Node-wise Pseudo-marginal Methods for Spatial Model Selection

Denishrouf Thesingarajah

A thesis submitted to the University of Warwick for the degree of
Doctor of Philosophy

Department of Statistics
Univeristy of Warwick
September 2021

Contents

List of Figures

List of Tables

List of Common Abbreviations

AIC Akaike's Information Criterion. 22

BIC Bayesian Information Criterion. 26

ESS effective sample size. 34

IS Importance Sampling. 32

MH Metropolis-Hastings. 38

MRF Markov Random field. 87

NWPM Node-wise Pseudo-marginal. 99

PET Positron Emission Tomography. 65

SMC Sequential Monte Carlo. 51

List of PET Symbols

$\boldsymbol{C}_{\mathrm{T}}$	Vector of target tissue concentrations for each compartment	$\mathrm{kBq} \cdot \mathrm{mL}^{-1}$
$\circledast$	Convolution operator	
$\phi_{1:m}$	Micro–parameter, function of the rate constants	
$\vartheta_{1:m}$	Micro–parameter, function of the rate constants	
C_{P}	Plasma time-activity function (input function)	$\mathrm{kBq} \cdot \mathrm{mL}^{-1}$
C_{T}	Target (total) tissue concentration	$\mathrm{kBq} \cdot \mathrm{mL}^{-1}$
H	Target tissue impulse response function, with respect to the plasma	$(\mathrm{mL}) \cdot \mathrm{s}^{-1} \cdot \mathrm{cm}^{-3}$
K_1	Plasma to brain tissue transport constant	$(\mathrm{mL}).\mathrm{s}^{-1}.\mathrm{cm}^{-3}$
k_2	Brain tissue to plasma transport constant	s^{-1}
k_3	First order association rate constant for specific binding	s^{-1}
k_4	Disassociation rate constant for specific binding	s^{-1}
k_5	Association rate constant for non-specific binding	s^{-1}
k_6	Disassociation rate constant for non-specific binding	s^{-1}
V_{B}	Fractional blood volume	
V_{D}	Total volume of distribution of the target tissue	$(\mathrm{mL}) \cdot \mathrm{cm}^{-3}$

Acknowledgements

First and foremost, I express my deepest gratitude to my supervisor Adam Johanson for his patience, encouragement and enthusiasm. I am greatly indebted to him for the opportunity to do this challenging but rewarding work.

Many more people have helped me, directly or indirectly. I would like to extend my sincere thanks to James, Stefan, Cida, Albert and Arham for making it easy to feel at home in Warwick. More broadly, I would like to thank members of the Statistics Department for contributing to a warm, inspiring academic environment — particularly the discussions in the Young Researchers' Meeting and other reading groups.

I extend my thanks to John Aston for providing the PET data set used in this study.

I am grateful to Pouviji for the love, support and unfaltering belief in me, especially over the last few years. Finally, I would like to thank my Appa, without whom I probably would not have been able to start this project; and my Amma, without whom I certainly would not have been able to finish it.

மாதா, பிதா, குரு, தெய்வம்

Declaration

I hereby declare that this thesis is the results of my own original work and research. It is submitted to the University of Warwick in support of the my application for the degree of Doctor of Philosophy. The thesis is the sole work of the author alone and has not been submitted for examination for any other degree. Parts of the thesis has been submitted for publication.

Denishrouf Thesingarajah
20/09/2021

To Kabilan.

Abstract

Motivated by problems from statistical analysis of neuroimaging data where current approaches make use of "mass univariate" analysis which neglects spatial structure entirely; A novel framework for incorporating spatial dependence within a large class of model-selection problems is introduced. Spatial dependence is encoded through a Markov random field model, enabling a variant of the pseudo-marginal Markov chain Monte Carlo algorithm to be developed. This method can then extended by a further augmentation of the underlying state space. The approach allows existing unbiased marginal likelihood estimator, used in settings in which spatial independence is assumed, to be readily exploited. This, therefore, allows the incorporation of spatial dependence using non-spatial estimates, with very minimal additional development effort.

Numerical investigation on measured PET image data show notable improvements in revealing underlying spatial structure, when compared to current methods that assume spatial independence. This novel, accessible algorithm can be realistically used for analysis of smaller subsets of large image data sets such as $2-$D slices of whole $3-$D dynamic PET brain images or other regions of interest. Principled approximations of the proposed method, together with the simple extensions based on the augmented spaces, are also investigated and shown to provide similar results to the full pseudo-marginal method. Such method variants allow the improved performance obtained by incorporating spatial dependence to be obtained at negligible additional cost.

Finally, software implementation of these proposed methods in the form of an R package is presented. This provides easy and direct access to these efficient novel algorithms for spatial model selection, without the requirement of proficiency in advanced programming knowledge.

Chapter 1

Introduction

"Come Watson, come! The game is afoot! Not a word! Into your clothes and come!"
— Sherlock Holmes, *The Abbey Grange, 1904.*

The size and complexity of data sets in neuroimaging, and other image-like spatial data, give rise to a number of issues in statistical analysis. Accounting for spatial dependence in such analyses expands this problem further. In particular, incorporating spatial dependence can be a difficult task largely due to the computational requirements. Oftentimes, there also exists numerous competing models, *a priori*, that can adequately represent the data. An important example of such data sets include PET (Positron Emission Tomography) images of the brain, where a whole image typically requires analysis of up to 10^6 time series (Hammers et al., 2007). Current state of the art analysis of such data generally either assumes full spatial independence of pixels, or perform large-scale aggregation over space to overcome such restrictions. The main objective of this thesis is to present, explore and evaluate a novel framework for the incorporation of spatial dependence in the process of model selection.

The proposed method can be stated, in brief, as follows: There exists many current models and associated computational methods, where spatial independence is assumed, that can be successfully used to analyse these data sets. Thus, a natural strategy is to use these existing approaches, building on them to achieve the task at hand. That is, carefully construct a larger spatial model, from these non-spatial models of the sub-units (pixels), to describe the whole data set. Where needed, reasonable assumptions are imposed. Next, robust computational methods can be built to enable reliable inference and model selection. Importantly, prudent exploitation of the structure of the problem, gives rise to efficient algorithms that overcome computational limitations.

This thesis, where we motivate, present and evaluate the above approach, can be roughly split into two halves: The first half focuses primarily on existing approaches and methods towards model selection. The second half presents the proposed methodology for incorporating spatial dependence in the process of model selection. More specifically, we begin by first describing the notion of a statistical model and how it can be used to represent observed data. Then, various methods used to compare and select between competing candidates of related models is explored. Specifically, methods using the frequentist approach are explored very briefly first. Then, the Bayesian approach is explored — here, it becomes clear that computational methods must be used to approximate quantities used compare models. A large portion of the first half will be an in depth exploration of

these Monte Carlo computational methods. Also included in the first half, is detailed exploration of PET image data sets. Specifically, we discuss the mechanics of the PET scanner, technical information about the data produced and the models used to describe the data.

In the second half, a generic hierarchical model that encodes spatial dependence using these existing models is first constructed. In particular, this model uses the Potts distribution to incorporate spatial relations at the model selection level. Then, computational methods that extend the Monte Carlo methods, reviewed in the first half, are presented. This framework is then empirically evaluated in simulation studies and then used to analyse measured PET data. Finally, software implementation of the proposed algorithms is presented.

For the remainder of this chapter, the motivation of this work is first discussed. Next, a brief outline of the chapters that form this thesis is described; Before concluding with a summary of some of the general conventions and notations used in this work.

1.1 Context

Rapid technological advancement has quickly proven to be both a boon and a problem for statistical analysts. Greater accessibility to computational resources means that contemporary data sets are significantly larger and more challenging to accurately model. Fortunately, the same computational accessibility gives rise to more opportunities to realistically address these problems through careful investigation. Doing so, bears greater insight into nature and allows for the advancement of scientific progress.

A highly pertinent example of such a setting is PET neuroimaging data. This imaging modality allows us to study the functionality of the human brain — which is, arguably, the most complex object in the known universe. Currently, PET images have been used successfully in both clinical and research settings. However, given the relative recency of the PET imaging modality (and the field of neuroscience, itself), there is considerable potential for further advancement. As we will see in the sequel, PET images rely greatly on good statistical models and methods for accurate analysis. Subsequently, improvements in these processes will give rise to better understanding and greater use of PET images, as well other similar data sets.

In particular, a Bayesian approach allows for the use of existing prior knowledge. Specifically, in the case for PET images, there exists considerable information on the tracers. Recent studies, detailed later, have shown that performance can be improved by using this approach when compared to existing methods. Additionally, as has be exploited by the proposed approach in this thesis, a prior distribution can be carefully utilised to enable the incorporation of spatial dependence.

There exists considerable literature on spatial modelling in Bayesian data analysis, in general. However, some specialised and newer settings, such as PET images, are yet to be explored to the same extent. In these cases, there may be significantly large amounts of data to be analysed and may require relatively sophisticated models for meaningful investigation. However, there have also been significant development and improvement of methods for Bayesian computations in recent times. For instance, adaptive variants of sequential Monte Carlo methods allow us perform Bayesian model comparison with minimal manual tuning. Such methods can be readily exploited for application in analysis that account for spatial relations.

These considerations highlight both the major hurdles of the task at hand, and the potential directions towards overcoming them. This thesis presents a framework that attempts to solve

these problems. The results is an accessible method for incorporate spatial dependence in these complex settings.

1.2 Outline

As aforementioned, the overall structure of this thesis can be generally split in two parts: specifically, Part I and Part II. In particular, introduction of detailed formal concepts, terms and notations relating to spatial dependence will be delayed until Part II; though, they will often times be discussed informally where pertinent, beforehand. There are two reasons for this: Firstly, the novel model and the associated computational method presented in this thesis are constructed using models that assume spatial independence. Thus, these non-spatial models and the relevant generic computational methods, which are the primary discussion of Part I, do not require notions of spatial dependence. Secondly, suppressing formal references to spatial relations in Part I, means that the presentation and discussion of the mathematical concepts is cleaner, more focused and concise. Consequently, introducing and exploring models and methods for spatial dependence in Part II is made intuitive, natural and very accessible; since the exposition can more straightforwardly build upon and extend from Part I.

In brief, the structure and content of each chapter of the thesis is as follows:

- Chapter 2 introduces the formal concept of statistical models, and reviews approaches and methods for model selection. Specifically, common computational methods for Bayesian model selection are studied in detail.
- Chapter 3 reviews the sequential Monte Carlo method, detailing its practical and theoretical properties in the context of Bayesian model comparison. In particular, the different traits of this sampler which motivates its use in the proposed algorithm is discussed in detail.
- Chapter 4 discusses PET image data and the compartmental models used to describe them. The image acquisition mechanism and process is discussed in detail, before introducing the models and reviewing existing statistical methods for analysis of PET images.
- Chapter 5 introduces the proposed framework for incorporating spatial dependence. The Potts distribution is first reviewed, allowing for the introduction of the hierarchical model that encodes spatial relations. Next, the associated method for characterising this proposed model, together with extensions for efficiency, is discussed.
- Chapter 6 is an empirical evaluation of the framework presented in the previous chapter. These numerical studies include comparison between the different algorithms in different settings, together with a investigation of the trade off between different tuning parameters.
- Chapter 7 presents the results of applying the methodology to measured PET data. Once more, the different algorithm variants are discussed and compared to some current, existing methods for PET analysis.
- Chapter 8 is a presentation of the software implementation of the novel algorithms, together with a brief review of other existing libraries

Work based on shorter versions of Chapter 5, 6 and 7 has been submitted for publication:

D. Thesingarajah and A. M. Johansen. The Node-wise Pseudo-marginal Method. *arXiv preprint arXiv:2109.08573*, 2021. URL http://arxiv.org/abs/2109.08573

For ease of exposition, a list of technical terms for PET is provided, with units, at the beginning of this thesis. A list of important and most commonly used acronyms is also provided together with the page reference to where the term is introduced in detail. All diagrams presented in this work belong to the author.

1.3 Notations

Terms and notations will be defined in context when first introduced throughout the thesis; However, for clarity, a summary of some generic notational conventions is provided below.

Given a set A, its cardinality is denoted $|A|$. The empty set is denoted $\emptyset$. The symbols $\mathbb{N}, \mathbb{Z}, \mathbb{R}$ denote the set of natural numbers, integers and real numbers, respectively.

P is used to denote probability measures, in general. Uppercase Roman letters will be used for random variables, and lowercase Roman letters for their realised values. More formally, given a probability space $(\Omega, \mathcal{F}, P)$, denote X to be a $\mathcal{F}$-measurable random variable. Suppose that X takes value in some general measurable space $(\mathcal{X}, \mathcal{E})$, allowing for $P \circ X^{-1}$ to be the (push-forward) measure on $(\mathcal{X}, \mathcal{E})$ corresponding to the law, or probability distribution, of X. The realisations of X are denoted $x \in \mathcal{X}$.

In this thesis, it is typically assumed that probability distributions will have a density with respect to some natural dominating measure, such as the Lebesgue measure or counting measure. The reference measure will be denoted $\mathrm{d}x$. Given a probability measure P, on the measurable space $(\mathcal{X}, \mathcal{E})$, for which a density p exists, the notations

$$X \sim P \text{ and } X \sim p$$

are used to mean that the random variable $X \in \mathcal{X}$ is distributed according to the distribution P or a distribution with density p, respectively.

With reference to distributions used within the Bayesian framework, in general p is used to denote the density of the prior distributions; Similarly, f for the likelihood and π for the posterior. Likewise, when describing Monte Carlo methods, μ is used to denote the target density and ν the proposal densities. For simplicity and accessibility, these density notations are often overloaded and the same symbols are used to refer to both the distribution and density – particularly where Greek letters are used.

For simplicity, it will be assumed that in most cases that the state space, or observational space, $\mathcal{X}$ of the random variable will be a topological space. Allowing for the associated $\sigma-$algebra $\mathcal{F}$ to be a collection of Borel sets denoted $\mathcal{B}(\mathcal{X})$. Further, $\mathcal{X}$ will often times be a product space. As such, bold-face (heavy) font $\boldsymbol{X} = \boldsymbol{x}$, for some $\boldsymbol{x} \in \mathcal{X}$, is used to denote random vectors taking value in this space. Allow $\boldsymbol{x}_{p:q}$ to denote the vector comprising of components $x_p, x_{p+1}, \ldots, x_q$.

Data will be denoted by $\boldsymbol{y}$.

For various Monte Carlo Estimators, $X^{(i)}$, for $i = 1, \ldots,$ will be used to denote random samples.

The letter $\mathbb{E}$ is used to mean expectation, and given a distribution or density π, $\mathbb{E}_\pi$ to mean expectation with respect to the distribution or density π. That is, given a measurable function $\varphi : \mathcal{X} \to \mathbb{R}$ we have

$$\mathbb{E}_\pi[\varphi(X)] = \int_{\mathcal{X}} \varphi \mathrm{d}\pi = \int_{\mathcal{X}} \varphi(x)\pi(x)\mathrm{d}x,$$

such that it is clear what distribution is intended. In cases where the integrated variable is ambiguous the notation $\mathrm{d}\pi(x)$ may be used.

Part I

Literature Review

Chapter 2

Bayesian Model Selection using Monte Carlo and Other Methods

The limits of my language mean the limits of my world.

— Ludwig Wittgenstein, *Tractatus Logigo-Philosphicus, 1922.*

A primary objective of the discipline of Statistics is to allow for a principled interpretation of natural, random phenomena using observed data. Typically, this involves the reductive step of imposing a probability model; Then, performing inference on this model, in an objective manner. It is often the case that a class of many related models could be reasonably believed to give rise to sufficient representation or interpretation of the data. Thus, the task of selecting the "best" model becomes more critical.

The power of mathematical modelling lies centrally in abstraction. This is done using simplifications which allow us to focus on what is deemed important; While, simultaneously conceptualising a generalisation that is malleable and generic enough for application in many settings. Of course, this reductive formalisation often draws criticism. However, the upheld *interpretation* perspective, in addition to the use of *probabilistic* modelling addresses such issues (Robert, 2007). Essentially, statistical models attempt to bring together the available information from the observation, while simultaneously accounting for uncertainty.

This chapter will be an exploration of the various intriguing, elegant and powerful frameworks and methods that statistical analysts use to address the above considerations. We begin by first introducing the formal concept of statistical models in Section 2.1. Then, we turn to the task of model selection. First, simple, efficient methods that use the frequentist framework are discussed in Section 2.2. Next, the theory of Bayesian model selection is detailed, in Section 2.3, before turning to a review of some of the most widely used Bayesian computational methods. In particular, Section 2.4 will be a detailed study of Monte Carlo methods in the view of Bayesian model selection. In said section, we will study their theoretical properties and discuss their performance in application. This allows us to introduce terms and concepts to describe a more sophisticated and robust method in the next chapter. Specifically, Chapter 3 will be dedicated to the sequential Monte Carlo method.

Finally, we will also briefly discuss Monte Carlo methods for simple graphical models in this chapter. Graphical models allow us to intuitively encode spatial information, as such the proposed model, constructed in Part II, will be based on graphical representation of spatial data.

Ultimately, we seek to apply the discussed concepts and methods in realistic, challenging settings. In this thesis the primary context of application is in PET images of the brain — specifically, we are interested in accounting for spatial dependence. As aforementioned, for the moment we will use generic terms and notations that do not allude to spatial dependence in the observed data. In some cases we simply refer to data at a single location (pixel). This is for ease of presentation and simplicity, allowing us to focus on the concepts introduced and discussed.

2.1 Statistical Model Selection

The purpose of statistical model selection, given data, is to identify from a collection of candidate models the "best" model. The notion of a model can be formally captured using the following objects: an observational space $\mathcal{Y}$, the Borel $\sigma-$algebra of this space $\mathcal{B}(\mathcal{Y})$ and a family of probability measures P_θ on $(\mathcal{Y}, \mathcal{B}(\mathcal{Y}))$, indexed by the parameter θ with parameter space Θ. In other words, a *statistical model* is the tuple $(\mathcal{Y}, \mathcal{B}(\mathcal{Y}), \{P_\theta : \theta \in \Theta\})$. When we specify a model for a given set of data, denoted $\boldsymbol{y} = (y_1, \ldots, y_k)$, we assume that it is a realisation of a $\mathcal{B}(\mathcal{Y})$-measurable random variable $\boldsymbol{Y} \in \mathcal{Y}$ with law or distribution corresponding to a probability measure in the set $\{P_\theta : \theta \in \Theta\}$. The parameter of this particular distribution is called the true parameter value and is denoted θ_0. In brief, statistical inference is the task of identify θ_0 given $\boldsymbol{y}$. Estimators of θ_0 will be denoted, in general, as $\widehat{\theta} : \mathcal{Y} \to \Theta$.

For simplicity, we hereafter assume and restrict our attention to distributions that are dominated by a reference measure, denoted $\mathrm{d}\boldsymbol{y}$. Then, denote the density, up to some normalising constant, of the distribution P_θ by f_θ. Note that, models can then be also characterised by their densities. Suppressing the Borel sets for brevity, models may be represented by

$$S \doteq (\mathcal{Y}, \{f_\theta\}_{\theta \in \Theta}).$$

In particular, we are interested in parametric models. That is, it shall be considered in what follows that $\Theta \subset \mathbb{R}^d$ and $d < \infty$. When it is believed that the data could realistically arise from many plausible models, *a priori*, we need to consider a collection of statistical models. To formalise this, denote $\mathcal{S} = \{S_m : m \in \mathcal{M}\}$, where $\mathcal{M}$ is some countable index set. Here, each model S_m is characterised by the density $f_m(\cdot; \theta)$ and the associated parameter space Θ_m:

$$S_m \doteq (\mathcal{Y}, \{f_m(\cdot; \theta) : \theta \in \Theta_m\}).$$

In this thesis, we term the members of the index set $m \in \mathcal{M}$ by *model orders*, and the set $\mathcal{M}$ the *model order space*. Informally, the model order, also called the *model indicator* (Robert, 2007, Section 7.1), can be thought of as some artificial but often relevant label of a model. We take the preference to use the term model orders here, as it implies some measure of complexity which increases with order. For example, the model order could indicate the degree of a polynomial of the formulation of some models. Another, more pertinent example is the number of compartments in compartmental models (Gunn et al., 2001), the class of models that will be used to model PET data in Section 4.2 below. That is, the higher the number of compartments, or the model order, the more complex the model may be.

Note that we have chosen, for the sake of brevity, to use the convention of f_m, rather than $f(\boldsymbol{y}; \theta, S_m)$. This formalisation should also prove to be mathematically cleaner later; For example, in Chapter 5, we refer to selecting a model at random and then reference to the corresponding density f_M.

Another central component of statistical models is the likelihood. For the parametric models considered here; the likelihood of the model S_m, may be defined

$$\mathcal{L}_m(\theta; \boldsymbol{y}) \doteq f_m(\boldsymbol{y}; \theta) \text{ for } \theta \in \Theta_m.$$

The emphasis here is that: since the data is given, and thus known, the likelihood is a function of $\theta \in \Theta_m$. Where it is clear from the context, we will simply refer to f_m as the likelihood.

Statistical model selection can then be seen as the principled process of selecting a single model, denoted by S^*, from the collection $\mathcal{S}$. In principle, though countable, the size of $\mathcal{S}$ may be infinite. This often leads to many difficulties, for example under the Bayesian framework, specifying a prior over an infinite set may be harder to do (Robert, 2007, Section 7.2). In practice, $\mathcal{S}$ is typically finite, but in some cases it can often be very large — For instance, a single data point (at one location, say) could be explained by many models. In the setting of interest in this thesis, the problem of large number of plausible models arises in a different manner. Essentially, if we follow the aforementioned strategy of selecting a model for a whole image data set based on models used at individual pixels, the number of possible model combinations can become large very quickly. We will detail this further in the sequel.

Statistical model selection strategies in general can be, very coarsely, thought of as defining an objective criteria of selection (typically a metric for some desired characteristics) and making a choice based on this. Formally, this is accomplished by defining a loss function and selecting the model that minimises this loss function. In other words, given data $\boldsymbol{y}$, and loss function $L : \mathcal{S} \to \mathbb{R}$ the selected model is

$$S^* = \arg\min_{S \in \mathcal{S}} L(S; \boldsymbol{y}).$$

Once a selection criteria has been established, model selection can be relatively straightforward — if the number of models to be considered is small. In such simpler cases a naive, exhaustive search strategy through the space $\mathcal{S}$ is a viable option. The computational overhead of calculating the values of the loss function for each model is typically reasonable enough to simply compare each model to the other and choose the best candidate. However, for many cases of interest, $\mathcal{S}$ tends to be very large. Subsequently, a good search strategy is often an important part of a model selection method. As alluded to before, this is something that must (and will) be addressed, by the class of algorithms that we are proposing in this work. Several approaches for statistical model selection exists, each with its own advantages and disadvantages. We discuss some common approaches next, beginning with some classical methods, before describing the Bayesian model selection framework and methods.

2.2 Information-Theoretic Approaches

The information-theoretic approach to model selection is to identify the model which contains the probability distribution that is the closest to a hypothetical "true data generating" distribution. The loss function used here attempts to formalise the distance between the density of each candidate distribution and the true distribution.

Consider the Kullback–Leibler (KL) divergence (Kullback and Leibler, 1951), a popular measure of discrepancy between two probability distributions. Given two probability densities g_1 and g_2 on the same probability space $\mathcal{Y}$, the KL divergence is defined

$$\mathrm{KL}(g_1, g_2) \doteq \int g_1(\boldsymbol{y}) \log\left(\frac{g_1(\boldsymbol{y})}{g_2(\boldsymbol{y})}\right) \mathrm{d}\boldsymbol{y}.$$

Suppose that $Y \sim g$, for some, typically unknown, density g, called the hypothetical "true density". Consider performing model selection on Y using a set of models $S_m = (\mathcal{Y}, \{f_m(\cdot;\theta)\}_{\theta\in\Theta_m})$ for $m \in \mathcal{M}$. In this particular case, g need not even belong to one of the models in $\mathcal{S} = \{S_m\}_{m\in\mathcal{M}}$. Then, the KL divergence of g and the parametric model density f_m is

$$\begin{aligned}\mathrm{KL}(g, f_m) &= \int g(\boldsymbol{y}) \log\left(\frac{g(\boldsymbol{y})}{f_m(\boldsymbol{y};\theta)}\right) \mathrm{d}\boldsymbol{y} \\ &= \mathbb{E}_g[\log g(Y)] - \mathbb{E}_g[\log f_m(Y;\theta)], \end{aligned} \tag{2.1}$$

for $\theta \in \Theta_m$. Here, and henceforth, $\mathbb{E}_g$ is used to denote expectation with respect to the probability density g.

Importantly, the first term of (2.1) is a constant; the second term is called the relative Kullback-Leibler (rKL) divergence. Intuitively, this divergence can be thought of the "information lost" from making the assumption that the data is density according to f_m (as done under each model) rather then the true density g.

In order to use this quantity for model selection: Firstly, noting that each model consists of multiple distributions, the *negative* rKL must be *maximised* over the parameter space Θ_m within each model S_m; Then, the model with the smallest estimated KL, over the model space $\mathcal{S}$, may be chosen. This is discussed in detail in the sequel.

2.2.1 Akaike's Information Criterion

Among the many information-theoretic approaches, perhaps the most important and widely used is the Akaike's Information Criterion (AIC) introduced by Akaike (1973). The AIC is an example of a model selection methods based on minimising the rKL divergence. Denote by θ^* the parameter that maximises the rKL divergence over Θ_m for model S_m. In general, θ^* maybe be analytically intractable and so unknown; instead, we must estimate it.

Consider the following, allow $\boldsymbol{Y} = (Y_i)_{i=1}^k$ to be a random sample of size k from the postulated true probability density g. In other words, $\boldsymbol{Y}$ is a collection of k independent and identically distributed (i.i.d) realisations of $Y \sim g$. Here the model density f_m of model S_m of $\boldsymbol{Y}$ is

$$f_m(\boldsymbol{y};\theta) = \prod_{i=1}^{k} f_m(y_i;\theta), \text{ for } \theta \in \Theta_m.$$

Trivially, $g(\boldsymbol{y})$ maybe be defined in the same manner. The log-likelihood function, corresponding to model S_m, may then be written

$$\ell_m(\theta) \doteq \sum_{i=1}^{k} \log f_m(y_i;\theta), \text{ for } \theta \in \Theta_m.$$

Maximising ℓ_m over $\theta \in \Theta_m$ gives the maximum likelihood estimator (MLE) $\widehat{\theta}_m^{(\text{MLE})}$. For the sake of clarity, the superscript is suppressed for the remainder of this section. In summary, we use $\widehat{\theta}_m(\boldsymbol{y})$ to approximate θ^*.

Denote the fitted density, based on using the MLE as a measure of fit, by $\widehat{f}_m \doteq f_m(\cdot; \widehat{\theta}_m)$. Note that, $\widehat{f}_m$ could be considered an estimate of the true density g.

Similarly, another estimator of the hypothetical data generating distribution is the empirical distribution $\widehat{P}_Y \doteq k^{-1} \sum_{i=1}^k \delta_{Y_i}$ where δ_Y denotes the Dirac measure. This motivates an estimator for the rKL, formalised

$$\int \log \widehat{f}_m \mathrm{d}\widehat{P}_Y = \frac{1}{k} \sum_{i=1}^{k} \log f_m(Y_i; \widehat{\theta}_m) = \frac{\ell_m(\widehat{\theta}_m)}{k}. \tag{2.2}$$

This empirical average $k^{-1}\ell_m$ could be considered a reasonably good estimator of the rKL. Indeed, given a sample $Y_1, \ldots, Y_k \overset{\text{i.i.d.}}{\sim} g$, by the *Strong Law of Large Numbers* (SLLN), we have that

$$\frac{1}{k}\ell_m(\theta) \xrightarrow{a.s} \mathbb{E}_g[\log f_m(Y; \theta)] \text{ for } \theta \in \Theta_m.$$

However, Akaike (1973) showed that using data to estimate *both* $\widehat{f}_m$ and $\widehat{P}_Y$ leads to systematically upwards biased approximations. The bias is approximately $\dim(\Theta_m)/k$, where $\dim(\Theta_m)$ is the dimension of the parameter space Θ_m corresponding to model S_m. More specifically, Akaike (1973) approximates the bias term using the first order Taylor expansion of the discrepancy $\mathbb{E}_g(k^{-1}\ell(\widehat{\theta}) - \mathrm{KL}(g, \widehat{f}))$. This results in the adjusted estimate for the rKL, given by

$$-\frac{1}{k} l_m(\widehat{\theta}_m) + \frac{\dim(\Theta_m)}{k}.$$

Re-scaling then gives the AIC for a model S_m,

$$\mathrm{AIC}(S_m) \doteq -2 l_m(\widehat{\theta}_m) + 2\dim(\Theta_m).$$

Given data $\boldsymbol{y}$, the model selected by AIC can then be expressed as

$$S^*_{\mathrm{AIC}} \doteq \underset{S_m \in \mathcal{S}}{\arg\min}\, \mathrm{AIC}(S_m).$$

A more technical derivation of the AIC can be found in Claeskens and Hjort (2010, Section 2.3).

It is possible to generalise the AIC to many settings, to an extent; in principle, it can work for any situation where parametric models are used. Note that although i.i.d samples where assumed in showing the asymptotic results above, this is not required for applying the AIC. For example, the AIC is efficient for selecting the order of an auto-regressive process (Lee and Karagrigoriou, 2001). However, even simple departures from the i.i.d setting often involve further refinements to better reflect the intricacies of particular types of models. Furthermore, the notion of the dimension of the parameter space can become quite complicated and there are several definitions of "effective dimension" in this setting.

Another factor to consider is that when there are multiple models with minimum expected KL the model with more parameters is selected. This is formally explored by Sin and White (1996):

Briefly, the log-likelihood increases linearly with the sample size, while the penalty component $2\dim(\Theta_m)$ is not affected.

It is worth noting that in application and realistic settings the sample size may be small, as is the case for PET data. In such cases, a second order variant (derived in the same paper Akaike (1973)) called the corrected AIC is used.

2.2.2 Cross-validation and other Approaches

Suppose that rather than seeking the model closest to the true model, we instead want the model with the best predictive performance. To do so, ideally, we would require separate test sample – in many cases, such as PET, this is typically not practical. Alternatively, portions of the original data can be used to test the model and the remaining data for inference. This can be done a number of times, each time using different subsets of the data. This is known as cross-validation, which has been extensively studied in applied and theoretical statistics. Geisser (1975) and Stone (1976) give formal presentation of this approach.

Briefly, in a K-fold cross-validation, the data set is split into $K \in \mathbb{N}$ subsets of roughly equal size. Then, the $K-1$ sub-samples (called the training set) are used to infer the model parameters, before testing the fitted model on the remaining subset (called the validation set). This is process is repeated K times and the average validation error from the K folds is reported as the cross-validation error. After performing cross-validation for various models, we then select the one with the smallest cross-validation error.

More specifically the testing is done as follows: For each fitted model, a loss function is defined; The loss function is taken as a measurement of the fitness of the model. Commonly used loss function include the mean squared prediction error (e.g. for linear models) or the log-density function (Stone, 1977).

There are various different ways to split the data in cross-validation method, the most popular is the leave-one-out procedure. As the name suggests, training sets $\boldsymbol{y}/\{y_i\}$, for $i = 1, \ldots, k$, are used. Leave-one-out cross-validation has been shown by Shao (1993) to be inconsistent when the true data mechanism is included in the set of candidate models. The procedure was shown to select unnecessarily large models, even with large sample size. This deficiency applies to AIC, since it can be shown that AIC is asymptotically equivalent to the leave-one-out cross-validation method (Stone, 1977).

Next, we compare and contrast the different methods for model selection. There exists, many other refinements, extensions and variants of the information theoretic approach. One example, among many others, is the Kashyap information criterion (Kashyap, 1980), which uses the Fisher information matrix. Some examples within the Bayesian paradigm include: the Bayesian information criterion, discussed in Section 2.3.1, and the deviance information criterion (Spiegelhalter et al., 2002).

Recall that, in the AIC approach we did not assume that any of the candidate models are necessarily the true model — instead we attempted to find the model that is closest to true model. Thus, problems can arise when there is significant misspecification of the models, producing interpretations that by not be reliable. Takeuchi (1976) propose a general derivation of the KL distance, that is more model robust; An intermediate result of doing so is the Takeuchi information criterion.

The AIC is used ubiquitously in the analysis of PET data (Zhou et al., 2013), where it is possible

to compute point estimates of the parameters i.e. good estimators $\widehat{\theta}_m$ for each model exists. The aforementioned second order approximation, called the corrected AIC (Akaike, 1973), is used in this setting; since the data (time series) size is small, when compared to the number of parameters to be estimated. In brief, the parameters can be inferred using a non-linear least squares (NLS, see Zhou et al. (2013)); which, in turn are used to estimate the AIC in order to do model selection. However, there is difficulty in extending these approaches to allow for robust noise modelling and unknown model order. Specifically, Zhou et al. (2013), showed that using an Bayesian approach resulted in improvement in the average mean squared error in simulated data. More importantly, the use of Bayesian modelling resulted in outputs that revealed underlying spatial structure — despite assuming full spatial independence. This motivates the focus of this thesis on Bayesian approaches to model selection.

2.3 Bayesian Model Selection Framework

Under the Bayesian paradigm, it is assumed that one of the models is the true distribution generating the data; thus model selection is the process of identifying the model that is most probable to be true in the Bayesian sense. Sometimes this is referred to as the M-closed (rather than M-open) paradigm: there is a closed model universe, one of which is "true". Naturally, the Bayesian framework for model selection specifies a prior distribution on the unknown, in this case the statistical models themselves. In brief, Bayesian model selection is the extension of prior modelling from parameters to models (Robert, 2007, Section 7.1). This gives rise to a *hierarchical* Bayesian setup — that is, there are several layers of unknown quantities, where the highest level unknown quantity is which model describes the process by which data is generated.

To this end, consider a prior probability distribution on the space of models $\mathcal{S}$; denote the density of this prior $p(S_m)$ for $S_m \in \mathcal{S}$. Additionally, we refer to the prior distribution on the parameter space conditional upon the model by the density denoted $p(\theta|S_m)$, for $\theta \in \Theta_m$. It is important to note here that the condition is upon the random model order M; and, as before we denotes observations of the random variable M by $m \in \mathcal{M}$. Similarly, we have the model likelihood denoted $f(\cdot|\theta, S_m) = f_m(\cdot;\theta)$. Here, and henceforth, the symbol "|" rather than ";" is used, to emphasise that θ and m should be thought of as random. That is, we may associate a probability distribution over these quantities; In doing so, also encode any prior knowledge in a principled manner.

In what follows, assume probability distributions admit (Lebesgue) densities — bar the discrete prior over the models $p(S_m)$. Then, by Bayes' theorem, define the full posterior density

$$\pi(S_m, \theta|\boldsymbol{y}) \doteq \frac{f(\boldsymbol{y}|\theta, S_m)p(\theta|S_m)p(S_m)}{f(\boldsymbol{y})}; \tag{2.3}$$

where:

$$f(\boldsymbol{y}) \doteq \sum_{m\in\mathcal{M}} f(\boldsymbol{y}|S_m)p(S_m)$$

with

$$f(\boldsymbol{y}|S_m) \doteq \int_{\Theta_m} f(\boldsymbol{y}|\theta, S_m)p(\theta|S_m)\mathrm{d}\theta.$$

The density $f(\boldsymbol{y}|S_m)$ is called the *marginal likelihood* or the *model evidence*; and $f(\boldsymbol{y})$ can be considered the normalising constant, for given data $\boldsymbol{y}$. The marginal likelihood will play a central role in this thesis.

Since we are conditioning on the model order $M = m$, the reference to the symbol S, within the density, is somewhat redundant. Instead, *mutatis mutandis*, the above Eq.(2.3) can be more succinctly rewritten

$$\pi(S_m, \theta|\boldsymbol{y}) = \pi(m, \theta|\boldsymbol{y}) \doteq \frac{f(\boldsymbol{y}|\theta, m)p(\theta|m)p(m)}{f(\boldsymbol{y})}.$$

Likewise, by marginalising the parameter of the full posterior, we may define for model $S_m \in \mathcal{S}$, the distribution with density (with respect to counting measure)

$$\pi(m|\boldsymbol{y}) \propto \int_{\Theta_m} \pi(\theta, m|\boldsymbol{y})\mathrm{d}\theta = \frac{p(m)\int_{\Theta_m} f(\boldsymbol{y}|\theta, m)\pi(\theta|m)\mathrm{d}\theta}{\sum_{m'\in\mathcal{M}} p(m')\int_{\Theta_{m'}} f(\boldsymbol{y}|\theta, m')\pi(\theta|m')\mathrm{d}\theta},$$

called the *posterior model probability*. For clarity, depending on the context and where more appropriate, the alternative notation $\pi(S_m|\boldsymbol{y})$ will also be used to refer to the posterior model probability.

The posterior model probability, $\pi(m|\boldsymbol{y})$, could be thought to specify numerical summaries of the evidence in favour of model $S_m \in \mathcal{S}$. In other words, following the Bayesian decision theoretic framework of Robert (2007) and others, the Bayesian decision is the one which minimises the posterior expected loss. I.e. the expectation with respect to the posterior of some loss function. Given that the Bayesian framework involves simultaneously providing parameter estimation, model selection, model averaging and other inferences; it can be difficult to defined a principled criterion that chooses models best suited for all these purposes. As such, we focus here in the case where the true mode is of interest; See Robert (2007, Section 7.2.1), for further discussion of the Bayesian model choice problem. Thus, a (naive) Bayes decision rule, can be arrived under a 0-1 loss — incur a loss of 0 if one chooses the correct model and 1 if one chooses the wrong one. In this case, given data $\boldsymbol{y}$ the "best" model would be the *maximum a-posteriori* (MAP) model, defined

$$S^*_{\text{Bayes}} \doteq \arg\max_{S_m\in\mathcal{S}} \pi(S_m|\boldsymbol{y}).$$

In other words, the model S_m that maximises the expected posterior loss is the model with the highest posterior probability. It is often difficult to determine S^*_{Bayes} since computing $\pi(m|\boldsymbol{y})$ can be difficult. For instance, the marginal likelihood $f(\boldsymbol{y}|m)$, which we recall is an integral over the parameter space Θ_m, is typically analytically intractable. Consequently, most commonly used methods of Bayesian model selection aim to instead approximate $\pi(m|\boldsymbol{y})$ — and thus estimate S^*_{Bayes}.

2.3.1 Bayesian Information Criterion

In essence Bayesian Information Criterion (BIC), first introduced by Schwarz (1978), is a large sample approximation of the Bayesian MAP model. In other words, BIC aims to select the model with the maximum posterior model probability by approximating the posterior model probability $\pi(m|\boldsymbol{y})$ using Laplace's method. That is, using a second order Taylor series expansions and consequently arrive at a Gaussian approximation. We detail this below.

Let $\boldsymbol{Y} = (Y_i)_{i=1}^k$ be a random sample of size k from the true density (which is assumed to be contained in one of the models $S_m \in \mathcal{S}$, here). Firstly, note that the log of the posterior model

probability of $\boldsymbol{Y}$ can be written

$$\log \pi(m|\boldsymbol{y}) = \text{const} + \log \int_{\Theta_m} \exp\{\ell_m(\theta)\} p(\theta|m) \mathrm{d}\theta + \log p(m), \tag{2.4}$$

where we recall that $\ell_m \doteq \log f_m$ is the log-likelihood function associated to the model $S_m, m \in \mathcal{M}$.

Recall the MLE $\widehat{\theta}_m^{(\text{MLE})}(\boldsymbol{y})$ (henceforth suppressing superscript for clarity); denote the observed Fisher information matrix at the MLE $I_{\boldsymbol{Y}}(\widehat{\theta}_m) \doteq -\frac{1}{k}\nabla^2 l_m(\widehat{\theta}_m)$, and let $d_m = \dim(\Theta_m)$. Assuming the prior over the parameter space $p(\theta|m)$ to be smooth, the Taylor expansion of the marginal likelihood in (2.4) above, about the MLE, gives

$$\begin{aligned}
\int_{\Theta_m} \exp\{\ell_m(\theta)\} p(\theta|m) \mathrm{d}\theta &\approx \int_{\Theta_m} \exp\left\{\ell_m(\widehat{\theta}_m) + \frac{1}{2}(\theta - \widehat{\theta}_m)^T \nabla^2 \ell_m(\widehat{\theta}_m)(\theta - \widehat{\theta}_m)\right\} p(\theta|m) \mathrm{d}\theta \\
&\approx \exp\{\ell_m(\widehat{\theta}_m)\} p(\widehat{\theta}_m|m) \int_{\Theta_m} \exp\left\{\frac{1}{2}(\theta - \widehat{\theta}_m)^T \nabla^2 \ell_m(\widehat{\theta}_m)(\theta - \widehat{\theta}_m)\right\} \mathrm{d}\theta \qquad (2.5) \\
&= \exp\{\ell_m(\widehat{\theta}_m)\} p(\widehat{\theta}_m|m) \left(\frac{(2\pi)^{d_m}}{k^{d_m} \det I(\widehat{\theta}_m)}\right)^{1/2}. \qquad (2.6)
\end{aligned}$$

Here, the second approximation (2.5) follows from noting that $\frac{1}{2}(\theta - \widehat{\theta}_m)^T \nabla^2 \ell_m(\widehat{\theta}_m)(\theta - \widehat{\theta}_m)$ being a negative definite quadratic form in θ, and having a extremum at $\theta = \widehat{\theta}_m$. The final line (2.6) is due to the integrand being proportional to a Gaussian density.

Note that it is reasonable to neglect the higher order terms in the Taylor expansion here; Since these terms will be asymptotically negligible because of the behaviour of the gradient of ℓ_m as k increases, under the smoothness assumption.

Thus, following from (2.4), we have that the posterior model probability $\pi(m|\boldsymbol{y})$ can be approximated by

$$\text{const} + l_m(\widehat{\theta}_m) - \frac{1}{2}\dim(\Theta_m)\log k + \log p(\widehat{\theta}_m|m) + \frac{1}{2}\dim(\Theta_m)\log(2\pi) - \frac{1}{2}\log\det I_{\boldsymbol{Y}}(\widehat{\theta}_m) + \log p(m).$$

As $k \to \infty$, the last four terms will disappear; subsequently we define the BIC for model $S_m \in \mathcal{S}$ to be,

$$\text{BIC}(S_m) \doteq -2\ell_m(\widehat{\theta}_m^{(\text{MLE})}) + \dim(\Theta_m)\log k.$$

Thus, the *maximum a-posteriori* model is approximated by the model which minimises the BIC over $\mathcal{S}$, in other words define

$$S^*_{\text{BIC}} \doteq \arg\min_{S_m \in \mathcal{S}} \text{BIC}(S_m),$$

and we have that $S^*_{\text{Bayes}} \approx S^*_{\text{BIC}}$.

Since the BIC imposes a stronger penalty for each additional parameter, simpler models are chosen when compared to AIC. It is clear from the above, that similar to the AIC, BIC assumes that the sample size is large enough. Another assumption is the regularity of the likelihood function. These considerations restrict the application of BIC for some settings with irregular likelihoods, see Robert (2007, Section 7.2.3) and references therein. Further, the fact that BIC does not depend on the prior distribution has been criticised as eliminating subjective input using Bayesian modelling.

Alternatively, it can be argued that this is advantageous since it avoids the problems of priors which can typically be hard to specify.

As briefly discussed above, we saw that cross-validation and AIC can be inconsistent when the true model is included in the set of candidate models (Shao, 1993; Stone, 1977). In comparison, BIC does consistently select the true model, if it is included. In summary, AIC and cross-validation are more useful in finding the model with the best predictive performance, whereas the BIC method is useful for identifying the best explanatory model.

2.3.2 Bayes Factor

A popular approach in Bayesian model selection involves considering a simplified setting where we wish to select between only two models $S_i, S_j \in \mathcal{S}$. This leads to the evidence problem — what would be a good numerical metric of evidence in favour one model over the other?

In the Bayesian context one such measure of evidence is the posterior odds of the two models S_i and S_j. In fact, it is more common to consider the Bayes factor defined

$$\begin{aligned} B_{ij} &\doteq \frac{f(\boldsymbol{y}|M=i)}{f(\boldsymbol{y}|M=j)} \\ &= \frac{\int_{\Theta_i} f(\boldsymbol{y}|\theta,i)p(\theta|i)\mathrm{d}\theta}{\int_{\Theta_j} f(\boldsymbol{y}|\theta,j)p(\theta|j)\mathrm{d}\theta}, \end{aligned}$$

as first proposed by Jeffreys (1935), and later developed by Kass and Raftery (1995). To see one motivation for using the Bayes factor, consider the following

$$\begin{aligned} \text{posterior odds} &\doteq \frac{\pi(i|\boldsymbol{y})}{\pi(j|\boldsymbol{y})} \\ &= \frac{p(i)\int_{\Theta_i} f(\boldsymbol{y}|\theta,i)p(\theta|i)\mathrm{d}\theta}{p(j)\int_{\Theta_j} f(\boldsymbol{y}|\theta,j)p(\theta|j)\mathrm{d}\theta} \\ &= \frac{p(i)}{p(j)} \cdot B_{ij} \\ &= \text{ prior odds} \times B_{ij}. \end{aligned}$$

In words, we may think of the posterior odds as the transformation of the prior odds by the Bayes factor, or the Bayes factor as the ratio of the posterior odds of m to its prior odds.

In particular, when using a uniform prior distribution over the model space $\mathcal{S}$, the Bayes factor and the posterior odds are equivalent. In such cases, to compute the Bayes factor we require only the marginal likelihood, $f(\boldsymbol{y}|m)$ of each model $S_m \in \mathcal{S}$.

