82}$$

We put (2.81) together with (2.82) to obtain

$$\inf_{\gamma\in\mathcal{Y}} \int_{(a(x)\cap B(x,K))\setminus C_{\pi(x)}(\delta)} R_{\gamma,L,i}\left(x,y\right)dy \geq \frac{\tilde{\epsilon}_L}{2}$$

for $|x| > \max\left[R_0, R_1\right]$. Now recall that for each δ there exists R_2 such that for $|x| > R_2$ if $y \in (a(x)\cap B(x,K))\setminus C_{\pi(x)}(\delta)$ then $\frac{\pi(y)}{\pi(x)} \geq \tilde{K}$ for $\tilde{K}$ defined in (2.66). Therefore in particular $y \in a_{\gamma,i}(x)$ for each $\gamma \in \mathcal{Y}$. Finally, for $|x| > \max\left[R_0, R_1, R_2\right]$ we have

$$\inf_{\gamma\in\mathcal{Y}} \int_{a_{\gamma,i}(x)} R_{\gamma,L,i}\left(x,y\right)dy \geq \inf_{\gamma\in\mathcal{Y}} \int_{(a(x)\cap B(x,K))\setminus C_{\pi(x)}(\delta)} R_{\gamma,L,i}\left(x,y\right)dy \geq \frac{\tilde{\epsilon}_L}{2},$$

which yields the claim. □

Proof of Lemma 2.9

Proof. Fix any $\gamma \in \mathcal{Y}$ and $i \in \mathcal{I}$. To prove the required result, we will use analogous arguments to those from the proof of Theorem 4.3 of Jarner and Hansen (2000). Let $\epsilon > 0$ and R be such that for $|x| > R$

$$\frac{x}{|x|}\cdot\frac{\nabla\pi(x)}{|\nabla\pi(x)|} \leq -\epsilon.$$

Fix $K > 0$ and define the cone $W(x)$ as

$$W(x) := \left\{x - a\xi : 0 < a < K, \xi \in S^{d-1}, \left|\xi - \frac{x}{|x|}\right| \leq \frac{\epsilon}{2}\right\}.$$

We now refer to the proof of Theorem 4.3 of Jarner and Hansen (2000) to see that for x sufficiently large $W(x) \subset a(x)$. What is more,

$$\liminf_{|x|\to\infty} \int_{W(x)} R_{\gamma,L,i}(x,y)dy > 0 \quad \text{and so} \quad \liminf_{|x|\to\infty} \int_{a(x)} R_{\gamma,L,i}(x,y)dy > 0.$$

Hence, since i was chosen arbitrarily, assumption (2.24) is satisfied for $\gamma^* := \gamma$.

We would like to point out here that originally Theorem 4.3 of Jarner and Hansen (2000) was proved under a stronger assumption, that is, $R_{\gamma,L,i}(x,y) = R_{\gamma,L,i}(|x-y|)$. However, careful inspection of this proof shows that it is enough to assume that $R_{\gamma,L,i}(x,y) = R_{\gamma,L,i}(y,x)$, which is satisfied in our case as $R_{\gamma,L,i}$ follows an elliptical distribution. □

Proof of Corollary 2.10

Proof. We will again refer multiple times to Jarner and Hansen (2000). Firstly, by Theorem 4.4 of that paper, if π_1 and π_2 are super-exponential and satisfy (2.25), then also $a_1\pi_1 + a_2\pi_2$ is super-exponential and satisfies (2.25) for positive a_1 and a_2. By Theorem 4.6 of the same paper, each density of the form $\pi(x) \propto \exp(-p(x))$ is super-exponential and satisfies (2.25), if p is a polynomial of order ≥ 2. Therefore, the assumptions of Lemma 2.9 hold, as required. □

A .3 Other algorithms in the Auxiliary Variable Adaptive MCMC class

As mentioned in Section 3 , an instance of an algorithm in the Auxiliary Variable Adaptive MCMC class is adaptive parallel tempering introduced by Miasojedow et al. (2013). Indeed, let us consider $\Phi := \mathcal{X}^N$ and $\tilde{X} := \mathcal{X} \times \Phi = \mathcal{X} \times \mathcal{X}^N$ and

$$\tilde{\pi}_\gamma\left(x_N, (x_0, \ldots, x_{N-1})\right) := \prod_{i=0}^{N} \pi(x_i)^{\beta_{j,\gamma}}.$$

Then for any $B \in \mathcal{B}(\mathcal{X})$ we have

$$\begin{aligned}\tilde{\pi}_\gamma\left(B \times \Phi\right) &= \int_B \pi(x_N)^{\beta_{N,\gamma}} dx_N \int_\Phi \prod_{i=0}^{N-1} \pi(x_i)^{\beta_{j,\gamma}} dx_0 \ldots dx_{N-1} \\ &= \int_B \pi(x_N)^{\beta_{N,\gamma}} dx_N = \pi(B),\end{aligned}$$

where the last equality follows since $\beta_{N,\gamma} = 1$ for all γ. Additionally, the transition kernels used in adaptive parallel tempering $\{\tilde{P}_\gamma\}_{\gamma\in\mathcal{Y}}$ are defined in such a way that detailed balance holds.

Another example of a group of algorithms in the Auxiliary Variable Adaptive MCMC class is an adaptive version of pseudomarginal algorithms. Recall that pseudomarginal algorithms are a powerful tool used in situations when the target density $\pi(x)$ on $\mathcal{X}$ cannot be evaluated pointwise or this evaluation would be very expensive, but an unbiased estimator

of $\pi(x)$ is available. In the simplest setting an importance sampling estimator is used for this purpose. Then the pseudomarginal algorithm is equivalent to the Metropolis-Hastings algorithm targeting a distribution $\tilde{\pi}^N(x,z)$ on an augmented state space $\mathcal{X}\times\mathcal{Z}$, where $Z\in\mathcal{Z}$ is a vector representing N samples on which the importance sampling estimator is based. A remarkable property of the pseudomarginal algorithms is that $\pi(x)$ is the marginal distribution of $\tilde{\pi}^N(x,z)$ regardless of N (see Andrieu and Roberts (2009) and Andrieu and Vihola (2015)). The number of samples N and, in more complex settings, the amount of correlation between those samples (see Deligiannidis et al. (2018)), may follow an adaptive scheme. Therefore, condition (2.6) are satisfied for $\Phi=\mathcal{Z}^{\mathbb{N}}\times\mathbb{N}$, where $N\in\mathbb{N}$ corresponds to the number of samples used for estimation.

A .4 Spectral gap of a single JAMS kernel

The performance of a single JAMS sampling kernel heavily depends on how closely the covariance matrices used in proposal distributions for jumps approximate the actual covariance structures in the mode, as this determines the acceptance rate of movements between the modes. As mentioned in Section 4 , we study the following target density:

$$\pi(x)=\frac{1}{2}N\left(-\underbrace{(1,\ldots,1)}_{d},\sigma_1^2 I_d\right)+\frac{1}{2}N\left(\underbrace{(1,\ldots,1)}_{d},\sigma_2^2 I_d\right).$$

We consider a scenario in which all the modes of the target distribution have been identified and their corresponding covariance matrices have been approximated in the burn-in algorithm. The result below shows that given the locations of the modes and a certain rate of approximation of the proposal covariance matrices to the true ones, a single JAMS kernel with independent proposal jumps mixes rapidly on this target density.

Recall that the proposal matrix Σ_i is equal to $S_i+\beta I_d$, where S_i is an empirical covariance matrix based on samples collected in the burn-in algorithm and d denotes the dimension of the state space. We drop the subscript γ in the notation to indicate that no adaptation is involved.

Lemma 2.19. *Consider N densities $\pi_1,\ldots,\pi_N$ and define a target density*

$$\pi(x)=\sum_{i=1}^{N}p_i\pi_i(x)$$

positive on $\mathcal{X}$ *for* $\sum_{i=1}^{N} p_i = 1$*, with modes at* $\mu_1, \ldots, \mu_N$*. Let* $\tilde{P}_h$ *be a JAMS kernel targeting*

$$\tilde{\pi}(x, i) := \pi(x) \frac{w_{\gamma,i} Q_i(\mu_i, \Sigma_i)(x)}{\sum_{j \in \mathcal{I}} w_{\gamma,j} Q_j(\mu_j, \Sigma_j)(x)}$$

with independent proposal jumps such that $R_{J,i} = Q_i(\mu_i, \Sigma_i)$ *is positive on* $\mathcal{X}$ *for* $i \in \mathcal{I}$*, and with a holding probability 0.5. Then if*

$$\min_{i \in \{1,\ldots,N\}} \inf_{x \in \mathcal{X}} \frac{Q_i(\mu_i, \Sigma_i)(x)}{\pi_i(x)}$$

decreases polynomially in d*, so does the spectral gap of* $\tilde{P}_h$*.*

In particular, let π *be given by* (2.26)*. Let* $R_{J,i}$ *follows the normal distribution* $N(\mu_i, \Sigma_i)$ *or the* t *distribution* $t_\nu(\mu_i, \Sigma_i)$*. Assume additionally that:*

a) Let $\lambda_{i_1}, \ldots, \lambda_{i,d}$ *be the eigenvalues of* S_i*. The eigenvalues satisfy*

$$\sigma_i^2 \left(1 - \frac{c_1 \log d}{d}\right) < \lambda_{ij} < \sigma_i^2 \left(1 + \frac{c_1 \log d}{d}\right) \text{ for } i \in \{1, 2\}, j \in \{1, \ldots, d\}.$$

b) The parameter β *used in the algorithm satisfies*

$$c_1 \frac{\log d}{d} \max_{i \in \{1,2\}} \sigma_i^2 < \beta < c_2 \frac{\log d}{d} \max_{i \in \{1,2\}} \sigma_i^2, \quad \text{for some } c_1, c_2 > 0.$$

Then the spectral gap of $\tilde{P}_h$ *decreases polynomially in dimension.*

Proof. The proof will rely mainly on a result from Gåsemyr (2006) about the spectrum of an independence sampler.

Firstly, $\tilde{P}_h$ is defined as $\tilde{P}_h := \epsilon \frac{1}{2}\left(\tilde{P}_J + I\right) + (1 - \epsilon)\frac{1}{2}\left(\tilde{P}_L + I\right)$, where $\tilde{P}_J$ and $\tilde{P}_L$ are the jump and local kernels, respectively. Each of the kernels $\tilde{P}_J$, $\frac{1}{2}\left(\tilde{P}_J + I\right)$ and $\frac{1}{2}\left(\tilde{P}_L + I\right)$ is

non-negative, so the spectral gap of $\tilde{P}_h$ satisfies

$$\begin{aligned}
\mathbf{Gap}\left(\tilde{P}_h\right) &= \inf_{Var_{\tilde{\pi}}(f)=1} \int f(x) \int f(y) \left(I - \tilde{P}_h\right)(x, dy) \\
&\geq \epsilon \inf_{Var_{\tilde{\pi}}(f)=1} \int f(x) \int f(y) \left(I - \frac{1}{2}\left(\tilde{P}_J + I\right)\right)(x, dy) \\
&\quad + (1-\epsilon) \inf_{Var_{\tilde{\pi}}(f)=1} \int f(x) \int f(y) \left(I - \frac{1}{2}\left(\tilde{P}_L + I\right)\right)(x, dy) \\
&\geq \epsilon \mathbf{Gap}\left(\frac{1}{2}\left(\tilde{P}_J + I\right)\right) \\
&= \frac{1}{2}\epsilon \mathbf{Gap}\left(\tilde{P}_J\right).
\end{aligned}$$

Therefore, it is enough to show that $\mathbf{Gap}\left(\tilde{P}_J\right)$ decreases polynomially with dimension. Note that $\tilde{P}_J$ is in fact an independence sampler kernel on $\mathcal{X} \times \mathcal{I}$ targeting $\tilde{\pi}$, with the proposal distribution $R_{J,i}$ used with probability a_i for $i \in \mathcal{I}$. We can now use a result from Gåsemyr (2006) and an earlier result from Liu (1996). The spectral gap of $\tilde{P}_J$ is therefore equal to $1/w^*$, where

$$w^* = \operatorname*{ess\,sup}_{x \in \mathcal{X}, i \in \mathcal{I}} w_i(x) \quad \text{and} \quad w_i(x) := \frac{\tilde{\pi}\left((x,i)\right)}{a_i R_{J,i}(x)},$$

where the essential supremum is considered with respect to $\tilde{\pi}$. Therefore

$$\begin{aligned}
\mathbf{Gap}\left(\tilde{P}_J\right) &= 1/w^* \\
&= \operatorname*{ess\,inf}_{x \in \mathcal{X}, i \in \mathcal{I}} \frac{a_i R_{J,i}(x)}{\tilde{\pi}\left((x,i)\right)} \\
&= \operatorname*{ess\,inf}_{x \in \mathcal{X}, i \in \mathcal{I}} a_i R_{J,i}(x) \frac{w_1 Q_1(\mu_1, \Sigma_1)(x) + \ldots w_N Q_N(\mu_N, \Sigma_N)(x)}{w_i Q_i(\mu_i, \Sigma_i)(x) \pi(x)} \\
&= \operatorname*{ess\,inf}_{x \in \mathcal{X}, i \in \mathcal{I}} \frac{a_i}{w_i} \frac{w_1 Q_1(\mu_1, \Sigma_1)(x) + \ldots w_N Q_N(\mu_N, \Sigma_N)(x)}{p_1 \pi_1(x) + \ldots + p_N \pi_N(x)} \\
&\geq \min_{i,k \in \mathcal{I}} \inf_{x \in \mathcal{X}} \frac{a_i}{w_i} \frac{w_k Q_k(\mu_k, \Sigma_k)(x)}{p_k \pi_k(x)},
\end{aligned}$$

which completes the proof of the first part of the result.

To obtain the second part, by $\tilde{\lambda}_{i_1}, \ldots, \tilde{\lambda}_{i,d}$ denote the eigenvalues of Σ_i. Conditions a) and b) imply that

$$\sigma_i^2 < \tilde{\lambda}_{i,j} < \sigma_i^2 \left(1 + \frac{c_3 \log d}{d}\right) \tag{2.83}$$

for some $c_3 > 0$ and $j = 1, \ldots, d$. Without loss of generality we will assume that Σ_i is diagonal. We will first consider the case when the proposal distributions are Gaussian to

then proceed to a scenario with t-distributed proposals. Note that

$$\frac{Q_i(\mu_i,\Sigma_i)(x)}{\pi_i(x)} = \frac{\sigma_i^d}{\sqrt{\prod_{j=1}^d \tilde{\lambda}_{ij}}} \exp\left(\sum_{j=1}^d \left(\sigma_i^{-2} - \tilde{\lambda}_{ij}^{-1}\right)(x_j-\mu_{ij})^2/2\right), \tag{2.84}$$

where $\mu_i = (\mu_{i1},\ldots,\mu_{id}), x = (x_1,\ldots,x_d)$. By (2.83) and (2.84)

$$\inf_{x\in\mathcal{X}} \frac{Q_i(\mu_i,\Sigma_i)(x)}{\pi_i(x)} = \frac{\sigma_i^d}{\sqrt{\prod_{j=1}^d \tilde{\lambda}_{ij}}} > \left(1+\frac{c_3\log d}{d}\right)^{-d/2} \approx d^{-c_3/2}$$

and so the spectral gap decreases polynomially with dimension.

For the t-distributed proposals we have

$$\begin{aligned}&\frac{Q_i(\mu_i,\Sigma_i)(x)}{\pi_i(x)}\\ &= \frac{\sigma_i^d}{\sqrt{\prod_{j=1}^d \tilde{\lambda}_{ij}}} \frac{\Gamma\left(\nu/2+d/2\right)}{\Gamma\left(\nu/2\right)(\nu/2)^{d/2}} \frac{\left(1+\frac{1}{\nu}\sum_{j=1}^d \tilde{\lambda}_{ij}^{-1}(x_j-\mu_{ij})^2\right)^{-(\nu+d)/2}}{\exp\left(-\sum_{j=1}^d \sigma_i^{-2}(x_j-\mu_{ij})^2/2\right)}.\end{aligned}$$

Then for $\tilde{\lambda}_{ij}$ satisfying (2.83) we obtain

$$\begin{aligned}&\inf_{x\in\mathcal{X}} \frac{Q_i(\mu_i,\Sigma_i)(x)}{\pi_i(x)}\\ &\geq \frac{\sigma_i^d}{\sqrt{\prod_{j=1}^d \tilde{\lambda}_{ij}}} \frac{\Gamma\left(\nu/2+d/2\right)}{\Gamma\left(\nu/2\right)(\nu/2)^{d/2}} \inf_{x\in\mathcal{X}} \underbrace{\frac{\left(1+\frac{1}{\nu}\sum_{j=1}^d \tilde{\lambda}_{ij}^{-1}(x_j-\mu_{ij})^2\right)^{-(\nu+d)/2}}{\exp\left(-\sum_{j=1}^d \tilde{\lambda}_{ij}^{-1}(x_j-\mu_{ij})^2/2\right)}}_{:=A(x)}.\end{aligned}$$

Note that $A(x)$ is minimized when $\sum_{j=1}^d \tilde{\lambda}_{ij}^{-1}(x_j-\mu_{ij})^2 = d$. Using Stirling's formula to approximate $\Gamma\left(\nu/2+d/2\right)$ we get that

$$\frac{\Gamma\left(\nu/2+d/2\right)}{\Gamma\left(\nu/2\right)(\nu/2)^{d/2}} \frac{\left(1+\frac{d}{\nu}\right)^{-(\nu+d)/2}}{\exp\left(-d/2\right)} \approx \frac{c_\nu}{\sqrt{d}},$$

asymptotically for large d, where c_ν is a constant depending only on the number of degrees of freedom ν. Hence

$$\inf_{x\in\mathcal{X}} \frac{Q_i(\mu_i,\Sigma_i)(x)}{\pi_i(x)} \geq d^{-(c_3+1)/2}$$

for large d. □

A question of interest now is: how many iterations of the covariance matrix estimation are

needed to obtain the desired rate of approximation? If this number depends polynomially on the dimension, then given the locations of the modes, the whole algorithm based on covariance matrix estimation in the burn-in algorithm followed by a non-adaptive version of JAMS is polynomial-time with respect to the dimension of the state space. Below we give informal arguments why this should be the case.

If one could generate i.i.d. samples from the normal distribution, the number of samples needed to approximate the eigenvalues of the covariance matrix with the precision required in condition a) of Lemma 2.19, depends polynomially on the dimension (see Remark 4.7.2 of Vershynin (2018)). In the case of using samples generated by a Markov chain targeting a normal distribution, the required number of iterations additionally depends on the rate of convergence of the chain. From the Central Limit Theorem for Markov chains we know that

$$\sigma^2_{P,f} \leq \left(\frac{2}{\mathbf{Gap}(P)}\right)\sigma^2_f,$$

where σ^2_f is the asymptotic variance for a function f in the i.i.d. case and $\sigma^2_{P,f}$ is the asymptotic variance for a Markov chain using a transition kernel P (see Kipnis and Varadhan (1986) and Section 4.2 of Hairer et al. (2014)). Results from Kannan and Li (1996) show that under certain assumptions the spectral gap of a Random Walk Metropolis algorithm targeting the normal distribution decreases polynomially with dimension. This would indicate the overall polynomial dependence of the required number of iterations on the dimension of the state space.

Now recall that in the burn-in algorithm at the covariance estimation stage the target distribution of chain i is $\tilde{\pi}_\gamma(x, i)$ (for a fixed i). If π is given by (2.26), then $\tilde{\pi}_\gamma(x, i)$ is not exactly Gaussian, but if $Q_i(\mu_i, \Sigma_{\gamma,i})$, $i = 1, 2$, follow the normal distribution approximating well their corresponding elements of the mixture, then $\tilde{\pi}_\gamma(x, i)$ approximates well $Q_i(\mu_i, \Sigma_{\gamma,i})(x)$.

In practice our burn-in algorithm is adaptive, and both the transition kernels and the target distributions are updated while the algorithm runs. Hence the above reasoning does not provide formal results, but it is just an intuitive explanation why the dependence on the dimension should be polynomial. One can expect that adapting improves mixing of the chain, so it is likely to improve the rate of convergence. However, obtaining formal results about the performance of the adaptive algorithm (both at the burn-in stage and in the main algorithm) as a function of dimension remains an open question.

B Supplementary Material: methodology and numerical examples

Supplementary Material B consists of Sections B .1, B .2 and B .3. In Sections B .1 and B .2 we present details of the implementation of our method. An additional simulation example and settings of our numerical experiments are shown in Section B .3.

B .1 Updating $w_{\gamma,i}$ and $a_{\gamma,ik}$ in the main algorithm

Recall that N denotes the number of modes. The weights $w_{\gamma,i}$ are set to $1/N$ at the beginning of the algorithm and they are adapted while the algorithm runs in such a way that they represent the proportion of samples observed so far in each mode. At the same time we do not allow any of the weights to get below some pre-specified value ϵ_w; otherwise the target distribution $\tilde{\pi}_\gamma$ could run the risk of being severely distorted by weights very close or equal to 0. In particular we use the update scheme described below.

Let $n_{i,\text{obs}}$ be the number of samples in mode i for $i = 1, \ldots, N$ observed after n iterations of the main algorithm. Then $n = \sum_{i=1}^{N} n_{i,\text{obs}}$. Define

$$w_{\text{add}} := \frac{n}{\frac{1}{\epsilon_w} - N} \quad \text{and set} \quad w_i := \frac{n_{i,\text{obs}} + w_{\text{add}}}{n + N w_{\text{add}}} \quad \text{for } i = 1, \ldots, N.$$

It is easily checked that if there are no observations in mode i, i.e. $n_{i,\text{obs}}$ is equal to 0, then $w_i = \epsilon_w$. Since ϵ_w must satisfy $N\epsilon_w < 1$ and the number of modes N is typically unknown in advance, in our implementation the user provides $\tilde{\epsilon}_w$ and the algorithm sets $\epsilon_w := \tilde{\epsilon}_w / N$.

Even though the theory we present in Section 4 holds when $a_{\gamma,ik}$ follow some adaptive rule, we propose to keep these values fixed throughout the run of the algorithm, with a default choice $a_{\gamma,ii} = 0$ and $a_{\gamma,ik} = 1/(N-1)$ for $\gamma \in \mathcal{Y}$, $i, k \in \{1, \ldots, N\}$ and $i \neq k$. If N is not very large the benefit of adapting $a_{\gamma,ik}$ is rather marginal while it may add to finite sample instability. A natural alternative improving acceptance rates would be to keep $a_{\gamma,ii} = 0$ and $a_{\gamma,ik} = w_{\gamma,k} / \sum_{j \neq i} w_{\gamma,j}$. However, consider a scenario when a mode with a significant weight in the target distribution is particularly difficult to jump into (for example, because the covariance matrix estimation has not been run for long enough). The jumps to this mode will very likely get rejected many times before we observe the first sample in this mode and start adapting its covariance matrix. In such case proposing modes

proportionally to the number of samples collected so far in those modes would effectively make moves to this "difficult" mode even less frequent. Consequently we could face the risk of underestimating the weight of this mode for a fixed computational budget. Hence, we adopt the more conservative approach of keeping these values fixed to avoid the risk described above.

Note also that we in our implementation we use $a_{\gamma,ii} = 0$ even though formally we assumed in Section 4 that $a_{\gamma,ii} > \epsilon_a$. This is because in practice we do not want to propose jumps to the same mode. In the case of deterministic jumps this would mean proposing a move to the same state (recall equation (2.4)), which would have a negative impact on the mixing of the algorithm.

B .2 Burn-in algorithm

The pseudocode of the burn-in routine described in Section 2 .2 is presented in Algorithm 4.

Mode finding

For the mode finding part in our implementation we use the BFGS method from the `optimx` package in `R` (R Core Team, 2014). We only pass to the next stage of the burn-in algorithm (mode-merging) those vectors for which first and second order Kuhn, Karush, Tucker (KKT) optimality conditions are satisfied. Checking these conditions is necessary in order to avoid including points that are not local minima of $-\log(\pi)$ (but, for example, saddle points) in the list of modes. Besides, we recommend that the user codes up their own function for calculating the gradient and the Hessian, whenever possible, or uses packages that compute those values with high numerical precision. This will help ensure numerical stability of the optimization runs. What is more, working with variables with bounded support tends to be problematic – the optimization algorithm will typically struggle in the neighbourhood of the boundary. In such cases it is usually beneficial to work with transformed variables, defined on the whole space (see Section B .3).

Recall that for adding we allow newly discovered modes on the fly, while the main algorithm runs, and the ergodic theory presented in Section 4 holds. In practice the inference is based on a finite number of samples from the main algorithm, therefore when a new mode is added, the user may prefer to re-run the main algorithm to ensure that the weight of each mode is properly represented.

Algorithm 4 JAMS: burn-in algorithm

1: **Input:** end points of the intervals $L_1, U_1, \ldots, L_d, U_d$ (or the prior π_0), number of starting points n, small positive value q, threshold b_{acc}.
2: Generate $s_1, \ldots, s_n$ uniformly from $[L_1, U_1] \times \ldots \times [L_d, U_d]$ (or $s_i \sim \pi_0$ for $i = 1, \ldots, n$.)
3: Run BFGS from $s_1, \ldots, s_n$ to minimize $-\log(\pi(x))$. ▷ (in parallel)
4: Denote the optimum points by $m_1, \ldots, m_n$ and their corresponding Hessian matrices by $H_1, \ldots, H_n$.
5: Set $\mu_1 := m_1$, $H_{\mu_1} := H_1$, $N := 1$.
6: **for** $i = 2, \ldots, n$ **do**
7: **if** $\min_{j \in 1,\ldots,N} \frac{1}{2}\left((\mu_j - m_i)^T H_{\mu_j}(\mu_j - m_i) + (\mu_j - m_i)^T H_i(\mu_j - m_i)\right) < q$ **then**
8: $k := \arg\min_{j \in 1,\ldots,N} \frac{1}{2}\left((\mu_j - m_i)^T H_{\mu_j}(\mu_j - m_i) + (\mu_j - m_i)^T H_i(\mu_j - m_i)\right)$
9: **if** $\pi(\mu_k) < \pi(m_i)$ **then**
10: Set $\mu_k := m_i$ and $H_{\mu_k} := H_i$.
11: **end if**
12: **else**
13: Set $\mu_{N+1} := m_i$ and $H_{\mu_{N+1}} := H_i$.
14: Set $N := N + 1$.
15: **end if**
16: **end for**
17: Define $\Sigma_{0,i} := H_{\mu_i}^{-1}$ for $i = 1, \ldots, N$.
18: Set $k = 0$ and define $\tilde{\pi}(x, i) := \pi(x) \frac{\frac{1}{N} Q_i(\mu_i, \Sigma_{0,i})(x)}{\sum_{j=1}^{N} \frac{1}{N} Q_j(\mu_j, \Sigma_{0,j})(x)}$.
19: **while** inhomogeneity factors $b_{k,1}, \ldots, b_{k,N}$ satisfy: $\max_{i \in \{1,\ldots,N\}} b_{k,i} > b_{\text{acc}}$ **do**
20: k:= k+1
21: Run Algorithm 1 targeting $\tilde{\pi}$ with $\epsilon = 0$ from $\mu_1, \ldots \mu_N$. ▷ (in parallel)
22: Define $\Sigma_{k,i}$ for $i = 1, \ldots N$ as the matrix of mode μ_i updated in round k.
23: Update the target distribution $\tilde{\pi}$ by setting $\tilde{\pi}(x, i) := \pi(x) \frac{\frac{1}{N} Q_i(\mu_i, \Sigma_{k,i})(x)}{\sum_{j=1}^{N} \frac{1}{N} Q_j(\mu_j, \Sigma_{k,j})(x)}$.
24: **end while**
25: Set $K := k$ and $\Sigma_{\gamma_0,i} := \Sigma_{K,i}$, $\tilde{\Sigma}_i := \Sigma_{K,i}$, $S_i := \Sigma_{K,i}$ for $i = 1, \ldots, N$.
26: **return** List of $\{\mu_i, \Sigma_{\gamma_0,i}, \tilde{\Sigma}_i, S_i\}$ for $i = 1, \ldots, N$.
Target distribution $\tilde{\pi}_{\gamma_0}(x, i) := \pi(x) \frac{\frac{1}{N} Q_i(\mu_i, \Sigma_{\gamma_0,i})(x)}{\sum_{j=1}^{N} \frac{1}{N} Q_j(\mu_j, \Sigma_{\gamma_0,j})(x)}$ that will be used as input for Algorithm 1.

Covariance matrix estimation

Recall that the initial value of the matrix corresponding to mode i at the beginning of round 1 of the covariance matrix estimation is the inverse of the Hessian evaluated at μ_i (see line 17 of Algorithm 4). The heuristics behind this idea is that in the case of the Gaussian distribution the inverse of the Hessian of $-\log(\pi)$ would correspond to the covariance matrix, so intuitively for a large class of target distributions this will be a good starting value.

As mentioned in Section 2 .2, we propose a semi-automatic way of choosing the number of rounds of the covariance matrix estimation, denoted by K. The intuition for this procedure is to stop the burn-in algorithm when running an additional round does not yield much

improvement in the accuracy of the estimation of $\Sigma_1, \ldots, \Sigma_N$. Recall that $\Sigma_{k,i}$ is the matrix corresponding to mode i updated during round k. The choice of K is based on monitoring the following quantity, called inhomogeneity factor (see Roberts and Rosenthal (2009) and Rosenthal (2011)), given by

$$b_{k,i} := d \frac{\sum_{j=1}^{d} \lambda_j^{-1}}{\left(\sum_{j=1}^{d} \lambda_j^{-1/2}\right)^2}, \tag{2.85}$$

where d is the dimension of the state space of π and λ_j for $j = 1, \ldots, d$ are the eigenvalues of $\Sigma_{k-1,i}^{-1}\Sigma_{k,i}$. Note that this factor is always a real number even though $\Sigma_{k-1,i}^{-1}\Sigma_{k,i}$ does not need to be symmetric. If λ is a complex eigenvalue of $\Sigma_{k-1,i}^{-1}\Sigma_{k,i}$, its conjugate $\bar{\lambda}$ is also an eigenvalue of $\Sigma_{k-1,i}^{-1}\Sigma_{k,i}$ and so the imaginary components cancel both in the numerator and the denominator of (2.85). Moreover, by Jensen's inequality $b_{k,i}$ satisfies $b_{k,i} \geq 1$, and $b_{k,i} = 1$ if and only if $\Sigma_{k-1,i}$ and $\Sigma_{k,i}$ are proportional to each other. In particular, the value of $b_{k,i}$ is always equal to 1 in the scaling phase.

The procedure we propose is the following: perform AC_1 scaling steps for each matrix (perhaps split into several rounds). Then perform at least one covariance-based round for each mode. Continue running covariance-matrix rounds until the inhomogeneity factor drops below a certain threshold b_{acc} for all matrices. In other words, having performed AC_1 scaling steps and at least one covariance-based round, we set K to the smallest value of k satisfying $\max_{i \in \{1,\ldots,N\}} b_{k,i} \leq b_{\text{acc}}$. Note also that the values of the other parameters of the algorithm, such as α, β and AC_2, are set to be the same as in the main algorithm.

In the version in which the modes can be added when the main algorithm runs, one could consider stopping the burn-in algorithm separately for each mode and passing the covariance matrix to the main algorithm once its corresponding inhomogeneity factor goes below b_{acc}.

As for the choice of the lengths of the rounds, by default we use a geometric sequence with a common ratio 2. The number AC_1 should grow with the dimension of the state space d since the initial covariance matrix will be based on AC_1 samples for each mode. In our experiments AC_1 is equal to $\max(1000, d^2/2)$.

Note that this construction implies that adapting the matrices by scaling will happen only during the burn-in algorithm, as the number of samples in each mode at the beginning of the main algorithm will be equal to the total length of the number of iterations in the burn-in algorithm, so in particular this number will exceed AC_1.

The adaptation scheme of the main algorithm is based on updating the covariance ma-

	Mixture of Gaussians	Mixture of banana-shaped and t-distributions	Sensor network	LOH example
Main algorithm				
number of iterations	500,000	500,000	500,000	200,000
α	0.7	0.7	0.7	0.7
β	0.0001	0.0001	0.0001	0.0001
ϵ	0.1	0.1	0.1	0.1
$\tilde{\epsilon_w}$	0.01	0.01	0.01	0.01
AC_2	1000	1000	500	500
optimal acceptance rate	0.234	0.234	0.234	0.234
local proposal distributions Q_i	Gaussian t with 7 df	Gaussian t with 7 df	Gaussian t with 7 df	Gaussian/ t-distributed t with 7 df
df of the proposal (if t-distributed)	7	7	7	7
Burn-in algorithm				
number of BFGS runs	1500	40,000	10,000	500
b_{acc}	1.1	1.1	1.1	1.1

Table 2.3: Settings of the parameters used for the examples presented in this paper.

trices passed from the burn-in algorithm.

B .3 Examples – further details

Below we present one more example, a hierarchical Bayesian model for cancer data. We also discuss some further details related to the simulations described in Section 5 . The exact parameters settings of our experiments are summarized in Table 2.3. For all examples shown in this paper we used an implementation of the algorithm in which the burn-in algorithm runs only before the main algorithm (without adding modes on the fly).

Hierarchical Bayesian model for LOH data

The example presented here is based on the Seattle Barrett's Oesophagus study (see Barrett et al. (1996)) analysed later by Warnes, Gregory R (2000), Craiu et al. (2009) and Bai et al. (2011a). Loss of Heterozygosity is the process by which a region of the genome is deleted on either the paternal or maternal inherited chromosomes, leading to a loss of diversity. Loss of Heterozygosity (LOH) rates were collected from oesophageal cancers for 40 regions, each on a distinct chromosome arm. They are of interest since chromosome regions with high rates of

LOH are thought to contain so-called Tumour Suppressor Genes (TSGs) whose functionality is adversely affected by the reduction in genetic diversity. There exists also a proportion of "background" (not cancer-related) LOH. The aim of this study is to provide, for each LOH rate, the probability of being in the TSG group and in the "background" group. Following the approach adopted in the above papers, we consider the following mixture model:

$$x_i \sim \eta\text{Binomial}(N_i, \pi_1) + (1-\eta)\text{Beta-Binomial}(N_i, \pi_2, \gamma) \text{ for } i = 1, \dots 40,$$

where x_i is the number of events of interest (Loss of Heterozygosity) observed in region i, and N_i – the corresponding sample size. Besides, η denotes the probability of a location being a member of the binomial group, π_1 is the probability of LOH in the binomial group, π_2 is the probability of LOH in the beta-binomial group, and γ controls the variability of the beta-binomial group. That is, the likelihood function for this model is given by $\prod_{i=1}^{40} f(x_i, N_i|\eta, \pi_1, \pi_2, \gamma)$ for

$$\begin{aligned} & f(x_i, N_i|\eta, \pi_1, \pi_2, \gamma) \\ &= \eta \binom{N_i}{x_i} \pi_1^{x_i}(1-\pi_1)^{n_i - x_i} + (1-\eta)\binom{N_i}{x_i} \frac{\mathrm{B}\left(x_i + \frac{\pi_2}{\omega_2}, n_i - x_i + \frac{1-\pi_2}{\omega_2}\right)}{\mathrm{B}\left(\frac{\pi_2}{\omega_2}, \frac{1-\pi_2}{\omega_2}\right)}, \end{aligned}$$

where B denotes the beta function and

$$\omega_2 := \frac{\exp(\gamma)}{2\,(1+\exp(\gamma))}.$$

The following prior distributions were used for the parameters of interest:

$$\begin{aligned} \eta &\sim \text{Unif}(0,1) \\ \pi_1 &\sim \text{Unif}(0,1) \\ \pi_2 &\sim \text{Unif}(0,1) \\ \gamma &\sim \text{Unif}(-30, 30). \end{aligned}$$

The resulting target distribution has two non-symmetric and well-separated modes, as depicted in Figure 2.8, one of which has a significantly bigger weight than the other; below we denote them by mode 1 and 2, respectively. We based our analysis on 200,000 steps of the

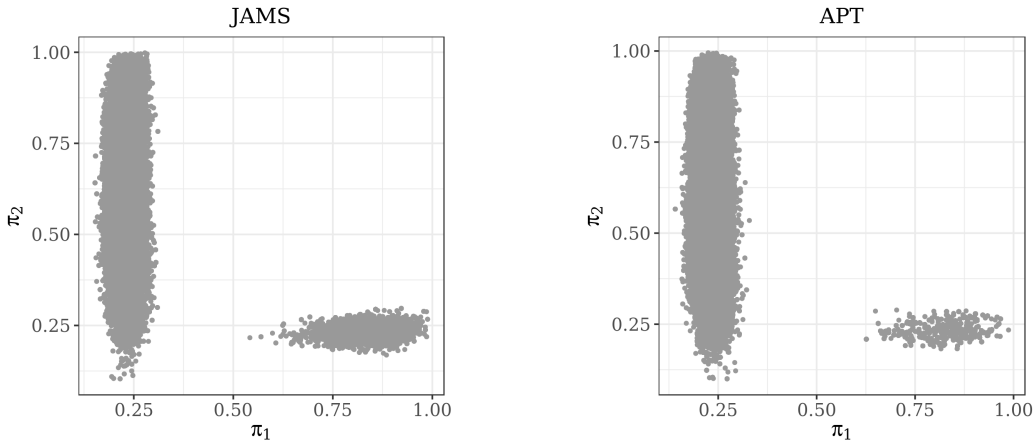


Figure 2.8: Scatterplots of coordinates π_1 and π_2 in the LOH study, obtained using JAMS with t-distributed jumps and APT with 4 temperature levels.

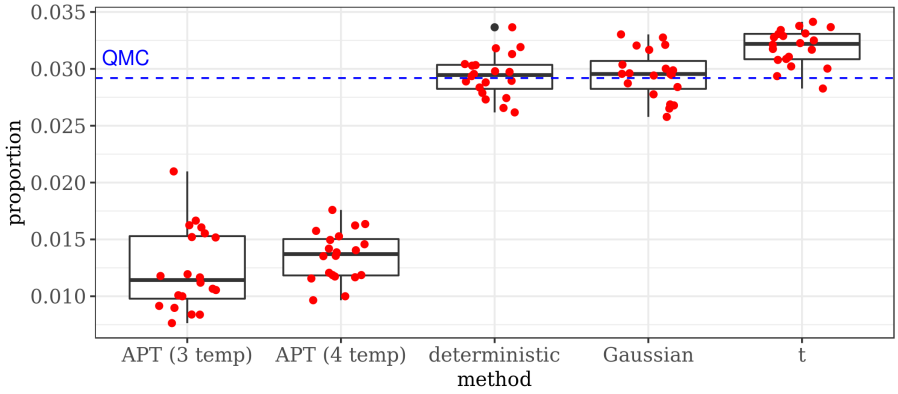


Figure 2.9: A boxplot of proportions of the smaller mode ($\pi_1 > 0.5$) for 20 runs of the experiment for APT (with 3 and 4 temperature levels) and JAMS with three different jump methods.

main algorithm and 500 BFGS runs for the mode-finding stage. The length of the covariance matrix estimation in burn-in algorithm was equal to 3000 iterations for each experiment (chosen automatically). Table 2.4 summarizes the acceptance rates of jumps between the modes for the three versions of the implementation of the algorithm.

As stated above, the prior distribution for all the variables has its support on a compact set. Since this typically has an adverse effect on both mode-finding and sampling, we decided to work with transformed variables, which live on the real line. For the first three variables we applied the logit transformation, i.e. we transformed them using a function $t_1(x) = \log(x) - \log(1-x)$. For the last variable we used the transformation given by $t_2(x) = \log(30+x) - \log(30-x)$.

The starting points for the optimization runs were sampled from the prior distribution and transformed the way described above. The number of function and gradient evaluations for the 20×500 BFGS runs varied between 27 and 428, with an average of 73.

Note that this example is low-dimensional (the dimension is only 4), so Quasi Monte Carlo methods are likely to be very successful at estimating expected values of functions applied to the posterior distribution. For this reason we calculated the weight of the smaller mode (defined as $\pi_1 > 0.5$) using JAMS and compared our results against the proportion of this mode estimated by QMC (with a Sobol sequence of length 20,000,000), used here as an approximation of the ground truth.

We also ran APT with 250,000 iterations, for 3 and 4 temperature levels. The starting points for APT were the 2 modes identified by BFGS runs (used 5 times each), and additionally 10 points sampled from the prior. As shown in Figure 2.9, JAMS gives results very close to those of QMC, consistently across all experiments, with a very slight overestimation of the weight for the t-distributed proposals. Adaptive parallel tempering tends to underestimate significantly the weight of the smaller mode (compare also the two scatterplots in Figure 2.8).

This example shows that multimodal distributions appear in applied problems and might be challenging for MCMC samplers, such as parallel tempering, even in low dimensions. Our method proved to work well on this posterior distribution, however, we admit that given the low dimensionality of the state space other approaches, such as Quasi Monte Carlo, may be more appropriate in this context.

	deterministic		Gaussian		t-distributed	
	Lowest	Highest	Lowest	Highest	Lowest	Highest
Gaussian local proposal						
mode 1 to mode 2	0.01	0.02	0.02	0.02	0.02	0.03
mode 2 to mode 1	0.44	0.71	0.70	0.76	0.71	0.77
t-distributed local proposal						
mode 1 to mode 2	0.01	0.03	0.02	0.03	0.02	0.02
mode 2 to mode 1	0.50	0.78	0.61	0.76	0.63	0.76

Table 2.4: The lowest and the highest value (across 20 runs of the experiment) of the acceptance rates of jump moves between the two modes of the posterior distribution in the LOH study for different jump methods (for the Gaussian and t-distributed local proposal).

Mixture of Gaussians

In Table 2.5 we gathered information about the number of the target density and its gradient evaluations (jointly) in the BFGS runs. We reported the minimum, the mean and the

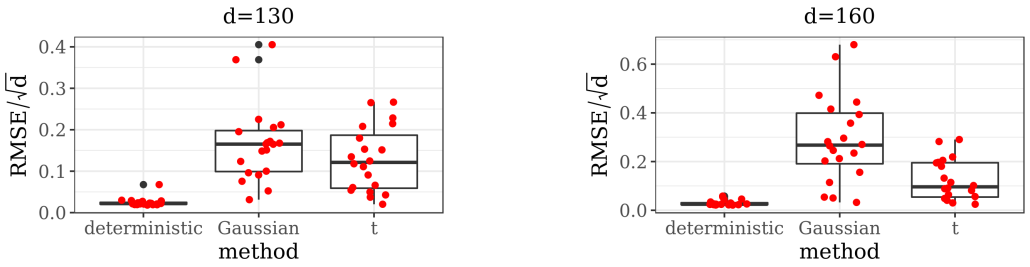


Figure 2.10: Boxplots of the values of RMSE/$\sqrt{d}$ for the mixture of Gaussians across 20 runs of the experiment, dimensions 130 and 160. Note different scales on the y-axis.

maximum value required for the optimization algorithm to converge. The last two columns show the minimum and the maximum number of iterations used for the estimation of the covariance matrices in the burn-in algorithm. These figures show that indeed the computational budget used by our method for dimensions $d = 10$ and $d = 20$ was significantly smaller than the budget of APT (see Section 5 .1).

Figures 2.10 and 2.11 illustrate good performance of our method in dimensions $d = 130$ and $d = 160$, especially for the deterministic jumps. Interestingly, the Gaussian proposal for jumps gives results of the poorest quality on this example.

	Optimization runs			Covariance matrix estimation	
	minimum	mean	maximum	minimum	maximum
d=10	9	11.39	42	3000	3000
d=20	9	10.61	39	3000	7000
d=80	6	7.36	27	255,000	511,000
d=130	8	8.07	24	511,000	511,000
d=160	6	8.02	22	511,000	511,000
d=200	6	6.85	23	1,023,000	1,023,000

Table 2.5: First part: number of the target density and its gradient evaluations in the optimization runs for the mixture of Gaussians. Second part: number of iterations used for the estimation of the covariance matrices in the burn-in algorithm.

Mixture of t and banana-shaped distributions

Let us first ecall the definition of the d-dimensional banana-shaped distribution introduced by Haario et al. (1999)[1]. Let f be the density of the centred t distribution with 7 degrees of freedom and shape matrix C, for $C = \text{diag}(100, \underbrace{1, \ldots, 1}_{d-1})$. Then the density of the banana-

[1]Originally in the paper by Haario et al. Haario et al. (1999) the function f was the density of the Gaussian distribution $N(\mathbf{0}, C)$.

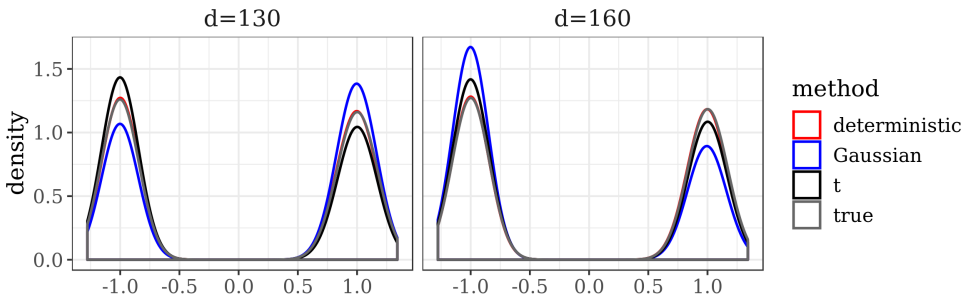


Figure 2.11: Density plots for the mixture of Gaussians, dimensions 130 and 160. The simulations chosen for the analysis correspond to the median value of RMSE across 20 experiments (the tenth largest value of RMSE).

	Optimization runs			Covariance matrix estimation	
	minimum	mean	maximum	minimum	maximum
d=10	21	49	220	7000	15,000
d=20	21	48	216	15,000	63,000
d=50	-	-	-	255,000	255,000
d=80	-	-	-	511,000	511,000

Table 2.6: First part: number of the target density and its gradient evaluations in the optimization runs for the mixture of banana-shaped and t-distributions. Second part: number of iterations used for the estimation of the covariance matrices in the burn-in algorithm.

shaped distribution (with t-tails) is given by

$$f_b = f \circ \phi_b,$$

where

$$\phi_b(x_1, \ldots, x_d) = (x_1, x_2 + bx_1^2 - 100b, x_3, \ldots, x_d). \tag{2.86}$$

In order to decrease the variance of the banana-shaped elements of the mixture, we used the following transformation of f_b (setting $b = 0.03$)

$$\tilde{f}_{0.03}(x_1, \ldots, x_d) = \left(20/\sqrt[4]{d}\right)^d f_{0.03}\left(20/\sqrt[4]{d} \cdot (x_1, \ldots, x_d)\right).$$

Furthermore, the formula on the second coordinate of (2.86) was assigned to coordinate 2, 4, 6, 8, 10 for mode 1, 2, 3, 4 and 5, respectively (this yields the shape of the banana for mode k in the projection onto coordinates 1 and $2k$, as described in Section 5 .2).

Table 2.6 presents analogous information to Table 2.5, for the mixture of t and banana-shaped distributions in different dimensions.

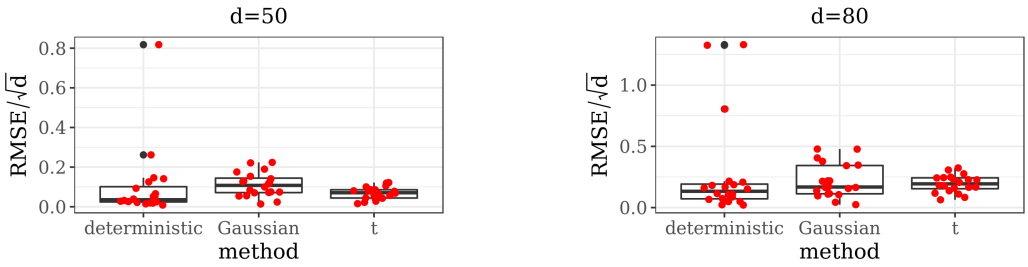


Figure 2.12: Boxplots of the values of RMSE/$\sqrt{d}$ for the mixture of banana-shaped and t-distributions across 20 runs of the experiment, dimensions 50 and 80. Note different scales on the y-axis.

	deterministic		Gaussian		t-distributed	
	Lowest	Highest	Lowest	Highest	Lowest	Highest
d=50	0.26	1.00	0.03	0.14	0.06	0.23
d=80	0.16	1.00	0.01	0.07	0.03	0.12

Table 2.7: The lowest and the highest value (across 20 runs of the experiment) of the acceptance rate of jumps from a given mode, for the mixture of banana-shaped and t-distributions, dimensions 50 and 80.

For dimensions $d = 50$ and $d = 80$ we did not run the mode-finding part, assuming the locations of the modes were known. Overall, our method in all its versions proved to perform well on this high-dimensional example, despite the complicated shapes of the modes. This is illustrated by the scatterplots in Figure 2.13; the skewed shapes of the banana are reflected even in such high dimensions, even though of course the "ends" of the bananas are explored with less accuracy then in dimension 20. Table 2.7 shows that, as before, the deterministic jump method ensures best between-mode mixing. However, it can be noticed that given 20 runs of the experiment, a few times this method delivered results that deviated significantly from the truth (see Figure 2.12).

Sensor network localization

Starting points for the BFGS procedures were sampled uniformly on $[0, 1]^{16}$. The number of function and gradient evaluations for these runs varied between 175 and 876, with an average of 400. The starting points for the APT simulations were the 14 modes identified by the BFGS optimizer and 6 points sampled uniformly on $[0, 1]^{16}$.

Recall that the results for APT presented in Section 5 .3 were based on 4 temperatures. In Figure 2.14 we present analogous results to those shown in Figure 2.7, with the number of temperatures increased to 5 (and the same number of iterations equal to 700,000). Under such settings the APT algorithm mixes better between the modes, however, as illustrated in Figure

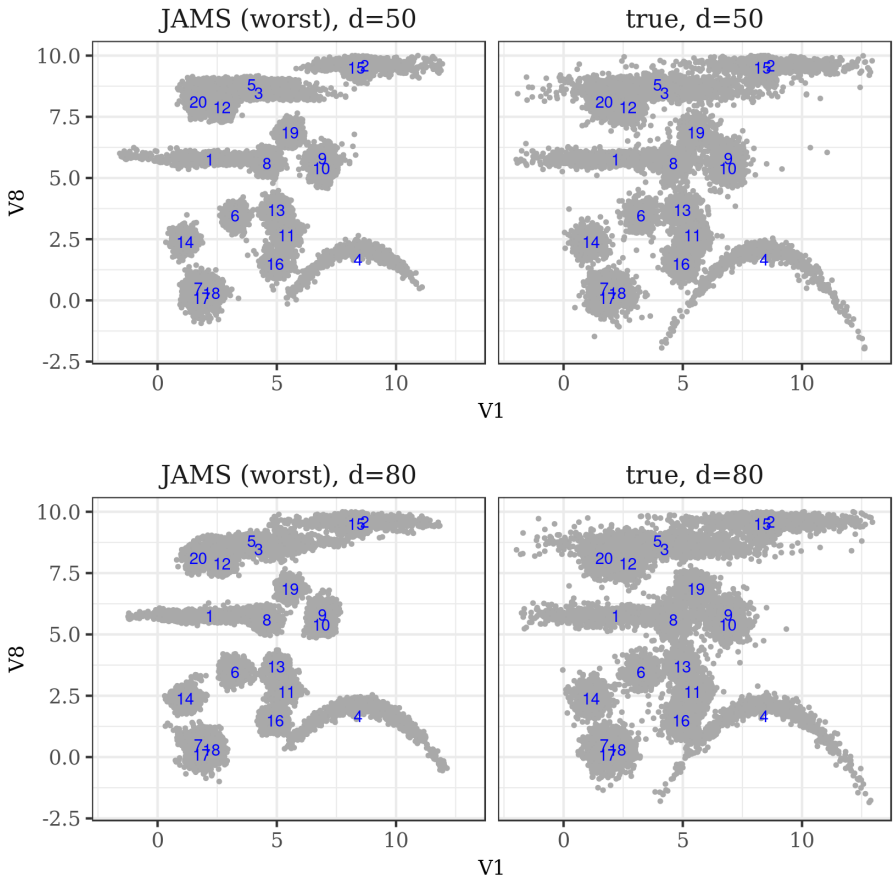


Figure 2.13: Scatterplots of the first and eighth coordinate of the mixture of t and banana-shaped distributions in dimension 50 and 80. Recall that modes 1-5 follow the banana-shaped distribution with t tails, where the shape of the banana for mode k is present in the projection onto coordinates 1 and $2k$. The blue numbers denote the centres of the modes. We compare the results obtained using JAMS (with t-distributed jumps) against samples generated from the true distribution. The simulations chosen for the analysis correspond to the maximum ("worst") value of RMSE across 20 experiments.

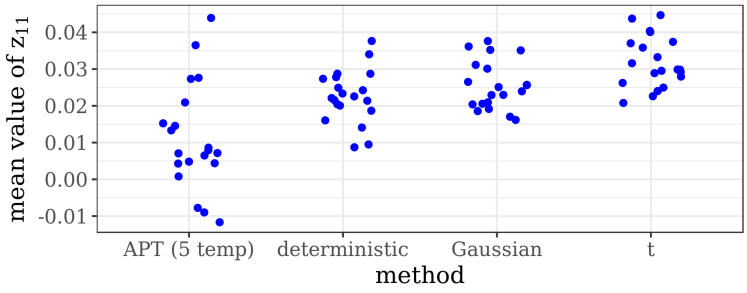


Figure 2.14: A boxplot of the mean value of the first coordinate of sensor 1 for 20 runs of the experiment for APT (with 5 temperature levels) and three versions of JAMS.

2.14, it still yields less stability than our method. Note that APT required $4 \times 700,000 = 2,800,000$ or $5 \times 700,000 = 3,500,000$ target evaluations, for 4 and 5 temperature levels, respectively. Assuming an implementation of JAMS on a standard desktop computer with 8 cores, the computational cost measured by the number of target and gradient evaluations per core would be at most:

- for mode finding: $10,000/8 \times 875$ (as for each BFGS run we had at most 875 such evaluations);
- for the burn-in-algorithm: $2 \times 15,000$ target evaluations (as the estimation of 14 covariance matrices needed to be split across 8 cores);
- for the main algorithm: $500,000$ target evaluations.

Altogether, this would give $1,623,750$ evaluations, and the additional overhead resulting from the communication between cores. This figure shows that APT required a larger computational cost than JAMS in our setup even though the above analysis was carried out under a pessimistic scenario. Firstly, on average there were 400 evaluations per BFGS procedure and plugging this value into the above calculations would decrease the overall number of evaluations to 1,000,030. What is more, typically a user would run our algorithm on a server or a cloud service, which we in fact did as well. This would allow splitting the computational cost (in particular, that of BFGS runs) across a much larger number of cores.

Statement of Authorship for joint/multi-authored papers for PGR thesis

To appear at the end of each thesis chapter submitted as an article/paper

The statement shall describe the candidate's and co-authors' independent research contributions in the thesis publications. For each publication there should exist a complete statement that is to be filled out and signed by the candidate and supervisor **(only required where there isn't already a statement of contribution within the paper itself).**

Title of Paper	A framework for adaptive MCMC targeting multimodal distributions
Publication Status	Published
Publication Details	Pompe, E., Holmes, C., Łatuszyński, K.: A framework for adaptive MCMC targeting multimodal distributions. The Annals of Statistics **48**(5),(2020)

Student Confirmation

Student Name:	Emilia Pompe		
Contribution to the Paper	• designed the main JAMS algorithm and the burn-in algorithm (the construction of the JAMS algorithm had significant input from Krys Łatuszyński) • proved theoretical results (some substantial parts of the proof the Weak Law of Large Numbers were done by Krys Łatuszyński; he also defined the interpolating processes that facilitated the construction of the Auxiliary Variable Adaptive MCMC class and our coupling construction; more generally, he gave feedback on theoretical results) • provided the literature review • coded up the simulations for all numerical examples included in the paper • wrote the manuscript (with feedback from both authors; the LOH example was partly written up by Chris Holmes)		
Signature	E Pompe	Date	21/10/2021

Supervisor Confirmation

By signing the Statement of Authorship, you are certifying that the candidate made a substantial contribution to the publication, and that the description described above is accurate.

Supervisor name and title:	Judith Rousseau, Professor		
Supervisor comments			
Signature		Date	21/10/2021

This completed form should be included in the thesis, at the end of the relevant chapter.

Chapter 3

Introducing prior information in Weighted Likelihood Bootstrap with applications to model misspecification

Abstract

We propose Posterior Bootstrap, a set of algorithms extending Weighted Likelihood Bootstrap, to properly incorporate prior information and address the problem of model misspecification in Bayesian inference. We consider two approaches to incorporating prior knowledge: the first is based on penalization of the Weighted Likelihood Bootstrap objective function, and the second uses pseudo-samples from the prior predictive distribution. We also propose methodology for hierarchical models, which was not previously considered for methods based on Weighted Likelihood Bootstrap. Edgeworth expansions guide the development of our methodology and allow us to provide greater insight on properties of Weighted Likelihood Bootstrap than were previously known. Our experiments confirm the theoretical results and show a reduction in the impact of model misspecification against Bayesian inference in the misspecified setting.

1 Introduction

1 .1 Model misspecification and robust inference in Bayesian statistics

Bayesian inference relies on an assumption that the true data-generating mechanism is known. However, in practice the models only provide an approximate reflection of the reality, so model misspecification is a frequent concern (De Blasi and Walker, 2013; Walker, 2013; Lyddon et al., 2019). Throughout the paper we write $x_1, \ldots, x_n \sim f(x|\theta)$ to denote that $x_1, \ldots, x_n$ are independent and identically distributed according to a distribution with density $f(x|\theta)$. We consider a generic parametric Bayesian model for parameter of interest $\theta \in \mathbb{R}^d$

$$x_1, \ldots, x_n \sim f(x|\theta), \quad \theta \sim \pi(\theta), \quad \theta \in \Theta, \tag{3.1}$$

with prior π and likelihood function f.

Misspecification of model (3.1) occurs when data is generated from a distribution P^* with density p^*, which is not equal to $f(x|\theta)$ for any $\theta \in \Theta$. In this case under regularity conditions the Bayesian posterior asymptotically concentrates around the pseudo-true parameter θ^* minimizing $D_{KL}(p^*||f)$, the Kullback–Leibler divergence between $f(\cdot|\theta)$, $\theta \in \Theta$, and p^* (Berk, 1966, 1970; Bunke and Milhaud, 1998). Applying standard Bayesian inference to misspecified models leads to misleading uncertainty estimates (Kleijn and Van der Vaart, 2012), so there has been increasing interest in robust inference schemes (Grünwald and van Ommen, 2017; Jewson et al., 2018; Hong and Martin, 2020).

One method that has gained a lot of popularity replaces the likelihood $f(x|\theta)$ by a power likelihood $f(x|\theta)^\eta$ for a parameter $\eta > 0$ (Jiang and Tanner, 2008; Grünwald, 2012; Bissiri et al., 2016; Grünwald and van Ommen, 2017). The interpretation of η is that it represents the learning rate, or the analyst's trust in the model. For $\eta < 1$ the data have less influence than in the standard Bayesian inference, while using $\eta > 1$ places more trust in the likelihood.

Resampling schemes are another tool for addressing Bayesian model misspecification (Waddell et al., 2002; Douady et al., 2003). The BayesBag algorithm (Bühlmann, 2014; Huggins and Miller, 2019, 2020) combines samples from posteriors on bootstrapped data sets, where each bootstrapped data set is created by drawing with replacement M out of n observations from the original data set. The algorithm utilizes a hyperparameter b defined

as $\lim_{n\to\infty} M/n$. Asymptotic theory for BayesBag was proved by Huggins and Miller (2019).

Another group of robust inference methods is based on Weighted Likelihood Bootstrap (Newton, 1991; Newton and Raftery, 1994; Lyddon et al., 2019), and this is also a starting point for our approach.

In this paper we develop Posterior Bootstrap, a set of scalable algorithms for performing principled inference in Bayesian models under possible model misspecification. We extend Weighted Likelihood Bootstrap to incorporate prior information. Several approaches to this task have been recently proposed in the literature (Lyddon et al., 2018; Fong et al., 2019; Newton et al., 2020), to our knowledge however, this is the first attempt to analyse theoretically the impact of various ways of incorporating the prior. Edgeworth expansions show that prior beliefs about the parameter of interest are properly represented in the inference in a similar way to standard Bayesian inference, and guide our choice of hyperparameters of the algorithms. Moreover, we propose a robust inference scheme for hierarchical models, which has not been done so far for methods based on Weighted Likelihood Bootstrap or power posteriors.

1 .2 Properties of Weighted Likelihood Bootstrap

Weighted Likelihood Bootstrap was originally proposed as a method for sampling approximately from a posterior distribution. Each sample is drawn by optimizing the weighted log likelihood

$$\arg\max_{\theta} \left\{ \sum_{i=1}^{n} w_i \log f(x_i|\theta) \right\}, \quad w_1, \ldots, w_n \sim \text{Exp}(1). \tag{3.2}$$

Combined with an adjustment by the Sequential Importance Resampling algorithm (Gordon et al., 1993; Doucet et al., 2000), which could correct for the fact that Weighted Likelihood Bootstrap is approximate, it was an alternative to then emerging Markov chain Monte Carlo methods. A key advantage of this algorithm is that it allows us to perform computations fully in parallel, therefore a recent resurgence of interest in this method (Newton et al., 2020; Lyddon et al., 2019; Ng and Newton, 2020; Nie and Ročková, 2020) can partly be explained by developments in parallel computing. Even though Weighted Likelihood Bootstrap appeared in the Bayesian literature, it does not incorporate a prior. A similar idea, called perturbing the minimand, was developed independently in the frequentist literature (Jin et al., 2001). The analogy between these two methods was noticed by Parzen and Lipsitz (2007). Some

frequentist approaches consider adding a lasso-type regularization term (Minnier et al., 2011; Das et al., 2019), which from the Bayesian point of view could be interpreted as adding a prior.

Asymptotic results for the Weighted Likelihood Bootstrap were proved by Newton (1991) under a correctly specified model and extended by Lyddon et al. (2019) to the misspecified case. Let the maximum likelihood estimator $\hat{\theta}$ be $\arg\max_{\theta\in\Theta}\sum_{i=1}^{n} f(x_i|\theta)$, and let $\tilde{\theta}$ be a draw from the Weighted Likelihood Bootstrap algorithm based on observations $x_1,\ldots,x_n$. Theorem 1 of Lyddon et al. (2019) states that under standard regularity conditions, for any Borel set A

$$\mathrm{pr}\left\{n^{1/2}\left(\tilde{\theta}-\hat{\theta}\right)\in A|x_{1:n}\right\}\to\mathrm{pr}\left(Z\in A\right),\quad Z\sim N\left\{0,J(\theta^*)^{-1}I(\theta^*)J(\theta^*)^{-1}\right\},\tag{3.3}$$

where

$$I(\theta)=E_{p^*}\left[\left\{\nabla_\theta\log f(X|\theta)\right\}\left\{\nabla_\theta\log f(X|\theta)\right\}^T\right],\quad J(\theta)=-E_{p^*}\left[D^2_\theta\left\{\log f(X|\theta)\right\}\right],\tag{3.4}$$

and θ^* is the unique pseudo-true parameter minimizing $D_{KL}(p^*||f)$. The asymptotic covariance matrix in (3.3) is the so-called sandwich covariance matrix known from the frequentist literature on model misspecification (Huber, 1967) as an asymptotic covariance matrix of the maximum likelihood estimator under a potentially misspecified model.

We let $\Pi(\cdot|x_{1:n})$ be the Bayesian posterior distribution given observations $x_1,\ldots,x_n$ under model (3.1), and by $\pi(\cdot|x_{1:n})$ we denote its density. Recall that the Bernstein–von Mises theorem for misspecified Bayesian models (Kleijn and Van der Vaart, 2012) states that for any Borel set A

$$\Pi\left\{n^{1/2}\left(\theta-\hat{\theta}\right)\in A|x_{1:n}\right\}\to\mathrm{pr}\left(Z'\in A\right),\quad Z'\sim N\left\{0,J(\theta^*)^{-1}\right\}.\tag{3.5}$$

Convergence in (3.5) holds in total variation, while Lyddon et al. (2019) considered convergence in distribution. We show however in Corollary 3.12 that convergence in (3.3) also holds in total variation. Note that under both standard Bayesian inference and Weighted Likelihood Bootstrap the asymptotic distribution is concentrated at the same mean, but with potentially different covariance matrices. Comparing (3.3) and (3.5), we conclude that for misspecified models under Weighted Likelihood Bootstrap credible sets are valid confidence

sets in the frequentist sense, which is not the case for standard Bayesian posteriors (Kleijn and Van der Vaart, 2012). For correctly specified models, the covariance matrices in (3.3) and (3.5) are equal.

While the above property is related to inference, the results obtained by Fushiki (2005, 2010) explain why Weighted Likelihood Bootstrap outperforms the standard Bayesian approach in terms of prediction. In particular, these two articles analyse the risk function

$$E_{x_{1:n}}\left[D_{KL}\left\{p^*||\hat{p}(\cdot|x_{1:n})\right\}\right] = \int p^*(x_{1:n})\left\{\int p^*(x_{n+1})\log\frac{p^*(x_{n+1})}{\hat{p}(x_{n+1}|x_{1:n})}\mathrm{d}x_{n+1}\right\}\mathrm{d}x_{1:n}, \quad (3.6)$$

where $\hat{p}(\cdot|x_{1:n})$ is the predictive distribution given data $x_{1:n}$ under a given method of inference. Proposition 3.1 below is a direct corollary of Theorem 1 of Fushiki (2010) and Theorems 1 and 3 of Fushiki (2005), see also Lyddon et al. (2018).

Proposition 3.1. *Consider model* (3.1) *and assume that* $\theta^* \in \Theta$ *minimizing* $D_{KL}(p^*||f)$ *is unique. Suppose that* $I(\theta^*)$ *and* $J(\theta^*)$ *are positive definite, and that* $I(\theta^*) \neq J(\theta^*)$. *Then the predictive distribution based on Weighted Likelihood Bootstrap asymptotically provides better prediction than the Bayesian predictive distribution, that is, the value of the risk in* (3.6) *is asymptotically smaller.*

In view of the robust inference methods discussed in Section 1 .1, a natural question to ask is whether the result stated in Proposition 3.1 can be extended from standard Bayesian inference to power posteriors. Theorem 3.2 below shows that the answer is negative since under certain conditions power posteriors yield an asymptotically smaller risk than Weighted Likelihood Bootstrap.

Theorem 3.2. *Consider model* (3.1) *and let* $\lambda_1, \ldots, \lambda_d$ *denote the eigenvalues of*

$$J(\theta^*)^{-1/2} I(\theta^*) J(\theta^*)^{-1/2},$$

where $\theta^* \in \Theta$ *is the unique pseudo-true parameter. Suppose that the eigenvalues satisfy either* $1 < \lambda_1, \ldots, \lambda_d$, *or* $\lambda_1, \ldots, \lambda_d < 1$. *Then the power* η *for a power posterior* π_η, *or the hyperparameter* b *in BayesBag can be chosen in such a way that the risk given by* (3.6) *associated with the predictive distribution under Weighted Likelihood Bootstrap is larger than the risk associated with the predictive distribution under* π_η *or under BayesBag, respectively.*

For prediction problems, Theorem 3.2 shows that in certain scenarios Weighted Likeli-

hood Bootstrap may be outperformed by power posteriors and by BayesBag, with an appropriate choice of hyperparameters. We present the proof in Supplementary Material A . If matrices $I(\theta^*)$ and $J(\theta^*)$ are diagonal, the above result has an intuitive interpretation. If $1 < \lambda_1, \ldots, \lambda_d$, or $\lambda_1, \ldots, \lambda_d < 1$, this means that the model exhibits either overdispersion or underdispersion on all coordinates. Then the hyperparameters for BayesBag or power posteriors can be chosen in such a way that these algorithms provide better prediction than Weighted Likelihood Bootstrap. This fact should not be surprising given that power posteriors and BayesBag have an additional calibration parameter, whereas Weighted Likelihood Bootstrap does not. What is more, Weighted Likelihood Bootstrap is designed to pursue the correct coverage, rather than predictive accuracy.

If, on the other hand, some eigenvalues are smaller, and some greater than 1, the model is overdispersed along some directions and underdispersed along others. Since BayesBag and power posteriors use one-dimensional parameters to scale the entire asymptotic covariance matrix, they are unable to correct for underdispersion and overdispersion at the same time. In these cases Weighted Likelihood Bootstrap, which can adjust the entire covariance matrix, provides asymptotically better prediction.

1 .3 Edgeworth expansions

In order to develop appropriate methodology for incorporating prior information into Weighted Likelihood Bootstrap, we first need to specify what tools can be used for measuring the impact of the prior on the draws $\tilde{\theta}$ from the resulting algorithm. To this end, we examine the way the prior influences the posterior in standard Bayesian inference. Recall that by the Bernstein–von Mises theorem the prior asymptotically does not influence the first order approximation, which necessitates considering higher order approximations to the posterior.

In classical asymptotic theory of parametric inference, the Edgeworth expansion of a density function is a well-established tool of analysing higher order approximations. For details we refer the reader to Bhattacharya and Ghosh (1978); Ghosh (1994); Hall (2013). The Gaussian approximation appearing in the central limit theorem is associated with an absolute approximation error of the order $O\left(n^{-1/2}\right)$. The Edgeworth expansion should be seen as a natural extension of this approximation, up to the order $O\left(n^{-k/2}\right)$ for an arbitrary natural number k. The coefficients of this expansion are expressed in terms of cumulants of the density of interest.

Letting $\ell(x,\theta) = \log f(x|\theta)$, we use

$$I_n = \frac{1}{n}\sum_{j=1}^{n} \nabla_\theta \ell(x_j,\hat{\theta}) \nabla_\theta \ell(x_j,\hat{\theta})^T, \quad J_n = -\frac{1}{n}\sum_{j=1}^{n} D^2_\theta \ell(x_j,\hat{\theta})$$

as empirical estimates of $I(\theta^*)$ and $J(\theta^*)$ respectively. Let ϕ denote the density of a d-variate standard normal distribution. Recall that in the case of well-specified models the Laplace approximation of the posterior density of $I_n^{1/2} n^{1/2} \left(\theta - \hat{\theta}\right)$ is

$$\phi(y)\left[1 + y^T I(\theta^*)^{-1/2} \frac{\{\nabla_\theta \log \pi(\theta^*)\}}{n^{1/2}} + I(\theta^*)^{-3/2} \frac{\sum_{p,q,r=1}^{d} c_{pqr} y_p y_q y_r}{6n^{1/2}}\right] + o\left(n^{-1/2}\right) \tag{3.7}$$

for $y = (y_1, \ldots, y_d) \in \mathbb{R}^d$, where $c_{pqr} = E_{p^*}\left\{\partial^3 \ell(\theta^*, x)/\partial\theta_p \partial\theta_q \partial\theta_r\right\}$. Since in Bayesian posteriors the impact of the prior appears in the second order Edgeworth expansion, we will keep this as our primary requirement. That is, we demand that the prior information appears in the second order Edgeworth expansion obtained for Posterior Bootstrap. Ideally, for well-specified models our method would incorporate the prior knowledge in the same way it is done in Bayesian inference, that is, through the same prior-dependent expression in the second order expansion. The reason for this is that in the case of well-specified models the Bayesian posterior offers optimal information about the parameter given the data and the prior distribution (Bernardo and Smith, 2009). Alternatively, as explained by Bissiri et al. (2016), the posterior $\Pi(\theta|x_{1:n})$ is the only probability measure minimizing the expected loss under the negative log likelihood loss function, which is an appropriate loss function to be used in the well-specified case, and hence provides a valid and coherent update of beliefs about θ.

On the other hand, in the case of misspecified models the first order expansions of Posterior Bootstrap and of the Bayesian posterior are different, as seen in (3.3) and (3.5), which is what allows Posterior Bootstrap to correct for model misspecification.

2 Posterior Bootstrap via penalization on the weighted log-likelihood

2 .1 Methodology for non-hierarchical models

We now present Algorithm 5, our Posterior Bootstrap method via prior penalization for model (3.1). The algorithm has only one parameter w_0 representing the strength of the penalization imposed on the weighted log likelihood with the prior distribution. For $w_0 = 0$, Algorithm 5 recovers Weighted Likelihood Bootstrap.

Algorithm 5 Posterior Bootstrap via prior penalization, returns samples from $P^{B,p}(\cdot|x_{1:n})$

1: **hyperparameters**: w_0.
2: **for** j = 1, . . . , N **do** ▷ (in parallel)
3: Draw $w_1^{(j)}, \ldots, w_n^{(j)} \sim \text{Exp}(1)$.
4: Set $\tilde{\theta}^{(j)} = \arg\max_\theta \left\{ \sum_{i=1}^n w_i^{(j)} \log f(x_i|\theta) + w_0^T \log \pi(\theta) \right\}$.
5: **end for**
6: **return** Samples $\{\tilde{\theta}^{(j)}\}_{j=1}^N$.

In the basic version of the algorithm we assume that w_0 is a non-negative real number. In this case Algorithm 5 resembles the Weighted Bayesian Bootstrap (Newton et al., 2020). Instead of a fixed parameter w_0, Newton et al. (2020) use a random variable $\lambda\text{Exp}(1)$ for a regularization constant λ, typically selected via cross-validation. If we additionally assume that the prior $\pi(\theta)$ factorizes with respect to all coordinates of $\theta = (\theta_1, \ldots, \theta_d)$ as

$$\pi(\theta) = \prod_{k=1}^{d} \pi_k(\theta_k), \tag{3.8}$$

we then treat w_0 as a d-dimensional vector $(w_{0,1}, \ldots, w_{0,d})$ with non-negative entries, rather than a non-negative real value. For simplicity of notation the penalization term $w_0^T \log \pi(\theta)$ of Algorithm 5 denotes $\sum_{k=1}^d w_{0,k} \log \pi_k(\theta_k)$. We defer a discussion about the appropriate choice of w_0 to Section 2 .3. We let $P^{B,p}$ be the distribution of $\tilde{\theta}$ drawn from Algorithm 5 and by $p^{B,p}$ we denote its density.

Instead of using the log likelihood function in Algorithm 5, the user may prefer to work with a loss function additive over $x_1, \ldots, x_n$, following the setting of Bissiri et al. (2016); Lyddon et al. (2019).

Remark 3.3. Algorithm 5 can be extended to the setting of Bissiri et al. (2016); Lyddon et al. (2019), where the negative log likelihood $-\log f(x|\theta)$ is replaced with any loss function

additive over $x_1, \ldots, x_n$. The asymptotic results will then be analogous to those stated in Section 2 .2 under the same regularity conditions.

This approach is particularly useful when the user believes the model is complex, with a high-dimensional nuisance parameter, while the parameter of interest is relatively low-dimensional. Examples of loss functions suitable for such scenarios are discussed for example by Bissiri et al. (2016). We stress out that the above remark applies to loss functions used for models with nuisance parameters of a potentially high, but fixed dimension. For results on scenarios where the dimension of the parameter grows with the number of observations, we refer the reader to Newton et al. (2020); Nie and Ročková (2020). We also touch upon the issue of a growing dimensionality of the parameter in Section 3 of this paper.

2 .2 Asymptotic results for Algorithm 5

In this section we provide the Edgeworth expansion of the output of Algorithm 5, which generalizes the results obtained by Lyddon et al. (2019) to the penalized version, and strengthens them to the second order expansion.

In what follows, if there is ambiguity, we write E_z to denote expectation with respect to z. We consider the following set of assumptions.

Assumption 3.4. (Log likelihood). The log likelihood function $\ell(x, \theta) = \log f(x|\theta)$ is measurable and bounded from above, with $E_{p^*}|\ell(X, \theta)| < \infty$ for all $\theta \in \Theta$.

Assumption 3.5. (Identifiability). There exists a unique maximizing parameter value

$$\theta^* = \arg\max_{\theta \in \Theta} \mathbb{E}_{p^*} \{\ell(X, \theta)\},$$

and further for all $\delta > 0$ there exists an $\epsilon > 0$ such that

$$\lim_{n \to \infty} P^* \left[\sup_{\|\theta - \theta^*\| \geq \delta} \frac{1}{n} \sum_{i=1}^{n} \{\ell(X_i, \theta) - \ell(X_i, \theta^*)\} \leq -\epsilon \right] = 1,$$

where P^* is the data-generating distribution.

Assumption 3.6. (Smoothness of the log likelihood function). There is an open ball B containing θ^* such that $\ell(\theta, x)$ is three times continuously differentiable with respect to $\theta \in B$ almost surely on x with respect to P^*. We additionally impose certain moment conditions on the partial derivatives. For details see Supplementary Material B .1.

Assumption 3.7. (Positive definiteness of information matrices and linear independence of partial derivatives). For B like in Assumption 3.6, the matrices $I(\theta)$ and $J(\theta)$ defined in (3.4) are positive definite for $\theta \in B$, with all elements finite. Besides, the first and second partial derivatives

$$\left\{ \frac{\partial \ell(x,\theta)}{\partial \theta_p}, \frac{\partial^2 \ell(x,\theta)}{\partial \theta_r \partial \theta_s} \right\}, \quad 1 \le p \le d,\, 1 \le r \le s \le d$$

are linearly independent as functions of x at θ^*.

Assumption 3.8. (Distribution of weights). The weights $w_{1:n} = (w_1, \ldots, w_n)$ are independent and identically distributed according to a continuous distribution W having all moments and such that $E(W) = 1$ and $\mathrm{var}(W) = 1$.

Assumptions 3.4 – 3.6, as well as positive definiteness of information matrices in Assumption 3.7 are direct analogues of regularity conditions considered by Lyddon et al. (2019). Linear independence of partial derivatives expressed in Assumption 3.7 is a common assumption for obtaining Edgeworth expansions of the maximum likelihood estimator. Assumptions 3.4 and 3.5 ensure that as $n \to \infty$, with probability going to 1 distribution $P^{B,p}(\cdot|x_{1:n})$ concentrates around θ^*. Assumptions 3.6 and 3.7 are used to show the validity of the Edgeworth expansion. Since the algorithms we present in the paper rely on the Weighted Likelihood Bootstrap formulation (3.2), by default the weights follow the exponential distribution Exp(1), thus satisfying Assumption 3.8, we allow however for other distributions of weights. We will use the above conditions throughout the paper to prove our theoretical results for various versions of Posterior Bootstrap.

In the case of Algorithm 5, we require the following smoothness assumption on $\pi(\theta)$ in the neighbourhood of θ^*.

Assumption 3.9. (Smoothness of log prior density). The function $\log \pi(\theta)$ is measurable, upper-bounded on Θ, and three times continuously differentiable in the neighbourhood of θ^*.

We denote by $w_0 \cdot \nabla_\theta \log \pi(\theta^*)$ a vector with pth entry $w_{0,p}\pi'(\theta_p^*)/\pi(\theta_p^*)$ if $w_0 \in \mathbb{R}^d$ and $w_0\pi'(\theta_p^*)/\pi(\theta_p^*)$ when $w_0 \in R$. When $\theta \in \mathbb{R}$, we adopt the following notation:

$$\mu_3 = E_{p^*}\left\{ D_\theta^3 \ell(X,\theta^*) \right\}, \quad A_3 = E_{p^*}\left\{ \nabla_\theta \ell(X,\theta^*) \right\}^3, \quad L_{1,2} = E_{p^*}\left\{ D_\theta^2 \ell(X,\theta^*) \nabla_\theta \ell(X,\theta^*) \right\}. \tag{3.9}$$

We denote by γ the third central moment of W, that is $\gamma = \mathbb{E}\{(W-1)^3\}$. In particular $\gamma = 2$ when $W \sim \text{Exp}(1)$.

Theorem 3.10. *Consider model (3.1) with a parameter of interest $\theta \in \mathbb{R}^d$ and suppose Assumptions 3.4 – 3.9 hold. As $n \to \infty$, with probability going to 1 the Edgeworth expansion of the density of $I_n^{-1/2} J_n n^{1/2}(\tilde{\theta} - \hat{\theta})$ for $\tilde{\theta}$ drawn from Algorithm 5 is*

$$\phi(y)\left[1 + y^T I(\theta^*)^{-1/2} \frac{\{w_0 \cdot \nabla_\theta \log \pi(\theta^*)\}}{n^{1/2}} + \frac{P_{3,n}(y)}{n^{1/2}}\right] + o\left(n^{-1/2}\right), \tag{3.10}$$

where $P_{3,n}(y)$ is an order three polynomial independent from the prior. When $d = 1$, formula (3.10) has the following more explicit form

$$\phi(y)\left\{1 + \kappa_{1,n} y + \kappa_{3,n} \frac{y^3 - 3y}{6}\right\} + o\left(n^{-1/2}\right), \tag{3.11}$$

where

$$\kappa_{1,n} = \frac{w_0 \nabla_\theta \log \pi(\theta^*)}{n^{1/2} I(\theta^*)^{1/2}} + \frac{\mu_3}{2n^{1/2}} \frac{I(\theta^*)^{1/2}}{J(\theta^*)^2} + \frac{L_{1,2}}{n^{1/2} I(\theta^*)^{1/2} J(\theta^*)},$$
$$\kappa_{3,n} = \frac{\gamma A_3}{n^{1/2} I(\theta^*)^{3/2}} + \frac{3\mu_3}{n^{1/2}} \frac{I(\theta^*)^{1/2}}{J(\theta^*)^2} + \frac{6 L_{1,2}}{n^{1/2} I(\theta^*)^{1/2} J(\theta^*)}.$$

Additionally, if the model is well-specified and $\gamma = 2$, formula (3.11) becomes

$$\phi(y)\left[1 + \frac{w_0 \nabla_\theta \log \pi(\theta^*) y}{n^{1/2} I(\theta^*)^{1/2}} - \frac{(A_3 + \mu_3)\, y}{3n^{1/2} I(\theta^*)^{3/2}} + \frac{\mu_3 y^3}{6n^{1/2} I(\theta^*)^{3/2}}\right] + o\left(n^{-1/2}\right). \tag{3.12}$$

The sketch of the proof of Theorem 3.10 is as follows. We first expand $\sqrt{n}\left(\tilde{\theta} - \hat{\theta}\right)$ as a function of $\sum_{i=1}^d b(x_i, \hat{\theta})$, where $b(x_i, \hat{\theta})$ is a vector containing derivatives of $\ell(x_i, \theta)$ with respect to θ. We then obtain the Edgeworth expansion of $\sum_{i=1}^d b(x_i, \hat{\theta})$ as a random variable in $w_{1:n}$. Results from Chapter 2 of Ghosh (1994) are used to justify that (3.11) is indeed a valid second order Edgeworth expansion, that is, that the incurred error is of order $o\left(n^{-1/2}\right)$. For the full proof with all calculations we refer the reader to Supplementary Material B . We also provide an Edgeworth expansion for the related method by Newton et al. (2020) in Supplementary Material B .6.

To derive formula (3.12) from (3.11), recall that for well-specified models we have $I(\theta^*) =$

$J(\theta^*)$ and by one of the Bartlett identities (Bartlett, 1953a,b)

$$-L_{1,2} = \frac{\mu_3 + A_3}{3}. \tag{3.13}$$

If additionally $\gamma = 2$, we can simplify the formulas for $\kappa_{3,n}$ and $\kappa_{1,n}$ to

$$\begin{aligned}
\kappa_{3,n} =& \frac{2A_3}{n^{1/2} I(\theta^*)^{3/2}} + \frac{3\mu_3}{2n^{1/2} I(\theta^*)^{3/2}} - \frac{\mu_3 + 2A_3}{n^{1/2} I(\theta^*)^{3/2}} = \frac{\mu_3}{n^{1/2} I(\theta^*)^{3/2}}, \\
\kappa_{1,n} =& \frac{w_0 \nabla_\theta \log \pi(\theta^*)}{n^{1/2} I(\theta^*)^{1/2}} + \frac{\mu_3}{6n^{1/2} I(\theta^*)^{3/2}} - \frac{A_3}{3n^{1/2} I(\theta^*)^{3/2}}.
\end{aligned}$$

For well-specified models, it is insightful to compare the Edgeworth expansions obtained for Posterior Bootstrap (3.10) with the Laplace approximations for Bayesian posteriors (3.7). Note that for $w_0 = 1$ the term corresponding to the prior, that is $n^{-1/2} I(\theta^*)^{-1/2} \nabla_\theta \log \pi(\theta^*)$ appears in the same way in both formulas. Firstly, this suggests using $w_0 = 1$ if the analyst believes the model is well-specified, which should in fact be intuitive given that for $w_0 = 1$

$$\mathbb{E}_{w_{1:n}} \left\{ \sum_{i=1}^{n} w_i \log f(x_i|\theta) + w_0 \log \pi(\theta) \right\} = \log \pi(\theta|x_{1:n}) + \mathrm{c},$$

where the term c does not depend on θ. This also has implications when the analyst wants to choose between two prior distributions $\pi_1(\theta)$ and $\pi_2(\theta)$. For $w_0 = 1$ we then have

$$\pi_1(\theta|x_{1:n}) - \pi_2(\theta|x_{1:n}) = p^{B,p}_{\pi_1}(\theta|x_{1:n}) - p^{B,p}_{\pi_2}(\theta|x_{1:n}) + o\left(n^{-1/2}\right),$$

where $p^{B,p}_{\pi}$ denotes a density $p^{B,p}$ under prior π. This facilitates understanding the impact of the prior and suggests making decisions about the choice of the prior following the guidelines developed for Bayesian inference.

At the same time it should be noted that in general the Posterior Bootstrap method does not yield second order correctness to Bayesian posteriors for well-specified models. This is due to the additional term $-n^{-1/2} I(\theta^*)^{-3/2} (A_3 + \mu_3)/3$ in formula (3.12), which is only equal to 0 in special cases. Further details about the consequences of this result are presented in Proposition 3.11.

Proposition 3.11. *Consider model (3.1) with a one-dimensional parameter θ. Suppose $\tilde{\theta}$ is drawn from Weighted Likelihood Bootstrap, that is, from Algorithm 5 with $w_0 = 0$. If the model is well-specified, then the Edgeworth expansion of $I_n(\hat{\theta})^{1/2} n^{1/2} (\tilde{\theta} - \hat{\theta})$ is second order*

correct to the corresponding expansion for Bayesian posterior utilizing Jeffreys prior on θ.

Proof. Suppose that $\pi(\theta)$ is a Jeffreys prior, that is $\pi(\theta) \propto \{I(\theta)\}^{1/2}$. Then

$$\nabla_\theta \log \pi(\theta^*) = \frac{I(\theta^*)^{-1/2}\nabla_\theta I(\theta^*)}{2I(\theta^*)^{1/2}} = \frac{\nabla_\theta I(\theta^*)}{2I(\theta^*)} = \frac{2L_{1,2}(\theta^*)}{2I(\theta^*)} = -\frac{A_3 + \mu_3}{3I(\theta^*)^{3/2}},$$

where the last equality follows from the Bartlett identity (3.13). Comparing formula (3.12) for $w_0 = 0$ and (3.7) completes the proof. □

The above result has an interesting and at the same time intuitive interpretation from the point of view of properties of Jeffreys priors. Recall that Jeffreys priors are designed to be non-informative and represent ignorance about the parameter of interest. Weighted Likelihood Bootstrap can be characterized by an analogous property, as it does not incorporate any prior information about the parameter. In fact the question of second order equivalence between Weighted Likelihood Bootstrap and Bayesian posteriors with Jeffreys priors was initially raised in Section 4 of Newton and Raftery (1994). Due to an omission in the Taylor expansion, the authors obtained an incomplete Edgeworth expansion, which lead to incorrect conclusions about second order correctness.

We present below Corollary 3.12 which can be directly inferred from the existence of Edgeworth expansions.

Corollary 3.12. *Consider model (3.1) and suppose $\tilde{\theta}$ is drawn from Algorithm 5. Under conditions of Theorem 3.10, $n^{1/2}(\tilde{\theta} - \hat{\theta})$ converges to the normal distribution*

$$N\left\{0, J(\theta^*)^{-1}I(\theta^*)J(\theta^*)^{-1}\right\}$$

in total variation for any choice of w_0.

Corollary 3.12 applied to $w_0 = 0$ generalizes Theorem 1 of Lyddon et al. (2019), and the convergence expressed in (3.3) can be stated in total variation, as it is typically done for the Bernstein–von Mises type results.

A reader may wonder why we advocate using Posterior Bootstrap rather than sampling from the approximating normal distribution. Firstly, the Posterior Bootstrap algorithm allows us to incorporate the prior information. What is more, it does not require the user to approximate matrices $I(\theta^*)$ and $J(\theta^*)$, which may be difficult for some models. We show

however in Section 2 .3 that approximating these matrices may be useful for setting w_0. Finally, Proposition 3.11 shows that is some cases Weighted Likelihood Bootstrap leads to second order correctness. In particular, this means that in the scenario of Proposition 3.11, Weighted Likelihood Bootstrap is able to capture correctly the skewness of the posterior distribution.

Corollary 3.13 results from the fact the Algorithm 5 provides the same first-order approximation as Weighted Likelihood Bootstrap, regardless of the choice of w_0. The proofs of Proposition 3.1 and Theorem 3.2 can therefore be easily extended to Posterior Bootstrap.

Corollary 3.13. *The statements of Proposition 3.1 and Theorem 3.2 remain true when we replace Weighted Likelihood Bootstrap with Algorithm 5, or Algorithm 7 introduced in Section 4 , for any choice of hyperparameters.*

2 .3 Choice of w_0 in Algorithm 5

We start by providing some intuition why, if the user simply sets $w_0 = 1$ under misspecification, their prior beliefs on the parameter may not be appropriately reflected in the inference. To this end, we consider a one-dimensional toy model

$$x_1, \ldots, x_n \sim N(\theta, 1), \quad \theta \sim \mathrm{N}(\mu_0, \sigma_0^2). \tag{3.14}$$

Suppose model (3.14) is misspecified and in fact $x_1, \ldots, x_n \sim N(0, \sigma^2)$ for $\sigma \neq 1$. The correct model would then be

$$x_1, \ldots, x_n \sim N(\theta, \sigma^2), \quad \theta \sim \mathrm{N}(\mu_0, \sigma_0^2). \tag{3.15}$$

Drawing a sample from Algorithm 5 is equivalent to computing

$$\arg\max_\theta \left\{ \sum_{i=1}^n \frac{-w_i(x_i-\theta)^2}{2} - \frac{w_0(\theta-\mu_0)^2}{2\sigma_0^2} \right\} = \arg\max_\theta \left\{ \sum_{i=1}^n \frac{-w_i(x_i-\theta)^2}{2\sigma^2} - \frac{w_0(\theta-\mu_0)^2}{2\sigma_0^2\sigma^2} \right\}. \tag{3.16}$$

If the user knew the true data generating mechanism, they would instead compute

$$\arg\max_\theta \left\{ \sum_{i=1}^n -\frac{w_i(x_i-\theta)^2}{2\sigma^2} - \frac{(\theta-\mu_0)^2}{2\sigma_0^2} \right\}. \tag{3.17}$$

Equation (3.16) is equivalent to (3.17) if $w_0 = \sigma^2$. In fact drawing samples according to (3.16) with $w_0 = 1$ would be equivalent to using the correct likelihood but having the prior belief encoded by $\theta \sim N(\mu_0, \sigma^2\sigma_0^2)$ rather than $\theta \sim N(\mu_0, \sigma_0^2)$.

We now make a connection between the above observation and the Edgeworth expansion obtained in Theorem 3.10, which will further guide our choice of w_0. Let $J(\theta^*)$ and $I(\theta^*)$ be the information matrices for model (3.14) at θ^*, that is $J(\theta^*) = 1$ and $I(\theta^*) = \sigma^2$. Let $I_{\text{true}}(\theta^*)$ be the information matrix at θ^* under the correct model (3.15). We thus have $I_{\text{true}}(\theta^*) = \sigma^{-2}$. In the second order expansion of the Bayesian posterior of $\sigma^{-1}n^{1/2}\left(\theta - \hat{\theta}\right)$ under the correct model, as shown in (3.7), the prior-dependent term is

$$n^{-1/2} I_{\text{true}}(\theta^*)^{-1/2} \pi'(\theta^*)/\pi(\theta^*) = n^{-1/2}\sigma\pi'(\theta^*)/\pi(\theta^*).$$

On the other hand, since $I(\theta^*)^{-1/2}J(\theta^*) = \sigma^{-1}$, by Theorem 3.10 we get that under model (3.14) and Algorithm 5 with $w_0 = \sigma^2$, the prior-dependent term in the second order Edgeworth expansion of $\sigma^{-1}n^{1/2}\left(\tilde{\theta} - \hat{\theta}\right)$ is

$$n^{-1/2} I(\theta^*)^{-1/2} w_0 \pi'(\theta^*)/\pi(\theta^*) = n^{-1/2}\sigma\pi'(\theta^*)/\pi(\theta^*).$$

To summarize, in this special case the prior-dependent term in the Edgeworth expansion of Posterior Bootstrap under the misspecified model with $w_0 = \sigma^2$ is the same as the prior-dependent term in Bayesian inference under the correct model.

The example described above, as well as the fact that Posterior Bootstrap corrects the covariance matrix of the asymptotic distribution, motivates the following methodology of setting w_0, for potentially multivariate models. Suppose the prior π satisfies (3.8) so that w_0 can be treated as a d-dimensional vector. The goal is to match the prior-dependent terms in Edgeworth expansion for Algorithm 5 and in the Laplace approximation for the Bayesian posterior with a corrected scaling $I(\theta^*)^{-1/2}J(\theta^*)$, rather than $I(\theta^*)^{1/2}$. Thus we want to set w_0 so that for all $y \in \mathbb{R}^d$ it satisfies

$$y^T I(\theta^*)^{-1/2} \begin{pmatrix} w_{0,1}\frac{\pi'(\theta_1^*)}{\pi(\theta_1^*)} \\ \vdots \\ w_{0,d}\frac{\pi'(\theta_d^*)}{\pi(\theta_d^*)} \end{pmatrix} = y^T \left\{ I(\theta^*)^{-1/2} J(\theta^*) \right\}^{-1} \begin{pmatrix} \frac{\pi'(\theta_1^*)}{\pi(\theta_1^*)} \\ \vdots \\ \frac{\pi'(\theta_d^*)}{\pi(\theta_d^*)} \end{pmatrix},$$

which is equivalent to

$$D(w_0)\begin{pmatrix}\frac{\pi'(\theta_1^*)}{\pi(\theta_1^*)}\\ \vdots \\ \frac{\pi'(\theta_d^*)}{\pi(\theta_d^*)}\end{pmatrix} = I(\theta^*)^{1/2}J(\theta^*)^{-1}I(\theta^*)^{1/2}\begin{pmatrix}\frac{\pi'(\theta_1^*)}{\pi(\theta_1^*)}\\ \vdots \\ \frac{\pi'(\theta_d^*)}{\pi(\theta_d^*)}\end{pmatrix}, \tag{3.18}$$

where $D(w_0)$ is a diagonal matrix with entries $w_{0,1}, \ldots, w_{0,d}$ on the diagonal. Note that for well-specified models formula (3.18) above would imply that $w_0 = (1, \ldots, 1)$, as expected. When $I(\theta^*)^{1/2}J(\theta^*)^{-1}I(\theta^*)^{1/2}$ is not diagonal, it is impossible to set w_0 so that (3.18) is satisfied. We therefore set w_0 as

$$w_0^* = \arg\min_{w_0} \|D(w_0) - I(\theta^*)^{1/2}J(\theta^*)^{-1}I(\theta^*)^{1/2}\|_F = \operatorname{diag}\left\{I(\theta^*)^{1/2}J(\theta^*)^{-1}I(\theta^*)^{1/2}\right\}, \tag{3.19}$$

where $\|\cdot\|_F$ denotes the Frobenius norm. Under Assumption 3.7 matrix $I(\theta^*)^{1/2}J(\theta^*)^{-1}I(\theta^*)^{1/2}$ is positive definite, which ensures that all entries of w_0 are positive. In practice, we approximate this formula with its empirical version

$$w_0^* = \operatorname{diag}\left(I_n^{1/2}J_n^{-1}I_n^{1/2}\right). \tag{3.20}$$

Since $I_n^{1/2}J_n^{-1}I_n^{1/2}$ converges in probability to $I(\theta^*)^{1/2}J(\theta^*)^{-1}I(\theta^*)^{1/2}$, using the empirical approximation of w_0^* does not affect the second order Edgeworth expansion obtained in (3.10). If the prior does not factorize, then we must constrain the diagonal elements of $D(w_0)$ in (3.18) to be equal. Then minimizing the Frobenius norm in (3.19) yields

$$\bar{w}_0^* = \frac{1}{d}\operatorname{tr}\left(I_n^{1/2}J_n^{-1}I_n^{1/2}\right). \tag{3.21}$$

In the case of factorizing priors, we advocate using formula (3.20) rather than (3.21) as it allows for capturing different kinds of misspecification, such as overdispersion or underdispersion, in different coordinates.

The above approach offers an automatic and principled way of setting w_0 so that the prior information is appropriately incorporated in the inference. It also has a natural extension to hierarchical models, as we demonstrate in Section 3 . We note, however, that the method aims to reflect faithfully the user's prior belief about the parameter. Thus, by its design the method will not be optimal, for example, when the user is interested in prediction, but

their prior beliefs are poorly supported by the data-generating mechanism. An alternative strategy would then be performing cross-validation.

3 Extending Posterior Bootstrap to hierarchical models

3 .1 Fixed number of groups K

We consider the following potentially misspecified hierarchical model.

$$\begin{aligned} x_{k,1}, \dots x_{k,n_k} &\sim f_k(x_k|\theta_k), \quad k = 1, \dots, K, \\ \theta_k &\sim g_k(\theta|\lambda), \qquad k = 1, \dots, K, \\ \lambda &\sim p(\lambda). \end{aligned} \tag{3.22}$$

Each group k has a sample size n_k for $k = 1, \dots, K$. We define $x_{1:K,1:n_k}$ as a vector of all observations $x_{k,j}$ for $j = 1, \dots, n_k$ and $k = 1, \dots, K$. The marginal prior on $\theta = (\theta_1, \dots, \theta_K)$ is

$$\tilde{\pi}(\theta) = \int \left\{ \prod_{k=1}^{K} g_k(\theta_k|\lambda) p(\lambda) \right\} \mathrm{d}\lambda. \tag{3.23}$$

If the integral in (3.23) is tractable, then Algorithm 5 can be applied to model (3.22), with the joint prior $\tilde{\pi}(\theta)$ on θ.

When this integral cannot be computed, we propose instead to use Algorithm 6, our Posterior Bootstrap method for hierarchical models. We let $P^{HB,p}(\cdot|x_{1:K,1:n_k})$ denote the distribution of $(\tilde{\theta}_1, \dots, \tilde{\theta}_K, \tilde{\lambda})$ under Algorithm 6 and by $p^{HB,p}(\cdot|x_{1:K,1:n_k})$ we denote its density. The prior information in Algorithm 6 is introduced via penalization, in the spirit of Algorithm 5. For each $k = 1, \dots, K$ parameter $w_{0,[k]}$ plays for data $x_{k,1}, \dots, x_{k,n_k}$ the role of w_0 used in Algorithm 5. The fact that the penalization term depends on $\bar{\lambda}$, which is kept the same across groups, induces correlation between $\theta_1, \dots, \theta_K$, in a similar way to what happens under standard Bayesian inference.

It is worth emphasizing here an additional advantage of using Algorithm 6 for model (3.22), even if the integral in (3.23) is tractable. Since the prior $\tilde{\pi}$ does not factorize, Algorithm 5 would need to be used with w_0 treated as a real number, whereas Algorithm 6 makes use of the conditional independence structure between $\theta_1, \dots, \theta_K$ given λ, and allows the user to set $w_{0,[k]}$ separately for each group. Arguments outlined in Section 2 .3 show that this approach accounts better for potential misspecification of $f_1, \dots, f_K$.

Algorithm 6 Posterior Bootstrap for hierarchical models via prior penalization, returns samples from $P^{HB,p}(\cdot|x_{1:K,1:n_k})$

hyperparameters: $w_{0,[1]}, \ldots, w_{0,[K]}$.
for k = 1, ..., K **do**
    Set $\hat{\theta}_k = \arg\max_{\theta_k} \sum_{i=1}^{n_k} \log f_k(x_{k,i}|\theta_k)$.
end for
for j = 1, ..., N **do** ▷ (in parallel)
    Sample $\bar{\lambda}^{(j)} \sim p\left(\lambda|\hat{\theta}_1, \ldots, \hat{\theta}_K\right)$.
    for k = 1, ..., K **do**
        Draw $w_1^{(j)}, \ldots, w_{n_k}^{(j)} \sim \text{Exp}(1)$.
        Set $\tilde{\theta}_k^{(j)} = \arg\max_{\theta_k} \left\{\sum_{i=1}^{n_k} w_i^{(j)} \log f_k(x_{k,i}|\theta_k) + w_{0,[k]}^T \log g_k\left(\theta_k|\bar{\lambda}^{(j)}\right)\right\}$.
    end for
    Sample $\tilde{\lambda}^{(j)} \sim p\left(\lambda|\tilde{\theta}_1^{(j)}, \ldots, \tilde{\theta}_K^{(j)}\right)$.
end for
return Samples $\left\{\left(\tilde{\lambda}^{(j)}, \tilde{\theta}_1^{(j)}, \ldots, \tilde{\theta}_K^{(j)}\right)\right\}_{j=1}^N$.

We let $\theta_k^* = \arg\max_\theta \mathbb{E}_{p_k^*}\{\log f_k(X|\theta)\}$, where $x_{k,1}, \ldots, x_{k,n_k}$ are generated from P_k^*. We treat here $\theta_1^*, \ldots, \theta_K^*$ as deterministic values, see Section 3 .2 for a scenario where $\theta_1^*, \ldots, \theta_K^*$ are themselves random variables. Firstly, as a direct consequence of Theorem 3.10, we get the first order approximation for the asymptotic distribution of $\tilde{\theta}_k$. For any Borel set $A \subseteq \Theta$ we have

$$P^{HB,p}\left\{n^{1/2}\left(\tilde{\theta}_k - \hat{\theta}_k\right) \in A|x_{1:K,1:n_k}\right\} \to \text{pr}\left(Z \in A\right), \quad Z \sim N\left\{0, J_k(\theta_k^*)^{-1} I_k(\theta_k^*) J_k(\theta_k^*)^{-1}\right\},$$

where I_k and J_k are given by (3.4), computed for f_k.

We explain below why $\bar{\lambda}$ in Algorithm 6 is drawn from the posterior $p\left(\lambda|\hat{\theta}_1, \ldots, \hat{\theta}_K\right)$ rather then from the prior $p(\lambda)$, even though the latter may seem more intuitive. To this end, let us analyse the impact of the prior on the second order Edgeworth expansion on θ_k. By construction of Algorithm 6 and by Theorem 3.10, if $n_1, \ldots, n_k$ are large, the prior-dependent term is

$$n_k^{-1/2} I_k(\theta_k^*)^{-1/2} \left[w_{0,[k]} \cdot \int \nabla_{\theta_k}\left\{\log g_k(\theta_k^*|\lambda)\right\} p(\lambda|\theta_1^*, \ldots, \theta_K^*) \text{d}\lambda\right]. \tag{3.24}$$

On the other hand, since in fact the prior on θ has a form (3.23), we would want the prior

impact term to be given by $n_k^{-1/2} I_k(\theta_k^*)^{-1/2} \left[w_{0,[k]} \cdot \nabla_{\theta_k} \{\log \tilde{\pi}(\theta^*)\} \right]$. We observe that

$$\begin{aligned}
\nabla_{\theta_k} \log \tilde{\pi}(\theta) =& \frac{\nabla_{\theta_k} \left\{ \int \prod_{j=1}^K g_k(\theta_j|\lambda) p(\lambda) \mathrm{d}\lambda \right\}}{\int \prod_{j=1}^K g_k(\theta_j|\lambda) p(\lambda) \mathrm{d}\lambda} \\
=& \frac{\int \left\{ \nabla_{\theta_k} g_k(\theta_k|\lambda) \prod_{j=1, j\neq k}^K g_k(\theta_j|\lambda) p(\lambda) \right\} \mathrm{d}\lambda}{\int \prod_{j=1}^K g_k(\theta_j|\lambda) p(\lambda) \mathrm{d}\lambda} \\
=& \int \left\{ \frac{\nabla_{\theta_k} g_k(\theta_k|\lambda)}{g_k(\theta_k|\lambda)} \frac{\prod_{j=1}^K g_k(\theta_j|\lambda) p(\lambda)}{\int \prod_{j=1}^K g_k(\theta_j|\lambda) p(\lambda) \mathrm{d}\lambda} \right\} \mathrm{d}\lambda \\
=& \int \nabla_{\theta_k} \{\log g_k(\theta_k|\lambda)\} \, p(\lambda|\theta_1, \ldots, \theta_K) \mathrm{d}\lambda,
\end{aligned}$$

so $n_k^{-1/2} I_k(\theta_k^*)^{-1/2} \left[w_{0,[k]} \cdot \nabla_{\theta_k} \{\log \tilde{\pi}(\theta^*)\} \right]$ is indeed equal to (3.24). We therefore conclude that if model (3.22) is well-specified, the impact of the prior is asymptotically the same under Bayesian inference and under Algorithm 6. Furthermore, this reasoning justifies setting $w_{0,[1]}, \ldots, w_{0,[K]}$ using rules outlined in Section 2 .3 for w_0 in Algorithm 5.