Suppose instead that we wish consider more than just two models. Since $B_{ij} = B_{im}B_{mj}$, for any $m \neq i, j$ — the ordering of models within the Bayes factors is transitive. So model selection, once more, reduces to finding the model with the highest marginal likelihood. Recall that

$$\pi(m|\boldsymbol{y}) \propto f(\boldsymbol{y}|m)p(m), \qquad \text{for } S_m \in \mathcal{S}.$$

In particular, the normalisation constant ($f(\boldsymbol{y})$) is the same for all the models.

In view of this, we may, as discussed above, minimise the posterior expected loss with a 0-1 loss function; We arrive at the objective of maximising the posterior model probability. Thus, this

approach can also be thought of as approximating the posterior model probability. If using uniform model priors, however, here we may do so by attempting to compute the marginal likelihood. This gives us:

$$S^*_{\text{BF}} \doteq \arg\max_{m \in \mathcal{M}} f(\boldsymbol{y}|m),$$

for given data set $\boldsymbol{y}$, and S^*_{BF} can be thought of as the model chosen using Bayes factors.

The Bayesian approach to model selection is appealing for a number of reasons. First, given the probabilistic interpretation of the results, it is very easy to account for model uncertainty. This naturally, leads to Bayesian model averaging, see Raftery et al. (1997) for example, in cases where more than one model is supported by the data.

More specifically, model averaging involves the computation of estimates under each candidate model; Then, calculating the weighted average of these estimates (Wasserman, 2000). Typically the weights are based on how likely each model is. An intuitive Bayesian approach is to use the (model) posterior probability, $f(\boldsymbol{y}|M)$ for each $M \in \mathcal{M}$, as the weights.

Bayesian model averaging is very useful in many contexts where there may be multiple candidate models that remain viable *a posteriori*. Of course, an important consideration when doing Bayesian model averaging is that the results will depend on the prior probabilities assigned to each candidate model (Hinne et al., 2020).

Next, a Bayesian model selection approach does not rely on asymptotic behaviour of the data, compared to the model selection methods reviewed earlier. Indeed, within the Bayesian framework a large sample size of the data can reduce uncertainty, however it is not necessary.

These advantages allow Bayesian model comparison to be used in a wide range of applications. In particular, these considerations are particularly appealing in challenging contexts, such as PET. Indeed, as Zhou et al. (2013) showed, using a Bayesian approach showed significant improvement in results and motivated further exploration of spatial dependence.

As previously mentioned, usually computing the marginal likelihood for each model is a non-trivial task; as, realistically speaking, they cannot be obtained analytically — instead we must turn to numerically approximating it. Fortunately, there exists a rapidly growing number of computational methods that are suited to this objective (and to which almost all of Bayesian analysis relies upon). These Bayesian computational methods are called Monte Carlo methods, we discuss them in detail for the remainder of this chapter.

2.4 Monte Carlo Approximations for Bayesian Model Selection

The Monte Carlo approach is notable for its simplicity, scalability and generality. Formally developed and named by Stainslaw Ulam and John von Neumann, Monte Carlo methods were initially used to study neutron diffusion as part of the Manhattan Project (Metropolis, 1987). However, with the widespread use of computers, this class of algorithms rapidly gained popularity and application in a wide variety of fields. They include finance, chemical physics, structural biology, operations research an queuing systems. Within Statistics, Monte Carlo methods become a mainstay with the development of the Bootstrap by Efron (1979), and later computer intensive methods for Bayesian analysis Geman and Geman (1984) [1].

[1] We note that, even this early on, Monte Carlo methods were used very successfully on image data sets

Broadly speaking, Monte Carlo methods are a class of algorithms that provide numerical solutions to analytically intractable problems, through the use of generated random samples. Within the sub-field of Bayesian statistical analysis, these numerical results typically approximate various characteristics of intractable probability distributions. The class of problems Monte Carlo methods address could be loosely categorised as: generating random samples from probability distributions, numerical integration and optimisation. In particular, the novel Bayesian spatial model constructed in this thesis, in Section 5.3, will give rise to *all* of these problems.

For instance, it is immediate that numerical integration would be a viable strategy for estimating the marginal likelihood $f(\boldsymbol{y}|m)$. Indeed, there are many numerical and simulation-based methods that could be used for approximating the marginal likelihood, see for example Green and Heikkinen (2003). However, as is usual under a Bayesian framework, Monte Carlo methods are by far the most common choice. Next, as alluded to before, the number of possible (spatial) models can be very large. As will be discussed later, posterior distributions constructed over this model space will tend to be intractable. Thus, in order to select a model, a method to reliably approximate such distributions will be required. Finally, selecting the "best" spatial model is in many ways a optimisation problem, as also previously suggested. These aspects, and other considerations will be discussed more formally and in detail later in Chapter 5, Part II.

Subsequently, the proposed novel Monte Carlo method for spatial model selection, introduced in Section 5.4.1, must address these problems. It does so using an amalgamation of existing, commonly used and extensively studied set of different Monte Carlo methods. In view of this, the remainder of this chapter will be a brief study of these Monte Carlo methods and their related relevant alternatives. As such, the sections that follow will be a formal discussion of the most popular Monte Carlo methods, the important theoretical properties that they posses, what role they play in Bayesian model selection, and relevant applicable strategies and extensions to improve performance and efficiency.

Furthermore, the discussion of Monte Carlo methods in this chapter will give intuition as well as introduce terms and concepts that can describe the sequential Monte Carlo sampler, in Chapter 3. This sampler will be the main estimator for the marginal likelihood, to be used in an applied setting in Part II. We conclude this section by first formally describing the generic problem that Monte Carlo methods attempt solve, and the basic principles used to do so.

To introduce some notation, denote $\mathcal{X}$ to be the state space, and $\mathcal{B}(\mathcal{X})$ the associated Borel $\sigma-$algebra. Let X be a random variable defined in this measurable space $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$. Finally, let the measure $\mu : \mathcal{B}(\mathcal{X}) \to [0, 1]$ denote the probability distribution of X, μ is typically called the *target distribution*. As before, assume that for μ that the density, with respect to some base measure such as the Lebesgue measure $\mathrm{d}x$, exists. With abuse of notation, the symbol μ will be overloaded to denote the density also.

Monte Carlo algorithms aim to approximate features of the target distribution μ, that are otherwise analytically intractable. For instance suppose that μ is only know up to some normalising constant. In other words, the target density can be written

$$\mu(x) = \frac{\gamma(x)}{Z} \text{ for all } x \in \mathcal{X},$$

with some density $\gamma : \mathcal{X} \to [0, \infty)$; and, where the normalising (or normalisation) constant

$$Z \doteq \gamma(\mathcal{X}) = \int_{\mathcal{X}} \gamma(x)\mathrm{d}x,$$

is an integral that cannot be easily computed using analytical methods. The marginal likelihood (which can be thought of as the normalising constant of the parameter posterior) is a pertinent example — in most cases of interest it cannot be easily computed using standard non-numerical methods.

Another, more commonly, approximated quantity is the expectation. Formally, given a measurable test function $\varphi : \mathcal{X} \to \mathbb{R}$ we seek the expectation with respect to μ, defined

$$\mathbb{E}_\mu[\varphi(X)] \doteq \int_{\mathcal{X}} \varphi \mathrm{d}\mu = \int_{\mathcal{X}} \varphi(x)\mu(x)\mathrm{d}x.$$

Typically, approximations of these expectations are used as Bayes estimators in statistical inferences of parameters. Again, these integrals are typically with respect to a posterior distributions that may not be easily solved using analytical approaches. Given their popularity, most literature on Monte Carlo methods focus primarily, or sometimes exclusively, on the properties of the estimates of the expectations. Indeed, the ability of Monte Carlo methods to estimate this expectation for *any* test function with accuracy, highlights its generality. In this work, given the primary task is model selection we prioritise attention on the normalising constant estimators — however, we will briefly discuss and make use of the expectation approximations as well.

The general strategy that all Monte Carlo methods use to solve these problems is to use simulations, from either the true or approximations of the target distribution. However, in almost all cases of practical interest, it is not possible to directly sample from the target distribution. Still, considering the simpler case where we can simulate from μ proves to be informative, as it serves to highlight the general principle which all Monte Carlo estimators are very loosely based on.

The concept can be summarised as follows: Given an $n-$sized i.i.d. random sample, from the target distribution μ, denoted $\{X^{(i)}\}_{i\geq 1}^n$, the Monte Carlo approximation of the target distribution can be written using the empirical distribution

$$\widehat{\mu} = n^{-1} \sum_{i=1}^{n} \delta_{X^{(i)}},$$

where we recall that δ_X denotes the Dirac measure. In particular, the estimator $\widehat{\mu}$ can be used to approximate the expectation of the test function φ with respect to the target distribution. That is, the Monte Carlo approximation of $\mathbb{E}_\mu[\varphi]$ can be written

$$\widehat{I}^{(\mathrm{MC})}(\varphi) \doteq \mathbb{E}_{\widehat{\mu}}[\varphi] = \int \varphi \mathrm{d}\widehat{\mu} = \frac{1}{n} \sum_{i=1}^{n} \varphi(X^{(i)}).$$

Importantly, this estimator is unbiased, and by SLLN converges almost surely (a.s.) to $\mathbb{E}_\mu[\varphi]$. Further, the variance of the error of this approximation decreases at the rate of $O(1/n)$, irrespective of the dimension of the state space. This is the characteristic advantage of the Monte Carlo method over other numerical methods — the versatility despite the minimal prerequisite of the target distribution.

In contrast to the setting of the Monte Carlo estimator above, in practice, it is often not possible to directly simulate samples from the target distribution. For instance, in Bayesian analysis, the posterior distribution is typically the target distribution. Indeed, posterior distributions can get unmanageable quickly in complex, challenging situations. The wide variety of Monte Carlo algo-

rithms developed over its history attempt to address this in many interesting ways. An important property of these more sophisticated Monte Carlo methods is that there is no assumption that it is possible to sample from the target distribution — we only need to know the target density point-wise. Instead, the general strategy is to use samples simulated from other tractable distributions. Once a correction for the error of sampling from another distribution is made, this sample can be used to approximate the above quantities of the original target distribution μ. In fact, even the assumption of being able to evaluate the target distribution point-wise may be also relaxed when we consider the so-called pseudo-marginal Monte Carlo methods, in Section 2.7.

2.5 Importance Sampling

Importance Sampling (IS) is a simple and useful Monte Carlo methods that allows straightforward estimation of both the normalising constant as well as expectations of test functions. In brief, IS samples from tractable probability distributions and corrects for doing so with weights. The generated random sample and the associated weights can then be used to approximate quantities of interest from the target distribution. Specifically, the weighted sample could be used to estimate expectations of test functions, and the weights themselves used to estimate the normalising constant of the target distribution.

More formally, denote by ν the distribution from which we can easily sample, called the *proposal distribution*. Specifically, assume that μ is absolutely continuous with respect to ν; and for which the Radon-Nikodym derivative

$$\rho(x) \doteq \frac{\mathrm{d}\mu}{\mathrm{d}\nu}(x)$$

is known. Suppose that we wish to compute $\mathbb{E}_\mu[\varphi(X')]$, recall this is expectation with respect to $X' \sim \mu$. Given $X \sim \nu$, and the assumptions above, it follows that

$$\mathbb{E}_\mu[\varphi(X')] = \mathbb{E}_\nu[\rho(X)\varphi(X)], \tag{2.7}$$

called the IS fundamental identity.

This motivates the following Monte Carlo estimator: given a random sample $\{X^{(i)}\}_{i=1}^n$ from the proposal distribution ν, approximate the expectation $\mathbb{E}_\mu[\varphi]$ by the estimator

$$\frac{1}{n}\sum_{i=1}^{n} \rho(X^{(i)})\varphi(X^{(i)}).$$

Doing so, gives that

$$\frac{1}{n}\sum_{i=1}^{n} \rho(X^{(i)})\varphi(X^{(i)}) \xrightarrow{\text{a.s.}} \mathbb{E}_\nu[\rho(X)\varphi(X)] = \mathbb{E}_\mu[\varphi(X')],$$

by SLLN. That is, the $\rho-$weighted empirical average of any test function, consisting of random sample from the proposal density ν, converges a.s. to the expectation with respect to the target density μ. Viewing IS in this somewhat formal manner, gives rise to the interpretation, as suggested by Chopin and Papaspiliopoulos (2020, Section 8.4), of the IS identity as a "change of measure" from the proposal measure to the target measure. This is a useful way to view some Monte Carlo methods, such as sequential Monte Carlo methods which uses similar ideas to IS, see Section 3.1.

In practice, we assume that the distributions μ, ν have densities with respect to the Lebesgue measure $\mathrm{d}x$; the Radon-Nikodym derivative simplifies to $\rho(x) = \frac{\mu(x)}{\nu(x)}$. Where, as above, we overload the distribution to denote the density as well. In particular, in the usual cases where $\mu \equiv \gamma/Z$ is only known up to a normalising constant, instead of ρ we may use the *unnormalised importance weights* defined

$$w(x) \doteq \frac{\gamma(x)}{\nu(x)} \text{ for } x \in \mathcal{X}.$$

Intuitively, the importance weights are the observed Radon-Nikodym derivatives of the unnormalised distribution γ with respect to ν.

Importantly, it follows from the fact

$$Z = \int_{\mathcal{X}} \gamma(x)\mathrm{d}x = \int_{\mathcal{X}} w(x)\nu(x)\mathrm{d}x,$$

that we may use the empirical average of the observed unnormalised weights as an estimator for the normalising constant. More precisely, given a random sample $\{X^{(i)}\}_{i=1}^n$ from the proposal distribution ν the IS estimator for the normalising constant Z of the target distribution $\mu = \gamma/Z$ is given by

$$\widehat{Z}^{(\mathrm{IS})} \doteq \frac{1}{n}\sum_{i=1}^{n} w(X^{(i)}). \tag{2.8}$$

It is immediate, once more, that by SLLN that this estimator will converge to Z a.s.. In addition, it is straightforward to show that $\widehat{Z}^{(\mathrm{IS})}$ is also unbiased, see Proposition 2.5.1.

Finally, noting that $\rho \equiv \frac{1}{Z}w$, and that the unnormalised weights will rarely sum to n — motivates the definition of the *normalised importance weights*

$$W^{(i)} \doteq \frac{w(X^{(i)})}{\sum_{j=1}^n w(X^{(j)})} \text{ for } i = 1, \ldots, n;$$

and, the self-normalising IS estimator of the expectation

$$\widehat{I}^{(\mathrm{IS})}(\varphi) \doteq \frac{\sum_{i=1}^n w(X^{(i)})\varphi(X^{(i)})}{\sum_{i=1}^n w(X^{(i)})} = \sum_{i=1}^{n} W^{(i)}\varphi(X^{(i)}).$$

2.5.1 Unbiased IS Estimator for Marginal Likelihoods

Recall, from Section 2.3.2, that we wish to compute the marginal likelihood $f(\boldsymbol{y}|m)$. Consider the following, for a pre-specified model $S_m \in \mathcal{S}$, given data $\boldsymbol{y}$, the posterior density of the parameter $\theta \in \Theta_m$ can be written

$$\pi(\theta|\boldsymbol{y}, m) = \frac{f(\boldsymbol{y}|\theta, m)p(\theta|m)}{\int_{\Theta_m} f(\boldsymbol{y}|\theta, m)p(\theta|m)\mathrm{d}\theta} = \frac{f(\boldsymbol{y}|\theta, m)p(\theta|m)}{f(\boldsymbol{y}|m)}.$$

In other words, if we target the unnormalised posterior density $f(\boldsymbol{y}|\theta, m)p(\theta|m)$; Monte Carlo methods that allow for approximations of the normalising constants, such as $\widehat{Z}^{(\mathrm{IS})}$, can be readily used to approximate the marginal likelihood itself $f(\boldsymbol{y}|m)$. In principle, this means IS can be used in Bayesian model selection directly.

However, the performance of IS depends on the choice of the proposal distribution (Robert and Casella, 2005, Section 3.3.2). Specifically, good performance is obtained when ν is close to μ — that

is, knowledge of the posterior distribution is required. Typically, this is not possible for complex models, hence the need to use Monte Carlo methods in the first place. Adaptive schemes that can address this problem, such as those introduced by Oh and Berger (1992), are available; however, they can be difficult to apply for high dimensional or multimodal target distributions. For these reasons, among other considerations, more sophisticated methods such as sequential Monte Carlo algorithms, as discussed in Chapter 3, are preferred.

In some other specialised settings, *multiple* estimates of a marginal (i.e. the marginal likelihood) may be needed. For example, in pseudo-marginal Monte Carlo methods, discussed in Section 2.7.1, multiple estimates of the marginal likelihood are used to approximate values within the algorithm. Specifically, in that context, these approximations are required to be unbiased. We now show that the IS estimator $\widehat{Z}^{(\mathrm{IS})}$ could be used as an unbiased estimator of the marginal likelihood.

Proposition 2.5.1. Given an $n-$sized random sample $X^{(1)}, \ldots, X^{(n)} \overset{i.i.d.}{\sim} \nu$ from the proposal distribution. We have that

$$\mathbb{E}_\nu[\widehat{Z}^{(\mathrm{IS})}] = Z,$$

where $Z = \int_{\mathcal{X}} \gamma(x)\mathrm{d}x$ is the normalising constant of the unnormalised density $\gamma : \mathcal{X} \to [0, \infty)$.

Proof. Straightforwardly,

$$\begin{aligned}
\mathbb{E}_\nu[\widehat{Z}^{(\mathrm{IS})}] &= \mathbb{E}_\nu\left[\frac{1}{n}\sum_{i=1}^{n} w(X^{(i)})\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\int \nu(x)w(x)\mathrm{d}x \\
&= \int_{\mathcal{X}} \gamma(x)\mathrm{d}x
\end{aligned} \tag{2.9}$$

□

2.5.2 Effective Sample Size

Clearly, there may be some loss of efficiency due to using weighted samples from a proposal distribution, instead of i.i.d samples from the target distribution. Perhaps surprisingly, there are (proposal) distributions which will give better expectations of the integral $\mathbb{E}_\mu[\varphi]$ than using μ itself; see, for example, Robert and Casella (2005, Section 3.3). However, obviously in terms of the approximation of the distribution itself there is a price to pay for sampling from a different one. The effective sample size (ESS), introduced by Kong (1992) and then later popularised by Liu (1996), is a statistic that is often used to quantify this loss. In Section 3.3.1, where sequential IS is discussed, the ESS is used as a criterion within an adaptive scheme. A detailed discussion the derivation of this statistic in the simpler IS context is given below, and further discussion held until the study of sequential methods.

Given an i.i.d. random sample $\{X^{(i)}\}_{i=1}^n$ from the proposal distribution ν, the ESS for the IS estimator $\widehat{I}^{(\mathrm{IS})}$, is defined

$$\mathrm{ESS}^{(\mathrm{IS})} \doteq \frac{n}{1 + \mathrm{Var}_\nu[W^{(1)}]}.$$

Here, the variance with respect to the distribution π is denoted

$$\mathrm{Var}_\pi[X] \doteq \mathbb{E}_\pi[X - \mathbb{E}_\pi[X]]^2.$$

In brief, the ESS can be thought of as the sample approximation of a low order Taylor expansion of the ratio between the variances of: the IS estimator $\widehat{I}^{(\mathrm{IS})}$; and the simple Monte Carlo estimator $\widehat{I}^{(\mathrm{MC})}$.

To be more precise, recall the Monte Carlo estimator $\widehat{I}^{(\mathrm{MC})}$ used n samples from the target distribution , $X^{(1)}, \ldots, X^{(n)} \overset{\text{i.i.d}}{\sim} \mu$, and was given by

$$\widehat{I}^{(\mathrm{MC})}(\varphi) = \frac{1}{n} \sum_{i=1}^{n} \varphi(X^{(i)}).$$

Similarity, the IS estimator $\widehat{I}^{(IS)}$ uses weighted samples from the proposal distribution $X^{(1)}, \ldots, X^{(n)} \overset{\text{i.i.d.}}{\sim} \nu$, written

$$\widehat{I}^{(\mathrm{IS})}(\varphi) = \sum_{i=1}^{n} W^{(i)} \varphi(X^{(i)}).$$

The loss in quality of the IS estimate due to the discrepancy of sampling from ν rather than μ can be studied using the ratio of the variance of the two estimates,

$$\frac{\mathrm{Var}_{\mu}\left[\widehat{I}^{(\mathrm{MC})}(\varphi)\right]}{\mathrm{Var}_{\nu}\left[\widehat{I}^{(\mathrm{IS})}(\varphi)\right]}.$$

It is straightforward to compute $\mathrm{Var}_{\mu}[\widehat{I}^{(\mathrm{MC})}]$, however $\mathrm{Var}_{\nu}[\widehat{I}^{(\mathrm{IS})}]$ has to be approximated. We are using the self-normalising estimator here, so complication arises due to taking the ratio of two random quantities. We momentarily assume that μ is normalised, the derivation follows the same for the unnormalised target, because the ratio $\widehat{I}^{(\mathrm{IS})}$ does not depend on the normalising constant. To this end, following Kong's original shorthand notations $W \doteq W^{(1)}$ and $H \doteq \varphi(X^{(1)})$, note the following statements:

$$\mathbb{E}_{\nu}[W] = \int \frac{\mu(x)}{\nu(x)} \nu(x) \mathrm{d}x = \int \mu(x) \mathrm{d}x = 1; \tag{2.10}$$

$$\mathbb{E}_{\nu}[HW] = \int \varphi(x) \frac{\mu(x)}{\nu(x)} \nu(x) \mathrm{d}x = \int \varphi(x) \mu(x) \mathrm{d}x = \mathbb{E}_{\mu}[H]; \tag{2.11}$$

$$\mathbb{E}_{\mu}[W] = \mathbb{E}_{\nu}[W^2] = \mathrm{Var}_{\nu}[W] + (\mathbb{E}_{\nu}[W])^2 = \mathrm{Var}_{\nu}[W] + 1; \tag{2.12}$$

and

$$\mathrm{Var}_{\mu}[H] = n \mathrm{Var}_{\mu}[\widehat{I}^{(\mathrm{MC})}]. \tag{2.13}$$

Using the asymptotic delta method approximation, see Liu (1996), gives us that,

$$\begin{aligned}
\mathrm{Var}_\nu\left[\widehat{I}^{(\mathrm{IS})}\right] &\approx \frac{1}{n}\left[\frac{\mathrm{Var}_\nu[HW]}{(\mathbb{E}_\nu[W])^2} - 2\frac{\mathbb{E}_\nu[HW]}{(\mathbb{E}_\nu[W])^3}\mathrm{Cov}_\nu(HW,W) + \frac{(\mathbb{E}_\nu[HW])^2}{(\mathbb{E}_\nu[W])^4}\mathrm{Var}_\nu[W]\right] \\
&= \frac{1}{n}\left(\mathrm{Var}_\nu[HW] - 2\mathbb{E}_\mu[H]\mathrm{Cov}_\nu(HW,W) + \mathbb{E}_\mu[H]^2\mathrm{Var}_\nu[W]\right) && \text{via (2.10) and (2.11)} \\
&\vdots && \text{see Appendix A} \\
&\approx \frac{1}{n}\Big(\mathbb{E}_\mu[H]^2\{1 + \mathrm{Var}_\nu[W] - \mathbb{E}_\mu W\} + \mathbb{E}_\mu[W]\mathrm{Var}_\mu[H]\Big) \\
&= \frac{1}{n}\Big(\mathbb{E}_\mu[H]^2\{1 + \mathrm{Var}_\nu(W) - \mathrm{Var}_\nu(W) - 1\} + \{\mathrm{Var}_\nu(W) + 1)\}\mathrm{Var}_\mu(H)\Big) && \text{using (2.12) above} \\
&= \mathrm{Var}_\mu[\widehat{I}^{(\mathrm{MC})}](1 + \mathrm{Var}_\nu(W)) && \text{using (2.13) above.}
\end{aligned}$$

Where the covariance, with respect to distribution π, of test functions φ and h is defined

$$\mathrm{Cov}_\pi(\varphi, h) \doteq \mathbb{E}_\pi[\varphi(X) - \mathbb{E}_\pi[\varphi(X)]]\mathbb{E}_\pi[h(X) - \mathbb{E}_\pi[h(X)]].$$

The intermediate steps, which includes further approximations using the delta method, can be found in detail in Appendix A.

Substituting this approximation into our quantity of the efficiency of the IS estimator compared to the direct Monte Carlo estimate, gives us

$$n \times \frac{\mathrm{Var}_\mu[\widehat{I}^{(\mathrm{MC})}]}{\mathrm{Var}_\nu[\widehat{I}^{(\mathrm{IS})}]} \approx \frac{n}{(1 + \mathrm{Var}_\nu[W])} = \mathrm{ESS}^{(\mathrm{IS})}.$$

Here, since the ratio (often called the relative ESS) is a proportion representing the loss of information — we multiply by the sample size n for scale.

Given that IS samples $X^{(1)}, \ldots, X^{(n)} \overset{\text{i.i.d.}}{\sim} \nu$; and that the associated unnormalised weights $W^{(i)}$ can be seen as a transformation of $X^{(i)}$, we may estimate the variance of W in the ESS using the sample variance. In other words, since

$$\mathrm{Var}_\nu W \approx \frac{1}{n}\sum_{i=1}^{n}\left(W^{(i)} - \frac{1}{n}\right)^2 = \frac{1}{n}\sum_{i=1}^{n}\left((W^{(i)})^2 - \frac{1}{n^2}\right),$$

we may estimate the $\mathrm{ESS}^{(\mathrm{IS})}$ using only the normalised weights,

$$\widehat{\mathrm{ESS}}^{(\mathrm{IS})} = \frac{n}{1 + \frac{1}{n}\sum_{i=1}^{n}\left((W^{(i)})^2 - \frac{1}{n^2}\right)} = \frac{1}{\sum_{i=1}^{n}(W^{(i)})^2}. \tag{2.14}$$

This Taylor series approximation is independent of the test function φ and there is information lost in making this approximation. Although the ESS is a useful heuristic, the actual variance ratio does depend on the function being considered and, indeed, the variance of the IS estimator can be smaller than that of the simple Monte Carlo one.

Another consideration is that, in addition to the issue of independence from the test function, one major weakness of the ESS is that it tends to be a bad overestimate in some extreme instances. For example, if the importance weights have very high variance, then they will be very heavily skewed (because they are necessarily non-negative) and so most will be very small. A sample, in which only very small but reasonably homogeneous values are seen, will have a reasonably high

ESS – but that is an artefact of the Monte Carlo approximation. Evaluating the actual $\text{ESS}^{(\text{IS})}$ above would give quite a different result.

The ESS provides a good measure of the quality of weighted samples, and its uses lie beyond verification of the quality of IS outputs. As we discuss later in Section 3.3.1: Since the sequential Monte Carlo framework is built on IS it is readily available for use in that context too. Typically, the ESS can be used as a measure of increasing discrepancy due to repeated and sequential sampling from proposal distributions. And so, ESS can be used as an indicator of when to process the collection of weighted samples to get rid of degenerate sample points.

Chen (2005) discusses a connection to the χ^2-distance and proposes using it as a discrepancy between measures. In this vein, Jasra et al. (2011) suggests another use of interpreting the ESS as a distance between two distributions. Specifically, it can be utilised in implementing inference of a Bayesian framework under sequential Monte Carlo. Typically, a form of simulated annealing process is used, for example initialising the particles in the prior distribution and sequentially moving to the posterior distribution. Having a quantity that can measure the distance between distributions can be utilised to design annealing schemes that allow for a more stable transition. These concepts will be discussed in detail later, in Section 3.3.

2.6 Markov Chain Monte Carlo

Markov Chain Monte Carlo (MCMC) methods are a class of algorithms that are based on using random samples generated from Markov chain constructed in a specific manner such that its invariant distribution is the target distribution μ. This correlated sample can then be used to give a Monte Carlo estimate of the integral $\mathbb{E}_\mu[\varphi]$, for some test function φ; if the Markov chain is ergodic then the estimator will satisfy the law(s) of large numbers. The important theoretical justifications and results that allows for this and the manner in which such a $\mu-$invariant Markov chain is constructed is discussed in detail below. Although MCMC methods can be used to perform Bayesian model selection, for example using the harmonic mean estimator; we are mainly interested in the aforementioned theoretical results. Exploring these results and concepts leads naturally to the pseudo-marginal MCMC methods, discussed in the sequel. This then allows us to straightforwardly present the proposed methodology, in Part II, which is an extension of the pseudo-marginal algorithm.

We begin by first presenting some standard concepts relating to (discrete time) Markov chains. For these sections, we use measure notations as it significantly simplifies presentation of the theoretical arguments — the specialised cases when the density exists will follow directly. Familiarity with the basics of Markov chains is assumed; however notations and terms that are more involved and technical will be defined here. A concise but detailed exposition of the essentials of Markov chains, particularly in relation to MCMC, can be found in Robert and Casella (2005, Chapter 6). For a more complete, through treatment of the subject with a deeper exploration of these concepts see Meyn and Tweedie (2009).

Recall that $\mathcal{X}$ denotes the state space, and $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ the measurable space equipped with the associated Borel $\sigma-$algebra. Denote the function $K : \mathcal{X} \times \mathcal{B}(\mathcal{X}) \to [0, 1]$ to be a Markov kernel, defined such that:

- for any state $x \in \mathcal{X}, K(x, \cdot)$ is a probability measure on $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$;
- and the mapping $x \to K(x, A)$ is $\mathcal{B}(\mathcal{X})-$measurable for any $A \in \mathcal{B}(\mathcal{X})$.

A sequence of random variables $(X^{(i)})_{i\geq 0}$ is called a *Markov chain* with *transition kernel* K, if the conditional probability distributions of the sequence are of the form

$$X^{(i)}|X^{(i-1)} \sim K(X^{(i-1)}, \cdot) \text{ for } i = 1, 2, \ldots.$$

Conventionally, the first member of the sequence $X^{(0)}$ is called the initial state, the distribution of $X^{(0)}$ is called the initial distribution. Intuitively, a Markov chain can be completely specified by its kernel K and the initial distribution $X^{(0)} \sim P$. By denoting $K^1(x, A) = K(x, A)$, the kernel for i transitions can be written

$$K^i(x, A) \doteq \int_{\mathcal{X}} K^{i-1}(y, A)\mathrm{d}K_x(y);$$

where, we use the shorthand $K_x(\cdot)$ to refer to the measure $K(x, \cdot)$, for $x \in \mathcal{X}$.

As before, if the density, with respect to some reference measure such as the Lebesgue measure $\mathrm{d}x$, exists; then the kernel notation K will be overloaded to denote the conditional density $K(x, y)$ for $y \in \mathcal{X}$ i.e. such that

$$P(X^{(i)} \in A|X^{(i-1)} = x) = \int_A \mathrm{d}K_x(y) = \int_A K(x, y)\mathrm{d}y \text{ for } A \in \mathcal{B}(\mathcal{X}).$$

For ease of presentation we use the convention of y to refer to some value in the state space $\mathcal{X}$, in this subsection — it should not be confused with the notation for data set $\boldsymbol{y}$.

For a set $A \in \mathcal{B}(\mathcal{X})$ the hitting time is defined

$$\tau_A \doteq \min\{i > 0 : X_i \in A\}.$$

Next, a set $C \in \mathcal{B}(\mathcal{X})$ is small if there exists constants $\lambda > 0$, $i \geq 1$ and some probability distribution π such that C satisfies the minorisation condition:

$$K^i(x, A) \geq \lambda\pi(A) \text{ for all } x \in C, A \in \mathcal{B}(\mathcal{X}).$$

In addition, overload K to be an operator on measures

$$\pi K(A) \doteq \int_{\mathcal{X}} K(x, A)\mathrm{d}\pi(x), A \in \mathcal{B}(\mathcal{X}),$$

in words if $X^{(i-1)} \sim \pi$ then $X^{(i)} \sim \pi K$. Similarly, overload K to also define operators on measurable functions φ,

$$K\varphi(x) \doteq \int_{\mathcal{X}} \varphi(y)\mathrm{d}K_x(y), x \in \mathcal{X}.$$

In regards to Markov chains, take $\mathbb{E}_\pi[X^{(i)}]$ to mean the expectation of $X^{(i)}$ given that the initial state $X^{(0)} \sim \pi$. Then, $K\varphi(x)$ is simply $\mathbb{E}_x[\varphi(X^{(1)})] \doteq \mathbb{E}[\varphi(X^{(1)})|X^{(0)} = x]$.

MCMC methods exploit a very strong stability property of Markov chains termed the *invariant probability distribution*. A transition kernel K is said to have an invariant probability distribution π if

$$\pi K = \pi;$$

or, in other words

$$\pi(A) = \int_{\mathcal{X}} K(x, A)\mathrm{d}\pi(x) \text{ for all } A \in \mathcal{B}(\mathcal{X}).$$

The intuition is that for all $i \geq 0$ if $X^{(i)} \sim \pi$ then $X^{(i+1)} \sim \pi$. For sake of brevity, we will also use the term that kernel K is $\pi-$invariant.

By careful construction of specific transition kernels, which depend on the target distribution μ, MCMC algorithm generate Markov chains which have the target distribution as their invariant distribution. The realisations of these constructed Markov chain can then be used to estimate characteristics of interest of the target distribution. The theoretical results that:

- guarantee that these constructed kernels will have the target distribution μ as their invariant distribution;
- the subsequent estimators will converge to the true value;
- and show at what rate of convergence to the invariant distribution;

will be the focus of the following subsection.

2.6.1 Metropolis-Hastings Algorithm

The Metropolis-Hastings (MH) algorithm is a powerful and widely-used MCMC method, first introduced by Metropolis et al. (1953) and Hastings (1970). The algorithm generates samples by mapping given transition kernel Q, to a $\mu-$invariant Markov chain. Treat Q as some tractable proposal distribution, similar to IS — as such, denote its density, when it exists, ν.

More precisely, given target distribution μ and proposal distribution Q, the MH algorithm specifies the kernel $K^{(\mathrm{MH})}$, such that for all $A \in \mathcal{B}(\mathcal{X})$

$$K^{(\mathrm{MH})}(x, A) \doteq \int_A \min\{1, R(x, y)\}\mathrm{d}Q_x(y) + \alpha_x \mathbf{1}_A(x). \tag{2.15}$$

Here: $R(x, y)$ is known as the acceptance ratio — that is, given proposal $y \sim Q(x, \cdot)$, with probability $\min\{1, R(x, y)\}$ let $X^{(i+1)} = y$; Next,

$$\alpha_x \doteq 1 - \int_{\mathcal{X}} \min\{1, R(x, y)\}\mathrm{d}Q_x(y)$$

is the probability of staying at current state x; Lastly, $\mathbf{1}_A$ denotes the indicator function of set A.

Importantly, note that R, similar to IS, is more formally the Radon-Nikodym derivative which satisfies

$$\int_{x\in A}\int_{y\in B} \varphi(x, y)R(x, y)\mathrm{d}\mu(x)\mathrm{d}Q_x(y) = \int_{x\in A}\int_{y\in B} \varphi(x, y)\mathrm{d}\mu(y)\mathrm{d}Q_y(x) \tag{2.16}$$

for any measurable A, B and function φ; see Tierney (1998) for further discussion. This property is important when showing that $K^{(\mathrm{MH})}$ is reversible, thus motivating why choosing proposals in this manner allows the algorithm to work. In fact, even (unbiased) estimates of this ratio can be used to give us exact samples — we postpone discussion of this till the study of pseudo-marginal algorithms in the sequel.

Markov chains created in this manner are called MH Markov chains. Although the kernel $K^{(\mathrm{MH})}$ may appear complex in its closed form; note that in practice, and henceforth, we assume that the

densities of the proposal and target distribution, with respect to some base measure, exist. In which case, the acceptance ratio simplifies to

$$R(x,y) = \frac{\mu(y)\nu(y,x)}{\mu(x)\nu(x,y)}; \tag{2.17}$$

where we recall that ν denotes the density of the proposal kernel Q. Thus, the algorithmic steps required to construct a Markov chain with kernel $K^{(\mathrm{MH})}$ are very accessible:

Algorithm 1 The Metropolis-Hastings Algorithm

1. Given $x^{(i-1)}$.
2. Sample $x^* \sim \nu(x^{(i-1)}, \cdot)$
3. Compute
$$r(x^{(i-1)}, x^*) := \frac{\mu(x^*)\nu(x^*, x^{(i-1)})}{\mu(x^{(i-1)})\nu(x^{(i-1)}, x^*)}$$
4. With probability $\min\{1, r\}$ let:
$$x^{(i)} = x^*;$$
Otherwise
$$x^{(i)} = x^{(i-1)}.$$

It is clear that each MH kernel is essentially dependent on the choice of the proposal ν; the most popular choices for the proposal is to simply set the proposed value incrementally. That is, at i-th step of the Markov chain the proposed $i+1$-th step would be set to $X^{(i+1)} = X^{(i)} + \epsilon^{(i)}$, where the increments $\epsilon^{(i)}$ are random variables distributed according to some symmetric distribution. In other words,

$$X^{(i+1)} \sim P_{X^{(i)}},$$

for some *symmetric* distribution centred on $X^{(i)}$. For example, this could be the normal distribution with mean equivalent to the current state of the chain and some (co-)variance $\sigma^2 > 0$, which we denote:

$$X^{(i+1)} \sim \mathcal{N}(X^{(i)}, \sigma^2).$$

Such choices lead to the symmetric random-walk Metropolis algorithm (RWM), see Brooks et al. (2011) for detailed discussion of various properties of this algorithm. Among other reasons, RWM is a popular choice due to its simplicity in application and we will use it when analysing PET data.

The simplicity of implementation results in the MH algorithms, such as the RWM, being a very popular choice. It is not immediately obvious, however, why these steps will indeed produce samples that can be used to approximate features of the target distribution μ. By showing that the MH kernel produces a chain which will asymptotically produce samples distributed according to μ, Proposition 2.6.1 below shows why this is a valid method.

Existence of invariant distribution

Firstly, it is straightforward to show that MH Markov chains targeting the distribution μ are $\mu-$reversible: Given a measure π, we say a Markov chain with kernel K is $\pi-$reversible if for any

pair of disjoint measurable sets $A, B \in \mathcal{B}(\mathcal{X})$, the detailed balance condition

$$\int_A K(x, B)\mathrm{d}\pi(x) = \int_B K(x, A)\mathrm{d}\pi(x)$$

is satisfied. Once detailed balance can be verified for $K^{(\mathrm{MH})}$, it is trivial to show that it is $\mu-$invariant.

Proposition 2.6.1. The Metropolis-Hastings kernel $K^{(\mathrm{MH})}$ with target distribution μ, as specified in Eq.(2.15), is:

i) $\mu-$reversible;

ii) thus, $\mu-$invariant.

Proof. i) We just need to check the detailed balance condition; we have that for all $A, B \in \mathcal{B}(\mathcal{X})$,

$$\int_A K^{(\mathrm{MH})}(x, B)\mathrm{d}\mu(x) = \int_{x\in A}\int_{y\in B} \min\{1, R(x, y)\}\mathrm{d}Q_x(y)\mathrm{d}\mu(x) + \int_{x\in A} \alpha_x \mathbf{1}_B(x)\mathrm{d}\mu(x).$$

Taking the first term,

$$\begin{aligned}
&\int_{x\in A}\int_{y\in B} \min\{1, R(x, y)\}\mathrm{d}Q_x(y)\mathrm{d}\mu(x) \\
&= \int_{x\in A}\int_{y\in B} \left(R(x, y)\mathbf{1}_{R(x,y)<1} + \mathbf{1}_{R(x,y)\geq 1}\right)\mathrm{d}Q_x(y)\mathrm{d}\mu(x) \\
&= \int_{x\in A}\int_{y\in B} \mathbf{1}_{R(x,y)<1}\mathrm{d}Q_y(x)\mathrm{d}\mu(y) + \int_{x\in A}\int_{y\in B} \mathbf{1}_{R(x,y)\geq 1}\mathrm{d}Q_x(y)\mathrm{d}\mu(x),
\end{aligned}$$

where the last line follows from Eq.(2.16). Next, since $R(x, y) = R(y, x)^{-1}$ we have $\mathbf{1}_{R(y,x)<1} = \mathbf{1}_{R(x,y)\geq 1}$, and the right hand side is symmetric in A and B.

Next, for the second term, it follows that:

$$\begin{aligned}
\int_{x\in A} \alpha_x \mathbf{1}_B(x)\mathrm{d}\mu(x) &= \int_{x\in\mathcal{X}} \mathbf{1}_A(x)\cdot\mathbf{1}_B(x)\int_{y\in\mathcal{X}} 1 - \min\{1, R(x, y)\}\mathrm{d}Q_x(y)\mathrm{d}\mu(x) \\
&= \int_{x\in B} \mathbf{1}_A(x)\int_{y\in\mathcal{X}} 1 - \min\{1, R(x, y)\}\mathrm{d}Q_x(y)\mathrm{d}\mu(x) \\
&= \int_{x\in B} \alpha_x \mathbf{1}_A(x)\mathrm{d}\mu(x).
\end{aligned}$$

Thus, the detailed balance equation is satisfied.

ii) Given that $K^{(\mathrm{MH})}$ is $\mu-$reversible, we have that for all $A \in \mathcal{B}(\mathcal{X})$,

$$\mu K^{(\mathrm{MH})}(A) = \int_{\mathcal{X}} K^{(\mathrm{MH})}(x, A)\mathrm{d}\mu(x) = \int_A K^{(\mathrm{MH})}(x, \mathcal{X})\mathrm{d}\mu(x) = \int_A \mathrm{d}\mu(x) = \mu(A).$$

That is, $K^{(\mathrm{MH})}$ is also $\mu-$invariant. □

Given the Monte Carlo context, it is not possible to sample from the target distribution μ — thus, the initial distribution $X^{(0)} \sim \pi$ of the MH Markov chain cannot be μ. Therefore, we must consider under what conditions the MH Markov chain "reaches" its invariant distribution μ, and at what rate. In order to answer these interesting questions, we must discuss the long-run behaviour of

the $K^{(\mathrm{MH})}$ kernel Markov chain: That is, interrupting the index i in the Markov chain $(X^{(i)})_{i\geq 1}$ from a temporal perspective, the limiting behaviour of the chain is explored. Indeed, in general the limiting distribution may not exist. In the cases where it does, a natural candidate would be the invariant distribution. The theoretical results presented below show under what conditions this can be guaranteed.

Convergence to target distribution

Recall the following properties of Markov chains, used towards presenting the main convergence theorems: Given a measure π, the Markov chain $(X^{(i)})_{i\geq 0}$ with transition kernel K is said to be π-*irreducible* if for every $A \in \mathcal{B}(\mathcal{X})$ with $\pi(A) > 0$ there exists an integer $i \geq 1$ such that

$$K^i(x, A) > 0 \text{ for every } x \in \mathcal{X}.$$

In the context of MCMC, irreducibly guarantees that the generated chain will explore all of the full support of the target distribution. In general the MH algorithm does not always yield irreducible chains, however restricting the proposal distribution ν to satisfy some mild conditions, see for example Roberts and Tweedie (1995), will produce an irreducible chain.

Next, a slightly stronger form of recurrence introduced by Harris (1956), that will be useful in verifying convergence. A $\pi-$irreducible Markov chain is called Harris recurrent if there exists a small set C such that

$$P(\tau_C < \infty | X^{(0)} = x) = 1 \text{ for any } x \in \mathcal{X};$$

That is, the hitting time is almost surely finite. Furthermore, we say that the chain is positive Harris recurrent if in addition, $\sup_{x\in\mathcal{X}} \mathbb{E}[\tau_C | X^{(0)} = x] < \infty$.

Harris recurrence is an important property: for Markov chains that possess this property, the empirical average of its observed values will converge to the true expected value. This is known as the so-called Law of Large Numbers for Markov chains (Robert and Casella, 2005, Thoerem 6.63). The specialised form, for the MH kernel, of this Law is formally stated below. Importantly, if the chain with a MH kernel is irreducible it is can be straightforwardly shown that it also is Harris recurrent (Robert and Casella, 2005, Lemma 7.3). In fact, since the MH kernel produces a chain with a invariant distribution the chain is also positive Harris recurrent.

To show the rate of convergence, we require a quantity to measure how different distribution, over a common state space, are. This is captured by the total variation (TV) norm, defined for two measures π_1 and π_2 as

$$||\pi_1 - \pi_2||_{\mathrm{TV}} \doteq \sup_{A\in\mathcal{B}(\mathcal{X})} |\pi_1(A) - \pi_2(A)|.$$

Note that convergence in total variation also implies convergence in distribution.

The final important condition is the following property. We say that a Markov chain $(X^{(i)})$ is a *aperiodic* if it has a cycle of length one. Specifically, a $\mu-$irreducible chain has a cycle of length l if there exists a small set C, an associated integer M, and a probability distribution ν_M such that l is the g.c.d. of

$$\{m \geq 1 : \exists \delta_m > 0 \text{ such that } C \text{ is small for } \nu_m \geq \delta_m \nu_M\}.$$

Straightforwardly, MH Markov chains can be made aperiodic by using proposal distributions that allow for events such as $\{X^{i+1} = X^i\}$. The two important convergence results alluded to above, can now be formally stated.

Theorem 2.6.2. (Robert and Casella, 2005, Theorem 6.51). If a MH Markov chain $(X^{(i)})_{i\geq 0}$ with kernel K (superscript suppressed for clarity) is aperiodic and $\mu-$irreducible, then

$$\lim_{i\to\infty} ||\pi K^i - \mu||_{\text{TV}} = 0 \tag{2.18}$$

for all initial distributions π. We say that K is *ergodic*.

(Aperiodicity not needed here) Additionally, for measurable $\varphi : \mathcal{X} \to \mathbb{R}$,

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \varphi(X^{(i)}) = \mathbb{E}_\mu[\varphi] \text{ a.s.}$$

for every starting value $X^{(0)}$.

Pragmatically speaking, these are two of the most important results as they motivate and validate the use of the MH estimator of the expectation. More precisely, correlated samples generated from the MH algorithm $\{X^{(i)}\}_{i=1}^n$, with target distribution μ, can be used to estimate the integral $\mathbb{E}_\mu[\varphi(X)]$; by the MCMC estimator:

$$\widehat{I}^{(\text{MH})} \doteq \frac{1}{n}\sum_{i=1}^{n} \varphi(X^{(i)}).$$

Rate of Convergence

Given that we are not able to initialise the MH Markov chain with the target distribution itself, we may also be interested in knowing how quickly the chain approaches the invariant distribution (i.e. the target distribution μ). The notion of the rate of convergence of $K^{(\text{MH})}$ to μ is quantified by the property of *geometric ergodicity*. In brief, a Markov kernel is geometrically ergodic if the total variation distance defined in Eq.(2.18) can be bounded by constants, which may depend on the starting value. More formally, a Markov kernel K is said to be geometrically ergodic if

$$||K^i(x,\cdot) - \mu||_{\text{TV}} \leq G(x)\rho(x)^i, \tag{2.19}$$

for x $\mu-$a.e. and functions $G : \mathcal{X} \to \mathbb{R}^+, \rho : \mathcal{X} \to [0,1)$.

Although many MCMC methods do satisfy this condition, it is difficult to show that they do directly. Instead, it is sufficient to construct the Lypanuov function $L : \mathcal{X} \to [1,\infty]$ such that the following *drift condition* is satisfied: There exists a small set C and constants $0 < \kappa < 1, 0 < b < \infty$ such that

$$KV(x) \geq \kappa L(x) + b\mathbf{1}_C(x).$$

If the set C and function L satisfy the drift function, with respect to the kernel K, then the following inequality holds (Hairer and Mattingly, 2011; Meyn and Tweedie, 2009):

$$||K^i(x,\cdot) - \mu(\cdot)||_L \leq GL(x)\rho^i, \tag{2.20}$$

where $G < \infty, \rho < 1$ (do not depend on x), $L : \mathcal{X} \to [1,\infty)$, and the norm

$$||\mu||_L = \sup\left\{\left|\int F(x)\mu(dx)\right| : \sup_{x\in\mathcal{X}} \frac{|F(x)|}{L(x)} \leq 1\right\}.$$

Importantly, under aperiodicity and $\mu-$irreducibility, this inequality (2.20) is in fact equivalent to

(2.19) and thus, K being geometrically ergodic.

2.6.2 Adaptive MCMC Methods

In practice, a crucial component for rapid convergence (and thus the MH Markov chain spending more time in regions of the distribution that we are interested in) of a MH algorithm is the choice of the proposal density ν. Theoretically the optimal choice would be $\nu(y, \cdot) \equiv \mu(\cdot)$, however this is obviously not possible. When using the RWM algorithm, the optimal proposal problem will translate to the choice of the scaling of the symmetric distribution we wish to use (Brooks et al., 2011). For example, suppose we wish to target density μ over the space $\mathcal{X}$, with dimension denoted d. Let $\mathbb{1}_d$ denote the $d \times d$ identity matrix. If we wish to use normally distributed increments, that is if $X^{(i+1)} \sim \mathcal{N}(X^{(i)}, \sigma^2 \mathbb{1}_d)$, the choice of the variance σ^2 will play an important role in the success of the RWM algorithm.

In particular, if σ is too small than almost all proposed states will be accepted. However, this will typically results in smaller movements, meaning the state space may not be sufficiently explored. In other words, the chain will not mix well overall. On the other had, if we allow σ is too large then most proposals will be rejected, which could result in the chain not moving at all. An *ad hoc* way to avoid these two extremes is to monitor the acceptance rate, the fraction of proposed states that are accepted, and ensuring the fraction is not too close to zero or one. This is could be done through time-consuming trail and error: Running the algorithm to compute the acceptance ratio and tuning σ to the appropriate value. For most applications this is not practical and for cases with high-dimensional problems this may be infeasible.

Alternatively, the incredible and very useful result presented by Roberts et al. (1997), can be used to find optimal values in an adaptive manner. By making strong assumptions on the target density, Roberts et al. proved using diffusion theorems that as $d \to \infty$ the optimal acceptance rate is $2\Phi(-2.38/2) = 0.234$, where Φ is the cumulative distribution function of a standard normal. Later works, see Roberts and Rosenthal (2001), relax these strong assumptions and resulted in the rise in popularity of these accessible adaptive variants of the algorithm. Adaptive MCMC (AMCMC), as discussed in Brooks et al. (2011, chapter 4), are algorithms based on these results that attempt to learn the scaling that corresponds to this optimal acceptance rate online, i.e. as the chain is running.

For instance, an adaptive RWM algorithm uses proposal distribution for the i-th iteration defined by

$$X^{(i+1)} \sim \mathcal{N}\left(X^{(i)}, \left[\frac{(2.38)^2}{d}\right] \widehat{\Sigma}_i\right),$$

where $\widehat{\Sigma}_i$ is the empirical co-variance matrix of $X^{(0)}, \ldots, X^{(i)}$. Important, note that $\widehat{\Sigma}_i$ is dependent on the values of the previous states of X — thus the chain is no longer Markovian and so the useful results discussed above no longer hold. Instead, this adaptive form is used in the initial period, and then the scaling is fixed; the samples from the adaptive part are discarded (similar to a burn–in period). As we will discuss in Section 3.2, AMCMC methods can often times be ideal proposal kernels within sequential Monte Carlo methods.

There exists even more sophisticated variants of MCMC methods. For example, the Metropolis-Adjusted Langevin algorithm (MALA; Roberts and Tweedie (1995)), where new states are proposed in the direction in which the target density μ is increasing. This is achieved through computing the discrete-time approximation to the continuous Langevin diffusion for μ. Under similar strong assumptions on the target density as the AMCMC case, Roberts and Rosenthal (1998)

proved that the optimal acceptance rate for MALA was 0.574. Given that this significantly larger than for AMCMC above, MALA has faster convergence. In practice, however, adaptive RWM is typically chosen over MALA. This is because MALA involves the computation of the gradient of the target density, which can be time-consuming or difficult to compute, in realistic settings.

Another example of the similar gradient-based approach is the Hamiltonian Monte Carlo (Duane et al., 1987). Similar to the standard non-adaptive MCMC, an important consideration when using these methods, in addition to the computation of the gradient, is that practitioners must adjust tuning parameters to ensure good performance of the algorithm. In some settings, this may require a lot of time. More recently, Livingstone and Zanella (2020) proposed the a gradient-based MCMC algorithm called the Barker proposal, that showed increased level of robustness to tuning, compared to current state-of-the-art gradient-based methods such as MALA or Hamiltonian Monte Carlo.

2.6.3 MCMC Estimators for Marginal Likelihood

MCMC algorithms can be used in numerous methods to perform Bayesian model selection, often via estimating the marginal likelihood $f(\boldsymbol{y}|m)$, see Green and Heikkinen (2003) for a detailed review. Many of these methods, however, are used in settings that require and justify additional complexity. Here, we briefly discuss a simple approach called the harmonic mean estimator, which has been successfully used in the applications of current interest (e.g. PET images, as presented by Zhou et al. (2013)).

Usually, MCMC methods are developed to sample from the parameter posterior density denoted $\pi(\theta|\boldsymbol{y}, m) \propto f(\boldsymbol{y}|\theta, m)p(\theta|m)$. Thus, a natural approach for model selection would be to use these MCMC samples, of the parameter θ, denoted $(\theta^{(1)}, \ldots, \theta^{(n)})$, from this posterior to approximate the marginal likelihood. Recall that the marginal likelihood was defined $f(\boldsymbol{y}|m) \doteq \int_{\Theta_m} f(\boldsymbol{y}|\theta, m)p(\theta|m)\mathrm{d}\theta$; importantly, also that the marginal likelihood is the normalising constant of the (parameter) posterior distribution.

Given some probability density function g, whose support is contained within the support of the posterior, we have the following identity

$$\int_{\Theta_m} g(\theta)\frac{\pi(\theta|\boldsymbol{y}, m)f(\boldsymbol{y}|m)}{f(\boldsymbol{y}|\theta, m)p(\theta|m)}\mathrm{d}\theta = \int_{\Theta_m} g(\theta)\frac{p(\theta, \boldsymbol{y}|m)}{p(\theta, \boldsymbol{y}|m)}\mathrm{d}\theta = 1.$$

We may then use this to estimate the marginal likelihood by noting that if we divide both sides by $f(\boldsymbol{y}|m)$, we have

$$\frac{1}{f(\boldsymbol{y}|m)} = \int_{\Theta_m} \pi(\theta|\boldsymbol{y}, m)\frac{g(\theta)}{f(\boldsymbol{y}|\theta, m)p(\theta|m)}\mathrm{d}\theta = \mathbb{E}_\pi\left[\frac{g(\theta)}{f(\boldsymbol{y}|\theta, m)p(\theta|m)}\middle|\boldsymbol{y}, m\right].$$

Using the sequence of MCMC samples of the parameter $(\theta^{(i)})_{i=1}^n$ from the posterior distribution, we may then estimate the RHS of the above using

$$\widehat{f(\boldsymbol{y}|m)} = \left[\frac{1}{n}\sum_{i=1}^{n}\frac{g(\theta^{(i)})}{f(\boldsymbol{y}|\theta^{(i)}, m)p(\theta^{(i)}|m)}\right]^{-1}; \tag{2.21}$$

an extension of the harmonic mean estimator, see Newton and Raftery (1994) and Gelfand and Dey (1994). Note that g should be chosen to have lighter tails than the posterior distribution (Congdon, 2006). Care must be taken when using this estimator as its variance can be infinite.

Another important property to consider here is that this estimator is not unbiased. As such, we will not be able to use this estimator within a pseudo-marginal algorithm, such as the methods we propose in Part II.

In general, MCMC algorithms are often the first choice of Monte Carlo methods used when implementing Bayesian model comparisons; as such, it is often worthwhile to consider extensions of MCMC and other Monte Carlo methods that may have unique advantages.

For example Reversible Jump MCMC (RJMCMC), proposed by Green (1995), is a widely used algorithm that is used for simulations where the dimension of the parameter space is not fixed. That is, in the context of Bayesian model selection, rather than simulating from each model as per the standard approach described above; RJMCMC proposes using a single algorithm that can *jump* between different models. Essentially, it can be used for inference of the full posterior $\pi(\theta, m|\boldsymbol{y})$, defined on the space $\cup_{m\in\mathcal{M}}(\{m\}\times\Theta_m)$. Zhou et al. (2016) investigated the use of the RJMCMC method in the PET setting, concluding that it was not particularly effective when compared to using their proposed automatic sequential Monte Carlo method.

In the immediate sequel, we will briefly explore an interesting extension of the MH algorithm which uses estimators of the target density — a class of MCMC methods called pseudo-marginal MCMC. Interestingly, see Andrieu and Roberts (2009, Section 7) for an investigation of how pseudo-marginal algorithms can be used with RJMCMC.

2.7 Pseudo-Marginal MCMC Methods

As alluded to before, we may extend the use of MCMC chains for target distributions where it may not be possible to even evaluate the distribution point-wise. Pseudo-marginal MH (Metropolis-Hastings), introduced by Beaumont (2003), is the based on the idea of using an approximation of the target density to compute the acceptance ratio of the MH algorithms. Remarkably, given the foundational assumption (and other mild conditions, see Andrieu and Roberts (2009)) that the estimates of the (usually marginal) density is unbiased, the Markov chain generated using a pseudo-marginal MH algorithm will have invariant distributions with marginals equivalent to intractable target density. Andrieu and Roberts (2009) , who studied the theoretical properties of these pseudo-marginal algorithms, showed that MH kernels that use these approximations have good convergent properties.

In the context of this thesis, this class of algorithms is particularly useful when we wish to study the properties of intractable model posterior distributions of the whole image data set. More precisely, a novel extension of this flexible method will allow us to characterise a somewhat complex spatial model constructed, in Section 5.3, to address the problem at hand. In view of this, this section introduces some of the basic ideas of the pseudo-marginal MCMC method.

2.7.1 Pseudo-marginal Metropolis-Hastings

Many intractable target densities, including those of interest here, can be expressed in terms of marginals of tractable densities: integrals that cannot be evaluated analytically. An obvious example is when we wish to use marginal likelihoods $f(\boldsymbol{y}|m) \doteq \int_{\Theta_m} f(\boldsymbol{y}|\theta, m)p(\theta|m)\mathrm{d}\theta$.

More formally, let the target density be of the form

$$\mu(x) = \int_{\Theta} \mu(x, \theta)\mathrm{d}\theta,$$

where the integrals cannot be solved analytically; and $\theta \in \Theta$ may be considered a latent (or nuisance) variable that is not of current interest.

Let $\widehat{\mu}(x)$ be an estimator of $\mu(x)$ and suppose that $\widehat{\mu}(x)$ is unbiased, for all $x \in \mathcal{X}$. That is, if we let $g(\cdot|x)$ denote the density of $\widehat{\mu}(x)$, then we have that $\mathbb{E}_g[\widehat{\mu}(x)] = \mu(x)$, where the expectation is taken with respect to g. Note, in general, g need not be known. The estimator $\widehat{\mu}(x)$ is random, and its variance will play an important part in producing accurate results — we postpone further discussion until Section 5.4.

The simplest example of such an approximation is the IS normalising constant estimator, $\widehat{Z}^{(\mathrm{IS})}$. Without loss of generality, for a fixed $x \in \mathcal{X}$, consider an IS estimator, denoted $\widehat{Z}^{(\mathrm{IS})}(x)$, approximating the marginal $\mu(x)$. We momentarily suppress the superscript for ease of presentation. That is, given a set of n IS samples of $\theta \in \Theta$, denoted $\{\theta^{(i)}\}_{i=1}^n$, generated from the appropriate IS proposal density denoted $\nu(\cdot|x)$; the unbiased IS estimator for marginal density $\mu(x)$ is simply the average of the unnormalised weights:

$$\widehat{\mu}(x) = \widehat{Z}(x) \doteq \frac{1}{n} \sum_{i=1}^{n} \frac{\mu(x, \theta^{(i)})}{\nu(\theta^{(i)}|x)}.$$

It follows from Proposition 2.5.1, that $\widehat{Z}(x)$ is a unbiased estimator of $\mu(x)$ for all $x \in \mathcal{X}$.

Beaumont (2003) first showed that such IS approximations could be successfully used to approximate intractable target densities in computation of acceptance probability of the MH updates. Alternatively, unbiased likelihood estimators are commonly constructed using sequential Monte Carlo methods, or particle filters, as in Andrieu et al. (2010). We discuss this computationally intensive method in more detail in Section 3.4.

In some cases, such as when estimating the marginal likelihood, the target density will be a normalising constant. As such, henceforth we use the symbol $\widehat{Z}$ to refer to estimators of these intractable marginal densities. Importantly, note that $\widehat{Z}(x)$ will be used to refer to any generic marginal estimator for $x \in \mathcal{X}$ — not just the IS normalising constant estimator.

We postpone proving that using an unbiased estimates of the target density produces a MH Markov chain with the correct invariant distribution until Section 5.5; where formal justification of the proposed approach of this is thesis is also discussed. Both arguments are essentially the same and relatively straightforward. However, studying the details of the pseudo-marginal MH algorithm will both be informative and provide insight into why this claim is true.

2.7.2 The GIMH Pseudo-marginal Algorithm

Andrieu and Roberts (2009) present the theoretical convergent properties of two approaches to using unbiased estimates in the MH algorithm:

- The Monte Carlo within Metropolis (MCWM) algorithm, where at each iteration estimates of the target density for both the current state and the proposed state are sampled;
- The Grouped independence Metropolis Hastings (GIMH) algorithm, where at each iteration only estimates of the target density at the proposed state is refreshed, but the estimate of the density of the current state from the previous iteration is used.

In what follows, we will only consider GIMH. Using the simpler reformulation given by Andrieu and Vihola (2015), the pseudo-code description of GIMH is presented below in **Algorithm** 2. The

symbol $\widehat{\mu}(x) = \widehat{Z}(x) \approx \mu(x)$ is momentarily used here to emphasis the similarity to the "marginal" MH algorithm, see **Algorithm** 1.

Algorithm 2 The GIMH Pseudo-Marginal Algorithm

1. Given $x^{(i-1)}$ and unbiased estimate $\widehat{\mu}_{(i-1)}(x^{(i-1)})$.
2. Sample:
   (a) $x^*|x^{(i-1)} \sim \nu(x^{(i-1)}, \cdot)$;
   (b) $\widehat{\mu}_*(x^*)|x^* \sim g(\cdot|x^*)$.
3. Compute
$$\widehat{r}(x^{(i-1)}, x^*) \doteq \frac{\widehat{\mu}_*(x^*)\nu(x^*, x^{(i-1)})}{\widehat{\mu}_{(i-1)}(x^{(i-1)})\nu(x^{(i-1)}, x^*)}.$$
4. With probability $\min\{1, \hat{r}\}$ let:
$$x^{(i)} = x^* \text{ and } \widehat{\mu}_{(i)}(x^{(i)}) = \widehat{\mu}_*(x^*);$$
   otherwise
$$x^{(i)} = x^{(i-1)} \text{ and } \widehat{\mu}_{(i)}(x^{(i)}) = \widehat{\mu}_{(i-1)}(x^{(i-1)}).$$

Here, we have used subscripts (i) and $*$ for $\widehat{\mu}$ to emphasis precisely where the estimate for the intractable density has been (re-)estimated. That is, the estimate of the marginal from the previous iteration $\widehat{\mu}_{(i)}(x^{(i)})$ and a new estimate for the proposed state $\widehat{\mu}_*(x^*)$ are used to compute the acceptance ratio. This is in contrast to the MCWM algorithm, where instead $\widehat{\mu}_{(i)}(x^{(i)})$ is discarded, and a new marginal estimate for the current state $\widehat{\mu}_*(x^{(i)})$ is sampled. Since we are only interested in the GIMH algorithm here, henceforth these subscripts will be suppressed for clarity of notation.

Within the context of this Monte Carlo algorithm, for each iteration i, with some abuse of notation denote $(\widehat{Z})^{(i)} \doteq \widehat{\mu}(x^{(i)})$ and $(\widehat{Z})^* \doteq \widehat{\mu}(x^*)$. As Andrieu and Roberts (2009) points out, the sequence of random variables $\{X^{(i)}\}_{i=1}^n$ generated according to Algorithm 2 is not a Markov chain. Instead, the sequence $\{X^{(i)}, (\widehat{Z})^{(i)}\}_{i=1}^n$ on an extended space, which includes the marginal estimate, is a Markov chain.

In other words, the above acceptance ratio $\widehat{r}$, which we may rewrite

$$\widehat{r} \doteq \frac{(\widehat{Z})^*\nu(x^*, x^{(i)})}{(\widehat{Z})^{(i)}\nu(x^{(i)}, x^*)},$$

is a simplification of the ratio on the extended space:

$$\widehat{R}((x^{(i)}, (\widehat{Z})^{(i)}), (x^*, (\widehat{Z})^*)) \doteq \frac{(\widehat{Z})^* g((\widehat{Z})^*|x^*)\nu(x^{(i)}, x^*)g((\widehat{Z})|x^{(i)})}{(\widehat{Z})^{(i)} g((\widehat{Z})^{(i)}|x^{(i)})\nu(x^*, x)g(\widehat{Z}_x^*|x^*)}.$$

This follows directly from the fact that $\widehat{Z}(x)$ is *distributed and sampled* according to $g(\cdot|x)$. Andrieu and Roberts (2009) showed that (a generalisation of) the MH kernel with such pseudo-marginal acceptance ratio will converge to the exact marginal (target) distribution μ.

Thus far, we have used the symbols $\widehat{\mu}$ and $\widehat{Z}$ interchangeably. However, in general we may need to estimate only part of the target distribution. For example, if targeting the model posterior distribution $\pi(m|\boldsymbol{y}) \propto f(\boldsymbol{y}|m)p(m)$; typically we only need to estimate the marginal likelihood — as the prior is usually known. Thus, following the above notations, given data $\boldsymbol{y}$ intuitively we

may denote the estimates $\widehat{Z}(m) \approx f(m|\boldsymbol{y})$ and the pseudo-marginal target distribution $\widehat{\mu}(m) = \widehat{Z}(m)p(m)$. We will explore such settings in more detail in Part II, Section 5.4.1.

2.8 Monte Carlo Methods for Graphical Models

Ultimately we are interested in modelling spatial data, whose spatial relations can often be modelled using graphs. Here we will very briefly review important Monte Carlo methods used to characterise some simple graphical models. More detailed terms and notations for graphical models will be introduced in Part II, Section 5.1 — since a graphical model is used of the spatial data. Here, we focus on the Monte Carlo methods and so keep graphical notations to the minimum.

The Ising model (and its extension, the Potts model, see Section 5.2) proposed by Ising (1925), is a popular model used when analysing binary images. See Hurn et al. (2003) for a review on using these models for Bayesian image analysis. More formally, consider a graph $G = (V, E)$ with set of vertices or nodes denoted V and set of edges denoted E. Each element of the edge-set E is denoted $\langle u, v\rangle \in E$ for some pair of elements of the node-set $u, v \in V$. Two nodes u, v are said to be connected by the edge $\langle u, v\rangle$ if $\langle u, v\rangle \in E$. We say that the two connected nodes u and v are neighbours, or adjacent, and denote this by the relation $u \sim v$. We will look only at undirected graphs, as such the relation $\sim$ will be symmetric and $\langle u, v\rangle$ is an unordered pair. Given a graph, G, a collection of random variables, $\boldsymbol{X} = (X_v : v \in V)$, indexed by the nodes-set, V, is called a random field on G. We say that $\boldsymbol{X}$ has a Ising distribution if it has probability mass function given by

$$p(\boldsymbol{x}|J, G) \propto \exp\left(J \sum_{v\sim u} \delta_{x_v, x_u}\right);$$

for coupling constant $J > 0$ and δ_{x_v,x_u} is the Kronecker delta notation, i.e. δ_{x_v,x_u} is one if $x_v = x_u$, and zero otherwise. Importantly, each component $X_v \in \mathcal{X}$ is binary and takes only one of two values — the Potts model is a generalisation to a finite (or even infinite) state space $\mathcal{X}$.

There exist efficient samplers for these models, they include most notably Glauber Dynamics Glauber (1963) corresponding to single-site Gibbs updates, and the elegant Swendsen-Wang algorithm (Swendsen and Wang, 1987).

Alternatively, it is standard practice to use a Gibbs sampler (Glauber, 1963; Geman and Geman, 1984) largely due to its simplicity and ease of implementation. Based on the Hammersley-Clifford theorem, this sampler sequentially updates each site(node) from its full conditionals to generate the Monte Carlo samples. The full conditionals are relatively straightforward to compute for the Ising model. To see this: let $\boldsymbol{x}_{-v}$ denote the vector equivalent to $\boldsymbol{x}$ without the $v-$th component x_v; Further, let $\partial(v)$ denote the neighbours of the node v. Then by the law of total probability we have that:

$$p(x_v|\boldsymbol{x}_{-v}) = \frac{p(\boldsymbol{x})}{\sum_{x'\in\mathcal{X}} p(x_v = x', \boldsymbol{x}_{-v})}.$$

The full conditionals can thus be shown to be

$$p(x_v|\boldsymbol{x}_{-v}) = \frac{\exp\left(J\sum_{v\in\partial(u)} \delta_{x_v,x_u}\right)}{\sum_{x'\in\mathcal{X}} \exp\left(J\sum_{u\in\partial(v)} \delta_{x_u,x'}\right)}.$$

Importantly note that, as is typical in Gibbs samplers, the use of full conditionals results in a simple node–wise update schedule — rather than relying on whole configuration proposals. More simply put, only the state of a single node is proposed to be changed at each iteration rather than

the whole image. This has significant simplification for both computation and for implementation, hence the popularity of this sampler for the Ising distribution. This motivates the use of single site update scheme in method that we propose in this thesis, that we will discuss in Part II.

An important consideration when sampling from the Ising distribution, is the presence of phase transition behaviour. The Swendsen-Wang algorithm is an alternative sampler typically used in simple settings, where there is an absence of a "strong external field" i.e. data in this context. In this regime, the Swendsen-Wang algorithm has appealing convergence properties for all values of the coupling strength, J, see Hurn et al. (2003). This is useful as this model exhibits phase transition behaviour at certain values of the coupling constant. However, this excellent performance deteriorates markedly in the presence of an external field (Higdon, 1998). We discuss critical values of the coupling constant in more detail in Section 5.2.1.

More broadly speaking, there exists considerable literature on approaches for spatial statistics (see, for example, Cressie (1992)) and spatial modelling in Bayesian data analysis (see, for example, Gelfand et al. (2010) and Banerjee et al. (2014)). However, many of these approaches may need to be application- or domain-specific — for instance Bezener et al. (2018), who propose a variable selection method that incorporate spatial information in MRI (Magnetic Resonance Imaging) data. In regards to PET imaging studies, analysis that incorporates spatial information are largely absent. This is largely due to the recency of the PET technology. One exception is a study by Zhou et al. (2002) — we discuss these studies further in Section 4.2.4. Bayesian approaches for incorporating spatial dependence for model selection and inference in specialised settings, such as those encountered in PET images, are yet to be explored to the same extent; In fact, studies by Zhou et al. (2013) and Zhou et al. (2016) were some of the first attempts at using this powerful framework for statistical analysis of PET images. In particular, the success of Zhou et al. (2016) in using more complex and sophisticated methods gives motivation for further exploration.

2.9 Summary

This chapter began by describing and formally defining the task of statistical model selection. Then, we briefly explored useful and readily applicable methods, based on the frequentist approach, that have proven to be useful when analysing complex data sets such as PET images. We also looked at Bayesian model selection, where an important central problem was the computation of the marginal likelihood. As is usually the case in the Bayesian framework, we must turn to Monte Carlo methods to address this problem. There are many different Monte Carlo methods that could be used. In exploring some of the important Monte Carlo methods, we saw that these numerical computational methods can also allow us to characterise the posterior model distribution itself. This will be particularly important when we construct and use graphical models with spatial dependence, that can be used to accurately infer complex spatial structures seen in measured data sets such as PET images.

In fact, in Part II we will use many of the above described methods and their properties to address the problem of incorporating spatial dependence in model selection. Importantly, pseudo-marginal algorithms allow us to relax the requirement of knowing the target distribution point-wise. Instead, the density value could also be estimated using a Monte Carlo method. In particular, given that it is standard to estimate the marginal likelihood using these methods; we can use such existing estimators within the pseudo-marginal algorithm to study the model posterior distribution for a spatial model of the whole image data set.

More precisely, the marginal likelihood of the model of each sub-unit of spatial data (for the case of images, this is a pixel), could potentially be used to approximate the model posterior distribution for the whole data set. We seek to use this strategy in the proposed approach, as we will discuss in the sequel. In order to so, we first need a robust method to estimate the intractable marginal likelihoods, which can be applied reliably in realistic, somewhat difficult settings such as PET images. The combination of IS and MCMC gives us the mathematical concepts to now describe sequential Monte Carlo methods. For instance, Zhou et al. (2016) used the normalising constant estimator of this method to perform non-spatial model comparison for PET data. Motivated by this, the sequential Monte Carlo estimator will play the "linchpin" role in the algorithm we propose in this thesis. In the next chapter we describe and study this interesting, popular technique.

Chapter 3

Sequential Monte Carlo

> *. . . but Thou hast ordered all things in measure and number and weight.*
>
> — Wisdom of Solomon 11:20, *King James Bible*

> *God created everything by number, weight and measure.*
>
> — Isaac Newton

Although Sequential Monte Carlo (SMC) methods have been available for many years, see for example Gordon et al. (1993), they are not commonly used in an Bayesian model comparison setting. Among other considerations, a primary reason for this is that when regarding computational cost, in some cases, it is more efficient to opt for an MCMC algorithm rather than use an SMC approach. However, as argued by Zhou et al. (2016), there are many advantages of using SMC over MCMC in this setting. In fact, as discussed in Aston and Johansen (2015), SMC-based approaches have proven to be of value when analysing a *wide* range of neuroimaging modalities — that is, not just for PET images.

The first, obvious and general advantage of SMC over MCMC is that SMC is more suited to the parallelization than conventional MCMC. Given that both of these methods must be implemented on a computer, and noting the current trend towards parallel computing, SMC-based approaches may prove to be more in-line with the future trajectory of technology. The second advantage is that SMC is more robust for simulating from complex distributions. Particularly for high-dimensional distributions, designing efficient MCMC algorithms can be considerably difficult. Another factor in favour of SMC is that we are able to obtain unbiased estimator of the normalising constant of the target distribution; This is especially important as we discuss in more detail in Section 5.4.1, where we seek to use pseudo-marginal algorithms to incorporate spatial information.

This chapter describes formally and in detail the main ideas behind generic SMC methods. The approach presented will be based upon the framework of Del Moral et al. (2006), though use of SMC dates as far back as Del Moral (1996) or Gordon et al. (1993). Most of these works describe SMC in its original use as a solution to the particle filtering problem, see Doucet and Johansen (2011). Naesseth et al. (2019) provide a recent, concise monogram on using SMC methods; For a more thorough treatment, see Chopin and Papaspiliopoulos (2020).

Intuitively, SMC can be thought of as an natural extension of IS, as discussed in Section 2.5, to

more complex settings. Essentially, this robust method arises naturally when seeking to apply IS to characterise a sequence of distributions rather than just a single one. Computational considerations and steps that account for sample degeneration due to sequential sampling lead to the standard SMC method. Further generalisations, that allow us to use the method for a greater range of sample domains, lead to the variant of the SMC sampler that will be used within the proposed method.

In view of this, we begin this chapter by formally describing the extension the IS sampler to a sequence of target distributions, in Section 3.1. Then, we look at how SMC methods, which were originally developed to solve particle filtering problems, can be used in Bayesian inference via SMC samplers in Section 3.2. Finally, in Section 3.4 we briefly look at the useful properties of the SMC normalising constant estimator.

In brief, SMC algorithms generate a large collections of weighted samples suitable for approximating each of a sequence of target distributions, denoted $\{\mu_t\}_{t\geq 1}$, defined on state spaces $\{\mathcal{X}_t\}_{t\geq 1}$. Denote $\boldsymbol{X}_{1:t} \doteq (X_1, \ldots, X_t)$ to be the random variable taking values in $\mathcal{X}_t$ — note, as such, $\mathcal{X}_t$ is a sequence of space with increasing dimensions. We will refer to the symbol t as the time index, though it will not have any relation to "real time".

The basic idea of SMC methods is to produce at time t a weighted sample of size N, denoted $\{W_t^{(i)}, X_{1:t}^{(i)}\}_{i=1}^N$; such that the empirical distribution of these weighted samples is a good approximation of μ_t. It is important to note that, with reference to the SMC method, we use the upper case letter N to refer to the SMC sample size rather than the lower case letter n, as we have done so far when describing Monte Carlo methods in the previous chapter. This convention is used to avoid confusion later on, as the SMC sampler estimator will be used within a pseudo-marginal MH algorithm (which itself will be used to generate a sample of size n) in the approach that we are proposing in this thesis. In particular, the sample size N of the SMC normalising constant estimator will affect the variance, as we will discuss in Section 3.4, and thus be a tuning parameter of the proposed method.

Within the SMC setting the sample points $X_{1:t}^{(i)}$ are called particles and the weights $W_t^{(i)}$ are typically normalised i.e. such that $\sum_{i=1}^N W_t^{(i)} = 1$ with $W_t^{(i)} > 0$ for all $i \in \{1, \ldots, N\}$. The set $\{W_t^{(i)}, X_{1:t}^{(i)}\}$ is sometimes called a particle system. Proceeding to the next time step $t+1$, SMC may be thought as extending the path of these particles, using a combination of *sequential IS* and a variance reduction technique called *resampling*.

Importantly, we will use the SMC method to estimate marginal likelihoods in the form of intractable integrals — which we recall once more is the normalising constant of the parameter posterior $\pi(\theta|\boldsymbol{y}, m)$. As such, we focus here on the SMC normalising constant estimator, which will be discussed in more detail in Section 3.4. Of course, similar to the IS estimator and the harmonic mean estimator described above; through the process of estimating the normalising constant, this approach also characterises the parameter posterior distribution itself. Thus this approach can also simultaneously be used to perform statistical inference. In other words, with the focus of model selection we use SMC to sample latent parameters – this sample them allows us to estimate the integral $\mathbb{E}_{\mu_t}[\varphi] = \int \varphi \mathrm{d}\mu_t$, using the Monte Carlo approximation of μ_t, as previously discussed in Section 2.4. Being able to perform statistical inference of these parameters will be a very useful by-product of this powerful method. We begin with a brief discussion of the main concepts used to describe SMC methods.

3.1 Sequential Importance Resampling

Sequential importance sampling (SIS) is a generalisation of IS: from sampling from a single target distribution with density $\mu = \frac{\gamma}{Z}$, known point-wise and up to some unknown normalising constant Z; To a sequence of target densities with densities

$$\left\{\mu_t = \frac{\gamma_t}{Z_t} : \gamma_t \text{ known point-wise}\right\}_{t\geq 1},$$

where once more the normalising constants Z_t may be unknown. In fact, we seek to sample *sequentially* from $\{\mu_t\}_{t\geq 1}$; i.e. first sampling from μ_1, then from μ_2 and so on. Note that an immediate problem we face is that sampling from $\mu_t(x_{1:t})$, for $t = 1, 2, \ldots$, with a conventional sampling scheme, such as IS, using proposal distributions $\{\nu_t(x_{1:t})\}_{t\geq 1}$ would typically have a computational cost that is at least linear in the number of variables t.

SIS addresses this by decomposing the proposal densities into conditional probabilities:

$$\begin{aligned}\nu_t(x_{1:t}) &= \nu_{t-1}(x_{1:t-1})\nu_t(x_t|x_{1:t-1}) \\ &= \nu_1(x_1)\prod_{s=2}^{t}\nu_s(x_s|x_{1:s-1}).\end{aligned}$$

Doing so immediately allows for a simpler sampling scheme at each iteration, that is at $t = 1$ we draw $\{X_1^{(i)}\}_{i=1}^N$ from ν_1 and then $\{X_t^{(i)}\}_{i=1}^N$ from $\nu_t(x_t|X_{1:t-1}^{(i)})$ for all time $t > 1$.

Secondly, we can compute the unnormalised weights, with respect to μ_t, recursively by

$$w_t(x_{1:t}) = w_{t-1}(x_{1:t-1}) \cdot \alpha_t(x_{1:t}), \tag{3.1}$$

where the *incremental importance weight* function is defined as

$$\alpha_t(x_{1:t}) \doteq \frac{\gamma_t(x_{1:t})}{\gamma_{t-1}(x_{1:t-1})\nu_t(x_t|x_{1:t-1})}. \tag{3.2}$$

Trivially, the normalised weights are

$$W_t^{(i)} \doteq \frac{w_t(X_{1:t}^{(i)})}{\sum_{j=1}^N w_t(X_{1:t}^{(j)})} \text{ for all } i = 1, \ldots, N.$$

This is particularly useful in implantation, as computing the weights in this manner is simpler, i.e. less prone to error.

At each iteration $t > 1$, we may estimate the target distributions μ_t and its normalising constant Z_t. Straightforwardly, similar to the IS estimator, see Eq.(2.8); At time t, the SIS normalising constant estimator is simply the average of the normalising weights:

$$\widehat{Z}_t^{(\text{SIS})} \doteq \frac{1}{N}\sum_{i=1}^{N} w_t(X_{1:t}^{(i)})$$

For implementation, using the incremental weights is preferable. Thus, we focus here on the unbiased (via Del Moral (2004, Proposition 7.4.1)) estimates of $\frac{Z_t}{Z_{t-1}}$, by first noting that

$$\int_{\mathcal{X}_t} \alpha_t(x_{1:t})\mu_{t-1}(x_{1:t-1})\nu_t(x_t|x_{1:t-1})\mathrm{d}x_{1:t} = \int_{\mathcal{X}_t} \frac{\gamma_t(x_{1:t})\mu_{t-1}(x_{1:t-1})\nu_t(x_t|x_{1:t-1})}{\gamma_{t-1}(x_{1:t-1})q_t(x_t|x_{1:t-1})}\mathrm{d}x_{1:t} = \frac{Z_t}{Z_{t-1}}.$$

Thus, motivating the estimators

$$\widehat{\frac{Z_t}{Z_1}}^{(\mathrm{SIS})} \doteq \prod_{p=2}^{t} \widehat{\frac{Z_p}{Z_{p-1}}}^{(\mathrm{SIS})} = \prod_{s=2}^{t} \sum_{i=1}^{N} W_{s-1}^{(i)} \alpha_s(X_{1:s}^{(i)}). \tag{3.3}$$

Typically, the initial distribution μ_1 is known (for example the prior, see Section 3.3).

Resampling

An important consideration, when using SIS, is that as t gets larger the discrepancy between ν_t and μ_t becomes larger. This causes the weights to become concentrated around a few particles, causing the SIS approach to fail — in extreme instances, distributions being approximated by a single particle with weight of $W_t^{(1)} = 1$. One way to avoid is by resampling, the second key step of SMC methods.

The basis of resampling is that given a sample and associated weights, we generate a new particle system, denoted $\{\overline{W}_t^{(i)}, \overline{X}_{1:t}^{i}\}_{i=1}^{N}$, such that the expectation of the estimate does not change; or in other words, such that

$$\mathbb{E}\left[\sum_{i=1}^{N} \overline{W}_t^{(i)} \varphi_t(\overline{X}_{1:t}^{(i)})\right] = \mathbb{E}\left[\sum_{i=1}^{N} W_t^{(i)} \varphi_t(X_{1:t}^{(i)})\right], \tag{3.4}$$

for measurable function γ_t. The rationale for resampling is to replicate particles with larger weights and discard those with smaller weights. Importantly, this property plays an essential role in the proof of the unbiasedness of the normalising constant estimator; see Proposition 3.4.1, Section 3.4.

A simple and popular resampling scheme is multinomial resampling, which works as follows. Given a particle system $\{W_t^{(i)}, X_{1:t}^{i}\}_{i=1}^{N}$ a random sample of non-negative integers $\{N_t^{(i)}\}_{i=1}^{N}$ is generated from a multinomial distribution of size N and weights $(W_t^{(1)}, \ldots, W_t^{(N)})$. These are integers are such that $\sum_{i=1}^{N} N_t^{(i)} = N$. Each particle $X_{1:t}^{(i)}$ is then replicated $N_t^{(i)}$ times, for all $i \in \{1, \ldots, N\}$, and the new weights are set to $\overline{W}_t^{(i)} = \frac{1}{N}$.

Specifically, at time t, with resampled particles $\left\{\overline{W}_t^{(i)} = \frac{1}{N}, \overline{X}_{t-1}^{(i)}\right\}_{i=1}^{N}$: the proposal $X_t^{(i)}$ is propagated from the distribution $\nu_t(\cdot|\overline{X}_{1:t-1}^{(i)})$; Then, each particle $X_{1:t}^{(i)}$ is formed by concatenating to the *resampled particle* $X_{1:t}^{(i)} \doteq (\overline{X}_{1:t-1}^{(i)}, X_t^{(i)})$. Importantly, doing so means that low weight particles may be discarded. The weight after this sampling step is calculated as before, using the incremental weights, as in the SIS scheme. Formally, this is written $W_t^{(i)} \propto \overline{W}_{t-1}^{(i)} \alpha_t(X_{1:t}^{(i)})$.

There are many other resampling scheme possible, Doucet and Johansen (2011, Section 3.4) provides a brief and concise survey, alternatively cf. Gerber et al. (2019) and Douc et al. (2005).

It is important to note that although resampling is particularly useful if we are interested in later "time marginals", i.e. $\mu_t(x_t)$; it will not give better approximation of the joint distribution of $X_{1:t}$. This is because we discard increasing many of the N points used to approximate X_1, say. However, resampling is useful in the setting that we are interested in, similarly for filtering, where we are interested in the final time marginal and its normalising constant.

SIS with resampling at every step is called Sequential Importance Resampling (SIR). In which case, the normalising constant estimators given in Eq.(3.3) is simpler; where all normalised weights $W_t^{(i)}$ are equivalent to $1/N$, for all $t = 1, 2, \ldots$ and $i = 1, \ldots, N$. However, it is not always necessary to

resample for every iteration t. Recall that the motivation for resampling was due to the discrepancy between the target and proposal distribution being too large — if this is not the case within the current iteration, it would be (computationally) wasteful to do so. Instead, it would be more effective to adaptively resample using a scheme based on the amount of discrepancy.

A principled way to measure the degeneracy of sequential weighted samples is to monitor the variability of the weights. As aforementioned, we can assess this using the ESS as a criterion, see for example Liu (2001). Recall the definition of ESS, from Eq.(2.14) in Section 2.5.2, and re-define it for the SMC context here, for iteration $t > 1$,

$$\widehat{\mathrm{ESS}}_t \doteq \frac{1}{\sum_{i=1}^{N} (W_t^{(i)})^2}; \tag{3.5}$$

Adaptive resampling can be achieved as follows: $\widehat{\mathrm{ESS}}_t$ takes value between 1 and N, and so resampling is done at iterations where it is below some pre-specified threshold N^* (typically $N^* = N/2$ is used). When implementing adaptive resampling schemes, it is important to keep a record of when resampling is done, as the correct weight values will be needed when computing the normalising constant estimator.

Following Doucet and Johansen (2011), this can be done through the use of the set of auxiliary weights denoted by overloading $\overline{W}_{t-1}^{(i)}$. More precisely: After the propagation step, calculate $\alpha_t(X_{1:t}^{(i)})$ and the weight $W_t^{(i)} \propto \overline{W}_{t-1}^{(i)} \alpha_t(X_{1:t}^{(i)})$.; If the resampling criterion is met, then resample to obtain N equally weighted particles and the weight is set to $\overline{W}_t^{(i)} = \frac{1}{N}$; Otherwise, set $\overline{W}_t^{(i)} = W_t^{(i)}$. We then have the normalising constant estimators for this adaptive regime:

$$\widehat{\frac{Z_t}{Z_{t-1}}}^{(\mathrm{SMC})} \doteq \sum_{i=1}^{N} \overline{W}_{t-1}^{(i)} \alpha_t(X_{1:t}^{(i)}) \doteq \begin{cases} \frac{1}{N} \sum_{i=1}^{N} \alpha_t(X_{1:t}^{(i)}) & \text{if resampling occurred at time } t; \\ \sum_{i=1}^{N} W_{t-1}^{(i)} \alpha_t(X_{1:t}^{(i)}) & \text{o/w.} \end{cases} \tag{3.6}$$

A pseudo-code description of SMC with adaptive resampling algorithm is provided below in Algorithm 3, Section 3.4 (note, this description uses Feynman-Kac models, also introduced in said subsection). Alternatively, see Doucet and Johansen (2011, Section 3.5) for a concise pseudo-code description of SMC with adaptive resampling.

As we will discuss in Section 3.3.1 , the $\widehat{\mathrm{ESS}}_t$ and other similar quantities, based on similar principles, could also be used to provide an automatic method for specifying an annealing scheme — used to specify a sequence of target distributions, as we will detail immediately below. For instance, see Jasra et al. (2011), who suggest basing the scheme on the condition of regular rate of $\widehat{\mathrm{ESS}}_t$ decay between adjacent distributions.

3.2 SMC Samplers

Consider the following, SIR in the standard case allows us to sample from a sequence of distributions over sample spaces of increasing dimensions. In contrast, for the case of interest (i.e. applications for PET data) the dimensions of the parameters are fixed, given the model order. More precisely, we seek to approximate the marginal likelihood thus want to target the parameter posterior distribution $\pi(\theta|\boldsymbol{y}, m)$, for a given model order. Therefore, we use the SMC method to

sample from an *annealing scheme*, to be discussed in Section 3.3. That is, instead of sampling from the sequence of densities $\{\mu_t(x_{1:t})\}_{t\geq 1}$, we wish to sample from sequence of (marginal) densities $\{\mu_t(x_t)\}_{t\geq 1}$.

An approach proposed by Del Moral et al. (2006), called *SMC samplers*, addresses this issue — giving us a method that allows for sampling from a much greater range of target distributions. The important property of SMC samplers is that it allows for target distributions defined on the same sample space (i.e. the dimension is fixed). For example, it is possible to sample sequentially from distributions where μ_t is defined on $\mathcal{X}$ and μ_{t+1} is also defined $\mathcal{X}$. In other words, the sequence of sample spaces do not need to increase in dimensions. Within the context of this subsection, for simplicity we assume no resampling occurs. SMC samplers can be formally described as follows:

Firstly, for the moment suppose that we wish to sample from $\{\mu_t(x_{1:t})\}_{t\geq 1}$; and note that in practice it can be very difficult to select appropriate proposal distributions, $\{\nu_t\}_{t\geq 1}$, within the SIS approach. Accordingly, (Del Moral et al., 2006) proposes instead move the particles using "local MCMC moves".

More formally, assume that μ_1 (the target density at time $t = 1$) is easy to approximate using IS. For instance, if μ_1 is tractable we simply sample from the proposal $\nu_1 \equiv \mu_1$ — in the case of a posterior distribution via an annealing scheme, see Section 3.3, μ_1 would be the prior distribution. Then, at time $t-1$ we will have a set of weighted samples $\{W_{t-1}^{(i)}, X_{1:t-1}^{(i)}\}$ distributed according to ν_{t-1}. At time t, the particles are carried forward via local MCMC moves; that is, the path of each particle is extended with some Markov kernel denoted K_t. For simplicity, we assume that the associated density of K_t, with reference to some base measure, exists and denote it $K_t(x, x')$.