Algorithm 6 bears a resemblance to Bayesian multiple imputation (Little, 1992) or modular inference (Liu et al., 2009; Jacob et al., 2017; Carmona and Nicholls, 2020), as $\bar{\lambda}$ is sampled first to be plugged into the objective function used for sampling $\tilde{\theta}_1, \ldots, \tilde{\theta}_K$. Finally, drawing $\tilde{\lambda}$ involves using $\tilde{\theta}_1, \ldots, \tilde{\theta}_K$ sampled earlier.

When presenting Algorithm 6, we assumed implicitly that the user can sample from the posterior $p(\lambda|\theta_1, \ldots, \theta_K)$. In conjugate examples one can draw λ directly, otherwise it would be necessary to employ a Markov chain Monte Carlo sampler for this step. We let $n = \min_{k=1,\ldots,K} n_k$. Proposition 3.14 demonstrates the asymptotic properties of $\tilde{\lambda}$ drawn from Algorithm 6, covering both the well-specified and misspecified cases. It shows that as $n \to \infty$, the distribution on $\tilde{\lambda}$ will tend to the same distribution as the Bayesian posterior on λ. A similar result was proved by Petrone et al. (2014) for marginal maximum likelihood estimators.

Proposition 3.14. *Consider model (3.22) with a fixed number of groups K, and suppose $\tilde{\lambda}$ is drawn from Algorithm 6. Suppose that Assumptions 3.4 – 3.7 hold for each likelihood function f_k and that Assumption 3.9 is satisfied for g_k, $k = 1, \ldots, K$. Assume also that for each λ the density $p(\lambda|\theta)$ is a continuous function of θ and that $p(\lambda|\theta)$ is jointly bounded as a function of θ and λ. Then as $n \to \infty$, we have the following convergence of densities*

$$p^{HB,p}\left(\tilde{\lambda}|x_{1:K,1:n_k}\right) \to p\left(\tilde{\lambda}|\theta^*\right), \tag{3.25}$$

with probability going to 1 with respect to $x_{1:K,1:n_k}$.

We present the proof in Supplementary Material C .1.

3 .2 Number of groups K going to infinity

We consider model (3.22) with $f_k(\cdot|\theta) = f(\cdot|\theta)$ and $g_k(\cdot|\lambda) = g(\cdot|\lambda)$ for all θ, λ and $k = 1, \ldots, K$. An alternative to sampling $\tilde{\lambda}$ using Markov chain Monte Carlo would be performing another round of Posterior Bootstrap, where $\tilde{\theta}_1, \ldots, \tilde{\theta}_K$ is used as data. The caveat is, however, that since only asymptotic results are known for Posterior Bootstrap, for a fixed K this strategy would lack a theoretical underpinning. We will assume instead that both the number of classes K, as well as the number of observations n_k in each class, go to infinity. The penalization strategy then yields

$$\tilde{\lambda} = \arg\max_{\lambda} \left\{ \sum_{k=1}^{K} v_k \log g(\tilde{\theta}_k|\lambda) + w_g^T \log p\,(\lambda) \right\}, \quad v_1, \ldots, v_K \sim \mathrm{Exp}(1). \tag{3.26}$$

We now describe the assumed data-generating mechanism of $x_{1:K,1:n_k}$ under this scenario. Let Q^* be a distribution on a space of distributions on $\mathbb{R}^s$, where s is the dimension of a single observation $x_{k,i}$ for $k = 1, \ldots, K, i = 1, \ldots, n_k$. We have

$$\begin{aligned} x_{k,1}, \ldots x_{k,n_k} &\sim P_k^*, \quad k = 1, \ldots, K, \\ P_1^*, \ldots, P_K^* &\sim Q^*. \end{aligned} \tag{3.27}$$

We use our standard notation for pseudo-true parameters θ_k^*, that is,

$$\theta_k^* = \arg\max_{\theta_k} E_{p_k^*} \left\{ \log f(X|\theta_k) \right\}.$$

Moreover, following the convention of denoting random variables with capital letters, we denote by Θ^* the random variable such that $\theta_1^*, \ldots, \theta_K^*$ are distributed according to Θ^*. Note that $\theta_1^*, \ldots, \theta_K^*$ depend only on Q^*. We additionally define $\lambda^* = \arg\max_\lambda E_{Q^*} \left\{ \log g(\Theta^*|\lambda) \right\}$ and $\hat{\lambda}_{\theta^*} = \arg\max_\lambda \left\{ \sum_{k=1}^{K} \log g(\theta_k^*|\lambda) \right\}$. The information matrices for λ are

$$I_g(\lambda) = E_{Q^*} \left[\left\{ \nabla_\lambda \log g(\Theta^*|\lambda) \right\} \left\{ \nabla_\lambda \log g(\Theta^*|\lambda) \right\}^T \right], \quad J_g(\lambda) = -E_{Q^*} \left[D_\lambda^2 \left\{ \log g(\Theta^*|\lambda) \right\} \right]. \tag{3.28}$$

Recall that $n = \min_{k=1,\ldots,K} n_k$. Theorem 3.15 shows that if $\tilde{\lambda}$ is drawn from (3.26) and

the rate at which $K \to \infty$ is slow enough compared with the rate at which $n \to \infty$, the distribution of $\tilde{\lambda}$ is asymptotically normal with the sandwich covariance matrix. We can therefore conclude that this algorithm can correct for the potential misspecification of $g(\theta|\lambda)$.

Theorem 3.15. *Suppose that the data generating mechanism follows* (3.27)*, and assume that* $\tilde{\lambda}$ *is drawn from* (3.26)*. Assume that* $K \to \infty$ *at a rate such that* $\lim_{K\to\infty} Kn^{-1/2} \log n = 0$*. Then under regularity conditions stated explicitly in Supplementary Material C .2 we have*

$$P^{HB,pp}\left\{K^{1/2}\left(\tilde{\lambda} - \hat{\lambda}_{\theta^*}\right) \in A | x_{1:K,1:n_k}\right\} \to \text{pr}\left(Z \in A\right), \quad Z \sim N\left\{0, J_g(\lambda^*)^{-1} I_g(\lambda^*) J_g(\lambda^*)^{-1}\right\},$$

with probability going to 1 with respect to $x_{1:K,1:n_k}$ *as* $n \to \infty$*, for matrices* I_g *and* J_g *defined in* (3.28)*.*

The regularity conditions used in Theorem 3.15 are analogues of Assumptions 3.4 – 3.9 for functions $\log f(x|\theta)$ and $\log g(\theta|\lambda)$. We additionally require $\log g(\theta|\lambda)$ to be sufficiently smooth both with respect to λ and θ. The proof of Theorem 3.15 is reported in Supplementary Material C .2.

Careful examination of the proofs of Proposition 3.14 and Theorem 3.15 shows that the same results would hold if we replaced $\tilde{\lambda}$ with $\bar{\lambda}$, where $\bar{\lambda}$ is the intermediate value of λ used to generate $\tilde{\theta}_1, \ldots, \tilde{\theta}_K$, see Algorithm 6. If the parameter of interest is λ, we however advocate using $\tilde{\lambda}$ for inference, as it gives more accurate results. We illustrate the difference between $\bar{\lambda}$ and $\tilde{\lambda}$ on a synthetic dataset in Supplementary Material D .3. On the other hand, if only parameters $\theta_1, \ldots, \theta_K$ are of interest, the user can skip the step of drawing $\tilde{\lambda}$.

To summarize, we have developed an algorithm for model (3.22) that demonstrates good asymptotic properties, both for inference on $\theta_1, \ldots, \theta_K$ and on λ, while being robust to misspecification. Furthermore, even when the model is well-specified, the fact that Algorithm 6 can be fully parallelized makes it an attractive alternative to Metropolis-within-Gibbs samplers typically used for (3.22).

4 Posterior Bootstrap via pseudo-samples from the prior predictive

4 .1 Methodology

Using Algorithm 5 is possible for most standard priors. However, combining certain prior distributions with Algorithm 5 would have adverse consequences for the inference, or could even be impossible from a computational point of view. Firstly, formulation of Algorithm 5 necessitates that evaluating $\log \pi(\theta)$ is possible for every θ, which is not the case, for example, for the original spike-and-slab prior (Mitchell and Beauchamp, 1988). Moreover, there is a widely used group of priors that are infinite on the boundary of the set, for example the Gamma distribution $\Gamma(\alpha, \beta)$ with $\alpha < 1$, Dirichlet prior $\mathcal{D}(\alpha_1, \cdots, \alpha_k)$ with some of the $\alpha_j < 1$ as advocated for instance by Rousseau and Mengersen (2011) for mixture models. For these priors Algorithm 5 would always return the boundary of the set, regardless of the data.

To overcome these drawbacks, we consider another version of Posterior Bootstrap, proposed originally by Fong et al. (2019). We summarize it in Algorithm 7. The prior information is incorporated into the algorithms via drawing so-called pseudo-samples from the prior predictive.

Algorithm 7 Posterior Bootstrap via pseudo-samples from the prior predictive

1: **hyperparameters**: c, T.
2: **for** j $= 1, \ldots,$ N **do** ▷ (in parallel)
3: Draw $\bar{\theta}_1^{(j)}, \ldots, \bar{\theta}_T^{(j)} \sim \pi(\theta)$.
4: Generate prior pseudo-samples $x_{n+i}^{(j)} \sim f\left(\cdot|\bar{\theta}_i^{(j)}\right)$ for $i = 1, \ldots, T$.
5: Draw $w_1^{(j)}, \ldots, w_n^{(j)} \sim \text{Exp}(1)$ and $w_{n+1}^{(j)}, \ldots, w_{n+T}^{(j)} \sim \text{Exp}(T/c)$.
6: Set $\tilde{\theta}^{(j)} = \arg\max_\theta \left\{\sum_{i=1}^n w_i^{(j)} \log f(x_i|\theta) + \sum_{i=1}^T w_{n+i}^{(j)} \log f\left(x_{n+i}^{(j)}|\theta\right)\right\}$.
7: **end for**
8: **return** Samples $\{\tilde{\theta}^{(j)}\}_{j=1}^N$.

To obtain the Edgeworth expansion of Algorithm 7, we no longer need Assumption 3.9 to hold so replace it with a milder condition. To this end we define the prior predictive density as $f_\pi(x) = \int f(x|\theta)\pi(\theta)\mathrm{d}\theta$.

Assumption 3.16. The quantity $\alpha_{\theta,\pi} = \int \nabla_\theta \{\log f(x|\theta)\} f_\pi(x)\mathrm{d}x$ is continuous as a function of $\theta \in B$.

Theorem 3.17. *Consider model* (3.1) *with a parameter of interest* $\theta \in \mathbb{R}^d$ *and suppose Assumptions 3.4 – 3.8, 3.16 hold. As* $n \to \infty$ *with probability going to 1 the Edgeworth expansion of the density of* $I_n^{-1/2} J_n n^{1/2}(\tilde{\theta} - \hat{\theta})$ *for* $\tilde{\theta}$ *drawn from Algorithm 7 is given by:*

$$\phi(y)\left\{1 + cy^T I(\theta^*)^{-1/2}\frac{\alpha_{\theta^*,\pi}}{n^{1/2}} + \frac{P_{3,n}(y)}{n^{1/2}}\right\} + o\left(n^{-1/2}\right) \tag{3.29}$$

for a constant $\alpha_{\theta^*,\pi}$ *defined like in Assumption 3.16 for* $\theta = \theta^*$*, and for* $P_{3,n}$ *defined in Theorem 3.10.*

The proof of Theorem 3.17 follows analogous steps to the proof of Theorem 3.10, for details see Supplementary Material B .5.

Since the prior information enters into the second order Edgeworth expansion via an intractable constant $\alpha_{\theta^*,\pi}$, in this case the impact of the prior has a slightly more cumbersome interpretation than in the case of Algorithm 5. However, to shed some light on possible interpretations of this constant, note that $\alpha_{\theta^*,\pi}$ is an integral of $\nabla_\theta \{\log f(x|\theta^*)\}$ with respect to the prior predictive $\int f(x|\theta)\pi(\theta)\mathrm{d}\theta$ and can also be written as

$$\alpha_{\theta^*,\pi} = \int \left[\nabla_\theta \{\log f(x|\theta^*)\}\{f_\pi(x) - f(x|\theta^*)\}\right] \mathrm{d}x.$$

since $\int \nabla_\theta \{\log f(x|\theta^*)\} f(x|\theta^*)dx = 0$. Therefore $\alpha_{\theta^*,\pi}$ measures the difference between the prior predictive and the predictive distribution under θ^*.

The form of $\alpha_{\theta^*,\pi}$ makes it difficult to provide detailed general guidelines on the appropriate choice of c. For well-specified models c should be interpreted as the effective sample size of the prior, which we make explicit in Section 4 .2 for conjugate priors in exponential families. Importantly, as opposed to the choice of w_0 in Algorithm 5, the optimal choice of c depends both on the likelihood and the prior. Parameter T does not appear in the second order Edgeworth expansion and we do not expect it to have much effect on the results.

Remark 3.18. Posterior Bootstrap via pseudo-samples can also be used for hierarchical models (3.22) within Algorithm 6. The procedure would involve drawing $\bar{\lambda}$ from $p(\lambda|\hat{\theta}_1, \ldots, \hat{\theta}_K)$, then drawing pseudo-samples given $\bar{\lambda}$ to finally compute $\tilde{\theta}_1, \ldots, \tilde{\theta}_K$. Using the same value of $\bar{\lambda}$ across all K groups induces correlation between $\tilde{\theta}_1, \ldots, \tilde{\theta}_K$. Inference on $\tilde{\lambda}$ given $\tilde{\theta}_1, \ldots \tilde{\theta}_K$ can then be performed using methods described in Section 3 .2, and its theoretical properties would be analogous to those presented in Proposition 3.14 and Theorem 3.15.

As mentioned in Section 1 .2, another method of incorporating prior information is Algorithm 1 of Lyddon et al. (2018). It is similar to Algorithm 7, but instead of drawing $\bar{\theta}$ from the prior $\pi(\theta)$, $\bar{\theta}$ is drawn from the Bayesian posterior $\pi(\theta|x_{1:n})$. Nevertheless, it turns out that the impact of the prior in that algorithm is negligible, as the prior information does not enter into the second order Edgeworth expansion.

To see this, note that the second order Edgeworth expansion for Algorithm 1 of Lyddon et al. (2018) can be obtained using analogous calculations to those required to prove Theorem 3.17. The expansion would thus be given by (3.29), where instead of the constant $\alpha_{\theta^*,\pi}$ we would have

$$\beta_{\theta^*,\pi} = \int \nabla_\theta \left\{\log f(x|\theta^*)\right\} \left\{\int f(x|\theta)\pi(\theta|x_{1:n})\mathrm{d}\theta\right\} \mathrm{d}x.$$

Since the posterior $\pi(\theta|x_{1:n})$ concentrates around θ^* as $n \to \infty$, we have $\beta_{\theta^*,\pi} = o_P(1)$ and therefore the prior information does not enter into the second order Edgeworth expansion. Details are provided in Supplementary Material B .6.

4 .2 Choice of c in Algorithm 7

We consider a special case of regular exponential families in a canonical form

$$f(x|\theta) = h(x) \exp\left\{T(x)^T\theta - A(\theta)\right\}.$$

The corresponding conjugate prior on θ is

$$\pi(\theta) = \pi(\theta|\tau, n_0) = H(\tau, n_0) \exp\left\{\tau^T\theta - n_0 A(\theta)\right\}. \tag{3.30}$$

for some hyperparameters τ and n_0. The resulting posterior follows $\pi(\theta|\tau + \sum_{i=1}^n T(x_i), n_0 + n)$. We can therefore interpret n_0 as the number of pseudo-samples quantifying the amount of information in the prior, called also effective sample size.

In this special case we can compute the expression $c\alpha_{\theta^*,\pi}$ appearing in the Edgeworth expansion for Algorithm 7. We have

$$\alpha_{\theta^*,\pi} = \int \{T(x) - \nabla_\theta A(\theta^*)\} \int f(x|\theta)\pi(\theta)\mathrm{d}\theta dx = \int \nabla_\theta A(\theta)\pi(\theta)\mathrm{d}\theta - \nabla_\theta A(\theta^*).$$

where the last equality follows from the property of exponential families that $\mathbb{E}_\theta \{T(X)\} = \nabla_\theta A(\theta)$. Since $\int \nabla_\theta \pi(\theta) \mathrm{d}\theta = 0$, as shown by Diaconis and Ylvisaker (1979), we get that

$$\tau - n_0 \int \nabla_\theta A(\theta) \pi(\theta) \mathrm{d}\theta = \int \{\tau - n_0 \nabla_\theta A(\theta)\} \pi(\theta) \mathrm{d}\theta = \int \nabla_\theta \pi(\theta) \mathrm{d}\theta = 0.$$

Thus $\alpha_{\theta^*,\pi} = \tau/n_0 - \nabla_\theta A(\theta^*)$. Recall that when we instead use Algorithm 5, the corresponding term in the Edgeworth expansion is given by $w_0^T \nabla_\theta \log \pi(\theta^*)$, which in the case of the prior (3.30) is equal to $w_0^T \{\tau - n_0 A(\theta^*)\}$. Therefore in this special case setting $c = n_0$ in Algorithm 7 yields the same second order Edgeworth expansion as setting $w_0 = 1$ in Algorithm 5. The above reasoning confirms our intuition that for well-specified models we should set parameter c to the effective sample size of the prior.

In fact using Posterior Bootstrap for well-specified exponential models with conjugate priors seems rather contrived, as one can in this case sample directly from the Bayesian posterior. However, under the misspecified scenario Posterior Bootstrap is more robust than Bayesian inference. In that case we propose to set the optimal c as

$$c^* = \bar{w}_0^* n_0. \tag{3.31}$$

for $\bar{w}_0^*$ defined in (3.21).

5 Illustrations

5 .1 Toy examples

We first analyse the following toy model

$$x_1, \ldots, x_n \sim N(\theta, 1), \quad \theta \sim \text{Gamma}(5, 3). \tag{3.32}$$

We consider three different scenarios by generating $x_1, \ldots, x_n \sim N(10, \sigma^2)$ for $n = 200$ with $\sigma^2 = 1$, $\sigma^2 = 2.8$ or $\sigma^2 = 0.6$. The correct model would then be

$$x_1, \ldots, x_n \sim N(\theta, \sigma^2), \quad \theta \sim \text{Gamma}(5, 3). \tag{3.33}$$

Recall from Section 2 .3 that w_0^* would in this case be equal to σ^2. We ran Algorithm 5 for

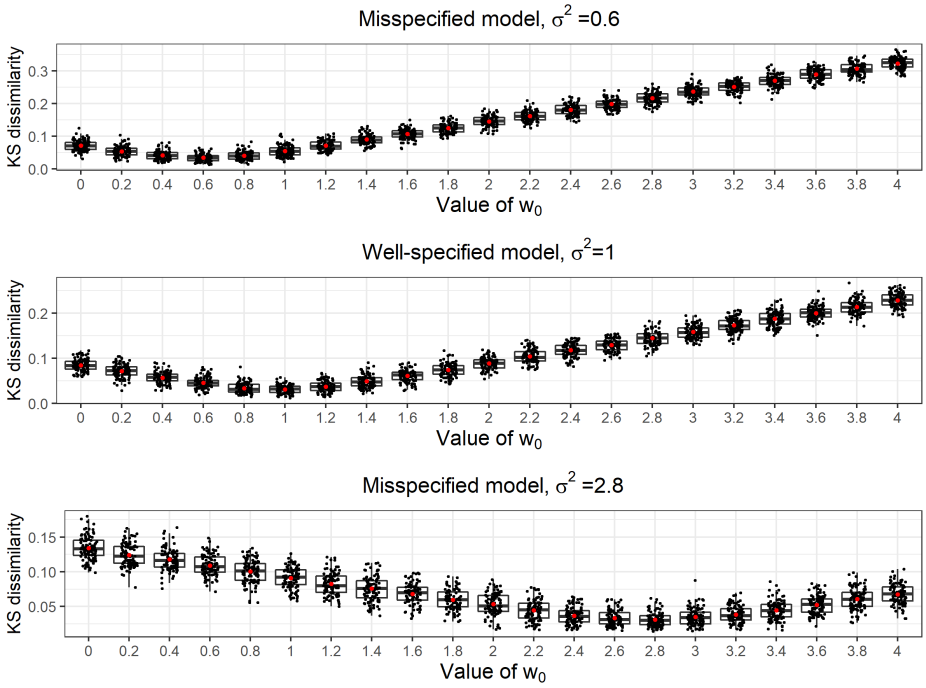


Figure 3.1: Kolmogorov–Smirnov dissimilarity between samples drawn from Posterior Bootstrap for model (3.32), and samples from the Bayesian posterior for the correct model (3.33). Three scenarios are considered, $\sigma^2 = 0.6$ (top panel), $\sigma^2 = 1$ (middle panel), and $\sigma^2 = 2.8$ (bottom panel). Posterior Bootstrap simulations were performed using Algorithm 5 for a range of values of w_0. Each experiment was repeated on 100 simulated datasets with $N = 2000$.

this model with different values of w_0 and measured the Kolmogorov–Smirnov dissimilarity between the empirical distribution of the obtained samples, and samples from the Bayesian posterior applied to the correct model. In this special case making such a comparison is justified since θ has the same interpretation of being the mean under both models, and the difference between these models lies only in the uncertainty around the mean. As shown in Figure 3.1, we indeed observe the shortest distance for $w_0 = w_0^*$. For additional plots and comparisons with other methods we refer the reader to Supplementary Material D .1.

We now analyse another toy example, with a misspecified model

$$x_1, \ldots, x_n \sim N(\theta, \Sigma_1), \quad \theta \sim N(\mu_0, \Sigma_0). \tag{3.34}$$

The correct model is

$$x_1, \ldots, x_n \sim N(\theta, \Sigma_2), \quad \theta \sim N(\mu_0, \Sigma_0). \tag{3.35}$$

We set

$$\mu_0 = (5,5), \quad \Sigma_0 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad \Sigma_1 = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 2 \end{bmatrix}, \quad \Sigma_2 = \begin{bmatrix} 0.7 & 0.6 \\ 0.6 & 3 \end{bmatrix},$$

so that the example is characterized by overdispersion of one coordinate, and underdispersion of the other. Moreover, following the discussion in Section 2 .3, in this example correlation between variables makes it impossible for (3.18) to hold. In our experiment the data $x_1, \ldots, x_n$ for $n = 200$ is generated from $N\{(0,0), \Sigma_2\}$. In Figure 3.2 we compare three versions of Posterior Bootstrap for model (3.34) with standard Bayesian posterior and two methods of robust Bayesian inference: BayesBag (Huggins and Miller, 2019) and power posteriors with the optimal power η^* as proposed by Lyddon et al. (2019). The plots show that Algorithm 5 outperforms the other methods on this example returning samples that are closer to the Bayesian posterior under the correct model (3.35). We notice that the power posterior approach governed by a single parameter η is unable to capture overdispersion of one coordinate and underdispersion of the other. BayesBag with the choice of hyperparameters advocated in Section 2.2.1 of Huggins and Miller (2019) tends to slightly overestimate the uncertainty. We also conclude that treating w_0 in Algorithm 5 as a two-dimensional vector rather than a single value helps to better incorporate the prior information.

We present an additional toy example for a hierarchical model in Supplementary Material D .2.

5 .2 Overdispersed and underdispersed Poisson regression

A common approach for modelling count data is the Poisson regression:

$$Y_i \sim \text{Poisson}(\lambda_i), \quad \lambda_i = \exp\left(\beta_0 + \beta_1 x_{1,i} + \cdots + \beta_p x_{p,i}\right), \quad i = 1, \ldots, n, \tag{3.36}$$

with a prior distribution on the vector of coefficients β. The above model assumes that $E\left(Y_i | x_i\right) = \text{var}\left(Y_i | x_i\right)$. In practice this assumption may be violated and the data exhibits either overdispersion or underdispersion relative to model (3.36). An alternative for the former case is using the negative binomial regression model, for the latter however no simple alternative is known.

The COM-Poisson regression (Shmueli et al., 2005; Sellers and Shmueli, 2010) was proposed as a generalization of the Poisson model, which can handle different levels of disper-

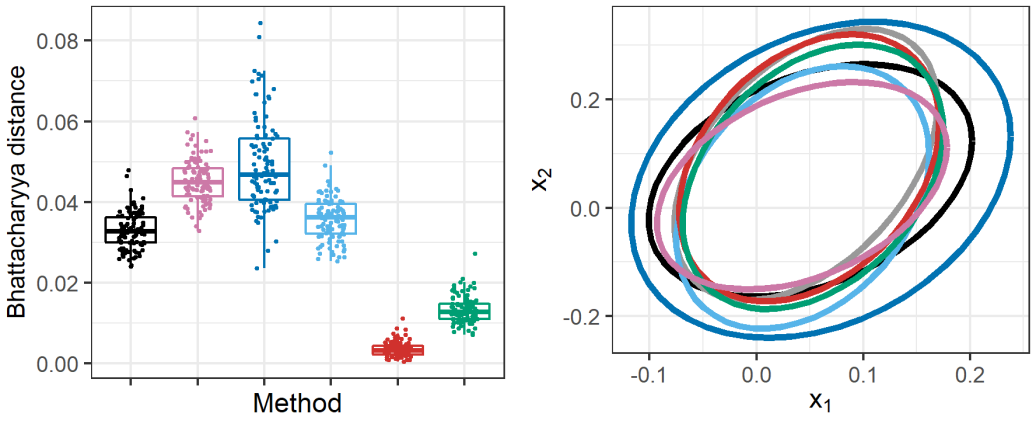


Figure 3.2: Left: Bhattacharyya distance between samples drawn from the Bayesian posterior for correct model (3.35) and standard Bayesian inference for the misspecified model (black), power posterior inference with optimal parameter η according to Lyddon et al. (2019) (pink), BayesBag (Huggins and Miller, 2019) (dark blue), Posterior Bootstrap Algorithm 7 with the optimal parameter $c^\star$ (light blue), Posterior Bootstrap Algorithm 5 with w_0 set according to (3.20) (red) and (3.21) (green). Each experiment was repeated on 100 simulated datasets with $N = 2000$ samples drawn from each method. Optimal c^* for Algorithm 7 was set as the value that empirically minimized the Bhattacharyya distance over a grid of valued. The size of each bootstrapped dataset in BayesBag was set according to Section 2.2.1 of Huggins and Miller (2019). Right: density contours for samples obtained using the above methods on a single synthetic dataset, and for samples from the Bayesian posterior for the correct model (3.35) (grey).

sion, including underdispersion. Since the COM-Poisson likelihood (Conway and Maxwell, 1962) has an intractable constant, Bayesian inference with the COM-Poisson regression requires non-standard and computationally costly Monte Carlo methods, based on a rejection sampler used within the exchange algorithm (Møller et al., 2006; Murray et al., 2012), as proposed by Chanialidis et al. (2018) and Benson and Friel (2021). Another issue with this approach is interpretability, since in the COM-Poisson model it is no longer true that $E\left(Y_i|x_i\right) = \beta_0 + \beta_1 x_{1,i} + \cdots + \beta_p x_{p,i}$.

An alternative could be to use Algorithm 5, which preserves the usual interpretability of parameter β, while being able to handle both underdispersion and overdispersion. The fact that Algorithm 5 can be fully parallelized and does not require tuning gives it a significant computational advantage over complex Monte Carlo methods proposed for the COM-Poisson regression. We illustrate the overdispersed Poisson regression with the `Articles` example from the `Rchoice` package (Sarrias, 2016) in `R`, which studies the number of articles published during scientists' Ph.D. studies. We preprocess the data following Chanialidis et al. (2018), who analysed the same example. For underdispersed Poisson models, we use as an illustration the `cottonbolls` example from the `mpcmp` package (Fung et al., 2020), modelling number of bolls produced by the cotton plants. Figure 3.3 shows that Algorithm 5 corrects

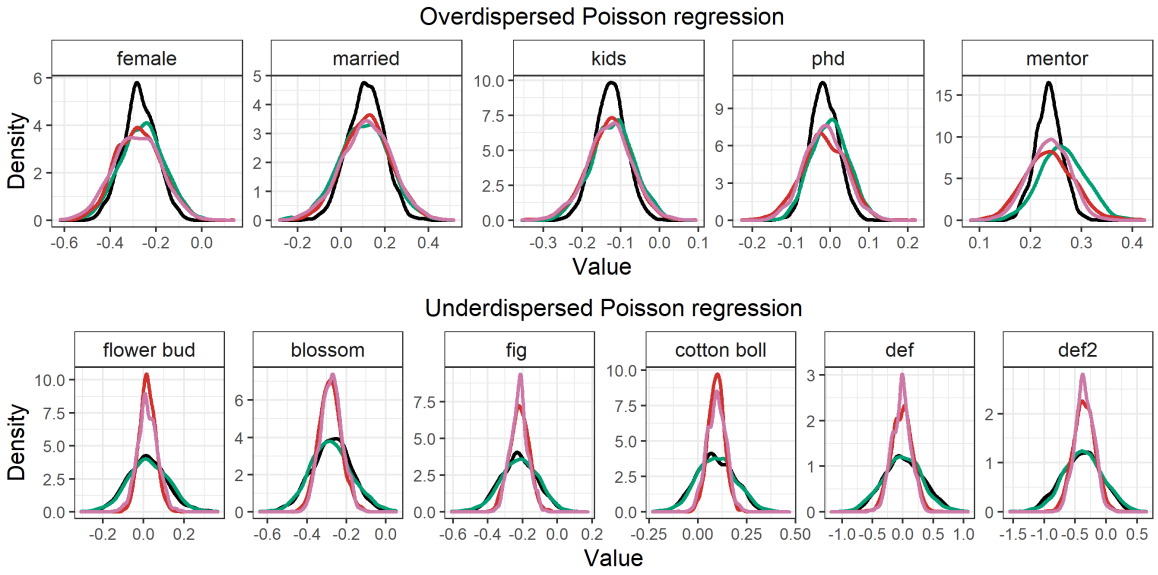


Figure 3.3: Kernel density estimates for the posterior on coefficients β in model (3.36) using four different methods: Bayesian inference for model (3.36) (black), Bayesian inference for the negative binomial regression model (green), power posterior inference with optimal parameter η according to Lyddon et al. (2019) (pink), and Algorithm 5 (red). We use the `Articles` dataset (top panel) and the `cottonbolls` dataset (bottom panel). For both examples we used the independent $N(0, 10^2)$ prior on parameters β.

both for overdispersion and underdispersion. Power posteriors also perform well on these examples, however in general this approach is less flexible in accounting for misspecification than Posterior Bootstrap.

5 .3 Dirichlet allocation model for opioid crisis data

We illustrate Posterior Bootstrap for hierarchical models, as well as the use of pseudosamples on a Dirichlet allocation model (Glynn et al., 2019) of deaths caused by drug overdoses in the United States (Ahmad et al., 2021). The data comprise 139 observations, where each observation is the numbers of deaths due to each of six drug types in a given state and year. Our primary interest lies in $\lambda = (\lambda_1, \ldots, \lambda_6)$, which describes the expected proportion of deaths in each category. For each coordinate of λ, we assume a $\mathrm{N}(0, \tau^2)$ prior truncated to the positive half-line, denoted by $\mathrm{TN}(0, \tau^2)$ below. For $m_k = (m_{k,1}, \ldots, m_{k,6})$, where $m_{k,l}$ is the deaths by drug type l in state-year k, we write the model:

$$\begin{aligned} m_k &\sim \text{Multinomial}(\theta_k, n_k), \quad && k = 1, \ldots, 139, \\ \theta_k &\sim \text{Dirichlet}(\lambda), && k = 1, \ldots, 139, \\ \lambda_l &\sim \mathrm{TN}(0, \tau^2), && l = 1, \ldots, 6. \end{aligned} \tag{3.37}$$

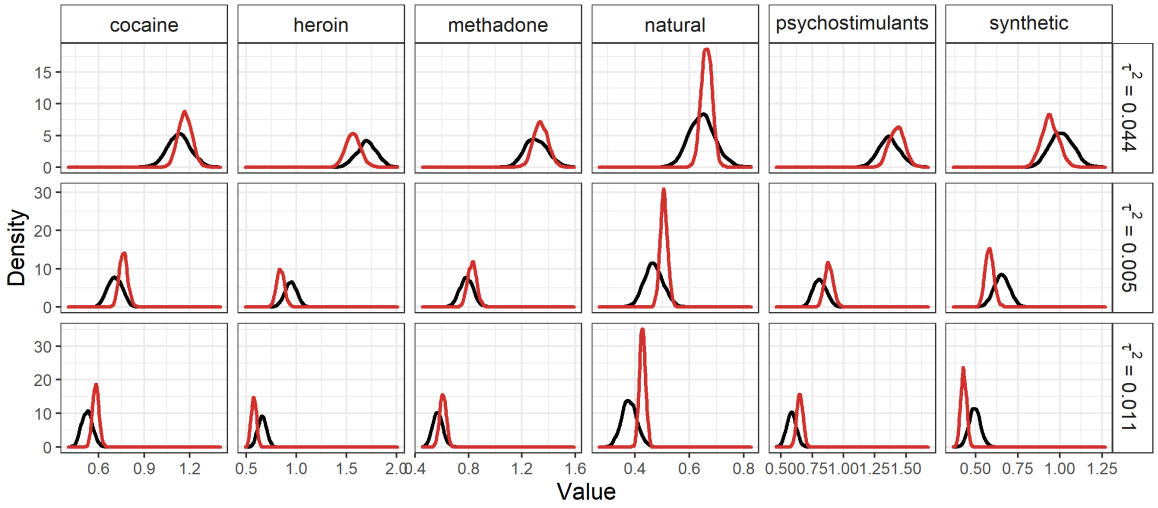


Figure 3.4: Kernel density estimates under standard Bayesian inference (black) and Posterior Bootstrap (red) for parameters λ_l in model (3.37), corresponding to six different types of drugs. We test this example on three prior distributions considered in Glynn et al. (2019).

By comparing I_g and J_g as defined in (3.28), we suspect that the Dirichlet model for θ_k given λ is misspecified. This, as well as the fact that the number of classes $K = 139$ is large motivates using (3.26) for inference on λ. As for the inference on θ_k, following the discussion in Section 4 .1, we cannot use the penalization-based approach due to the Dirichlet(λ) prior possibly with some components smaller than 1. We therefore follow the strategy based on pseudo-samples, and conjugacy of the model motivates setting c according to (3.31). We present the full resulting algorithm in Supplementary Material D .3.

Figure 3.4 displays the posterior distribution on parameter λ in (3.37) using Posterior Bootstrap and Bayesian inference performed with Metropolis-within-Gibbs Markov chain Monte Carlo. As τ decreases, the truncated normal prior concentrates on zero and we observe a shift in distributions under both methods, which confirms that our method incorporates the prior in a similar way to Bayesian inference. The results for the two methods are different, and we expect this to be due to model misspecification. In Supplementary Material D .3 we present an analogous analysis on a synthetic dataset generated from model (3.37) and show that the results delivered by these methods are then the same. In addition to correcting for model misspecification, Posterior Bootstrap runs in parallel with virtually no tuning and without the requirement for convergence diagnostics necessary in the case of Markov chain Monte Carlo.

6 Limitations of methods based on Weighted Likelihood Bootstrap

In previous sections we focused on positive aspects and properties of Posterior Bootstrap and its extensions, such as its computational advantages and its capability to correct the uncertainty in the case of model misspecification. The aim of this section is to answer some frequently asked questions about limitations of Weighted Likelihood Bootstrap and its extensions.

There are numerous situations when Posterior Bootstrap cannot be used for computational reasons, whereas standard Bayesian modelling combined with Markov chain Monte Carlo methods prove to be successful. Firstly, by construction of the algorithms it is implicitly assumed that one can numerically find a maximum of the randomized log likelihood function, which is typically done using gradient-based methods. However, if the parameter of interest live in a discrete space, or a mix of discrete and continuous state spaces, this will not be possible. There exists a vast literature on Gibbs samplers that deal effectively with such models, for example (Richardson and Green, 1997; Titsias and Yau, 2017; Zanella, 2020). Secondly, the Posterior Bootstrap algorithms cannot be used in the case of intractable likelihoods, that is, when the likelihood has a form $f(x|\theta)/Z(\theta)$ for some unknown function $Z(\theta)$. Several Markov chain Monte Carlo methods have been proposed for target distributions of that type, including Møller et al. (2006); Andrieu and Roberts (2009); Murray et al. (2012); Deligiannidis et al. (2018); Middleton et al. (2020).

We would also like to note a limitation related specifically to Algorithm 7. Recall that the guidance on setting parameter c is well-understood only for conjugate models due to its link with the effective sample size of the prior. Our simulations show that in non-conjugate scenarios choosing c may be problematic. Pseudo-samples drawn from the prior predictive are often outliers with respect to the original dataset, and this phenomenon is particularly pronounced for vague priors. Therefore assigning too large importance to the pseudo-samples leads to unstable results and consequently distorts the inference. Another issue is that a one-dimensional parameter c controls the impact of the prior on all coordinates, as we showed in Theorem 3.17.

There have been attempts in the Bayesian literature to extend the notion of the effective sample size beyond the conjugate scenario, and quantify the amount of information in the

prior relative to the assumed sampling distribution, for example Morita et al. (2008, 2012); Neuenschwander et al. (2020). Developing the appropriate methodology for the choice of c in general models, potentially based on the approaches cited above, as well as supporting it with theoretical results, is an interesting future work direction.

Moreover, we note that even though the algorithms presented in this paper have a mechanism of incorporating the user's prior knowledge about the parameter of interest, they do not necessarily share other properties of Bayesian inference. In particular, they do not provide a coherent belief update about the parameter in the sense of Bissiri et al. (2016), even when the model is well-specified.

Even though Posterior Bootstrap methods can alleviate the consequences of model misspecification, this approach is not always optimal. In particular, in certain cases BayesBag and power posteriors turn out to be provably better in terms of prediction. Furthermore, it is important to emphasize that Posterior Bootstrap, power posteriors and BayesBag rely on correcting automatically the uncertainty around parameters, while preserving the same asymptotic mean. There are cases when the user has some prior knowledge about the nature of misspecification and has more trust in certain parts of the model than others. The cut model approach, also called modular inference, relies on cutting feedback, or the information flow, from the part of the model which we put less trust in to the one we trust more. This way of tackling misspecification allows the user to make better use of their understanding of the model than when applying more automatic approaches listed earlier. Bayesian literature on the cut model approach includes Liu et al. (2009); Plummer (2015); Jacob et al. (2017); Carmona and Nicholls (2020).

Acknowledgements

This work is supported by the EPSRC and MRC Centre for Doctoral Training in Next Generation Statistical Science: the Oxford-Warwick Statistics Programme, EP/L016710/1, and the Clarendon Fund. The author thanks Judith Rousseau, who helped enormously with all parts of the paper, and without whom this work would never exist. The author also thanks a person who asked the question of extending Weighted Likelihood Bootstrap to hierarchical models, but wishes to remain anonymous, as well as Luke J. Kelly, Pierre E. Jacob, Roman D. Stasiński, Stephen G. Walker and Jonathan Huggins for helpful discussions and feedback.

A Proof of Theorem 3.2

Proof. Let $\hat{p}_{\hat{\theta}}$, $\hat{p}_B$, $\hat{p}_{\pi_\eta}$ and $\hat{p}_{BB,b}$denote the predictive distributions under $\hat{\theta}$, Efron's bootstrap (Efron, 1979) , power posterior π_η, and BayesBag with hyperparameter b, respectively.

Firstly, note that for the risk associated with $\hat{p}_B$, we can use directly the formula given by Fushiki (2005) as follows:

$$E_{x_{1:n}}\left\{D_{KL}\left(p^*||\hat{p}_B\right)\right\} = E_{x_{1:n}}\left\{D_{KL}(p^*||\hat{p}_{\hat{\theta}})\right\} - \frac{1}{2n}\left[\operatorname{tr}\left\{I(\theta^*)J^{-1}(\theta^*)I(\theta^*)J^{-1}(\theta^*)\right\} - \operatorname{tr}\left\{I(\theta^*)J^{-1}(\theta^*)\right\}\right] + o\left(1/n\right).$$

To obtain an analogous formula for the risk associated with $\hat{p}_{\pi_\eta}$, we can use Lemma 5 of Shimodaira (2000) with

$$K_w^{[1\cdot 1]} = I(\theta^*), \quad K_w^{[2]} = J(\theta^*), \quad H_w = \eta J(\theta^*), \quad K_w^{[1]} = 0,$$

where the last equality holds because the expectation of the score function at θ^* is 0. We therefore have

$$E_{x_{1:n}}\left\{D_{KL}\left(p^*||\hat{p}_{\pi_\eta}\right)\right\} = E_{x_{1:n}}\left\{D_{KL}(p^*||\hat{p}_{\hat{\theta}})\right\} - \frac{1}{2n}\left[\eta^{-1}\operatorname{tr}\left\{I(\theta^*)J^{-1}(\theta^*)\right\} - \eta^{-1}d\right] + o\left(1/n\right).$$

Recall that $\lambda_1, \ldots, \lambda_d$ denote the eigenvalues of $J^{-1/2}(\theta^*)I(\theta^*)J^{-1/2}(\theta^*)$. By modifying calculations from Section 2 of Fushiki (2005) we get

$$\begin{aligned}
&E_{x_{1:n}}\left\{D_{KL}\left(p^*||\hat{p}_{\pi_\eta}\right)\right\} - E_{x_{1:n}}\left\{D_{KL}\left(p^*||\hat{p}_B\right)\right\} \\
=&\frac{1}{2n}\left[\operatorname{tr}\left\{I(\theta^*)J^{-1}(\theta^*)I(\theta^*)J^{-1}(\theta^*)\right\} - \operatorname{tr}\left\{I(\theta^*)J^{-1}(\theta^*)\right\}\right] \\
&- \frac{1}{2n}\left[\eta^{-1}\operatorname{tr}\left\{I(\theta^*)J^{-1}(\theta^*)\right\} - \eta^{-1}d\right] + o\left(1/n\right) \\
=&\frac{1}{2n}\operatorname{tr}\left[\left\{J^{-1/2}(\theta^*)I(\theta^*)J^{-1/2}(\theta^*)\right\}^2\right] - \frac{(1+\eta^{-1})}{2n}\operatorname{tr}\left\{J^{-1/2}(\theta^*)I(\theta^*)J^{-1/2}(\theta^*)\right\} \\
&+ \frac{\eta^{-1}d}{2n} + o\left(1/n\right) \\
=&\frac{1}{2n}\sum_{i=1}^{d}\left\{\lambda_i^2 - (1+\eta^{-1})\lambda_i + \eta^{-1}\right\} + o\left(1/n\right) \\
=&\frac{1}{2n}\sum_{i=1}^{d}(\lambda_i - 1)(\lambda_i - \eta^{-1}) + o\left(1/n\right).
\end{aligned}$$

Finally, we use Theorem 1 of Fushiki (2010) which captures the relationship between the risk associated with Weighted Likelihood Bootstrap and the risk associated with Efron's bootstrap, that is

$$E_{x_{1:n}}\left\{D_{KL}\left(p^*||\hat{p}_{\mathrm{WLB}}\right)\right\} = E_{x_{1:n}}\left\{D_{KL}\left(p^*||\hat{p}_B\right)\right\} + o\left(n^{-3/2}\right),$$

where $\hat{p}_{\mathrm{WLB}}$ denotes the predictive distribution under Weighted Likelihood Bootstrap. Therefore also

$$E_{x_{1:n}}\left\{D_{KL}\left(p^*||\hat{p}_{\pi_\eta}\right)\right\} - E_{x_{1:n}}\left\{D_{KL}\left(p^*||\hat{p}_{\mathrm{WLB}}\right)\right\} = \frac{1}{2n}\sum_{i=1}^{d}(\lambda_i - 1)(\lambda_i - \eta^{-1}) + o\left(1/n\right).$$

We finally observe that the expression $\sum_{i=1}^{d}(\lambda_i - 1)(\lambda_i - \eta^{-1})$ is negative when all eigenvalues are between 1 and η^{-1}. This completes the proof of the result for power posteriors.

We will now use similar calculations to obtain the result for BayesBag. By Theorem 3.2 of Huggins and Miller (2020), under regularity conditions, the asymptotic distribution of $\sqrt{n}\left(\theta - \hat{\theta}\right)$ for θ drawn from the BayesBag algorithm with hyperparameter b, is normal with mean 0 and covariance matrix

$$b^{-1}J^{-1} + b^{-1}J^{-1}(\theta^*)I(\theta^*)J^{-1}(\theta^*).$$

Using again the notation from Lemma 5 of (Shimodaira, 2000), we have

$$K_w^{[1\cdot1]} = I(\theta^*), \quad K_w^{[2]} = J(\theta^*), \quad H_w = b^{-1}J^{-1} + b^{-1}J^{-1}(\theta^*)I(\theta^*)J^{-1}(\theta^*), \quad K_w^{[1]} = 0.$$

Therefore

$$
\begin{aligned}
&E_{x_{1:n}}\left\{D_{KL}\left(p^*||\hat{p}_{\pi_{BB,b}}\right)\right\} - E_{x_{1:n}}\left\{D_{KL}\left(p^*||\hat{p}_B\right)\right\} \\
=&\frac{1-b^{-1}}{2n}\left[\operatorname{tr}\left\{I(\theta^*)J^{-1}(\theta^*)I(\theta^*)J^{-1}(\theta^*)\right\} - \operatorname{tr}\left\{I(\theta^*)J^{-1}(\theta^*)\right\}\right] \\
&- \frac{b^{-1}}{2n}\left[\operatorname{tr}\left\{I(\theta^*)J^{-1}(\theta^*)\right\} - d\right] + o\left(1/n\right) \\
=&\frac{1-b^{-1}}{2n}\operatorname{tr}\left[\left\{J^{-1/2}(\theta^*)I(\theta^*)J^{-1/2}(\theta^*)\right\}^2\right] - \frac{1}{2n}\operatorname{tr}\left\{J^{-1/2}(\theta^*)I(\theta^*)J^{-1/2}(\theta^*)\right\} \\
&+ \frac{b^{-1}d}{2n} + o\left(1/n\right) \\
=&\frac{1}{2n}\sum_{i=1}^{d}\left\{(1-b^{-1})\lambda_i^2 - \lambda_i + b^{-1}\right\} + o\left(1/n\right) \\
=&\frac{1}{2n}\sum_{i=1}^{d}(\lambda_i - 1)\left\{(1-b^{-1})\lambda_i - b^{-1}\right\} + o\left(1/n\right),
\end{aligned}
$$

and consequently

$$
\begin{aligned}
&E_{x_{1:n}}\left\{D_{KL}\left(p^*||\hat{p}_{\pi_{BB,b}}\right)\right\} - E_{x_{1:n}}\left\{D_{KL}\left(p^*||\hat{p}_{\mathrm{WLB}}\right)\right\} \\
=&\frac{1}{2n}\sum_{i=1}^{d}(\lambda_i - 1)\left\{(1-b^{-1})\lambda_i - b^{-1}\right\} + o\left(1/n\right).
\end{aligned}
$$

If $1/(b-1) < \lambda_1, \ldots, \lambda_d < 1$ or $b < 1 < \lambda_1, \ldots, \lambda_d$, then the expression

$$
\sum_{i=1}^{d}(\lambda_i - 1)\left\{(1-b^{-1})\lambda_i - b^{-1}\right\}
$$

is negative. This completes the proof.

□

B Proofs of theoretical results presented in Sections 2 and 4

B .1 Summary and Assumption 3.6

The proofs of Theorems 3.10 and 3.17 require similar reasoning and calculations. We therefore present below a detailed version of the proof of Theorem 3.10 and refer to it when proving Theorem 3.17. The plan of action is as follows. We start by obtaining an expansion for $n^{1/2}(\tilde{\theta} - \hat{\theta})$. Then applying Lemma 3.19, we show that the Edgeworth expansion of this

expression is indeed valid. Finally, we compute the coefficients of the Edgeworth expansion using the formula provided in Chapter 2 of Ghosh (1994). We start by presenting the full version of Assumption 3.6 and introducing the notation used in the proofs.

(Smoothness of the log likelihood function). There is an open ball B containing θ^* such that $\ell(x,\theta)$ is three times continuously differentiable with respect to $\theta \in B$ almost surely on x. Furthermore, there exist measurable functions G_1, G_2 and G_3 such that for $\theta \in B$

$$\begin{aligned}
\left|\frac{\partial \ell(x,\theta)}{\partial \theta_j}\right| &\leq G_1(x), & \mathbb{E}_{p^*}\left\{G_1(X)^4\right\} &< \infty, \\
\left|\frac{\partial^2 \ell(x,\theta)}{\partial \theta_j \partial \theta_k}\right| &\leq G_2(x), & \mathbb{E}_{p^*}\left\{G_2(X)^4\right\} &< \infty, \\
\left|\frac{\partial^3 \ell(x,\theta)}{\partial \theta_j \partial \theta_k \partial \theta_l}\right| &\leq G_3(x), & \mathbb{E}_{p^*}\left\{G_3(X)\right\} &< \infty.
\end{aligned}$$

B .2 Notation

We will adopt the following notation. Suppose $C \in \mathbb{R}^{d\times d\times d}$ and v is a vector in $\mathbb{R}^d$. Then $v^3 \in \mathbb{R}^{d\times d\times d}$ and the element ijk of v^3 is given by

$$v^3_{[ijk]} = v_i v_j v_k.$$

We also denote

$$Cv^3 = \sum_{i=1}^{d}\sum_{j=1}^{d}\sum_{k=1}^{d} C_{[ijk]} v_i v_j v_k.$$

Furthermore, for $v \in \mathbb{R}^d$ and $w \in \mathbb{R}^d$ we define $z = Cvw$ as a d-dimensional vector such that

$$z_i = \sum_{j=1}^{d}\sum_{k=1}^{d} C_{[ijk]} v_j w_k.$$

In particular, if $z = Cv^2$, then

$$z_i = \sum_{j=1}^{d}\sum_{k=1}^{d} C_{[ijk]} v_j v_k.$$

The above notation simplifies to standard multiplication for $d = 1$.