Consequently, it is clear that, at time t, the set of particles $\{X_{1:t}^{(i)}\}_{i=1}^N$ will be distributed according to a (joint) proposal density given by

$$\nu_t(x_{1:t}) = \nu_1(x_1) \prod_{s=2}^{t} K_s(x_{s-1}, x_s), \tag{3.7}$$

where ν_1 is the density of the initial distribution of the particles. The importance weights can then be computed by,

$$w_t(x_{1:t}) = \frac{\gamma_t(x_{1:t})}{\nu_1(x_1)\prod_{s=2}^{t} K_s(x_{s-1}, x_s)},$$

where we recall that γ is unnormalised version of the target distribution μ. It follows that if we wish to characterise the target $\mu_t(x_t)$ in the manner described above (i.e. using MCMC moves), we would eventually need to compute the following (marginal) importance weights:

$$w_t(x_t) = \frac{\gamma_t(x_t)}{\int_{\mathcal{X}^{s-1}} \nu_1(x_1)\prod_{s=2}^{t} K_s(x_{s-1}, x_s)\mathrm{d}x_{1:s-1}}.$$

That is,

$$\nu_t(x_t) = \int_{\mathcal{X}^{s-1}} \nu_1(x_1) \prod_{s=2}^{t} K_s(x_{s-1}, x_s)\mathrm{d}x_{1:s-1},$$

must be known point-wise in order to compute the weights. This is generally not possible for large t — it involves a possibly intractable high-dimensional integration with respect to $x_{1:t-1}$.

As shown by Del Moral et al. (2006), this can be avoided through the use of an auxiliary variable

technique. More specifically, SMC samplers use these local MCMC moves within the SIS frame work by performing iterative IS between the proposal distribution $\nu_t(x_{1:t})$ and a carefully constructed auxiliary distribution. The auxiliary distribution is defined such that the density is

$$\widetilde{\mu}_t(x_{1:t}) = \frac{\widetilde{\gamma}_t(x_t)}{\widetilde{Z}_t};$$

where

$$\widetilde{\gamma}_t(x_{1:t}) \doteq \gamma_t(x_t) \prod_{s=1}^{t-1} L_s(x_{s+1}, x_s).$$

Here, the Markov kernels L_{t-1}, density denoted $L_{t-1}(x_t, x_{t-1})$, are termed the artificial backward kernels. These backward kernels formally arbitrary but influences the estimator variance Del Moral et al. (2006). The important property of $\tilde{\mu}_t$, to note here, is that it has been constructed to admit $\mu_t(x_t)$ as a marginal density.

Thus, the new (overloaded) weights can be recursively computed using $\tilde{\gamma}_t$ in Eq.(3.1) and Eq.(3.2), given by:

$$\begin{aligned} w_t(x_{1:t}) &\doteq \frac{\widetilde{\gamma}_t(x_{1:t})}{\nu_t(x_{1:t})} \\ &= \frac{\gamma_t(x_t) \prod_{s=1}^{t-1} L_s(x_{s+1}, x_s)}{\nu_1(x_1) \prod_{k=1}^{t} K_k(x_{k-1}, x_k)} \\ &= w_{t-1}(x_{1:t-1}) \widetilde{w}_t(x_{t-1}, x_t); \end{aligned}$$

Where the *unnormalised incremental weights* are calculated

$$\widetilde{w}_t(x_{t-1}, x_t) \doteq \frac{\gamma_t(x_t) L_{t-1}(x_t, x_{t-1})}{\gamma_{t-1}(x_{t-1}) K_t(x_{t-1}, x_t)}. \tag{3.8}$$

Importantly, note that it is only required that $\gamma_t(x_t)$ is known and not $\nu_t(x_t)$.

Optimal Backward Kernels

As aforementioned, the choice of the backward kernel plays a critical role the variance of the estimator; In fact, Del Moral et al. (2006) further showed that the optimal choice is

$$L_{t-1}^*(x_t, x_{t-1}) \doteq \frac{\nu_{t-1}(x_{t-1}) K_t(x_{t-1}, x_t)}{\nu_t(x_t)}.$$

Unfortunately, as previously mentioned, the marginal densities $\nu_t(x_t)$ is not usually tractable; An alternative is to instead substitute μ_{t-1} for ν_{t-1}, and use

$$\widehat{L}_{t-1}(x_t, x_{t-1}) \doteq \frac{\mu_{t-1}(x_{t-1}) K_t(x_{t-1}, x_t)}{\int_{\mathcal{X}_{t-1}} \mu_{t-1}(x_{t-1}) K_t(x_{t-1}, x_t) \mathrm{d}x_{t-1}}$$

as an approximation for L_{t-1}^*. If we specify the transition kernel K_t to be a μ_t-invariant MCMC kernel and when $\mu_{t-1} \approx \mu_t$ we have that

$$\begin{aligned} \widehat{L}_{t-1}(x_t, x_{t-1}) &= \frac{\mu_t(x_{t-1}) K_t(x_{t-1}, x_t)}{\int_{\mathcal{X}_{t-1}} \mu_t(x_{t-1}) K_t(x_{t-1}, x_t) \mathrm{d}x_{t-1}} \\ &= \frac{\mu_t(x_{t-1}) K_t(x_{t-1}, x_t)}{\mu_t(x_t)}. \end{aligned}$$

To compute the unnormalised incremental weights, we simply substitute $\widehat{L}_{t-1}$ into Eq.(3.8) for the case where we know μ_t only up to a constant and so use γ_t instead; Giving

$$\begin{aligned}\widetilde{w}_t(x_{t-1}, x_t) &= \frac{\gamma_t(x_t)L_{t-1}(x_t, x_{t-1})}{\gamma_{t-1}(x_{t-1})K_t(x_{t-1}, x_t)} \\ &= \frac{\gamma_t(x_{t-1})}{\gamma_{t-1}(x_{t-1})}.\end{aligned}$$

It is noteworthy here that in order to compute the weights at iteration t, we do not need the samples $\{X_t^{(i)}\}_{i=1}^N$ from iteration t. Practically, this means that we can compute the weights before the sampling step (i.e. where the particles are moved according to the kernel K_t). In cases where γ_t (thus μ_t) need be specified, for example in adaptive annealing schemes (see Section 3.3.1); this can also be done before the sampling step of iteration t. Additionally, it is also possible to resample with these weights, if required, before the sampling step and hence retain a greater degree of sampler diversity.

The SMC sampler framework is an ideal place to use AMCMC methods, should they be needed. To see this, note that μ_t-invariant MCMC kernels K_t are used and μ_{t-1}, in many cases, is similar to μ_t (see annealing schemes below.). That is, at time t we have weighted samples that can be used to approximate μ_t and can be used to estimate the empirical variance required to compute optimal scaling, as discussed in Section 2.6.2. This motivates the use of AMCMC kernels: For example, Zhou et al. (2016) successfully used it in settings such as PET data analysis. In this vein, a potential direction of further investigation and development is to use the Barker proposal, as proposed by Livingstone and Zanella (2020).

Straightforwardly, these weight quantities can then be used to adapt the SIS normalising constant estimator Eq.(3.3) to this the SMC samplers approach (Del Moral et al., 2006, Equation 14)):

$$\widehat{\frac{Z_t}{Z_1}}^{(\text{SMC})} \doteq \prod_{s=2}^{t}\sum_{i=1}^{N} W_{s-1}^{(i)}\widetilde{w}_s(x_{s-1:s}^{(i)}), \tag{3.9}$$

where we assume, for simplicity, no resampling has occurred. Note: here and henceforth, we assume that K_t is constructed to be μ_t-invariant, μ_t is absolutely continuous with respect to μ_{t-1} and the associated time-reversal kernel. We detail the appealing properties (towards use within a pseudo-marginal algorithm) of this SMC estimator in Section 3.4.

We complete this subsection, by noting that each particle can be propagated by these MCMC kernels multiple times. That is, at each iteration t, the local MCMC move can used for a MH chain of length bigger than one. We investigate this factor in numerical studies in Section 6.2.2, Part II.

3.3 Sequential Bayesian Inference

It is possible to apply SMC samplers to a Bayesian inference context in a number of ways. An immediate approach would be to introduce each data point one by one to the posterior distribution, this is called data tempering and it is somewhat similar to the approach as particle filters (for which SMC was originally designed). More formally, suppose we have data of the form $\boldsymbol{y} = (y_1, \ldots, y_k)$; We wish to infer some parameter of interest θ using the (parameter) posterior distribution $\pi(\theta|\boldsymbol{y}, m)$ for give model order m. Additionally, we may use the normalising constant estimator to estimate

the marginal likelihood $f(\boldsymbol{y}|m)$ for a given model order $m \in \mathcal{M}$. Thus we construct a sequence of target distributions $\{\mu_t\}_{t=1}^{k}$ with each member specified to be

$$\mu_t(\theta) = \pi(\theta|m, y_1, \ldots, y_t) \text{ for all } t = 1, \ldots, k;$$

to be used in the SMC samplers method as described above. It would also be possible to introduce data in batches; Both variants of data tempering schemes have been shown to have beneficial effects by (Chopin, 2001). However, among many other factors, the order in which and the amount of data introduced will have an effect.

As an alternative to data tempering, Neal (2001) proposes using an annealing schedule as follows: Begin at an easy-to-sample distribution as the first member of the sequence of target distributions μ_1; Then, move the particles through a sequence of artificial intermediate distributions to the distribution of interest, represented by μ_t. For instance, a simple annealing strategy would be to begin sampling from the parameter prior $p(\theta|m)$, for the given model. Next, sweep through a sequence of $T-1$ intermediate distributions (constructed to be similar adjacently), for some integer $T > 1$, to the posterior $\pi(\theta|\boldsymbol{y}, m) \propto f(\boldsymbol{y}|\theta, m)p(\theta|m)$. That is, we can construct for each model $m \in \mathcal{M}$ the sequence $\{\mu_t\}_{t=1}^{T}$ defined

$$\mu_t(\theta) \propto p(\theta|m)f(\boldsymbol{y}|\theta, m)^{\rho(t/T)},$$

with

$$\mu_1(\theta) = p(\theta|m) \quad \text{and} \quad \mu_T(\theta) \propto \pi(\theta|\boldsymbol{y}, m);$$

and the function $\rho(t/T) : [0,1] \to [0,1]$, such that $\rho(1/T) \doteq 0$ and $\rho(1) \doteq 1$, is known as the *annealing scheme*.

The specification of the intermediate distributions via the annealing scheme can be specified in a number of ways, with different levels of sophistication. For example, Zhou et al. (2016) studied different schemes to analyse measured PET data, including the following simpler (non-adaptive) schemes: $\rho(t/T) = t/T$(linear), $\rho(t/T) = (t/T)^5$(prior) and $\rho(t/T) = 1-(1-t/T)^5$(posterior). In particular the scheme $\rho(t/T) = (t/T)^5$ was shown to be very effective in the PET settings. We term this scheme the "Prior 5" annealing scheme and apply it in the numerical studies in Part II. In addition, they also investigated an adaptive scheme based on estimating the conditional ESS, we detail this below in Section 3.3.1.

We complete this subsection by drawing attention to two of the tuning parameters for the SMC samplers approach for Bayesian inference and model selection via annealing schemes: Firstly, as mentioned before the number of particles used N; Second, the total number of intermediate distribution in the annealing scheme T. In particular, it should be noted that even when using an adaptive annealing scheme, T can be roughly specified via different values of the criterion threshold.

3.3.1 Adaptive Annealing Schemes using Conditional ESS

Recall that, a principled way to measure the degeneracy of sequential weighted samples is to monitor the variability of the weights. We can assess this using the ESS as a criterion, which we

recall, from Eq.(3.5); Which we re-write here for clarity:

$$\widehat{\text{ESS}}_t \doteq \frac{1}{\sum_{i=1}^{N}(W_t^{(i)})^2}.$$

As aforementioned, $\widehat{\text{ESS}}$ is typically used in adaptive resampling to decide whether to resample or not, at each iteration (Doucet and Johansen, 2011). However, Jasra et al. (2011) used $\widehat{\text{ESS}}$ to provide an automatic method for specifying an annealing schedule based on the criterion of regular rate of $\widehat{\text{ESS}}$ decay between adjacent distributions.

Specifically, Jasra et al. (2011) argues that that we may use the ESS as a measure of distance (specifically the χ^2-distance, see Chen (2005)) between distributions and propose intermediate distributions in annealing scheme based on this. SMC samplers based on this principle also include algorithms that move the particle system only when resampling happens — this results in essentially in the same context as resampling at every step.

Zhou et al. (2016) proposes, instead, a more general adaptive scheme that allows for better properties when adaptive resampling is employed. To see this, by Eq.(2.12), we have the following

$$\text{ESS}_t = \frac{N}{(1 + Var_{\nu_t}[W_t])} \overset{Eq.(2.12)}{\approx} \frac{N}{\mathbb{E}_{\mu_t}[W_t]};$$

which they approximate using the empirical approximation:

$$\widehat{\text{CESS}}_t = \frac{N}{\sum_{i=1}^{N} W_{t-1}^{(i)} \left(\frac{w_t^{(i)}}{\sum_{j=1}^{N} W_t^{(j)} w_t^j} \right)^2},$$

termed the *conditional* effective sample size (CESS).

More formally, these quantities can be used to adaptively determine an optimal annealing scheme as follows: Within the present setting, denote the annealing scheme to be dependent on t, and at time t denote it ρ_t. Jasra et al. (2011), and Zhou et al. (2016), argue that we seek to make $\rho_t - \rho_{t-1}$ as large as possible whilst ensuring that μ_t and μ_{t-1} are sufficiently similar — thus maximising computational efficiency. The discrepancy between the adjacent μ_t and μ_{t-1} are approximated using estimates of CESS (or ESS); That is, how good an importance proposal would μ_{t-1} be of μ_t. Then, ensure that the CESS between adjacent intermediate distributions in an annealing scheme stay constant throughout the scheme. Intuitively, this means that there are no large jumps at some parts of the sequences and smaller jumps in other parts.

When implementing these principles, set a fixed target denoted CESS^* to be the minimum distance between adjacent distributions. Then, at time t of an adaptive SMC sampler algorithm, use a binary search to determine the next distribution $t+1$ in the annealing scheme. The binary search is to find ρ_{t+1} such that using μ_t as importance proposals of μ_{t+1} gives a CESS of at least CESS^*. CESS^* is usually some threshold of the total sample size — denote specification by $\text{CESS}^* = \tau N$ for $0 < \tau \leq 1$. Henceforth, we call annealing schemes that use this approach "CESS-adaptive annealing schemes".

An important factor to consider when using these adaptive approaches is that the normalising constant estimator will no longer be unbiased. One possible way to address this is to generate the annealing scheme using $\widehat{\text{CESS}}$ via a pilot study; Then, use this predetermined scheme within a non-adaptive scheme SMC sampler. In their numerical studies using annealing schemes on

PET image analysis, Zhou et al. (2016) showed that the Prior 5 scheme $\rho(t/T) = (t/T)^5$, had similar performance (with regards to the variance of the normalising constant estimates) to a CESS-adaptive annealing scheme. As such, where possible a simple non-adaptive scheme could be used.

There exists many further techniques based on the SMC approach. The success of the SMC sampler in the PET setting (Zhou et al., 2016), motivates the use of other SMC sampler based methods. For example, in contrast to the approach proposed in this work, the divide and conquer SMC approach (Lindsten et al., 2017) could also be used to incorporate spatial dependence for PET and other image data sets. This wold serve to be a interesting direction for future investigations.

3.4 A Robust Unbiased Normalising Constant Estimator

We now briefly explore theoretical and asymptotic properties of the SMC methods normalising constant estimators discussed above. Firstly, the simple SIS method generates an estimator with variance given:

$$\mathrm{Var}[\widehat{Z}_t^{(\mathrm{SIS})}] = \frac{Z_t^2}{N}\left(\int \frac{\mu_t^2(x_{1:t})}{q_t(x_{1:t})}\mathrm{d}x_{1:t} - 1\right),$$

note the $1/N$ rate of decrease of the variance. If multinomial resampling is used at every iteration, the estimator $\widehat{Z}^{(\mathrm{SIR})}/Z$ satisfies CLT(Central Limit Theorem). For an elegant proof of this, see Del Moral et al. (2006, Proposition 9.4.1); Alternatively, see also Chopin (2004) or Chopin and Papaspiliopoulos (2020, Proposition 11.2). In particular, the (relative) variance of the estimator $\widehat{Z}_n^{(\mathrm{SIR})}/Z_n$ is given by

$$\frac{1}{N}\left(\int \frac{\mu_1^2(x_1)}{\nu_1(x_1)}\mathrm{d}x_1 - 1 + \sum_{k=2}^{t}\frac{\mu_t^2(x_{1:k})}{\mu_{k-1}(x_{1:k-1})\nu_k(x_k|x_{1:k-1})}\mathrm{d}x_{k-1:k} - 1\right).$$

It is now immediate why N is an important tuning parameter when this sampler is used within a pseudo-marginal algorithm (Doucet et al., 2015). The variance of this normalising constant estimator can be estimated using the weighted particles, see for example Lee and Whiteley (2018) and Du and Guyader (2021). Such variance estimators can naturally be used within pseudo-marginal algorithms, such as the proposed method, to straightforwardly extend to an adaptive setting.

Next, we turn to the important property of unbiasedness. For the simpler case, it is straightforward to extend the proof for the IS estimator to SIS. For the SIR estimator, an elegant proof can be found in Del Moral (2004, Proposition 7.4.1). Alternatively, Naesseth et al. (2019, Section 4.A) provides a proof that is slightly more accessible.

However, it is often the case (as will be in this thesis too) that adaptive resampling is used. The proof for this regime is also relatively straightforward; Albeit, some further abstractions are required. Specifically, the description of the SMC method using the Feynman-Kac models is used — we introduce some notations and terms, towards this end.

Using Markov kernels as proposals for SMC samplers, as discussed above, naturally leads to the abstraction given by Feynman-Kac models. Similar to the intuition of the IS identity, Eq.(2.7), we may interpret SMC algorithms as Monte Carlo approximations of some underlying Feynman-Kac model. Specifically, the Feynman-Kac model is obtained from changes of measure from the

proposal distributions (Chopin and Papaspiliopoulos, 2020, Chapter 5).

More formally, let ν_T be a Markov probability law defined on (a common, (as the setting described in Section 3.2)) state space $\mathcal{X}$, with initial distribution ν_1 and transition kernels denoted $Q_2, \ldots, Q_T$:

$$\int_A \mathrm{d}\nu_T(x_{1:T}) \doteq \int_A \mathrm{d}\nu_1(x_1) \prod_{t=2}^{T} \mathrm{d}Q_{t,x_{t-1}}(x_t), \text{ for all } A \in \mathcal{B}(\mathcal{X})^T;$$

where $Q_{t,x_{t-1}}(\cdot)$ is the shorthand for the measure $Q_t(x_{t-1}, \cdot)$ for all $x_{t-1} \in \mathcal{X}$ and $t = 2, \ldots, T$.

Next, consider a sequence of so-called potential functions, denoted $G_1 : \mathcal{X} \to \mathbb{R}^+$, and $G_t : \mathcal{X}^2 \to \mathbb{R}^+$, for $2 \leq t \leq T$. A sequence, for $1 \leq t \leq T$, of *Feynman-Kac* models is given by probability measures $\mu_t^{(\mathrm{FK})}$ on $(\mathcal{X}^t, \mathcal{B}(\mathcal{X})^t)$, attained as the following changes of measure from proposals ν_t:

$$\int_A \mathrm{d}\mu_t^{(\mathrm{FK})}(x_{1:t}) \doteq \frac{1}{Z_t^{(\mathrm{FK})}} \int_A G_1(x_1) \left\{ \prod_{s=1}^{t} G_s(x_{s-1}, x_s) \right\} \mathrm{d}\nu_t(x_{1:t}) \text{ for all } A \in \mathcal{B}(\mathcal{X})^t;$$

With normalising constant denoted

$$Z_t^{(\mathrm{FK})} \doteq \int_{\mathcal{X}^t} G_1(x_1) \prod_{s=2}^{t} G_s(x_{s-1}, x_s) \mathrm{d}\nu_t(x_{1:t}) = \mathbb{E}_{\nu_t} \left[G_1(X_1) \prod_{s=2}^{t} G_s(X_{s-1}, X_s) \right]. \tag{3.10}$$

Specifically, in the context of this thesis, $Z_T^{(\mathrm{FK})}$is the marginal likelihood.

We are interested in adaptive resampling, as such the genealogy of the particles will prove to be useful. As such, denote $\{\overline{A}_t^{(i)}\}_{i=1}^N$ to be a generalisation of ancestor variables (introduced by Andrieu et al. (2010)), in the adaptive resampling setting. Specifically, these variables represent the index of the parent node at resampling steps, and are equivalent to i, otherwise. Given the set of normalised weights $\{W_t^{(i)}\}_{i=1}^N$, for time t; Let $F(W_t^{(1)}, \ldots, W_t^{(N)})$ denote some discrete distribution of $\{\overline{A}_{t+1}^{(i)}\}_{i=1}^N$; such that the unbiased resampling condition Eq.(3.4) is satisfied. The SMC algorithm, with adaptive resampling, can now be stated in pseudo-code form for this more abstract Feynman-Kac formalisation:

Algorithm 3 The SMC algorithm for a given Feynman-Kac model

1. To initialise:
   (a) Sample
   $$X_1^{(i)} \sim \nu_1.$$
   (b) Compute
   $$W_1^{(i)} \propto G_1(X_1^{(i)}).$$
2. For time $t \geq 2$:
   (a) If the resampling criterion is met (e.g. $\widehat{\mathrm{ESS}}_t \leq \mathrm{ESS}^*$):
      i. Sample
      $$\overline{A}_t^{(i)} \sim F(W_{t-1}^{(1)}, \ldots, W_{t-1}^{(N)}).$$
      ii. Set
      $$\overline{W}_t^{(i)} = \frac{1}{N}.$$
   (b) Otherwise,
      i. Set
      $$\overline{A}_t^{(i)} = i.$$
      ii. Set
      $$\overline{W}_t^{(i)} = W_t^{(i)}.$$
   (c) Propagate $X_t^{(i)} \sim Q_t(X_{t-1}^{(\overline{A}_t^{(i)})}, \cdot)$ and concatenate $X_{1:t}^{(i)} \doteq (X_{t-1}^{\overline{A}_t^{(i)}}, X_t^{(i)})$.
   (d) Calculate
   $$W_t^{(i)} \propto \overline{W}_{t-1}^{(i)} G_t(X_{t-1}^{(i)}, X_t^{(i)}).$$

Using these conventions, we may adapt Eq.(3.6), to give normalising constant estimators of the form:

$$\widehat{\frac{Z_t}{Z_{t-1}}}^{(\mathrm{FK})} \doteq \sum_{i=1}^{N} \overline{W}_{t-1}^{(i)} G_t(X_{t-1}^{(\overline{A}_t^{(i)})}, X_t^{(i)}) \doteq \begin{cases} \frac{1}{N}\sum_{i=1}^{N} G_t(X_{t-1}^{(\overline{A}_t^{(i)})}, X_t^{(i)}) & \text{if resampling occurred at time } t; \\ \sum_{i=1}^{N} W_{t-1}^{(i)} G_t(X_{t-1}^{(i)}, X_t^{(i)}) & \text{o/w.} \end{cases}$$

Straightforwardly, we may use IS to estimate $Z_1^{(\mathrm{FK})}$; it then follows that we obtain unbiased estimates of the normalising constant:

Proposition 3.4.1. The estimator

$$\widehat{Z}_T^{(\mathrm{FK})} = \left(\frac{1}{N}\sum_{i=1}^{N} G_1(X_1^{(i)})\right) \prod_{t=2}^{T} \sum_{i=1}^{N} \overline{W}_t^{(i)} G_t(X_{t-1}^{\overline{A}_t^i}, X_t^i),$$

attained by following **Algorithm** 3, is an unbiased estimator of $Z_t^{(\mathrm{FK})}$, Eq.(3.10), for any time $T \geq 1$.

Proof. We extend the proof for the SIR case, given by Chopin and Papaspiliopoulos (2020, Proposition 16.3), to the adaptive resampling regime.

Take $T = 2$, the general case follows in a similar manner. Let $\mathcal{F}_1 \doteq \sigma(X_1^{(1)}, \ldots, X_1^{(N)})$ be the σ-algebra generated by the random variables $X_1^{(1)}, \ldots, X_1^{(N)}$; $Q_2(x_1, G_2) \doteq \int_{\mathcal{X}} G_2(x_1, x_2) \mathrm{d}Q_{2,x_1}(x_2)$, for each $x_1 \in \mathcal{X}_1$; and $R \subseteq \mathcal{X}$ be the event that resampling occurs.

Then, noting that $W_1^{(i)} = G_1(X_1^{(i)}) / \sum_{j=1}^N G_1(X_1^{(j)})$, we have

$$
\begin{aligned}
&\mathbb{E}[\widehat{Z}_2^{(\mathrm{FK})} | \mathcal{F}_1] \\
&= \left(\frac{1}{N} \sum_{i=1}^N G_1(X_1^{(i)}) \right) \mathbb{E}\left[\left(\mathbf{1}_R \frac{1}{N} \sum_{i=1}^N Q_2(X_1^{(\overline{A}_2^{(i)})}, G_2) \right) + \left(\mathbf{1}_{R^C} \sum_{i=1}^N W_1^{(i)} Q_2(X_1^{(i)}, G_2) \right) \middle| \mathcal{F}_1 \right] \\
&= \left(\frac{1}{N} \sum_{i=1}^N G_1(X_1^{(i)}) \right) \left\{ \mathbb{E}\left[\mathbf{1}_R \frac{1}{N} \sum_{i=1}^N Q_2(X_1^{(\overline{A}_2^{(i)})}, G_2) \middle| \mathcal{F}_1 \right] + \mathbb{E}\left[\mathbf{1}_{R^C} \sum_{i=1}^N W_1^{(i)} Q_2(X_1^{(i)}, G_2) \middle| \mathcal{F}_1 \right] \right\} \\
&= \left(\frac{1}{N} \sum_{i=1}^N G_1(X_1^{(i)}) \right) \left\{ \mathbb{E}\left[\mathbf{1}_R \sum_{i=1}^N W_1^{(i)} Q_2(X_1^{(i)}, G_2) \middle| \mathcal{F}_1 \right] + \mathbb{E}\left[\mathbf{1}_{R^C} \sum_{i=1}^N W_1^{(i)} Q_2(X_1^{(i)}, G_2) \middle| \mathcal{F}_1 \right] \right\} \\
&= \left(\frac{1}{N} \sum_{i=1}^N G_1(X_1^{(i)}) \right) \left\{ \mathbb{E}\left[\sum_{i=1}^N W_1^{(i)} Q_2(X_1^{(i)}, G_2) \middle| \mathcal{F}_1 \right] \right\} \\
&= \left(\frac{1}{N} \sum_{i=1}^N G_1(X_1^{(i)}) \right) \left\{ \sum_{i=1}^N W_1^{(i)} Q_2(X_1^{(i)}, G_2) \right\} \\
&= \left(\frac{1}{N} \sum_{i=1}^N G_1(X_1^{(i)}) \right) \left\{ \frac{\sum_{i=1}^N G_1(X_1^{(i)}) Q_2(X_1^{(i)}, G_2)}{\sum_{j=1}^N G_1(X_1^{(j)})} \right\} \\
&= \frac{1}{N} \sum_{i=1}^N G_1(X_1^{(i)}) Q_2(X_1^{(i)}, G_2).
\end{aligned}
$$

In words: In the first equality we integrated out each $X_2^{(i)}$; Split the integral over the two partitioning events in the second equality; Applied the unbiased resampling condition, see Eq.(3.4), in the third equality; Rejoined the integral in the fourth equality; Applied linearity of expectation for conditionally independent variables in the fifth equality; Finally, expanded each $W_1^{(i)}$ in the second to last equality.

Next, noting that each $X_1^{(i)} \sim \nu_1$, by the tower property of conditional expectation

$$
\mathbb{E}\left[\widehat{Z}_2^{(\mathrm{FK})} \right] = \int_{\mathcal{X}} Q_2(x_1, G_1) G_1(x_1) \mathrm{d}\nu_1 = Z_2^{(\mathrm{FK})}.
$$

□

3.5 Summary

SMC methods have enabled robust statistical inference and model comparison in increasingly complex and challenging contexts. Here, we reviewed the standard SMC framework and its properties, motivated by the interest in using the SMC normalising constant estimator. We also looked at how the standard SMC framework can be adapted, for example using SMC samplers with annealing schemes, for use in an Bayesian inference settings. Additionally, we studied different extensions and refinements when implementing this relatively sophisticated method.

Importantly, the unbiased property of the robust SMC normalising constant estimator means that it is an ideal marginal estimator for use within the pseudo-marginal framework. Specifically, in Chapter 5 we will use the SMC sampler to estimate the marginal likelihood of the model at each sub-unit(or pixel) of spatial data. In particular, SMC allows us to sample from parameter posterior distributions for complex data sets such as PET images. In the immediate sequel we will explore PET data sets — the primary motivating problem and setting of this thesis.

Chapter 4

Compartmental Models for Positron Emission Tomography

Knowledge rests not upon truth alone, but upon error also.

— Carl Jung, *Modern Man in Search of a Soul, 1933.*

Incorporating spatial dependence in the process of analysing large image, and other spatial image–like, data sets can be a difficult problem largely due to the computational requirements. An important example of such data sets include Positron Emission Tomography (PET) images of the brain, where a whole image typically requires analysis of up to 10^6 time series (Hammers et al., 2007). Towards this objective, this chapter will be a review of the mechanics of the PET instrumentation and the important mathematical models used to describe the measured PET signal. Additionally, recent advances in Bayesian inferential methodology for PET images is also briefly described. Primarily, we argue here that current state–of-the–art analysis of such data generally either assumes spatial independence of pixels or performs large-scale aggregation over space to overcome computational restrictions. An important auxiliary goal of this section is to also introduce terms, models and methods (that are primarily non–spatial) to be used within the proposed spatial methodology presented in Chapter 5.

PET (Phelps et al., 2006) is an increasingly important dynamical imaging modality which can provide key insight into a plethora of neural biochemical processes. The theory and methodology behind the acquisition and construction of the 3-D (3-dimensional) dynamical PET images will be briefly discussed in Section 4.1. In studying the mechanisms behind the PET machinery many sources of noisy measurements become apparent. The methods and technique used to address such errors are briefly discussed in Section 4.1.2. Data from a PET study will also be briefly explored in Section 4.1.3. The large remaining proportion of this chapter, Section 4.2, is dedicated to an in depth introduction to one of the most popular class of mathematical models used in PET analysis, namely the compartmental models.

This flexible class of models is based on simplifications of the complex underlying pharmacokinetics of the radioactive tracer used in PET. However, it is often difficult to select, in a principled manner, a single model (out of this class of models) that provides the most accurate description. Indeed, due to the differences in the macro– and micro–structures of the various regions of the brain, the

model used in one sub-unit of the PET image may not be adequate for another. Thus, any good statistical procedure for PET data analysis must not only be accurate, but also computationally efficient in the dual tasks of model selection and parameter inference.

In fact, even when attempting to perform parameter inference, it becomes quickly clear that the macro–parameter of primary inferential interest, the volume of distribution, is dependent on the model order. Similarly, micro-parameters, such as the rate constants, are affected by the model choice due to identifiability problems. These concepts, obstacles and methods in which these obstacles can be addressed is discussed in Section 4.2.4.

There are two, somewhat immediate, problems when attempting to analyse PET data with these statistical and computational objectives. Firstly PET data is limited; this is a consequence of the invasive and radioactive nature of the procedure, resulting in relatively small number of measurements (at each location). In contrast, the data is also somewhat high-dimensional; the number of parameters to be inferred can be as numerous as 10 or more, for example see (Mankoff et al., 1998). Recent works of (Zhou et al., 2013, 2016) (and also (Peng et al., 2008) for parameter estimation) showed that algorithms based on a Bayesian approach can outperform current standard (non-Bayesian) approaches. This motivates further investigation of incorporating spatial dependence within the Bayesian approach.

An underlying, often unstated, assumption amongst almost all current methods for quantitative analysis of PET data is that each sub-unit or voxel (a 3-dimensional pixel) of the image is independent. However, in order to begin considering the general task at hand as an inferential problem on a *spatio-temporal* data set, it is important to avoid this oversimplification of the vastly sophisticated structure of the brain. Though there is still much knowledge to be discovered about these neural structures, it is important to begin to take the underlying relationships in these structures into account in order to provide better quantitative conclusions. In this vein, statistical methods must begin to incorporate spatial information in a meaningful manner.

For clarity and to aid intuition a list of technical or specialised symbols used to describe these models of PET data is provided at the beginning of this document.

4.1 Positron Emission Tomography

Neuroimaging (or brain imaging), a sub-field of neuroscience, is concerned with imaging the structure and function of the brain. Most modern neuroimaging modalities and procedures place emphasis on acquiring images that illustrate the functions of the brain; in particular, to show regions which are related to the task or brain state of interest. Thus, these modalities usually capture sequences of images which dynamically show responses in the brain over the course of a stimulus. Such contemporary functional neuroimaging techniques include: functional Magnetic Resonance Imaging (fMRI), Magnetoencephalography (MEG), Single-Photon Emission Computed Tomography (SPECT) and others. Among these, PET has a unique potential due to its specificity.

PET is an invasive neuroimaging technique that uses the detection of high-energy photons, released as a consequence of the tracer's positron emissions, as a signal to form three-dimensional images. Given that there will be greater radioactivity in regions where there are higher amounts of the tracer, the signal intensity in each voxel will be proportional to the concentration of the tracer. Thus, by taking a sequence of measurements over time, the constructed image represents the time course of the tissue concentration of the tracer *in vivo*. This section is a brief description of the (nuclear) physics of PET imaging; it is a basic description of the physics and mechanics behind

the most simplest type of PET scanner. This is then followed by a discussion on potential errors that arise from the image measurement procedure; as well as how these are corrected for in their raw data form before constructed into a final PET image. For a detailed treatment of the concepts discussed here, see (Phelps et al., 2006).

4.1.1 PET Physics, Instrumentation and Image Acquisition

The PET image acquisition procedure begins with, first, creating the tracer. This is usually a radionuclide (a radioactive isotope of an element) attached to the molecule of scientific interest called the tracee. It is the tracee's biochemical and physiological fate that is the primary objective of a PET study. Creating the radioactive tracer involves the use of a cyclotron and is usually done on-site. The tracer is then introduced into the subject via intravenous injection; the dosage is a trace amount (hence the apt term — tracer). Atrial measurements for the plasma input function, detailed in Section 4.2.1, may begin a short period before the injection. The tracer will then be physiologically distributed into the tissues based upon its biochemical properties.

The atoms of the radionuclide in the tracer contain an excess number of protons, causing the nucleus to be radioactively unstable. This results in the emission of a positron (the antiparticle of an electron, i.e. a positively charged electron) and a neutrino from the atom (specifically, from a proton), in a process known as β^+ decay[1] . Since the tissue is an electron rich environment, the positron will be inevitably absorbed and there will be a positron-electron interaction in a process known as *annihilation*. In annihilation, the mass of the positron and electron is converted to electromagnetic energy, and subsequently will results in its release in the form of two photons. These photons will be emitted simultaneously at almost 180° apart (in opposite directions). Importantly, these two photons will carry the exact same energy of 511 keV (kilo - electron Volts); this is due to the assumption that there is no net momentum and thus arises from the rest mass of the positron-electron pair. These photon exit the subject's body and are then detected by the PET scanner that surround the subject to be imaged. The PET cameras consists of detectors known as scintillator detectors [2] . A scintillator is a material that absorbs photons and coverts it into electrical signals; in PET the electrical signals are then used to construct the image.

The above properties of the annihilation process form the basis around which the PET instrumentation is designed. Firstly, the annihilation releases photons that are highly energetic. This means that they are penetrative enough to leave the body to be detected by the scintillator – it is the photons that are detected, not the positrons (PET is somewhat of a misnomer here). In addition, the fact that these photons will always carry 511 keV of energy is an advantage in its own right; this means that the instruments can be designed and optimised to detect this specific energy level. Note also that the energy will be 511 keV independent of the tracer used. Although this means that the same scanner can be used regardless of the tracer and thus the biochemical process of interest; It also means that multi–tracer studies, involving two or more tracers simultaneously, are not possible. Finally, and perhaps most importantly, the precise geometric relation of the two photons suggests that: once the photons have been detected and localised, the point of annihilation must lie on the line joining the path of these two detected photons. This is known as electronic collimation. Thus, the location [3] of the annihilation (and so the radionuclide, from which the emitted positron travels a microscopically small distance from) can identified. This also

[1] There also exists another decaying process known as *electron capture*, which we will not discuss here.

[2] Other detectors, that use different mechanics, do exists; but, for brevity, we will not discuss them in detail here.

[3] Note there are many ways of doing this. Noting the fact that one of the two photons will be detected before the other one, due to proximity, it becomes clear that it is possible to use this time-difference together with the electronic collimation to localise the annihilation e.g. time-of-flight. An example is shown in Figure 4.1.

means that events (decays) can be detected in many directions simultaneously.

In a typical PET scan procedure, as many as 10^6 to 10^9 of such events will be detected. Through measuring the total radioactivity along lines that pass at many different angles through the object, cross-sectional images of the concentration of the tracer in tissues throughout the body can be generated. This can be done, for example, by arranging the scintillator detectors in a circular manner, in the axial plane, around the subject. This piece of apparatus is called the detector ring. Several of these cross-sectional images, from adjacent measurements of detector rings in the transverse direction, constitutes a full PET image.

Figure 4.1 provides a diagrammatic summary of the described chain of events above; included are two sources of errors. The first source of error is due to the tortuous path of the positron in the tissue before annihilation. The length of this path is called the positron range and the spatial error is defined here to be the perpendicular distance between the radionuclide and the point of annihilation. Essentially this is a hard limit on the spatial resolution of PET imaging, though admittedly it is a microscopically small one. Indeed, the detector resolution is a much bigger limiting factor.

The second source of error is known as non-colinearity: Due to the fact that the positron and electron are not completely at rest when they interact, the angle of the two photons may not be exactly 180° — zero net-momentum is an assumption here. A small net momentum of the particles means that the photons will be emitted at a (normal) distribution of angles around 180°; The error this causes will be proportional to the diameter of the PET scanner. Note once more, this error is usually very small and is not a huge limiting factor.

Although the spatial measurement errors due to positron range and non-colinearity are small enough that they rarely have a significant effect on the quality of the final image; other sources of error in the acquisition process can affect the location and radioactivity measurements drastically and lead to poor image quality and thus quantification. In order to make good statistical statements and quantification, these sources of such noise and the methods use to correct them must be taken into consideration — they are discussed next.

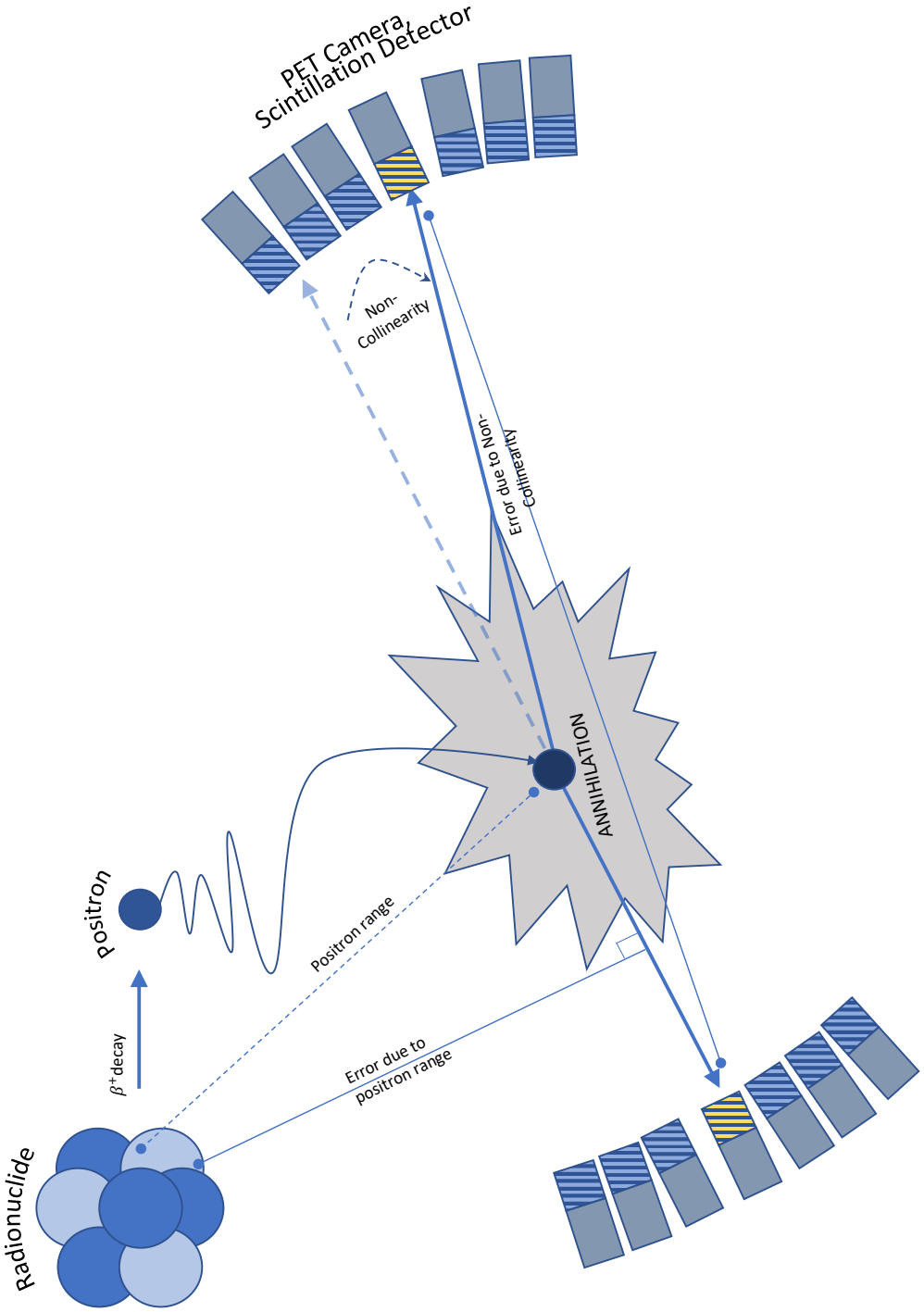


Figure 4.1: The chain of radioactive events that lead to the signal detected by PET scanner. The radioactive decay process of positron emission and subsequent annihilation, due to the positron-electron interaction, results in two 511 keV annihilation photons emitted in opposite directions. Since the emitted positrons rapidly lose their energy in tissue, the distance between the radionuclide and the point of annihilation is usually very small. Two possible sources of errors, positron range and non-colinearity, are also shown — though, usually neither is a limiting factor. Also shown here is part of a circular PET scanner, which then detect the high-energy photons. Note that the distance travelled by the photon going to the left is smaller than that of the right. This time difference, together with electronic collimation, can be used to localise the point of annihilation.

4.1.2 Data Correction

The technique used by PET cameras to record an event, i.e. an isotope decay, is so called "*coincidence*" detection. That is, a pair of annihilation photons must be detected by the pair of detectors in the detector ring. Recall that this, therefore, means that the point of annihilation, thus also the radionuclide, can be localised to within a straight path in space. The path that a pair of detectors detect is called the line of response (LOR). A coincidence is detected on the LOR every time both the pair of detectors detect a photon. A true coincidence is a coincidence that is a result of annihilation photons emitted due to the chain of events described above, and in Figure 4.1, and so correspond to a true event. However, since both of the photon pair must leave the body and carry 511 keV of energy, not every coincidence that is detected is a true coincidence.

Of course, before each photon leaves the body and reaches the detector, it may interact with tissue or other materials, such as lead and tungsten, that may be part of the PET instrument. Two mechanisms by which the photon may interact with matter are called the photoelectric effect and Compton scattering (Compton, 1923). Of the two, Compton scattering is the more important. Briefly, a Compton scattering interaction is when the photon scatters off an electron (free or loosely bound) in the medium; Then, in the process, transfers some of its energy and changes direction.

Over 90% of the events detected by the PET camera are single event (Phelps et al., 2006, page 35), this is where only one of the two photons are detected. This often happens due to the partner photon not reaching the other associated detector, or because it did not have enough energy to be detected — usually as a result of Compton scattering. Thus, Compton scattering together with the limitations of the PET camera lead to the following types of noisy events that are detected in addition to true coincidences:

1. Random coincidences. This occurs when two photons from *two different* annihilations are detected by the detectors. Here, their partner photons may have been absorbed or not detected due to Compton scattering. In this case, there is no spatial information about the radioactive event(s). An event is recorded erroneously in the LOR of these detectors; this produces background noise in the images.

2. Scattered coincidences. Here, one or both of the photons from the annihilation change direction due to Compton scattering. Once more, there is a loss of spatial information of the event.

3. Multiple coincidences. This occurs when there is a large amount of isotope decay, more than two detectors may detect an event within a short time window. This adds ambiguity to the position of the events. This is usually accounted for within the statistical model, see Section 4.2.4.

These types of undesirable events contaminate the data and lead to the degradation of the quality of the final image. Corrections must be made to the data as, ideally, the PET image should be an image volume in which the value of each voxel is representative of the tissue concentration of tracer. Many data correction method do exist and they are usually based around these types of coincidence. Most techniques involve the application of a series of multiplicative factors to the sinogram (a raw data format, discussed further below). For example: non-uniformities between the detectors are corrected by normalisation; Errors in attenuation can be corrected using calculated or measured attenuation correction; Finally, error due to scatter can be corrected using analytical or simulation based methods. A good review and in–depth discussion of these methods can be found

in Phelps et al. (2006, pg.51-70). We focus here on correction for random coincidence.

Note that it is impossible to distinguish between a true coincidence and random coincidence. Instead correction methods such as these give us a statistical approximation of the rate of random coincidences. Let E_1 and E_2 be the individual photon detection rates (counts per second) of a pair of detectors — this is the rate at which the detector detects single photons, not pairs. It can be shown that the rate of random coincidences, denoted E_R is given by:

$$E_R = 2\tau E_1 E_2,$$

where τ is the width of the logic pulses produced when an photon is detected — take τ to be given. In other words, it is possible to determine the rate of random coincidences based on the singles photon detection rates of the two detectors. In order to determine E_1 and E_2, additional detectors can be added to the scanner.

Next, letting the total number of counts measured be called prompt counts; we may, then, subtract the total counts of random events to give us the net true counts. In other words,

$$\mathcal{E}_{\text{true}} = \mathcal{E}_{\text{prompt}} - \mathcal{E}_{\text{random}},$$

where $\mathcal{E}_{\text{true}}$, $\mathcal{E}_{\text{prompt}}$ and $\mathcal{E}_{\text{random}}$ denote the number of true, prompt and random counts, respectively.

Data correction is applied when the image is in a raw data form called a sinogram. A sinogram can be thought of as matrix with elements which correspond to the measurement of radioactivity at a given location. The measurement of radioactivity here is essentially the number of events recorded by a detector pair. The rows of the this matrix are arranged according to the projection angle in the axial plane; and the columns according to the radial offset from the centre of the scanner. See (Phelps et al., 2006, pg.44) for more details. It is important to note that since the correction takes place prior to image reconstruction, this may lead to strange values (e.g. negative concentration values) in final image.

Image reconstruction methods such as the re-projection algorithm (Kinahan and Rogers, 1989; Phelps et al., 2006) are then applied to the sinogram to give back the counts of each measured voxel. The dynamic PET scans record the counts of events between a set of (often predetermined) time intervals; these counts are then averaged to give a measurement for each interval. This can then be used to create an image, where each voxel value is the concentration of the tracer, in the tissue, in the corresponding time interval.

4.1.3 Measured PET Data: [^{11}C]-diprenorphine Data for Opioid Receptor Quantification

Although the ultimate aim is to produce statistical methods and approaches that are applicable to a large variety of problems that can be studied by PET; the simulation studies in Chapters 6 and 7, Part II will focus on a specific type of PET imaging study. This allows for both focused motivation and clarity in interpretation of the models when first introduced in Section 4.2.

Specifically, the computer simulations studies in Section 6.2 will be representations of the data from a [^{11}C]-diprenorphine opioid ligand-receptor PET study data set. The aim of such PET studies is to measure opioid receptor concentrations in the brain of normal subjects allowing a baseline to be found for subsequent studies on diseases such as epilepsy. In Chapter 7, we analyse a data set using

the proposed novel method and provide further relevant details. Here, to provide some context we briefly explore a measured PET image from an opioid receptor study. First, some general, useful information from the meta-data is discussed next.

The measured data is the dynamic scan of the concentration of the tracer [^{11}C]-diprenorphine in a normal subject, for which an arterial input function was available. [^{11}C]-diprenorphine is a tracer that binds to the neural opioid (pain) receptor system. The subject underwent 95−min dynamic [^{11}C]-diprenorphine PET baseline scan. The subject was injected with 185 MBq (Mega Becquerel) of [^{11}C]-diprenorphine. PET scans are acquired in 3D mode on a Siemens/CTI ECAT EXACT3D PET camera, with a spatial resolution after image reconstruction of approximately 5mm. Technical details including the data correction and normalisation methods of the ECAT EXACT3D PET camera can be found in (Spinks et al., 2000). The PET data was reconstructed using the re-projection algorithm (Kinahan and Rogers, 1989) with ramp and Colsher filters cutoff at the Nyquist frequency. Reconstructed voxel size were 2.096mm × 2.096mm × 2.43mm. Acquisition was performed in listmode (event-by-event) and scans were rebinned into 32 time frames of increasing duration. The lengths of these periods, in seconds, are: $(27.5, 32.5, 2 \times 10, 20, 6 \times 30, 75, 11 \times 120, 210, 5 \times 300, 450$ and $2 \times 600)$. Frame-by-frame movement correction was performed on the PET images. Overall this resulted in images of dimensions/size $128 \times 128 \times 95$ voxels. This gives a total of $1,556,480$ separate times series respectively to be analysed. Note that this is an upper-bound, since masking over the brain regions will result in (often, orders of magnitude) fewer time series.

A single time series from a voxel from this data set is shown in Figure 4.2. Note that there are negative values at the beginning of the measurements, this may occur due to the fact that prior to administration of the tracer there would be little–to–no events detected by the PET scanners (which would then be corrected). Also note that the time intervals of measurements increase in size; this is to account for the fact that there is a rapid increase in the initial perfusion stages followed by a slow decay. In other words, there should be better time resolution at the beginning, when there is more changes in the concentration of the tracer.

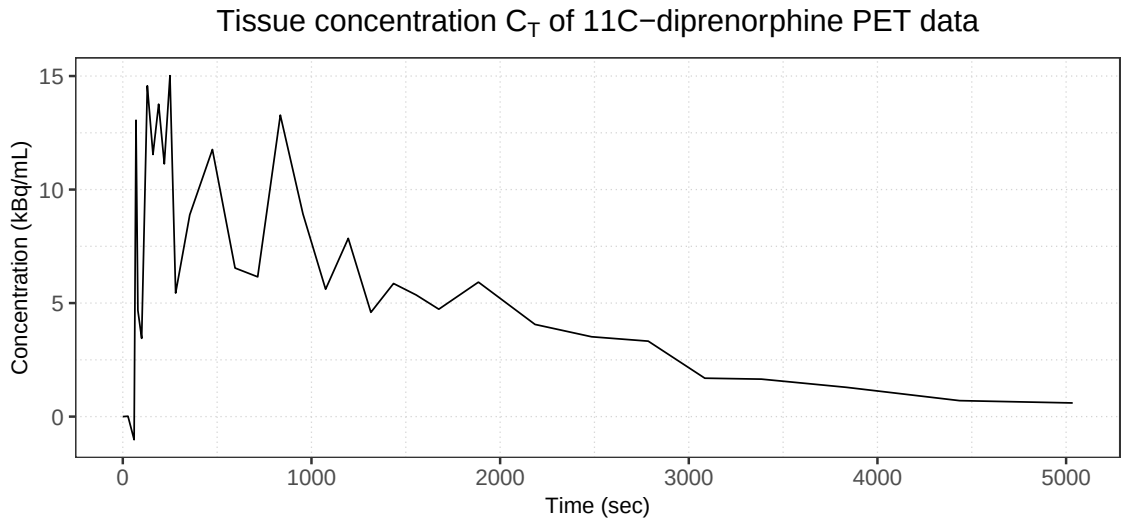


Figure 4.2: A tracer concentration time series of a voxel from measured [^{11}C]-diprenorphine PET data.

This PET data set has been previously analysed by Jiang et al. (2009) and Peng et al. (2008);

However, both works focused on parameter estimation rather than model selection. Zhou et al. (2013, 2016) are more recent works, which do focus on model selection.

Some neural diseases, such as epilepsy, tend to have a patho-physiology which involves changes in the brain receptor concentrations or occupancy levels (Sarikaya, 2015). This may be either due to physical lesions in the brain structure or other biochemically relevant differences from normal controls. By analysing PET data, such as the one presented in this section, it is possible to compute important pharmacological quantities about the tracer, which then allow researchers to derive insights about factors such as receptor concentrations and/or occupancy levels.

Interestingly, both the main advantage and disadvantage of PET lie in the use of its characteristic component – the positron-labelled tracer. It is the tracer, with its increasing variety and accessibility, that enables the specificity of the imaging modality. However, it is also due of the *radioactive* tracer, which must be introduced into the body in an invasive manner via injection, that designs of PET imaging studies are often limited. Another disadvantage, which is now more obvious having explored the machinery, is the cost of the PET scan.

There are many other creative and exciting applications of PET due to the increasing number and availability of tracers. For example, see Morris et al. (2004) and reference therein for description of using PET to study gene transcription and gene therapy. However, scientific progress and insights using PET data are often limited by modelling and analysis methods, underlining the importance of developing better methods.

4.2 Compartmental Models

In order to relate the radioactivity concentration measured with PET, to the underlying neural physiology or biochemistry; mathematical models that adequately describe the tracer kinetics must be applied. In a clinical setting, methods such as the Standard Uptake Value (SUV) method (Kubota et al., 1985) or tissue-to-plasma ratio (RATIO) method (Lehtio et al., 2003) are the most regularly used. In contrast, in research setting, methods such as Spectral Analysis (SA) (Cunningham and Jones, 1993) have existed for over 20 years. A important property of SA is that it does not require any information on the model structure *a priori*; this is an advantages shared by graphical methods such as Patlak (Patlak S. and Blasberg, 1986) and the popular Logan (Logan et al., 1990) analysis.

Although there are many ways to model data from PET, *compartmental models* are among the most widely used (Gunn et al., 2001). In this thesis, we focus on model selection and inference on compartmental models at each of the smallest sub-unit of PET image — the voxel. This is motivated largely due to the model's simplicity, ease of use and flexibility. However, another important property of this class of model is that it aligns well with the concept of model orders. Compartmental models are a collection of related models with varying complexity. That is, an increase in the number of compartments, represented by the model order, results in the increase in the complexity of the model. This gives a natural interpretation, as well as intuition both in spatial and non-spatial statistical methods. As we will discuss further in Section 5.3, a Potts model can be readily and intuitively used to model the number of compartments (more generally, the model order) at each voxel over the whole image. Essentially, using compartmental models to describe the observed data at the voxel–level; We can then model the spatial relationship over the number of compartments at each voxel for the whole image. This is based on the rationale that two adjacent voxels are more likely to be represented by models with the same number of compartments, due

to proximity — something that can be modelled well by the Potts model. Towards this end, we explore this useful class of models in detail.

In the field of Medicine, compartmental models are often used as approximations of the pharmacokinetics of a chemical of interest inside the body. We will restrict our attention to these types of compartmental models, though similar notions can be applied to other contexts. In the context of interest, compartmental models will be used to approximate the (pharmaco)kinetics of the tracer in the body. We may then use this information to make quantitative statements about the PET image; In particular, we will focus on computing the volume of distribution.

The term kinetics can be thought of as the scientific study of the fate of a chemical substance administered to a living organism, as well as how the body interacts with such substance. Indeed, it will become evident that much of the notation used for compartmental models originate from the study of chemical reactions. As such, we begin by formally describing a generic form of compartmental models, thus make use of more precise, mathematical notation; Standard, conventional notation will then be used, once the specific type of model we are interested in is introduced. This may seems somewhat of a round-about way of defining notations — but, as it will soon become clear, this enables us to be both precise and concise.

Compartmental models are a class of mathematical models characterised by the assumption that at any given time, post administration, the chemical of interest must exist in one of many compartments (Morris et al., 2004). Obviously, in the present context, we are interested in modelling the kinetics of the tracer used in PET imaging. This class of models does have the flexibility to allow us, if needed, to construct complex forms consisting of large number of compartments. However, due to the sparsity (at each voxel) and noisy nature of PET data, we restrict our attention to more simpler and pragmatic forms of the model.

Before we begin, it is noteworthy that, within the context of modelling PET images, the compartments need not necessarily be a physical locations. Instead, as with all models (Robert, 2007, Chapter 1), they are modelling constructs approximating phenomena within a complex underlying system.

We introduce concepts and notation for representing compartmental models; They will be based on the framework presented, in much greater detail, by Gunn et al. (2001). When considering a mathematical representation we must account for the following factors: the amount of tracer in each compartment, the flow of tracer between each compartment and the flow of tracer between the compartments and the outside (or the environment). A compartment can be thought to be a state, real or conceptual, of the tracer. For example, this may be some biochemical state (Morris et al., 2004).

Suppose the tracer can take on m different states in the system we wish to study. The model should therefore have some positive $m \in \mathbb{N}$ compartments; We term this model an m–compartment model. Doing so, intuitively allows us to assign an artificial but relevant label to each model. In other words, we can naturally let the model order (as discussed in Section 2.1) of each of these model be equivalent to the number of compartments — hence the use of the letter m here.

Next, we may refer to each compartment in an $m-$compartmental model by $i \in \{0, 1, \ldots, m\}$. Importantly, the index i can take on the value of 0. This is due to the fact that we refer to the environment, or the outside of the system we wish to model, as the 0-th (pseudo-)compartment. In the context of PET images, the 0-th compartment is the plasma. Importantly, this compartment is usually treated in a different manner to the other compartments.

Each compartment will be characterised by the amount of tracer it contains. More formally, we are interested in the concentration of the tracer in each i-th compartment at time $t \geq 0$, denoted here by $C_i(t)$. Over time, compartments interact through the flowing of the tracer between each compartment. A quantity that describes such interaction is the transfer rate, denoted $r_{i,j}(t)$; It is the rate at which tracers flow from one compartment to another, at time t. Here, straightforwardly $r_{i,j}(t)$ corresponds to the rate of transfer from compartment j to compartment $i \neq j$.

Let $q_i(t)$ denote the quantity of tracer in compartment i, at time t. Since q_i is different for each compartment; we may standardise the notion of tracer flow by using, instead, the transfer coefficient, defined

$$\varrho_{i,j}(t) \doteq \frac{r_{i,j}(t)}{q_i(t)}.$$

For PET data, we make the assumption that $\varrho_{i,j}(t) \geq 0$ is constant over time t; Thus, we simply use the notation $\varrho_{i,j}$. Such a model is called a *linear compartmental model*, the term $\varrho_{i,j}$ is then called the rate constant. To aid visualisation, an simple example is shown in Example 4.2.1 below.

Example 4.2.1. Consider a $m = 4$-compartment model, conventionally this (general) model can be diagrammatically represented by Figure 4.3

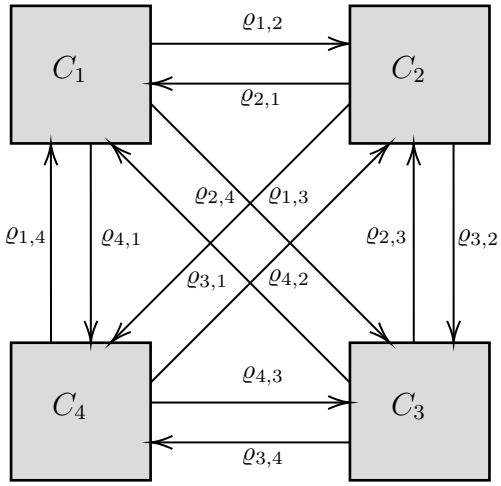


Figure 4.3: A diagram of a 4–compartment model. Here the flow between compartments are represented by arrows and the constants $\varrho_{i,j}$ are the rate of flow. Note that the C_i's here are the concentrations in the compartments. For simplicity the 0-th compartment is not included here.

General models like these are not particularly good at describing tracer kinetics. For example, if we consider the tracer of a ligand (a molecule that binds to a protein in the tissue), it would be physiologically impossible for the tracer to transition directly from certain compartments to other compartments. For example, ligands do not transition from being bound to one type of receptor to another type of receptor. $\triangle$

Next, we would like to describe the dynamics of this system — that is, how does the tracer amount in each compartment change over time. In order to do so, we must also take into consideration inflow of the tracer into the system. We must quantify the amount and rate of the tracer entering the compartments from the outside or environment.

To this end, consider a linear m-compartment model which we wish to construct: Let $\boldsymbol{f}(t) \in \mathbb{R}^m$ be a positively valued m–vector whose j-th component corresponds to the concentration of the tracer in compartment $j \in \{1, \ldots, m\}$, at time t. Similarly, let $\boldsymbol{b}(t) \in \mathbb{R}^m$ be a non-negatively

valued vector with j-th component, representing the transfer rate of the inflow of tracer from the environment (the 0-th compartment) into the j-th compartment.

Finally, define the *transition matrix* by letting $A \in \mathbb{R}^{m\times m}$ with elements corresponding to the transfer rates, or rate constants, between compartments. That is, $A_{i,j} \geq 0$ is the rate of tracer flow from compartment j to compartment i; Or more formally,

$$A_{i,j} = \varrho_{i,j}, \text{ for } i \neq j$$

and

$$A_{i,i} = -\sum_{j'=0:j'\neq i}^{m} \varrho_{j',i}.$$

It is important to note here that the 0-th compartment is not represented in any of the vectors or the transition matrix — hence the term m-compartment model rather than an $(m+1)$-compartment model. However, the rate constants for tracer transitions (transfer between compartments) relating to the environment *are* included. Specifically, note that the inflow $\boldsymbol{b}(t)$ consists of such rate constants; Additionally, the summation for $A_{i,i}$ is over the set containing $j' = 0$.

The dynamics of this system can now be expressed concisely as the following collection of ordinary differential equations (ODE):

$$\frac{\mathrm{d}\boldsymbol{f}}{\mathrm{d}t}(t) = A\boldsymbol{f}(t) + \boldsymbol{b}(t) \text{ with } \boldsymbol{f}(0) = \boldsymbol{\zeta}, \tag{4.1}$$

where $\boldsymbol{\zeta} \in \mathbb{R}^m_+$ is the initial condition. In words, the change in the amount of tracer in the compartments is a function of the inflow of the tracer from the environment in addition to the current amount transformed by the transition matrix A. By solving this system of ODEs we may determine the concentration of tracer in each compartment at time t; That is, we would like an explicit form of $\boldsymbol{f}(t)$.

Before we solve the ODE Eq.(4.1), let us make some restrictions on the general linear compartmental models described above to better describe the PET tracer kinetics. We will see that doing so will then result in the tractable solution that we seek.

4.2.1 Plasma Input Compartmental Models

Conventionally, practical adjustments can be made to the notations introduced above to allow for a more accessible and interpretable representation (with respect to the context of PET tracer kinetics). Firstly, when considering models for PET data, rather than referring to each compartment using ambiguous number indices; it would be more descriptive to refer to them by the state of the tracer that they are representing. For example, the tracer enters the (neural) tissue through the atrial blood, or more specifically the plasma. Therefore, the plasma can be treated as the environment or the outside i.e. the 0-th compartment. In particular for PET, it is possible to observe, through atrial measurement, the tracer concentrations in the plasma over time. These plasma concentration observations are treated as a function, called the plasma time-activity function, or just *input function*; We denote this function $C_{\mathrm{P}}(t)$.

To emphasis: Rather than using C_0, we use C_{P} since it is the concentration in the "plasma" compartment. In addition, all non-plasma compartments will be categorised together as *tissue compartments*. As is usual in analysis of PET data, see for example Buck et al. (1996) and Turkheimer et al. (2003), we consider models containing up to $m = 3$ (tissue) compartments. In

some works, these types of compartmental models are sometimes referred to as m-tissue compartmental models. Since we are only interested in the tissue compartmental models, we use the term m-compartmental model for brevity.

One possible interpretation[4], among many, of the 3–compartment model is the following. The three-compartmental model contains a (tissue) compartment for:

- Tracers in a freely diffusible state in the tissue, labelled F.
- Tracers bound to specific receptors, usually the receptor under study, labelled SP.
- Tracers that are non-specifically bound, labelled NS.

The corresponding concentrations of these compartments can therefore be denoted as $C_{\mathrm{F}}, C_{\mathrm{SP}}$ and C_{NS}, respectively.

Finally, interactions between certain compartments that are physiologically not possible are not considered. Mathematically, this is equivalent to fixing the rate constants ϱ to be 0 for certain physiologically-incompatible compartments; Diagrammatically, arrows representing such flow of tracers are not shown. Consequently, in standard conventions simpler notations are used for the smaller subset of rate constants. That is, rather than using $\varrho_{i,j}$ for each i-th and j-th compartments, we use k_l for $l \in \{1, \ldots, 2m\}$ [5]. The manner in which these k rate constants are defined and organised, together with the notations and conventions described above, are shown in Figure 4.4. For clarity, we state the definitions of the new rate constants:

$$K_1 \doteq \varrho_{\mathrm{P,F}} = \varrho_{0,1}; \qquad k_2 \doteq \varrho_{\mathrm{F,P}} = \varrho_{1,0}; \qquad k_3 \doteq \varrho_{\mathrm{F,SP}} = \varrho_{1,2};$$
$$k_4 \doteq \varrho_{\mathrm{SP,F}} = \varrho_{2,1}; \qquad k_5 \doteq \varrho_{\mathrm{F,NS}} = \varrho_{1,3}; \qquad k_6 \doteq \varrho_{\mathrm{NS,F}} = \varrho_{3,1}.$$

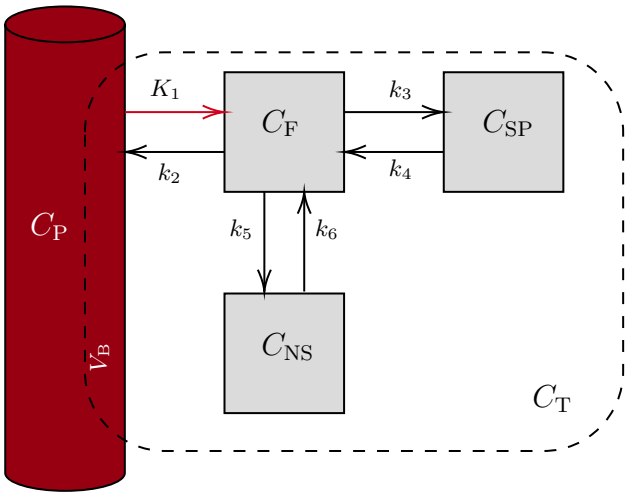


Figure 4.4: A diagram of the linear 3-compartment (plasma input) model, where the flow between compartments are represented by arrows and the constants K_1 and k_i's are the rate of flow. The environment (0-th compartment) here is the plasma. The model of the system (the tissue) is represented here by the tissue compartments. The tracer flows into the system from the environment (plasma) in a freely diffusible state *only*. The rate constant corresponding to this ($\varrho_{\mathrm{P,F}}$) is denoted by K_1, conventionally capitalised since its units differ to the other rate constants.

It should be noted that we have included this interpretation for completeness and exposition *only*.

[4] Particularly for ligand-binding PET studies.
[5] This notation style originate from the notations used to represent chemical reactions.

As will be made clear later, in Section 4.2.2, this interpretation may not always be plausible for actual data. In fact, it is immediately apparent here that this is a simplification of the true system; realistically speaking, tracers bind non-specifically to the molecular by-layer or to many other sites with much lower affinity, thus accurate modelling would require the addition of many more compartments. Unfortunately, observed PET data does not allow for these more sophisticated models. We consider only up to $m = 3$ tissue compartment models. The 1–compartment and 2–compartment model are diagrammatically represented in Figure B.1 and Figure B.2, in Appendix B.1 and Appendix B.2, respectively.

Following Gunn et al. (2001), we will classify these rate constants (except for K_1[6]) as micro-parameters. As we will see later, micro–parameters are generally less stable with respect to parameter estimation from dynamic PET data. Henceforth, we will use the standard convention that the concentration is measured in $\mathrm{kBq/mL}^{-1}$ (kilo Becquerel per millilitre — Becquerel is a unit of radioactivity, and so suitable here for PET data); Subsequently, the rate constants will have units of s^{-1} (per second). The exception is K_1, as it is the inflow from the plasma (liquid state) to tissue (solid state), which will have units $\mathrm{mL} \cdot \mathrm{cm}^{-3} \cdot \mathrm{s}^{-1}$ — note it is possible to use s^{-1} as the unit for K_1, conventionally the former is used for emphasis. It is also convention to capitalise K_1, due to this small difference.

Tracer in the blood from the vascular volume fraction of the tissue will contribute to the total radioactivity concentration of the tissue measured by PET. We denote this V_B, as shown in Figure 4.4. In most cases, V_B is very small (Zhou et al., 2013), accordingly we will assume that $V_\mathrm{B} = 0$.

4.2.2 The Tissue Time-Activity Function

Let the vectors $\mathbf{1}$ and $\mathbf{0}$ denote m-vectors with all components equal to 1 and 0, respectively. Use of a (plasma) input function results in the assumption that the signal component of PET data measurements in each voxel is the sum of the concentrations in all tissue compartments. That is, if $\boldsymbol{C}_\mathrm{T}(t)$ is an m–vector with i-th component corresponding to the tracer concentration in tissue compartment i, then the observed PET signal (i.e. noise-free) is then the scalar function $C_\mathrm{T}(t) \doteq \mathbf{1}^\top \boldsymbol{C}_\mathrm{T}(t)$. We call C_T the *tissue time-activity function*.

Following from Eq.(4.1), *mutatis mutandis*, the dynamics of the m tissue compartments with a plasma input function $C_\mathrm{P}(t)$ can then be written as:

$$\frac{\mathrm{d}\boldsymbol{C}_\mathrm{T}}{\mathrm{d}t}(t) = A\boldsymbol{C}_\mathrm{T}(t) + \widetilde{\boldsymbol{b}}C_\mathrm{P}(t), \tag{4.2}$$

and initial conditions,

$$\boldsymbol{C}_\mathrm{T}(0) = \mathbf{0}.$$

In other words, the availability of the plasma input function means we have that the inflow vector $\boldsymbol{b}(t) = \widetilde{\boldsymbol{b}}C_\mathrm{P}(t)$, where we have fixed $\widetilde{\boldsymbol{b}} \doteq (K_1, 0, \ldots, 0)^\top$; Furthermore, the transition matrix has components consisting of the appropriate rate constants from $K_1, k_2, \ldots, k_6$ (i.e. with respect to the arrangement shown in Figure 4.4).

Consider the following interesting and very useful result, a direct consequence of the assumptions we have made above. Firstly, the restrictions on the (physiologically impossible) rate constants

[6]This should not be confused with the Markov kernel notation K.

results in a non-cyclic system. Thus, as noted by (Schmidt, 1999), the transition matrix A will now be negative semi-definite. That is, A will constitute of non-positive diagonal elements and non-negative off-diagonal elements. More importantly, it can be further shown that A is diagonisable, see Friedberg et al. (2003, Chapter 5); as well as having spectral decomposition

$$A = S\Xi S^{-1},$$

where $\Xi \doteq \text{diag}(\xi_1, \ldots, \xi_m)$ is a diagonal matrix of eigenvalues and the $S \in \mathbb{R}^{m\times m}$ is a matrix consisting of eigenvectors. To be precise: Ξ consists of the negatively–valued eigenvalues of A as its diagonal; with the i-th column of the matrix S and the i-th row of the matrix S^{-1} are the right and left eigenvectors of A , respectively, corresponding to each eigenvalue ξ_i. Consequently, we then have that,

$$A = \sum_{j=1}^{m} \xi_j \boldsymbol{u}_j \boldsymbol{v}_j^\top, \tag{4.3}$$

where $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_m$ and $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_m$ are the columns and rows of S and S^{-1}, respectively.

Returning to the task at hand, it is now possible to attain the explicit solution to the ODE Eq.(4.1) above. First, as per Seber and Wild (2003), we have that:

$$\boldsymbol{C}_\text{T}(t) = e^{At}\boldsymbol{0} + \int_0^t e^{A(t-s)}\boldsymbol{b}(s)\text{d}s; \tag{4.4}$$

where e^{At} is the matrix exponential, defined

$$e^{At} \doteq \sum_{j=0}^{\infty} \frac{(At)^j}{j!}.$$

But, by Eq.(4.3) above, we also have that

$$\begin{aligned} e^{At} &= Se^{\Xi t}S^{-1} \\ &= \sum_{j=1}^{m} e^{\xi_j t}\boldsymbol{u}_j\boldsymbol{v}_j^\top. \end{aligned}$$

Noting the first term of Eq.(4.4) is 0, we rewrite the second term as,

$$\int_0^t e^{A(t-s)}\boldsymbol{b}(s)\text{d}s = \sum_{j=1}^{m} K_1 \int_0^t e^{\xi_j(t-s)}\boldsymbol{u}_j\boldsymbol{v}_j^\top \boldsymbol{e}_1 C_\text{P}(s)\text{d}s,$$

where $\boldsymbol{e}_1 \doteq (1, 0, 0, \ldots, 0)$.

Thus, we obtain the solution to the ODE Eq.(4.1), given here in the form of the tissue time–activity function (the signal which we will attempt to estimate),

$$\begin{aligned} C_\text{T}(t) &= \boldsymbol{1}^\top \boldsymbol{C}_\text{T}(t) \\ &= \boldsymbol{1}^\top \sum_{j=1}^{m} K_1 \int_0^t e^{\xi_j(t-s)}\boldsymbol{u}_j\boldsymbol{v}_j^\top \boldsymbol{e}_1 C_\text{P}(s)\text{d}s. \end{aligned}$$

More concisely, we may rewrite the immediate above as a convolution,

$$\begin{aligned} C_{\mathrm{T}}(t) &= H(t) \circledast C_{\mathrm{P}}(t) \\ &\doteq \int_0^t C_{\mathrm{P}}(t) H(t-s) \mathrm{d}s; \end{aligned}$$

where,

$$H(t) \doteq \sum_{j=1}^{m} \phi_j e^{-\vartheta_j t},$$

and

$$\begin{aligned} \phi_j &\doteq K_1 \mathbf{1}^T \boldsymbol{u}_j \boldsymbol{v}_j^T \boldsymbol{e}_1, \\ \vartheta_j &\doteq -\xi_j, \end{aligned}$$

for all $j \in \{1, \ldots, m\}$. The function H is called the impulse response function (IRF) and can be interpreted as the tissue tracer concentration curve that would be measured after an ideal instantaneous bolus injection. A formal proof of this solution is given by Gunn et al. (2001, Theorem 2.2).

Define $\phi_{1:m} \doteq (\phi_1, \ldots, \phi_m)$ and $\vartheta_{1:m} \doteq (\vartheta_1, \ldots, \vartheta_m)$. Here, the components of these vectors can be thought of as functions of the rate constants; Accordingly, they will be treated as micro–parameters. Expression of the explicit form of the functions relating the rate constants to the micro–parameters can be found in Gunn et al. (2001); They are also include in Appendices B.1, B.2 and B.3 for clarification and completeness. It is now possible to use the following notation for the tissue time-activity function,

$$C_{\mathrm{T}}(t; \phi_{1:m}, \vartheta_{1:m}) \doteq \sum_{j=1}^{m} \phi_j \int_0^t C_{\mathrm{P}}(s) e^{-\vartheta_j (t-s)} \mathrm{d}s. \tag{4.5}$$

Given this form, an alternative interpretation, to the one given for Figure 4.4, which may be more representative of the true system is the following: There exists an arbitrary unknown decay function for the tracer concentration in each voxel; the equation above is an approximation with a linear m–compartment model with exponential decay. A similar interpretation, and approach, is used in the non-negative non-linear least squares (NNLS, Cunningham and Jones (1993)) method for PET analysis. We discuss the NNLS (also called Spectral Analysis) method for PET analysis in the sequel.

In view of this, each ϑ_j can be thought of as the rate of loss due to the combination of radioactive dissipation in the tissue and the transient phenomena (e.g. circulation through tissue vasculature) in compartment j. Similarly, ϕ_j can be interpreted as a quantity related to the rate of tracer inflow into the compartment — indeed, in the simple setting of linear $m = 1$–compartment model, $\phi_1 = K_1$. However, for a larger number of compartments this relationship is complex, see Appendix B; On the other hand, the relationship can still be thought to be governed by the transient matrix A and its eigenvectors and eigenvalues. Recall that the input function C_{P} is assumed to be continuously measured; since C_{T} is measured discretely, this leads to the measured values of the integral of the signal over each n consecutive, non-overlapping time intervals. See Figure 4.5 and Figure 4.6 for an example of C_{P} and C_{T}.

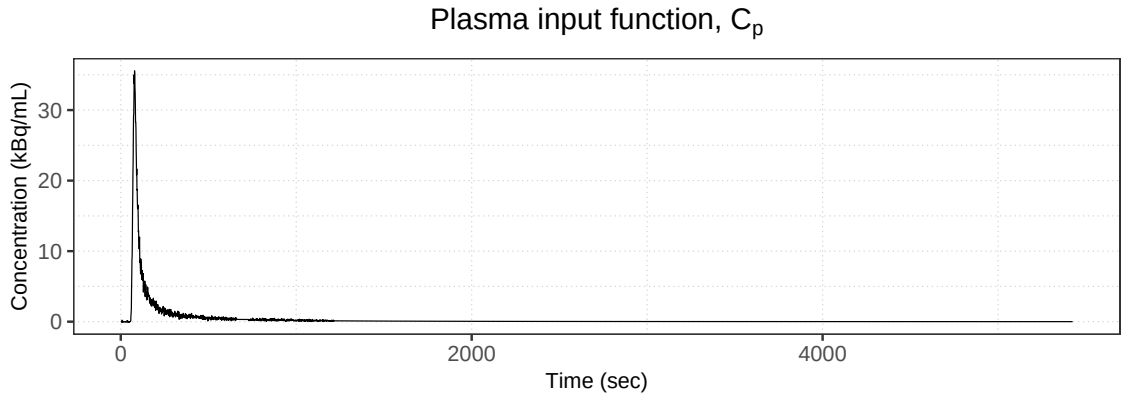


Figure 4.5: The plasma input function, C_{P}. This is measured through continuous arterial blood sampling using an online monitor. The function represents the cumulative availability of the tracer in the arterial plasma.

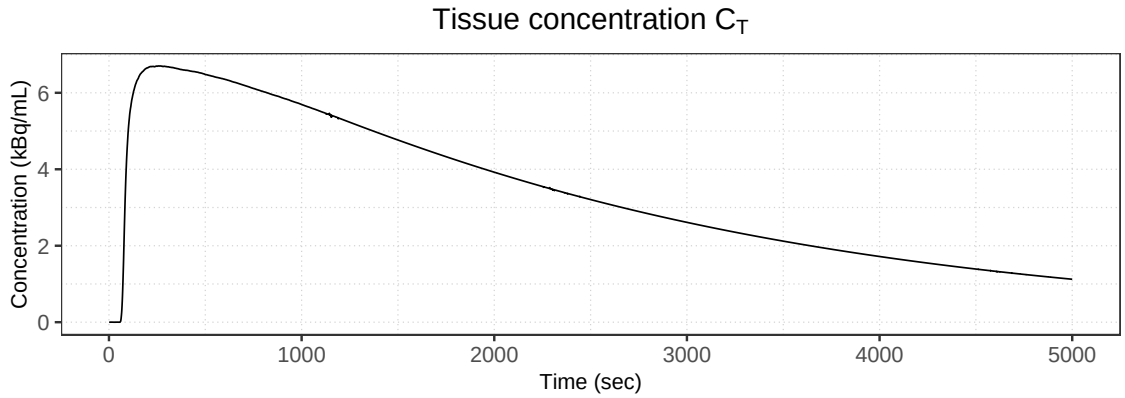


Figure 4.6: Function showing the tissue concentration, C_{T}, for simulated noise free data generated from a 3–compartment plasma input model.

Importantly, in this form, as shown in Eq.(4.5), it is also easy to see that each ϕ_j and ϑ_j is exchangeable with any other ϕ_k and ϑ_k, for $k \neq j$, respectively. Note that the ϕ's and ϑ's must correspond, hence the fixed subscript. For example, take $m = 2$ and fix $\phi_{1:2}$ and $\vartheta_{1:2}$, then

$$C_{\mathrm{T}}(\cdot; (\phi_1, \phi_2), (\vartheta_1, \vartheta_2)) \equiv C_{\mathrm{T}}(\cdot; (\phi_2, \phi_1), (\vartheta_2, \vartheta_1)).$$

The notion that parameters can be estimated uniquely from noise-free input-output date (i.e. C_{T}) is captured by the notion of *identifiability*. Here, we can see that the micro–parameters in this context are not identifiable. Of course, in order to make meaningful conclusions from any model of actual data we must do so through identifiable parameters. Furthermore, it is now apparent that it is not very meaningful to impose any interpretations of these models (such as Figure 4.4) with any certainty.

4.2.3 Volume of Distribution

Consider the following important quantity

$$V_{\mathrm{D}} \doteq \int_0^\infty H(t)\mathrm{d}t,$$

called the *volume of distribution*. It is defined as the ratio of the tracer concentration in target tissue to the tracer concentration in plasma at equilibrium. In other words, in the hypothetical situation where a tracer injection is made into the plasma such that the plasma concentration remained constant over time, then the ratio of tracer concentration in the tissue to that in the plasma after infinite time has passed would be exactly V_{D}. Roughly speaking, V_{D} is correlated to the amount of tracer distributed into the tissue; high V_{D} represents greater tissue distribution. Thus, the V_{D} of a tracer identifies the degree to which a tracer has been distributed in body tissue rather than the plasma.

It is trivial to see that

$$V_{\mathrm{D}} = \sum_{j=1}^{m} \frac{\phi_j}{\vartheta_j};$$

subsequently, note that V_{D} is uniquely identifiable from perfect input-output data. In other words, regardless of any permutations of the order of the micro–parameters, since V_{D} is a sum of the ratio of corresponding micro–parameters will be identifiable from the data. V_{D} (and $K_1 = \sum_{j=1}^{m} \phi_j$) are classified as a macro parameter and in general they are more stable — evidently, almost all works in PET data analysis including: Zhou et al. (2013, 2016), Gunn et al. (2001), Peng et al. (2008) and Cunningham and Jones (1993); place V_{D} as the *parameter of principal inferential interest*.

Although being able to identify the rate constants would give a more sophisticated understanding of the data; the V_{D} still allows inferences of many important quantities of interest including drug distribution, receptor density and other physiological/pharmacological quantities.

Having now formally defined the volume of distribution, the importance of model selection becomes more clear. The V_{D} depends on the model order m and since the best model is unknown *a priori* we must perform both model selection and parameter estimation to get good estimates of this important macro parameter.

4.2.4 A Statistical Model for PET Tracer Kinetics

Recall, from Section 4.1.2, that many factors of the PET image machinery and acquisition process leads to noise in the measurement of tracer radioactivity. In contrast, the model given by the solution the linear compartmental model ODE, namely C_{T}, in Eq.(4.5) is deterministic. Therefore, we must accommodate the uncertainty caused by the noise using statistical models. We characterise this uncertainty to complete the construction of this model.

Suppose that we have positive $k \in \mathbb{N}$ number of measurements: Let $t_1, \ldots, t_k$ be the *end points* of the time frames at which the tissue concentrations were measured; Similarly, denote $y_1, \ldots, y_k$ to be the observed data. We will need to take into consideration the physical characterisation of the PET instrumentation and its system when we model the noise. Firstly, the time frames are irregularly spaced. The length of the interval affects the amount of uncertainty present, since the measurements at the end point of the time frame is derived by averaging from the measured radiation within that interval. Additionally, due to multiple coincidences, as discussed in Section 4.1.2, higher concentration of the tracer often results in increased noise.

These factors results in the assumption that the error is white and additive, with zero mean and variance proportional to the activity divided by length of time frames. Furthermore the PET data

is generated through independent Poisson decay of radioisotopes — a normally distributed error is a good plausible approximation of this Poisson nature.

Thus, combing the deterministic evolution model with this stochastic measurement, yields the following model:

$$y_i = C_{\mathrm{T}}(t_i; \phi_{1:m}, \vartheta_{1:m}) + \sqrt{\frac{C_{\mathrm{T}}(t_i; \phi_{1:m}, \vartheta_{1:m})}{t_i - t_{i-1}}}\epsilon_i, \text{ for all } i = 1, \ldots, k; \tag{4.6}$$

where: for clarity, we re-state from Eq.(4.5)

$$C_{\mathrm{T}}(t_i; \phi_{1:m}, \vartheta_{1:m}) \doteq \sum_{j=1}^{m} \phi_j \int_0^{t_i} C_{\mathrm{P}}(s) e^{-\vartheta_j (t_i - s)} \mathrm{d}s,$$

for all $i \in \{1, \ldots, k\}$, model order $m = 1, 2$ or 3 indicates the number of tissue compartments, fix $t_0 \doteq 0$ and $\{\epsilon_i\}_{i=1}^{k}$ are i.i.d zero–mean random variables.

Conventionally, in the PET analysis literature, the innovations ϵ_j are considered to be normally distributed. More recently, Zhou et al. (2013) showed that, when the Bayesian approach, modelling the error with a Student's t–distribution yielded evidence for better fitting of actual observed data. In view of this, we may consider both types of error structures:

$$\epsilon_i \sim \mathcal{N}(0, \sigma^2) \qquad \text{or} \qquad \epsilon_i \sim \mathcal{T}(0, \tau, \nu) \qquad \forall i \in \{1, \ldots, k\};$$

where $\mathcal{N}(0, \sigma^2)$ is the normal distribution with mean zero and variance σ^2 and $\mathcal{T}(0, \tau, \nu)$ is the Student's t–distribution with location zero, scale τ, and ν degrees of freedom.

4.2.5 Current Statistical Methods for PET data analysis

Given the relative recency of PET scans, studies that attempt to model spatial relationships in PET images are sparse and tend not to use a Bayesian approach. One such example is Zhou et al. (2002) who proposes a method that extends on current standard PET data analysis to allow for spatial dependence. That is, non-linear ridge regression, with local spatial constraints in the parameter space, is used to improve upon parametric images produced by conventional weighted NLS methods. The proposed method, termed nonlinear simple ridge regression with spatial constraints (NLRRSC), can be summarised as follows: Firstly, cluster analysis is applied to the dynamic PET image. Next, compartmental tracer kinetic models are fitted to the kinetics of each cluster from the cluster analysis. The inferred parameter estimates are then used to extract the components of the parameter space. Subsequently, component representation model analysis is used to generate initial estimates and constraints. These initial estimates are then modified through optimisation using the non-linear regression with the spatial constraints. Zhou et al. (2002) showed that using the NLRRSC method reduced the percentage mean square error by $60 - 80\%$ in simulation studies. These promising results motivate the incorporation of spatial dependence when analysing PET images, particularly in Bayesian frameworks where such studies have been largely absent.

An important factor to also consider is that approaches for encoding spatial information may need be application-specific. For example, Bezener et al. (2018) presents a method of variable selection, that models spatial dependence using hierarchical spatial priors and parcellation, for the analysis

of MRI images. It is interesting to note that this study did not use the popular Potts model approach for modelling spatial dependence.

On the other hand, there exists many computationally efficient methods for complex settings, such as PET data analysis, albeit they typically involve assumptions of spatial independence. These include, among many others: Gunn et al. (2002), Peng et al. (2008), Jiang et al. (2009), and Zhou et al. (2013, 2016).

The aforementioned NNLS method (Cunningham and Jones, 1993) is an effective and commonly used method for analysis of PET data. This method involves minimal modelling assumptions and imposes no prior assumptions on the number of components. Instead, as discussed in Section 4.2.2, the time-series is interpreted as a noisy measurement of exponential decay. The method involves the formulation of a constrained linear optimisation problem, where: m will represent the maximum number of terms to be included, for example Cunningham and Jones (1993) use $m = 100$; Each $\theta_j, j = 1, \ldots, m$ is fixed and predetermined to lie within a physiologically meaningful range and takes m possible values. Finally, the optimal value (in the least square sense) of ϕ_j for each of θ_j is calculated using some numerical method, subject to $\phi_j \geq 0, j = 1, \ldots, m$.

In contrast, Zhou et al. (2013) argue for a Bayesian approach for statistical analysis of PET image: Given that there exists considerable information and data on the tracer, as well as previous PET imaging studies. They investigated an application-specific MCMC approach to Bayesian model comparison for this class of models; Studying both vague non–informative and biologically informed priors. In their work, samples from a MH chain were used in the harmonic mean estimator, Eq.(2.21), to approximate the model evidence in order to perform model comparison, selection and averaging. This work also demonstrated the possibility that Monte Carlo approaches could be used to investigate and meaningfully compare more complex models, containing up to $m = 3$ compartments, rather than the typical up to $m = 2$ compartmental models. Zhou et al. (2013) further showed that within a Bayesian framework, a $t-$distributed error structure may be far more plausible. In the numerical studies in Part II, Sections 6.2 and 7.2 , we will make use of these Bayesian models, given by Zhou et al.. As such, expressions for prior and posterior densities, of both cases of error structures, can be found in Appendix C.

Importantly, with regard to model selection within this context, Zhou et al. (2013) also showed that a Bayesian approach *exhibited some spatial structure* despite assuming spatial independence. This was in contrast to using the AIC (Akaike, 1973), as explored in Section 2.2.1, for model selection with the NLS methods used to approximate the MLE, as well as the NNLS method — these showed no obvious spatial structure, see also Section 7.2.

Even more recently, Zhou et al. (2016) used an adaptive extension of the SMC sampler algorithm to estimate the model evidence (marginal likelihood). The class of algorithms presented in this work, as explored in detail in Chapter 3, uses adaptive MCMC kernels together with adaptive annealing schemes to minimise tuning requirements to further the accessibility of Monte Carlo approaches to this problem. Once the model evidence is estimated, Bayes factors can be used to facilitate Bayesian model selection. The computational method proposed in this thesis, builds upon this approach and thus inherits many of these advantages.