B .3 Auxiliary results

Our main tools for obtaining the Edgeworth expansion from the Taylor expansion are the two formulas below. The first one links the characteristic function of a variable $Y \in \mathbb{R}^d$ with its density, that is if $|\psi_Y(t)| < \infty$,

$$g_Y(y) = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \exp(-it^T y)\psi_Y(t)\mathrm{d}t_1, \dots \mathrm{d}t_d. \tag{3.38}$$

The second one is useful for computing the integral in equation (3.38):

$$\frac{1}{(2\pi)^d} \int_{-\infty}^{\infty} \exp(-zy) \exp(-z^2/2)(iz)^k \mathrm{d}z = H_k(y)\phi(y), \tag{3.39}$$

where $H_k(y)$ is the k-th Hermite polynomial and $\phi(y)$ is the density of the standard normal distribution. In particular, we can infer from (3.39) the following properties, which will be useful in our calculations. For $1 \leq q \leq d$ we have

$$\frac{1}{(2\pi)^d} \int \left\{ \exp\left(-t^T y\right) \exp\left(-t^T t/2\right) i^3 t_q^3 \right\} \mathrm{d}t_1, \dots \mathrm{d}t_d = H_3(y_q)\phi(y_1) \cdot \ldots \cdot \phi(y_d). \tag{3.40}$$

Furthermore, for $1 \leq q, r \leq d, q \neq r$ we get

$$\frac{1}{2\pi} \int_{\mathbb{R}^d} \left\{ \exp\left(-t^T y\right) \exp\left(-t^T t/2\right) i^3 t_q^2 t_r \right\} dt_1, \dots dt_d = H_2(y_q)H_1(y_r)\phi(y_1) \cdot \ldots \cdot \phi(y_d). \tag{3.41}$$

Finally, for $1 \leq q, r, s \leq d$, for pairwise different indices q, r, s we obtain the following equality

$$\frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \left\{ \exp\left(-t^T y\right) \exp\left(-t^T t/2\right) i^3 t_q t_r t_s \right\} \mathrm{d}t_1, \dots \mathrm{d}t_d = H_1(y_q)H_1(y_r)H_1(y_s)\phi(y_1) \cdot \ldots \cdot \phi(y_d). \tag{3.42}$$

Our proof of existence of valid Edgeworth expansions will rely on Lemma 3.19 below. Before stating the result, we introduce some additional notation. Let $a_j = (a_{j,1}, \dots, a_{j,d}) \in \mathbb{R}^d$ and define $\bar{a}_{2,n} \in \mathbb{R}^{d \times d}$ as

$$\bar{a}_{2,n} = \frac{1}{n} \sum_{j=1}^{n} a_j a_j^T.$$

Hence element (q, r) of $\bar{a}_{2,n}$ is equal to $\frac{1}{n} \sum_{j=1}^{n} a_{j,q} a_{j,r}$. Similarly, let $\bar{a}_{3,n} \in \mathbb{R}^{d \times d \times d}$ be such that element (q, r, s) of $\bar{a}_{3,n}$ is given by $\frac{1}{n} \sum_{j=1}^{n} a_{j,q} a_{j,r} a_{j,s}$. Furthermore, we let $b_j = \bar{a}_{2,n}^{-1/2} a_j$,

and similarly set $\bar{b}_{3,n} \in \mathbb{R}^{d\times d\times d}$ so that element (q,r,s) of $\bar{b}_{3,n}$ is given by $\frac{1}{n}\sum_{j=1}^{n} b_{j,q}b_{j,r}b_{j,s}$, where $b_{j,q}$ denotes element q of b_j. Note that a_j and b_j may depend on n.

Lemma 3.19. *We consider $\mathbb{R}$-valued independent random variables $Z_1,\ldots,Z_n$ distributed according to some continuous distribution Z having all moments and satisfying $E(Z)=1$ and $var(Z)=1$. Let $(a_j)_{j=1}^{n} \in \mathbb{R}^d$, be a sequence of vectors in $\mathbb{R}^d$, possibly depending on n and satisfying conditions a) – c) below.*

Assume that there exists a sequence $\epsilon_n = o(1)$ such that

a) There exist $A_2 \in \mathbb{R}^{d\times d}$, $A_2 \succ 0$ and $A_3 \in \mathbb{R}^{d\times d\times d}$ such that $\|\bar{a}_{2,n} - A_2\| \le \epsilon_n$ and $\|\bar{a}_{3,n} - A_3\| \le \epsilon_n$ for $n \ge 1$.

b) We have $\sup_n \frac{1}{n}\sum_{j=1}^{n} \|a_j\|^4 < \infty$.

c) For any $n \ge 1$ we have $\sup_{j\le n} \|a_j\| \le \epsilon_n n^{1/2}$.

Then S_n defined as

$$S_n = \frac{\bar{a}_{2,n}^{-1/2}}{n^{1/2}} \sum_{j=1}^{n} a_j Z_j,$$

has a valid Edgeworth expansion. Moreover the expansion of the characteristic function ψ_{S_n} of S_n up to the term $o\left(n^{-1/2}\right)$ is given by

$$\psi_{S_n}(t) = \exp\left(-\frac{t^T t}{2}\right)\left(1 + i^3\gamma\frac{\sum_{1\le q,r,s\le d} B_{q,r,s}t_q t_r t_s}{6n^{1/2}}\right) + o\left(n^{-1/2}\right),$$

where $\gamma = \mathbb{E}\left(Z^3\right)$ and $B = \lim_{n\to\infty}\bar{b}_{3,n}$.

Proof. Before proceeding with the proof, observe that the limit $\lim_{n\to\infty}\bar{b}_{3,n}$ exists by definition of b_j and assumption a), thus B is well-defined with all elements finite.

The characteristic function of S_n is given by:

$$\psi_{S_n}(t) = \mathbb{E}\left\{\exp\left(it^T S_n\right)\right\} = \mathbb{E}\left[\exp\left\{it^T\left(\frac{\bar{a}_{2,n}^{-1/2}}{n^{1/2}}\sum_{j=1}^{n} a_j Z_j\right)\right\}\right] = \prod_{j=1}^{n}\psi_Z\left(\frac{t^T b_j}{n^{1/2}}\right),$$

where ψ_Z is the characteristic function of Z. Using the fact that $\mathbb{E}(Z)=0$ and $\mathbb{E}\left(Z^2\right)=1$

we have

$$\begin{aligned}
\psi_Z\left(\frac{t^T b_j}{n^{1/2}}\right) =& \mathbb{E}\left\{\exp\left(i\frac{t^T b_j}{n^{1/2}}Z\right)\right\} \\
=& \mathbb{E}\left\{1+\frac{i\left(t^T b_j\right)Z}{n^{1/2}}+\frac{i^2\left(t^T b_j\right)^2 Z^2}{2n}+\frac{i^3\left(t^T b_j\right)^3 Z^3}{6n^{3/2}}\right\}+O\left(\frac{\|b_j\|^4\mathbb{E}Z^4}{n^2}\right) \\
=& 1-\frac{\left(t^T b_j\right)^2}{2n}+\frac{i^3\left(t^T b_j\right)^3\gamma}{6n^{3/2}}+O\left(\frac{\|b_j\|^4\mathbb{E}Z^4}{n^2}\right) \\
=& 1-\frac{\left(t^T b_j\right)^2}{2n}+\frac{i^3\left(t^T b_j\right)^3\gamma}{6n^{3/2}}+O\left(\frac{\|b_j\|^4}{n^2}\right)
\end{aligned}$$

since $\mathbb{E}\left(Z^4\right)<\infty$.

Note that the definition of b_j yields the following simple equality:

$$\begin{aligned}
\frac{1}{n}\sum_{j=1}^n (t^T b_j)^2 =& \frac{1}{n}\sum_{j=1}^n \left(t^T \bar{a}_{2,n}^{-1/2} a_j a_j^T \bar{a}_{2,n}^{-1/2} t\right) \\
=& t^T \bar{a}_{2,n}^{-1/2}\left(\frac{1}{n}\sum_{j=1}^n a_j a_j^T\right)\bar{a}_{2,n}^{-1/2} t = t^T t.
\end{aligned} \tag{3.43}$$

Therefore

$$\begin{aligned}
\prod_{j=1}^n \psi_Z\left(\frac{t^T b_j}{n^{1/2}}\right) &= \exp\left[\sum_{j=1}^n \log\left\{1-\frac{\left(t^T b_j\right)^2}{2n}+\frac{i^3\left(t^T b_j\right)^3\gamma}{6n^{3/2}}+O\left(\frac{\|b_j\|^4}{n^2}\right)\right\}\right] \\
&= \exp\left(\sum_{j=1}^n\left[-\frac{\left(t^T b_j\right)^2}{2n}+\frac{i^3\left(t^T b_j\right)^3\gamma}{6n^{3/2}}+O\left(\frac{\|b_j\|^4}{n^2}\right)+O\left\{\frac{\left(t^T b_j\right)^4}{4n^2}\right\}\right]\right) \\
&= \exp\left\{-\frac{t^T t}{2}+\frac{i^3\sum_{j=1}^n\left(t^T b_j\right)^3\gamma}{6n^{3/2}}+O\left(n^{-1}\right)\right\} \\
&= \exp\left(-\frac{t^T t}{2}\right)\left(1+i^3\gamma\frac{\sum_{1\le q,r,s\le d} B_{q,r,s}t_q t_r t_s}{6n^{1/2}}\right)+o\left(n^{-1/2}\right).
\end{aligned}$$

The second equality above comes from the Taylor expansion of $\log(1-x)$ around $x=0$ and the fact that for some constant c and any n we have $\sup_{j\le n}\|b_j\|\le c\epsilon_n n^{1/2}$ so that $\sup_{j\le n}\|b_j\|^2/n = o(1)$, which follows from convergence of $\bar{a}_{2,n}$ to A_2 and assumption c). The third equality above results from (3.43) and the fact that $\sup_n\sum_{j=1}^n\|b_j\|^4/n<\infty$ by convergence of $\bar{a}_{2,n}$ to A_2 and assumption b). The last equality comes from $\|\bar{b}_{3,n}-B\|=o(1)$. Applying formulas (3.38) as well as (3.40), (3.41) and (3.42) we get the Edgeworth expansion

up to $o\left(n^{-1/2}\right)$. □

Lemma 3.20. *Consider model (3.1) and one of the two scenarios. In the first case we consider $\tilde{\theta}$ drawn from Algorithm 5 and suppose that Assumptions 3.4 – 3.9 hold. In the second case we consider $\tilde{\theta}$ drawn from Algorithm 7 and suppose that Assumptions 3.4 – 3.8, 3.16 hold. Under any of the scenarios we have $\|\tilde{\theta} - \hat{\theta}\| = o_P(1)$, as well as $\|\tilde{\theta} - \theta^*\| = o_P(1)$ as n goes to infinity. Moreover, $-n^{-1}\sum_{j=1}^n w_j D_\theta^2 \ell(x_j, \hat{\theta})$ converges to $J(\theta^*)$ in probability.*

Lemma 3.20 can be obtained by repeating the reasoning presented in Chapter 3 of Newton (1991), in particular Theorem 6 and lemmas used to prove it; see also Supplementary Material of Lyddon et al. (2019). The difference between Chapter 3 of Newton (1991) and our case is that we allow for the model to be misspecified, and therefore consider a pseudo-true parameter rather than the true one. Moreover, in the first scenario instead of considering the function $n^{-1}\sum_{j=1}^n w_j \ell(x_i, \theta)$, we analyse $n^{-1}\left\{\sum_{j=1}^n w_j \ell(x_i, \theta) + w_0^T \log \pi(\theta)\right\}$. The smoothness assumption on the prior and the fact that $n^{-1} w_0^T \log \pi(\theta) \to 0$ as $n \to \infty$ ensure that we can use directly the arguments from Chapter 3 of Newton (1991). In the second scenario we consider a function $n^{-1}\left\{\sum_{j=1}^n w_j \log f(x_i|\theta) + \sum_{i=1}^T w_{n+i} \log f(x_{n+i}|\theta)\right\}$, where $x_{n+1}, \ldots, x_{n+T}$ are pseudosamples. Since the number of pseudosamples is finite, again, we can repeat the reasoning from Newton (1991).

B .4 Proof of Theorem 3.10

Taylor expansion

We begin the proof by writing the Taylor expansion as follows. Let

$$L_n(\theta) = \sum_{j=1}^n w_j \ell(x_i, \theta) + w_0^T \log \pi(\theta).$$

Then by Assumptions 3.6 and 3.9, using notation from Sections 1 .3 and 2 .2 we have

$$0 = \nabla_\theta L_n(\tilde{\theta}) = \nabla_\theta L_n(\hat{\theta}) + D_\theta^2 L_n(\hat{\theta})(\tilde{\theta} - \hat{\theta}) + D_\theta^3 L_n(\hat{\theta})\frac{(\tilde{\theta} - \hat{\theta})^2}{2} + o_P\left(n\|\tilde{\theta} - \hat{\theta}\|^2\right)$$

and consequently

$$\left\{-\frac{1}{n} D_\theta^2 L_n(\hat{\theta})\right\} n^{1/2}(\tilde{\theta} - \hat{\theta}) = \frac{\nabla_\theta L_n(\hat{\theta})}{n^{1/2}} + \frac{D_\theta^3 L_n(\hat{\theta})(\tilde{\theta} - \hat{\theta})^2}{2n^{1/2}} + o_P\left(n^{1/2}\|\tilde{\theta} - \hat{\theta}\|^2\right) \quad (3.44)$$

We now use Assumptions 3.4 and 3.5 to state that $(\tilde{\theta} - \hat{\theta}) = o_P(1)$. We use the following notation

$$\begin{aligned}
S_{1,w,n} =& \frac{\sum_{j=1}^n (w_j - 1)\nabla_\theta \ell(x_j, \hat{\theta})}{n^{1/2}}, & S_{2,w,n} =& \frac{\sum_{j=1}^n (w_j - 1)\left\{-D_\theta^2 \ell(x_j, \hat{\theta})\right\}}{n^{1/2}}, \\
J_{w,n} =& -\frac{1}{n}\sum_{j=1}^n w_j D_\theta^2 \ell(x_j, \hat{\theta}), & \mu_{3,w,n} =& \frac{1}{n}\sum_{j=1}^n w_j D_\theta^3 \ell(x_j, \hat{\theta}), \\
\mu_{3,n} =& \frac{1}{n}\sum_{j=1}^n D_\theta^3 \ell(x_j, \hat{\theta}), & \mu_3 =& \mathbb{E}_{p^*}\left\{D_\theta^3 \ell(X, \theta^*)\right\}, \\
A_{3,n} =& \frac{1}{n}\sum_{j=1}^n \left\{\nabla_\theta \ell(x_j, \hat{\theta})\right\}^3, & A_3 =& \mathbb{E}_{p^*}\left\{\nabla_\theta \ell(X, \theta^*)\right\}^3, \\
\tilde{L}_{1,2,n} =& \frac{1}{n}\sum_{j=1}^n D_\theta^2 \ell(x_j, \hat{\theta}) J_n^{-1} \nabla_\theta \ell(x_j, \hat{\theta}), & \tilde{L}_{1,2} =& \mathbb{E}_{p^*}\left\{D_\theta^2 \ell(X, \theta^*) J(\theta^*)^{-1} \nabla_\theta \ell(X, \theta^*)\right\}.
\end{aligned}$$

We observe that since $\|J_n - J(\hat{\theta})\| = o_P(1)$ and $\|I_n - I(\hat{\theta})\| = o_P(1)$, then by Assumption 3.7 matrices I_n, J_n as well as $J_{w,n}$ are positive definite and hence invertible for large n. Furthermore, Assumption 3.6 ensures that A_3, $\tilde{L}_{1,2}$ and μ_3 are finite. In passing, A_3 and μ_3 are defined above for an arbitrary dimension d, they are however consistent with the one-dimensional definitions presented in (3.9).

Note also that since $\sum_{j=1}^n \nabla_\theta \ell(x_i, \hat{\theta}) = 0$, then $S_{1,w,n} = n^{-1/2}\left\{\sum_{j=1}^n w_j \nabla_\theta \ell(x_i, \hat{\theta})\right\}$. We can therefore rewrite (3.44) as

$$n^{1/2}(\tilde{\theta} - \hat{\theta}) = J_{w,n}^{-1}\left\{S_{1,w,n} + \frac{w_0 \cdot \nabla_\theta \log \pi(\hat{\theta})}{n^{1/2}}\right\} + J_{w,n}^{-1}\frac{\mu_{3,w,n}(\tilde{\theta} - \hat{\theta})^2}{2n^{1/2}} + o_P\left(n^{1/2}\|\tilde{\theta} - \hat{\theta}\|^2\right). \tag{3.45}$$

Let $\mathbf{I}$ denote the $d \times d$ identity matrix. We have $J_{w,n} = J_n + S_{2,w,n}/n^{1/2}$, and so

$$\begin{aligned}
J_{w,n}^{-1} &= \left\{J_n\left(\mathbf{I} + J_n^{-1} S_{2,w,n}/n^{1/2}\right)\right\}^{-1} \\
&= \left(\mathbf{I} + J_n^{-1} S_{2,w,n}/n^{1/2}\right)^{-1} J_n^{-1} \\
&= \left\{\mathbf{I} - J_n^{-1} S_{2,w,n}/n^{1/2} + o\left(n^{-1/2}\right)\right\} J_n^{-1} \\
&= J_n^{-1} - J_n^{-1} S_{2,w,n} J_n^{-1}/n^{1/2} + o\left(n^{-1/2}\right) \\
&= J_n^{-1}\left(\mathbf{I} - S_{2,w,n} J_n^{-1}/n^{1/2}\right) + o\left(n^{-1/2}\right),
\end{aligned} \tag{3.46}$$

where the third equality comes from the Taylor expansion of $1/(1+x)$ at $x = 0$.

The first order approximation of $n^{1/2}(\tilde{\theta} - \hat{\theta})$ is thus given by

$$n^{1/2}(\tilde{\theta} - \hat{\theta}) = J_n^{-1} S_{1,w,n} + o(1). \tag{3.47}$$

Plugging (3.47) and (3.46) into (3.45) we get

$$\begin{aligned} n^{1/2}(\tilde{\theta} - \hat{\theta}) =& J_n^{-1} \left\{ S_{1,w,n} + \frac{w_0 \cdot \nabla_\theta \log \pi(\hat{\theta})}{n^{1/2}} - \frac{S_{2,w,n} J_n^{-1} S_{1,w,n}}{n^{1/2}} \right\} \\ &+ J_{w,n}^{-1} \frac{\mu_{3,w,n} \left(J_n^{-1} S_{1,w,n}\right)^2}{2n^{1/2}} + o\left(n^{-1/2}\right) \\ =& J_n^{-1} \left\{ S_{1,w,n} + \frac{w_0 \cdot \nabla_\theta \log \pi(\hat{\theta})}{n^{1/2}} - \frac{S_{2,w,n} J_n^{-1} S_{1,w,n}}{n^{1/2}} \right\} \\ &+ J_n^{-1} \frac{\mu_3 \left(J_n^{-1} S_{1,w,n}\right)^2}{2n^{1/2}} + o\left(n^{-1/2}\right), \end{aligned}$$

where the last equality results from $J_{w,n} = J_n + S_{2,w,n}/n^{1/2}$ and $\mu_{3,w,n} = \mu_3 + o_P(1)$. We finally get

$$\begin{aligned} I_n^{-1/2} J_n n^{1/2} (\tilde{\theta} - \hat{\theta}) =& I_n^{-1/2} \left\{ S_{1,w,n} + \frac{w_0 \cdot \nabla_\theta \log \pi(\hat{\theta})}{n^{1/2}} - \frac{S_{2,w,n} J_n^{-1} S_{1,w,n}}{n^{1/2}} \right\} \\ &+ I_n^{-1/2} \frac{\mu_3 \left(J_n^{-1} S_{1,w,n}\right)^2}{2n^{1/2}} + o\left(n^{-1/2}\right). \end{aligned} \tag{3.48}$$

We will use the above result to obtain the Edgeworth expansion.

Edgeworth expansion – multidimensional case

Proof. Our proof will rely mainly on results from Chapter 2 of Ghosh (1994). Firstly, we apply Lemma 3.19 to

$$Z_j = w_j - 1 \quad \text{and} \quad a_j = \left\{ \frac{\partial \ell(x_j, \hat{\theta})}{\partial \theta_p}, \frac{\partial^2 \ell(x_j, \hat{\theta})}{\partial \theta_r \partial \theta_s} \right\}, \quad 1 \le p \le d, 1 \le r \le s \le d,$$

that is, a_j is a vector of all first and second partial derivatives of $\ell(x_j, \hat{\theta})$, to get that $(S_{1,w,n}, S_{2,w,n})$ has a joint valid Edgeworth expansion. We justify below briefly why the conditions of Lemma 3.19 are satisfied. Observe that the assumptions about Z_j hold by Assumption 3.8. Conditions b) and c), as well as convergence to A_2 and A_3 in condition a) of Lemma 3.19, hold with probability 1 with respect to the observations $x_1, \ldots, x_n$ by Assumption 3.6, and the strong law of large numbers. We also have $A_2 \succ 0$ as required in condition a) by Assumption 3.7.

Now let X_n denote the right hand side of equation (3.48). Since conditioning on the observations $x_1, \ldots, x_n$ the variable X_n is a polynomial of $(S_{1,w,n}, S_{2,w,n})$, it also has a valid Edgeworth expansion, and we can apply the theory outlined in Chapter 2 of Ghosh (1994) to obtain the coefficients of this expansion.

By equation 2.7 of Ghosh (1994) the characteristic function of X_n satisfies

$$\psi_{X_n} = \exp\left(-\frac{t^T t}{2}\right)\left\{1 + \kappa_{1,n}(it) + \frac{\kappa_{3,n}(it)^3}{6}\right\} + o\left(n^{-1/2}\right), \tag{3.49}$$

where $\kappa_{1,n}$ and $\kappa_{3,n}$ are the first and the third cumulant, respectively, calculated up to $o\left(n^{-1/2}\right)$.

We also define $Z_n := (Z_{n,1}, \ldots, Z_{n,d})$ where

$$Z_{n,q} := \frac{1}{n}\sum_{j=1}^{n} \nabla\ell(x_j, \hat{\theta})^T J_n^{-1} \mu_{3,q} J_n^{-1} \nabla\ell(x_j, \hat{\theta}), \quad q = 1, \ldots, d,$$

and $\mu_{3,q} \in \mathbb{R}^{p\times p}$ is a matrix such that element $[r, s]$ of $\mu_{3,q}$ is equal to element $[q, r, s]$ of μ_3. In an analogous way we define $Z = (Z_1, \ldots, Z_d)$, where

$$Z_q = \mathbb{E}_{p^*}\left\{\nabla\ell(X, \theta^*)^T J(\theta^*)^{-1} \mu_{3,q} J(\theta^*)^{-1} \nabla\ell(X, \theta^*)\right\}, \quad q = 1, \ldots, d.$$

We observe that by Assumption 3.6 Z_q is finite for each $q = 1, \ldots, d$. We obtain

$$\mathbb{E}(X_n) = I_n^{-1/2}\left\{\frac{w_0 \cdot \nabla_\theta \log\pi(\hat{\theta})}{n^{1/2}} + \frac{Z_n}{2n^{1/2}} + \frac{\tilde{L}_{1,2,n}}{n^{1/2}} + o\left(n^{-1/2}\right)\right\}.$$

Therefore we can write

$$\kappa_{1,n} = I(\theta^*)^{-1/2}\left(\frac{w_0 \cdot \nabla_\theta \log\pi(\theta^*)}{n^{1/2}} + \frac{Z}{2n^{1/2}} + \frac{\tilde{L}_{1,2}}{n^{1/2}}\right).$$

Recall $\mathbb{E}(X_n - \mathbb{E}X_n)^3 = \kappa_{3,n} + o\left(n^{-1/2}\right)$. Since the term $n^{-1/2} I(\theta^*)^{-1/2}\left\{w_0 \cdot \nabla_\theta \log\pi(\hat{\theta})\right\}$ cancels in the expression $X_n - \mathbb{E}X_n$, we get that $\kappa_{3,n}$ does not depend on the prior. Finally, using formula (3.49), and equations (3.40), (3.41) and (3.42) we conclude with the general form of the Edgeworth expansion

$$\phi(y)\left[1 + y^T I(\theta^*)^{-1/2}\frac{\{w_0 \cdot \nabla_\theta \log\pi(\theta^*)\}}{n^{1/2}} + \frac{P_{3,n}(y)}{n^{1/2}}\right] + o\left(n^{-1/2}\right),$$

where $P_{3,n}(y)$ is an order three polynomial independent from the prior. This completes the proof of equation (3.10). □

Edgeworth expansion – one-dimensional case

Proof. When $d = 1$, we can obtain a more explicit form of the Edgeworth expansion. In particular, $\kappa_{1,n}$ becomes

$$\kappa_{1,n} = \frac{w_0 \nabla_\theta \log \pi(\theta^*)}{n^{1/2} I(\theta^*)^{1/2}} + \frac{\mu_3}{2n^{1/2}} \frac{I(\theta^*)^{1/2}}{J(\theta^*)^2} + \frac{L_{1,2}}{n^{1/2} I(\theta^*)^{1/2} J(\theta^*)},$$

for $L_{1,2}$ defined in (3.9). We additionally define

$$L_{1,2,n} = \frac{1}{n} \sum_{j=1}^{n} D_\theta^2 \ell(x_j, \theta^*) \nabla_\theta \ell(x_j, \theta^*).$$

We then get

$$\begin{aligned}\mathbb{E}\left(X_n - \mathbb{E}X_n\right)^3 = \mathbb{E}\left(\frac{S_{1,w,n}^3}{I_n^{3/2}}\right) &+ 3\mathbb{E}\left\{\frac{S_{1,w,n}^2}{I_n}\left(\frac{\mu_3}{2n^{1/2}}\frac{S_{1,w,n}^2}{I_n^{1/2}J_n^2} - \frac{\mu_3}{2n^{1/2}}\frac{I_n}{I_n^{1/2}J_n^2}\right)\right\} \\ &- 3\mathbb{E}\left\{\frac{S_{1,w,n}^2}{I_n}\left(\frac{L_{1,2,n}}{n^{1/2}I_n^{1/2}J_n} + \frac{S_{1,w,n}S_{2,w,n}}{n^{1/2}I_n^{1/2}J_n}\right)\right\} + o\left(n^{-1/2}\right).\end{aligned}$$

We have

$$\begin{aligned}\mathbb{E}\left(S_{1,w,n}^3\right) &= \frac{\gamma}{n^{1/2}} A_{3,n}, \\ \mathbb{E}\left(S_{1,w,n}^4\right) &= \frac{3}{n^{1/2}} I_n^2 + o\left(n^{-1/2}\right), \\ \mathbb{E}\left(S_{1,w,n}^3 S_{2,w,n}\right) &= -\frac{3}{n^{1/2}} I_n L_{1,2,n} + o\left(n^{-1/2}\right),\end{aligned}$$

therefore

$$\begin{aligned}\mathbb{E}\left(X_n - \mathbb{E}X_n\right)^3 = \frac{\gamma A_{3,n}}{n^{1/2} I_n^{3/2}} &+ \left(\frac{9\mu_3}{2n^{1/2}}\frac{I_n^2}{I_n^{3/2}J_n^2} - \frac{3\mu_3}{2n^{1/2}}\frac{I_n^2}{I_n^{3/2}J_n^2}\right) \\ &- \left(\frac{3I_n L_{1,2,n}}{n^{1/2}I_n^{3/2}J_n} - \frac{9I_n L_{1,2,n}}{n^{1/2}I_n^{3/2}J_n}\right) + o\left(n^{-1/2}\right).\end{aligned}$$

We get the following value of $\kappa_{3,n}$

$$\kappa_{3,n} = \frac{\gamma A_3}{n^{1/2} I(\theta^*)^{3/2}} + \frac{6\mu_3}{2n^{1/2}} \frac{I(\theta^*)^{1/2}}{J(\theta^*)^2} + \frac{6L_{1,2}}{n^{1/2} I(\theta^*)^{1/2} J(\theta^*)}$$

Using equation (3.49) we get

$$\psi_{X_n} = \exp\left(-\frac{t^2}{2}\right)\left[1 + \left\{\frac{w_0\nabla_\theta \log\pi(\theta^*)}{n^{1/2}I(\theta^*)^{1/2}} + \frac{\mu_3}{2n^{1/2}}\frac{I(\theta^*)^{1/2}}{J(\theta^*)^2} + \frac{L_{1,2}}{n^{1/2}I(\theta^*)^{1/2}J(\theta^*)}\right\}(it) + \left\{\frac{\gamma A_3}{6n^{1/2}I(\theta^*)^{3/2}} + \frac{\mu_3}{2n^{1/2}}\frac{I(\theta^*)^{1/2}}{J(\theta^*)^2} + \frac{L_{1,2}}{n^{1/2}I(\theta^*)^{1/2}J(\theta^*)}\right\}(it)^3\right] + o\left(n^{-1/2}\right),$$

and finally equations (3.39) and (3.38) yield the Edgeworth expansion of the density

$$g_{X_n}(y) = \phi(y)\left[1 + \left\{\frac{w_0\nabla_\theta \log\pi(\theta^*)}{n^{1/2}I(\theta^*)^{1/2}} + \frac{\mu_3}{2n^{1/2}}\frac{I(\theta^*)^{1/2}}{J(\theta^*)^2} + \frac{L_{1,2}}{n^{1/2}I(\theta^*)^{1/2}J(\theta^*)}\right\}H_1(y) + \left\{\frac{\gamma A_3}{6n^{1/2}I(\theta^*)^{3/2}} + \frac{\mu_3}{2n^{1/2}}\frac{I(\theta^*)^{1/2}}{J(\theta^*)^2} + \frac{L_{1,2}}{n^{1/2}I(\theta^*)^{1/2}J(\theta^*)}\right\}H_3(y)\right] + o\left(n^{-1/2}\right),$$

which completes the proof of (3.11). □

B .5 Proof of Theorem 3.17

Proof. Analogous steps that we took to obtain the Edgeworth expansion for Algorithm 5 will lead us to the result for Algorithm 7. When considering the Taylor expansion, instead of equation (3.48) we get

$$\tilde{X}_n = I_n^{-1/2}\left\{S_{1,w,n} + \frac{\sum_{j=n+1}^{n+T} w_j\nabla_\theta\ell(x_j,\hat{\theta})}{n^{1/2}} - \frac{S_{2,w,n}J_n^{-1}S_{1,w,n}}{n^{1/2}}\right\} + I_n^{-1/2}\frac{\mu_3\left(J_n^{-1}S_{1,w,n}\right)^2}{2n^{1/2}} + o\left(n^{-1/2}\right).$$

Notice that $n^{-1/2}\sum_{j=n+1}^{n+T} w_j\nabla_\theta\ell(x_j,\hat{\theta})$ is independent from $(S_{1,w,n}, S_{2,w,n})$ conditionally on the data, and has a valid Edgeworth expansion, therefore there exists a valid joint expansion of

$$\left\{S_{1,w,n}, S_{2,w,n}, \frac{1}{n^{1/2}}\sum_{j=n+1}^{n+T} w_j\nabla_\theta\ell(x_j,\hat{\theta})\right\}. \tag{3.50}$$

Since $\tilde{X}_n$ is a polynomial of (3.50), similar reasoning to that used for Theorem 3.10 yields an expansion

$$\psi_{\tilde{X}_n} = \exp\left(-\frac{t^Tt}{2}\right)\left\{1 + \tilde{\kappa}_{1,n}(it) + \tilde{\kappa}_{3,n}\frac{(it)^3}{6}\right\} + o\left(n^{-1/2}\right),$$

where $\tilde{\kappa}_{1,n}$ and $\tilde{\kappa}_{3,n}$ are the first and the third cumulant, respectively, calculated up to $o\left(n^{-1/2}\right)$. Since $\mathbb{E}\left(w_j\right) = c/T$ for $j = n+1, \ldots, n+T$ we have

$$\begin{aligned}\mathbb{E}\left\{\sum_{j=n+1}^{n+T} w_j \nabla_\theta \ell(x_j, \hat{\theta})\right\} &= \sum_{j=n+1}^{n+T} \mathbb{E}_{\pi(\theta)}\left[\mathbb{E}_{f(x_j|\theta)}\left\{w_j \nabla_\theta \ell(x_j, \hat{\theta})\middle|\theta\right\}\right] \\ &= c\int \nabla_\theta \left\{\log f(x|\hat{\theta})\right\} f_\pi(x)\mathrm{d}x,\end{aligned} \tag{3.51}$$

for the prior predictive $f_\pi(x)$ defined in Section 4 . By Assumption 3.16 we have

$$\int \nabla_\theta \left\{\log f(x|\hat{\theta})\right\} f_\pi(x)\mathrm{d}x \to \alpha_{\theta^*,\pi}$$

as $n \to \infty$, which leads to the following formula for $\tilde{\kappa}_{1,n}$

$$\tilde{\kappa}_{1,n} = I(\theta^*)^{-1/2}\left(\frac{c\alpha_{\theta^*,\pi}}{n^{1/2}} + \frac{Z}{2n^{1/2}} + \frac{\tilde{L}_{1,2}}{n^{1/2}}\right).$$

Since the quantities

$$I_n^{-1/2}\frac{\sum_{j=n+1}^{n+T} w_j \nabla \ell(x_j, \hat{\theta})}{n^{1/2}}, \quad \tilde{X}_n - I_n^{-1/2}\frac{\sum_{j=n+1}^{n+T} w_j \nabla \ell(x_j, \hat{\theta})}{n^{1/2}}$$

are independent, we get that the third cumulant does not depend on the prior. We therefore get $\tilde{\kappa}_{3,n} = \kappa_{3,n}$ for $\kappa_{3,n}$ defined in the proof of Theorem 3.10. This completes the proof. □

B .6 Edgeworth expansions for other methods based on Weighted Likelihood Bootstrap

Our results obtained in Theorem 3.10 and 3.17 can easily be extended to obtain Edgeworth expansions for other methods based on Weighted Likelihood Bootstrap, in particular Newton (1991) and Lyddon et al. (2018).

We start by discussing the method proposed by Newton et al. (2020) presented in Algorithm 8. The regularization parameter λ plays a similar role to w_0 in Algorithm 5 if $w_0 \in \mathbb{R}$, it is however multiplied by a random variable $\tilde{w}$ drawn from the exponential distribution. When the prior factorizes, Newton et al. (2020) alternatively propose to assign a different random weight $\tilde{w}_i \sim \mathrm{Exp}(1)$ to each coordinate $i = 1, \ldots, d$. The penalization term would

then be given by

$$\sum_{i=1}^{d} \lambda \tilde{w}_i \log \pi_i(\theta_i).$$

Note that since the regularization parameter stays fixed for all coordinates, this approach is of different nature to the way we proposed to set w_0 in Section 2 .3. In our case the idea is to assign potentially different regularization parameters to different coordinates, accounting for situations when overdispersion or underdispersion is more severe for some coordinates than others.

Algorithm 8 Weighted Bayesian Bootstrap (Newton et al., 2020)

1: **hyperparameters**: λ.
2: **for** j = 1, . . . , N **do** ▷ (in parallel)
3: Draw $\tilde{w}^{(j)}, w_1^{(j)}, \ldots, w_n^{(j)} \sim \text{Exp}(1)$.
4: Set $\tilde{\theta}^{(j)} = \arg\max_\theta \left\{ \sum_{i=1}^n w_i \log f(x_i|\theta) + \lambda \tilde{w}^{(j)} \log \pi(\theta) \right\}$.
5: **end for**
6: **return** Samples $\left\{\tilde{\theta}^{(j)}\right\}_{j=1}^N$.

Assuming that $\tilde{w}$ is a one-dimensional variable drawn from Exp(1), we will show that Algorithm 5 and Algorithm 8 have the same Edgeworth expansion as long as $w_0 = \lambda$. To this end, we use very similar arguments to those used in the proof of Theorem 3.10. Firstly, note that the Taylor expansion of $I_n^{-1/2} J_n n^{1/2}(\tilde{\theta} - \hat{\theta})$ for $\tilde{\theta}$ drawn from Algorithm 8 is given by

$$I_n^{-1/2} \left(S_{1,w,n} + \frac{\lambda \tilde{w} \nabla_\theta \log \pi(\hat{\theta})}{n^{1/2}} - \frac{S_{2,w,n} J_n^{-1} S_{1,w,n}}{n^{1/2}} \right) + I_n^{-1/2} \frac{\mu_3 \left(J_n^{-1} S_{1,w,n} \right)^2}{2n^{1/2}}$$

Repeating the arguments used in the proofs and Theorems 3.10 and 3.17 we conclude that

$$\left(S_{1,w,n}, S_{2,w,n}, \frac{1}{n^{1/2}} \lambda \tilde{w} \nabla_\theta \log \pi(\hat{\theta}) \right)$$

has a valid Edgeworth expansion. Since $\mathbb{E}\lambda\tilde{w} = \lambda$, then the corresponding cumulants $\kappa_{1,n}$ and $\kappa_{3,n}$ are the same as in the proof of Theorem 3.10 for $\lambda = w_0$.

As for the method proposed by Lyddon et al. (2018), which we present in Algorithm 9, its second order Edgeworth expansion was discussed in Section 4 . To make the statement more precise, let $\tilde{\theta}$ be drawn from Algorithm 9, and θ be drawn from standard Weighted Likelihood Bootstrap (Newton and Raftery, 1994). Then up to the term of order $o\left(n^{-1/2}\right)$ the Edgeworth expansion of $I_n^{-1/2} J_n n^{1/2}(\tilde{\theta} - \hat{\theta})$ is the same as the Edgeworth expansion $I_n^{-1/2} J_n n^{1/2}(\theta - \hat{\theta})$, for any choice of fixed parameters c and T.

Algorithm 9 Posterior Bootstrap (Lyddon et al., 2018) (with pseudo-samples from the posterior predictive)

1: **for** j = 1, ..., N **do** ▷ (in parallel)
2: Draw $\theta^{(j)}_{\text{post}}$ from the Bayesian posterior $\pi(\theta|x_1,\dots,x_n)$.
3: Generate posterior pseudo-samples $x^{(j)}_{n+1},\dots,x^{(j)}_{n+T} \sim f(\cdot|\theta^{(j)}_{\text{post}})$.
4: Draw $w^{(j)}_1,\dots,w^{(j)}_n \sim \text{Exp}(1)$ and $w^{(j)}_{n+1},\dots,w^{(j)}_{n+T} \sim \text{Exp}(T/c)$
5: Set $\tilde{\theta}^{(j)} = \arg\max_\theta \left\{\sum_{i=1}^n w^{(j)}_i \log f(x_i|\theta) + \sum_{i=1}^T w^{(j)}_{n+i} \log f(x^{(j)}_{n+i}|\theta)\right\}$.
6: **end for**
7: **return** Samples $\{\tilde{\theta}^{(j)}\}_{j=1}^N$.

To prove the above statement, we follow the steps from the proof of Theorem 3.17 in Supplementary Material B .5. However, instead of the calculation in (3.51), in the case of Algorithm 9 we have

$$\begin{aligned}\mathbb{E}\left\{\sum_{j=n+1}^{n+T} w_j \nabla_\theta \ell(x_j,\hat{\theta})\right\} =& \mathbb{E}_{\pi(\theta|x_{1:n})} \sum_{j=n+1}^{n+T} \left[\mathbb{E}_{f(x_j|\theta)}\left\{w_j \nabla_\theta \ell(x_j,\hat{\theta})\Big|\theta\right\}\right] \\ =& c\int \nabla_\theta \left\{\log f(x|\bar{\theta})\right\}\left\{\int f(x|\theta)\pi(\theta|x_{1:n})\mathrm{d}\theta\right\}\mathrm{d}x = o_P(1).\end{aligned}$$

The last equality follows from the fact that the posterior concentrates around the maximum likelihood estimator as $n \to \infty$.

C Proofs of theoretical results presented in Section 3

C .1 Proof of Proposition 3.14

Proof. Fix $\epsilon > 0$. Recall that we denote by $P^{HB,p}(\cdot|x_{1:K,1:n_k})$ the distribution of samples drawn from Algorithm 6. Let $p^{HB,p}(\cdot|x_{1:K,1:n_k})$ denote its corresponding density. Let $\tilde{\theta} = (\tilde{\theta}_1,\dots,\tilde{\theta}_K)$ and $\theta^* = (\theta^*_1,\dots,\theta^*_K)$, and let $p(\lambda|\tilde{\theta})$ be the Bayesian posterior of λ given $\tilde{\theta}$ in model 3.22. We have

$$\begin{aligned}p^{HB,p}(\tilde{\lambda}|x_{1:K,1:n_k}) &= \int p(\tilde{\lambda}|\tilde{\theta}) p^{HB,p}(\tilde{\theta}|x_{1:K,1:n_k})\mathrm{d}\tilde{\theta} \\ &= \int_{A_\epsilon} p(\tilde{\lambda}|\tilde{\theta}) p^{HB,p}(\tilde{\theta}|x_{1:K,1:n_k})\mathrm{d}\tilde{\theta} + \int_{A^c_\epsilon} p(\tilde{\lambda}|\tilde{\theta}) p^{HB,p}(\tilde{\theta}|x_{1:K,1:n_k})\mathrm{d}\tilde{\theta} \quad (3.52)\end{aligned}$$

where $A_\epsilon = \{\tilde{\theta} : \|\tilde{\theta} - \theta^*\| \le \epsilon\}$. Since $\tilde{\theta}$ under $P^{HB,p}$ converges weakly to δ_{θ^*}, then by the continuity assumption for any λ the first integral in (3.52) converges to $p(\tilde{\lambda}|\theta^*)$. Since $\sup_{\lambda\in\Lambda,\theta\in\Theta} p(\lambda|\theta) < \infty$, the second integral converges to 0. Finally we get that $p^{HB,p}(\tilde{\lambda}|\tilde{\theta}) \to$

$p(\tilde{\lambda}|\theta^*)$ as $n \to \infty$. By similar arguments the Bayesian posterior $p\left(\lambda|x_{1:K,1:n_k}\right)$ has the same limiting distribution. □

C .2 Additional information for Section 3 .2 and proof of Theorem 3.15

Recall that in Section 3 .2 we considered the following hierarchical model

$$\begin{aligned} x_{k,1}, \dots x_{k,n_k} &\sim f(x_{k,i}|\theta_k), \qquad x_{k,i} \in \mathbb{R}^s, \quad k = 1, \dots, K, \\ \theta_k &\sim g(\theta_k|\lambda), \qquad \theta_k \in \Theta \subseteq \mathbb{R}^d, \quad k = 1, \dots, K, \\ \lambda &\sim p(\lambda). \end{aligned}$$

We present the method in Algorithm 10. We proposed using for this model a version of Posterior Bootstrap, in which $\tilde{\lambda}$ is, as opposed to Algorithm 6, obtained via bootstrapping. We present this method in Algorithm 10.

Algorithm 10 Posterior Bootstrap for hierarchical models via prior penalization under the large K scenario, returns samples from $P^{HB,pp}(\cdot|x_{1:K,1:n_k})$

hyperparameters: $w_g, w_{0,[1]}, \dots, w_{0,[K]}$.
for $k = 1, \dots, K$ **do**
  Set $\hat{\theta}_k = \arg\max_{\theta_k} \sum_{i=1}^{n_k} \log f(x_{k,i}|\theta_k)$.
end for
for j = 1, . . . , N **do** ▷ (in parallel)
  Draw $z_1^{(j)}, \dots, z_K^{(j)} \sim \text{Exp}(1)$.
  Set $\bar{\lambda}^{(j)} = \arg\max_\lambda \left\{ \sum_{k=1}^K z_k^{(j)} \log g(\hat{\theta}_k|\lambda) + w_g^T \log p\left(\lambda\right) \right\}$.
  for $k = 1, \dots, K$ **do**
    Draw $w_1^{(j)}, \dots, w_{n_k}^{(j)} \sim \text{Exp}(1)$.
    Set $\tilde{\theta}_k^{(j)} = \arg\max_{\theta_k} \left\{ \sum_{i=1}^{n_k} w_i^{(j)} \log f(x_{k,i}|\theta_k) + w_{0,[k]}^T \log g\left(\theta_k|\bar{\lambda}^{(j)}\right) \right\}$.
  end for
  Draw $v_1^{(j)}, \dots, v_K^{(j)} \sim \text{Exp}(1)$.
  Set $\tilde{\lambda}^{(j)} = \arg\max_\lambda \left\{ \sum_{k=1}^K v_k^{(j)} \log g(\tilde{\theta}_k^{(j)}|\lambda) + w_g^T \log p\left(\lambda\right) \right\}$.
end for
return Samples $\left\{ \left(\tilde{\lambda}^{(j)}, \tilde{\theta}_1^{(j)}, \dots, \tilde{\theta}_K^{(j)} \right) \right\}_{j=1}^N$.

Before proceeding with the proof, we present the regularity conditions for Theorem 3.15.

Assumption 3.21. (Data-generating mechanism and log likelihood function ℓ) The mechanism of generating data $x_{1:K,1:n_k}$ follows (3.27). The log likelihood function $\ell(x, \theta) = \log f(x|\theta)$ is measurable and bounded from above, and

$$Q^* \left\{ E_{p^*} \left| \ell(X, \theta) \right| < \infty, \theta \in \Theta \right\} = 1$$

Moreover, let $\mathcal{P}$ be a space of probability measures on $\mathbb{R}^s$. We assume that the map $\theta^* : \mathcal{P} \to \Theta \subseteq \mathbb{R}^d$ such that

$$P^* \to \theta^*(P^*) = \arg\max_{\theta} \{E_{p^*} \log f(X|\theta)\}$$

is measurable with respect to the σ-field of Borel sets on Θ. We assume also that θ^* has a density $q^*_{\theta^*}$ with respect to the Lebesgue measure, defined on a compact subset of Θ.

Assumption 3.22. (Log likelihood function ℓ_g and identifiability of λ) The log likelihood function $\ell_g(\theta, \lambda) = \log g(\theta|\lambda)$ is measurable and bounded from above, with $E_{Q^*}|\ell_g(\Theta^*, \lambda)| < \infty$ for all $\lambda \in \Lambda$. Moreover, there exists a unique maximizing parameter value λ^*

$$\lambda^* = \arg\max_{\lambda} E_{Q^*} \{\ell_g(\Theta^*|\lambda)\},$$

and further for all $\delta > 0$ there exists an $\epsilon > 0$ such that

$$\lim_{K\to\infty} Q^* \left[\sup_{\|\lambda-\lambda^*\|\geq\delta} \frac{1}{K} \sum_{k=1}^{K} \{\ell_g(\theta^*_k, \lambda) - \ell_g(\theta^*_k, \lambda^*)\} \leq -\epsilon \right] = 1.$$

Assumption 3.23. (Identifiability of θ). For all $\delta > 0$ there exists an $\epsilon > 0$ such that

$$\lim_{n\to\infty} P^* \left[\sup_{\|\theta-\theta^*\|\geq\delta} \frac{1}{n} \sum_{i=1}^{n} \{\ell(x_i, \theta) - \ell(x_i, \theta^*)\} \leq -\epsilon \right] = 1.$$

almost surely with respect to Q^*.

Assumption 3.24. (Smoothness of l_g). There is an open ball B containing λ^* such that $\ell_g(\theta, \lambda)$ is three times continuously differentiable with respect to $\lambda \in B$ for almost all $\theta \in \Theta$. Furthermore, there exist measurable functions G_1, G_2 and G_3 such that for $\lambda \in B$

$$\begin{aligned}
\left|\frac{\partial \ell_g(\theta, \lambda)}{\partial \lambda_j}\right| &\leq G_1(\theta) \qquad \mathbb{E}_{q*}\left\{G_1(\Theta^*)^4\right\} < \infty, \\
\left|\frac{\partial^2 \ell_g(\theta, \lambda)}{\partial \lambda_j \partial \lambda_l}\right| &\leq G_2(\theta) \qquad \mathbb{E}_{q*}\left\{G_2(\Theta^*)^4\right\} < \infty, \\
\left|\frac{\partial^3 \ell_g(\theta, \lambda)}{\partial \lambda_j \partial \lambda_l \partial \lambda_m}\right| &\leq G_3(\theta) \qquad \mathbb{E}_{q*}\{G_3(\Theta^*)\} < \infty.
\end{aligned}$$

Assumption 3.25. (Smoothness of ℓ_g with respect to θ). For any $\lambda \in \Lambda$, the function $\ell_g(\theta, \lambda)$ is measurable as a function of θ, and three times continuously differentiable with respect to θ. Besides, $\ell_g(\theta, \lambda)$ is jointly bounded as a function of $\lambda \in \Lambda, \theta \in \Theta$. Furthermore,

there exist $L_1, L_2, L_3 > 0$ such that for any $\lambda \in \Lambda$ and any $\theta_1, \theta_2 \in \Theta$

$$
\begin{aligned}
|\ell_g(\theta_1, \lambda) - \ell_g(\theta_2, \lambda)| &\leq L_1 \|\theta_1 - \theta_2\|, \\
\|\nabla_\lambda \ell_g(\theta_1, \lambda) - \nabla_\lambda \ell_g(\theta_2, \lambda)\| &\leq L_2 \|\theta_1 - \theta_2\|, \\
\|D^2_\lambda \ell_g(\theta_1, \lambda) - D^2_\lambda \ell_g(\theta_2, \lambda)\| &\leq L_3 \|\theta_1 - \theta_2\|.
\end{aligned}
$$

Assumption 3.26. (Smoothness of ℓ). The function $\ell(x, \theta)$ is three times continuously differentiable with respect to $\theta \in \Theta$ almost surely on x. We assume that there exist measurable functions H_1, H_2 and H_3, and a constant $C > 0$ such that for $\theta \in \Theta$

$$
\begin{aligned}
\Big|\frac{\partial \ell(x,\theta)}{\partial \theta_j}\Big| &\leq H_1(x) \quad \mathbb{E}_{Q^*}\left[\mathbb{E}_{P*}\left\{H_1(X)^4 | P^*\right\}\right] < C, \\
\Big|\frac{\partial^2 \ell(x,\theta)}{\partial \theta_j \partial \theta_l}\Big| &\leq H_2(x) \quad \mathbb{E}_{Q^*}\left[\mathbb{E}_{P*}\left\{H_2(X)^4 | P^*\right\}\right] < C, \\
\Big|\frac{\partial^3 \ell(x,\theta)}{\partial \theta_j \partial \theta_l \partial \theta_m}\Big| &\leq H_3(x) \quad \mathbb{E}_{Q^*}\left[\mathbb{E}_{P*}\left\{H_3(X) | P^*\right\}\right] < C.
\end{aligned}
$$

for any $k = 1, 2, \ldots$, and the constant C does not depend on k.

Assumption 3.27. (Positive definiteness of information matrices for λ). The matrices $I_g(\lambda)$ and $J_g(\lambda)$ defined in (3.28) are positive definite for $\lambda \in B$ with all elements finite.

Assumption 3.28. (Positive definiteness of information matrices for θ and linear independence of partial derivatives). The matrices $I(\theta)$ and $J(\theta)$, defined as

$$
\begin{aligned}
I(\theta) =& E_{Q^*}\left(E_{P^*}\left[\{\nabla_\theta \log f(X|\theta)\}\{\nabla_\theta \log f(X|\theta)\}^T | P^*\right]\right), \\
J(\theta) =& -E_{Q^*}\left(E_{P^*}\left[D^2_\theta \{\log f(X|\theta)\} | P^*\right]\right),
\end{aligned}
$$

are positive definite for $\theta \in \Theta$ with all elements finite. Furthermore, $\det J(\theta)$ and $\det I(\theta)$ are bounded away from 0 on any compact subset of Θ. Besides, the first and second partial derivatives

$$
\left\{\frac{\partial \ell(x,\theta)}{\partial \theta_p}, \frac{\partial^2 \ell(x,\theta)}{\partial \theta_r \partial \theta_s}\right\}, \quad 1 \leq p \leq d, 1 \leq r \leq s \leq d
$$

are linearly independent as functions of x at any $\theta \in \Theta$.

Assumption 3.29. (Smoothness of log prior density). The function $\log p(\lambda)$ is measurable, upper-bounded on λ, and three times continuously differentiable in the neighbourhood of λ^*.

The strategy of the proof is as follows. We first prove consistency of $\tilde{\lambda}$, and then its

asymptotic normality. The key element in establishing consistency of $\tilde{\lambda}$ is obtaining the upper bound on the rate of convergence of $\tilde{\theta}_k$ to θ_k^*. We obtain this upper bound in Lemma 3.31, which relies largely on the multivariate Berry–Esseen theorem with explicit constants, proved by Raič (2019).

In the proof of Lemma 3.31 we use the following auxiliary result.

Lemma 3.30. *Let ϵ_n be a deterministic sequence depending on n such that $\epsilon_n = o(1)$ and*

$$\epsilon_n n^{1/2}/(\log n)^{1/2} \to \infty$$

as $n \to \infty$. Then for any constant C we have

$$1 - \Phi_1\left(C\epsilon_n n^{1/2}\right) = o\left(n^{-1/2}\right),$$

where Φ_1 denotes a cumulative distribution function of a univariate standard normal distribution.