There continues to be current interest and further development of Bayesian approaches in analysis of PET data, albeit in almost all cases they are non-spatial approaches. The most recent example is Fan et al. (2021), who propose a simple and intuitive algorithm, based on Approximate Bayesian Computation(ABC; Rubin (1984)), for analysis of PET data. As with the above Bayesian ap-

proaches, this method, termed PET-ABC, assumes voxels are spatially independent. PET-ABC works by simulating from the parameter space using a simple rejection scheme: Firstly proposals are sampled from the prior distribution, typically this is the uniform density with physiologically meaningful range. This proposal is used as a trail value to estimate the signal C_{T}. Next, the error between summary statistics of the estimated signal and the denoised data $\boldsymbol{y}$ is computed. In their study, Fan et al. (2021) suggest using spline smoothed estimates of C_{T} as the summary statistic for both the estimated and observed signal. The proposed value of the parameter is accepted if the above error is below a predetermined threshold. The generated parameter sample can then be use for point estimation and to quantify uncertainty. Bayesian model selection follows the above method, with the additional step of proposing a model at each iteration.

Finally we note the following: Zhou et al. (2013) compared an MCMC-based method to the NLS(AIC) approach and the popular NNLS method, showing that in simulated study the Bayesian approach produced more accurate parameter estimates. Importantly, as mentioned above, when applied to measured PET data, analysis using the Bayesian approach revealed spatial structures not seen in the other methods. Zhou et al. (2016) produced similar results as the MCMC approach, but presented a near automated SMC-based method with minimal tuning requirements. Given these considerations, we will treat the method proposed by Zhou et al. (2016) as state of the art and use this SMC approach for empirical comparison with the proposed method.

4.3 Summary

As stated in the beginning of this chapter, many of the current methods for PET image analyses assume spatial independence to overcome computational limitations. We saw that this arose from the size and complexity of typical PET image data. This was made clear from the exploration of the PET camera mechanism and the metadata of data set it self.

Furthermore, we saw that computational simplicity in the above mentioned statistical methods is typically attained through a “mass univariate” approach. Importantly, these methods have shown that even with spatially independence, compartmental models can be successfully used to model PET data. Likewise, using a Bayesian approach revealed that even when spatial independence is assumed, the inferred model orders configuration output suggest the existence of some underlying spatial structures.

One natural strategy is, therefore, to adapt existing non-spatial approaches to a model that does incorporate spatial dependence. This strategy is particularly natural when modelling spatial dependence via a Potts model as it can be efficiently targeted by MCMC methods with single-site update schemes. That is, to be more precise, we encode spatial dependence at the level of the number of compartments only —Or, at the level of model orders rather than the model parameters. We may now turn to exploring this interesting strategy in more formal detail, next.

Part II

Methodology

Chapter 5

The Node-wise Pseudo-marginal Algorithm

Know a drop of seawater and all the sea is known.

— Sri Ramakrishna Paramahamsa

Having reviewed the necessary background and relevant existing computational methods, we can now turn to the construction, presentation and evaluation of the proposed novel approach for incorporating spatial dependence. The technique presented in this chapter is an extension of the pseudo-marginal MCMC algorithm originally presented by Beaumont (2003), and characterised by Andrieu and Roberts (2009).

More specifically, the Potts model (Potts, 1952) is first utilised as a prior distribution to encode spatial dependence. Imposing further assumptions, of conditional spatial independence for a subset of the parameters in the model likelihood, gives rise to significant simplifications in computations. In other words, this Markov random field model is used on the discrete space of model orders to describe spatial dependence between neighbouring nodes (pixels). Finally, a standard component-wise MCMC updating scheme is used in conjunction with the above assumptions. This then allows for the use of node-level marginal likelihood estimators to be used within a pseudo-marginal MCMC method. The end result is a tractable, accessible computational method that uses model selection at the individual (node/pixel) level to perform spatially dependent model selection on the whole (graph/image) data set. This method is a flexible but efficient algorithm that can be readily implemented. In fact, the proposed methods allows for the use of existing unbiased non-spatial estimators to be used within a framework which incorporates spatial dependence.

In addition, the proposed method can be further extended through augmentation of the state space of the generated Markov chain. This gives rise to a specialised setting, where careful specification of the proposal distributions can lead to many further computational approaches and techniques.

The remainder of this chapter is structured as follows. The Potts model, together with relevant points on the coupling constant, is briefly reviewed in Section 5.2. The hierarchical model that incorporates the Potts prior, together with the important assumption of spatial independence

is presented in Section 5.3. Section 5.4 then presents the proposed method, with Section 5.5 a discussion on theoretical considerations. Included in Section 5.5, is formal argument for an augmentation in the state space that allows for the innovation of further novel techniques.

The method's numerical performance will be evaluated in Chapter 6 and applied in realistic settings in Chapter 7. Empirical studies which show notable improvements in revealing spatial structures, in measured PET data, are presented in the latter. Also investigated in this empirical study are methods using simple proposal schemes based on the multiple augmented space, as well as straightforward approximation of the proposed algorithm.

In this chapter, examples are provided to aid intuition and visualisation, but can be ignored as the notation and concepts are kept generic and self-contained. Finally, as aforementioned, work based on shorter versions of the chapters in this Part, bar Chapter 8, have been submitted for publication, see Thesingarajah and Johansen (2021).

5.1 Graphs and notation

The generic hierarchical Markov Random field model we wish to construct, and the Potts model prior that it incorporates, are graphical models, thus require terms and concepts referring to graphs. We begin by first introducing and defining the necessary notations. Note, some of these terms were briefly introduced in Section 2.8, we re-state them here for completion and clarity.

A graph $G = (V, E)$ is the pair of sets of nodes, or vertices, denoted V and edges denoted E. In particular each element of the edge-set E must be some pair of elements of the nodes-set $u, v \in V$, denoted $\langle u, v\rangle \in E$. Two nodes u, v are said to be connected by the edge $\langle u, v\rangle$ if $\langle u, v\rangle \in E$. In this case we may say that u and v are neighbours, or adjacent, and denote this by the relation $u \sim v$. Here, we will look only at undirected graphs, so the relation $\sim$ will be symmetric and $\langle u, v\rangle$ is an unordered pair. Denote by $\partial(v) \doteq \{u \in V : v \sim u\}$ the set of neighbours of v, by convention v is not a neighbour of itself.

Given a graph, G, a collection of random variables, $\boldsymbol{X} = (X_v : v \in V)$, indexed by the nodes-set, V, is called a random field on G. Let P be the law or probability distribution of $\boldsymbol{X}$; and define, for any $A \subset V$, $\boldsymbol{X}_A = \{X_v : v \in A\}$ and $\boldsymbol{X}_{-A} = \{X_v : v \in V \setminus A\}$. With a slight abuse of this notation, we will write $\boldsymbol{X}_{-v}$ to denote $\boldsymbol{X}_{V\setminus\{v\}}$ for $v \in V$. $\boldsymbol{X}$ is called a Markov Random field (MRF) (Besag, 1974) on a discrete graph if and only if we have that:

$$P(X_v|\boldsymbol{X}_{-v}) = P(X_v|\boldsymbol{X}_{\partial(v)}),$$

where each component X_v takes value in some set $\mathcal{X}$. Thus, $\boldsymbol{X}$ takes values in $\mathcal{X}^V$, the collection of maps from V to $\mathcal{X}$. In this thesis, we are interested in model selection (and further extend, by augmenting, the state space in Section 5.5.2); Thus, $\mathcal{X}$ will be restricted to be finite.

Example 5.1.1. Let d be a positive integer, and recall that $\mathbb{Z}^d$ is the set of all $d-$vectors $\boldsymbol{v} = (v_1, v_2, \ldots, v_d)^\top$ with integer valued coordinates. We define a notion of distance between two vectors $\boldsymbol{v}, \boldsymbol{u} \in \mathbb{Z}^d$ by the L_1 (Manhattan) distance

$$\delta(\boldsymbol{u}, \boldsymbol{v}) \doteq \sum_{i=1}^{d} |u_i - v_i|.$$

This then allows us to represent $\mathbb{Z}^d$ as a graph. Consider a graph with vertices $V = \mathbb{Z}^d$, and

associated edge-set defined

$$E^d = \{\langle \boldsymbol{u}, \boldsymbol{v}\rangle : \boldsymbol{u}, \boldsymbol{v} \in \mathbb{Z}^d, \delta(\boldsymbol{u}, \boldsymbol{v}) = 1\}.$$

Such graphs are called $d-$dimensional integer lattice graphs and are denoted $L^d \doteq (\mathbb{Z}^d, E^d)$. In particular, finite subsets of the 2-dimensional lattice L^2 called the square lattice could be used to represent a 2D images such as a slice of dynamic PET image. $\triangle$

Example 5.1.2. The elements of the vertices-set will be used to index coordinates of random vectors. When using lattices to represent images, for ease of readability, the following convention can be used instead. Let $[l] \doteq \{0, \ldots, l-1\} \subset \mathbb{Z}$, and consider the Cartesian product of finite sets of positive integers denoted by

$$\mathbb{Z}^d_{[l_1,\ldots,l_d]} \doteq [l_1] \times \cdots \times [l_d] = \{(v_1, \ldots, v_d) : v_1 \in [l_1], \ldots, v_d \in [l_d]\}.$$

Note that this *finite non-negative* integer grid $\mathbb{Z}^d_{[l_1,\ldots,l_d]}$ is a subset of $\mathbb{Z}^d$, and so, in this space, the distance δ can remain the same; similarly the definition of an edge-set $E^d_{[l_1,\ldots,l_2]}$ for a lattice graph pertaining to $\mathbb{Z}^d_{[l_1,\ldots,l_d]}$ follows in the same fashion.

Each element of the finite grid $\mathbb{Z}^d_{[l_1,\ldots,l_d]}$ may be transformed into a distinct scalar form using the lexicographic order. Formally this is done as follows, given some element of $\mathbb{Z}^d_{[l_1,\ldots,l_d]}$ denoted $\boldsymbol{v} = (v_1, \ldots, v_d)^\top$, define the function $f : \mathbb{Z}^d_{[l_1,\ldots,l_d]} \to \mathbb{Z}^+$ given by

$$f(\boldsymbol{v}) = 1 + v_d + (l_d)v_{d-1} + \{(l_d)(l_{d-1})v_{d-2}\} + \cdots + \{l_d \times \cdots \times (l_2) \cdot v_1\}.$$

For example, a simple 4×4 finite grid $\mathbb{Z}^d_{[4,4]}$ can be transformed into a lexicographic form, as shown in Figure 5.1 below.

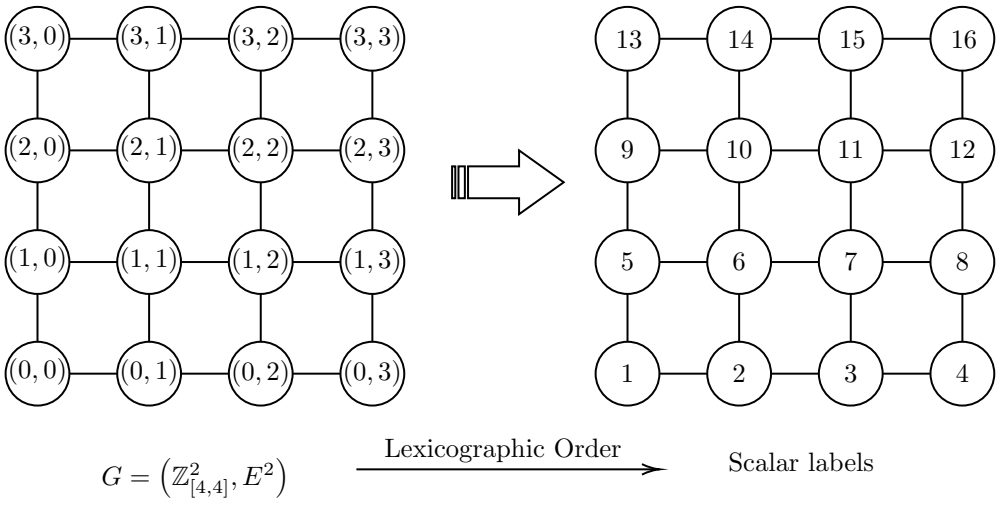


Figure 5.1: The lexicographic order transformation of the 4×4 integer lattice graph.

In other words, the lexicographic order gives a natural way to label the vectors of finite subsets of $\mathbb{Z}^d$ using scalar form. These lattice and lexicographic order graphs provide an simple way to visualise and label images date when associating them with graphs. Henceforth, where appropriate, we will suppress the dimensional superscript.

Finally, consider an MRF $\boldsymbol{X}$ with associated lexicographic order graph $G = (V, E)$. Suppose,

each component of the random vector takes value in the state space $\mathcal{X}$. Then, $\boldsymbol{X}$ takes value in the space $\mathcal{X}^V = \mathcal{X}^{|V|}$. In words, straightforwardly, with the labelling $V = \{1, ..., |V|\}$; the usual $|V|$-dimensional vector space coincides with $\mathcal{X}^V$. $\triangle$

Having introduced the necessary, relevant notations, we may now briefly discuss the problem at hand in more formal detail. Recall the strategy of using existing simpler models (used to describe data at the node-level) to build a larger model, that describes the whole spatial data (or the whole graph). That is, we will use finite MRFs with components representing the model orders of models at each of the node or pixel.

For example, consider compartment models which can be successfully used to model the data at each node in the graph; Then, spatial information can be incorporated in a number of ways. One elaborate way to do so is to allow compartments at different nodes to interact. However, we propose a simpler, arguably more natural, strategy with no inter-node compartment interactions; Instead, we use the finite MRF to represent the number of compartment at each node. That is, the model order at a given node will be dependent on the model orders of neighbouring nodes. We detail this further, in the sequel. It is noteworthy that, even when considering up to $m = 3$ compartments, the number of possible model order configurations, that can describe the whole image, is $3^{|V|}$. Essentially, the task of model selection here is that we seek the best model within a collection of $3^{|V|}$ models. In the PET image settings, $|V|$ can be thought of as the total number of voxels — typically, this may be as large as 10^6 (Hammers et al., 2007).

Of course, we require a distribution of this model order MRF, that can encode the spatial relationship in a accurate, principled manner — we discuss this next.

5.2 The Potts Model

The Potts model (Potts, 1952), a generalisation of the Ising model (Ising, 1925), see also Section 2.8 , was used originally to model interacting spins on a lattice. However, these models have been shown to be also very effective in analysis of image data; for a detailed discussion and review of such applications see Winkler (1995), Geman (1990) and references therein. For example, Geman and Geman (1984), is an early study demonstrating the effectiveness of the Ising model for restoration of images under a Bayesian framework.

More recently, the Potts model has been used very successfully within the more broader but still growing sub-field of Bayesian image analysis. (Besag, 1993) showed that MRFs can be successfully used in Bayesian image analysis. As such, the Potts model has become a standard model for categorical images (Hurn et al., 2003). However, typically the Gibbs sampler or a standard exact marginal MH algorithms are used to characterise models based on this distribution. Additionally, the use of these methods for model selection in this specialised setting is not very well studied. Using these standard approaches in the setting of current interest is possible, however doing so would fail to exploit the structure of the problem — as done so by the method proposed here.

In the present context, the Potts model is an ideal prior for model orders, due to its discrete state space, minimal parametrisation and ability to encode the general principle that nearby vertices are *a priori* likely to be best described by the same model. Following the Bayesian approach, in this study, the model order of the data will be an MRF with a Potts prior distribution. For a review of Potts distributions, particularly in the image analysis context, see Hurn et al. (2003).

Given a graph $G = (V, E)$, finite state space $\mathcal{X}$ and coupling constant $J > 0$, the Potts model

specifies a family of parametric probability distributions on $\mathcal{X}^V$; characterised by the joint mass function:

$$p(\boldsymbol{x}|J,G) = \frac{1}{\zeta(J)} \exp\left(J \sum_{v\sim u} \delta_{x_v,x_u}\right).$$

Here, recall $v \sim u$ denotes neighbouring pairs and δ_{x_v,x_u} is the Kronecker delta, i.e. δ_{x_v,x_u} is one if $x_v = x_u$, and zero otherwise. The intractable normalising constant (or partition function), written

$$\zeta(J) \doteq \sum_{\boldsymbol{x}'\in\mathcal{X}^V} \exp\left(J \sum_{v\sim u} \delta_{x'_v,x'_u}\right),$$

is a function of J.

The parameter J dictates how likely neighbouring random variable X_v and X_u are to have the same value. Given that the normalising function is a function of this parameter, J is typically very difficult to infer (Everitt, 2012); and will be treated as known in this thesis. This will be discussed in the sequel.

Given some graph $G = (V, E)$, and the associated random variable $\boldsymbol{X} = (X_v \in \mathcal{X} : v \in V)$, we write

$$\boldsymbol{X} \sim \text{Potts}(J, G, \mathcal{X})$$

to mean that the random variable $\boldsymbol{X}$, taking values in state space $\mathcal{X}^V$, is distributed according to the Potts model with given coupling constant J. Where no ambiguity arises, the parameters will be omitted from notation in the interests of clarity.

Many Monte Carlo methods, such as the Gibbs sampler, used to characterise the Ising distribution (as discussed in Section 2.8) will also readily extend to the Potts distribution. However, typically it is not possible to compute the full conditionals for posterior densities, such as those of interest here and discussed later; because the likelihood cannot be evaluated even up to a normalising constant. So we must turn to pseudo-marginal Monte Carlo methods to tackle this problem.

5.2.1 Critical Values of the Coupling Constant, J

Straightforwardly, a random variable with a Potts distribution is an MRF. In addition to simplifying the joint distribution of image (or more generally spatial) data to more manageable tasks; MRFs also readily lends themselves to component-wise update schemes in MCMC methods. The method presented in this thesis will use a component-wise MH proposals; another example of this is the Gibbs sampler, as already mentioned. However, when employing such approaches to target these models, it is important to give careful considerations of the coupling constant J.

An important property of the Potts model, similarly the Ising model, is that it exhibits phase transition behaviour. When the parameter $J = 0$, the random field is essentially a collection of independent random variables with uniform distribution, and all configurations are equally distributed. As $J > 0$ increases, the variables X_v has an increased probability to take the value of the most common state among its neighbours. Intuitively, as J tends to infinity, X_v will be the same value for all $v \in V$. The value could be any of the elements in the state space $\mathcal{X}$, showing clearly that this distribution is multi-modal.

This has important implications when using single-site update scheme MCMC methods, which could result in slower mixing; particularly for cases above critical values of J. More specifically, if the Markov chain is in a configuration such as the case described above, changes to single variables

will mean proposing to move to states of lower probability.

More specifically, phase transition behaviour happens close to or higher than critical values of the coupling constant, which we denote as J_{critical}. The phase transition is the sharp change in the macroscopic properties of the model with a small change in the value of the model's parameter, in this case J. That is, there is qualitatively different behaviour either side of this critical value. In the Monte Carlo setting, there is a sharp transition from fast mixing below J_{critical} and slow mixing above it. Strictly speaking, the phase transition behaviour described here is a property of Potts models on the *infinite lattice*, although it gives a good qualitative characterisation of the behaviour on finite lattices of the sort of interest here.

Onsager (1944) showed that for the Ising model on the two dimensional first order square lattice the exact value was

$$J_{\text{critical}} = \log\left(1 + \sqrt{2}\right) \approx 0.881.$$

More recently, see for instance Matveev and Shrock (1996) or Wu (1982), this has been extended: Let $D = |\mathcal{X}|$, then for a $D-$state Potts model,

$$J_{\text{critical}} = \log\left(1 + \sqrt{D}\right).$$

In applications, since the intractable normalising constant $\zeta(J)$ is dependent on J, it is difficult to infer in the model fitting process (Everitt, 2012). Within the Monte Carlo context, these types of distributions are referred to as *doubly-intractable*. Møller et al. (2006) introduced a ingenious method to address such problems for the Ising model. For a more recent work, and for a concise but detailed review of developments since Møller et al. (2006), see Moores et al. (2020) and references therein.

Additionally, J is a parameter of a prior distribution, thus under the Bayesian analysis framework it would not be too unreasonable for it to be specified beforehand. As such, in what follows, we will treat J as known. For example, in the numerical studies below, see Section 6.1.3, we use the value that roughly gives the best model selection performance. More specifically, the proposed algorithm is applied in a toy setting to perform model selection on a simulated image using a range of fixed values for J. Next, the performance of the algorithm for these varying values are calculated and the value of J which gives the best model selection performance is used for the numerical studies that follow.

5.3 Hidden Potts Model for Spatial Model Selection

We may now turn our attention to formally constructing and presenting the model used to effectively encode spatial dependence for image and image-like data. As aforementioned, a natural strategy towards this objective is to adapt existing non-spatial methods to a model that does incorporate spatial dependence. This strategy is particularly appealing when modelling spatial dependence via a Potts model as it can be efficiently targeted by MCMC methods with single-site update schemes. Below, we construct a class of models that allows for such computational methods.

5.3.1 Generic Model for Spatial Inference

Recall that a standard, intuitive way of modelling spatial or image data in a principled manner is to represent spatial relations using a graph. More formally, given spatial data (or spatial-temporal data, such as PET) $\boldsymbol{Y}$, associate a graph $G_{\boldsymbol{Y}} = (V_{\boldsymbol{Y}}, E_{\boldsymbol{Y}})$ that encodes the geometry of spatial dependence. That is, each component Y_v of $\boldsymbol{Y}$ is indexed by a node $v \in V_{\boldsymbol{Y}}$. Each data point Y_u is conditionally independent of Y_v given $Y_{-\{u,v\}}$ unless $\langle u, v\rangle \in E_{\boldsymbol{Y}}$. We will call Y_v the node data point at the node $v \in V$ and the whole data set $\boldsymbol{Y}$ as the image.

Example 5.3.1. For image data, a simple lattice graph would suffice. Consider a 2-dimensional image, then a square lattice encodes the notion that adjacent pixels are dependent; here we associate each pixel to a vector in the finite grid vertex-set $\mathbb{Z}^2_{[l_1,l_2]}$, and their dependence by the edges in the edge-set $E^2_{[l_1,l_2]}$. As before a lexicographic form may be used to simplify label notations, see Figure 5.2.

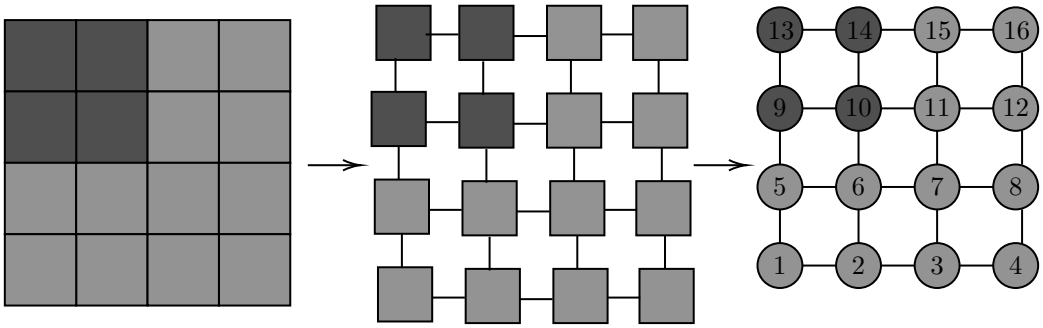


Figure 5.2: Representing a simple image using an MRF on a lexicographic order graph.

Of course, higher order neighbours could also be used see; for example Tjelmeland and Besag (1998) investigate the effects of doing so. In the interest of simplicity, here we will just assume first order neighbours only. △

Example 5.3.2. In cases such as PET images, the data set may be spatio-temporal, thus at each node the observed measurement will be a time series. For simplicity, we write this by letting the observational space be a subset of the Euclidean space, $Y_v \in \mathbb{R}^k$. △

Building good statistical models for the (whole) image $\boldsymbol{Y}$ can be difficult; model selection would involve selecting among many complex and computationally intensive models from the model space. Analysis in this setting can be simplified by using existing simpler models such as those that assume spatial independence and model data at the node-level.

Such an approach has two prominent strengths. Firstly, as aforementioned, many existing analysis and methodology would naturally make spatial-independence assumptions. This approach can readily exploit these technique, and allow them to be readily used to incorporate spatial dependence. Secondly, using this approach allows for considerable computational simplifications as will be seen later.

We briefly describe the construction of a generic hierarchical model based on this approach, we will detail with relevant specifics in the sequel.

Formally, consider parametric models for the node data point Y_v, for all $v \in V_{\boldsymbol{Y}}$, rather than the

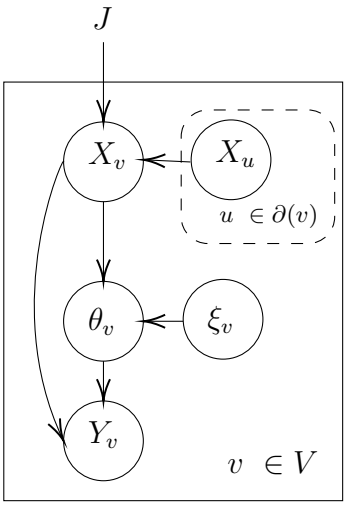


Figure 5.3: Graphical plates diagram of generic hierarchical model.

whole data set $\boldsymbol{Y}$. Denote the parameter of these models by θ_v, with parametric space Θ_v for each node $v \in V$.

It is immediate that a Potts model can be used to encode spatial relationships in any number of ways. One possibility is to use a Potts prior on the collection of parameters, however this is too restrictive — typically, parameters do not take values in small finite sets. Alternatively, introduce a new Potts random variable between models at different locations in space; define $\boldsymbol{X} = (X_v \in \mathcal{X} : v \in V_{\boldsymbol{Y}})$ a Potts MRF with respect to the graph $G_{\boldsymbol{Y}}$. Here, the coupling constant J is treated as known and constant.

More specifically, suppose we specify a (sub-)model for the nodal data point Y_v. Then, the parametric distribution for Y_v has parameters (θ_v, X_v). Importantly, X_v is a component of the Potts random variable $\boldsymbol{X}$, and so has spatial relations dictated by $E_{\boldsymbol{Y}}$. In the approach proposed here, the parameter X_v will determine the model associated with node v — more precisely, the model order at node v is given by X_v. However, for the remainder of this subsection, we will treat X_v as a generic parameter. Given a family of parameterised prior densities for each possible model denoted generically p, over the parameter space; we may compactly summarise this hierarchical model as:

$$\boldsymbol{X}|J, G_{\boldsymbol{Y}} \sim \text{Potts}(J, G_{\boldsymbol{Y}}), \tag{5.1}$$

$$\theta_v|X_v, \xi_v \sim p(\cdot|\xi_v, X_v) \text{ for all } v \in V_{\boldsymbol{Y}}, \tag{5.2}$$

$$Y_v|\theta_v, X_v \sim f(\cdot|\theta_v, X_v) \text{ for all } v \in V_{\boldsymbol{Y}}. \tag{5.3}$$

This generic model is represented as a plate diagram in Figure 5.3. Here, ξ_v is a parameter of the distribution p. In principle a hyper-prior could be attached to this parameter with no particular difficulty; we work with this simple setting in the interest of parsimony. To this end, we fix ξ_v to be a known hyper-parameter common to all nodes, v, and so we write ξ instead of ξ_v, in what follows.

In realistic settings, it may be that the parameter $\boldsymbol{\theta} = (\theta_v)_{v \in V}$ also exhibits spatial dependence. The above hierarchical model, makes the underlying assumption that it does not and this setting is the focus of the present work, where we have found the incorporation of spatial structure at the level of model sufficient to improve upon the state of the art for problems of interest. This is a by-product

of the strategy to build a spatial model, using simpler individual spatially independent models of each unitary data points. Making this assumption allows for the re-use of existing computational methods for the simpler models, and so results in significant reduction in computational overhead. Further generalisation would be possible and provides an interesting direction for further work. We encode this in Assumption 1 which will be a standing assumption throughout this thesis and is discussed in Section 5.3.2.

Assumption 1 (Conditional Independence)**.**

$$\theta_v | X_v, \xi \perp \boldsymbol{X}_{-v}, \boldsymbol{\theta}_{-v} \text{ for all } v \in V.$$

The primary interest of the proposed methodology is to infer $\boldsymbol{X}$. Although $\boldsymbol{\theta}$ *is* inferred as a by-product of the proposed method, it can be thought of as a latent variable in this framework and may be a nuisance parameter in some settings which mitigates the impact of modelling the parameters in this way.

Example 5.3.3. We describe here a simple toy model, based on the hierarchical model above, that will be used in simulation studies below. Essentially both the prior and model likelihood are specified to be normal distributions, at every node. A Potts distribution is used over the mean hyper-parameter of the prior distributions.

More precisely, given image $\boldsymbol{Y} = (y_v \in \mathbb{R} : v \in V)$, suppose each y_v is normally distributed with mean μ_v and known variance σ^2. Specify a prior over μ_v to be a normal distribution with mean $\mu_0^{(X_v)}$ and known variance σ_0^2. Here, the hyper-parameter $\mu_0^{(X_v)} \in \{-5, 5\}$ will be determined by X_v. Specifically, let $\boldsymbol{X} = (X_v \in \mathcal{X} : v \in V_{\boldsymbol{Y}})$ be a Potts random variable on a 2-dimensional square lattice. For instance, letting $\mathcal{X} = \{\mathrm{A}, \mathrm{B}\}$ and specify $\mu_0^{(\mathrm{A})} \doteq -5$ and $\mu_0^{(\mathrm{B})} \doteq 5$.

These choices then allow for the marginal likelihood at each node to be straightforwardly evaluated. In fact, it is simply equivalent to $\mathcal{N}(y_v; \mu_0^{(X_v)}, \sigma^2 + \sigma_0^2)$; That is, the value of the density of the normal distribution evaluated at y_v, with mean $\mu_0^{(X_v)}$ and variance $\sigma^2 + \sigma_0^2$. $\triangle$

Note that $\boldsymbol{X}$ need not only be an additional parameter of the generative model or likelihood, as shown in the example above. Even when used in addition to more general collection of parameter $\boldsymbol{\theta}$, a Potts distribution over most parameters would usually be too restrictive for application in realistic settings. However, this general framework would be pertinent and useful if used for *model selection* by specifying this distribution over the model orders at each node. This is discussed in the sequel.

5.3.2 Spatial Bayesian Model Selection and Assumption 1

To exploit the generic framework presented above for model selection for spatial data, denote by $\mathcal{S}_v \doteq \{S_{v,m_v} : m_v \in \mathcal{M}\}$ the model space used at each nodal data point Y_v. I.e. every node in the graph is associated with a model m_v from a set common to all nodes, $\mathcal{M}$. Thus a statistical model is associated with each node in the graph and hence each data point,

$$S_{v,m_v} \doteq \{f_{m_v}(\cdot|\theta_v) : \theta_v \in \Theta_{v,m_v}\}.$$

Recall, from Section 2.3, that we use $f_m(\cdot|\theta) = f(\cdot|\theta, m)$ to denote the likelihood pertaining to a statistical model with model order m. The notation Θ_{v,m_v} is used here to emphasise that since

m_v is the model order at node v, the parameter space will be dependent on it. For example, for compartmental models, the model order dictates the dimension of the parameter space. These collections of models can be used to generate a model space for the whole data set $\boldsymbol{Y}$. Specifically, a set of candidate models for $\boldsymbol{Y}$ can be formulated as

$$\mathcal{S}_{\boldsymbol{Y}} \doteq \{(S_{v,m_v}) : m_v \in \mathcal{M}, v \in V\}.$$

Writing $\boldsymbol{m} = (m_v)_{v\in V} \in \mathcal{M}^V$, Bayesian model selection of spatial data $\boldsymbol{Y}$, in this setting, can be thought of as inference of the model order parameter $\boldsymbol{M}$ in the space $\mathcal{M}^V$. Each realisation $\boldsymbol{m}$ of $\boldsymbol{M}$ is called a configuration; as aforementioned, there are $|\mathcal{M}|^{|V|}$ candidate models for $\boldsymbol{Y}$.

As mentioned before, the Potts model is a natural choice for the prior distribution over model order. For model selection, we adapt the hierarchical model above such that:

1. $\boldsymbol{M} = (M_v)_{v\in V}$ is a Markov random field with a Potts distribution, with spatial dependence represented by $E_{\boldsymbol{Y}}$ as before;
2. Each θ_v, for all $v \in V$, is dependent on M_v (e.g. model order determines parameter dimension in some cases, accordingly we denote the prior $p(\cdot|\xi, M_v) = p_{M_v}(\cdot|\xi)$) and known constant hyper-parameter ξ, importantly it is spatially independent (given the model order M_v);
3. Y_v, the observed data, has likelihood $f(\cdot|\theta_v, M_v) = f_{M_v}(\cdot|\theta_v)$, where the model M_v dictates which model is selected, for all $v \in V$.

Formally, the above can be summarised:

$$\boldsymbol{M} \sim \text{Potts}(J, G_{\boldsymbol{Y}}) \tag{5.4}$$
$$\theta_v|M_v, \xi \sim p_{M_v}(\cdot|\xi) \text{ for all } v \in V, \tag{5.5}$$
$$Y_v|\theta_v, M_v \sim f_{M_v}(\cdot|\theta_v) \text{ for all } v \in V. \tag{5.6}$$

Following the convention of f_{M_v}, we use the notation $p_{M_v}(\cdot|\xi)$ to emphasis that under the Bayesian framework a different prior distribution must be used for each model order.

Example 5.3.4. When considering compartmental models, straightforwardly we let the state space of the components of the Potts random variable be the collection of all possible number of compartments: For instance, when considering up to $m = 3-$compartment model, $\mathcal{M} = \{1, 2, 3\}$. Then, given image $\boldsymbol{y} = (y_v \in \mathbb{R}^k)$; each nodal data point (or time series) is modelled using an M_v-compartment model. Intuitively, spatial dependence, based on image $\boldsymbol{y}$, is encoded by specifying that $\boldsymbol{M} = (M_v \in \mathcal{M} : v \in V_{\boldsymbol{y}})$ is a Potts random variable on a graph with edges-set $E_{\boldsymbol{y}}$. △

Note that, at the nodal(pixel) level, when using parametric models Bayesian model selection would often involve the computation of the marginal likelihood; Given the new graphical setting, we may restate the marginal likelihood

$$f(y_v|m_v) = \int_{\Theta_{v,m_v}} f(y_v|\theta_v, m_v)p(\theta_v|m_v, \xi)\mathrm{d}\theta_v.$$

Finally, note that we are interested in the marginal likelihood of the whole image $\boldsymbol{Y}$, which in general may involve multiple intractable high-dimensional integrals. As a direct consequence of

the above model and Assumption 1, the marginal likelihood of $\boldsymbol{Y}$ can be written as the product of the marginal likelihoods of Y_v, over all $v \in V$. We discuss this further in Section 5.3.3 below.

Discussion of Assumption 1

Assumption 1 may not hold in general: if there is spatial dependence between the generative model which describes each vertex, then there may also be dependence between the parameters of those models. In some settings it can be viewed as an approximation which may be tolerable in order to allow inference under a model which is, at least, better than the assumption of full spatial independence. However, in some settings, such as model selection, this assumption is not too unrealistic. It is also noteworthy that, if the model order dictates the dimensions of the parameter space at that node, and the space contains dependent parameters, it may be difficult to incorporate spatial dependence at a the parameter level in a meaningful manner.

For example, if considering compartmental models: if two adjacent pixels were to contain different number of compartments, it may not be possible to say anything meaningful about the spatial dependence of the micro-parameters, such as the transfer rates. Of course, one possibility, then, is to instead consider macro-parameters. Doing so may mean that model selection may no longer be required — thus the opportunity to gain information from the model order output image may be lost. Additionally, in both of settings it may not be possible to use a Potts distribution, which is typically defined over a discrete space. Such considerations make it difficult to impose some more general MRF over the parameters.

Furthermore, incorporating spatial dependence at a model order level will typically take some precedence over spatial dependence at the parameter level. For instance, in the PET setting, generally we would like to know if two adjacent pixels have the same compartments (i.e. whether or not the localised area contains receptors) before we think about the transfer rates.

To summarise the problem of interest in general terms: the image $\boldsymbol{Y}$ is modelled node-wise using the distribution $F(\theta_v, M_v)$, with priors over the parameters $\boldsymbol{\theta}$ and $\boldsymbol{M}$. In particular, we propose the use of the Potts model as a prior over discrete parameters $\boldsymbol{M}$ to allow for tractable incorporation of spatial dependence in images. As mentioned above, $\boldsymbol{\theta}$ will be treated as a latent variable — this is discussed in detail next.

5.3.3 Inference from the Proposed Model

Having specified this class of models, we now turn to the task of inference. Let $p(\boldsymbol{m})$ denote the Potts prior probability mass function over $\boldsymbol{M} \in \mathcal{M}^V$. Then, the model posterior density, for $\boldsymbol{Y}$, is denoted

$$\pi(\boldsymbol{m}|\boldsymbol{y}) \propto f(\boldsymbol{y}|\boldsymbol{m})p(\boldsymbol{m}).$$

Here, $f(\boldsymbol{y}|\boldsymbol{m})$ will be called the *graph* marginal likelihood of $\boldsymbol{y}$; and represents the data generative function of the proposed model.

The proposed method allows us to incorporate spatial dependence at the level of the discrete parameter $\boldsymbol{M}$ only. Subsequently, for computational tractability we impose Assumption 1: given the model order M_v the random variable Y_v at each node $v \in V$ is independent. In other words, as per Eq.(5.2) and Eq.(5.3), we have that

$$f(\boldsymbol{y}|\boldsymbol{m}) = \prod_{v \in V} f(y_v|m_v).$$

Here $f(y_v|m_v)$ denotes the likelihood at each node v, and will be henceforth called the *node-wise* likelihood. This assumption is important as it makes the graph likelihood tractable, and provides computational simplifications discussed later. Typically, as described above, a parametric model is used, thus $f(y_v|m_v)$ will in fact be a marginal likelihood or model evidence (i.e. an integral, which is likely to be analytically intractable).

We finally have the probability mass function of primary computational interest

$$\begin{aligned}\pi(\boldsymbol{m}|\boldsymbol{y}) &\propto \left\{\prod_{v\in V} f(y_v|m_v)\right\} p(\boldsymbol{m}) \\ &\propto \left\{\prod_{v\in V}\int_{\Theta_v} f(y_v|\theta_v, m_v)p(\theta_v|m_v,\xi)\mathrm{d}\theta_v\right\} \times \exp\left(J\sum_{v\sim u}\delta_{m_v,m_u}\right). \end{aligned} \tag{5.7}$$

Recall the mild assumption that hyper-parameter ξ is known and henceforth suppressed in notation.

Even in this general formulation, this mass function is not particularly attractive. There are some positives: It is straightforward to evaluate up to the intractable normalising constant; and Assumption 1 does seemingly lead to a simpler computation even at this stage, as only the node-wise marginal likelihoods $f(y_v|m_v)$ need to be computed rather than the high-dimensional integration of the full marginal likelihood $f(\boldsymbol{y}|\boldsymbol{m})$.

However, even in this simpler setting, the graph marginal likelihood poses two computational difficulties: i) it is a product of $|V|$ integrals and ii) these integrals in general are analytically intractable in most cases of interest.

5.4 A Pseudo-marginal Algorithm for Graphical Model Selection

Having presented the model that allows for effective incorporation of spatial dependence, we can now present the proposed computational method that exploits this model. The algorithm presented is a natural extension of the generic pseudo-marginal approach, discussed in Section 2.7.2 above, to this specialised setting and class of models. The relevant differences in theoretical justifications of this novel computational method, and further extensions through space augmentation will also be discussed next in Section 5.5. Also discussed in the sequel is a straightforward approximation of the proposed algorithm.

Now, the current objective is to characterise the posterior density $\pi(\boldsymbol{m}|\boldsymbol{y})$ of the model given in Eq.(5.4)-(5.6). Suppose, for now, that a generic MH algorithm, as discussed in Section 2.6.1, Algorithm 1, could be used in this setting. That is, to generate a MH Markov chain that targets the posterior density. This then gives the acceptance ratio (adapting Eq.(2.17) to the current context):

$$R(\boldsymbol{m}, \boldsymbol{m}^*) = \frac{\pi(\boldsymbol{m}^*|\boldsymbol{y})\nu(\mathbf{m}^*, \boldsymbol{m})}{\pi(\boldsymbol{m}|\boldsymbol{y})\nu(\boldsymbol{m}, \boldsymbol{m}^*)}$$

Even with a tractable marginal likelihood, computing this acceptance ratio would be costly. Under Assumption 1 (given $\boldsymbol{M}$ the parameters $\boldsymbol{\theta} = (\theta_v)_{v\in V}$, of the model over the data $\boldsymbol{y}$, are independent) the computational complexity of the problem is reduced considerably; now, only the node-wise marginal likelihoods need to be computed. However, every time a new configuration is

proposed, up to $|V|$ new integrals would then need to be computed and the acceptance probability is likely to be very small. Given that the number of possible Potts configurations is $|\mathcal{M}|^{|V|}$, growing geometrically with the number of nodes (i.e. pixels, for image data), the computational costs are not appealing.

Under Assumption 1 it is natural to employ a Metropolis–within–Gibbs approach to mitigate these difficulties and to obtain significant computational simplifications. Suppose that the MCMC sampler is currently at some configuration $\boldsymbol{m} = (m_v)_{v\in V}$. Under a single variable updating schedule: *only* the model order m_v, at some node $v \in V$, will be proposed to be changed in a given step. We will refer to this as the node-wise updating schedule, since we update the model order at a single node v only. Denote the proposed state to be m_v^*; and $\boldsymbol{m}_{[v]}^*$ to be the vector equivalent to $\boldsymbol{m}$ but with m_v^* as the v-th coordinate. For simplicity, assume that the proposal distribution is symmetric, and thus the acceptance ratio is

$$\begin{aligned}\frac{\pi(\boldsymbol{m}_{[v]}^*|\boldsymbol{y})}{\pi(\boldsymbol{m}|\boldsymbol{y})} &= \frac{f(\boldsymbol{y}|\boldsymbol{m}_{[v]}^*)p(\boldsymbol{m}_{[v]}^*)}{f(\boldsymbol{y}|\boldsymbol{m})p(\boldsymbol{m})} \\ &= \frac{f(y_v|m_v^*)p(\boldsymbol{m}_{[v]}^*)}{f(y_v|m_v)p(\boldsymbol{m})} \\ &= \frac{\int f(y_v|\theta_v, m_v^*)p(\theta_v|m_v^*)\mathrm{d}\theta_v p(\boldsymbol{m}_{[v]}^*)}{\int f(y_v|\theta_v, m_v)p(\theta_v|m_v)\mathrm{d}\theta_v p(\boldsymbol{m})}.\end{aligned}$$

Thus, the acceptance probability of each Metropolis-within-Gibbs step requires the computation of only a single integral. A complete sweep over the graph thus requires $|V|$ such integrals but one could expect a reasonable proportion of these proposed moves to be accepted, in contrast to a global proposal which sought to update every node simultaneously.

Typically, even the node-wise marginal likelihood $f(y_v|m_v)$ will be difficult or impossible to evaluate analytically. A pseudo-marginal MH algorithm is a natural choice in such cases, accordingly its extension to this context explored next.

5.4.1 Graph Model Selection using Node-wise Marginal Likelihood Estimates

The node-wise update schedule, gives rise to nested iterations. For clarity, term the outer iteration, indexed with i, of a full pass through the whole graph as the *graphical iteration*.The inner iteration over V will be termed the *node-wise iteration*. To be precise, one graphical iteration would involve computing the acceptance ratio for all $v \in V$.

Suppose now that the MH Markov chain denoted, $(\boldsymbol{M}^{(i)})_{i=1}^n$, is at graphical iteration i and that the state at node v is being proposed to be changed. Since the node-wise marginal likelihood is a normalising constant, following notations of Section 2.7, denote

$$Z_v(m) \doteq f(y_v|M_v = m) \text{ for all } m \in \mathcal{M}.$$

That is, $Z_v(m)$ is the normalising constant of the parameter posterior density at node v when the model order $M_v = m$. For clarity of presentation, we use the convention of denoting the shorthand for a approximations of $Z_v(m)$ by $\widehat{Z}_v(m)$.

More specifically, within the aforementioned MH chain setting, at graphical iteration i, node v, for given model order proposal $m_v^{(i)}$, let $(\widehat{Z}_v)^{(i)}$ denote the approximation of the marginal $f(y_v|M_v =$

$m_v^{(i)}$). Associate with this generic estimator, the computation cost parameter N (for the SMC sampler this was the number of particles, see Eq.(3.9), Section 3.2). As mentioned before, this is a tuning parameter for the pseudo-marginal, thus also the algorithm presented below.

It is important to note the slight abuse of notation here — $(\widehat{Z}_v)^{(i)}$ is dependent on the model order proposal $m_v^{(i)}$. As previously stated, we require that $(\widehat{Z}_v)^{(i)}$ is unbiased. Additionally, recall we denote $g(\cdot|M_v = m_v^{(i)})$ the possibly unknown density of $(\widehat{Z}_v)^{(i)}$.

Following the node-wise updating schedule, whenever a new state $m_v^* \sim \nu(m_v^{(i-1)}, \cdot)$ is proposed, a new node-wise marginal estimation, denoted $(\widehat{Z}_v)^*$, is recomputed. The node-wise acceptance ratio is, then,

$$\widehat{r}(\boldsymbol{m}^{(i-1)}, \boldsymbol{m}_{[v]}^{(i-1)*}) \doteq \frac{p(\boldsymbol{m}_{[v]}^*)(\widehat{Z}_v)^* \nu(m_v^*, m_v^{(i-1)})}{p(\boldsymbol{m}^{(i-1)})(\widehat{Z}_v)^{(i-1)} \nu(m_v^{(i-1)}, m_v^*)}.$$

We emphasise that the computation of $\widehat{r}$, precisely speaking, involves not just the standard MH proposal of a new configuration in $\mathcal{M}^V$, but also a proposal of the random variable $\widehat{Z}_v$: it is a pseudo-marginal algorithm in the sense of Andrieu and Roberts (2009), with the added subtlety that the unbiased estimates of the marginal likelihood associated with every other node in the graph are also retained as a part of the extended state. The justification for using this acceptance ratio will be detailed in Section 5.5.1

The above choices and assumptions lead to the algorithm which we term the Node-wise Pseudo-marginal (NWPM) algorithm; presented in pseudo-code in **Algorithm 4** below.