Proof. We use the following property of the univariate standard normal distribution:

$$1 - \Phi_1(y) \leq \frac{1}{y(2\pi)^{1/2}} \exp\left(-y^2/2\right), \quad y > 0. \tag{3.53}$$

Therefore

$$1 - \Phi_1\left(C\epsilon_n n^{1/2}\right) \leq n^{-1/2} \frac{\exp\left(-C^2\epsilon_n^2 n/2 - \log \epsilon_n\right)}{C(2\pi)^{1/2}}. \tag{3.54}$$

By the assumption that $\epsilon_n n^{1/2} \to \infty$ as $n \to \infty$, we have $\epsilon_n > n^{-1/2}$ for large enough n. Thus as n goes to infinity

$$C^2\epsilon_n^2 n/2 + \log \epsilon_n > C^2\epsilon_n^2 n/2 - (\log n)/2 = \log n/2 \left\{C^2\epsilon_n^2 n/\log n - 1\right\} \to \infty,$$

which implies that the right hand side of (3.54) is $o\left(n^{-1/2}\right)$. □

Lemma 3.31. *Let ϵ_n be a deterministic sequence depending on n such that $\epsilon_n = o(1)$ and*

$$\epsilon_n n^{1/2}/(\log n)^{1/2} \to \infty$$

as $n \to \infty$. Let Then under Assumptions 3.21–3.29 for any $k = 1, \ldots, K$ we have

$$pr\left(\|\tilde{\theta}_k - \theta_k^*\| > \epsilon_n\right) = O\left(n^{-1/2}\right), \tag{3.55}$$

where $n = \min_{k=1,\ldots,K} n_k$, and the probability in (3.55) is defined as

$$pr\left(\|\tilde{\theta}_k - \theta_k^*\| > \epsilon_n\right) = E_{Q^*}\left[E_{P^*_{1:K}}\left\{P^{HB,pp}\left(\|\tilde{\theta}_k - \theta_k^*\| > \epsilon_n | x_{1:K,1:n_k}\right) | P^*_{1:K}\right\}\right].$$

Proof. We fix $k \leq K$. We notice that $\tilde{\theta}_k$ depends also on data $x_{j,1:n_j}$ for $j \neq k$. In the remaining part of the proof for simplicity of notation we write $P^* = P^*_{1:K}$, as well as $x = x_{1:K,1:n_k}$, denoting by X its corresponding random variable. We also drop the subscript k. We have

$$\begin{aligned}
\text{pr}\left(\|\tilde{\theta} - \theta^*\| > \epsilon_n\right) \leq & \text{pr}\left(\|\tilde{\theta} - \hat{\theta}\| > \epsilon_n/2\right) + \text{pr}\left(\|\hat{\theta} - \theta^*\| > \epsilon_n/2\right) \\
= & E_{Q^*}\left[E_{P^*}\left\{P^{HB,pp}\left(\|\tilde{\theta} - \hat{\theta}\| > \epsilon_n/2 | x\right) | P^*\right\}\right] \\
& + E_{Q^*}\left[E_{P^*}\left(\|\hat{\theta} - \theta^*\| > \epsilon_n/2\right) | P^*\right].
\end{aligned}$$

We start by dealing with the first term in the above equation. To this end we expand $n^{1/2}\left(\tilde{\theta} - \hat{\theta}\right)$ as follows. We define

$$\tilde{L}_n(\theta) = \sum_{j=1}^{n} w_j \ell(x_j, \theta) + w_0^T \ell_g(\theta, \bar{\lambda}).$$

Using the Taylor expansion with the integral remainder we get

$$0 = n^{-1/2}\nabla_\theta \tilde{L}_n(\tilde{\theta}) = n^{-1/2}\nabla_\theta \tilde{L}_n(\hat{\theta}) - \tilde{J}_{w,n} n^{1/2}(\tilde{\theta} - \hat{\theta}),$$

where

$$\begin{aligned}
\tilde{J}_{w,n} = & -\int_0^1 (1-u)\frac{D^2_\theta \tilde{L}_n(u\hat{\theta} + (1-u)\tilde{\theta})}{n}\mathrm{d}u = \sum_{j=1}^{n} w_j \int_0^1 (1-u)\frac{D^2_\theta \ell(x_j, u\hat{\theta} + (1-u)\tilde{\theta})}{n}\mathrm{d}u \\
& + w_0^T \int_0^1 (1-u)\frac{D^2_\theta \ell_g(u\hat{\theta} + (1-u)\tilde{\theta}, \bar{\lambda})}{n}\mathrm{d}u.
\end{aligned}$$

Consequently

$$I_n^{-1/2}\tilde{J}_{w,n}n^{1/2}(\tilde{\theta}-\hat{\theta}) = I_n^{-1/2}\frac{\sum_{j=1}^n(w_j-1)\nabla_\theta\ell(x_j,\hat{\theta}) + w_0\cdot\nabla_\theta\ell_g(\hat{\theta},\bar{\lambda})}{n^{1/2}}, \tag{3.56}$$

where

$$I_n = \frac{1}{n}\sum_{j=1}^n \nabla_\theta\ell(x_j,\hat{\theta})\nabla_\theta\ell(x_j,\hat{\theta})^T.$$

Let $H_{w,n} = I_n^{-1/2}\tilde{J}_{w,n}$. Then

$$\|H_{w,n}^{-1}\|^2 = \sup_{\|y\|=1} y^T\tilde{J}_{w,n}^{-1}I_n\tilde{J}_{w,n}^{-1}y = \|\tilde{J}_{w,n}^{-1}I_n\tilde{J}_{w,n}^{-1}\|. \tag{3.57}$$

For any $c_1 > 0$ we have

$$\begin{aligned}
P^{HB,pp}\left(\|\tilde{\theta}-\hat{\theta}\| > \epsilon_n/2|x\right) =& P^{HB,pp}\left(n^{1/2}\|\tilde{\theta}-\hat{\theta}\| > n^{1/2}\epsilon_n/2|x\right)\\
=& P^{HB,pp}\left(\|H_{w,n}^{-1}H_{w,n}n^{1/2}(\tilde{\theta}-\hat{\theta})\| > n^{1/2}\epsilon_n/2|x\right)\\
\leq& P^{HB,pp}\left\{\|H_{w,n}n^{1/2}(\tilde{\theta}-\hat{\theta})\| > n^{1/2}\epsilon_n/(2c_1)|x\right\}\\
&+ P^{HB,pp}\left(\|H_{w,n}^{-1}\| > c_1\right).
\end{aligned}$$

Using (3.56) and (3.57), we have

$$\begin{aligned}
&P^{HB,pp}\left(\|\tilde{\theta}-\hat{\theta}\| > \epsilon_n/2|x\right)\\
\leq& P^{HB,pp}\left\{I_n^{-1/2}\frac{\sum_{j=1}^n(w_j-1)\nabla_\theta\ell(x_j,\hat{\theta}) + w_0\cdot\nabla_\theta\ell_g(\hat{\theta},\bar{\lambda})}{n^{1/2}} > n^{1/2}\epsilon_n/(2c_1)\Big|x\right\}\\
&+ P^{HB,pp}\left(\|\tilde{J}_{w,n}^{-1}I_n\tilde{J}_{w,n}^{-1}\| > c_1^2\Big|x\right)
\end{aligned} \tag{3.58}$$

In order to deal with the first part of the above expression, we let $C(\Theta)$ denote the set

of measurable convex sets $A \subseteq \Theta$, and let Z follow the d-variate normal distribution. Then

$$
\begin{aligned}
& P^{HB,pp}\left\{ I_n^{-1/2} \frac{\sum_{j=1}^n (w_j - 1)\nabla_\theta \ell(x_j, \hat{\theta}) + w_0 \cdot \nabla_\theta \ell_g(\hat{\theta}, \bar{\lambda})}{n^{1/2}} > n^{1/2}\epsilon_n/(2c_1) \Big| x \right\} \\
\leq & \sup_{A \in C(\Theta)} \left| P^{HB,pp}\left\{ I_n^{-1/2} \frac{\sum_{j=1}^n (w_j - 1)\nabla_\theta \ell(x_j, \hat{\theta}) + w_0 \cdot \nabla_\theta \ell_g(\hat{\theta}, \bar{\lambda})}{n^{1/2}} \in A \Big| x \right\} \right. \\
& \qquad \left. - \operatorname{pr}\left\{ Z + I_n^{-1/2} \frac{w_0 \cdot \nabla_\theta \ell_g(\hat{\theta}, \bar{\lambda})}{n^{1/2}} \in A \Big| x \right\} \right| \\
& + \operatorname{pr}\left\{ Z + I_n^{-1/2} \frac{w_0 \cdot \nabla_\theta \ell_g(\hat{\theta}, \bar{\lambda})}{n^{1/2}} > n^{1/2}\epsilon_n/(2c_1) \Big| x \right\} \\
= & \sup_{A \in C(\Theta)} \left| P^{HB,pp}\left\{ I_n^{-1/2} \frac{\sum_{j=1}^n (w_j - 1)\nabla_\theta \ell(x_j, \hat{\theta})}{n^{1/2}} \in A \Big| x \right\} - \operatorname{pr}(Z \in A) \right| \qquad (3.59) \\
& \qquad + \operatorname{pr}\left[Z > \left\{ n^{1/2}\epsilon_n/(2c_1) - I_n^{-1/2} \frac{w_0 \cdot \nabla_\theta \ell_g(\hat{\theta}, \bar{\lambda})}{n^{1/2}} \right\} \Big| x \right] \qquad (3.60)
\end{aligned}
$$

To obtain the upper bound on (3.59), we now apply Theorem 1.1 of Raič (2019) to the expression

$$
I_n^{-1/2} \frac{\sum_{j=1}^n (w_j - 1)\nabla_\theta \ell(x_j, \hat{\theta})}{n^{1/2}}.
$$

Recall that d is the dimension of space Θ. We get that for a constant $c_0 = 42d^{1/4} + 16$ we have

$$
\begin{aligned}
& \sup_{A \in C(\Theta)} \left| P^{HB,pp}\left\{ I_n^{-1/2} \frac{\sum_{j=1}^n (w_j - 1)\nabla_\theta \ell(x_j, \hat{\theta})}{n^{1/2}} \in A \Big| x \right\} - \operatorname{pr}(Z \in A) \right| \\
\leq & \frac{c_0}{n^{1/2}} \frac{1}{n} \sum_{j=1}^n \mathbb{E}_{w_{1:n}} \left\{ \| (w_j - 1) I_n^{-1/2} \nabla_\theta \ell(x_j, \hat{\theta}) \|^3 \right\} \qquad (3.61) \\
\leq & \frac{2c_0}{n^{1/2}} \frac{\| I_n^{-1/2} \|^3}{n} \sum_{j=1}^n \| \nabla_\theta \ell(x_j, \hat{\theta}) \|^3,
\end{aligned}
$$

where constant 2 in the last equation comes from the fact that $E\{(w_j - 1)^3\} = 2$. To bound (3.60), note that on a set such that $\| I_n^{-1/2} \| \leq c_2$, we have for some $c_3 > 0$

$$
\operatorname{pr}\left[Z > \left\{ n^{1/2}\epsilon_n/(2c_1) - I_n^{-1/2} \frac{w_0 \cdot \nabla_\theta \ell_g(\hat{\theta}, \bar{\lambda})}{n^{1/2}} \right\} \Big| x \right] \leq d\left\{ 1 - \Phi_1\left(c_3 \epsilon_n n^{1/2} \right) \right\} = o(n^{-1/2}), \qquad (3.62)
$$

where the last equality follows from Lemma 3.30.

To show that $\operatorname{pr}\left(\| \tilde{\theta} - \hat{\theta} \| > \epsilon_n/2 \right) = o(n^{-1/2})$, we will now combine (3.58), (3.59) and (3.60), and the upper bounds obtained in (3.61) and (3.62). Using additionally Assump-

tions 3.26 and 3.28, we have

$$
\begin{aligned}
& E_{Q^*}\left[E_{P^*}\left\{P^{HB,pp}\left(\|\tilde{\theta}-\hat{\theta}\|>\epsilon_n/2|x\right)|P^*\right\}\right] \\
\leq & E_{Q^*}\left[E_{P^*}\left\{\frac{2c_0}{n^{1/2}}\frac{\|I_n^{-1/2}\|^3}{n}\sum_{j=1}^{n}\|\nabla_\theta \ell(X_j,\hat{\theta})\|^3\Big|P^*\right\}\right] \\
& \quad + E_{Q^*}\left[E_{P^*}\left\{P^{HB,pp}\left(\|\tilde{J}_{w,n}^{-1}I_n\tilde{J}_{w,n}^{-1}\|>c_1^2\Big|X\right)|P^*\right\}\right] \\
& \quad + E_{Q^*}\left[E_{P^*}\left\{\|I_n^{-1/2}\|\geq c_2|P^*\right\}\right]+o(n^{-1/2}) \\
\leq & \frac{2c_0}{n^{1/2}}E_{Q^*}\left[E_{P^*}\left\{\|I_n^{-1/2}\|^3 H_1(X)^3|P^*\right\}\right] \\
& \quad + E_{Q^*}\left[E_{P^*}\left\{P^{HB,pp}\left(\|\tilde{J}_{w,n}^{-1}I_n\tilde{J}_{w,n}^{-1}\|>c_1^2\Big|X\right)|P^*\right\}\right] \\
& \quad + E_{Q^*}\left[E_{P^*}\left\{\|I_n^{-1/2}\|\geq c_2|P^*\right\}\right]+o(n^{-1/2}) \\
\leq & E_{Q^*}\left[E_{P^*}\left\{P^{HB,pp}\left(\|\tilde{J}_{w,n}^{-1}I_n\tilde{J}_{w,n}^{-1}\|>c_1^2\Big|X\right)|P^*\right\}\right] \\
& \quad + E_{Q^*}\left[E_{P^*}\left\{\|I_n^{-1/2}\|\geq c_2|P^*\right\}\right]+O(n^{-1/2}).
\end{aligned}
$$

It remains to show that for some constant $c_4 > 0$ the probability that $\|\tilde{J}_{w,n}^{-1}\| c >_4$, as well as the probabilities of $\|I_n^{-1/2}\| > c_4$ and $\|I_n\| > c_4$, are of order $o(n^{-1/2})$. Recalling the definitions of $\tilde{J}_{w,n}$ and I_n, and using the Chebyshev's inequality, we get that under Assumptions 3.26 and 3.28 each of these events is of order $O(n^{-1})$. This completes the proof of the claim for $\text{pr}\left(\|\tilde{\theta}-\hat{\theta}\|>\epsilon_n/2|X\right)$.

Using similar reasoning we now find an upper bound for $\text{pr}\left(\|\hat{\theta}-\theta_*\|>\epsilon_n/2\right)$. To this end, similarly to the previous case, we define $L_n(\theta)=\sum_{j=1}^n \ell(x_j,\theta)$ and write

$$
0 = n^{-1/2}\nabla_\theta L_n(\hat{\theta}) = n^{-1/2}\nabla_\theta L_n(\theta^*) - \tilde{J}_n n^{1/2}(\theta^*-\hat{\theta}),
$$

where

$$
\tilde{J}_n = -\int_0^1 (1-u)\frac{D_\theta^2 \tilde{L}_n(u\theta^*+(1-u)\hat{\theta})}{n}\mathrm{d}u
$$

Therefore

$$
I(\theta^*)^{-1/2}\tilde{J}_n n^{1/2}(\hat{\theta}-\theta^*) = I(\theta^*)^{-1/2}n^{-1/2}\sum_{j=1}^{n}\nabla_\theta \ell(x_i,\theta^*).
$$

For Z_n following the normal distribution $N\left\{0, \tilde{J}_n^{-1} I(\theta^*) \tilde{J}_n^{-1}\right\}$ we have

$$\begin{aligned}
\operatorname{pr}\left(\|\theta^* - \hat{\theta}\| > \epsilon_n/2\right) =& \operatorname{pr}\left(n^{1/2}\|\hat{\theta} - \theta^*\| > \epsilon_n n^{1/2}/2\right) \\
&\leq \sup_{A \in C(\Theta)} \left|\operatorname{pr}\left\{n^{1/2}\left(\hat{\theta} - \theta^*\right) \in A\right\} - \operatorname{pr}\left(Z_n \in A\right)\right| \\
&\quad + \operatorname{pr}\left(\|Z_n\| > \epsilon_n n^{1/2}/2\right).
\end{aligned}$$

Let Z' follow the d-variate standard normal distribution. We then have

$$\begin{aligned}
&\sup_{A \in C(\Theta)} \left|\operatorname{pr}\left\{n^{1/2}\left(\hat{\theta} - \theta^*\right) \in A\right\} - \operatorname{pr}\left(Z_n \in A\right)\right| \\
=& \sup_{A \in C(\Theta)} \left|\operatorname{pr}\left\{I(\theta^*)^{-1/2}\tilde{J}_n n^{1/2}\left(\hat{\theta} - \theta^*\right) \in A\right\} - \operatorname{pr}\left(I(\theta^*)^{-1/2}\tilde{J}_n Z_n \in A\right)\right| \\
=& \sup_{A \in C(\Theta)} \left|\operatorname{pr}\left\{I(\theta^*)^{-1/2} n^{-1/2} \sum_{j=1}^{n} \nabla_\theta \ell(X_j, \theta^*) \in A\right\} - \operatorname{pr}\left(Z' \in A\right)\right| \\
\leq& \frac{c_0}{n^{1/2}} \frac{1}{n} \sum_{j=1}^{n} \mathbb{E}_{Q^*}\left\{\|I(\theta^*)^{-1/2} \nabla_\theta \ell(X_j, \theta^*)\|^3\right\} \\
\leq& \frac{c_0}{n^{1/2}} \mathbb{E}_{Q^*}\left\{\|I(\theta^*)^{-1/2} \nabla_\theta \ell(X, \theta^*)\|^3\right\} = o(n^{-1/2}),
\end{aligned}$$

where the second last equality is again obtained by applying Theorem 1.1 of Raič (2019). To bound $\operatorname{pr}\left(\|Z_n\| > \epsilon_n n^{1/2}/2\right)$ by $o\left(n^{-1/2}\right)$, we again use Lemma 3.30 on the set such that $\|\tilde{J}_n\| < c_5$ for some constant c_5. Similarly to the previous case, by Chebyshev's inequality the probability that $\|\tilde{J}_n\| < c_5$ is $O(n^{-1})$. Finally, since the constants appearing in the proof do not depend on k, for any $k = 1, \ldots, K$ we have

$$\operatorname{pr}\left(\|\tilde{\theta}_k - \theta_k^*\| > \epsilon_n\right) = O\left(n^{-1/2}\right),$$

as required. □

Proof of Theorem 3.15

Proof. The strategy of the proof is as follows. We first prove consistency of $\tilde{\lambda}$, that is, we will show that $\|\tilde{\lambda} - \lambda^*\| = o_P(1)$. Then we will prove asymptotic normality.

Let ϵ_n be a deterministic sequence depending on n such that $\epsilon_n n^{1/2}/(\log n)^{1/2} \to \infty$ as $n \to \infty$ and $\epsilon_n K^{1/2} = o(1)$, which is possible by the assumption that $K = O\left(n^{1/2}/\log n\right)$. Consider a set Ω_n such that on Ω_n we have $\sup_{1 \leq k \leq K} \|\tilde{\theta}_k - \theta_k^*\| \leq \epsilon_n$. By Lemma 3.31 we

get that

$$\mathrm{pr}\left(\Omega_n^c\right) \leq K \sup_{1\leq k\leq K} \mathrm{pr}\left(\|\tilde{\theta}_k - \theta_k^*\| \leq \epsilon_n\right) \leq KO\left(n^{-1/2}\right) = o(1).$$

We define

$$l(\lambda) = \frac{1}{K}\left\{\sum_{k=1}^{K} v_k \ell_g(\tilde{\theta}_k, \lambda) + w_g^T \log p(\lambda)\right\}, \quad l^*(\lambda) = \frac{1}{K}\left\{\sum_{k=1}^{K} v_k \ell_g(\theta_k^*, \lambda) + w_g^T \log p(\lambda)\right\}$$

and $\tilde{\lambda}_{\theta^*} = \arg\max_\lambda l^*(\lambda)$. Recall that $\tilde{\lambda} = \arg\max_\lambda l(\lambda)$. Since $\theta_1^*, \ldots, \theta_K^*$ are i.i.d. with respect to Q^*, we can use Assumption 3.22, as well as analogous arguments to those used in Lemma 3.20 to get that $\|\tilde{\lambda}_{\theta^*} - \lambda^*\| = o_P(1)$ and $\|\tilde{\lambda}_{\theta^*} - \hat{\lambda}_{\theta^*}\| = o_P(1)$. Therefore, to prove identifiability of $\tilde{\lambda}$, it is enough to show that $\|\tilde{\lambda}_{\theta^*} - \tilde{\lambda}\| = o_P(1)$.

Note that by Assumption 3.25 on Ω_n we have

$$|l(\lambda) - l^*(\lambda)| \leq \frac{1}{K}\sum_{k=1}^{K} v_k \left| \left\{\ell_g(\tilde{\theta}_k, \lambda) - \ell_g(\theta_k^*, \lambda)\right\} \right| \leq \frac{L_1}{K}\sum_{k=1}^{K} v_k \epsilon_n.$$

The random variable $L_1/K \sum_{k=1}^{K} v_k \epsilon_n$ has expected value $L_1\epsilon_n$ and variance going to 0. Thus, for any fixed $\delta > 0$ we have

$$\mathrm{pr}\left(\sup_{\lambda\in\Lambda} |l(\lambda) - l^*(\lambda)| < \delta\right) \to 1$$

as K goes to infinity. What is more, by Assumption 3.22 probability of the event that $\arg\max_\lambda l^*(\lambda)$ is uniquely defined, goes to 1.

Now fix $\epsilon > 0$ and define a ball B centred at $\tilde{\lambda}_{\theta^*}$ with radius ϵ. Suppose we are on the event that the supremum of l^* is attained only at $\tilde{\lambda}_{\theta^*}$. By smoothness of l^*, for δ small enough, we can find $\epsilon_2 < \epsilon$ such that for a ball B_2 centred at $\tilde{\lambda}_{\theta^*}$ with radius ϵ_2 we have

$$\sup_{B_2} l^*(\lambda) - \sup_{\Lambda\setminus B} l^*(\lambda) > 3\delta.$$

Thus on the event such that $\sup_{\lambda\in\Lambda} |l(\lambda) - l^*(\lambda)| < \delta$ we have

$$\sup_{B_2} l(\lambda) - \sup_{\Lambda\setminus B} l(\lambda) > \delta.$$

Therefore $\tilde{\lambda}$ must belong to B. Since ϵ was arbitrary, we have $\|\tilde{\lambda}_{\theta^*} - \tilde{\lambda}\| = o_P(1)$, as required. This property is useful in the proof of asymptotic normality, which we present below.

We define $L_K(\lambda) = \sum_{k=1}^K v_k \ell_g(\tilde{\theta}_k, \lambda) + w_g^T \log p(\lambda)$. We can write $L_K(\lambda)$ defined above as

$$L_K(\lambda) = \sum_{k=1}^K v_k \ell_g(\theta_k^*, \lambda) + w_g^T \log p(\lambda) + h_K(\lambda),$$

where

$$h_K(\lambda) = \sum_{k=1}^K v_k \left\{ \ell_g(\tilde{\theta}_k, \lambda) - \ell_g(\theta_k^*, \lambda) \right\}.$$

Recall that $\|\tilde{\lambda} - \hat{\lambda}_{\theta^*}\| = o_P(1)$. We get the following expansion

$$0 = \frac{\nabla_\lambda L_K(\hat{\lambda}_{\theta^*})}{K^{1/2}} + \tilde{J}_{v,K}^{-1} K^{1/2} \left(\tilde{\lambda} - \hat{\lambda}_{\theta^*} \right),$$

where

$$\tilde{J}_{v,K} = -\int_0^1 (1-u) \frac{D_\lambda^2 \tilde{L}_K(u\hat{\lambda}_{\theta^*} + (1-u)\tilde{\lambda})}{K} \mathrm{d}u$$

Therefore

$$K^{1/2} \left(\tilde{\lambda} - \hat{\lambda}_{\theta^*} \right) = \tilde{J}_{v,K}^{-1} \frac{1}{K^{1/2}} \nabla_\lambda L_K(\hat{\lambda}_{\theta^*}). \tag{3.63}$$

Since $\sum_{k=1}^K \nabla_\lambda \ell_g(\theta_k^*, \hat{\lambda}_{\theta^*}) = 0$, we have

$$\frac{\nabla_\lambda L_K(\hat{\lambda}_{\theta^*})}{K^{1/2}} = \frac{1}{K^{1/2}} \sum_{k=1}^K (v_k - 1) \nabla_\lambda \ell_g(\theta_k^*, \hat{\lambda}_{\theta^*}) + \frac{1}{K^{1/2}} \left\{ w_g \cdot \nabla_\lambda \log p(\hat{\lambda}_{\theta^*}) + \nabla_\lambda h_K(\hat{\lambda}_{\theta^*}) \right\}.$$

What is more,

$$\begin{aligned} \tilde{J}_{v,K} = &- \sum_{k=1}^K v_k \int_0^1 (1-u) \frac{D_\lambda^2 \ell_g(\theta_k^*, u\hat{\lambda}_{\theta^*} + (1-u)\tilde{\lambda})}{K} \mathrm{d}u \\ &- \int_0^1 (1-u) \frac{D_\lambda^2 w_g^T \log p(u\hat{\lambda}_{\theta^*} + (1-u)\tilde{\lambda})}{K} \mathrm{d}u \\ &- \int_0^1 (1-u) \frac{D_\lambda^2 h_K(u\hat{\lambda}_{\theta^*} + (1-u)\tilde{\lambda})}{K} \mathrm{d}u \end{aligned} \tag{3.64}$$

By Assumptions 3.25 and 3.29 we have on Ω_n

$$\frac{1}{K^{1/2}} \left\{ w_g^T \nabla_\lambda \log p(\hat{\lambda}_{\theta^*}) + \nabla_\lambda h_K(\hat{\lambda}_{\theta^*}) \right\} = O_P\left(K^{-1/2}\right) + O_P\left(K^{1/2}\epsilon_n\right) = o_P(1), \tag{3.65}$$

where the last equality holds since we assumed that $K\epsilon_n = o(1)$, and the probability is

considered with respect to $v_{1:K}$. By the central limit theorem, as K goes to infinity, we have

$$\frac{1}{K^{1/2}}\sum_{k=1}^{K}(v_k-1)\nabla_\lambda \ell_g(\theta_k^*,\hat{\lambda}_{\theta^*}) \to N\left\{0, I_g(\lambda^*)\right\},$$

and therefore $K^{-1/2}\nabla_\lambda L_K(\hat{\lambda}_{\theta^*})$ by (3.65) also converges to $N\left\{0, I_g(\lambda^*)\right\}$ in distribution.

We similarly have in (3.64)

$$\int_0^1 (1-u)\frac{D_\lambda^2 w_g^T \log p(u\hat{\lambda}_{\theta^*}+(1-u)\tilde{\lambda})}{K}\mathrm{d}u + \int_0^1 (1-u)\frac{D_\lambda^2 h_K(u\hat{\lambda}_{\theta^*}+(1-u)\tilde{\lambda})}{K}\mathrm{d}u = o_P(1).$$

Thus, using the law of large numbers and the fact that $\|\tilde{\lambda}-\lambda^*\| = o_P(1)$ we get that $\tilde{J}_{v,K}$ converges to $J_g(\lambda^*)$ in probability. Finally, we use the Slutsky's theorem to get that (3.63) converges in distribution to $N\left\{0, J_g(\lambda^*)^{-1}I_g(\lambda^*)J_g(\lambda^*)^{-1}\right\}$ as required.

□

D Illustrations – further details

D .1 Additional information for Section 5 .1

Throughout the paper the Kolmogorov–Smirnov dissimilarity is computed using the `KS.diss` function from the `provenance` package (Vermeesch et al., 2016) in `R`, while the Bhattacharyya distance is computed using the `bhattacharyya.dist` function from the `fpc` package (Hennig and Imports, 2015). All our computations for the BayesBag algorithm are based on 50 bootstrapped datasets.

In Figures 3.5 – 3.7 we present additional results for model (3.32). Figure 3.5 confirms, using another measure of distance, our conclusion drawn from Figure 3.1 that it is optimal to set $w_0 = \sigma^2$, which coincides with formula (3.20). In Figures 3.6 and 3.7 we present also comparisons with BayesBag. For each experiment the size of the bootstrapped datasets was set according to the guidelines for the Gaussian location model provided in Section 2.2.1 of Huggins and Miller (2019). According to results presented in Figure 3.6, both Posterior Bootstrap and BayesBag capture well the uncertainty around the parameter, and under misspecification perform better than standard Bayesian inference. Figure 3.7 shows that our method outperforms BayesBag, which in turn performs better than Weighted Likelihood Bootstrap. We do not make comparisons with power posteriors, as we did for model (3.34). The reason for this is that in this example using power posteriors with parameter η set to

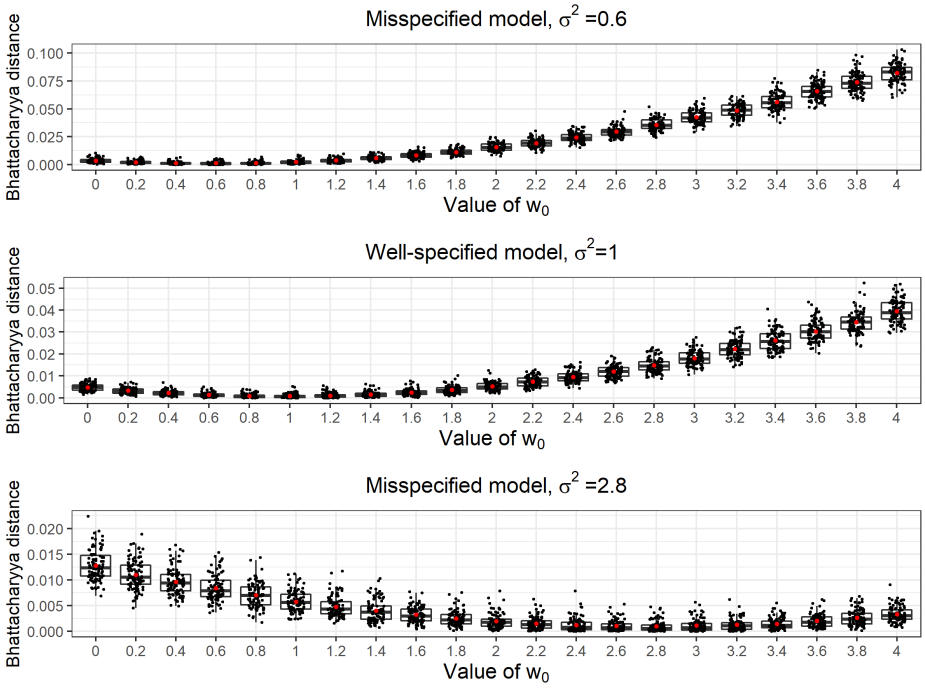


Figure 3.5: Bhattacharyya distance between samples drawn from Posterior Bootstrap for model (3.32), and samples from the Bayesian posterior for the correct model (3.33). Three scenarios are considered, $\sigma^2 = 0.6$ (top panel), $\sigma^2 = 1$ (middle panel), and $\sigma^2 = 2.8$ (bottom panel). Posterior Bootstrap simulations were performed using Algorithm 5 for a range of values of w_0. Every experiment was repeated on 100 simulated datasets with $N = 2000$ samples generated from Algorithm 5 for each w_0.

η^* proposed by Lyddon et al. (2019) is equivalent with drawing samples from the Bayesian posterior for the correct model.

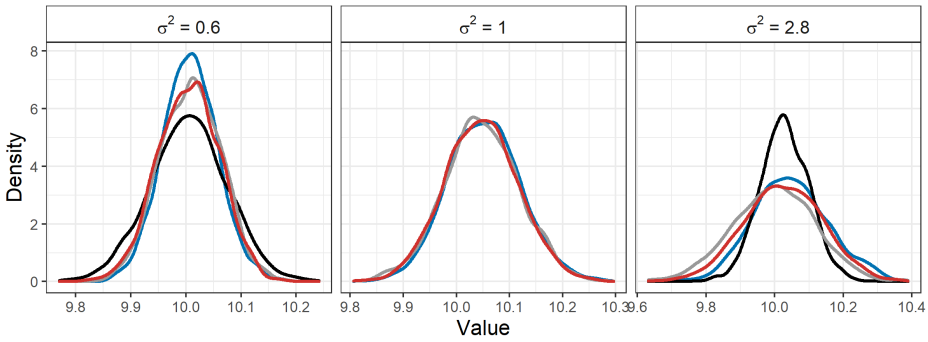


Figure 3.6: Kernel density estimates of posterior distribution for θ based on a single synthetic dataset, under standard Bayesian inference for the misspecified model (3.14) (black), Algorithm 5 with $w_0 = w_0^*$ (red), BayesBag (dark blue), and standard Bayesian inference for the correct model (3.15) (grey). We consider three scenarios $\sigma^2 = 0.6$ (left panel), $\sigma^2 = 1$ (middle panel), and $\sigma^2 = 2.8$ (right panel).

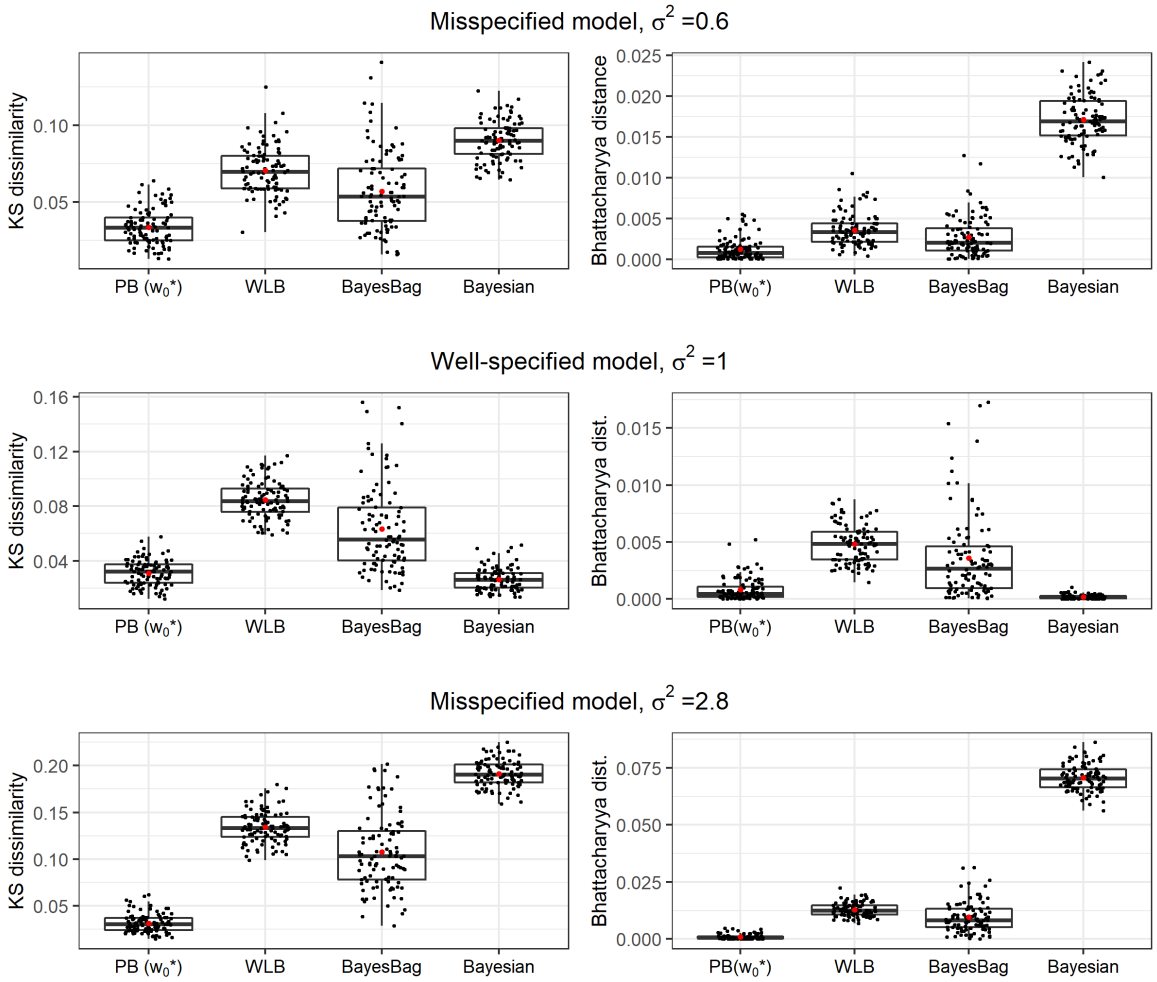


Figure 3.7: Kolmogorov–Smirnov dissimilarity (left panels) and Bhattacharyya distance (right panels) between samples from the Bayesian posterior for the correct model (3.33) and samples obtained using the following four methods: (from left to right) Algorithm 5 with $w_0 = w_0^*$, denoted by PB(w_0^*), Weighted Likelihood Bootstrap (WLB), BayesBag and standard Bayesian posterior for model (3.14). Three scenarios are considered, $\sigma^2 = 0.6$ (top panels), $\sigma^2 = 1$ (middle panels), and $\sigma^2 = 2.8$ (bottom panels). Each experiment was repeated on 100 simulated datasets for $N = 2000$.

D .2 Toy example of a hierarchical model

We consider the following hierarchical model with parameters λ, $\theta_1, \theta_2, \theta_3$:

$$\begin{aligned} x_{k,1}, \ldots, x_{k,n_k} &\sim \text{Poisson}(\theta_k), \qquad k = 1, 2, 3, \\ \theta_1, \ldots, \theta_3 &\sim \text{Gamma}(\alpha, \lambda), \qquad k = 1, 2, 3 \\ \lambda &\sim \text{Gamma}(\alpha_0, \beta_0). \end{aligned} \tag{3.66}$$

We consider two scenarios. In the first one data $x_{k,1}, \ldots, x_{k,n_k}$ is generated from the Poisson(θ_k) distribution with $\theta_1 = 1$, $\theta_2 = 2$ and $\theta_3 = 4$, so we have a well-specified model.

In the second scenario the correct model is

$$\begin{aligned} x_{k,1}, \ldots, x_{k,n_k} &\sim \text{NB}(\theta_k, 0.7), \qquad k = 1, 2, 3, \\ \theta_1, \ldots, \theta_3 &\sim \text{Gamma}(\alpha, \lambda), \qquad k = 1, 2, 3 \\ \lambda &\sim \text{Gamma}(\alpha_0, \beta_0). \end{aligned} \tag{3.67}$$

where $\text{NB}(\theta_k, \omega)$ denotes the negative binomial distribution with mean θ_k and size ω. Recall that variance of the $\text{NB}(\theta, \omega)$ distribution is $\theta + \theta^2/\omega$, so the negative binomial is often used for modelling overdispersed Poisson data. In the second scenario we generate $x_{k,1}, \ldots, x_{k,n_k}$ from $\text{NB}(\theta_k, 0.7)$, for $\theta_1 = 1$, $\theta_2 = 2$ and $\theta_3 = 4$. In both scenarios we set $\alpha_0 = 9, \beta_0 = 3$ and $\alpha = 2$, and the number of data points in each class is $n_1 = n_2 = n_3 = 100$. In this misspecified scenario we compare results obtained with different versions of Posterior Bootstrap performed on model (3.66) with the Bayesian posterior for model (3.67).

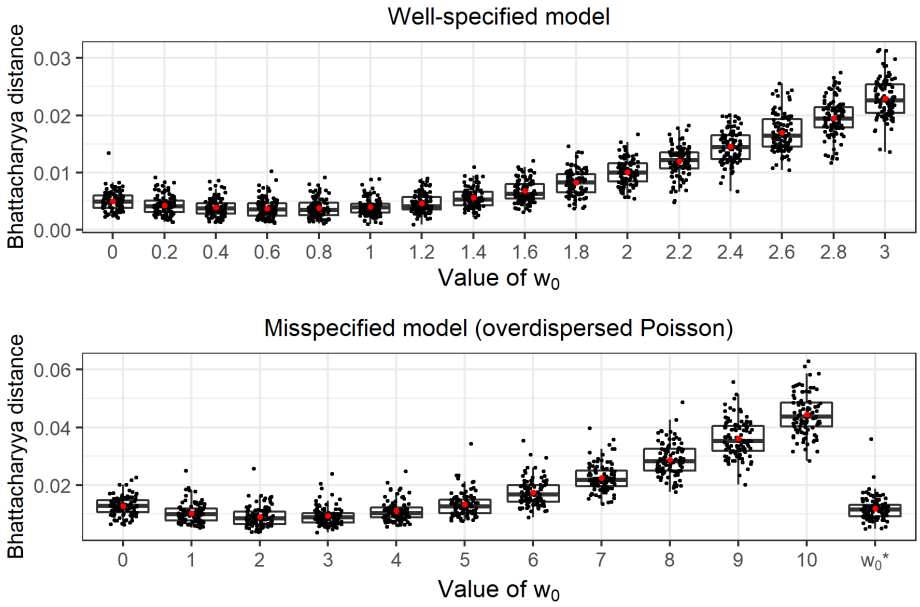


Figure 3.8: Bhattacharyya distance between samples from the Bayesian posterior for the correct model and samples obtained using Algorithm 6 for a range of values of w_0. We set the same value of w_0 for each class, that is, we use $w_{0,[1]} = w_{0,[2]} = w_{0,[3]} = w_0$, for w_0 displayed on the x-axis of the plot. We consider two scenarios: well-specified (top panel), where (3.66) is the correct model, and misspecified (bottom panel), for which (3.67) is the correct model. In the case of the misspecified model we also include the results for Algorithm 6 with $w_0^* = \left(w^*_{0,[1]}, w^*_{0,[2]}, w^*_{0,[3]}\right)$. Each experiment was repeated on 100 simulated datasets for $N = 2000$.

We tested Algorithm 6 on this example for a range of values of w_0 and the results of these experiments are presented in Figure 3.8. As expected, in the well-specified case it is

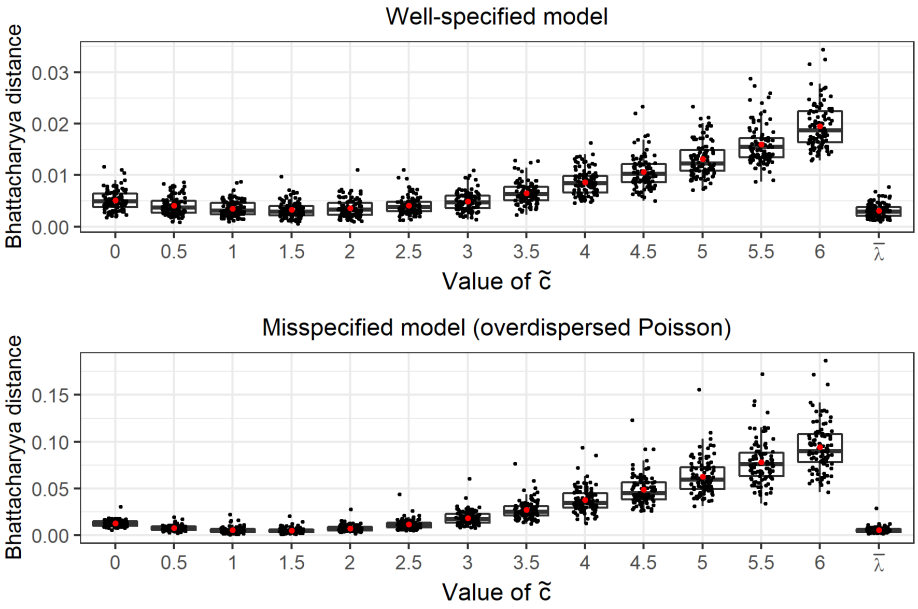


Figure 3.9: Bhattacharyya distance between samples from the Bayesian posterior for the correct model and samples obtained using Posterior Bootstrap with pseudo-samples for a range of values of $\tilde{c}$. We use the following notation: parameter c in Algorithm 7 is set to $c = (c_1, c_2, c_3) = \tilde{c}\left(w^*_{0,[1]}, w^*_{0,[2]}, w^*_{0,[3]}\right)$. We also present results for $\tilde{c} = \bar{\lambda}$, which is equivalent to setting c according to (3.31). We consider two scenarios: well-specified (top panel), where (3.66) is the correct model, and misspecified (bottom panel), for which (3.67) is the correct model. Each experiment was repeated on 100 simulated datasets for $N = 2000$.

optimal to set w_0 to 1 for each class, whereas in the overdispersed Poisson case setting w_0 according to (3.20) for each class yields results that are closest to the posterior under the correct model.

For a fixed value of λ model (3.66) is conjugate with the effective sample size of the prior equal to λ. This motivates testing Posterior Bootstrap with pseudo-samples within Algorithm 6, as mentioned in Remark 3.18. We expect to obtain best results setting parameter c according to (3.31) as

$$c = (c_1, c_2, c_3) = \bar{\lambda}\left(w^*_{0,[1]}, w^*_{0,[2]}, w^*_{0,[3]}\right), \tag{3.68}$$

which is confirmed by the results of our experiments shown in Figure 3.9. The pseudo-samples approach appears to have a slight advantage over the penalization-based approach on this example, as shown in Figure 3.10. In particular, both methods perform well when it comes to inference on λ, as shown in Figure 3.11. Both Posterior Bootstrap approaches give significantly better results than standard Bayesian inference in the misspecified case, as shown in the right panel of Figure 3.10.

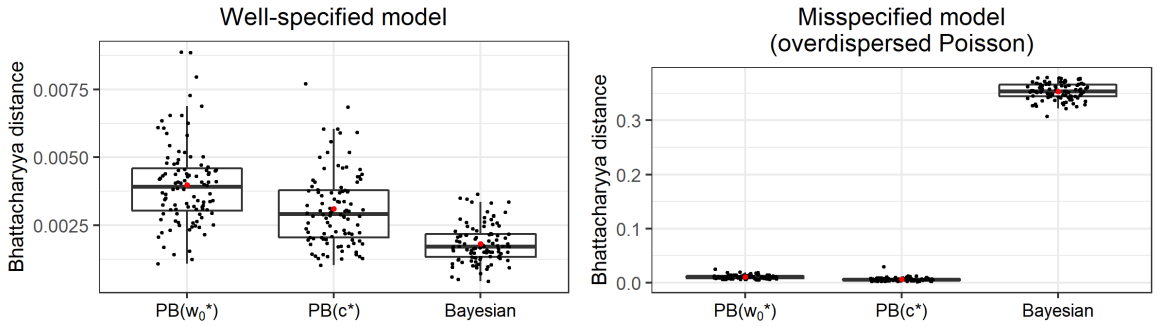


Figure 3.10: Bhattacharyya distance between samples from the Bayesian posterior for the correct model and samples obtained using four different methods: (from left to right): Algorithm 6 with w_0 set according to (3.20) for each class, Posterior Bootstrap with pseudo-samples with c set according to (3.31), and standard Bayesian posterior for model (3.66). We consider two scenarios: well-specified (left panel), where (3.66) is the correct model, and misspecified (right panel), for which (3.67) is the correct model. Each experiment was repeated on 100 simulated datasets for $N = 2000$. Note the difference in the scales of the y-axes.

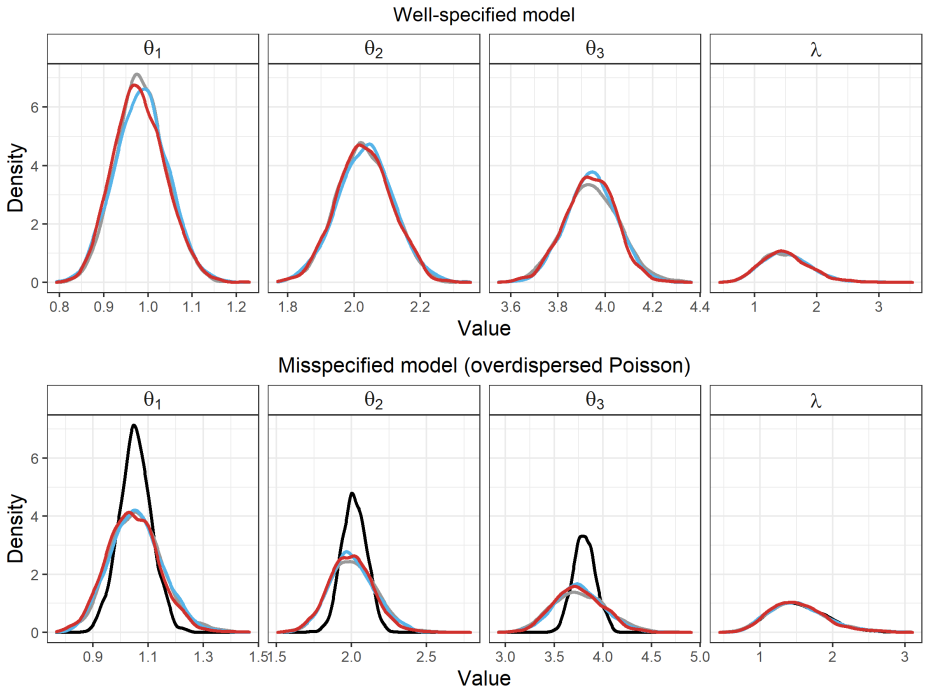


Figure 3.11: Kernel density estimates of posterior distribution for $\theta_1, \theta_2, \theta_3$ and λ based on a single synthetic dataset, under standard Bayesian inference for model (3.66) (black), Algorithm 6 with $w_0 = w_0^*$ (red), Posterior Bootstrap with pseudo-samples with c set according to (3.31) (light blue). We consider two scenarios: well-specified (top panel) and misspecified (bottom panel). For the misspecified case we include also the kernel density estimate for standard Bayesian inference under the correct model (3.67) (grey).

D .3 Additional information on the Dirichlet allocation model

We used Algorithm 11 to obtain results described in Section 5 .3. By $g(\theta|\lambda)$ we denote here the density of the Dirichlet distribution with parameter λ. We let $f(x_{k,i}|\theta)$ denote the probability mass function of a Multinomial$(\theta, 1)$ distribution, where $x_{k,i} = p$ for $p \in \{1, \ldots, 6\}$ denotes that the reason of death for person i in group k was overdosing drug of type p, for $k = 1, \ldots, K$ and $i = 1, \ldots, n_k$.

Algorithm 11 Posterior Bootstrap for the Dirichlet allocation model (3.37)

for $k = 1, \ldots, K$ **do**
  Set $\hat{\theta}_k = (m_{k1}/n_k, \ldots, m_{k6}/n_k)$.
end for
for j = 1, ..., N **do** ▷ (in parallel)
  Draw $z_1^{(j)}, \ldots, z_K^{(j)} \sim \text{Exp}(1)$.
  Set $\bar{\lambda}^{(j)} = \arg\max_\lambda \left\{ \sum_{k=1}^K z_k^{(j)} \log g(\hat{\theta}_k|\lambda) + w_g^T \log p\,(\lambda) \right\}$.
  Set $c = \bar{\lambda}_1^{(j)} + \cdots + \bar{\lambda}_6^{(j)}$.
  for $k = 1, \ldots, K$ **do**
    Draw $\bar{\theta}_{k,1}^{(j)}, \ldots, \bar{\theta}_{k,T}^{(j)} \sim \text{Dirichlet}(\bar{\lambda})$.
    Generate prior pseudo-samples $x_{k,n_k+i}^{(j)} \sim \text{Multinomial}\left(\bar{\theta}_{k,i}^{(j)}, 1\right)$ for $i = 1, \ldots, T$.
    Draw $w_1^{(j)}, \ldots, w_{n_k}^{(j)} \sim \text{Exp}(1)$ and $w_{n_k+1}^{(j)}, \ldots, w_{n_k+T}^{(j)} \sim \text{Exp}(T/c)$.
    Set $\tilde{\theta}_k^{(j)} = \arg\max_{\theta_k} \left\{ \sum_{i=1}^{n_k} w_i^{(j)} \log f(x_{k,i}|\theta_k) + \sum_{i=n_k+1}^{n_k+T} w_i^{(j)} \log f\left(x_{k,i}^{(j)}|\theta_k\right) \right\}$.
  end for
  Draw $v_1^{(j)}, \ldots, v_K^{(j)} \sim \text{Exp}(1)$.
  Set $\tilde{\lambda}^{(j)} = \arg\max_\lambda \left\{ \sum_{k=1}^K v_k^{(j)} \log g(\tilde{\theta}_k^{(j)}|\lambda) + w_g^T \log p\,(\lambda) \right\}$.
end for
return Samples $\left\{ \left(\tilde{\lambda}^{(j)}, \tilde{\theta}_1^{(j)}, \ldots, \tilde{\theta}_K^{(j)}\right) \right\}_{j=1}^N$.

We additionally tested Algorithm 11 on a synthetic dataset generated from model (3.37)with $\lambda = (12, 12, 12, 10, 10, 10)$, $K = 300$ and $n_k = 1000$ for $k = 1, \ldots, K$. Figure 3.12 shows that Bayesian inference and Algorithm 11 return very similar results, which suggests the difference between the two methods in Figure 3.4 is caused by model misspecification. In the same figure we plotted kernel density estimates of $\bar{\lambda}$. We notice that despite the fact that asymptotically $\bar{\lambda}$ and $\tilde{\lambda}$ have the same distributions, for this example $\tilde{\lambda}$ performs much better. We also observe that, as expected, using more concentrated priors causes a shift towards 0 in posterior distributions.

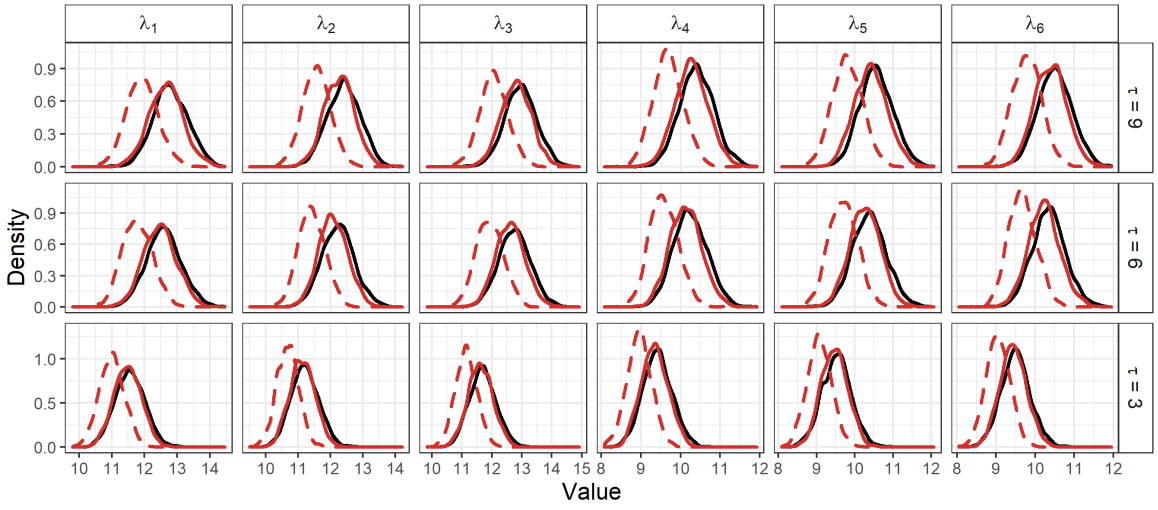


Figure 3.12: Kernel density estimates of posterior distribution on concentration parameter λ for a synthetic dataset generated from model (3.37) with $\lambda = (12, 12, 12, 10, 10, 10)$, $K = 300$ and $n_k = 1000$ for $k = 1, \ldots, K$. We present results for Bayesian inference (black), $\tilde{\lambda}$ drawn from Algorithm 11 (red solid line) and $\bar{\lambda}$ drawn from Algorithm 11 (red dashed line).

Chapter 4

Asymptotics of cut distributions and robust modular inference using Posterior Bootstrap

Abstract

Bayesian inference provides a framework to combine an arbitrary number of model components with shared parameters, allowing joint uncertainty estimation and the use of all available data sources. However, misspecification of any part of the model might propagate to all other parts and lead to unsatisfactory results. Cut distributions have been proposed as a remedy, where the information is prevented from flowing along certain directions. We consider cut distributions from an asymptotic perspective, find the equivalent of the Laplace approximation, and notice a lack of frequentist coverage for the associated credible regions. We propose algorithms based on the Posterior Bootstrap that deliver credible regions with the nominal frequentist asymptotic coverage. The algorithms involve numerical optimization programs that can be performed fully in parallel. The results and methods are illustrated in various settings, such as causal inference with propensity scores and epidemiological studies.

1 Introduction

1 .1 Cutting feedback in Bayesian models

Bayesian models are often built of different components or "modules", which can represent different sources of information about the studied phenomenon. Standard Bayesian inference relies on samples from the posterior distribution over the joint model combining all modules, and this approach is optimal when the model is well-specified (Zellner, 1988; Bissiri et al., 2016). In many applications, however, the analyst suspects that some or all modules may be misspecified. Such cases motivate departures from the joint model. In particular, it may be beneficial to cut the flow of information from the module that we put less trust in, to the ones that we trust more.

The issue described above often arises in pharmacology. Pharmacokinetics-pharmacodynamics (PK/PD) models combine data from two sources. They have a PK component, considering the body's effect on a drug, and a PD module, focusing on the effect the drug has on the body. Pharmacokinetics is believed to be better understood scientifically, which gives researchers more trust in the PK part of the model, while PD components are often suspected to be misspecified. Therefore, instead of relying on the full model, researchers doing PK/PD analysis often first fit the PK model, and then use the estimated parameters as input for the PD module, thus cutting the feedback from the PD component to the PK one (Zhang et al., 2003a,b; Lunn et al., 2009). We will refer to this approach as modular Bayesian inference, but it is also known in the literature as "modularized" approach (Liu et al., 2009) or "cutting feedback" (Plummer, 2015). In econometrics similar situations are commonly tackled with "two-stage" procedures (Murphy and Topel, 2002).

Apart from the PK/PD models, the modular approach is commonly used in econometrics, for example, when errors from one regression model are used as a covariate in another model. It has also been useful for epidemiological models (Maucort-Boulch et al., 2008), studies on health effects of air pollution (Blangiardo et al., 2011), analysis of computer models and meta-analysis (Liu et al., 2009). Another situation where two-stage estimation is routinely used is Bayesian inference based on a dataset where missing data were first imputed, e.g. with "multiple imputation" (Little, 1992; Rubin, 2004; Zhou and Reiter, 2010).

Furthermore, cutting feedback is an established practice in causal inference. To estimate the average treatment effect from observational data, the user needs to account for con-

founding variables in the dataset, which can be done using propensity scores. Recall that in randomized trials the treatment assignment happens before the outcome is known, whereas in observational studies propensity scores are used to compensate for the lack of randomized treatment assignment. This, apart from potential misspecification, is another reason for choosing the two-step approach. The first stage of the analysis aims to estimate propensity scores, given a set of covariates but not using the outcome, whereas the second stage focuses on estimating the treatment effect on the outcome, given the covariates and stratified by propensity scores. As shown by Zigler et al. (2013), performing the full Bayesian inference is possible, but leads to bias in the estimation of causal effects; see also Zigler (2016). A review of various applications of modular inference is presented in Jacob et al. (2017).

Throughout the paper we will write $y_1, \ldots, y_n \sim f(y|\theta)$ to denote that observations $y_1, \ldots, y_n$ are independent and identically distributed according to a distribution with density $f(y|\theta)$. Formally, we will study the following model:

$$\begin{aligned} x_{1,1}, \ldots, x_{1,n_1} &\sim f_1(x_1|\theta_1), \qquad \theta_1 \sim \pi_1(\theta_1), \quad \theta_1 \in \Theta_1 \subseteq \mathbb{R}^{d_1}, \\ x_{2,1}, \ldots, x_{2,n_2} &\sim f_2(x_2|\theta_1, \theta_2), \quad \theta_2 \sim \pi_2(\theta_2), \quad \theta_2 \in \Theta_2 \subseteq \mathbb{R}^{d_2}. \end{aligned} \tag{4.1}$$

In the first module the likelihood f_1 depends only on θ_1, whereas the second likelihood f_2 depends both on θ_1 and θ_2. Distributions π_i denote the priors on θ_i, for $i = 1, 2$. The analyst may alternatively consider the prior distribution in the second module as dependent on θ_1, that is, to define it as $\pi_2(\theta_2|\theta_1)$; the findings of the present article are easily adapted to that more general setup. By $x_{i,1:n_i}$, $i = 1, 2$, we denote the vector of observations $x_{i,1}, \ldots, x_{i,n_i}$.

The full posterior distribution associated with model (4.1) is given by

$$\pi(\theta_1, \theta_2|x_{1,1:n_1}, x_{2,1:n_2}) \propto \pi_1(\theta_1) \prod_{j=1}^{n_1} f_1(x_{1,j}|\theta_1)\pi_2(\theta_2) \prod_{j=1}^{n_2} f_2(x_{2,j}|\theta_1, \theta_2). \tag{4.2}$$

The marginal distribution of (4.2) with respect to θ_1 is then

$$\pi(\theta_1|x_{1,1:n_1}, x_{2,1:n_2}) \propto \pi(\theta_1|x_{1,1:n_1})\pi(x_{2,1:n_2}|\theta_1),$$

where

$$\pi(x_{2,1:n_2}|\theta_1) = \int \prod_{j=1}^{n_2} f_2(x_{2,j}|\theta_1, \theta_2)\pi_2(\theta_2)\mathrm{d}\theta_2$$

is called the feedback term of $x_{2,1:n_2}$ on θ_1. The modular approach removes that feedback,

by defining the "cut" posterior as

$$\pi_{\text{cut}}(\theta_1, \theta_2) \propto \frac{\pi(\theta_1, \theta_2 | x_{1,1:n_1}, x_{2,1:n_2})}{\pi(x_{2,1:n_2} | \theta_1)}. \tag{4.3}$$

Under the cut posterior, the inference on θ_1 is based only on the first module and the first marginal is

$$\pi_{1,\text{cut}}(\theta_1) \propto \pi(\theta_1 | x_{1,1:n_1}). \tag{4.4}$$

As shown in Lemma 1 of (Yu et al., 2021), the cut posterior $\pi_{\text{cut}}(\theta_1, \theta_2)$ minimizes the Kullback–Leibler divergence between the families of densities $q(\theta_1, \theta_2)$ having $\pi_{1,\text{cut}}$ as their marginal distribution with respect to θ_1, and the full posterior $\pi(\theta_1, \theta_2)$.

Both Jacob et al. (2017) and Yu et al. (2021) propose data-driven methods to assess whether to cut feedback or not. Carmona and Nicholls (2020) extend the modular methodology to "semi-modular" inference (SMI), where the user can regulate the amount of feedback passed on from the second module to the first one.

1 .2 Computational challenges of Bayesian modular inference

Drawing samples from the cut posterior is challenging due to the intractable constant $\pi(x_{2,1:n_2} | \theta_1)$ in the denominator of (4.3). As a result there is no known way of designing a Markov chain directly targeting the cut posterior distribution, and Plummer (2015) showed that the cut function implemented in the OpenBUGS software (Spiegelhalter et al., 2007) does not converge to the cut posterior in (4.3). Typically, sampling for Bayesian modular inference involves two stages, where the user first runs an MCMC algorithm targeting (4.4). Then, for each θ_1 from the (possibly thinned) chain, the user runs a chain targeting

$$\pi_2(\theta_2 | \theta_1, x_{2,1:n_2}) \propto \pi_2(\theta_2) \prod_{j=1}^{n_2} f_2(x_{2,j} | \theta_1, \theta_2).$$

This approach, in its simplest form, may be costly and rather inconvenient as it involves many MCMC runs at the second stage, each targeting a different distribution depending on θ_1, and thus requires tedious diagnostics of convergence. In particular, sampling from the cut posterior poses a challenge when it is difficult to draw samples from $\pi_2(\theta_2 | \theta_1, x_{2,1:n_2})$, or when the number of MCMC iterations needed to reach stationarity heavily depends on the value of θ_1. Finding more convenient and more accurate computational methods for

modular Bayesian inference is an active area of research. The unbiased MCMC approach developed in Jacob et al. (2020) uses the lack-of-bias property to remove the need to monitor convergence of each chain in the second stage. The user obtains unbiased estimators of functionals with respect to π_{cut}. An alternative approach was proposed by Liu and Goudie (2020), who designed an adaptive MCMC method that asymptotically approximates the cut posterior (4.3). Variational inference for modular inference is developed in Yu et al. (2021).

1 .3 Related work

The topic of Bayesian inference under model misspecification has attracted a lot of attention in recent years. Consider a generic Bayesian model,

$$x_1, \ldots, x_n \sim f(x|\theta), \quad \theta \sim \pi(\theta), \quad \theta \in \Theta \subseteq \mathbb{R}^d. \tag{4.5}$$

We assume that observations $x_1, \ldots, x_n$ are generated from a distribution with density p^*, which does not necessarily correspond to $f(x|\theta)$ for any $\theta \in \Theta$. We therefore allow for misspecification of model (4.5). We define the maximum likelihood estimator $\hat{\theta} = \arg\max_{\theta\in\Theta} \sum_{i=1}^{n} \log f(x_i|\theta)$ and the pseudo-true parameter $\theta^* = \arg\max_{\theta\in\Theta} E_{p^*} \log f(X|\theta)$. We also define information matrices I^* and J^* as

$$I^* = E_{p^*}\left\{\frac{\partial \log f(X|\theta^*)}{\partial\theta}\left(\frac{\partial \log f(X|\theta^*)}{\partial\theta}\right)^T\right\}, \quad J^* = -E_{p^*}\left\{\frac{\partial^2 \log f(X|\theta^*)}{\partial\theta\partial\theta^T}\right\}.$$

The Bernstein–von Mises result for misspecified models (Kleijn and Van der Vaart, 2012) states that the limiting distribution of $\sqrt{n}(\theta - \hat{\theta})$, where θ is drawn from the posterior, is $N\{0, J^{*-1}\}$. For comparison, standard asymptotic theory states that $\sqrt{n}(\hat{\theta} - \theta^*)$ converges to $N\{0, J^{*-1}I^*J^{*-1}\}$. Therefore, Bayesian inference in the misspecified scenario yields misleading uncertainty estimations in the sense that credible sets are not valid confidence sets for the pseudo-true parameter. Robust approaches, such as power posteriors (Grünwald, 2012; Bissiri et al., 2016; Grünwald and van Ommen, 2017) or BayesBag (Huggins and Miller, 2019, 2020) are aimed at correcting the uncertainty around the mean. Another approach to addressing misspecification is based on Weighted Likelihood Bootstrap (Newton, 1991; Newton and Raftery, 1994) and its extensions (Lyddon et al., 2018, 2019; Fong et al., 2019; Newton et al., 2020; Nie and Ročková, 2020; Pompe, 2021).

The methodology that we propose in Section 3 builds on Posterior Bootstrap via prior

penalization (Newton et al., 2020; Pompe, 2021), which we review briefly below. Drawing a sample $\tilde{\theta}$ from the Posterior Bootstrap for model (4.5) relies on computing

$$\tilde{\theta} = \arg\max_{\theta \in \Theta} \sum_{i=1}^{n} w_i f(x_i|\theta) + w_0^T \log \pi(\theta), \quad w_1, \ldots, w_n \sim \text{Exp}(1). \tag{4.6}$$

If the prior $\pi(\theta)$ factorizes as $\prod_{i=1}^{d} \pi_i(\theta_i)$, hyperparameter w_0 is treated as a vector and $w_0^T \log \pi(\theta) = \sum_{i=1}^{d} w_{0,i} \log \pi_i(\theta_i)$ for $w_0 = (w_{0,1}, \ldots, w_{0,d})$. Otherwise w_0 is a non-negative real number.

Note that this is slightly different than what is called the Posterior Bootstrap in Lyddon et al. (2018); Fong et al. (2019). In Section 3 .3 we will also consider the version of Posterior Bootstrap proposed by Fong et al. (2019).

The appeal of methods based on Weighted Likelihood Bootstrap discussed above is that the asymptotic distribution of $\sqrt{n}(\tilde{\theta} - \hat{\theta})$ is $N\left\{0, J^{*-1} I^* J^{*-1}\right\}$, so credible sets are asymptotically valid confidence sets even under misspecification (Fushiki, 2010; Lyddon et al., 2019). Computationally the samples can be obtained by solving optimization programs in parallel, using readily-available optimization methods for a wide range of models.

We emphasize that the methods mentioned above correct for misspecification in a different way than cut posteriors. For example the Posterior Bootstrap and BayesBag concentrate asymptotically on the same mean as the standard posterior, but with a possibly different covariance matrix. The modular approach, on the other hand, can shift the location of the posterior distribution. Moreover, while the aforementioned methods take a semi-automatic approach to correcting for misspecification, modularization directly takes advantage of the user's understanding of the working model and of its potential flaws. As we demonstrate later in this paper, in certain settings combining Posterior Bootstrap with cutting feedback provides further benefits over using either of these approaches.

1 .4 Contribution

The remaining part of this paper is organized as follows. In Section 2 we find the equivalent of the Laplace approximation for the cut posterior, and compare it to the asymptotic distribution of two-stage point estimators. We notice that the credible sets of the cut posterior may not asymptotically be the confidence sets in the frequentist sense. The difference between these two asymptotic results may stem from two sources. Firstly, like in one-stage inference,

it may result from model misspecification. Secondly, it may come from both modules relying on the same dataset, as cut posteriors fail to account for the correlation between the two likelihoods. In Section 3 we propose an approach to sampling in modular inference, called Posterior Bootstrap for Modular Inference (PBMI), which extends the ideas from Weighted Likelihood Bootstrap and Posterior Bootstrap. We prove that it helps overcome the above issue by ensuring that asymptotically the credible sets have the nominal coverage. Moreover, we discuss the asymptotic prediction properties of PBMI and those of cut posteriors. The proposed algorithms involve computations that can easily be distributed over parallel processors. In Section 4 we illustrate our methodology with toy and real data examples, including an epidemiological model (Maucort-Boulch et al., 2008; Plummer, 2015) and a causal inference study measuring the effect of the labour training on the post-intervention income (LaLonde, 1986; Dehejia and Wahba, 1999). We conclude with a discussion in Section 5 , where we summarize advantages and limitations of our approach, and indicate potential future research directions.

2 Asymptotics of Bayesian versus point estimation in modular settings

We obtain and compare the asymptotic distributions in modular inference, as the number of observations goes to infinity, of maximum likelihood and Bayesian approaches. We work with model (4.1) and allow for possible misspecification. We assume that for $i = 1, 2$ observations $x_{i,1}, \ldots, x_{i,n_i}$ are independent and identically distributed with density p_i^*. We define the maximum likelihood estimators $\hat{\theta}_1$ and $\hat{\theta}_2$ as

$$\hat{\theta}_1 = \arg\max_{\theta_1} \sum_{j=1}^{n_1} \log f_1(x_{1,j}|\theta_1), \quad \hat{\theta}_2 = \arg\max_{\theta_2} \sum_{j=1}^{n_2} \log f_2(x_{2,j}|\hat{\theta}_1, \theta_2).$$

Similarly, we define pseudo-true parameters as follows

$$\theta_1^* = \arg\max_{\theta_1} E_{p_1^*} \log f_1(X_1|\theta_1), \quad \theta_2^* = \arg\max_{\theta_2} E_{p_2^*} \log f(X_2|\theta_1^*, \theta_2).$$

Note that we define these maximizers separately for each module, so $(\hat{\theta}_1, \hat{\theta}_2)$ is in general different from the maximum likelihood estimator for the full model, given by

$$(\theta_{1,\mathrm{MLE}}, \theta_{2,\mathrm{MLE}}) = \arg\max_{\theta_1,\theta_2} \left\{ \sum_{j=1}^{n_1} \log f_1(x_{1,j}|\theta_1) + \sum_{j=1}^{n_2} \log f_2(x_{2,j}|\theta_1, \theta_2) \right\}.$$

While the standard posterior for the full model concentrates asymptotically around $(\theta_{1,\mathrm{MLE}}, \theta_{2,\mathrm{MLE}})$, the cut posterior concentrates around $(\hat{\theta}_1, \hat{\theta}_2)$. We will prove the latter result formally later in this section. This justifies a statement we made in Section 1 .3: cutting feedback has a potential of shifting the asymptotic mean, which is not possible in more automatic robust inference approaches, such as power posteriors or Weighted Likelihood Bootstrap.

Let us also define the following matrices:

$$\begin{aligned}
I_1^* =& E_{p_1^*} \left\{ \frac{\partial \log f_1(X_1|\theta_1^*)}{\partial \theta_1} \left(\frac{\partial \log f_1(X_1|\theta_1^*)}{\partial \theta_1} \right)^T \right\} \\
J_1^* =& - E_{p_1^*} \left\{ \frac{\partial^2 \log f_1(X_1|\theta_1^*)}{\partial \theta_1 \partial \theta_1^T} \right\} \\
I_2^* =& E_{p_2^*} \left\{ \frac{\partial \log f_2(X_2|\theta_1^*, \theta_2^*)}{\partial \theta_2} \left(\frac{\partial \log f_2(X_2|\theta_1^*, \theta_2^*)}{\partial \theta_2} \right)^T \right\} \\
J_2^* =& - E_{p_2^*} \left\{ \frac{\partial^2 \log f_2(X_2|\theta_1^*, \theta_2^*)}{\partial \theta_2 \partial \theta_2^T} \right\} \\
R_I^* =& E_{p^*} \left\{ \frac{\partial \log f_1(X_1|\theta_1^*)}{\partial \theta_1} \left(\frac{\partial \log f_2(X_2|\theta_1^*, \theta_2^*)}{\partial \theta_2} \right)^T \right\} \\
R_J^* =& - E_{p_2^*} \left\{ \frac{\partial^2 \log f_2(X_2|\theta_1^*, \theta_2^*)}{\partial \theta_1 \partial \theta_2^T} \right\}.
\end{aligned}$$

The theoretical results we state in this section, as well as asymptotic results we present in Section 3 for Posterior Bootstrap for Modular Inference, hold under standard regularity conditions, which we state explicitly in Appendix A .1. They are analogous to the regularity conditions considered typically for asymptotic normality of maximum likelihood estimators, as well as to conditions used for example by Lyddon et al. (2019); Pompe (2021) to obtain asymptotic results about Weighted Likelihood Bootstrap and its extensions.

We consider two distinct scenarios.

Scenario 1. We consider model (4.1), assuming that the datasets $x_{1,1}, \ldots, x_{1,n_1}$ and $x_{2,1} \ldots, x_{2,n_2}$ are generated independently, that is

$$p^*(x_{1,1:n_1}, x_{2,1:n_2}) = p_1^*(x_{1,1:n_1}) p_2^*(x_{2,1:n_2}).$$

Scenario 2. We consider model (4.1), assuming that $n_1 = n_2 = n$. We assume that observations $(x_{1,j}, x_{2,j})$ for $j = 1, \ldots, n$ are generated i.i.d. with density p^* and

$$p^*(x_{1,j}, x_{2,j}) \neq p_1^*(x_{1,j}) p_2^*(x_{2,j}), \quad j = 1, \ldots, n. \tag{4.7}$$

In Sections 2 .1 and 2 .2 we discuss in more detail Scenario 1 and Scenario 2 respectively, and present asymptotic results for the associated point estimators. In Section 2 .3 we present the equivalent of the Laplace approximation for the cut posterior (4.3) under any of the two scenarios.

2 .1 Scenario 1

Scenario 1 refers to situations where the data come from independent sources, such as the "biased data" example, introduced by Liu et al. (2009) and analysed later by Jacob et al. (2017); Carmona and Nicholls (2020). Suppose we are given a good quality dataset $x_{1:1:n_1} \sim N(\theta_1, 1)$ and our task is to estimate θ_1. We also have access to another dataset, of worse quality, as we suspect the data is biased. The second module is given by $x_{2:1:n_2} \sim N(\theta_1 + \theta_2, 1)$, where θ_1 represents the same location as in the first module, and θ_2 represents the "bias". This example was used by Liu et al. (2009) to show that in certain cases it is beneficial to use a modular approach rather than a joint model. We analyse this example in more detail in Section 4 .2.

Proposition 4.1 gives the asymptotic distribution of $(\hat{\theta}_1, \hat{\theta}_2)$ under Scenario 1. The working assumptions are listed in Appendix A .1.

Proposition 4.1. *Consider Scenario 1 and assume that the number of observations in both modules goes to infinity at rates such that $n_1/n_2 \to \alpha$ for some $\alpha > 0$. Then $\sqrt{n_2}(\hat{\theta}_1 - \theta_1^*, \hat{\theta}_2 - \theta_2^*)$ converges in distribution to $N(0, \Sigma)$, where*

$$\Sigma = \begin{pmatrix} 1/\alpha S_1 & -1/\alpha S_1 R_J^* J_2^{*-1} \\ -1/\alpha J_2^{*-1} R_J^{*T} S_1 & J_2^{*-1}\left(I_2^* + R_J^{*T} 1/\alpha S_1 R_J^*\right) J_2^{*-1} \end{pmatrix} \tag{4.8}$$

and $S_1 = J_1^{*-1} I_1^* J_1^{*-1}$.

The proof follows directly by repeating the steps from Section 4 of Murphy and Topel (2002). The only difference between the two settings is that Murphy and Topel (2002) did

not take into account potential misspecification of the modules, thus having $I_1^* = J_1^*$ and $I_2^* = J_2^*$.

2 .2 Scenario 2

This scenario represents situations where datasets feeding both modules contain measurements of the same individuals. Scenario 2 is common for example in causal inference with propensity scores. We next introduce formally some basic definitions related to propensity scores.

Let Z denote the binary treatment variable, such that $Z_i = 1$ if treatment has been given to individual i, and $Z_i = 0$ otherwise. By Y_i we denote the outcome variable for individual i, and we let $X_i = (X_{i1}, \dots X_{ip})$ be the vector of covariates. The object of causal inference is the estimation of the influence of Z onto Y. To estimate the treatment effect from observational studies, the propensity score methodology can be used to compensate for the lack of randomization of the treatment assignment. The propensity score is defined as $P(Z_i = 1|X_i)$ (Rosenbaum and Rubin, 1983). Note that some approaches extend the propensity score methodology to continuous treatment variables (Imai and Van Dyk, 2004; Hirano and Imbens, 2004).

The first module of the cut model estimates propensity scores via regressing Z_i on covariates X_i. The second module entails regressing the outcome Y_i on propensity scores and Z_i. That is $Z_{1:n}$ plays the role of $x_{1,1:n}$ in model (4.1), and $Y_{1:n}$ plays the role of $x_{2,1:n}$. A standard assumption made in causal inference with propensity scores is unconfoundedness, which stipulates that

$$Y(z) \perp\!\!\!\perp Z|X, \quad z = 0, 1, \tag{4.9}$$

where $Y(z)$ is the potential outcome that would have been observed under treatment z (see, for example, Antonelli et al. (2020)). Condition (4.9) means that X contains all common causes of treatment Z and outcome Y, and that there are no unmeasured confounders. Note that under this condition we would in fact be in Scenario 1. However, in practice the researchers can rarely be sure that all confounders have been included, which is why settings with missing confounders are of great interest in the literature (McCandless et al., 2012; Bonvini and Kennedy, 2021). Having unmeasured confounders implies that we are under Scenario 2. We illustrate our PBMI methodology with an example from causal inference

with propensity scores in Section 4 .3.

The result presented below is an analogue of Proposition 4.1 under Scenario 2.

Proposition 4.2. *Consider Scenario 2. We define* $S_2 = R_I^{*T} J_1^{*-1} R_J^* + R_J^{*T} J_1^{*-1} R_I^*$ *and* $S_3 = J_1^{*-1} R_I^* J_2^{*-1} - S_1 R_J^* J_2^{*-1}$. *Then as* n *in model* (4.7) *goes to infinity,* $\sqrt{n}(\hat{\theta}_1 - \theta_1^*, \hat{\theta}_2 - \theta_2^*)$ *converges in distribution to* $N(0, \tilde{\Sigma})$, *where*

$$\tilde{\Sigma} = \begin{pmatrix} S_1 & S_3 \\ S_3^T & J_2^{*-1}\left(I_2^* + R_J^{*T} S_1 R_J^* - S_2\right) J_2^{*-1} \end{pmatrix} \tag{4.10}$$

for S_1 *defined in Proposition 4.1.*

The proof again follows directly by repeating the arguments presented in Section 4 of (Murphy and Topel, 2002).

2 .3 Laplace approximation for the cut posterior

Proposition 4.3 presents the asymptotic result for Bayesian modular inference, under any of the two scenarios defined above. The proof can be found in Appendix A .2.

Proposition 4.3. *Consider model* (4.1) *and suppose that* n_1 *and* n_2 *go to infinity at rates such that* $n_1/n_2 \to \alpha$ *for some* $\alpha > 0$. *Then under regularity conditions for* (θ_1, θ_2) *drawn from the cut posterior* (4.3), *the limiting distribution of* $\sqrt{n_2}(\theta_1 - \hat{\theta}_1, \theta_2 - \hat{\theta}_2)$ *is* $N(0, \Sigma_B)$, *where*

$$\Sigma_B = \begin{pmatrix} 1/\alpha J_1^{*-1} & -1/\alpha J_1^{*-1} R_J^* J_2^{*-1} \\ -1/\alpha J_2^{*-1} R_J^{*T} J_1^{*-1} & J_2^{*-1} + 1/\alpha J_2^{*-1} R_J^{*T} J_1^{*-1} R_J^* J_2^{*-1} \end{pmatrix}. \tag{4.11}$$

The difference between asymptotic covariance matrices appearing in Propositions 4.1 and 4.2 and Proposition 4.3 results from two sources. First, as in standard Bayesian inference, there is a difference between the asymptotic posterior variance and the "sandwich" variance associated with valid confidence intervals. The second difference appears under Scenario 2, since the cut posterior does not take into account the correlation between the likelihoods of the two modules, and as a result the matrix R_I^* does not appear in the limiting covariance matrix. This means that cut posteriors may not provide the correct frequentist coverage even when each of the modules is well-specified. Asymptotically the credible regions of the cut posterior may be smaller or larger than the corresponding frequentist confidence regions.

Even though under standard Bayesian inference with well-specified models, credible sets have the correct frequentist coverage, this property does not always hold for generalized posteriors (Miller, 2021). To give another example of an inference scheme without this property, let us recall the definition of composite likelihoods (Lindsay, 1988; Varin et al., 2011). Suppose $s_j(x)$ and $t_j(x)$ for $j = 1, \ldots, k$ denote functions of data x. When $X \sim f(|\theta)$, let $p_\theta(s_j|t_j)$ denote the density of the conditional distribution of $s_j(X)$ given $t_j(X)$. Then the composite likelihood is given by

$$L^{\mathrm{CL}}(\theta) = \prod_{j=1}^{k} p_\theta(s_j|t_j).$$

The composite likelihood posterior is proportional to $L^{\mathrm{CL}}(\theta)\pi(\theta)$, where π denotes the prior on θ. Theoretical properties of posteriors based on composite likelihoods were studied, among others, by Pauli et al. (2011); Ribatet et al. (2012); Miller (2021). In generality, they do not have the correct frequentist coverage even in well-specified cases. For details, we refer the reader to Section 5.2 of Miller (2021).

We mainly focus in this paper on the situations where $n_1/n_2 \to \alpha > 0$. However, other asymptotic regimes, such as $n_2 = o(n_1)$, are also of interest. This may be relevant, for example, when the second module is a regression model with covariates estimated from the first module. An example of such a setting is the epidemiological study of Maucort-Boulch et al. (2008), which we consider in Section 4 .4.

3 Methodology and theoretical results

3 .1 Algorithms and asymptotic results for modular inference

To address the issues outlined above, we design a methodology based on Posterior Bootstrap (Pompe, 2021), which we further call Posterior Bootstrap for Modular Inference (PBMI). We will consider separately the two scenarios outlined in Section 2 .

We note that the asymptotic distributions elicited in the previous section (Propositions 4.1 and 4.2) could be approximated directly by multivariate Normal distributions, plugging $\hat{\theta}_1$ and $\hat{\theta}_2$ as the mean and, for the variance, using sample averages to approximate expectations with respect to p_1^* and p_2^* in the definitions of the various matrices appearing in Σ and $\tilde{\Sigma}$. Our methodology presented below is asymptotically first-order equivalent, but

comes with additional benefits. Firstly, it allows the user to incorporate prior information. Secondly, it does not require the direct approximation of covariance matrices, which may be cumbersome for complex models.

When the datasets for both modules are generated independently, that is, we are under Scenario 1, we propose to use Algorithm 12. A single iteration of this algorithm relies on performing Posterior Bootstrap twice: for the first module, and then for the second module, given the sample $\tilde{\theta}_1$ drawn for the first module.

Algorithm 12 Posterior Bootstrap for Modular Inference (PBMI) for Scenario 1

1: **for** k = 1, ..., N **do** ▷ (in parallel)
2: Draw $(w_1^{(k)}, \ldots, w_{n_1}^{(k)}) \sim \text{Exp}(1)$.
3: Set $\tilde{\theta}_1^{(k)} = \arg\max_{\theta_1} \left\{ \sum_{j=1}^{n_1} w_j^{(k)} \log f_1(x_{1,j}|\theta_1) + w_0^T \log \pi_1(\theta_1) \right\}$.
4: Draw $(v_1^{(k)}, \ldots, v_{n_2}^{(k)}) \sim \text{Exp}(1)$.
5: Set $\tilde{\theta}_2^{(k)} = \arg\max_{\theta_2} \left\{ \sum_{j=1}^{n_2} v_j^{(k)} \log f_2(x_{2,j}|\tilde{\theta}_1^{(k)}, \theta_2) + v_0^T \log \pi_2(\theta_2) \right\}$.
6: **end for**
7: **return** Sample $\{(\tilde{\theta}_1^{(k)}, \tilde{\theta}_2^{(k)})\}_{k=1}^N$.

For Scenario 2 we propose Algorithm 13, which also requires performing Posterior Bootstrap twice at each iteration. The difference between the two procedures is that in Algorithm 13 at any given iteration we use the same set of weights $w_1, \ldots, w_n$ at both sampling stages. By p_{cut}^{PB} we denote the density of the output of Algorithm 12 or Algorithm 13, under Scenario 1 or Scenario 2, respectively.

Algorithm 13 Posterior Bootstrap for Modular Inference (PBMI) for Scenario 2

1: **for** k = 1, ..., N **do** ▷ (in parallel)
2: Draw $(w_1^{(k)}, \ldots, w_n^{(k)}) \sim \text{Exp}(1)$.
3: Set $\tilde{\theta}_1^{(k)} = \arg\max_{\theta_1} \left\{ \sum_{j=1}^{n} w_j^{(k)} \log f_1(x_{1,j}|\theta_1) + w_0^T \log \pi_1(\theta_1) \right\}$.
4: Set $\tilde{\theta}_2^{(k)} = \arg\max_{\theta_2} \left\{ \sum_{j=1}^{n} w_j^{(k)} \log f_2(x_{2,j}|\tilde{\theta}_1^{(k)}, \theta_2) + v_0^T \log \pi_2(\theta_2) \right\}$.
5: **end for**
6: **return** Sample $\left\{(\tilde{\theta}_1^{(k)}, \tilde{\theta}_2^{(k)})\right\}_{k=1}^N$.

Let us summarize the main properties of these algorithms. Firstly, they are presented for a model built of two modules, but can be easily extended to more modules, by adding more optimization steps. In both algorithms the "for loop" can be executed fully in parallel and the cost of each iteration scales linearly with the number of modules. Recall from Section 1 that Bayesian modular approaches can be considered more computationally demanding than standard Bayesian ones, due to the difficulty in intractable feedback terms in the definition of the cut distribution. However, the situation is different with the Posterior Bootstrap:

modular inference is as easy as standard inference with the Posterior Bootstrap, the only difference is the addition of an optimization program for each module. The following results establish asymptotic properties of Algorithm 12 under Scenario 1 and of Algorithm 13 under Scenario 2, respectively.

Theorem 4.4. *Consider Scenario 1 and suppose that n_1 and n_2 go to infinity at rates such that $n_1/n_2 \to \alpha$ for some $\alpha > 0$. Assume that $(\tilde{\theta}_1, \tilde{\theta}_2)$ is drawn from Algorithm 12 and that the regularity conditions hold. Then for any Borel set $A \subseteq \mathbb{R}^{d_1+d_2}$ we have*

$$P\left\{\sqrt{n_2}(\tilde{\theta}_1 - \hat{\theta}_1, \tilde{\theta}_2 - \hat{\theta}_2) \in A | x_{1,1:n_1}, x_{2,1:n_2}\right\} \to P(Z \in A),$$

where $Z \sim N(0, \Sigma)$ for Σ given by (4.8), with probability going to 1 with respect to $x_{1,1:n_1}, x_{2,1:n_2}$.

Theorem 4.5. *Consider Scenario 2. Assume that $(\tilde{\theta}_1, \tilde{\theta}_2)$ is drawn from Algorithm 13 and that the regularity conditions hold. Then as $n \to \infty$ for any Borel set $A \subseteq \mathbb{R}^{d_1+d_2}$ we have*

$$P\left\{\sqrt{n}(\tilde{\theta}_1 - \hat{\theta}_1, \tilde{\theta}_2 - \hat{\theta}_2) \in A | x_{1,1:n}, x_{2,1:n}\right\} \to P(Z \in A),$$

where $Z \sim N(0, \tilde{\Sigma})$ for $\tilde{\Sigma}$ given by (4.10), with probability going to 1 with respect to $x_{1,1:n}, x_{2,1:n}$.

The above theorems show that under both scenarios introduced in Section 2 our methodology provides credible sets that are (asymptotically) confidence sets. In the case of Scenario 2, the difference between results obtained using Algorithm 13 and Bayesian cut posteriors is that Algorithm 13 accounts for the dependence between the likelihoods by using the same weights at both stages of inference. The proofs are presented in Appendix A .3.

In some cases the user may not know whether their model corresponds to Scenario 1 or 2. For example, when using causal inference with propensity scores, researchers do not usually know whether all confounders have been observed or not. In Algorithm 13 using the same weights serves the purpose of recovering R_I^* in the asymptotic covariance matrix, which is clear upon inspection of the proof of Theorem 4.5. Therefore, when both modules are based on the same dataset (like in causal inference with propensity scores), and the user does not know whether condition

$$p^*(x_{1,j}) = p_1^*(x_{1,j}) p_2^*(x_{2,j}), \quad j = 1, \ldots, n, \tag{4.12}$$

is satisfied, then Algorithm 13 should be used. Under condition (4.12) we have $R_I^* = 0$, and hence Algorithms 12 and 13 asymptotically give first order-equivalent results. If however condition (4.12) does not hold, using Algorithm 12 the user risks falsely setting $R_I^* = 0$, as would be done under cut posteriors (see Section 2 .3). We illustrate it on a toy example in Section 4 .1.

In Appendix A .5 we show how the above methodology can be extended to semi-modular inference (Carmona and Nicholls, 2020), where instead of choosing between the cut approach and the full model, the user can regulate the impact on the second module on the inference.

3 .2 Prediction under Posterior Bootstrap for Modular Inference

The results presented above show the advantage of Posterior Bootstrap in modular inference regarding the coverage of confidence sets, but it may also be of interest to analyse its properties from the point of view of prediction

The work done by T. Fushiki (Fushiki, 2005, 2010) laid the foundations for comparing the predictive properties of various versions of bootstrapping methods with those of Bayesian inference, for a simple Bayesian model (4.5). In particular, these papers consider the following risk function (which we define using the notation introduced in Section 1 .3):

$$E_{x_{1:n}}\left[D_{KL}\left\{p^*||\hat{p}(\cdot|x_{1:n})\right\}\right] = \int p^*(x_{1:n}) \int p^*(x_{n+1}) \log \frac{p^*(x_{n+1})}{\hat{p}(x_{n+1}|x_{1:n})} \mathrm{d}x_{n+1} \mathrm{d}x_{1:n}, \tag{4.13}$$

where $\hat{p}(\cdot|x_{1:n})$ is the predictive distribution given data $x_{1:n}$ under a given method of inference.

Weighted Likelihood Bootstrap turns out to outperform asymptotically Bayesian inference in terms of prediction as long as $J^* \neq I^*$ for model (4.5), that is, asymptotically the risk in (4.13) is smaller under Weighted Likelihood Bootstrap (Fushiki, 2005, 2010). A careful inspection of the proof shows Posterior Bootstrap, used for model (4.5), shares this advantageous property of Weighted Likelihood Bootstrap. Thus, a simple corollary from this fact is that Posterior Bootstrap approach yields better prediction for the first module in model (4.1). In Proposition 4.6 below we obtain the value of the expected risk (4.13) for the second module, under the cut Bayesian approach, and under PBMI.

Let $\hat{p}_B$ and $\hat{p}_{PB}$ denote the predictive distribution on x_2 given $x_{1,1:n_1}, x_{2,1:n_2}$ under the modular Bayesian approach, and under Algorithm 12 or Algorithm 13, given Scenario 1

or Scenario 2, respectively. Recall that p_2^* denotes the true data-generating distribution of $x_{2,1:n_2}$. We define

$$a_B = \lim_{n_2 \to \infty} n_2 \int (\theta - \hat{\theta}) \pi_{\text{cut}}(\theta | x_{1,1:n_1}, x_{2,1:n_2}) \mathrm{d}\theta,$$
$$a_{PB} = \lim_{n_2 \to \infty} n_2 \int (\theta - \hat{\theta}) p_{\text{cut}}^{PB}(\theta | x_{1,1:n_1}, x_{2,1:n_2}) \mathrm{d}\theta,$$

where the above limits are taken in probability with respect to $x_{1,1:n_1}, x_{2,1:n_2}$. In particular a_B and a_{PB} are deterministic. We also introduce matrices

$$I_{f_2}^* = E_{p_2^*}\left\{ \frac{\partial \log f_2(x_2|\theta_1^*, \theta_2^*)}{\partial \theta} \left(\frac{\partial \log f_2(x_2|\theta_1^*, \theta_2^*)}{\partial \theta} \right)^T \right\}, \quad J_{f_2}^* = -E_{p_2^*}\left\{ \frac{\partial^2 \log f_2(x_2|\theta_1^*, \theta_2^*)}{\partial \theta \partial \theta^T} \right\}. \tag{4.14}$$

Proposition 4.6. *Let Σ_B be the asymptotic covariance matrix for Bayesian modular inference given by (4.11). Then under regularity conditions the risk associated with $\hat{p}_B$ is asymptotically expanded as*

$$\begin{aligned} E_{p^*}\left[D_{KL}\left\{ p_2^* || \hat{p}_B(\cdot | x_{1,1:n_1}, x_{2,1:n_2}) \right\} \right] &= E_{p^*}\left[D_{KL}\left\{ p_2^* || f_2(\cdot | \hat{\theta}_1, \hat{\theta}_2) \right\} \right] \\ - E_{p_2^*}\left\{ \frac{\partial \log f_2(x_{2,n_2+1} | \theta_1^*, \theta_2^*)}{\partial \theta} \right\} \frac{a_B}{n_2} &- \frac{1}{2n_2} tr\left\{ \left(I_{f_2}^* - J_{f_2}^* \right\} \Sigma_B \right) + o(n_2^{-1}). \end{aligned} \tag{4.15}$$

Suppose that the assumptions of Theorem 4.4 are satisfied and let Σ_{PB} be given by (4.8), or consider the assumptions of Theorem 4.5 and let Σ_{PB} be given by (4.10). In both cases the risk associated with $\hat{p}_{PB}$ is asymptotically expanded as

$$\begin{aligned} E_{p^*}\left[D_{KL}\left\{ p_2^* || \hat{p}_{PB}(\cdot | x_{1,1:n_1}, x_{2,1:n_2}) \right\} \right] &= E_{p^*}\left[D_{KL}\left\{ p_2^* || f_2(\cdot | \hat{\theta}_1, \hat{\theta}_2) \right\} \right] \\ - E_{p_2^*}\left\{ \frac{\partial \log f_2(x_{2,n_2+1} | \theta_1^*, \theta_2^*)}{\partial \theta} \right\} \frac{a_{PB}}{n_2} &- \frac{1}{2n_2} tr\left\{ \left(I_{f_2}^* - J_{f_2}^* \right\} \Sigma_{PB} \right) + o(n_2^{-1}). \end{aligned} \tag{4.16}$$

The proof of the above result largely relies on following the steps of Shimodaira (2000) and we report it in Appendix A .4. One may be wondering how the risk for the second module under PBMI (4.16) compares with that under Bayesian inference (4.15). It turns out any of the two approaches may be better, depending on the example at hand. In Appendix A .4 we show an example illustrating this phenomenon.

3 .3 Incorporating prior information

Even though the choice of hyperparameters w_0 and v_0 does not affect the first order asymptotic approximation expressed in Theorems 4.4 and 4.5, it plays a role in calibrating the impact of the priors. We propose to follow the strategy developed by Pompe (2021) for one-stage Posterior Bootstrap, guided by Edgeworth expansions. If the priors on θ_1 and θ_2 factorize, we propose setting w_0 and v_0 as follows:

$$w_0 = \operatorname{diag}\left(I_1^{*1/2} J_1^{*-1} I_1^{*1/2}\right), \quad v_0 = \operatorname{diag}\left(I_2^{*1/2} J_2^{*-1} I_2^{*1/2}\right). \tag{4.17}$$

Otherwise we set

$$w_0 = \operatorname{tr}\left(I_1^{*1/2} J_1^{*-1} I_1^{*1/2}\right), \quad v_0 = \operatorname{tr}\left(I_2^{*1/2} J_2^{*-1} I_2^{*1/2}\right).$$

Since matrices I_i^* and J_i^* are rarely known, in practice we use their empirical approximations. Note that the choice of w_0 and v_0 is important in the case of informative priors. When the prior is vague and meant to represent lack of prior information about the parameter, we can for simplicity set these values to 0 or 1.

As discussed by Pompe (2021), Posterior Bootstrap with prior penalization cannot be used with priors that are not bounded on the set, as the maximum in equation (4.6) would always be attained at the boundary of the set, regardless of the weights $w_1, \ldots, w_n$ and the data. An example of such a prior is the Dirichlet distribution $\operatorname{Dir}(\alpha_1, \ldots, \alpha_k)$, where any of the components α_j is smaller than 1. In these cases, an alternative way of incorporating prior information is via so-called pseudosamples generated from the prior predictive distribution (Fong et al., 2019). Meaningful and theoretically justified calibration of the the impact of the prior in Posterior Bootstrap with pseudosamples is possible for conjugate priors. At the same time, one should be cautious when using this method with non-conjugate priors, in particular weakly informative ones. In such cases Posterior Bootstrap with pseudosamples often yields unstable results that distort the inference; see Sections 4 and 6 of Pompe (2021) for a related discussion.

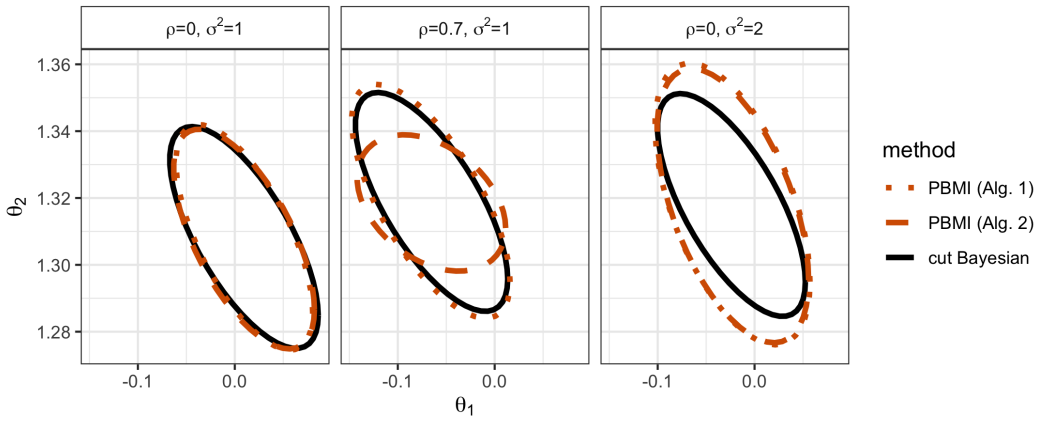


Figure 4.1: Density contours, corresponding to the 95% mass region, for samples obtained using PBMI (Algorithm 12 and Algorithm 13) and Bayesian modular inference for model (4.18). We considered three settings, depending on the choice of ρ and σ^2 in the data-generating mechanism. Each plot is based on 2000 samples.

4 Illustrations

4 .1 Toy example

We consider the following model:

$$\begin{aligned} z_i &\sim N(\theta_1, 1), \quad i = 1, \ldots, n \\ y_i &\sim N(\theta_1 + \theta_2 x_i, 1), \quad i = 1, \ldots, n. \end{aligned} \tag{4.18}$$

Independent $N(0, 10^2)$ priors are set on θ_1 and θ_2. We generate $x_1, \ldots, x_n \sim N(3, 1)$ with $n = 800$. The observations (z_i, y_i) are generated from the bivariate Normal distribution with mean $(0, 1 + x_i)$ and covariance matrix

$$\begin{pmatrix} 1 & \rho \\ \rho & \sigma^2 \end{pmatrix}.$$

We consider three different settings, depending on the values of ρ and σ^2. Note that setting $\rho = 0$ makes z_i and y_i independent, as in Scenario 1; setting $\sigma^2 \neq 1$ makes the second module misspecified.

Figure 4.1 presents results for these three settings, under Bayesian modular inference and under PBMI. Since both modules have the same number of observations, in the case of PBMI we used both Algorithm 12 and Algorithm 13 in order to compare the results. In the first scenario, depicted in the leftmost panel, both modules are well-specified and there is

no dependence between datasets $z_1, \ldots, z_n$ and $y_1, \ldots, y_n$. As expected from Proposition 4.3 and Theorems 4.4 and 4.5, all three methods provide very similar results. Note that the distributions on θ_2 are located around 4/3, since the distribution on θ_1 is located around zero, and since y_i is centered around 4 and x_i around 3.

The middle panel illustrates a situation when both modules separately are well-specified, while the whole model is misspecified, as it does not represent correctly the correlation between the error terms. We notice that the cut posterior fails to account for correlation between the considered likelihoods, and so does Algorithm 12. Algorithm 13 has the capacity to adjust for it, thanks to the use of the same weights at both sampling stages. The rightmost panel shows that Posterior Bootstrap corrects the uncertainty when the second module is misspecified.

4 .2 Biased data

We revisit the biased data example of Liu et al. (2009), introduced in Section 2 .1:

$$\begin{aligned} x_{1,i} &\sim N(\theta_1, 1), \quad i = 1, \ldots, n_1, \\ x_{2,i} &\sim N(\theta_1 + \theta_2, 1), \quad i = 1, \ldots, n_2. \end{aligned} \tag{4.19}$$

The priors for θ_1 and θ_2 are independent $N(0, 1)$ and $N(0, 0.1^2)$. The data $x_{1,1:n_1}$ and $x_{2:1:n_2}$ is generated independently, from distributions $N(0, \sigma_1^2)$ and $N(1, \sigma_2^2)$, respectively. Thus the magnitude of the "bias" parameter θ_2 is poorly anticipated a priori. In what follows we consider different settings, depending on the values of $\sigma_1^2, \sigma_2^2, n_1$ and n_2. Note that in this example the pseudo-true parameter is the same as the true parameter under the above data-generating model.

Here we assume that the interest lies in the prediction using the first module, that is, prediction on x_{1,n_1+1}. To this end we empirically measured the expected log pointwise predictive density for a new dataset (elpd) (Bernardo and Smith, 2009; Gneiting and Raftery, 2007; Vehtari et al., 2017). Recall that, using notation from model (4.5), and denoting the new dataset by $\tilde{x}_1, \ldots, \tilde{x}_n$, elpd is defined as

$$\text{elpd} = \sum_{i=1}^{n} \int p^*(\tilde{x}_i) \log \hat{p}(\tilde{x}_i | x_{1:n}) \mathrm{d}\tilde{x}_i, \tag{4.20}$$

where $\hat{p}$ is the predictive distribution under a given method of inference.

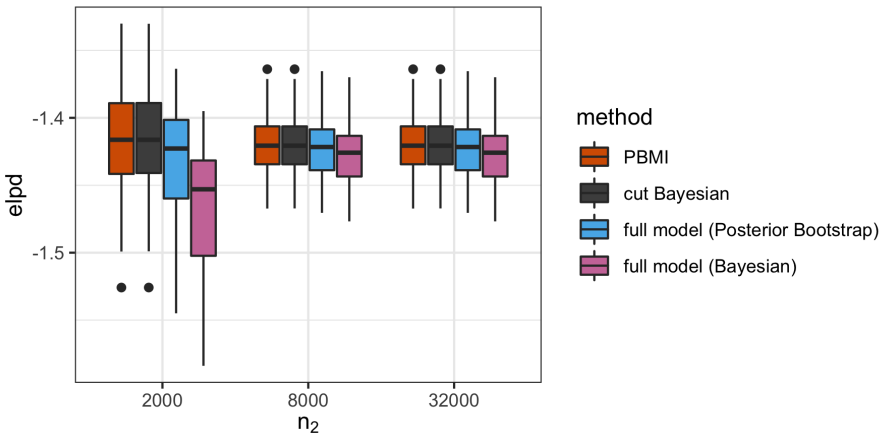


Figure 4.2: Boxplots of elpd values obtained for model (4.19). Each plot is based on 50 datasets generated with $\sigma_1^2 = 1$, $\sigma_2^2 = 0.5$, n_2 presented on the x-axis and $n_1 = n_2/10$. We use 2000 samples drawn using each method.