Algorithm 4 The Node-wise Pseudo-marginal Algorithm

Given: (i) target density $\mu(\boldsymbol{m}) \doteq \pi(\boldsymbol{m}|\boldsymbol{y})$; (ii) unbiased node-wise marginal likelihood estimators $\widehat{Z}_v(m) \approx \int f(y_v|m, \theta_v)p(\theta_v|m, \xi)d\theta_v$ for all $m \in \mathcal{M}$, $v \in V$; and (iii) coupling constant J, tuning parameter N.

1. At $i = 1$:
   (a) Initialise $\boldsymbol{m}^{(1)}$.
   (b) Sample $(\widehat{Z}_v)^{(1)} \sim g(\cdot|m_v^{(1)})$ for each $v \in V$.
2. For $i = 2, \ldots, n$:
   (a) For $v \in V$:
      i. Sample: $m_v^* \sim \nu(m_v^{(i-1)}, \cdot)$.
      ii. Sample: $(\widehat{Z}_v)^*|m_v^* \sim g(\cdot|m_v^*)$.
      iii. Compute:
$$\hat{r} = \frac{p(\mathbf{m}_{[v]}^*)(\widehat{Z}_v)^* \nu(m_v^*, m_v^{(i-1)})}{p(\mathbf{m}^{(i-1)})(\widehat{Z}_v)^{(i-1)} \nu(m_v^{(i-1)}, m_v^*,)}$$
      iv. With probability $\min\{1, \widehat{r}\}$ take:
$$M_v^{(i)} = m_v^* \text{ and } (\widehat{Z}_v)^{(i)} = (\widehat{Z}_v)^*$$
   otherwise,
$$M_v^{(i)} = m_v^{(i-1)} \text{ and } (\widehat{Z}_v)^{(i)} = (\widehat{Z}_v)^{(i-1)}$$

Within the context of this study, the NWPM algorithm proposed here is a broadly applicable extension of the GIMH algorithm described above in **Algorithm 2**.

The output of this algorithm, the Monte Carlo sample $(\boldsymbol{M}^{(i)})_{i=1}^n$, can be used to perform model

selection on image $\boldsymbol{y}$ in a principled manner. For example, this could be done using the node-wise marginal modal model order : That is, at each node $v \in V$ the model order which occurs the most in the (marginal) chain $(M_v^{(i)})_{i=1}^n$ is selected. Additionally, samples of the latent variables $\boldsymbol{\theta}$ are typically available as by-products of the marginal likelihood estimates and can be used to perform parameter inference as demonstrated in the numerical studies below.

An obvious example of application is PET data model selection; a field in which there are still many open and unanswered questions that could be further studied using this methods. However, this approach could also be generalised to other similar settings and problems. In theory, it can be used for spatial dependent model selection for any image and image-like data. An advantage of this pseudo-marginal based approach in this setting is its flexibility. It can be used with essentially any unbiased estimator of marginal likelihood and hence allows existing technology from any application domain to be employed.

The incorporation of spatial dependence via the modelling framework and algorithm developed here can come at the very little computational cost by approximating or making pragmatic adaptations to **Algorithm** 4. We explore such techniques in the simulation studies presented in Chapter 6 and discuss in more detail in Section 5.6.1 below.

Finally, note that for convenience we assume that the tuning parameter N is fixed to the same value for all $v \in V$. This parameter typically determines the variance of the estimator $\widehat{Z}_v(m)$ and additional flexibility could be obtained by allowing it to vary between vertices. Unsurprisingly, a larger N should give a better performance at the expense of greater computational cost. Indeed Andrieu and Roberts (2009), and more recently Andrieu and Vihola (2016), showed that when the dispersion of the marginal likelihood estimates are larger the mixing rate of any pseudo-marginal MH algorithm decreases. Of course, there is a higher computational cost when using larger N to reduce the estimate variance. The effects of this tuning parameter, and the trade off, on the mixing of the Markov chain in a standard pseudo-marginal algorithm have been studied by Sherlock et al. (2015); Doucet et al. (2015) and Sherlock (2016). These works give theoretical justification and numerical studies for the recommendation that N should be chosen such that the variance of the log-likelihood estimator is close to one.

5.5 Theoretical Considerations

5.5.1 Marginal Invariant Distribution from Approximate Marginal Likelihood

Here, we establish the formal validity of a class of algorithms which includes the NWPM, allowing for more general moves than those described in **Algorithm** 4. We prove a slightly more general result than the NWPM setting. For further discussion on specifying proposals see Section 5.5.2. We now show that the invariant distribution, of the Markov chain generated when following NWPM algorithm above, is the posterior density π, as shown in Eq.(5.7).

Recalling Assumption 1, we may write the target distribution of the NWPM algorithm as

$$\mu(\boldsymbol{m}) \propto p(\boldsymbol{m}) \prod_{v \in V} \int_{\Theta_v} f(y_v|\theta_v, m_v) p(\theta_v|m_v, \xi) \mathrm{d}\theta_v,$$

for $\boldsymbol{m} = (m_v)_{v \in V} \in \mathcal{M}^V$.

Given $\boldsymbol{M}$, denote the random vector of the unbiased node-wise marginal likelihood estimators as

$\widehat{\boldsymbol{Z}}|\boldsymbol{M} \sim g(\cdot|\boldsymbol{M})$, i.e.,

$$\widehat{\boldsymbol{Z}} \doteq (\widehat{Z}_v(M_v) : v \in V).$$

Henceforth using the shorthand $\widehat{Z}_v^m \doteq \widehat{Z}_v(m)$ for all $m \in \mathcal{M}$; we note that, due to independence,

$$g(\widehat{\boldsymbol{Z}}|\boldsymbol{M} = \boldsymbol{m}) \propto \prod_{v \in V} g(\widehat{Z}_v^{m_v}|m_v).$$

Finally, define a pseudo-marginal target density

$$\widehat{\mu}(\boldsymbol{m}, \widehat{\boldsymbol{Z}}) = \frac{p(\boldsymbol{m}) \prod_{v \in V} \widehat{Z}_v^{m_v} g(\widehat{Z}_v^{m_v}|m_v)}{\sum_{\boldsymbol{m}' \in \mathcal{M}^V} p(\boldsymbol{m}') \prod_{v \in V} f(y_v|m'_v)} \tag{5.8}$$

$$= \frac{p(\boldsymbol{m}) \prod_{v \in V} \widehat{Z}_v^{m_v} g(\widehat{Z}_v^{m_v}|m_v)}{f(\boldsymbol{y})}. \tag{5.9}$$

Here, we let $f(\mathbf{y})$ denote the marginal probability of the observed data, integrating out unknown parameters and unknown model orders; Given observation $\boldsymbol{Y} = \boldsymbol{y}$, we treat it as a normalising constant. The NWPM algorithm has similar theoretical properties as GIMH. As such, the remainder of the argument of formal justification then follows the same as the GIMH case, see Andrieu and Roberts (2009).

Proposition 5.5.1. Let $\boldsymbol{M}$ denote a random variable in the graph model order space $\mathcal{M}^V$. Let $\widehat{\boldsymbol{Z}}|\boldsymbol{M}$, g, μ and $\widehat{\mu}$ be defined as above.

For any $U \subset V$, let ν_U denote the proposal density, a Markov kernel on $(\mathcal{M} \times \mathbb{R}_+)^V$, with the form

$$\nu_U((\boldsymbol{m}, \widehat{\boldsymbol{Z}}), (\boldsymbol{m}^*, \widehat{\boldsymbol{Z}}^*)) = \delta_{(\boldsymbol{m}_{-U}, \widehat{\boldsymbol{Z}}_{-U})}(\boldsymbol{m}^*_{-U}, \widehat{\boldsymbol{Z}}^*_{-U}) \times \nu_U^{\mathcal{M}}(\boldsymbol{m}_U, \boldsymbol{m}^*_U) \prod_{v \in U} g(\widehat{Z}_v^*|m_v^*),$$

where $\nu_U^{\mathcal{M}}$ denotes a Markov kernel on $\mathcal{M}^U$ and we slightly abuse density notation using the Dirac delta functions to indicate that variables associated with nodes outside U are unchanged (absolute continuity of the numerator and denominator of the Metropolis-Hastings ratio is ensured by the symmetry of this singular part of the kernel).

The standard Metropolis-Hasting acceptance probability with target distribution $\widehat{\mu}$ can be expressed as:

$$1 \wedge \frac{p(\boldsymbol{m}^*)(\prod_{v \in U} \widehat{Z}_v^*)\nu_U^{\mathcal{M}}(\boldsymbol{m}^*_U, \boldsymbol{m}_U)}{p(\boldsymbol{m})(\prod_{v \in U} \widehat{Z}_v)\nu_U^{\mathcal{M}}(\boldsymbol{m}_U, \boldsymbol{m}^*_U)} \tag{5.10}$$

and the marginal distribution of $\boldsymbol{M}$ under $\widehat{\mu}$ is μ.

Proof. First, consider the acceptance ratio of a Metropolis-Hastings algorithm with target distribution $\widehat{\mu}$ and proposal kernel ν_U:

$$\hat{r} = \frac{\widehat{\mu}(\boldsymbol{m}^*, \widehat{\boldsymbol{Z}}^*)\nu_U((\boldsymbol{m}^*, \widehat{\boldsymbol{Z}}^*), (\boldsymbol{m}, \widehat{\boldsymbol{Z}}))}{\widehat{\mu}(\boldsymbol{m}, \widehat{\boldsymbol{Z}})\nu_U((\boldsymbol{m}, \widehat{\boldsymbol{Z}}), (\boldsymbol{m}^*, \widehat{\boldsymbol{Z}}^*))}$$

upon inserting the definition of ν_U one observes that the numerator is absolutely continuous with respect to the denominator and the singular elements simply impose that $\boldsymbol{m}_{-U} = \boldsymbol{m}^*_{-U}$ and

$\widehat{\boldsymbol{Z}}_{-U} = \widehat{\boldsymbol{Z}}^*_{-U}$, and we have:

$$\begin{aligned}\hat{r} =& \frac{\widehat{\mu}(\boldsymbol{m}^*, \widehat{\boldsymbol{Z}}^*)\nu_U^{\mathcal{M}}(\boldsymbol{m}^*_U, \boldsymbol{m}_U) \prod_{v\in U} g(\widehat{Z}_v|m_v)}{\widehat{\mu}(\boldsymbol{m}, \widehat{\boldsymbol{Z}})\nu_U^{\mathcal{M}}(\boldsymbol{m}_U, \boldsymbol{m}^*_U) \prod_{v\in U} g(\widehat{Z}^*_v|m^*_v)} \\ =& \frac{p(\boldsymbol{m}^*)}{p(\boldsymbol{m})} \prod_{v\in V} \frac{\widehat{Z}^*_v g(\widehat{Z}^*_v|m^*_v)}{\widehat{Z}_v g(\widehat{Z}_v|m_v)} \times \frac{\nu_U^{\mathcal{M}}(\boldsymbol{m}^*_U, \boldsymbol{m}_U)}{\nu_U^{\mathcal{M}}(\boldsymbol{m}_U, \boldsymbol{m}^*_U)} \prod_{v\in U} \frac{g(\widehat{Z}_v|m_v)}{g(\widehat{Z}^*_v|m^*_v)} \\ =& \frac{p(\boldsymbol{m}^*)}{p(\boldsymbol{m})} \prod_{v\in U} \frac{\widehat{Z}^*_v}{\widehat{Z}_v} \times \frac{\nu_U^{\mathcal{M}}(\boldsymbol{m}^*_U, \boldsymbol{m}_U)}{\nu_U^{\mathcal{M}}(\boldsymbol{m}_U, \boldsymbol{m}^*_U)} \underbrace{\prod_{v\notin U} \frac{\widehat{Z}^*_v g(\widehat{Z}^*_v|m^*_v)}{\widehat{Z}_v g(\widehat{Z}_v|m_v)}}_{=1},\end{aligned}$$

where the final factor is equal to one almost surely under the proposal distribution, and the result follows.

The marginal distribution of $\boldsymbol{m}$ follows by the essentially the same argument as in the standard pseudo-marginal context.

$$\begin{aligned}\int_{\mathbb{R}_+^V} \widehat{\mu}(\boldsymbol{m}, \widehat{\boldsymbol{Z}})\mathrm{d}\widehat{\boldsymbol{Z}} &= \frac{p(\boldsymbol{m})}{f(\boldsymbol{y})} \prod_{v\in V} \int_{\mathbb{R}_+} \widehat{Z}_v^{m_v} g(\widehat{Z}_v^{m_v}|m_v)\mathrm{d}\widehat{Z}_v^{m_v} \\ &= \frac{p(\boldsymbol{m})}{f(\boldsymbol{y})} \prod_{v\in V} \mathbb{E}_g[\widehat{Z}_v(m_v)] \\ &= \frac{p(\boldsymbol{m}) \prod_{v\in V} f(y_v|m_v)}{f(\boldsymbol{y})} \\ &= \mu(\boldsymbol{m})\end{aligned}$$

□

Clearly, the node-wise proposals described previously fit this framework with $U = \{u\}$, for $u \in V$. being the single node being updated, although this framework would allow a somewhat broader class of proposals and blocked Metropolis-within-Gibbs type strategies to be explored. Standard arguments mean that a mixture or cycle of such kernels will also preserve $\widehat{\mu}$ as the invariant distribution.

Next, we discuss some immediate extensions and innovations of this proposed methodology; resulting in further algorithms, approximations and future approaches.

5.5.2 Multiple Augmentation Pseudo-marginal Algorithms

In this section we will explore the potential of further augmentation of the state space. The specialised setting of the problem, allows approaches which to the author's knowledge, have not been previously studied. The strategy developed here allows for a number of further extensions. Some simple approaches are discussed below — methods based on these approaches will also be evaluated in the numerical studies.

The augmentation of the state space used to justify node-wise pseudo-marginal algorithms can be further extended by adding an estimate of the marginal likelihood associated with *every* possible model at *every* node to the state space. Although doing so may seem counter-intuitive and leads to a rather large state space, it allows a number of algorithmic innovations. In particular, by

considering the following extended state space it will be possible to use a variety of standard MH moves in order to explore this space.

Let $\bar{\boldsymbol{Z}} \doteq (\widehat{Z}_v(m) : v \in V, m \in \mathcal{M})$ be a vector of marginal likelihood estimator for each model order $m \in \mathcal{M}$, at every node $v \in V$. Consider, next, an extended form of the pseudo-marginal target density (5.9) :

$$\bar{\mu}(\boldsymbol{m}, \bar{\boldsymbol{Z}}) \doteq \frac{p(\boldsymbol{m})}{f(\boldsymbol{y})} \prod_{v \in V} \widehat{Z}_v^{m_v} \left(\prod_{m' \in \mathcal{M}} g(\widehat{Z}_v^{m'} | m') \right).$$

The extended joint density $\bar{\mu}$ is constructed such that its marginal over $\boldsymbol{m}$ coincides exactly with the *correct posterior distribution.* Indeed, using essentially the same argument as in Proposition 5.5.1, we have:

$$\begin{aligned}
\int_{\mathbb{R}_+^{V \times \mathcal{M}}} \bar{\mu}(\boldsymbol{m}, \bar{\boldsymbol{Z}}) d\bar{\boldsymbol{Z}} &= \frac{p(\boldsymbol{m})}{f(\boldsymbol{y})} \prod_{v \in V} \int_{\mathbb{R}_+^{\mathcal{M}}} \widehat{Z}_v^{m_v} \left(\prod_{m' \in \mathcal{M}} g(\widehat{Z}_v^{m'} | m') \mathrm{d}\widehat{Z}_v^{m'} \right) \\
&= \frac{p(\boldsymbol{m})}{f(\boldsymbol{y})} \prod_{v \in V} \int_{\mathbb{R}_+} \widehat{Z}_v^{m_v} g(\widehat{Z}_v^{m_v} | m_v) \mathrm{d}\widehat{Z}_v^{m_v} \\
&= \mu(\boldsymbol{m}).
\end{aligned}$$

This further extended target distribution allows for some generalisations of the standard pseudo-marginal algorithm in the context of interest; it depends fundamentally on the fact that the variables associated with each node of the graph take values within a small finite set. It is possible to consider a variety of MH like moves applied to this extended target density. For simplicity we will consider only two types of proposal here: one which changes a single node's associated model and another which refreshes the likelihood estimates for a single node.

Proposal Densities for the Augmented Space

Denote the proposal of the random vector $\bar{\boldsymbol{Z}} \doteq (\widehat{Z}_v^m : v \in V, m \in \mathcal{M})$ by $\bar{\boldsymbol{Z}}^* \doteq ((\widehat{Z}_v^m)^* : v \in V, m \in \mathcal{M})$. Intuitively, $\bar{\boldsymbol{Z}}^*$ denotes a re-computation of all the components of $\bar{\boldsymbol{Z}}$.

First consider a proposal q_u^1 for the model associated with node u. Recalling that $\boldsymbol{m}_{[u]}^*$ is equivalent to $\boldsymbol{m}$ with component $m_u^* \neq m_u$, we have

$$\begin{aligned}
q_u^1\left((\boldsymbol{m}, \bar{\boldsymbol{Z}}), (\boldsymbol{m}^*, \bar{\boldsymbol{Z}}^*)\right) &\doteq \left(\nu_u^1(m_u, m_u^*) \prod_{v \neq u} \delta_{m_v, m_v^*} \right) \prod_{v \in V} \prod_{m \in \mathcal{M}} \delta_{\widehat{Z}_v^m}((\widehat{Z}_v^m)^*) \\
&= \underbrace{\delta_{\boldsymbol{m}_{-u}, \boldsymbol{m}_{-u}^*} \nu_u^1(m_u, m_u^*)}_{\text{Gibbs-like state proposal}} \underbrace{\prod_{v \in V} \prod_{m \in \mathcal{M}} \delta_{\widehat{Z}_v^m}((\widehat{Z}_v^m)^*)}_{\text{Same marginal estimates}}. \qquad (5.11)
\end{aligned}$$

Here $\delta_{x,y}$ is the Kronecker delta, taking value 1 if $x = y$ and 0 otherwise and $\delta_z(z^*)$ is, with the obvious abuse of notation, the singular measure concentrated at z evaluated over an infinitesimal neighbourhood of z^*. In this proposal, only a change of m_u^* is proposed — in particular, $\bar{\boldsymbol{Z}}$ and

$\boldsymbol{m}_{-u}$ do not change. Subsequently, the usual MH acceptance probability for this move is:

$$\begin{aligned} &1 \wedge \frac{\bar{\mu}(\boldsymbol{m}^*_{[u]}, \bar{\boldsymbol{Z}})\nu^1_u(m^*_u, m_u)}{\bar{\mu}(\boldsymbol{m}, \bar{\boldsymbol{Z}})\nu^1_u(m_u, m^*_u)} \\ =&1 \wedge \frac{\widehat{Z}_u^{m^*_u} p(\boldsymbol{m}^*_{[u]})\nu^1_u(m^*_u, m_u)}{\widehat{Z}_u^{m_u} p(\boldsymbol{m})\nu^1_u(m_u, m^*_u)}. \end{aligned}$$

For simplicity we have assumed that $\nu^1_u(m_u, m^*_u)$ is independent of the remaining state variables, but this is not necessary.

Similarly, consider a proposal q^2_u which refreshes the augmenting variables associated with node u:

$$q^2_u\left((\boldsymbol{m}, \bar{\boldsymbol{Z}}), (\boldsymbol{m}^*, \bar{\boldsymbol{Z}}^*)\right) = \delta_{\boldsymbol{m},\boldsymbol{m}^*} \prod_{m \in \mathcal{M}} \left(g((\widehat{Z}_u^m)^*|m) \prod_{v \neq u} \delta_{\widehat{Z}_v^m}((\widehat{Z}_v^m)^*) \right), \tag{5.12}$$

for which the MH acceptance probability is simply

$$1 \wedge \frac{(\widehat{Z}_u^{m_u})^*}{\widehat{Z}_u^{m_u}}$$

by exploiting the same cancellation as in the standard pseudo-marginal setting. Clearly moves which update only some of the augmenting variables associated with node u could be justified in the same way.

Standard arguments allow the combination of moves of these types, and others, within mixtures or cycles to provide an irreducible chain allowing considerable flexibility. In particular, it is no longer necessary to sample a marginal likelihood estimate for every proposed move in the state space. In the numerical studies below, we evaluate the performance of using simple combinations of such proposals.

5.6 Approximations of the NWPM Algorithm

We finish this section with a brief exploration of approximations of the exact pseudo-marginal algorithm described above, with the aim of obtaining inference which is almost as good at a fraction of the computational cost by allowing for a small bias in those estimates.

Clearly, a significant portion of the computational load is used to compute the marginal likelihood; this is especially the case when using more sophisticated methods such as the SMC sampler. For the NWPM method, as presented in Algorithm 4, the pseudo-marginal MH Markov chain of length n requires $n|V|$ marginal likelihood estimates. This quickly becomes infeasible with large data sets such as PET images, which may contain up to 10^6 (voxels) time series to be analysed. It is, therefore, worthwhile to consider strategies that reduce the number of marginal likelihood estimates that are required.

5.6.1 NWSE : Single Estimation Approximation of the NWPM Algorithm

The multiple augmentation approach to the NWPM algorithm decouples the estimation of marginal likelihoods from moves within the state space. Taken to the extreme, one can significantly reducing the frequency of calling SMC sampler is to only call it once for each model at each node. Naturally, this approach leads to a Markov chain which is not irreducible on the extended space and which can no-longer be considered a pseudo-marginal algorithm.

Such an algorithm would simply sample each of the marginal likelihood of all the model orders only once, for each node, before starting the chain; then, follow Algorithm 4 otherwise, using only these initial single estimates. This amounts to making a stochastic approximation which will change the invariant distribution of the resulting Markov chain.

Doing so leads to a (now marginal) MH Markov chain, that targets an approximation of the target density rather than the target density itself. As such, it would not be formally justified under the pseudo–marginal framework, and there would be a loss of accuracy with any subsequent results, since it is an approximation. Since we only do a single estimation of the marginal likelihoods we will refer to this method as the Node-wise Single-Estimation (NWSE) algorithm.

The biggest appeal of the NWSE is that, for a chain of the same length n, the sampler would only need to be used $|V||\mathcal{M}|$ times, reducing the costs by $n/|\mathcal{M}|$ times. Any reduction in accuracy and performance could be adjusted for by using some of the saved residual computational resources towards reducing the variance of the sampler marginal likelihood estimates.

This approximation would allow at least preliminary spatial analyses to be conducted with little additional computational cost beyond that required for the associated mass-univariate analysis: if existing analysis has been done, and thus some estimates of the marginal likelihood have already been obtained then these can be readily used within the algorithm to incorporate spatial dependence.

The following elementary proposition demonstrates that in the context of small discrete spaces the error introduced by approximating marginal likelihoods can be controlled under reasonable conditions.

Proposition 5.6.1. For each model $m \in \mathcal{M}$, let $\widehat{Z}^m := \widehat{Z}(m)$ be a RV with expectation $Z^m := f(y|m)$, corresponding to the associated marginal likelihood, and variance $\sigma_m^2 < \sigma_*^2 < \infty$. If the $(\widehat{Z}^m)_{m\in\mathcal{M}}$ are mutually independent, then, letting $m^* = \arg\max\{Z^m\}_{m\in\mathcal{M}}$ which we assume to be unique the probability of selecting the correct model via maximization of the marginal likelihood (equivalently, the posterior mode of the distribution over models given a uniform prior over $\mathcal{M}$) is at least:

$$1 - (|\mathcal{M}| - 1)\left(1 + \frac{\Delta^2}{2\sigma_*^2}\right)^{-1},$$

where $\Delta = \min_{l\neq m^*} Z^{m^*} - Z^l$ and $\sigma_*^2 = \max_{l\in\mathcal{M}} \sigma_l^2$.

Proof. Let $m^* = \arg\max\{Z^m\}_{m\in\mathcal{M}}$, then:

$$\begin{aligned}
&\mathbb{P}(\{\widehat{Z}^{m^*} > \widehat{Z}^l \forall l \neq m^*\}) \\
&\quad = 1 - \mathbb{P}(\cup_{l\neq m^*}\{\widehat{Z}^{m^*} \leq \widehat{Z}^l\}) \\
&\quad \geq 1 - \sum_{l\neq m^*} \mathbb{P}(\{\widehat{Z}^{m^*} \leq \widehat{Z}^l\}) \\
&\quad \geq 1 - (|\mathcal{M}| - 1)\left(1 - \min_{l\neq m^*} \mathbb{P}(\widehat{Z}^{m^*} > \widehat{Z}^l)\right),
\end{aligned}$$

by the Union bound. Applying Cantelli's inequality, noting that $Z^{m^*} - Z^l > 0$:

$$\begin{aligned}
&\mathbb{P}(\{\widehat{Z}^{m^*} > \widehat{Z}^l\}) \\
&\quad = \mathbb{P}(\{\widehat{Z}^{m^*} - \widehat{Z}^l - (Z^{m^*} - Z^l) > -(Z^{m^*} - Z^l)\}) \\
&\quad \geq 1 - \frac{\sigma^2_{m^*} + \sigma^2_l}{\sigma^2_{m^*} + \sigma^2_l + (Z^{m^*} - Z^l)^2} \\
&\quad = 1 - \left(1 + \frac{(Z^{m^*} - Z^l)^2}{\sigma^2_{m^*} + \sigma^2_l}\right)^{-1}.
\end{aligned}$$

Combining these expressions yields:

$$\begin{aligned}
&\mathbb{P}(\{\widehat{Z}^{m^*} > \widehat{Z}^l \forall l \neq m^*\}) \\
&\geq 1 - (|\mathcal{M}| - 1) \max_{l\neq m^*}\left[1 - 1 + \left(1 + \frac{(Z^{m^*} - Z^l)^2}{\sigma_{m^*} + \sigma^2_l}\right)^{-1}\right] \\
&\geq 1 - (|\mathcal{M}| - 1)\left(1 + \frac{\min_{l\neq m^*}(Z^{m^*} - Z^l)^2}{2\max_{l\in\mathcal{M}} \sigma^2_l}\right)^{-1} \\
&= 1 - (|\mathcal{M}| - 1)\left(1 + \frac{\Delta^2}{2\sigma^2_*}\right)^{-1}.
\end{aligned}$$

□

This bound can be made arbitrarily close to 1 by choosing estimators with sufficiently small variance. As the variance of the normalising constant estimates needs to be assessed to allow pseudo-marginal algorithms to be tuned, it is reasonable to suppose that this information could be obtained in settings in which this type of method is used and hence that a reasonable degree of confidence can be obtained that the use of this approximation does not substantially influence the resulting inference. Naturally, this result suggests that if estimating each marginal likelihood once rather than for every algorithmic step as in a pseudo-marginal algorithm, a relatively small variance might be required to obtain good performance.

In the numerical study, presented below, we investigate, evaluate and compare the effects when using the NWSE algorithm in different settings.

5.7 Summary

The general strategy used here was essentially the standard philosophy of breaking larger tasks into smaller tasks — or rather, to build a bigger, complex model using a smaller, simpler model. Combining this with a distribution that sufficiently encodes spatial relations in a simple, intuitive manner readily gives rise to the associated computational method. Here, using a single site update

scheme means that the resulting NWPM algorithm is relatively easy to implement and thus the approach is very accessible.

In exploring the theoretical justifications and concepts of this algorithm, a multiple augmentation space was introduced. This consideration resulted in variants of the algorithm that have significantly reduced computation requirements; as well as the potential for the development of further techniques. Given that they are based on the same approach, these techniques will also inherit many of the advantages of the NWPM algorithm.

Finally, we emphasise that algorithmic variants based on the multiple augmentation space are formally justified (i.e. not approximations, but valid pseudo-marginal methods) but also require significantly fewer computational overhead (similar to the approximation). In other words, multiple augmentation space algorithms attain the advantages of both NWPM and NWSE, while simultaneously avoiding the disadvantages of both (when compared to each other).

We now turn to the empirical evaluation of these methods.

Chapter 6

Simulation Studies

Be a philosopher; but, amidst all your philosophy, be still a man.
— David Hume, *An Enquiry Concerning Human Understanding, 1748.*

In this chapter, the NWPM algorithm together with the SE approximation and a straightforward variant of the multiple augmentation extension are applied in different settings in order to study numerical properties and performance. The first application is a simple toy model, briefly discussed in Example 5.3.3, where the true marginal likelihood is known — This is the focus of Section 6.1. Next, in Section 6.2, simulated PET data from the plasma input linear compartmental models, as discussed in Section 4.2.1, with highly stylised ground truth spatial structures of model orders, will be analysed. In addition, we will look also briefly investigate the effects of using the CESS-adaptive annealing scheme in these algorithms. Finally, in the next chapter the algorithms will be applied to measured PET data.

The software implementation of these methods can be found in the R package `bayespetr`, available at: `https://github.com/dt448/bayespetr`. See Chapter 8, also, for further discussion on this software implementation.

In these simulation studies, for both the toy and PET models, the 20×20 image displayed in Figure 6.1 will be used to generate the ground truth Potts configuration. The overall structure pattern of the image was adapted from studies by Bezener et al. (2018), it includes a range of spatial structures of varying complexity. The image is split into four regions, labelled $R_0, \ldots, R_3$. The image is split into four regions, labelled $R_0, \ldots, R_3$; all pixels within a given region have a common model order; the details vary between experiments and are given below.

Whenever the NWPM algorithms are used in the numerical studies below, $\boldsymbol{M}$ was initialised using the Gibbs a sampler targeting the prior (Potts) distribution. However, when analysing measured PET data in the next chapter, Section 7.2, the algorithms were initialised from the output of the spatially independent SMC sampler method.

Unless stated otherwise, we used the SMC sampler together with the Prior 5 annealing scheme, as in Zhou et al. (2016), see also Section 3.3. As above, we will use the notations N (number of particles/samples) and T (number of intermediate distributions), to refer to the tuning parameters within the SMC sampler. CESS-adaptive annealing schemes were also briefly studied, results are shown in Appendix E.

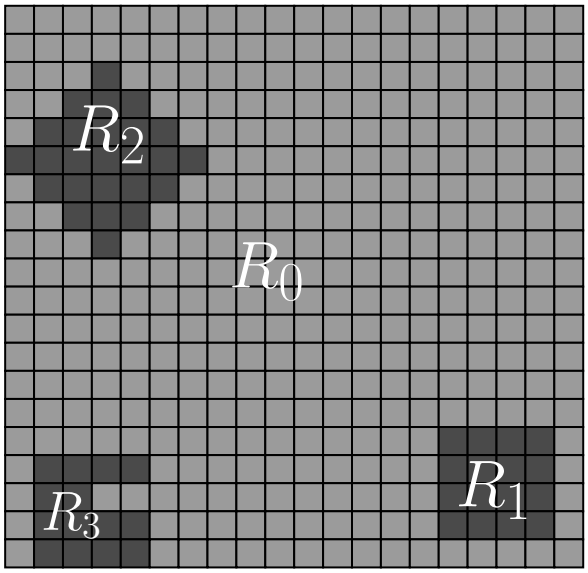


Figure 6.1: Ground Truth Configuration: Image of the spatial structure used to produce ground truth Potts configuration for the studies on simulated data. A 20×20 Potts configuration, with four regions. The model order of these regions will be changed to be appropriate to the setting and models studied.

Additionally, we investigated a simple proposal strategy based on the multiple augmentation space of the NWPM algorithm, as discussed in Section 5.5.2. In brief, the proposal kernel q^1, see Eq.(5.11), together with q^2, from Eq.(5.12), at every positive $\kappa \in \mathbb{N}$ graphical iteration, is used. Essentially, upon initialising the chain, only the change of the model at the node is proposed. After every $\kappa-$th graphical iterations, changes in the auxiliary variable or marginal likelihood estimates for every model order at every node is proposed to change. In this section, we will refer to this variant as the NWMA algorithm, with tuning parameter κ.

Finally, we note the following : Parallel computations are appealing and important in cases where there may be a fixed computational budget. Additionally, they take advantage of the future trajectory of processing technology. An important advantage of SMC sampler is the possibility for parallel computation implementations. Indeed, software packages such as the `vSMC` library (Zhou, 2015) make this feature readily accessible. In the studies below, for simplicity, we investigated sequential (i.e. non-parallel) SMC samplers only. Similarly, it is also possible to exploit the structure of the Potts distribution, where each node's state is only dependant on the state of adjacent nodes, to introduce parallel computation of the marginal likelihood. For instance, multiple marginal likelihoods can be estimated in a "chequerboard" manner — i.e. computing at locations of nodes that are not dependent. Once again, in the interest of simplicity we will not explore such strategies here.

6.1 Simulation Studies: Toy Model

We begin with the simpler toy setting. Importantly, we will use this model to heuristically determine the value of the coupling constant, J, that yields the best performance of the algorithm.

6.1.1 Preliminaries

Toy Model: Consider the simple toy model introduced in Example 5.3.3, in which both the prior and model likelihood are normal, at every node. We used this simple setting to validate the proposed methodology. We restate the model here for clarity.

The toy data set comprises a 2-dimensional digital image, each pixel having a scalar intensity. The graph $G = (V, E)$ used to represent the spatial structure is a finite square lattice, i.e., it has vertex set $V \subset \mathbb{Z}^2$ with nodes $v = (v_1, v_2)^\top$ and edge set $E = \{\langle u, v\rangle : \delta(u, v) = 1\}$, where $\delta(u, v) = \sum_i |u_i - v_i|$ is the L^1 (Manhattan) distance.

In the notation of Section 5.3.2, given an image $\boldsymbol{Y} = (y_v \in \mathbb{R} : v \in V)$, the model studied here is:

$$\begin{aligned} \boldsymbol{M} &\sim \text{Potts}(J, G_{\boldsymbol{Y}}), \\ \mu_v | M_v &\sim \mathcal{N}(\mu_0^{(M_v)}, \sigma_0^2) \text{ for all } v \in V, \\ Y_v | M_v, \mu_v &\sim \mathcal{N}(\mu_v, \sigma^2) \text{ for all } v \in V; \end{aligned}$$

where $\mathcal{N}(\mu, \sigma^2)$ denotes the normal distribution of mean $\mu \in \mathbb{R}$ and variance $\sigma^2 > 0$.

In words: At each pixel, v, we model the data point Y_v as a normal random variable with mean μ_v and variance σ^2, which is assumed known and is common to all pixels. The prior over the latent variable μ_v is normal with mean $\mu_0^{(M_v)}$ determined entirely by the model order M_v and fixed variance σ_0^2. As before, the model order configuration will be the random variable $\boldsymbol{M}$ with a Potts prior distribution.

The node-wise marginal likelihood, momentarily suppressing node index subscript, can be straightforwardly evaluated:

$$\begin{aligned} f(y|M = m) &= \int_{\mathbb{R}} \mathcal{N}(\mu; \mu_0^{(m)}, \sigma_0^2)\mathcal{N}(y; \mu, \sigma^2)\mathrm{d}\mu \\ &= \mathcal{N}(y; \mu_0^{(m)}, \sigma^2 + \sigma_0^2), \end{aligned}$$

for all $m \in \mathcal{M}$. This is trivial to compute since all these terms are known.

The spatial configuration of model order used to simulate the data was specified to represent simple spatial structures; the ground truth value of $\boldsymbol{M}$ was fixed to be the Potts configuration shown in Figure 6.1.

Toy Model Ground Truth: Ground Truth Configuration, as shown in Figure 6.1 was used to generate a ground truth model order image. For this toy model, this is a 20×20 Potts configuration with model order space $\mathcal{M} = \{\mathrm{A}, \mathrm{B}\}$. Region R_0 was fixed to model order $M_v = \mathrm{A}$ and the remaining three regions R_1, R_2 and R_3 where fixed to model order $M_v = \mathrm{B}$. At each node $v \in V$, the mean parameter $\mu_0^{(M_v)}$ will depend on M_v. In other words, the lighter pixels have nodes with hyper-parameter $\mu_0^{(A)} \doteq +5$ and the darker pixels have nodes with hyper-parameter $\mu_0^{(B)} \doteq -5$. Finally, we specified $\sigma_0^2 = 5^2$ and $\sigma^2 = 1^2$.

6.1.2 Toy Pilot Study : Variance Log of Marginal Likelihood Estimates

To begin, we investigate the variance of the SMC marginal likelihood estimator, to provide a range of suitable tuning parameters that could be used for the normalising constant estimator as part

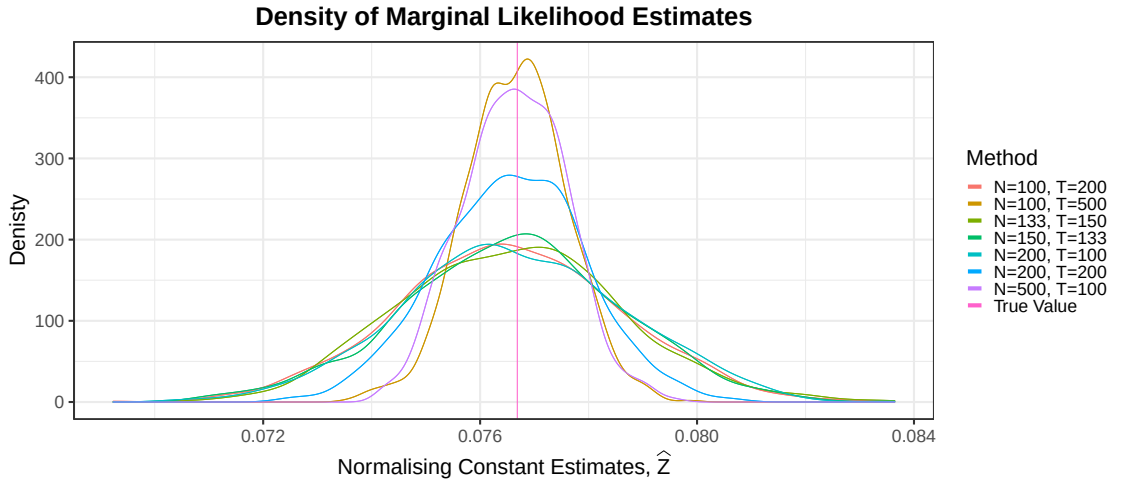


Figure 6.2: Kernel density (Gaussian kernel estimator, using Silverman's "rule of thumb" method (Silverman, 1986), bandwidth = 0.00205) for various parameters of the SMC sampler. The true value of the marginal likelihood $f(y_1|M_1 = 0) = 0.07668543$ shown by the vertical line.

of the NWPM. In particular, the trade-off between the number of particles N and intermediate distributions T is studied here. Particle numbers (or sample size) and distributions numbers both ranging from 100 particles/distributions to 500 particles/distributions were used. This experiment design was based on preliminary studies and numerical studies reported by Zhou et al. (2016).

As such, for simplicity, consider a single pixel from the toy model (i.e. single 1-dimensional data point) with parameters specified to

$$M_1 = 0, \sigma_0 = 5 \text{ and } \sigma = 1.$$

The datum drawn was $y_1 = 1.021248$. Using the above calculations, the true value of the marginal likelihood is $f(y_1|M_1 = 0) = 0.07668543$.

The normalising constant of the posterior given this synthetic data was estimated using the SMC sampler. This was done for 1000 replicates, a kernel density plots of the estimations is shown in Figure 6.2. The mean absolute error and variance log (Var[log $\widehat{Z}$]) of these estimates are shown in the table 6.1.

Tuning Parameter Values	Mean Absolute Error	Variance Log, Var[log $\widehat{Z}$]
$N = 100, T = 500$	2.2×10^{-5}	1.4×10^{-4}
$N = 100, T = 200$	1.5×10^{-4}	7.0×10^{-4}
$N = 133, T = 150$	1.0×10^{-4}	6.9×10^{-4}
$N = 150, T = 133$	9.1×10^{-5}	6.8×10^{-4}
$N = 200, T = 100$	5.6×10^{-5}	7.0×10^{-4}
$N = 200, T = 200$	1.2×10^{-4}	3.1×10^{-4}
$N = 500, T = 100$	4.96×10^{-5}	1.5×10^{-4}

Table 6.1: The MAE of the estimated marginal likelihood and its log variance, for the toy model, for varying tuning parameter values of the SMC sampler.

Firstly, even with relatively low numbers of particles N and distributions T, the variance of the log-

likelihood estimates are small and well below the required threshold for use in a pseudo-marginal setting (Sherlock, 2016; Doucet et al., 2015). Unsurprisingly, higher computational costs yields estimates closer to the true value and with less error. It is evident here that there does not seem to be a significant discrepancy when looking at the trade–off between the two tuning parameters. In some disagreement with the findings of Zhou et al. (2016), the number of distributions seems to have a slight edge over the number of particles particularly when looking at the mean absolute error. However, this is settings is much simpler than analysing PET images and thus may not require as many intermediate distributions.

Having established the range of tuning parameters that is suitable for use in a pseudo-marginal algorithm, we now turn to looking at simulated 2-D images.

6.1.3 Toy Simulation Study 1: Altering the Coupling Constant J

As discussed in Section 5.2.1, typically it is difficult to infer the coupling constant J of the Potts distribution (Møller et al., 2006; Moores et al., 2020). In this work, we make the assumption that J is known when using the NWPM algorithm — this is not too unreasonable as J is a parameter of the prior distribution. In this vein, here we evaluate the performance of the algorithm for various values of J.

The following design of experiment was used: An image data set of $20 \times 20 = 400$ pixels(nodes) from toy model and model order ground truth, as dictated by the Toy Model Ground Truth above, was generated. This simulated image was then analysed using the NWPM algorithm, for varying values of the coupling constant J. Coupling constants ranging from $J = 0$ to $J = 5$ were investigated. Based on pilot studies, an SMC sampler with $N = 200$ particles and $T = 500$ distributions was used within the NWPM algorithms. The pseudo-marginal MH Markov chain was ran for $n = 100$ iterations. As mentioned above, the pseudo-marginal MH Markov chain was initialised using a Gibbs sampler, with the respective coupling constant. This was done for 50 replicates of the NWPM Markov chain.

Model selection was carried out by selecting the modal model order marginally at each node $v \in V$. In other words, the modal state in which each $M_v^{(i)}$, for $i = 1, \ldots, 100$, was selected for each pixel $v \in V$. The empirical averages, over the 50 replicates was calculated. The mean percentages of the nodes for which the correct model order was selected, for the different variations of the NWPM algorithm, is shown in Figure 6.3.

It is evident that, in this simulation study, the performance of the algorithm peaks for coupling constant $J = 0.4$. This is smaller than the critical threshold J_{critical}, as discussed in Section 5.2.1. This is noteworthy, as there is typically a sharp decrease in speed of mixing of the Markov chain above this critical value. In fact, the graph shows that the variance of the percentage of correctly selected model orders increases with the coupling constant; Importantly, there is a jump in the rate of increase between $J = 1$ and $J = 1.5$.

An important question to consider here is how much of the performance difference is due to poor approximation of the posterior by the Markov chain, and how much to the posterior itself. It is reasonable to argue that below the critical value, reasonably good mixing is obtained and this probably reflects a good approximation of the posterior as indicated by the low variability. In contrast, above this critical value the results are largely dominated by the poor mixing of the Markov chain over this rather short period. We also note that the graph shows good performance for even this relatively short chain, for the appropriate coupling constant values. We use short

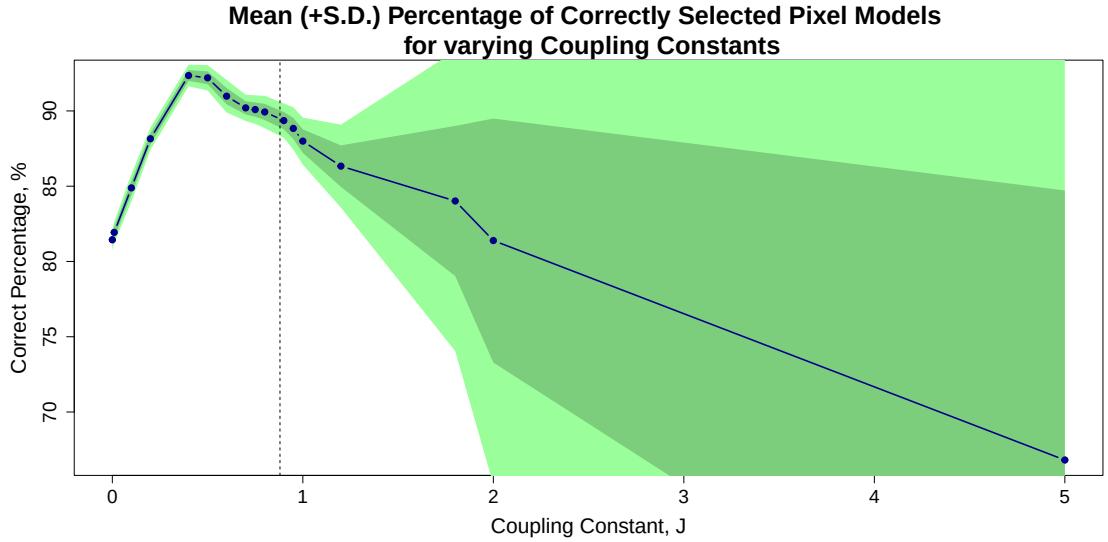


Figure 6.3: Empirical average, and standard deviations (s.d.), of the percentages(%) of the total (400) pixels where the correct model order was selected, for different values of the coupling constant J. Mode selection was done on simulated toy model data using NWPM with an SMC estimator (with $N = 200$, and $T = 500$) for the marginal likelihood. The configuration selected for estimation by looking at the percentage of time each node spent in each model order state. The line plot shows the empirical average of the correct percentage of pixels. The lighter and darker region show the $2\times$ s.d. and $1\times$ s.d. error bands, respectively. J_{critical} is shown as the dashed vertical line.

number of graphical iterations here, to evaluate the performance; This is due to the fact that in resource intensive settings, such as PET images below, it is not plausible to run the full NWPM algorithm for extended lengths of the chain.