We used four different methods: PBMI (Algorithm 12), cut posterior, joint model using Bayesian inference and joint model with Posterior Bootstrap. We set $\sigma_1^2 = 1$ and $\sigma_2^2 = 0.5$, and considered different values of n_2 while keeping $n_1/n_2 = 1/10$. The results are presented in Figure 4.2. We notice that Posterior Bootstrap outperforms Bayesian inference on the full model yielding higher elpd, the best results are however achieved by PBMI and cut posteriors. We do not observe a significant difference between the latter two because the first module is well-specified.

The fact that the prior on θ_2 is strongly informative motivates also an analysis of the choice of hyperparameter v_0. Moreover, the interpretation of (θ_1, θ_2) does not change depending on σ_1^2 and σ_2^2, so it is justified to compare our samples drawn from PBMI to results obtained using the cut posterior for the correct model,

$$\begin{aligned} x_{1,i} &\sim N(\theta_1, \sigma_1^2), \quad i = 1, \ldots, n_1, \\ x_{2,i} &\sim N(\theta_1 + \theta_2, \sigma_2^2), \quad i = 1, \ldots, n_2. \end{aligned} \tag{4.21}$$

Note that this cut distribution is different from the 'Bayesian cut' depicted in Figure 4.2, which was based on the misspecified model (4.19). According to formula (4.17) we should set $w_0 = \sigma_1^2$ and $v_0 = \sigma_2^2$. In order to check empirically which value of v_0 yields results closest to the ones obtained with a Bayesian modular approach with the correct distribution (4.21), we measured Kolmogorov–Smirnov dissimilarity between samples θ_2 from the cut posterior for the correct model (4.21) and samples obtained using PBMI (see Figure 4.3). Recall that Kolmogorov–Smirnov dissimilarity measures the maximum distance between two empirical

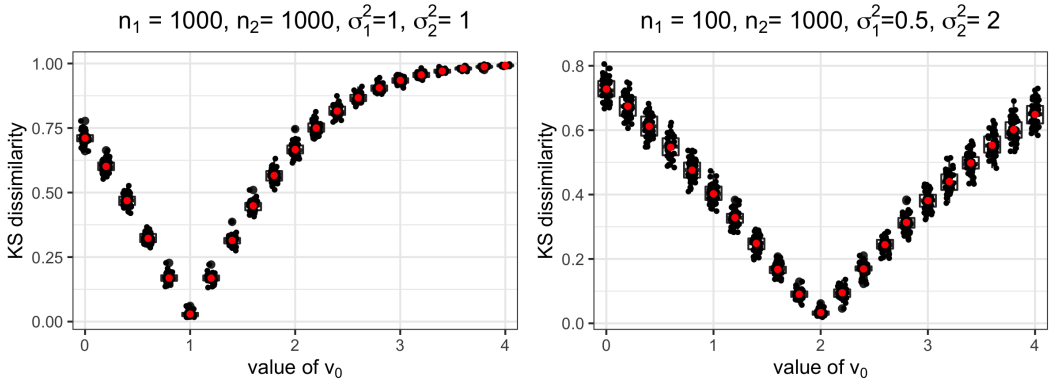


Figure 4.3: Kolmogorov–Smirnov dissimilarity between samples θ_2 from the cut posterior for the correct model (4.21) and samples obtained using PBMI (Algorithm 12) for model (4.19). We consider two different settings of $\sigma_1^2, \sigma_2^2, n_1$ and n_2. Each plot is based on 50 simulated datasets and $N = 2000$ samples drawn using each method.

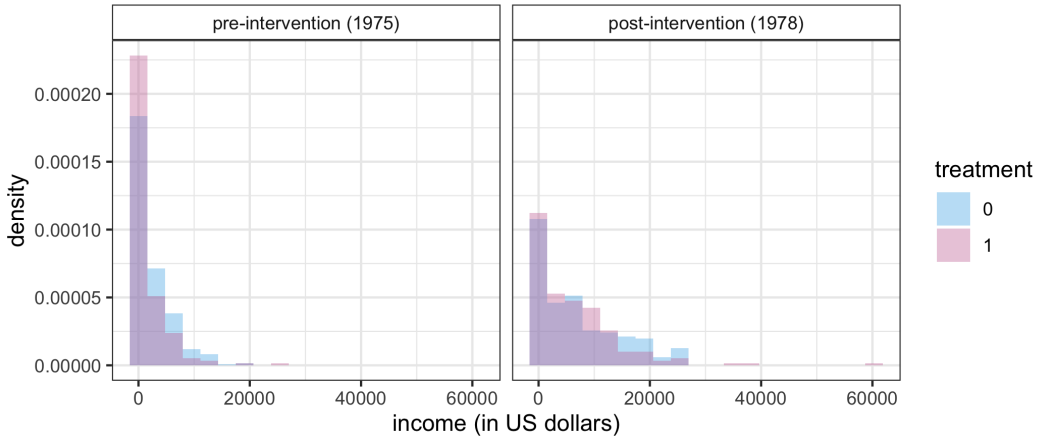


Figure 4.4: Histograms of income of the participants of the study, pre-intervention (left panel) and post-intervention (right panel) for the treated and the non-treated group.

cumulative distribution functions. The plots in Figure 4.3 confirm that setting v_0 according to formula (4.17) yields the best performance on this example.

4 .3 Causal inference with propensity scores

To illustrate the performance of Algorithm 13 on causal inference models described in Section 2 .2, we revisit the example studied extensively by LaLonde (1986) and later by Dehejia and Wahba (1999), who measured the effect of a labour training for the post-intervention income. The dataset comes from National Supported Work Demonstration and the Population Survey of Income Dynamics, and is currently available as the `lalonde` dataset in `MatchIt` `R` package. The treatment assignment was not randomized, in particular the assignment to the intervention was highly dependent on the pre-intervention income (see Figure 4.4).

Following Dehejia and Wahba (1999), we define the first module as a logistic regression

with treatment as the dependent variable, and the following covariates: age, age squared, number of years of schooling, race (black, Hispanic or white) an indicator for being married, an indicator for whether the individual has a high school degree, income in 1974 and income in 1975. We standardized all continuous variables so that they have mean 0 and variance 1. The second module is a linear regression model, where the outcome variable Y_i is the income of individual i after the intervention (in 1978), standardized to have expectation 0 and variance 1. The dependent variables are the indicator of treatment and quintiles of propensity scores, obtained in the first part of the model. To this end, we will write $\text{ps}_i = k$ to denote that the propensity score of individual i falls into the kth quintile. Using the notation from Section 2 .2, we can summarize this model as follows:

$$\begin{aligned} \text{logit}P(Z_i = 1|X_i) =& \beta_{1,0} + \beta_{1,1}X_{i,1} + \ldots + \beta_{1,p}X_{i,p} \\ Y_i =& \beta_{\text{treat}}Z_i + \beta_{2,0} + \sum_{k=2}^{5} \beta_{2,k}\delta_k(\text{ps}_i) + \epsilon, \quad \epsilon \sim N(0, \sigma^2), \end{aligned} \tag{4.22}$$

where $\delta_k(\text{ps}_i) = 1$ when $\text{ps}_i = k$ and 0 otherwise. Since the dependent variable has many zeros, as shown in Figure 4.4, we suspect that the second module is misspecified.

We compare the results obtained using Bayesian modular inference with those obtained using our Algorithm 13. In the case of Bayesian inference, we used vague priors for all parameters involved, that is, $\beta_{1,1}, \ldots, \beta_{1,p}$ in the logistic regression in the first module, as well as parameters β and σ in the linear regression in the second module. In the case of PBMI we set $w_0 = v_0 = 0$. The parameter of particular interest is β_{treat}, depicted in Figure 4.5. We notice that PBMI yields more narrow credible intervals around the mean, suggesting that the intervention had a positive impact on the income with increased certainty. We believe that this difference mainly stems from misspecification of the second module, which in the case of Bayesian inference leads to wrong uncertainty estimation.

4 .4 Epidemiological study

The example we present below comes originally from Maucort-Boulch et al. (2008), who investigated the international correlation between human papillomavirus (HPV) prevalence and cervical cancer incidence. It was used as an example in Plummer (2015); Jacob et al. (2017, 2020); Carmona and Nicholls (2020).

Cervical cancer is known to be caused by around 20 "high-risk" types of HPV. This

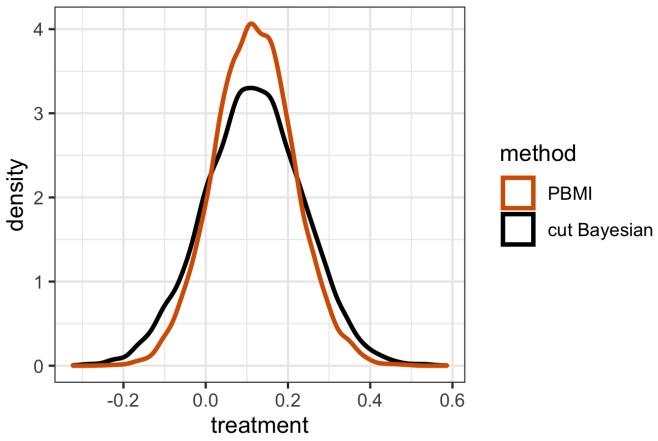


Figure 4.5: Density plot of β_{treat} drawn from the cut posterior under model (4.22), and from Algorithm 13. Each plot is based on 10000 samples.

epidemiological study considers two modules, based on two different data sources. The high-risk HPV prevalence dataset was obtained via a series of surveys carried out in 13 different countries. Incidence data come from cancer registries in the same populations. In each population i, for $i = 1, \ldots, 13$, we let Z_i be the number of women infected with high-risk HPV in a sample of size $n_{1,i}$ from the same population. Moreover, the outcome Y_i denotes the number of cancer cases arising from T_i woman-years of follow-up. We consider the following model.

$$\begin{aligned} Z_i \sim & \text{Binomial}(n_{1,i}, \theta_{1,i}) \\ \log \mu_i = & \theta_{2,1} + \theta_{2,2}\theta_{1,i} + T_i \\ Y_i \sim & \text{Poisson}(\mu_i). \end{aligned} \tag{4.23}$$

We set independent priors $\theta_{1,i} \sim \text{Beta}(1,1)$ for $i = 1, \ldots, 13$. The prior for θ_2 is $\theta_{2,1} \sim N(0, 10^3)$, $\theta_{2,2} \sim N(0, 10^3)$. As mentioned by Plummer (2015), the log-linear relationship $\log \mu_i = \theta_{2,1} + \theta_{2,2}\theta_{1,i} + T_i$ is speculative, which suggests the misspecification of the second module.

It is worth emphasizing that the small number of countries considered in this study ($n_2 = 13$) casts some doubts on whether the estimators have reached their asymptotic regime. Moreover, in this kind of studies, where the first module is based on individual-level data, whereas the second module is build around country-level data, the number of observations at the first level will be an order of magnitude larger than the number of countries. In particular, this setting does not correspond to any of the cases considered in Theorems 4.4

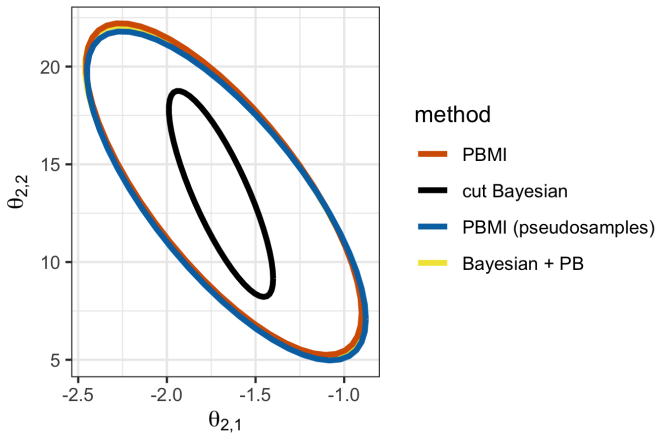


Figure 4.6: Density contour plots, corresponding to the 95% mass region, for $(\theta_{2,1}, \theta_{2,2})$ in model (4.23) obtained using four different methods. We used Algorithm 12 (PBMI), PBMI where at the first stage the prior information is incorporated via pseudosamples (PBMI pseudosamples), a combination of Bayesian inference at the first stage and Posterior Bootstrap at the second stage (Bayesian+PB), and Bayesian modular inference (cut Bayesian). Each plot is based on 5000 samples.

and 4.5.

We nevertheless applied PBMI to this example, using a slightly modified version of Algorithm 12, where at the first stage of inference we use separate Posterior Bootstrap procedures to draw $\tilde{\theta}_{1,i}$ for $i = 1, \ldots, 13$. We also applied to this model a version of PBMI where instead of penalization we use the pseudosamples strategy to incorporate the prior information in the first module. Note that the use of pseudosamples is justified thanks to conjugacy of the Beta prior. Finally, since we believe the first module is well-specified, and the use of conjugate priors allows for drawing directly samples from the Bayesian posterior, we combined the use of Bayesian inference for the first module with Posterior Bootstrap, given the draws of $\theta_{1,i}$, $i = 1, \ldots, 13$, for the second module. We compared the above approaches to the standard Bayesian modular inference. To sample from the cut posterior, we first drew directly 5000 samples from the posterior distribution for the first module, and then for each of these samples we ran a Markov chain targeting the posterior associated with the second module given θ_1. We assumed that the burn-in period of 5000 was enough for each chain to reach stationarity.

In Figure 4.6 we present results obtained for $(\theta_{2,1}, \theta_{2,2})$ using the four methods described above. We notice that the credible regions for the cut distribution are much smaller than the credible regions under any of the other methods. It is worth pointing out that credible regions under PBMI may be either narrower, as in Figure 4.6, or wider, as in Figure 4.5, than credible regions derived from cut posteriors.

We also calculated elpd (4.20) for PBMI and the cut posterior, using leave-one-out cross-validation. The results were -9.22 for the cut posterior and -6.27 for PBMI. In summary, these results suggest that Posterior Bootstrap corrects for misspecification of the second module by adjusting the uncertainty, and outperforms the cut posterior in terms of prediction.

5 Discussion

The proposal of this article participates in an emerging range of alternatives to standard Bayesian inference, designed to retain the benefits of fully probabilistic approaches while addressing the issue of model misspecification, and specifically here the setting of models built of multiple modules (Liu et al., 2009; Jacob et al., 2017; Leonelli et al., 2018; Goudie et al., 2019; Manderson and Goudie, 2020).

The main contributions of this work are two-fold. Firstly, we found the Laplace approximation of the cut posteriors and explained its lack of the appropriate frequentist coverage. As a remedy to this issue, we designed Posterior Bootstrap for Modular Inference, and proved its theoretical properties, regarding inference and prediction. In particular, we showed that its credible regions are asymptotically confidence sets in the frequentist sense. Our methodology is robust to misspecification and allows for performing the computations fully in parallel. Even though each of the module in this article was presented as a non-hierarchical model, our approach can be extended to certain types of hierarchical models, following the strategies developed by Pompe (2021). Moreover, the methodology could be extended to more than two modules.

An interesting research direction, not considered in this paper, is developing a principled way of choosing between Posterior Bootstrap for the full model or Posterior Bootstrap for Modular Inference. Alternatively, the user could combine Posterior Bootstrap with semi-modular inference, as we suggest in Appendix A .5 – this still however requires choosing the hyperparameter regulating the impact of the second module.

Our methodology inherits the limitations of Weighted Likelihood Bootstrap and Posterior Bootstrap. In particular, performing optimization typically involves computing gradients, therefore Posterior Bootstrap is not well-suited for distributions defined on discrete or mixed state spaces. Moreover, only asymptotic theory is known for methods based on Posterior Bootstrap. However, there are recent approaches looking into establishing finite sample

properties of methods based on Weighted Likelihood Bootstrap (Walker, 2021).

In asymptotic regimes, an alternative to our methodology would be to obtain the Normal distributions in Propositions 4.1 and 4.2, by replacing the asymptotic covariance matrices by empirical estimates. Our approach allows the use of prior knowledge and does not require computing information matrices, but a thorough comparison of these types of approaches would be of interest, perhaps using higher-order asymptotics.

Acknowledgements

We thank Judith Rousseau and Chris Carmona for feedback and helpful discussions. EP is supported by the EPSRC and MRC Centre for Doctoral Training in Next Generation Statistical Science: the Oxford-Warwick Statistics Programme, EP/L016710/1, and the Clarendon Fund.

A Proofs of theoretical results

A .1 Assumptions

Following (Newton, 1991; Lyddon et al., 2019; Pompe, 2021) we consider the following regularity conditions. The data-generating distributions of $x_{1,1:n_1}$ and $x_{2,1:n_2}$ are denoted by P_1^* and P_2^*.

Assumption 4.7. (Log likelihood functions). The log likelihood functions $\ell_1(x_1, \theta_1) = \log f_1(x_1|\theta_1)$ and $\ell_2(x_2, \theta_1, \theta_2) = \log f_2(x_2|\theta_1, \theta_2)$ are measurable and bounded from above, with $E_{p_1^*}|\ell_1(X_1, \theta_1)| < \infty$, $E_{p_2^*}|\ell_2(X_2, \theta_1, \theta_2)| < \infty$ for all $\theta_1 \in \Theta_1$, $\theta_2 \in \Theta_2$.

Assumption 4.8. (Identifiability). There exists a unique maximizing parameter value

$$\theta_1^* = \arg\max_{\theta_1 \in \Theta_1} \mathbb{E}_{p_1^*} \{\ell_1(X_1, \theta_1)\}.$$

Moreover, there is an open ball $B \subseteq \Theta_1$ containing θ_1^* such that for each $\theta_1 \in B$ we have a unique maximizer

$$\theta_2^*(\theta_1) = \arg\max_{\theta_2 \in \Theta_2} \mathbb{E}_{p_2^*} \{\ell_2(X_2, \theta_1, \theta_2)\}.$$

Furthermore, for all $\delta > 0$ there exists an $\epsilon > 0$ such that

$$\lim_{n\to\infty} P_1^* \left[\sup_{\|\theta_1 - \theta_1^*\| \geq \delta} \frac{1}{n} \sum_{j=1}^{n} \{\ell_1(X_{1,j}, \theta) - \ell_1(X_{1,j}, \theta_1^*)\} \leq -\epsilon \right] = 1, \tag{4.24}$$

and

$$\lim_{n\to\infty} P_2^* \left[\sup_{\|(\theta_1 - \theta_1^*, \theta_2 - \theta_2^*)\| \geq \delta} \frac{1}{n} \sum_{j=1}^{n} \{\ell_2(X_{2,j}, \theta_1, \theta_2) - \ell_2(X_{2,j}, \theta_1^*, \theta_2^*)\} \leq -\epsilon \right] = 1. \tag{4.25}$$

Assumption 4.9. (Smoothness of the log likelihood function). There is an open ball $B_1 \subseteq \Theta_1$ containing θ_1^* such that $\ell_1(x_1, \theta_1)$ is three times continuously differentiable with respect to $\theta_1 \in B_1$ almost surely on x_1 with respect to P_1^*. Moreover, there is an open ball $B_2 \subseteq \Theta_2$ containing θ_2^* such that $\ell_2(x_2, \theta_1, \theta_2)$ is three times continuously differentiable with respect to θ_1 and θ_2 for $(\theta_1, \theta_2) \in B_1 \times B_2$, almost surely on x_2 with respect to P_2^*.

We additionally impose certain moment conditions on the partial derivatives. There exist measurable functions G_k for $k = 1, \ldots, 6$ such that for $(\theta_1, \theta_2) \in B_1 \times B_2$ we have

$$\begin{aligned}
\left|\frac{\partial \ell_1(x_1, \theta_1)}{\partial \theta_j}\right| &\leq G_1(x_1), & 1 \leq j \leq d_1, \quad & \mathbb{E}_{p_{1*}}\left\{G_1(X_1)^4\right\} < \infty, \\
\left|\frac{\partial^2 \ell_1(x_1, \theta_1)}{\partial \theta_j \partial \theta_k}\right| &\leq G_2(x_1), & 1 \leq j \leq k \leq d_1, \quad & \mathbb{E}_{p_{1*}}\left\{G_2(X_1)^4\right\} < \infty, \\
\left|\frac{\partial^3 \ell_1(x_1, \theta_1)}{\partial \theta_j \partial \theta_k \partial \theta_l}\right| &\leq G_3(x_1), & 1 \leq j \leq k \leq l \leq d_1 \quad & \mathbb{E}_{p_{1*}}\{G_3(X_1)\} < \infty \\
\left|\frac{\partial \ell_2(x_2, \theta_1, \theta_2)}{\partial \theta_j}\right| &\leq G_4(x_2), & 1 \leq j \leq (d_1 + d_2), \quad & \mathbb{E}_{p_{2*}}\left\{G_1(X_2)^4\right\} < \infty, \\
\left|\frac{\partial^2 \ell_2(x_2, \theta_1, \theta_2)}{\partial \theta_j \partial \theta_k}\right| &\leq G_5(x_2), & 1 \leq j \leq k \leq (d_1 + d_2), \quad & \mathbb{E}_{p_{2*}}\left\{G_5(X_2)^4\right\} < \infty, \\
\left|\frac{\partial^3 \ell_2(x_2, \theta_1, \theta_2)}{\partial \theta_j \partial \theta_k \partial \theta_l}\right| &\leq G_6(x_2), & 1 \leq j \leq k \leq l \leq (d_1 + d_2), \quad & \mathbb{E}_{p_{2*}}\{G_6(X_2)\} < \infty,
\end{aligned}$$

where in the above equations θ_m, $m = 1, \ldots, d_1 + d_2$, denotes the m-th coordinate of vector $\theta = (\theta_1, \theta_2)$.

Assumption 4.10. (Positive definiteness of information matrices and linear independence of partial derivatives). For B_i like in Assumption 4.9, $i = 1, 2$, the matrices I_i and J_i defined

as

$$I_1(\theta_1) = \mathbb{E}_{p_1^*}\left\{\frac{\partial \ell_1(X_1,\theta_1)}{\partial \theta_1}\left(\frac{\partial \ell_1(X_1,\theta_1)}{\partial \theta_1}\right)^T\right\}, \qquad J_1(\theta_1) = -\mathbb{E}_{p_1^*}\left\{\frac{\partial^2 \ell_1(X_1,\theta_1)}{\partial \theta_1 \partial \theta_1^T}\right\},$$
$$I_2(\theta_1,\theta_2) = \mathbb{E}_{p_2^*}\left\{\frac{\partial \ell_2(X_2,\theta_1,\theta_2)}{\partial \theta_2}\left(\frac{\partial \ell_2(X_2,\theta_1,\theta_2)}{\partial \theta_2}\right)^T\right\}, \quad J_2(\theta_1,\theta_2) = -\mathbb{E}_{p_2^*}\left\{\frac{\partial^2 \ell_2(X_2,\theta_1,\theta_2)}{\partial \theta_2 \partial \theta_2^T}\right\},$$

are positive definite for $\theta_i \in B_i$ with all elements finite.

Assumption 4.11. (Smoothness of log prior density). For $i = 1, 2$ the function $\log \pi_i(\theta_i)$ is measurable, upper-bounded on Θ_i, and three times continuously differentiable at θ_i^*.

We note that (4.24) implies that $\hat{\theta}_1$ converges to θ_1^*, like in standard asymptotic analysis for one module. To show that $\hat{\theta}_2$ converges to θ_2^*, we notice that

$$\begin{aligned}
&\frac{1}{n}\sum_{j=1}^{n}\left\{\ell_2(X_{2,j},\hat{\theta}_1,\hat{\theta}_2) - \ell_2(X_{2,j},\theta_1^*,\theta_2^*)\right\} \\
=&\frac{1}{n}\sum_{j=1}^{n}\left\{\ell_2(X_{2,j},\hat{\theta}_1,\hat{\theta}_2) - \ell_2(X_{2,j},\hat{\theta}_1,\theta_2^*)\right\} + \frac{1}{n}\sum_{j=1}^{n}\left\{\ell_2(X_{2,j},\hat{\theta}_1,\theta_2^*) - \ell_2(X_{2,j},\theta_1^*,\theta_2^*)\right\}.
\end{aligned}$$

Thus by equation (4.25), the fact that $\hat{\theta}_1$ converges to θ_1^*, and the smoothness of ℓ_2 with respect to θ_1 in the neighbourhood of θ_1^* we infer that $\hat{\theta}_2$ converges to θ_2^*.

To show that $(\tilde{\theta}_1, \tilde{\theta}_2)$ converges to (θ_1^*, θ_2^*), we can therefore follow the steps of Newton (1991).

A .2 Proof of Proposition 4.3

Proof. In this proof we will use the fact that, as noticed by Plummer (2015), there are two equivalent ways of defining cut posteriors. The first one relies on writing its density in the form (4.3). Alternatively, one can define the cut posterior for (θ_1, θ_2) through its sampling scheme: we first sample θ_1 from $\pi(\theta_1|x_{1,1:n_1})$, and then θ_2 given θ_1 from $\pi(\theta_2|x_{2,1:n_2}, \theta_1)$.

We will now follow standard steps to obtain the Laplace approximation of posterior densities, utilizing the Taylor expansion of the logarithm of the density; see for example Van der Vaart (2000). Firstly, the second definition above justifies why we choose $(\hat{\theta}_1, \hat{\theta}_2)$ as the centring point. The assumptions listed in Section A .1 ensure that the incurred error in the Taylor expansion is small enough so that $\sqrt{n_2}(\theta_1 - \hat{\theta}_2, \theta_1 - \hat{\theta}_2)$ converges in distribution to a Normal distribution. It remains to find the limiting covariance matrix Σ_B.

We have

$$\log \pi_{\text{cut}}(\theta_1, \theta_2) = \sum_{j=1}^{n_1} \log f_1(x_{1,j}|\theta_1) + \sum_{j=1}^{n_2} \log f_2(x_{2,j}|\theta_1, \theta_2) + \log \pi_1(\theta_1) + \log \pi_2(\theta_2) - \log \int \prod_{j=1}^{n_2} f_2(x_{2,j}|\theta_1, \theta_2)\pi_2(\theta_2)d\theta_2 + \text{constant}.$$

We thus obtain

$$-\frac{1}{n_2}\frac{\partial^2 \log \pi_{\text{cut}}(\hat{\theta}_1, \hat{\theta}_2)}{\partial\theta_2\partial\theta_2^T} = -\frac{1}{n_2}\sum_{j=1}^{n_2}\frac{\partial^2 \log f_2(x_{2,j}|\hat{\theta}_1, \hat{\theta}_2)}{\partial\theta_2\partial\theta_2^T},$$

and

$$-\frac{1}{n_2}\frac{\partial^2 \log \pi_{\text{cut}}(\hat{\theta}_1, \hat{\theta}_2)}{\partial\theta_1\partial\theta_2^T} = -\frac{1}{n_2}\sum_{j=1}^{n_2}\frac{\partial^2 \log f_2(x_{2,j}|\hat{\theta}_1, \hat{\theta}_2)}{\partial\theta_1\partial\theta_2^T},$$

which converge in distribution to J_2^* and R_J, respectively. Therefore

$$\Sigma_B^{-1} = \begin{pmatrix} A & R_J^* \\ R_J^{*T} & J_2^* \end{pmatrix}$$

for some matrix A. To find this matrix, we use the formula for inverting a block matrix,

$$\Sigma_B = \begin{pmatrix} (A - R_J^* J_2^{*-1} R_J^T)^{-1} & -(A - R_J^* J_2^{*-1} R_J^{*T})^{-1} R_J^* J_2^{*-1} \\ -J_2^{*-1} R_J^{*T} (A - R_J^* J_2^{*-1} R_J^{*T})^{-1} & J_2^{*-1} + J_2^{*-1} R_J^{*T} (A - R_J^* J_2^{*-1} R_J^{*T})^{-1} R_J^* J_2^{*-1} \end{pmatrix}.$$

Recall that θ_1 is drawn from $\pi(\theta_1|x_{1,1:n_1})$, so the limiting distribution of $\sqrt{n_1}(\theta_1 - \hat{\theta}_1)$ is $N(0, J_1^{*-1})$. Therefore we must have

$$(A - R_J^* J_2^{*-1} R_J^{*T})^{-1} = \frac{1}{\alpha} J_1^{*-1},$$

which provides the expression of A and completes the proof. □

A .3 Proofs of Theorems 4.4 and 4.5

We first present the proof of Theorem 4.4. We will then use a similar reasoning to prove Theorem 4.5.

Proof. Let $\ell_1(x_{1,j},\theta_1)=\log f_1(x_{1,j}|\theta_1)$ and $\ell_2(x_{2,j},\theta_1,\theta_2)=\log f_2(x_{2,j}|\theta_1,\theta_2)$. We also define

$$L_1^w(\theta_1)=\sum_{j=1}^{n_1} w_j\ell_1(x_{1,j},\theta_1)+w_0^T\log\pi_1(\theta_1)$$

and

$$L_2^v(\theta_1,\theta_2)=\sum_{j=1}^{n_2} v_j\ell_2(x_{2,j},\theta_1,\theta_2)+v_0(\theta_1)\log\pi_2(\theta_2).$$

Then

$$0=\frac{\partial L_1^w(\tilde\theta_1)}{\partial\theta_1}=\frac{\partial L_1^w(\hat\theta_1)}{\partial\theta_1}+\frac{\partial^2 L_1^w(\hat\theta_1)}{\partial\theta_1\partial\theta_1^T}(\tilde\theta_1-\hat\theta_1)+O_P\left(n_2\|\tilde\theta_1-\hat\theta_1\|^2\right),$$

and therefore

$$-\frac{1}{\sqrt{n_2}}\frac{\partial L_1^w(\hat\theta_1)}{\partial\theta_1}=\frac{1}{n_2}\frac{\partial^2 L_1^w(\hat\theta_1)}{\partial\theta_1\partial\theta_1^T}\sqrt{n_2}(\tilde\theta_1-\hat\theta_1)+O_P\left(\sqrt{n_2}\|\tilde\theta_1-\hat\theta_1\|^2\right). \tag{4.26}$$

For the second module we have

$$0=\frac{\partial L_2^v(\tilde\theta_1,\tilde\theta_2)}{\partial\theta_2}=\frac{\partial L_2^v(\hat\theta_1,\hat\theta_2)}{\partial\theta_2}+\frac{\partial^2 L_2^v(\hat\theta_1,\hat\theta_2)}{\partial\theta_1^T\partial\theta_2}(\tilde\theta_1-\hat\theta_1)+\frac{\partial^2 L_2^v(\hat\theta_1,\hat\theta_2)}{\partial\theta_2\partial\theta_2^T}(\tilde\theta_2-\hat\theta_2)+O_P\left(n_2\|\tilde\theta-\hat\theta\|^2\right),$$

where $\tilde\theta=(\tilde\theta_1,\tilde\theta_2)$ and $\hat\theta=(\hat\theta_1,\hat\theta_2)$. Similarly, we have

$$-\frac{1}{\sqrt{n_2}}\frac{\partial L_2^v(\hat\theta_1,\hat\theta_2)}{\partial\theta_2}-\frac{1}{n_2}\frac{\partial^2 L_2^v(\hat\theta_1,\hat\theta_2)}{\partial\theta_1^T\partial\theta_2}\sqrt{n_2}(\tilde\theta_1-\hat\theta_1)=\frac{\partial^2 L_2^v(\hat\theta_1,\hat\theta_2)}{\partial\theta_2\partial\theta_2^T}\frac{\sqrt{n_2}}{n_2}(\tilde\theta_2-\hat\theta_2)+O_P\left(\sqrt{n_2}\|\tilde\theta-\hat\theta\|^2\right). \tag{4.27}$$

Note that

$$-\frac{1}{n_2}\frac{\partial^2 L_1^w(\hat\theta_1)}{\partial\theta_1\partial\theta_1^T},\quad -\frac{1}{n_2}\frac{\partial^2 L_2^v(\hat\theta_1,\hat\theta_2)}{\partial\theta_1^T\partial\theta_2},\quad -\frac{1}{n_2}\frac{\partial^2 L_2^v(\hat\theta_1,\hat\theta_2)}{\partial\theta_2\partial\theta_2^T}$$

converge in probability to αJ_1^*, R_J^{*T} and J_2^*, respectively. Therefore using (4.26) and (4.27) we get that

$$\begin{aligned}\sqrt{n_2}(\tilde\theta_2-\hat\theta_2)=&J_2^{*-1}\frac{1}{\sqrt{n_2}}\frac{\partial L_2^v(\hat\theta_1,\hat\theta_2)}{\partial\theta_2}-J_2^{*-1}R_J^{*T}\sqrt{n_2}(\tilde\theta_1-\hat\theta_1)+O_P\left(\sqrt{n_2}\|\tilde\theta-\hat\theta\|^2\right)\\ =&J_2^{*-1}\frac{1}{\sqrt{n_2}}\frac{\partial L_2^v(\hat\theta_1,\hat\theta_2)}{\partial\theta_2}-J_2^{*-1}R_J^{*T}J_1^{*-1}\frac{1}{\alpha\sqrt{n_2}}\frac{\partial L_1^w(\hat\theta_1)}{\partial\theta_1}+O_P\left(\sqrt{n_2}\|\tilde\theta-\hat\theta\|^2\right).\end{aligned} \tag{4.28}$$

We will now find the asymptotic distribution of

$$\left(\frac{1}{\sqrt{n_2}}\frac{\partial L_1^w(\hat\theta_1)}{\partial\theta_1},\frac{1}{\sqrt{n_2}}\frac{\partial L_2^v(\hat\theta_1,\hat\theta_2)}{\partial\theta_2}\Big|x_{1,1:n_1},x_{2,1:n_2}\right). \tag{4.29}$$

Since

$$\sum_{j=1}^{n_1} \frac{\partial \ell_1(x_{1,j}, \hat{\theta}_1)}{\partial \theta_1} = 0, \quad \sum_{j=1}^{n_2} \frac{\partial \ell_2(x_{2,j}, \hat{\theta}_1, \hat{\theta}_2)}{\partial \theta_2} = 0$$

we have

$$\frac{\partial L_1^w(\hat{\theta}_1)}{\partial \theta_1} = \sum_{j=1}^{n_1} \frac{(w_j - 1)\partial \ell_1(x_{1,j}, \hat{\theta}_1)}{\partial \theta_1} + w_0^T \frac{\partial \log \pi_1(\hat{\theta}_1)}{\partial \theta_1}.$$

Similarly,

$$\frac{\partial L_2^v(\hat{\theta}_1, \hat{\theta}_2)}{\partial \theta_2} = \sum_{j=1}^{n_2} \frac{(v_j - 1)\partial \ell_2(x_{2,j}, \hat{\theta}_1, \hat{\theta}_2)}{\partial \theta_2} + v_0^T \frac{\partial \log \pi_2(\hat{\theta}_2)}{\partial \theta_2}.$$

Therefore, applying the central limit theorem we get that the asymptotic distribution of (4.29) is Normal with mean 0 and covariance matrix

$$\begin{pmatrix} \alpha I_1^* & 0 \\ 0 & I_2^* \end{pmatrix},$$

due to independence of $w_{1:n_1}$ and $v_{1:n_2}$. Finally, using Slutsky's theorem, we get that the asymptotic distribution of $\sqrt{n_2}(\tilde{\theta}_1 - \hat{\theta}_1, \tilde{\theta}_2 - \hat{\theta}_2)$ given $x_{1,1:n_1}, x_{2,1:n_2}$ is Normal with mean 0 and covariance matrix (4.8). □

In what follows we present the proof of Theorem 4.5.

Proof. We let $\ell_1(x_{1,j}, \theta_1) = \log f_1(x_{1,j}|\theta_1)$ and $\ell_2(x_{2,j}, \theta_1, \theta_2) = \log f_2(x_{2,j}|\theta_1, \theta_2)$. We also define

$$L_1^w(\theta_1) = \sum_{j=1}^{n} w_j \ell_1(x_{1,j}, \theta_1) + w_0^T \log \pi_1(\theta_1)$$

and

$$L_2^w(\theta_1, \theta_2) = \sum_{j=1}^{n} w_j \ell_2(x_{2,j}, \theta_1, \theta_2) + v_0^T \log \pi_2(\theta_2).$$

We can repeat the proof of Theorem 4.4 until (4.28), replacing n_1 and n_2 with n, and v_j with w_j. Similarly to that proof, we have

$$\frac{\partial L_1^w(\hat{\theta}_1)}{\partial \theta_1} = \sum_{j=1}^{n} \frac{(w_j - 1)\partial \ell_1(x_{1,j}, \hat{\theta}_1)}{\partial \theta_1} + w_0^T \frac{\partial \log \pi_1(\hat{\theta}_1)}{\partial \theta_1},$$

and

$$\frac{\partial L_2^w(\hat{\theta}_1, \hat{\theta}_2)}{\partial \theta_2} = \sum_{j=1}^{n} \frac{(w_j - 1)\partial \ell_2(x_{2,j}, \hat{\theta}_1, \hat{\theta}_2)}{\partial \theta_2} + v_0^T \frac{\partial \log \pi_2(\hat{\theta}_2)}{\partial \theta_2}.$$

Therefore, applying the central limit theorem and the law of large numbers we get that the

asymptotic distribution of

$$\left(\frac{1}{\sqrt{n}}\frac{\partial L_1^w(\hat{\theta}_1)}{\partial \theta_1}, \frac{1}{\sqrt{n}}\frac{\partial L_2^w(\hat{\theta}_1, \hat{\theta}_2)}{\partial \theta_2}\Big| x_{1,1:n}, x_{2,1:n}\right)$$

is Normal with mean 0 and covariance matrix

$$\begin{pmatrix} I_1^* & R_I^* \\ R_I^{*T} & I_2^* \end{pmatrix},$$

since random variables $(w_j - 1)$ for $j = 1, \ldots, n$ are i.i.d. with expectation 0 and variance 1. This completes the proof. □

A .4 Prediction under Posterior Bootstrap for Modular Inference versus Bayesian modular inference

In what follows we first prove Proposition 4.6. We then design an example showing that each of the two approaches, Bayesian modular inference and PBMI, may yield better prediction for x_{2,n_2+1}, depending on the exact setting.

Proof. Recall that x_2 denotes the new observations sampled from p_2^*. In order to simplify the notation, we will denote in this proof $(x_{1,1:n_1}, x_{2,1:n_2})$ by x. Moreover, by p^* we will denote the density of the true underlying distribution of x. Analogously to (4.13), we define the risk associated with prediction for the second module, as

$$E_{p^*}\left[D_{KL}\left\{p_2^* || \hat{p}(\cdot|x)\right\}\right] = \int p^*(x)\left\{\int p_2^*(x_2)\log\frac{p_2^*(x_2)}{\hat{p}(x_2|x)}\mathrm{d}x_2\right\}\mathrm{d}x,$$

for a predictive distribution $\hat{p}$, which in our case will be either $\hat{p}_B$ or $\hat{p}_{PB}$. Firstly, notice that

$$E_{p^*}\left[D_{KL}\left\{p_2^* || \hat{p}(\cdot|x)\right\}\right] = -\int p^*(x)\left\{\int p_2^*(x_2)\log\hat{p}(x_2|x)\mathrm{d}x_2\right\}\mathrm{d}x + c, \tag{4.30}$$

where the constant c does not depend on the prediction method.

In order to find asymptotic expansions of

$$E_{p^*}\left[D_{KL}\left\{p_2^* || \hat{p}_B(\cdot|x)\right\}\right]$$

and

$$E_{p^*}\left[D_{KL}\left\{p_2^*||\hat{p}_{PB}(\cdot|x)\right\}\right]$$

up to $o(n_2^{-1})$, we will follow the steps of (Fushiki, 2010) and (Shimodaira, 2000). We have

$$\hat{p}_B(x_2|x) = \int f_2(x_2|\theta_1,\theta_2)\pi_{\text{cut}}(\theta|x)\mathrm{d}\theta \tag{4.31}$$

and

$$\hat{p}_{PB}(x_2|x) = \int f_2(x_2|\theta_1,\theta_2)p_{\text{cut}}^{PB}(\theta|x)\mathrm{d}\theta.$$

We start by expanding asymptotically equation (4.31) around $(\hat{\theta}_1,\hat{\theta}_2)$. Setting $\theta=(\theta_1,\theta_2)$, we have

$$\begin{aligned}
\hat{p}_B(x_2|x) =& f_2(x_2|\hat{\theta}_1,\hat{\theta}_2)\int \pi_{\text{cut}}(\theta|x)\mathrm{d}\theta \\
&+ \frac{\partial f_2(x_2|\hat{\theta}_1,\hat{\theta}_2)}{\partial\theta}\int(\theta-\hat{\theta})\pi_{\text{cut}}(\theta|x)\mathrm{d}\theta \\
&+ \int(\theta-\hat{\theta})^T\frac{\partial^2 f_2(x_2|\hat{\theta}_1,\hat{\theta}_2)}{\partial\theta\partial\theta^T}(\theta-\hat{\theta})\pi_{\text{cut}}(\theta|x)\mathrm{d}\theta + o_P(n_2^{-1}) \\
=& f_2(x_2|\hat{\theta}_1,\hat{\theta}_2) + \frac{\partial f_2(x_2|\hat{\theta}_1,\hat{\theta}_2)}{\partial\theta}\int(\theta-\hat{\theta})\pi_{\text{cut}}(\theta|x)\mathrm{d}\theta \\
&+ \frac{1}{2n}\mathrm{tr}\left\{\frac{\partial^2 f_2(x_2|\hat{\theta}_1,\hat{\theta}_2)}{\partial\theta\partial\theta^T}\Sigma_B\right\} + o_P(n_2^{-1}),
\end{aligned}$$

and therefore

$$\begin{aligned}
\hat{p}_B(x_2|x) =& f_2(x_2|\hat{\theta}_1,\hat{\theta}_2)\left[1 + \frac{\partial\log f_2(x_2|\hat{\theta}_1,\hat{\theta}_2)}{\partial\theta}\int(\theta-\hat{\theta})\pi_{\text{cut}}(\theta_1,\theta_2|x)\mathrm{d}\theta_1\mathrm{d}\theta_2\right. \\
&+ \left.\frac{1}{2n}\mathrm{tr}\left\{\frac{1}{f_2(x_2|\hat{\theta}_1,\hat{\theta}_2)}\frac{\partial^2 f_2(x_2|\hat{\theta}_1,\hat{\theta}_2)}{\partial\theta\partial\theta^T}\Sigma_B\right\}\right] + o_P(n_2^{-1}).
\end{aligned}$$

Following Shimodaira (2000), we will use the identity

$$\frac{1}{g}\frac{\partial^2 g}{\partial\theta\partial\theta^T} = \frac{\partial\log g}{\partial\theta}\frac{\partial\log g}{\partial\theta^T} + \frac{\partial^2\log g}{\partial\theta\partial\theta^T}.$$

Applying this identity, and the Taylor expansion of $\log(1+y)$ around $y=0$, we get

$$\begin{aligned}
\log\hat{p}_B(x_2|x) =& \log f_2(x_2|\hat{\theta}_1,\hat{\theta}_2) + \frac{\partial\log f_2(x_2|\hat{\theta}_1,\hat{\theta}_2)}{\partial\theta}\int(\theta-\hat{\theta})\pi_{\text{cut}}(\theta_1,\theta_2|x)\mathrm{d}\theta_1\mathrm{d}\theta_2 \\
&+ \frac{1}{2n_2}\mathrm{tr}\left[\left\{I_{f_2}^{\hat{\theta}}(x_2) - J_{f_2}^{\hat{\theta}}(x_2)\right\}\Sigma_B\right] + o_P(n_2^{-1}),
\end{aligned}$$

where

$$I^{\theta}_{f_2}(x_2) = \frac{\partial \log f_2(x_2|\theta_1,\theta_2)}{\partial \theta}\left(\frac{\partial \log f_2(x_2|\theta_1,\theta_2)}{\partial \theta}\right)^T, \quad J^{\theta}_{f_2}(x_2) = -\frac{\partial^2 \log f_2(x_2|\theta_1,\theta_2)}{\partial\theta\partial\theta^T}.$$

Finally, we use equality (4.30) to get that

$$\begin{aligned} E_{p^*}\left[D_{KL}\left\{p_2^*||\hat{p}_B(\cdot|x)\right\}\right] =& E_{p^*}\left[D_{KL}\left\{p_2^*||f_2(\cdot|\hat{\theta}_1,\hat{\theta}_2)\right\}\right] - E_{p_2^*}\left\{\frac{\partial \log f_2(x_2|\theta_1^*,\theta_2^*)}{\partial\theta}\right\}\frac{a_B}{n_2} \\ &- \frac{1}{2n_2}\mathrm{tr}\left\{\left(I^*_{f_2} - J^*_{f_2}\right)\Sigma_B\right\} + o(n_2^{-1}), \end{aligned}$$

for matrices $I^*_{f_2}$ and $J^*_{f_2}$ defined in (4.14). Performing analogous calculations for Posterior Bootstrap, we obtain

$$\begin{aligned} E_{p^*}\left[D_{KL}\left\{p_2^*||\hat{p}_{PB}(\cdot|x)\right\}\right] =& E_{p^*}\left[D_{KL}\left\{p_2^*||f_2(\cdot|\hat{\theta}_1,\hat{\theta}_2)\right\}\right] - E_{p_2^*}\left\{\frac{\partial \log f_2(x_2|\theta_1^*,\theta_2^*)}{\partial\theta}\right\}\frac{a_{PB}}{n_2} \\ &- \frac{1}{2n_2}\mathrm{tr}\left\{\left(I^*_{f_2} - J^*_{f_2}\right\}\Sigma_{PB}\right) + o(n_2^{-1}), \end{aligned}$$

which completes the proof. □

As mentioned earlier, we will now use the above expansions to show that Posterior Bootstrap in some settings yields asymptotically better prediction for the second module, whereas in other settings Bayesian modular inference performs better. Let us consider the following misspecified model:

$$\begin{aligned} &x_{1,1:n} \sim N(\theta_1, 1), \quad \theta_1 \sim N(0,1), \\ &x_{2,1:n} \sim N\left((\theta_1,\theta_2),\Sigma_1\right), \quad \theta_2 \sim N(0,1), \end{aligned} \tag{4.32}$$

where

$$\Sigma_1 = \begin{pmatrix} 4/3 & -2/3 \\ -2/3 & 4/3 \end{pmatrix} \quad \text{and therefore} \quad \Sigma_1^{-1} = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}.$$

In fact the data are generated independently according to

$$x_{1,1:n} \sim N(0,\sigma^2), \quad x_{2,1:n} \sim N(0,\Sigma_2) \tag{4.33}$$

where

$$\Sigma_2 = \Sigma_1 \begin{pmatrix} 0.8 & 0.3 \\ 0.3 & 0.8 \end{pmatrix} \Sigma_1.$$

We consider below different scenarios for the value of σ.

Firstly, note that $(\theta_1^*, \theta_2^*) = (0, 0)$, and that the pseudo-true parameter for θ_1 with respect to f_2 is also equal to 0, thus

$$E_{p_2^*}\left\{\frac{\partial \log f_2(x_2|\theta_1^*, \theta_2^*)}{\partial \theta}\right\} = 0.$$

Consequently

$$\begin{aligned} &E_{p^*}\left[D_{KL}\left\{p_2^*||\hat{p}_{PB}(\cdot|x)\right\}\right] - E_{p^*}\left[D_{KL}\left\{p_2^*||\hat{p}_B(\cdot|x)\right\}\right] \\ =&\frac{1}{2n}\mathrm{tr}\left\{\left(I_{f_2}^* - J_{f_2}^*\right)\left(\Sigma_B - \Sigma_{PB}\right)\right\} + o(n^{-1}). \end{aligned}$$

For this example we have

$$I_{f_2}^* = \begin{pmatrix} 0.8 & 0.3 \\ 0.3 & 0.8 \end{pmatrix}, \quad J_{f_2}^* = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}$$

and $R_J^* = 0.5$. Thus $J_1^{*-1} = 1$, $J_2^{*-1} = 1$, $I_1^* = \sigma^2$ and $I_1^* = 2.5$. We also have

$$\Sigma_B = \begin{pmatrix} 1 & -0.5 \\ -0.5 & 1.25 \end{pmatrix}, \quad \Sigma_{PB} = \begin{pmatrix} \sigma^2 & -0.5\sigma^2 \\ -0.5\sigma^2 & 0.8 + 0.25\sigma^2. \end{pmatrix}$$

Since

$$I_{f_2}^* - J_{f_2}^* = \begin{pmatrix} -0.2 & -0.2 \\ -0.2 & -0.2 \end{pmatrix},$$

we have $\mathrm{tr}\{(I_{f_2}^{\theta^*} - J_{f_2}^{\theta^*})\Sigma_B\} = -0.25$. As for the corresponding trace for prediction based on Posterior Bootstrap, for $\sigma = 1$ we have $\mathrm{tr}\{(I_{f_2}^{\theta^*} - J_{f_2}^{\theta^*})\Sigma_{PB}\} = -0.21$, whereas for $\sigma = 2$ the value of the trace is -0.36.

To summarize, we have shown that Algorithm 12 in model (4.32) does not necessarily yield asymptotically better prediction for the second module than Bayesian inference, and that it depends on the value of σ^2 in the data-generating mechanism (4.33).

A .5 Extensions to semi-modular inference

In what follows we show how PBMI can be extension to Posterior Bootstrap for semi-modular inference developed by (Carmona and Nicholls, 2020). We will show that the constructed algorithm requires only one optimization run per iteration, and that it provides the correct frequentist coverage. In what follows, we assume we work under Scenario 1. However, analogous procedures and calculations could be extended to Scenario 2.

Recall that semi-modular inference utilizes a parameter $\eta \in [0,1]$, which calibrates the impact of the second module. For $\eta = 0$ semi-modular inference boils down to the cut posterior, and for $\eta = 1$ it is equivalent to using the full model.

Firstly, let us introduce some additional notation. We assume that n_1, n_2 go to infinity at the rate such that $n_1/n_2 \to \alpha$ and suppose that all regularity conditions hold. We fix $\eta > 0$. Let

$$f_\eta(X_1, X_2|\theta) = f_\eta(X_1, X_2|\theta_1, \theta_2) = \{\alpha \log f_1(X_1|\theta_1) + \eta \log f_2(X_2|\theta_1, \theta_2)\}.$$

Then the pseudo-true parameter $\theta^*_\eta = (\theta^*_{1,\eta}, \theta^*_{2,\eta})$ is defined as

$$(\theta^*_{1,\eta}, \theta^*_{2,\eta}) = \arg\max_{\theta_1,\theta_2} E_{p^*_1,p^*_2} f_\eta(\theta_1, \theta_2). \tag{4.34}$$

We also define matrices

$$I_i^{j,k} = E_{p_i^*}\left\{\frac{\partial \log f_i(X_i|\theta^*_\eta)}{\partial \theta_j}\left(\frac{\partial \log f_i(X_i|\theta^*_\eta)}{\partial \theta_k}\right)^T\right\}, \quad i,j,k = 1,2$$

and

$$J_i^{j,k} = -E_{p_i^*}\left\{\frac{\partial^2 \log f_i(X_i|\theta^*_\eta)}{\partial \theta_j \partial \theta_k^T}\right\} \quad i,j,k = 1,2.$$

$$I^*_\eta = \begin{pmatrix} \alpha/\eta^2 I_1^{1,1} + I_2^{1,1} & I_2^{1,2} \\ I_2^{2,1} & I_2^{2,2} \end{pmatrix}, \quad J^*_\eta = \begin{pmatrix} \alpha/\eta J_1^{1,1} + J_2^{1,1} & J_2^{1,2} \\ J_2^{2,1} & J_2^{2,2} \end{pmatrix}. \tag{4.35}$$

We now consider the following point estimation procedure, returning $\hat{\theta}_{1,\eta}, \bar{\theta}_{2,\eta})$. We compute

$$(\hat{\theta}_{1,\eta}, \bar{\theta}_{2,\eta}) = \arg\max_{\theta_1,\theta_2}\left\{\sum_{j=1}^{n_1} \log f_1(x_{1,j}|\theta_1) + \eta \sum_{j=1}^{n_2} \log f_2(x_{2,j}|\theta_1, \theta_2)\right\},$$

followed by

$$\hat{\theta}_{2,\eta} = \arg\max_{\theta_2} \sum_{j=1}^{n_2} \log f_2(x_{2,j}|\hat{\theta}_{1,\eta}, \theta_2). \tag{4.36}$$

We will show that $\hat{\theta}_{2,\eta} = \bar{\theta}_{2,\eta}$, so in other words, step (4.36) is redundant.

Note that by definition (4.36), we have

$$\sum_{j=1}^{n_2} \log f_2(x_{2,j}|\hat{\theta}_{1,\eta}, \hat{\theta}_{2,\eta}) \geq \sum_{j=1}^{n_2} \log f_2(x_{2,j}|\hat{\theta}_{1,\eta}, \bar{\theta}_{2,\eta}).$$

Suppose that

$$\sum_{j=1}^{n_2} \log f_2(x_{2,j}|\hat{\theta}_{1,\eta}, \hat{\theta}_{2,\eta}) > \sum_{j=1}^{n_2} \log f_2(x_{2,j}|\hat{\theta}_{1,\eta}, \bar{\theta}_{2,\eta}).$$

Since $\eta > 0$, we would then have

$$\sum_{j=1}^{n_1} \log f_1(x_{1,j}|\hat{\theta}_{1,\eta}) + \eta \sum_{j=1}^{n_2} \log f_2(x_{2,j}|\hat{\theta}_{1,\eta}, \hat{\theta}_{2,\eta}) > \sum_{j=1}^{n_1} \log f_1(x_{1,j}|\hat{\theta}_{1,\eta}) + \eta \sum_{j=1}^{n_2} \log f_2(x_{2,j}|\hat{\theta}_{1,\eta}, \bar{\theta}_{2,\eta}), \tag{4.37}$$

which contradicts (A .5). Thus we have

$$\sum_{j=1}^{n_2} \log f_2(x_{2,j}|\hat{\theta}_{1,\eta}, \hat{\theta}_{2,\eta}) = \sum_{j=1}^{n_2} \log f_2(x_{2,j}|\hat{\theta}_{1,\eta}, \bar{\theta}_{2,\eta}),$$

and hence by the identifiability assumption $\hat{\theta}_{2,\eta} = \bar{\theta}_{2,\eta}$.

Therefore, step (4.36) above is redundant, and we can instead directly set

$$(\hat{\theta}_{1,\eta}, \hat{\theta}_{2,\eta}) = \arg\max_{\theta_1,\theta_2} \left\{ \sum_{j=1}^{n_1} \log f_1(x_{1,j}|\theta_1) + \eta \sum_{j=1}^{n_2} \log f_2(x_{2,j}|\theta_1, \theta_2) \right\}. \tag{4.38}$$

The above calculations suggest a method for performing Posterior Bootstrap for semi-modular inference, which we present as Algorithm 14.

Algorithm 14 Posterior Bootstrap for Semi-modular Inference for Scenario 1

1: **for** k = 1, ..., N **do** ▷ (in parallel)
2: Draw $w_1^{(k)}, \ldots, w_{n_1}^{(k)}, v_1^{(k)}, \ldots, v_{n_2}^{(k)} \sim \text{Exp}(1)$.
3: Set $(\tilde{\theta}_{1,\eta}^{(k)}, \tilde{\theta}_{2,\eta}^{(k)}) = \arg\max_{\theta_1,\theta_2} \{ \sum_{j=1}^{n_1} w_j^{(k)} \log f_1(x_{1,j}|\theta_1) + \eta \sum_{j=1}^{n_2} v_j^{(k)} \log f_2(x_{2,j}|\theta_1, \theta_2) + w_0^T \log \pi_1(\theta_1) + v_0^T \log \pi_2(\theta_2) \}$.
4: **return** Sample $\left\{ \tilde{\theta}_{1,\eta}^{(k)}, \tilde{\theta}_{2,\eta}^{(k)} \right\}_{k=1}^{N}$.

We will show that Algorithm 14 provides the correct frequentist coverage with respect to the procedure in equation (4.38).

Firstly, we will show that

$$n_2^{1/2}(\hat{\theta}_\eta - \theta_\eta^*) \to N(0, J_\eta^{*-1} I_\eta^* J_\eta^{*-1}) \tag{4.39}$$

in distribution. Next, we will show that the asymptotic distribution of $n_2^{1/2}(\tilde{\theta}_\eta - \hat{\theta}_\eta)$ given $x_{1,1:n_1}, x_{2,1:n_2}$ is also $N(0, J_\eta^{*-1} I_\eta^* J_\eta^{*-1})$.

To obtain the first result, let $L(\theta) = \sum_{j=1}^{n_1} \log f_1(x_{1,j}|\theta_1) + \eta \sum_{j=1}^{n_2} \log f_2(x_{2,j}|\theta_1, \theta_2)$. Then since

$$\hat{\theta}_\eta = (\hat{\theta}_{1,\eta}, \hat{\theta}_{2,\eta}) = \arg\max_{\theta_1, \theta_2} L(\theta_1, \theta_2),$$

we have $\nabla_\theta L(\hat{\theta}_\eta) = 0$. Using the Taylor expansion we have

$$0 = \nabla_\theta L(\hat{\theta}_\eta) = \nabla_\theta L(\theta_\eta^*) + D_\theta^2 L(\theta_\eta^*)(\hat{\theta}_\eta - \theta_\eta^*) + O_P(n_2 \|\hat{\theta}_\eta - \theta_\eta^*\|^2),$$

and therefore

$$0 = n_2^{-1/2} \nabla_\theta L(\theta_\eta^*) + \frac{D_\theta^2 L(\theta_\eta^*)}{n_2} n_2^{1/2}(\hat{\theta}_\eta - \theta_\eta^*) + O_P(n_2^{1/2} \|\hat{\theta}_\eta - \theta_\eta^*\|^2),$$

which is equivalent to

$$-\frac{D_\theta^2 L(\theta_\eta^*)}{n_2} n_2^{1/2}(\hat{\theta}_\eta - \theta_\eta^*) = n_2^{-1/2} \nabla_\theta L(\theta_\eta^*) + O_P(n_2^{1/2} \|\hat{\theta}_\eta - \theta_\eta^*\|^2).$$

Note that $-n_2^{-1} D_\theta^2 L(\theta_\eta^*)$ converges in probability to ηJ_η^* and that $n_2^{1/2} \|\hat{\theta}_\eta - \theta_\eta^*\|^2$ converges in probability to 0 as $n_2 \to \infty$. What is more, since $E_{p_1^*, p_2^*} f_\eta(X_1, X_2|\theta_\eta^*) = 0$ and $n_1/n_2 \to \alpha$, then $n_2^{-1/2} \nabla_\theta L(\theta_\eta^*)$ converges to $N(0, \eta^2 I_\eta^*)$ in distribution. Finally, using Slutsky's theorem we obtain (4.39).

Let us now find the asymptotic distribution of $n_2^{1/2}(\tilde{\theta}_\eta - \hat{\theta}_\eta)$ given $x_{1,1:n_1}, x_{2,1:n_2}$. To this end, similarly we define

$$\tilde{L}(\theta) = \sum_{j=1}^{n_1} w_j \log f_1(x_{1,j}|\theta_1) + \eta \sum_{j=1}^{n_2} v_j \log f_2(x_{2,j}|\theta_1, \theta_2) + w_0^T \log \pi_1(\theta_1) + v_0^T \log \pi_2(\theta_2).$$

We thus have using the Taylor expansion

$$0 = \nabla_\theta \tilde{L}(\tilde{\theta}_\eta) = \nabla_\theta \tilde{L}(\hat{\theta}_\eta) + D_\theta^2 \tilde{L}(\hat{\theta}_\eta)(\tilde{\theta}_\eta - \hat{\theta}_\eta) + O_P(n_2 \|\tilde{\theta}_\eta - \hat{\theta}_\eta\|^2),$$

and similarly to the previous proof we have

$$-\frac{D_\theta^2 \tilde{L}(\hat{\theta}_\eta)}{n_2} n_2^{1/2}(\tilde{\theta}_\eta - \hat{\theta}_\eta) = n_2^{-1/2}(\nabla_\theta \tilde{L}(\hat{\theta}_\eta) + O_P(n_2^{1/2}\|\tilde{\theta}_\eta - \hat{\theta}_\eta\|^2). \tag{4.40}$$

Analogously to the previous case, we have $-D_\theta^2 \tilde{L}(\hat{\theta}_\eta)/n_2$ converge to ηJ_η^* in probability and $n_2^{1/2}\|\tilde{\theta}_\eta - \hat{\theta}_\eta\|^2$ converge to 0 in probability.

Recall $L(\theta)$ defined in the previous case. We have $\nabla_\theta L(\hat{\theta}_\eta) = 0$, and therefore

$$\begin{aligned}\nabla_\theta \tilde{L}(\hat{\theta}_\eta) &= \nabla_\theta \tilde{L}(\hat{\theta}_\eta) - \nabla_\theta L(\hat{\theta}_\eta) \\ &= \sum_{j=1}^{n_1}(w_j - 1)\nabla_\theta \log f_1(x_{1,j}|\theta_1) + \eta \sum_{j=1}^{n_2}(v_j - 1)\nabla_\theta \log f_2(x_{2,j}|\theta_1, \theta_2) + o_P(1).\end{aligned}$$

Recall that $E(w_j - 1) = 0, E(w_j - 1)^2 = 1$ for $j = 1, \ldots, n_1$ and $E(v_j - 1) = 0, E(v_j - 1)^2 = 1$ for $j = 1, \ldots, n_2$, so the right hand side of (4.40) converges in distribution to $N(0, \eta^2 I_\eta^*)$. Consequently $n_2^{1/2}(\tilde{\theta}_\eta - \hat{\theta}_\eta)$ given the data converges in distribution to $N(0, J_\eta^{*-1} I_\eta^* J_\eta^{*-1})$. Comparing this result with (4.39), we conclude that Algorithm 14 gives us asymptotically the correct frequentist coverage.

Statement of Authorship for joint/multi-authored papers for PGR thesis

To appear at the end of each thesis chapter submitted as an article/paper

The statement shall describe the candidate's and co-authors' independent research contributions in the thesis publications. For each publication there should exist a complete statement that is to be filled out and signed by the candidate and supervisor **(only required where there isn't already a statement of contribution within the paper itself).**

Title of Paper	Asymptotics of cut distributions and robust modular inference using Posterior Bootstrap
Publication Status	Unpublished and unsubmitted work written in a manuscript style
Publication Details	Pompe E. and Jacob P. E., Asymptotics of cut distributions and robust modular inference using Posterior Bootstrap. arXiv preprint arXiv:2110.11149

Student Confirmation

Student Name:	Emilia Pompe		
Contribution to the Paper	• initiated the project • designed algorithms presented in the paper • proved theoretical results (with feedback on these results) • provided the literature review • coded up the numerical examples included in the paper • wrote up the manuscript (with detailed feedback and edits from Pierre Jacob)		
Signature	E Pompe	Date	22/10/2021

Supervisor Confirmation

By signing the Statement of Authorship, you are certifying that the candidate made a substantial contribution to the publication, and that the description described above is accurate.

Supervisor name and title:	Judith Rousseau, Professor		
Supervisor comments	I am satisfied by the listed authors		
Signature	JRB	Date	21/10/2021

This completed form should be included in the thesis, at the end of the relevant chapter.

Chapter 5

Conclusions and discussion

In this chapter we summarize the main contributions of this thesis, and discuss potential future work directions.

In Chapter 2 we introduced the Jumping Adaptive Multimodal Sampler, which is a framework for targeting multimodal distributions, using an adaptive Markov chain Monte Carlo method. There are two main methodological contributions in this chapter.

Firstly, we constructed an adaptive MCMC algorithm, which, given the locations of the local maxima of the target π, can move efficiently between the modes. The algorithm operates on an extended state space $\mathcal{X} \times \mathcal{I}$, where $\mathcal{X}$ is the original space of the target of interest, and $\mathcal{I}$ is a finite set of modes. The adaptive kernel $\tilde{P}_\gamma$, indexed by a design parameter of the algorithm $\gamma \in \mathcal{Y}$, is invariant with respect to the target distribution $\tilde{\pi}(x, i)$, where $\sum_{i \in \mathcal{I}} \tilde{\pi}(x, i) = \pi(x)$ for any $\gamma \in \mathcal{Y}$. The algorithm relies on two types of moves, local ones, for within-mode mixing, and jumps between modes. Both types of moves utilize the knowledge of the local geometries of regions associated with different modes. Adaptation of the parameters (typically the covariance matrices) is performed separately for each of these regions, and can follow standard strategies known from the literature on optimizing MCMC algorithms.

This is possible thanks to the auxiliary variable approach, which allows the sampler to assign each iteration of the chain to a mode, and consequently to learn the parameters based on previous samples, separately for each mode.

Another methodological contribution is designing the burn-in algorithm, which is crucial for the good performance of our method in practice. Having relatively good estimates of the parameters at the beginning ensures that jumps to any particular mode are not rejected too

often, as otherwise the sampler is unable to start learning the geometry of the associated mode. While the burn-in algorithm imposes an additional cost on the entire procedure, its advantage is that it can be parallelized over modes of π, with only a few exchanges of information between cores.

The construction of the algorithm is particularly well-suited for target distributions with several well-separated modes. At the same time, JAMS is of limited usefulness for target distributions with a strongly jagged surfaces, rather than several distinct regions associated with local maxima. The main bottleneck, however, is the mode-finding phase of the method, and in particular sampling the starting points for optimization runs. Ideally, we would like to find all modes of the target, but it is also beneficial to avoid wasting the computational budget on rediscovering the same mode many times. This is an interesting area of future work, and the ongoing research on optimization can be helpful for that purpose. For example, the Basin Hopping algorithm, or Basin Hopping with Skipping (Wales and Doye, 1997; Goodridge et al., 2021), despite being originally developed to solve the issue of global optimization, creates a path of local maxima and encourages exploration of distant basins of attraction.

We compared our JAMS algorithm against adaptive parallel tempering on several examples. One of its advantages over APT is that JAMS has a mechanism of remembering the locations of the modes. Therefore, having narrow and wide modes in the same target distributions is not problematic, as long as the user knows their locations.

Our work on JAMS motivated also some theoretical developments. In Chapter 2 we introduced a new class of algorithms, Auxiliary Variable Adaptive MCMC, which generalizes Adaptive MCMC. We showed that it can be analysed in a very similar way to standard adaptive MCMC using minorization and drift conditions. Since JAMS belongs to Auxiliary Variable Adaptive MCMC, these results allowed us to prove its theoretical validity. There are other algorithms in this class, which makes our results useful more generally, for analysing theoretical properties of other adaptive methods.

The main methodological contribution of Chapter 3 is two-fold. Firstly, we developed a principled method of incorporating prior information into Weighted Likelihood Bootstrap, based on pseudosamples or on penalization. Furthermore, we extended these two strategies of incorporating the prior to hierarchical methods, considering scenarios when the number of classes is fixed, or when it goes to infinity. The theoretical analysis carried out in Chapter 3 allowed us to measure appropriately the impact of the prior, using Edgeworth expansions,

and guided the choice of hyperparameters. Moreover, our theoretical results are useful for understanding the properties of Weighted Likelihood Bootstrap and its extensions. In particular, we showed that in the one-dimensional case the distribution of the samples based on Weighted Likelihood Bootstrap is second order correct to the Bayesian posterior with the Jeffreys prior.

We also explained in which situations the user should rely on the penalization method to incorporate the prior information, and when they should resort to the method based on pseudosamples. In particular, the method based on prior penalization cannot be used for priors that are infinite on the boundary of the set, which is the case, for example, for Dirichlet$(\alpha_1, \ldots, \alpha_k)$ if any $\alpha_j < 1$, $j = 1, \ldots, k$. This was illustrated on a real-data application, for the opioid crisis study. As for methods based on pseudosamples, when the prior is non-conjugate, finding an appropriate value of the scaling parameter c is often an issue. This is particularly problematic for vague priors, because pseudosamples from the prior predictive are typically outliers with respect to the original dataset, and may severely distort the analysis. In these situations we advise using the penalization-based version of Posterior Bootstrap.

It is important to emphasize that all algorithms developed in Chapter 3 can be performed fully in parallel and require almost no tuning, which makes them very efficient computationally and convenient from the point of view of implementation.

Chapter 3 also helps to understand the strengths and weaknesses of different robust inference methods: generalized posteriors, BayesBag and Weighted Likelihood Bootstrap with its extensions. Out of these three approaches, only the last one has the property that credible set are valid asymptotic confidence sets in the frequentist sense. When the user is interested in inference and in particular, in having a reliable uncertainty estimation, Weighted Likelihood Bootstrap outperforms the other methods.

When it comes to prediction, asymptotically Weighted Likelihood Bootstrap performs better than Bayesian inference in misspecified settings. However, we show in Chapter 3 that Weighted Likelihood Bootstrap is outperformed by generalized posterior methods and by BayesBag (for certain choices of hyperparametrs of these algorithms), when all eigenvalues of $J(\theta^*)^{-1/2} I(\theta^*) J(\theta^*)^{-1/2}$ are greater than 1, or when all these eigenvalues are smaller than 1.

Chapter 4 largely builds on work from Chapter 3, focusing on Posterior Bootstrap for

modular Bayesian inference. We study there a model built of several modules and consider a scenario when, due to misspecification, the analyst wants to cut feedback flowing from one module to the other. Firstly, we show that under the Bayesian approach to modular inference, credible sets may not be valid confidence sets in the frequentist sense. The difference in the asymptotic distribution between the frequentist and the Bayesian has two reasons. Firstly, like in one-stage inference, it may result from misspecification of any involved modules. Secondly, it may appear when the modules rely on the same dataset, as cut posteriors fail to take into account the correlation between the involved likelihoods. The latter setting corresponds, for example, to causal inference with propensity scores in a situation of unobserved confounders.

In Chapter 4 we propose new methodology based on Posterior Bootstrap, which overcomes the problem described above. The algorithms we propose can be performed fully in parallel and can be easily implemented using the readily-available optimization procedures. This gives them a significant computational advantage over existing approaches to Bayesian modular inference, which are known to be costly and inconvenient from the point of view of implementation and diagnostics (Plummer, 2015; Jacob et al., 2017).

Methods based on Weighted Likelihood Bootstrap have also their limitations. Firstly, they require finding the maximum of the weighted log likelihood function, which is typically done using gradient-based methods. This may in practice limit algorithms based on WLB to continuous state spaces. Moreover, there is no known way of adapting Posterior Bootstrap to intractable likelihoods (Møller et al., 2006; Murray et al., 2012). Finally, theoretical properties of Weighted Likelihood Bootstrap are only well understood for large sample sizes. Understanding finite-sample properties of WLB is an active area of research (Walker, 2021).

Bibliography

Ahmad, F., Rossen, L., and Sutton, P. (2021). Provisional drug overdose death counts. `https://healthdata.gov/dataset/vsrr-provisional-drug-overdose-death-counts`. Data retrieved from National Center for Health Statistics on 25th February 2021.

Ahn, S., Chen, Y., and Welling, M. (2013). Distributed and adaptive darting Monte Carlo through regenerations. In *Artificial Intelligence and Statistics*, pages 108–116.

Alquier, P. (2020). Approximate Bayesian Inference. *Entropy*, 22(11).

Andricioaei, I., Straub, J. E., and Voter, A. F. (2001). Smart darting Monte Carlo. *The Journal of Chemical Physics*, 114(16):6994–7000.

Andrieu, C. and Moulines, É. (2006). On the ergodicity properties of some adaptive MCMC algorithms. *The Annals of Applied Probability*, 16(3):1462–1505.

Andrieu, C. and Roberts, G. O. (2009). The pseudo-marginal approach for efficient Monte Carlo computations. *The Annals of Statistics*, 37(2):697–725.

Andrieu, C. and Thoms, J. (2008). A tutorial on adaptive MCMC. *Statistics and computing*, 18(4):343–373.

Andrieu, C. and Vihola, M. (2015). Convergence properties of pseudo-marginal Markov chain Monte Carlo algorithms. *The Annals of Applied Probability*, 25(2):1030–1077.

Antonelli, J., Papadogeorgou, G., and Dominici, F. (2020). Causal Inference in high dimensions: A marriage between Bayesian modeling and good frequentist properties. *Biometrics*.

Atchadé, Y. and Fort, G. (2010). Limit theorems for some adaptive MCMC algorithms with subgeometric kernels. *Bernoulli*, 16(1):116–154.

Atchadé, Y. F., Roberts, G. O., and Rosenthal, J. S. (2011). Towards optimal scaling of Metropolis-coupled Markov chain Monte Carlo. *Statistics and Computing*, 21(4):555–568.

Atchadé, Y. F. and Rosenthal, J. S. (2005). On adaptive Markov chain Monte Carlo algorithms. *Bernoulli*, 11(5):815–828.

Bai, Y., Craiu, R. V., and Di Narzo, A. F. (2011a). Divide and conquer: a mixture-based approach to regional adaptation for MCMC. *Journal of Computational and Graphical Statistics*, 20(1):63–79.

Bai, Y., Roberts, G. O., and Rosenthal, J. S. (2011b). On the containment condition for adaptive Markov chain Monte Carlo algorithms. *Advances and Applications in Statistics*, 21(1):1–54.

Barber, D., Cemgil, A. T., and Chiappa, S. (2011). *Bayesian time series models*. Cambridge University Press.

Barrett, M. T., Galipeau, P. C., Sanchez, C. A., Emond, M. J., and Reid, B. J. (1996). Determination of the frequency of loss of heterozygosity in esophageal adenocarcinoma by cell sorting, whole genome amplification and microsatellite polymorphisms. *Oncogene*, 12(9):1873–1878.

Bartlett, M. S. (1953a). Approximate confidence intervals. *Biometrika*, 40(1/2):12–19.

Bartlett, M. S. (1953b). Approximate confidence intervals. II. More than one unknown parameter. *Biometrika*, 40(3/4):306–317.

Bennett, J. and Wakefield, J. (2001). Errors-in-variables in joint population pharmacokinetic/pharmacodynamic modeling. *Biometrics*, 57(3):803–812.

Benson, A. and Friel, N. (2021). Bayesian inference, model selection and likelihood estimation using fast rejection sampling: the Conway-Maxwell-Poisson distribution. *Bayesian Analysis*.

Berk, R. H. (1966). Limiting behavior of posterior distributions when the model is incorrect. *The Annals of Mathematical Statistics*, pages 51–58.

Berk, R. H. (1970). Consistency a posteriori. *The Annals of Mathematical Statistics*, pages 894–906.

Bernardo, J. M. and Smith, A. F. (2009). *Bayesian theory*, volume 405. John Wiley & Sons.

Beskos, A., Pillai, N., Roberts, G., Sanz-Serna, J.-M., and Stuart, A. (2013). Optimal tuning of the hybrid Monte Carlo algorithm. *Bernoulli*, 19(5A):1501–1534.

Bhattacharya, R. N. and Ghosh, J. K. (1978). On the validity of the formal Edgeworth expansion. *Ann. Statist*, 6(2):434–451.

Bissiri, P. G., Holmes, C. C., and Walker, S. G. (2016). A general framework for updating belief distributions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5):1103–1130.

Blangiardo, M., Hansell, A., and Richardson, S. (2011). A Bayesian model of time activity data to investigate health effect of air pollution in time series studies. *Atmospheric Environment*, 45(2):379–386.

Bonvini, M. and Kennedy, E. H. (2021). Sensitivity analysis via the proportion of unmeasured confounding. *Journal of the American Statistical Association*, pages 1–11.

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.

Brockwell, A. E. and Kadane, J. B. (2005). Identification of regeneration times in MCMC simulation, with application to adaptive schemes. *Journal of Computational and Graphical Statistics*, 14(2):436–458.

Bühlmann, P. (2014). Discussion of Big Bayes Stories and BayesBag. *Statistical science*, 29(1):91–94.

Bunke, O. and Milhaud, X. (1998). Asymptotic behavior of Bayes estimates under possibly incorrect models. *The Annals of Statistics*, 26(2):617–644.

Carmona, C. and Nicholls, G. (2020). Semi-Modular Inference: enhanced learning in multi-modular models by tempering the influence of components. In *International Conference on Artificial Intelligence and Statistics*, pages 4226–4235. PMLR.

Chanialidis, C., Evers, L., Neocleous, T., and Nobile, A. (2018). Efficient Bayesian inference for COM-Poisson regression models. *Statistics and computing*, 28(3):595–608.

Chernozhukov, V. and Hong, H. (2003). An MCMC approach to classical estimation. *Journal of Econometrics*, 115(2):293–346.

Chimisov, C., Łatuszyński, K., and Roberts, G. (2018). Air Markov Chain Monte Carlo. *arXiv preprint arXiv:1801.09309*.

Conway, R. W. and Maxwell, W. L. (1962). A queuing model with state dependent service rates. *Journal of Industrial Engineering*, 12(2):132–136.

Craiu, R., Rosenthal, J., and Yang, C. (2009). Learn from thy neighbor: parallel-chain and regional adaptive MCMC. *Journal of the American Statistical Association*, 104(488):1454–1466.

Das, D., Gregory, K., and Lahiri, S. (2019). Perturbation bootstrap in adaptive lasso. *The Annals of Statistics*, 47(4):2080–2116.

Dawid, A. P. (1984). Present position and potential developments: Some personal views statistical theory the prequential approach. *Journal of the Royal Statistical Society: Series A (General)*, 147(2):278–290.

De Blasi, P. and Walker, S. G. (2013). Bayesian asymptotics with misspecified models. *Statistica Sinica*, pages 169–187.

Dehejia, R. H. and Wahba, S. (1999). Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American statistical Association*, 94(448):1053–1062.

Deligiannidis, G., Doucet, A., and Pitt, M. K. (2018). The correlated pseudomarginal method. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(5):839–870.

Dharamshi, A., Ngo, V., and Rosenthal, J. S. (2021). Sampling by Divergence Minimization. *arXiv preprint arXiv:2105.00520*.

Diaconis, P. and Ylvisaker, D. (1979). Conjugate priors for exponential families. *The Annals of Statistics*, 7(2):269–281.

Douady, C. J., Delsuc, F., Boucher, Y., Doolittle, W. F., and Douzery, E. J. (2003). Comparison of Bayesian and maximum likelihood bootstrap measures of phylogenetic reliability. *Molecular biology and evolution*, 20(2):248–254.

Doucet, A., Godsill, S., and Andrieu, C. (2000). On sequential Monte Carlo sampling methods for Bayesian filtering. *Statistics and computing*, 10(3):197–208.

Drton, M. and Richardson, T. S. (2004). Multimodality of the likelihood in the bivariate seemingly unrelated regressions model. *Biometrika*, 91(2):383–392.

Duane, S., Kennedy, A. D., Pendleton, B. J., and Roweth, D. (1987). Hybrid Monte Carlo. *Physics letters B*, 195(2):216–222.

Efron, B. (1979). Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, pages 1–26.

Feroz, F., Hobson, M., and Bridges, M. (2009). MultiNest: an efficient and robust Bayesian inference tool for cosmology and particle physics. *Monthly Notices of the Royal Astronomical Society*, 398(4):1601–1614.

Feroz, F., Hobson, M., Cameron, E., and Pettitt, A. (2013). Importance nested sampling and the MultiNest algorithm. *arXiv preprint arXiv:1306.2144*.

Finucane, M. M., Rowley, C. F., Paciorek, C. J., Essex, M., and Pagano, M. (2016). Estimating the prevalence of transmitted HIV drug resistance using pooled samples. *Statistical methods in medical research*, 25(2):917–935.

Fong, E., Lyddon, S., and Holmes, C. (2019). Scalable nonparametric sampling from multimodal posteriors with the posterior bootstrap. In *International Conference on Machine Learning*, pages 1952–1962. PMLR.

Fort, G., Moulines, E., and Priouret, P. (2011). Convergence of adaptive and interacting Markov chain Monte Carlo algorithms. *Annals of Statistics*, 39(6):3262–3289.

Fung, T., Alwan, A., Wishart, J., and Huang, A. (2020). Package 'mpcmp'.

Fushiki, T. (2005). Bootstrap prediction and Bayesian prediction under misspecified models. *Bernoulli*, 11(4):747–758.

Fushiki, T. (2010). Bayesian bootstrap prediction. *Journal of statistical planning and inference*, 140(1):65–74.

Gåsemyr, J. (2006). The spectrum of the independent Metropolis–Hastings algorithm. *Journal of Theoretical Probability*, 19(1):152–165.

Gelfand, A. E. and Smith, A. F. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American statistical association*, 85(410):398–409.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian data analysis*. CRC press.

Gelman, A., Gilks, W. R., and Roberts, G. O. (1997). Weak convergence and optimal scaling of random walk Metropolis algorithms. *The annals of applied probability*, 7(1):110–120.

Gelman, A. and Vehtari, A. (2021). What are the most important statistical ideas of the past 50 years? *Journal of the American Statistical Association*, (just-accepted):1–29.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence*, (6):721–741.

George, E. I. and McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889.

Germain, P., Bach, F., Lacoste, A., and Lacoste-Julien, S. (2016). PAC-Bayesian theory meets Bayesian inference. *arXiv preprint arXiv:1605.08636*.

Geyer, C. J. (1991). Markov chain Monte Carlo maximum likelihood.

Ghosh, J. K. (1994). Higher order asymptotics. In *NSF-CBMS Regional Conference Series in Probability and Statistics*, pages i–111. JSTOR.

Gilks, W. R., Roberts, G. O., and Sahu, S. K. (1998). Adaptive Markov chain Monte Carlo through regeneration. *Journal of the American statistical association*, 93(443):1045–1054.

Giordani, P. and Kohn, R. (2010). Adaptive independent Metropolis–Hastings by fast estimation of mixtures of normals. *Journal of Computational and Graphical Statistics*, 19(2):243–259.

Girolami, M. and Calderhead, B. (2011). Riemann manifold langevin and Hamiltonian Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(2):123–214.

Glynn, C., He, J., Polson, N. G., and Xu, J. (2019). Bayesian Inference for Polya Inverse Gamma Models. *arXiv preprint arXiv:1905.12141*.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 102(477):359–378.

Goodridge, M., Moriarty, J., Vogrinc, J., and Zocca, A. (2021). Hopping between distant basins. *arXiv preprint arXiv:2108.05229*.

Gordon, N. J., Salmond, D. J., and Smith, A. F. (1993). Novel approach to nonlinear/non-Gaussian Bayesian state estimation. In *IEE proceedings F (radar and signal processing)*, volume 140, pages 107–113. IET.

Goudie, R. J., Presanis, A. M., Lunn, D., De Angelis, D., and Wernisch, L. (2019). Joining and splitting models with Markov melding. *Bayesian analysis*, 14(1):81.

Green, P. J., Łatuszyński, K., Pereyra, M., and Robert, C. P. (2015). Bayesian computation: a summary of the current state, and samples backwards and forwards. *Statistics and Computing*, 25(4):835–862.

Grünwald, P. (2012). The Safe Bayesian. In *International Conference on Algorithmic Learning Theory*, pages 169–183. Springer.

Grünwald, P. and van Ommen, T. (2017). Inconsistency of Bayesian inference for misspecified linear models, and a proposal for repairing it. *Bayesian Analysis*, 12(4):1069–1103.

Guedj, B. (2019). A primer on PAC-Bayesian learning. *arXiv preprint arXiv:1901.05353*.

Haario, H., Saksman, E., and Tamminen, J. (1999). Adaptive proposal distribution for random walk Metropolis algorithm. *Computational Statistics*, 14(3):375–396.

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

Haario, H., Saksman, E., and Tamminen, J. (2005). Componentwise adaptation for high dimensional MCMC. *Computational Statistics*, 20(2):265–273.

Hairer, M., Stuart, A. M., and Vollmer, S. J. (2014). Spectral gaps for a Metropolis–Hastings algorithm in infinite dimensions. *The Annals of Applied Probability*, 24(6):2455–2490.

Hall, P. (2013). *The bootstrap and Edgeworth expansion*. Springer Science & Business Media.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications.

Hennig, C. and Imports, M. (2015). Package 'fpc'. *Available at: https://cran. r-project. org/web/packages/fpc/index. html ENT*, 91.

Hirano, K. and Imbens, G. W. (2004). The propensity score with continuous treatments. *Applied Bayesian modeling and causal inference from incomplete-data perspectives*, 226164:73–84.

Holmes, C. and Walker, S. (2017). Assigning a value to a power likelihood in a general Bayesian model. *Biometrika*, 104(2):497–503.

Hong, L. and Martin, R. (2020). Model misspecification, Bayesian versus credibility estimation, and Gibbs posteriors. *Scandinavian Actuarial Journal*, 2020(7):634–649.

Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 221–233. University of California Press.

Huggins, J. H. and Miller, J. W. (2019). Robust inference and model criticism using bagged posteriors. *arXiv preprint arXiv:1912.07104*.

Huggins, J. H. and Miller, J. W. (2020). Robust and reproducible model selection using bagged posteriors. *arXiv preprint arXiv:2007.14845*.

Ibrahim, J. G., Chen, M.-H., and Sinha, D. (2014). Bayesian Survival Analysis. *Wiley StatsRef: Statistics Reference Online*.

Ihler, A. T., Fisher, J. W., Moses, R. L., and Willsky, A. S. (2005). Nonparametric belief propagation for self-localization of sensor networks. *IEEE Journal on Selected Areas in Communications*, 23(4):809–819.

Imai, K. and Van Dyk, D. A. (2004). Causal inference with general treatment regimes: Generalizing the propensity score. *Journal of the American Statistical Association*, 99(467):854–866.

Jacob, P. E., Murray, L. M., Holmes, C. C., and Robert, C. P. (2017). Better together? Statistical learning in models made of modules. *arXiv preprint arXiv:1708.08719*.

Jacob, P. E., O'Leary, J., and Atchadé, Y. F. (2020). Unbiased Markov chain Monte Carlo methods with couplings. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(3):543–600.

Jarner, S. and Hansen, E. (2000). Geometric ergodicity of Metropolis algorithms. *Stochastic Processes and Their Applications*, 85(2):341–361.

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

Jasra, A., Paulin, D., and Thiery, A. H. (2015). Error Bounds for Sequential Monte Carlo Samplers for Multimodal Distributions. *arXiv preprint arXiv:1509.08775*.

Jewson, J., Smith, J. Q., and Holmes, C. (2018). Principles of Bayesian inference using general divergence criteria. *Entropy*, 20(6):442.

Jiang, W. and Tanner, M. A. (2008). Gibbs posterior for variable selection in high-dimensional classification and data mining. *The Annals of Statistics*, pages 2207–2231.

Jin, Z., Ying, Z., and Wei, L. (2001). A simple resampling method by perturbing the minimand. *Biometrika*, 88(2):381–390.

Kannan, R. and Li, G. (1996). Sampling according to the multivariate normal density. In *Proceedings of 37th Conference on Foundations of Computer Science*, pages 204–212. IEEE.

Kipnis, C. and Varadhan, S. S. (1986). Central limit theorem for additive functionals of reversible Markov processes and applications to simple exclusions. *Communications in Mathematical Physics*, 104(1):1–19.

Kleijn, B. J. K. and Van der Vaart, A. W. (2012). The Bernstein-von-Mises theorem under misspecification. *Electronic Journal of Statistics*, 6:354–381.

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

LaLonde, R. J. (1986). Evaluating the econometric evaluations of training programs with experimental data. *The American economic review*, pages 604–620.

Lan, S., Streets, J., and Shahbaba, B. (2014). Wormhole Hamiltonian Monte Carlo. In *AAAI*, pages 1953–1959.

Łatuszyński, K. and Rosenthal, J. S. (2014). The containment condition and AdapFail algorithms. *Journal of Applied Probability*, 51(4):1189–1195.

LeCam, L. (1953). On some asymptotic properties of maximum likelihood estimates and related Bayes estimates. *Univ. California Pub. Statist.*, 1:277–330.

Leonelli, M., Barons, M. J., and Smith, J. Q. (2018). A conditional independence framework for coherent modularized inference. *arXiv preprint arXiv:1807.10628*.

Lindsay, B. G. (1988). Composite likelihood methods. *Contemporary mathematics*, 80(1):221–239.

Little, R. J. (1992). Regression with missing X's: a review. *Journal of the American Statistical Association*, 87(420):1227–1237.

Liu, F., Bayarri, M., and Berger, J. (2009). Modularization in Bayesian analysis, with emphasis on analysis of computer models. *Bayesian Analysis*, 4(1):119–150.

Liu, J. S. (1996). Metropolized independent sampling with comparisons to rejection sampling and importance sampling. *Statistics and Computing*, 6(2):113–119.

Liu, Y. and Goudie, R. J. (2020). Stochastic Approximation Cut Algorithm for Inference in Modularized Bayesian Models. *arXiv preprint arXiv:2006.01584*.

Loaiza-Maya, R., Martin, G. M., and Frazier, D. T. (2021). Focused Bayesian Prediction. *Journal of Applied Econometrics*, 36(5):517–543.

Luengo, D. and Martino, L. (2013). Fully adaptive Gaussian mixture Metropolis-Hastings algorithm. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 6148–6152. IEEE.

Lunn, D., Best, N., Spiegelhalter, D., Graham, G., and Neuenschwander, B. (2009). Combining MCMC with 'sequential'PKPD modelling. *Journal of Pharmacokinetics and Pharmacodynamics*, 36(1):19–38.

Lyddon, S., Holmes, C., and Walker, S. (2019). General Bayesian updating and the loss-likelihood bootstrap. *Biometrika*, 106(2):465–478.

Lyddon, S., Walker, S., and Holmes, C. C. (2018). Nonparametric learning from Bayesian models with randomized objective functions. In *Advances in Neural Information Processing Systems*, pages 2071–2081.

Maire, F., Friel, N., Mira, A., and Raftery, A. E. (2019). Adaptive Incremental Mixture Markov chain Monte Carlo. *Journal of Computational and Graphical Statistics*, 28(4):790–805.

Manderson, A. A. and Goudie, R. J. (2020). A numerically stable algorithm for integrating Bayesian models using Markov melding. *arXiv preprint arXiv:2001.08038*.

Mangoubi, O., Pillai, N. S., and Smith, A. (2018). Does Hamiltonian Monte Carlo mix faster than a random walk on multimodal densities? *arXiv preprint arXiv:1808.03230*.

Marinari, E. and Parisi, G. (1992). Simulated tempering: a new Monte Carlo scheme. *EPL (Europhysics Letters)*, 19(6):451.

Maucort-Boulch, D., Franceschi, S., and Plummer, M. (2008). International correlation between human papillomavirus prevalence and cervical cancer incidence. *Cancer Epidemiology and Prevention Biomarkers*, 17(3):717–720.

McCandless, L. C., Douglas, I. J., Evans, S. J., and Smeeth, L. (2010). Cutting feedback in Bayesian regression adjustment for the propensity score. *The international journal of biostatistics*, 6(2).

McCandless, L. C., Richardson, S., and Best, N. (2012). Adjustment for missing confounders using external validation data and propensity scores. *Journal of the American Statistical Association*, 107(497):40–51.

McLeish, D. L. (1975). A maximal inequality and dependent strong laws. *The Annals of Probability*, 3(5):829–839.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6):1087–1092.

Metropolis, N. and Ulam, S. (1949). The Monte Carlo method. *Journal of the American statistical association*, 44(247):335–341.

Meyn, S. P. and Tweedie, R. L. (2012). *Markov chains and stochastic stability*. Springer Science & Business Media.

Miasojedow, B., Moulines, E., and Vihola, M. (2013). An adaptive parallel tempering algorithm. *Journal of Computational and Graphical Statistics*, 22(3):649–664.

Middleton, L., Deligiannidis, G., Doucet, A., and Jacob, P. E. (2020). Unbiased Markov chain Monte Carlo for intractable target distributions. *Electronic Journal of Statistics*, 14(2):2842–2891.

Miller, J. W. (2021). Asymptotic Normality, Concentration, and Coverage of Generalized Posteriors. *Journal of Machine Learning Research*, 22(168):1–53.

Miller, J. W. and Dunson, D. B. (2018). Robust Bayesian inference via coarsening. *Journal of the American Statistical Association*.

Minnier, J., Tian, L., and Cai, T. (2011). A perturbation method for inference on regularized regression estimates. *Journal of the American Statistical Association*, 106(496):1371–1382.

Mitchell, T. J. and Beauchamp, J. J. (1988). Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, 83(404):1023–1032.

Møller, J., Pettitt, A. N., Reeves, R., and Berthelsen, K. K. (2006). An efficient Markov chain Monte Carlo method for distributions with intractable normalising constants. *Biometrika*, 93(2):451–458.

Morita, S., Thall, P. F., and Müller, P. (2008). Determining the effective sample size of a parametric prior. *Biometrics*, 64(2):595–602.

Morita, S., Thall, P. F., and Müller, P. (2012). Prior effective sample size in conditionally independent hierarchical models. *Bayesian Analysis*, 7(3).

Müller, U. K. (2013). Risk of Bayesian inference in misspecified models, and the sandwich covariance matrix. *Econometrica*, 81(5):1805–1849.

Murphy, K. M. and Topel, R. H. (2002). Estimation and inference in two-step econometric models. *Journal of Business & Economic Statistics*, 20(1):88–97.

Murray, I., Ghahramani, Z., and MacKay, D. (2012). MCMC for doubly-intractable distributions. *arXiv preprint arXiv:1206.6848*.

Mykland, P., Tierney, L., and Yu, B. (1995). Regeneration in Markov chain samplers. *Journal of the American Statistical Association*, 90(429):233–241.

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

Neal, R. M. (2011). MCMC using Hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, 2(11).

Nemeth, C., Lindsten, F., Filippone, M., and Hensman, J. (2019). Pseudo-extended Markov chain Monte Carlo. In *Advances in Neural Information Processing Systems*, pages 4312–4322.

Neuenschwander, B., Weber, S., Schmidli, H., and O'Hagan, A. (2020). Predictively consistent prior effective sample sizes. *Biometrics*, 76(2):578–587.

Newton, M. A. (1991). *The weighted likelihood bootstrap and an algorithm for prepivoting*. PhD thesis, University of Washington.

Newton, M. A., Polson, N. G., and Xu, J. (2020). Weighted Bayesian bootstrap for scalable posterior distributions. *Canadian Journal of Statistics*.

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)*, 56(1):3–26.

Ng, T. L. and Newton, M. A. (2020). Random weighting to approximate posterior inference in LASSO regression. *arXiv preprint arXiv:2002.02629*.

Nie, L. and Ročková, V. (2020). Bayesian Bootstrap Spike-and-Slab LASSO. *arXiv preprint arXiv:2011.14279*.

Nocedal, J. and Wright, S. J. (2006). Numerical optimization: Springer series in operations research and financial engineering. *Springer, New York*.

Nummelin, E. (1978). A splitting technique for Harris recurrent Markov chains. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 43(4):309–318.

Parzen, M. and Lipsitz, S. R. (2007). Perturbing the minimand resampling with Gamma (1, 1) random variables as an extension of the Bayesian Bootstrap. *Statistics & probability letters*, 77(6):654–657.

Pauli, F., Racugno, W., and Ventura, L. (2011). Bayesian composite marginal likelihoods. *Statistica Sinica*, pages 149–164.

Petrone, S., Rousseau, J., and Scricciolo, C. (2014). Bayes and empirical Bayes: do they merge? *Biometrika*, 101(2):285–302.

Plummer, M. (2015). Cuts in Bayesian graphical models. *Statistics and Computing*, 25(1):37–43.

Pompe, E. (2021). Introducing prior information in Weighted Likelihood Bootstrap with applications to model misspecification. *arXiv preprint arXiv:2103.14445*.

Pompe, E., Holmes, C., and Łatuszyński, K. (2020). A framework for adaptive MCMC targeting multimodal distributions. *Annals of Statistics*, 48(5):2930–2952.

Pompe, E. and Jacob, P. E. (2021). Asymptotics of cut distributions and robust modular inference using Posterior Bootstrap. *arXiv preprint arXiv:2110.11149*.

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Raič, M. (2019). A multivariate Berry–Esseen theorem with explicit constants. *Bernoulli*, 25(4A):2824–2853.

Ribatet, M., Cooley, D., and Davison, A. C. (2012). Bayesian inference from composite likelihoods, with an application to spatial extremes. *Statistica Sinica*, pages 813–845.

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.

Roberts, G. and Rosenthal, J. (2001). Optimal scaling for various Metropolis–Hastings algorithms. *Statistical Science*, 16(4):351–367.

Roberts, G. and Rosenthal, J. (2004). General state space Markov chains and MCMC algorithms. *Probability Surveys*, 1:20–71.

Roberts, G. and Rosenthal, J. (2007). Coupling and ergodicity of adaptive Markov chain Monte Carlo algorithms. *Journal of Applied Probability*, 44(2):458.

Roberts, G. and Rosenthal, J. (2009). Examples of adaptive MCMC. *Journal of Computational and Graphical Statistics*, 18(2):349–367.

Roberts, G. O. and Rosenthal, J. S. (2013). A note on formal constructions of sequential conditional couplings. *Statistics & Probability Letters*, 83(9):2073–2076.

Roberts, G. O. and Tweedie, R. L. (1996). Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, 2(4):341–363.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.

Rosenthal, J. S. (1995). Minorization conditions and convergence rates for Markov chain Monte Carlo. *Journal of the American Statistical Association*, 90(430):558–566.

Rosenthal, J. S. (2011). Optimal proposal distributions and adaptive MCMC. *Handbook of Markov Chain Monte Carlo*, 4(10.1201).

Rousseau, J. and Mengersen, K. (2011). Asymptotic behaviour of the posterior distribution in overfitted mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(5):689–710.

Rubin, D. B. (2004). *Multiple imputation for nonresponse in surveys*, volume 81. John Wiley & Sons.

Sarrias, M. (2016). Discrete choice models with random parameters in R: The Rchoice Package. *Journal of Statistical Software*, 74(10):1–31.

Sellers, K. F. and Shmueli, G. (2010). A flexible regression model for count data. *The Annals of Applied Statistics*, pages 943–961.

Shen, M.-L., Keenlyside, N., Selten, F., Wiegerinck, W., and Duane, G. S. (2016). Dynamically combining climate models to "supermodel" the tropical Pacific. *Geophysical Research Letters*, 43(1):359–366.

Shimodaira, H. (2000). Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 90(2):227–244.

Shmueli, G., Minka, T. P., Kadane, J. B., Borle, S., and Boatwright, P. (2005). A useful distribution for fitting discrete data: revival of the Conway–Maxwell–Poisson distribution. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54(1):127–142.

Sminchisescu, C. and Welling, M. (2011). Generalized darting Monte Carlo. *Pattern Recognition*, 44(10):2738–2748.

Spiegelhalter, D., Thomas, A., Best, N., and Lunn, D. (2007). OpenBUGS user manual. *Version*, 3(2):2007.

Syed, S., Bouchard-Côté, A., Deligiannidis, G., and Doucet, A. (2019). Non-reversible parallel tempering: a scalable highly parallel MCMC scheme. *arXiv preprint arXiv:1905.02939*.

Syed, S., Romaniello, V., Campbell, T., and Bouchard-Côté, A. (2021). Parallel Tempering on Optimized Paths. *arXiv preprint arXiv:2102.07720*.

Syring, N. and Martin, R. (2019). Calibrating general posterior credible regions. *Biometrika*, 106(2):479–486.

Tak, H., Meng, X.-L., and van Dyk, D. A. (2017). A Repelling-Attracting Metropolis Algorithm for Multimodality. *Journal of Computational and Graphical Statistics*.

Tawn, N. G. and Roberts, G. O. (2019). Accelerating parallel tempering: Quantile tempering algorithm (QuanTA). *Advances in Applied Probability*, 51(3):802–834.

Tawn, N. G., Roberts, G. O., and Rosenthal, J. S. (2020). Weight-preserving simulated tempering. *Statistics and Computing*, 30(1):27–41.

Titsias, M. K. and Yau, C. (2017). The Hamming ball sampler. *Journal of the American Statistical Association*, 112(520):1598–1611.

Tjelmeland, H. and Hegstad, B. (2001). Mode jumping proposals in MCMC. *Scandinavian Journal of Statistics*, 28(1):205–223.

Van der Vaart, A. W. (2000). *Asymptotic statistics*, volume 3. Cambridge university press.

Varin, C., Reid, N., and Firth, D. (2011). An overview of composite likelihood methods. *Statistica Sinica*, pages 5–42.

Vehtari, A., Gelman, A., and Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and computing*, 27(5):1413–1432.

Vermeesch, P., Resentini, A., and Garzanti, E. (2016). An R package for statistical provenance analysis. *Sedimentary Geology*, 336:14–25.

Vershynin, R. (2018). *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge University Press.

Vihola, M. (2011). On the stability and ergodicity of adaptive scaling Metropolis algorithms. *Stochastic Processes and their Applications*, 121(12):2839–2860.

Waddell, P. J., Kishino, H., and Ota, R. (2002). Very fast algorithms for evaluating the stability of ML and Bayesian phylogenetic trees from sequence data. *Genome Informatics*, 13:82–92.

Wales, D. J. and Doye, J. P. (1997). Global optimization by basin-hopping and the lowest energy structures of Lennard-Jones clusters containing up to 110 atoms. *The Journal of Physical Chemistry A*, 101(28):5111–5116.

Walker, S. and Hjort, N. L. (2001). On Bayesian consistency. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(4):811–821.

Walker, S. G. (2013). Bayesian inference with misspecified models. *Journal of Statistical Planning and Inference*, 143(10):1621–1633.

Walker, S. G. (2021). On an Asymptotic Distribution for the MLE. *arXiv preprint arXiv:2106.06597*.

Wang, C., Blei, and M, D. (2018). A general method for robust Bayesian modeling. *Bayesian Analysis*, 13(4):1163–1191.

Wang, F. and Landau, D. (2001a). Determining the density of states for classical statistical models: A random walk algorithm to produce a flat histogram. *Physical Review E*, 64(5):056101.

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

Wang, Y., Kucukelbir, A., and Blei, D. M. (2017). Robust probabilistic modeling with Bayesian data reweighting. In *International Conference on Machine Learning*, pages 3646–3655. PMLR.

Warnes, Gregory R (2000). *The Normal Kernel Coupler: An adaptive Markov Chain Monte Carlo method for efficiently sampling from multi-modal distributions*. PhD thesis, University of Washington.

Woodard, D., Schmidler, S., and Huber, M. (2009a). Sufficient conditions for torpid mixing of parallel and simulated tempering. *Electronic Journal of Probability*, 14:780–804.

Woodard, D. B., Schmidler, S. C., and Huber, M. (2009b). Conditions for rapid mixing of parallel and simulated tempering on multimodal distributions. *The Annals of Applied Probability*, 19(2):617–640.

Yu, X., Nott, D. J., and Smith, M. S. (2021). Variational inference for cutting feedback in misspecified models. *arXiv preprint arXiv:2108.11066*.

Zanella, G. (2020). Informed proposals for local MCMC in discrete spaces. *Journal of the American Statistical Association*, 115(530):852–865.

Zellner, A. (1988). Optimal information processing and Bayes's theorem. *The American Statistician*, 42(4):278–280.

Zhang, L., Beal, S. L., and Sheiner, L. B. (2003a). Simultaneous vs. sequential analysis for population PK/PD data I: best-case performance. *Journal of pharmacokinetics and pharmacodynamics*, 30(6):387–404.

Zhang, L., Beal, S. L., and Sheiner, L. B. (2003b). Simultaneous vs. sequential analysis for population PK/PD data II: robustness of methods. *Journal of pharmacokinetics and pharmacodynamics*, 30(6):405–416.

Zhou, Q. (2011). Multi-Domain Sampling With Applications to Structural Inference of Bayesian Networks. *Journal of the American Statistical Association*, 106(496):1317–1330.

Zhou, X. and Reiter, J. P. (2010). A note on Bayesian inference after multiple imputation. *The American Statistician*, 64(2):159–163.

Zigler, C. M. (2016). The Central Role of Bayes' Theorem for Joint Estimation of Causal Effects and Propensity Scores. *The American Statistician*, 70(1):47–54.

Zigler, C. M., Watts, K., Yeh, R. W., Wang, Y., Coull, B. A., and Dominici, F. (2013). Model feedback in Bayesian propensity score estimation. *Biometrics*, 69(1):263–273.