Based on evaluating the performance for this range of coupling constant and due to the similar qualitative features of the different settings; for all the following studies we will henceforth fix $J = 0.4$.

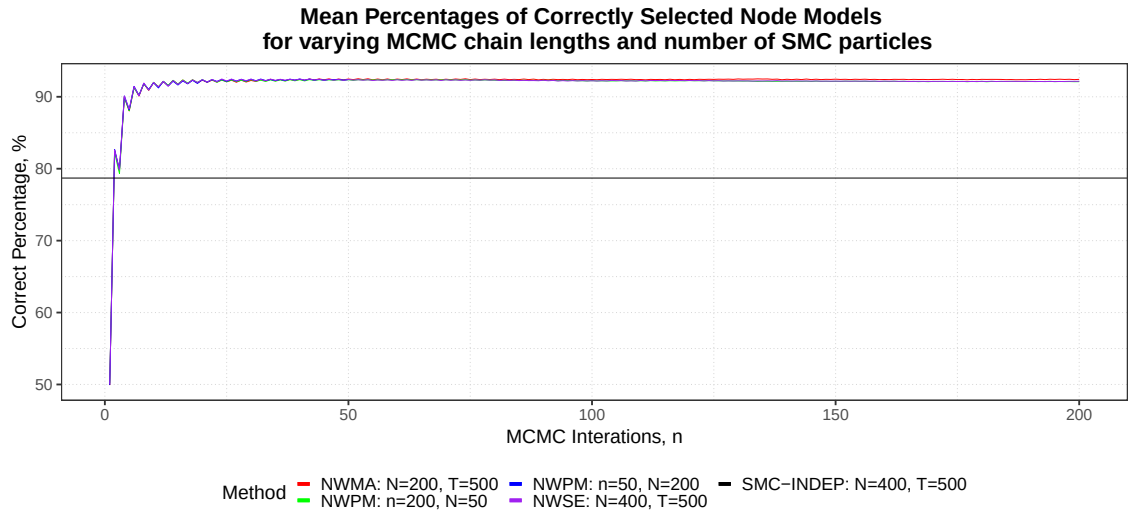
6.1.4 Toy Simulation Study 2: Estimator Sample Size vs Markov Chain Length Trade-off

Next, the computational trade–off between the number of particles N used in the SMC sampler and the length of the pseudo-marginal MH Markov chain , or graphical iterations, n was investigated. The same design as the study in Section 6.1.3 was used, with the exception of the Markov Chain length varying from $n = 50$ to $n = 200$, and the number of particles varying between $N = 50$ and $N = 200$.

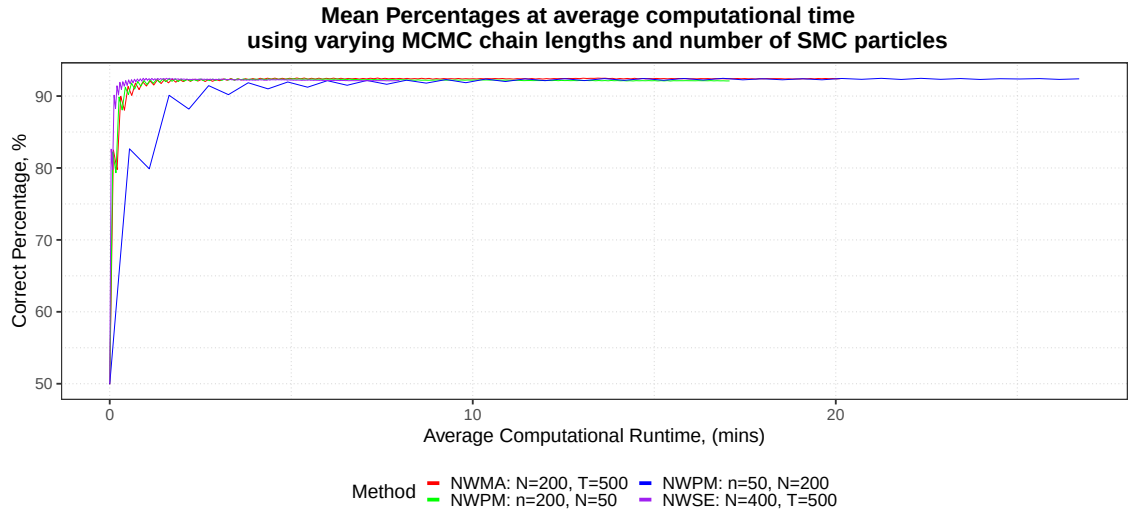
Additionally, spatially *independent* model selection, using an SMC sampler to estimate the model evidence was also investigated. The NWSE approximation was also used to analyse the simulated 2-D image. For these two methods, marginal likelihoods were estimated using the SMC sampler with $N = 400, T = 500$. Similarly, the NWMA algorithm, as described above, with updates to the marginal likelihood estimates every $\kappa = 10$ graphical iterations was also used. Here, marginal likelihoods were estimated using the SMC sampler with $N = 200, T = 500$; For the two, less resource–intensive, Markov chain methods, $n = 200$ graphical iterations were used.

Model selection was carried out in the same manner as described above, and the percentage of correctly selected nodes was used as the metric of performance. This was done for 100 replicates,

for each method used. Results are shown in Figure 6.4.



(a) Mean percentages at each graphical (MCMC) iteration.



(b) Mean computational runtime (averaged over each graphical iteration).

Figure 6.4: Part (a) shows the average percentages(%) of the whole toy model image ($20 \times 20 = 400$ pixels) where the correct model order was selected at each iteration of the MCMC chain; Using NWPM (and NWSE approximation) for varying Markov Chain length, n, and number of particles, N, in the SMC sampler. The configuration was selected using the marginal modal state at each node, cumulatively over $i = 1, \ldots, n$ iterations. The horizontal line shows the average percentage when using spatially independent SMC model selection. Average computational runtimes are shown is shown part (b).

These results suggest that, at least in this simple setting, the algorithm is not particularly sensitive to the allocation of computational resources investigated here. Importantly, we see that the selected graph model/configuration stabilises fairly quickly for all the combination of tuning parameters. In particular, by looking at the longer $n = 200$ chain we can see that the convergence is to around 92% of nodes having the correct model order selected, when using the NWPM algorithm. This is considerably higher than the performance of the spatially independent SMC sampler, which was 79%; confirming that exploiting spatial structure can significantly improve estimation in this type of simple model. We also observe that the NWSE approximation method and the simple NWMA algorithm achieved almost identical results to the full NWPM algorithm: suggesting that, in this

simple setting, there is enough signal to allow for the use of the approximation to achieve very similar results, with very little additional computational cost. A version of Figure 6.4a, where only graphical iterations $n = 10$ and on wards are shown is presented in Figure D.1, Appendix D — This graph allows for better discernment of the difference in performance across the different methods. Intermediate chain lengths not shown in the above graphs, i.e. $n = 75$ and $N = 134$, also produced similar results.

Figure 6.4b shows the average computational runtime for each method. The runtimes were computed by looking at the average CPU time per iteration - that is the total run time for each trail was recorded and then averaged over the total graphical iteration. We can draw similar conclusions as above, with the exception that the NWPM method with $N = 200$ particles seemed to take a longer time (in minutes) to reach stability.

Given the significantly reduced computational costs, long term behaviour for the NWSE and NWMA MCMC chains were also briefly verified, results are shown in Figure D.3, Appendix D.

6.2 Simulation Studies: Compartmental Models for PET Data

Before we investigate the performance of the proposed algorithms on measured PET data, we verify and investigate the tuning parameters when applied to simulated PET data. We first evaluate the SMC sampler at a single pixel, in various designs and settings. Then, we evaluate the NWPM and its variants and approximations on simulated 2-D dynamical PET images. Finally, we briefly look at the effects of using CESS-based adaptive SMC schemes.

6.2.1 Preliminaries

Consider a time series from a single voxel site of a dynamical 3-D PET image, denoted $y = (y_1, \ldots, y_k)^\top$. Recall Eq.(4.6), from Section 4.2.4, that given model order $m \in \{1, 2, 3\}$, the m-compartmental model for this data point can be written:

$$y_i = C_T(t_j; \phi_{1:m}, \vartheta_{1:m}) + \sqrt{\frac{C_T(t_j; \phi_{1:m}, \vartheta_{1:m}}{t_j - t_{j-1}}} \epsilon_j,$$
$$C_T(t_j; \phi_{1:m}, \vartheta_{1:m}) = \sum_{i=1}^{m} \phi_i \int_0^{t_j} C_{\mathrm{P}}(s) e^{-\vartheta(t_j - s)} \mathrm{d}s;$$

for all $j = 1, \ldots, k$. In particular, the latent parameters $\phi_{1:m} = (\phi_1, \ldots, \phi_m)$ and $\vartheta_{1:m} = (\vartheta_1, \ldots, \vartheta_m)$ determined the dynamics of this model.

For the PET simulation study in this section, we simulate data from this collection of $m-$compartment models. Specifically, we will use the normal distribution as the error structures (see Section 4.2.4): $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ for all $i \in \{1, \ldots, k\}$; Where, as before, $\mathcal{N}(0, \sigma^2)$ is the normal distribution with mean zero and variance σ^2.

For simplicity, we restrict the design of these simulation experiments to using the same weight and exponential pairs, respective to the model order, at every voxel site. These parameter values are selected from Jones et al. (1994) (and subsequently used for simulation studies by Gunn et al. (2002); Peng et al. (2008); Jiang et al. (2009) and Zhou et al. (2013)). More specifically, these

micro–parameter values are specified as follows: For the $m = 1$ and $m = 2$ compartmental models we use $\phi_1 = 4.9 \times 10^{-3}s^{-1}$, $\phi_2 = 1.8 \times 10^{-3}$, $\vartheta_1 = 5 \times 10^{-4}s^{-1}$ and $\vartheta_2 = 0.011s^{-1}$; For the $m = 3$–compartmental model we use $\phi_{1:3} = (4.4 \times 10^{-3}s^{-1}, 1 \times 10^{-4}s^{-1}, 1.4 \times 10^{-3}s^{-1})$ and $\vartheta_{1:3} = (4.5\times10^{-4}s^{-1}, 2.7\times10^{-3}s^{-1}, 1\times10^{-2}s^{-1})$. In each of these cases the volume of distributions is roughly $V_{\mathrm{D}} = 10$. An exception is the small full factorial study in Section 6.2.3, the specific micro–parameter values used will be detailed further within that section.

The simulated synthetic data was obtained using the plasma input function and integration periods used in the measured PET data set studied in Chapter 7. Similarly, normally–distributed noise, with zero mean, was added to the synthetic data. The variance was such that it was proportional to the true time activities divided by the length of the frame duration; and scaled such that the highest variance in the sequence is equal to the noise level (this is discussed further in Section 4.2.1). A noise level of 0.5, which lies within the representative range (between 0.01 and 1.28) of real data (Jiang et al., 2009), was used when simulating the 2-D PET image.

We used non-informative uniform priors over ϕ and ϑ (see Zhou et al. (2013), who concluded that their proposed modelling was not sensitive to prior specification, particularly comparing non-informative to biologically informed priors.). The ϕ micro–parameter was constrained to lie within the range $[10^{-5}, 10^{-1}]$, and the ϑ micros–parameters within $[10^{-4}, 10^{-1}]$. These ranges are based on pilot and previous studies, and also ensure that the parameters are physiologically meaningful (Cunningham and Jones, 1993).

For the normally distributed error model, the precision parameter is denoted $\lambda = 1/\sigma^2$. A gamma distribution, with both parameters equal to 10^{-3}, was used as the prior for λ – a proper approximation to the improper Jeffry's prior.

Expressions of all the relevant distributions, in analytical form, can be found in Appendix C.

PET Simulation Ground Truth: As before, the ground truth of the model order configuration were based on spatial structure pattern of Ground Truth Configuration, shown in Figure 6.1. In this PET simulation study, a 20×20 Potts configuration with model order space $\mathcal{M} = \{1, 2, 3\}$, representing the number of compartments was generated. Region R_0 was specified to model order $M_v = 2$; region R_2 and R_3 to be model order $M_v = 1$ and region R_1 to be model order $M_v = 3$. In other words, the lighter pixels have time series generated from the $m = 2$ compartment model; regions R_2 and R_3 of the darker pixels generated from $m = 1$ compartment model and R_1 from the $m = 3$ compartment model.

To begin, pilot studies using the SMC sampler for data at a single voxel were carried out: To identify the appropriate tuning parameter values to be used for the simulation studies as well as the application on measured PET data below.

6.2.2 PET Pilot Study 1 : Tuning Parameters for SMC Likelihood Estimator

In evaluating the SMC normalising constant estimator, Zhou et al. (2016) investigated the trade–off between the number of particles N and the number of intermediate distributions T (within the annealing scheme) when analysing PET data at a single voxel site. The objective of this pilot study is similar to this work; However, here we will look at simulated PET data from a larger range of noise level. Additionally, in contrast to Zhou et al. (2016), here we will investigate performance with regards to correct model selection, as well as the variance of the normalising

constant estimator. Similarly, in addition to the two tuning parameters N and T, here we will also investigate the effect of using a range of local MCMC steps.

Variance of the Log of the Marginal Likelihood Estimates

We first investigated the trade–off between the number of particles and the number of local MCMC moves at each iteration t within the SMC sampler. For the number of particles : $N = 200$, $N = 400$, $N = 800$, $N = 1000$ and $N = 2000$ were investigated. For each of these specified numbers of particles, SMC samplers with: one, two, four and eight local MCMC moves were used. As briefly discussed in Section 2.6.2 and 3.2, we used adaptive MCMC. The number of distributions was fixed to $T = 500$, for all of the above combinations.

To study the variance of the log of the likelihood estimator, a single time series from a $m = 2-$compartment model for each noise levels $0.01, 0.1, 0.2, 0.5, 1.28$ and 5.12 were drawn. SMC samplers, with all the above combinations of tuning parameters, were used to estimate the marginal likelihood. As mentioned above, we used the Prior 5 annealing scheme. This was done for 200 replicates; For each combination of tuning parameter and noise levels. The results are shown in Figure 6.5.

Next, the above was repeated for the number of intermediate distributions T. That is, the number of particles was fixed to $N = 200$. Instead, for the number of intermediate distributions $T = 500, T = 1000, T = 2000, T = 2500$ and $T = 3000$ were used. The range of local MCMC moves was the same as above. The results are shown in Figure 6.6

In both of these studies, the results show that, generally speaking, a higher noise level results in a smaller empirical variance of the normalising constant estimator. This is expected as, roughly speaking, the likelihood density for a higher noise level is flatter, and so the weights will have a smaller variance. Similarly, the biggest improvement in performance, when looking at the effects using multiple MCMC moves, is seen between when using one move and two moves. This increase in performance decreases notably when using more than two moves, suggesting that there is diminishing returns when resource is allocated to multiple MCMC kernel applications. Similar results are seen when looking at the effects of increasing N or T. That is, the biggest improvements are seen early on i.e. when comparing $N = 200$ to $N = 400$ particles and similarly $T = 500$ to $T = 1000$ distributions. Beyond these two values, we see a decrease in this improvement for higher values of these tuning parameters. This is especially pronounced when using SMC samplers with a single local MCMC move at each iteration.

From the graphs, is difficult to make conclusive remarks about the trade-off between these tuning parameters. It is clear that the biggest effect is seen when using multiple MCMC moves rather than just a single one. However, increasing this beyond even two MCMC moves shows a rapid decrease in the amount of improvement. Pragmatically, when there is a need to be economical with computational resources, as is the setting of interest here: It would be ideal to use a small number of MCMC moves rather than allocating resources to higher numbers of particles and/or distributions (i.e. more than $N = 800$ particles or $T = 1000$).

An important point to note that, for the noise levels that we are particularly interested in (that is, the noise level 0.5), the variances of the log-likelihood estimates, for all the combinations of tuning parameters investigated, are smaller than one. As such, SMC samplers with these tuning parameters can be readily used within the NWPM algorithm (Doucet et al., 2015; Sherlock, 2016) when investigating simulated 2-D images below.

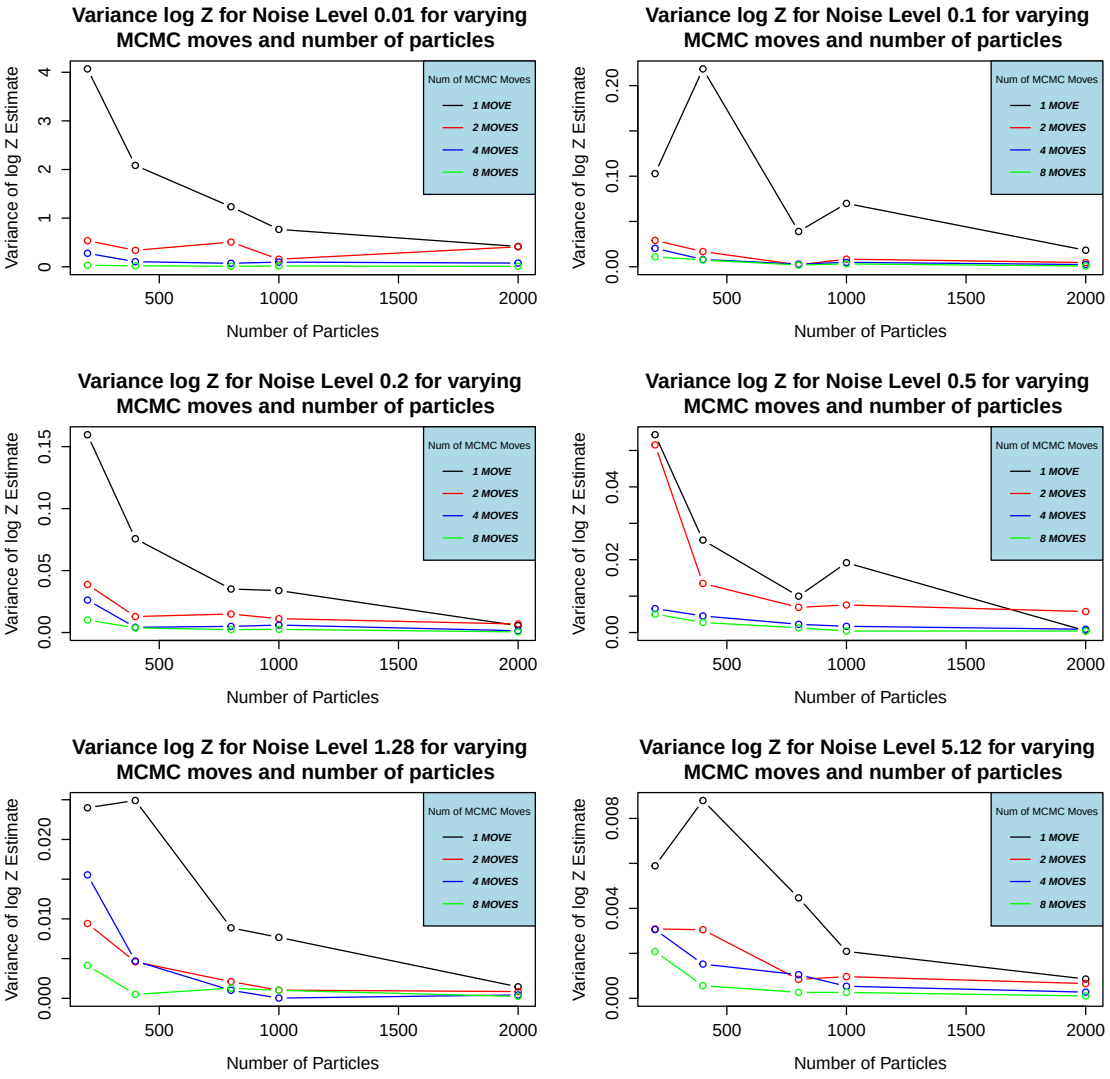


Figure 6.5: Variance log of SMC normalising constant estimator for different number of particles and MCMC moves, for different noise levels.

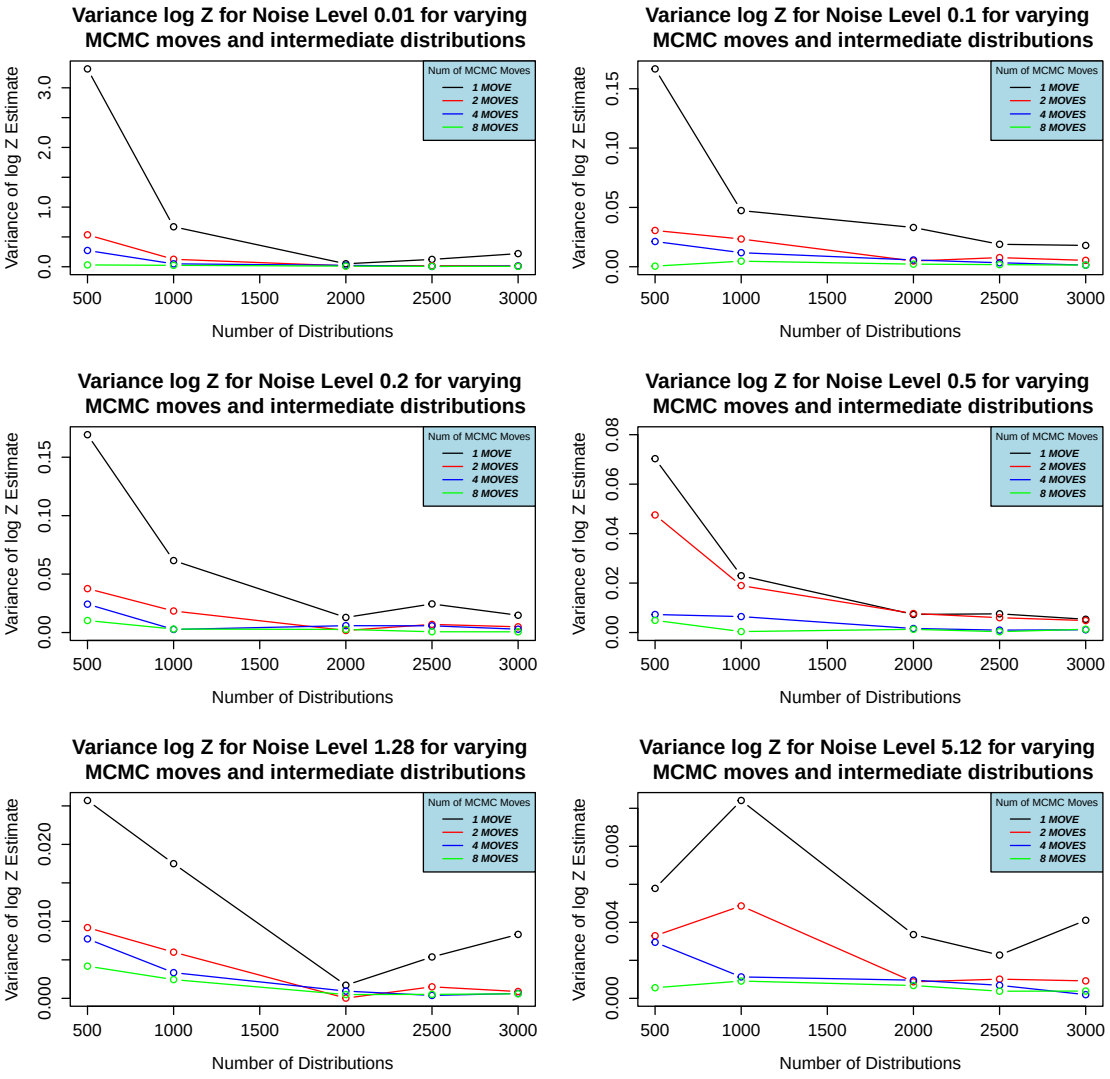


Figure 6.6: Variance log of SMC normalising constant estimator for different number of intermediate distributions and MCMC moves.

Model Selection

Next, the above two experiments were repeated, where instead the model selection was evaluated. Furthermore, 200-replicates from a $m = 2$–compartment model with noise level 0.5, were drawn. We focus here on the noise level 0.5, as we will use this for the studies on the stylised simulated 2–D image, in the sequel. For each of these synthetic time series, SMC samplers from each combination of tuning parameters above were used to estimate the marginal likelihood for given model orders $m = 1, 2$ and 3. Bayesian model selection was performed based on these estimates. The results are shown in Figure 6.7.

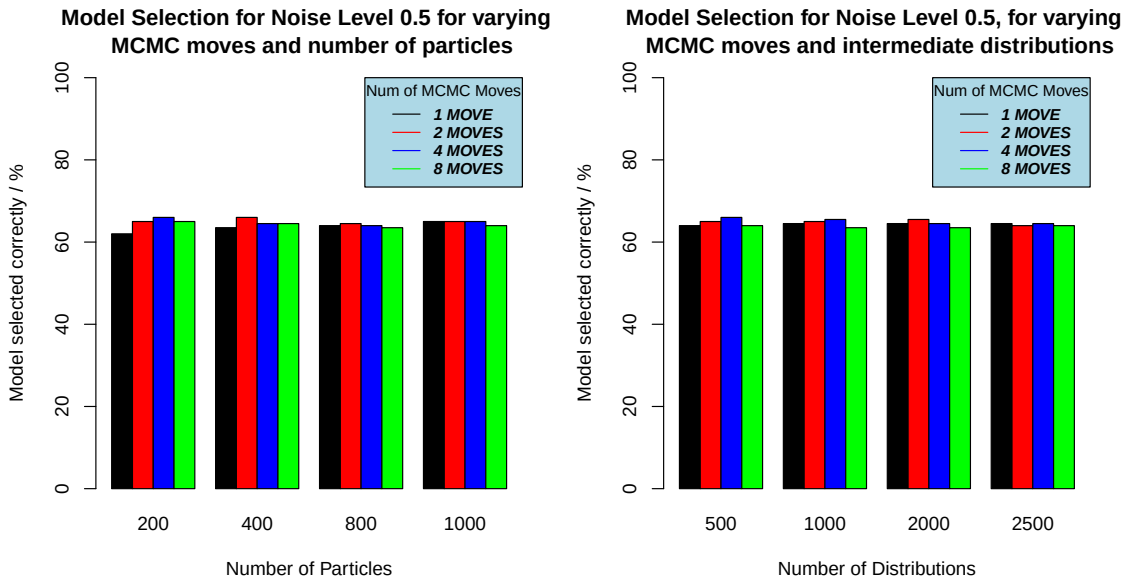


Figure 6.7: Model selection for simulated PET data (with noise level 0.5) for varying number of MCMC moves and number of particles, using SMC sampler.

The above results show the diminishing returns effect of increases use of computational resources, for model selection. We see that, although there is some differences, they are insignificant and do not show any improvement in performance with increased values of tuning parameters. In particular, the highest performance is , for tuning parameter $N = 200$ and $T = 500$ and three local MCMC moves.

Henceforth, based on these empirical results, we will use two local MCMC moves within the SMC sampler at each step $t = 1, \ldots, T$.

6.2.3 Pilot Study 2: A Full Factorial Experiment

Numerical studies that use large designs of experiments in PET simulated studies are sparse, due to the computationally intensive requirements of analysing this large data set. In particular, factorial designs for compartmental models would require looking at a large number of parameter combinations. As such most, empirical studies focus on designs containing a range of noise levels rather than investigating micro–parameter values. However, it is feasible to study a small full factorial design for the rate constants if looking at the $m = 1$–compartment model.

In view of this, to investigate the effect of data generated from models with different values of the rate constants on the performance of the SMC algorithm. A $2 \times 2 \times 3$ factorial design experiment will be carried out using one-dimensional simulated data from a $m = 1$–compartment model.

Recall that all rate constants must lie within $[5 \times 10^{-4}, 10^{-2}]$ (Jones et al., 1994) — we conduct a simulation study using the 5–percentile and 95–percentile of this range for levels of each rate constant. Define:

$$K_1^{\text{Up}} = k_2^{\text{Up}} \doteq 9.525 \times 10^{-3},$$

and

$$K_1^{\text{Low}} = k_2^{\text{Low}} \doteq 9.75 \times 10^{-4}.$$

The simulation experiment is carried out using the above values for the two levels of rate constants; The range of noise levels 0.01, 0.5 and 1.28, were used.

Similar to the pilot study above, we evaluate the performance by looking at the variance of the normalising constant estimates and model selection. For the variance investigation, for each noise level and rate constant parameter combination a synthetic time series was drawn. An SMC sampler with $N = 200$ and $T = 500$ was then used to estimate the marginal likelihood for given model order $m = 1$. As before a Prior 5 annealing scheme was used. This was done for 200 replicates of the SMC sampler. The variance of these estimates are shown in Table 6.2.

Similarly, to evaluate model selection, 200 times series were drawn for each combination of the factorial design. An SMC sampler with $N = 200$ and $T = 500$ was then used to estimate the model evidence for model orders $m = 1, 2$ and 3. The percentage of correctly selected model, when using Bayesian model selection, is shown in Table 6.3

Variance of log marginal likelihood, $\text{Var}(\ln \widehat{Z})$

	Noise level		
Design Matrix	0.01	0.5	1.28
$(K_1^{\text{Up}}, k_2^{\text{Up}})$	0.00362	0.00394	0.00331
$(K_1^{\text{Up}}, k_2^{\text{Low}})$	0.02865	0.01092	0.00777
$(K_1^{\text{Low}}, k_2^{\text{Up}})$	0.00603	0.00748	0.00583
$(K_1^{\text{Low}}, k_2^{\text{Low}})$	0.03210	0.00551	0.00366

Table 6.2: Variance of log-likelihood estimates for full factorial simulated 1-compartment PET data, for noise levels $0.01, 0.5$ and 1.28.

Percentage of models correctly selected, %

	Noise level		
Design Matrix	0.01	0.5	1.28
$(K_1^{\text{Up}}, k_2^{\text{Up}})$	0	0	0
$(K_1^{\text{Up}}, k_2^{\text{Low}})$	100	99.5	100
$(K_1^{\text{Low}}, k_2^{\text{Up}})$	100	100	100
$(K_1^{\text{Low}}, k_2^{\text{Low}})$	100	100	100

Table 6.3: Percentage of correctly selected models, for full factorial simulated 1-compartment PET data, for noise levels $0.01, 0.5$ and 1.28.

Looking at the numerical results above, we see evidence of the robustness of the SMC sampler. In particular, the variance of the normalising constant estimates are relatively low, for all the noise levels studied and for each combination of the design. Specifically, these estimates suggest that the variance is low enough to be used within a pseudo-marginal algorithm, albeit for the marginal likelihood a 1–compartment model. Typically, simpler (low dimensional) models will

have a smaller estimator variance. However, combining these results with the other pilot studies provide good evidence and range of tuning parameters for using the SMC sampler within the proposed framework.

In regards to the model selection performance, we see that the SMC sampler is accurate in almost all cases studied. The exception is for the $(K_1^{\text{Up}}, k_2^{\text{Up}})$ combination, where a higher model order (> 1) was selected for all data points. Given that this is a combination of upper-bounds values, it is not unreasonable to assume that these extreme rate constant values will only occur in realistic settings in very rare occasions.

Having briefly explored the SMC algorithm as well as simulated data from the model; we may now turn to the main objective of evaluating the NWPM algorithm on data with spatial structures.

6.2.4 PET Simulation Study 1: Algorithm Comparison

Presented in this sub-section is analysis of a simulated 2-D PET image using the NWPM algorithm(s). Recall that the 2-D image uses the "PET Simulation Ground Truth" described above as the spatial structure patterns to simulate the time-series at each pixel.

From the above pilot studies, we saw that: For the variance of the SMC normalising constant estimates, in agreement with numerical studies by Zhou et al. (2016), the numerical study showed that, there was a decrease linearly for higher values of N and similarly in T. However, beyond a certain threshold, the reduction were no longer useful. Similar results were seen for model selection performance, for 1-D data simulated from just the $m = 2$–compartment model with noise level 0.5, performance could not be increased beyond 65% regardless of any increase in tuning parameter values. Here, we are interested in investigating whether the computational resource could be better allocated towards incorporating spatial structures using the NWPM algorithm and the result of doing so.

Based on the above pilot studies, an SMC sampler with the Prior 5 annealing scheme, $T = 400$ intermediate distributions, was used as the estimator of the normalising constants at all pixels for all models. These values, and the values of the number of particles N in the design below, were chosen as they allow for the variance of the likelihood estimates to be within a suitable range to allow for optimal scaling (Doucet et al., 2015; Sherlock, 2016). Chain lengths of $n = 50$, $n = 75$, $n = 100$ and $n = 200$ were investigated, with particles $N = 200$, $N = 134$, $N = 100$ and $N = 50$, respectively.

These, relatively smaller, graphical iteration lengths where chosen due to the computational overhead of this algorithm in this resource intensive setting. Note that, since model selection here uses the marginal modal state from a space of just 3 model orders — these shorter chains can still produce reliable results.

Additionally, the image was also analysed using spatially independent SMC sampler and the NWSE approximation; using an SMC sampler with $N = 400$ particles and $T = 600$. The NWMA algorithm, with SMC sampler using $N = 200$ and $T = 600$, at every $\kappa = 40$ iterations, was also evaluated. Given their significantly reduced computational costs, for the two pseudo-marginal methods, $n = 500$ iterations were used.

Following Algorithm 4, all the chains where initialised using the Gibbs sampler to target the prior distribution. The simulated image was analysed by these variations of the NWPM algorithm for

30 replicates. Additionally, the image was also analysed using spatially independent SMC sampler and the NWSE approximation, using $N = 400$ particles and $T = 600$ distributions. The results are shown in Figure 6.8 and Table 6.4 (performance for the intermediate ranges of the tuning parameters of the NWPM algorithm are shown in Figure 6.9).

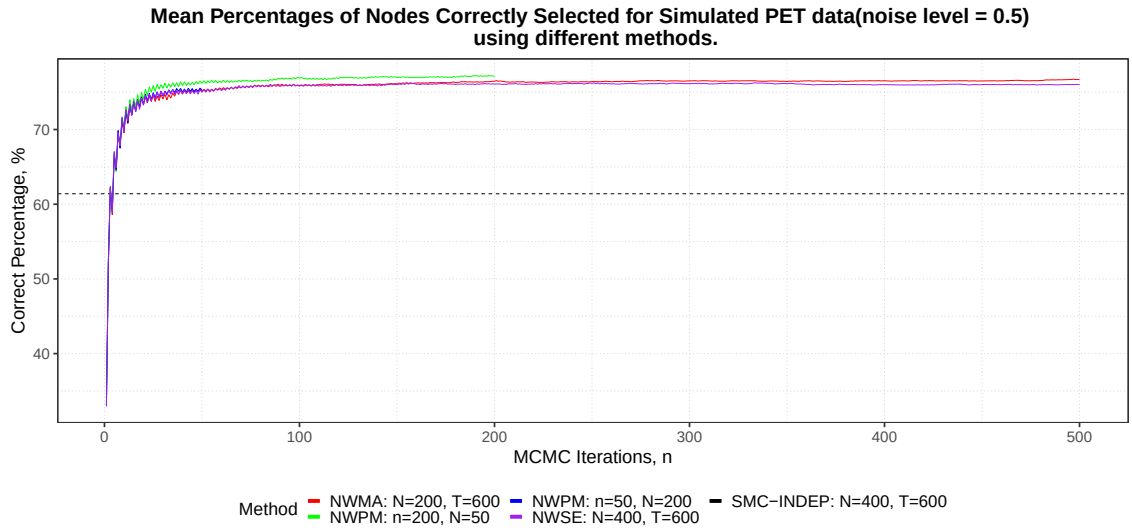
Looking at Figure 6.8, the results show notable improvements in model selection performance when using the different variants of the NWPM algorithms compared to spatially independent model selection using the SMC sampler. Analysing the simulated 2-D PET image using the spatially independent method gave 61.4% pixels correctly selected, compared to the NWPM algorithms all giving around 76% even for as short as $n = 50$ graphical iterations. On average, this is a 23.4% increase in the number of pixels selected correctly when using the proposed NWPM approaches compared to the spatially independent SMC sampler method. On large data sets, like PET images, this is a considerable number of pixels. Furthermore since this is spatially dependent model selection the improvement in correctly selected pixels, when compared to SMC independent, will be towards revealing the underlying spatial structures. These results suggest that incorporating spatial dependence does results in better model selection performance.

Specifically, Figure 6.8a shows that, for the chains where higher iterations were used, we can see that this increase in results was maintained. As before, a version of this figure with shorter graphical iteration ($n = 10$, on wards) can be found in Figure D.2, Appendix D. Further, to verify long term behaviour of the algorithms: long-run MCMC chains, of the less expensive NWSE and NWMA algorithms, are shown in Figure D.4, Appendix D. Note these larger iteration numbers are not feasible nor required in realistic applied settings.

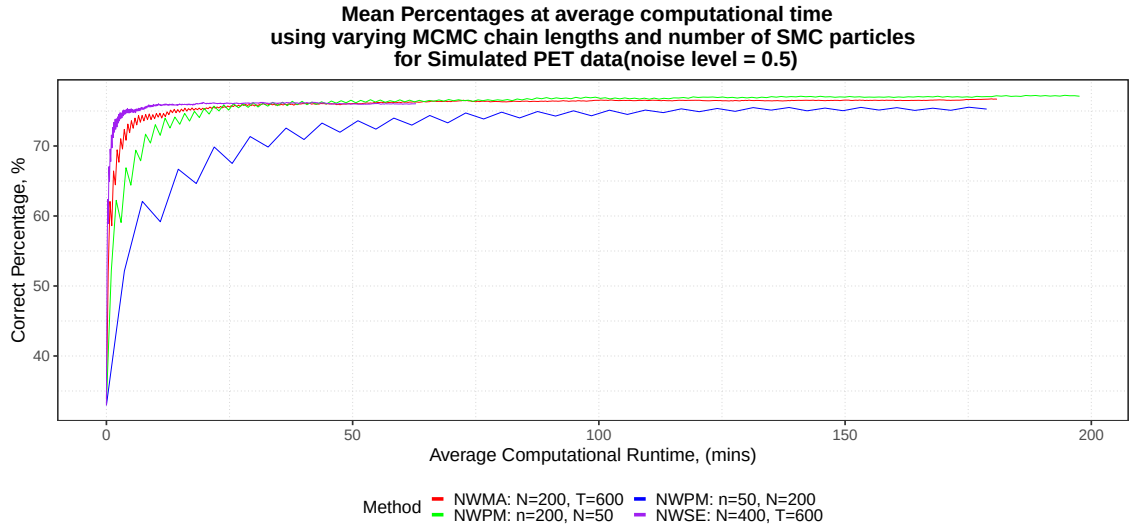
Figure 6.8 also shows that similar performance can be achieved using the NWSE approximation and the NWMA variant of the proposed NWPM algorithm. Pragmatically, this is very promising as it suggests that even with an approximation current state of the art methods can be outperformed with very little additional costs. Alternatively, rather than using an approximation, algorithms based on the multiple augmented state space with significantly reduced computational overhead can be used. For example, the simple NWMA algorithm studied here. Similar to the toy model, there does not seem to be any notable difference when using higher chain length and lower particles. Additionally, there is only a small increase in performance when using more a computationally intensive algorithm, on average. That is, the full NWPM performs slightly better than the NWMA variant; which in turn also has a minute increase in performance compared to the NWSE approximation.

As before, Figure 6.4b shows the average computational runtime for each method. The runtimes were computed in the same manner as described in Section 6.1.4. Similarly, as also seen in this aforementioned section, the NWPM method with $n = 200$ seems to be the slowest in regards to mixing (in minutes).

Table 6.4 shows the average root mean squared error (RMSE) of the V_{D} estimates for each variant of computational method. The quantity was calculated by taking the average RMSE over all the pixels in the 2-D PET image, and then over all the replicates (the empirical variance of these quantities are shown in the parenthesis). We can see that the non-spatial SMC sampler method has the highest empirical RMSE. The NWPM algorithms, including all chain lengths and the SE variant, produce a noticeably smaller RMSE compared to the spatially independent SMC method. There is a small improvement when using the full pseudo-marginal NWPM, compared to its approximation, in this case but it is within the range that could be plausibly explained by random fluctuations.



(a) Mean percentages at each graphical (MCMC) iteration.



(b) Mean computational runtime (averaged over each graphical iteration).

Figure 6.8: Part (a) shows the average percentages(%) of the whole image ($20 \times 20 = 400$ pixels) where the correct model order was selected at each iteration of the pseudo-marginal MH chain, using: NWPM for varying Markov Chain length, n, and number of particles, N, in the SMC sampler; NWSE with $N = 400$ particles and $T = 600$ distributions and NWMA using SMC sampler with $N = 200$ and $T = 600$ at every $\kappa = 40$. Corresponding average computational runtimes is shown in part (b). The dashed line shows the average percentage when using spatially independent SMC($N = 400, T = 600$) model selection.

Method	Model Averaging V_{D}	Posterior V_{D}
NWPM n =50	0.1319 (0.0061)	0.1321 (0.0061)
NWPM n =75	0.1321 (0.0061)	0.1323 (0.0061)
NWPM n =100	0.1320 (0.0061)	0.1323 (0.0061)
NWPM n =200	0.1322 (0.0061)	0.1324 (0.0061)
NWMA	0.1319 (0.0061)	0.1323 (0.0061)
NWSE	0.1322 (0.0061)	0.1324 (0.0061)
INDEP-SMC	0.1341 (0.0066)	0.1341 (0.0063)

Table 6.4: Average RMSE (and s.e.) for simulated PET data, analysed using different algorithms and tuning parameters.

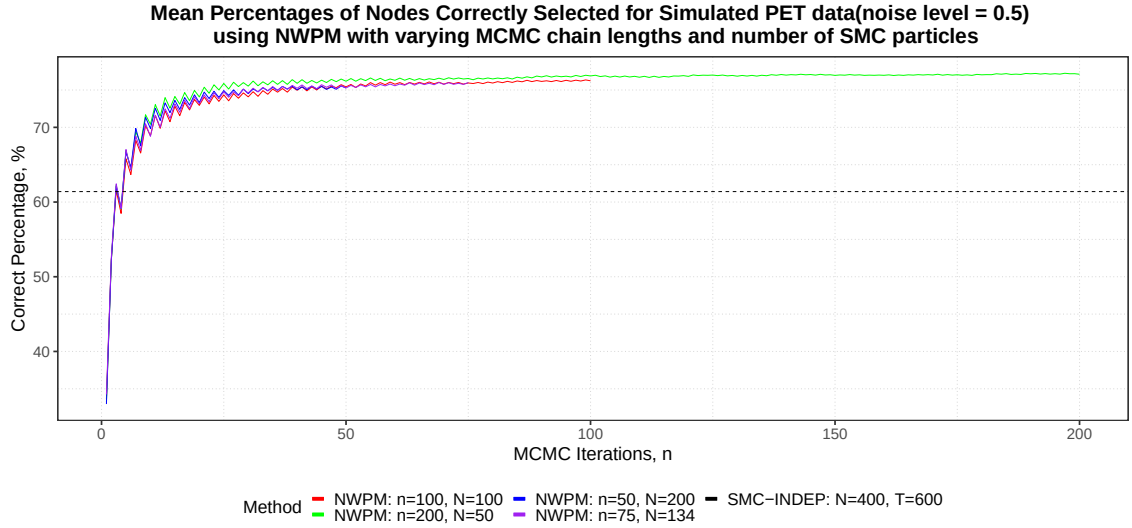


Figure 6.9: Average percentages(%) of the whole image ($20 \times 20 = 400$ pixels) where the correct model order was selected at each iteration of the pseudo-marginal MH chain, using the NWPM algorithm for varying Markov Chain length, n, and number of particles, N, in the SMC sampler. The dashed line shows the average percentage when using spatially independent SMC($N = 400, T = 600$) model selection.

An image showing the average RMSE at each pixel, over the 30 replicates, for the NWPM algorithm with $n = 200$, is shown in figure 6.10 (the other algorithms showed almost identical outputs). This parametric image shows that, in general the RMSE is relatively small and uniform over all most all the pixels. However, interestingly the highest values of the RMSE is found at nodes which have neighbours that are of different states to them. This can be seen by comparing this figure to the PET Ground Truth configuration (or overlaying them). One possible explanation is that that the wrong model is selected (i.e. $m = 2$ or $m = 1$ rather than the true model order $m = 3$).

CESS-adaptive Schemes Study

Lastly, the effects of using the CESS to determine the annealing scheme, see Section 3.3.1, was also investigated. In brief, the above pilot study (as in Section 6.2.2), using a single pixel time series, was also carried out using CESS-adaptive annealing scheme with various CESS thresholds and number of particles. Firstly, the effect of using different thresholds CESS* was investigated. The design used for the CESS* were chosen such that the number of intermediate distributions were roughly (corresponding CESS* in parenthesis): $T = 500(0.9997), 1000(0.99992), 2000(0.99997), 2500(0.999985)$ and $3000(0.99999)$. For this experiment, $N = 200$ particles where used.

Similarly, the effects of change in the number particles was investigated; CESS$^* = 0.9997$ used here. For both studies, 200 replicates were used. The variance log of the normalising constant estimator is shown in Figure E.1, Appendix E.

Next, the performance of model selection was studied using identical design to the experiment immediately above. The results are shown in Figure E.2, Appendix E.

Lastly, the simulated 2-D image was analysed. For simplicity, only the NWPM and NWSE algorithms were investigated. In particular for the NWPM, $N = 200$ was studied with $n = 200$ graphical iterations. For NWSE, $N = 500$ and $n = 500$ was used. Results are shown in E.3, Appendix E.

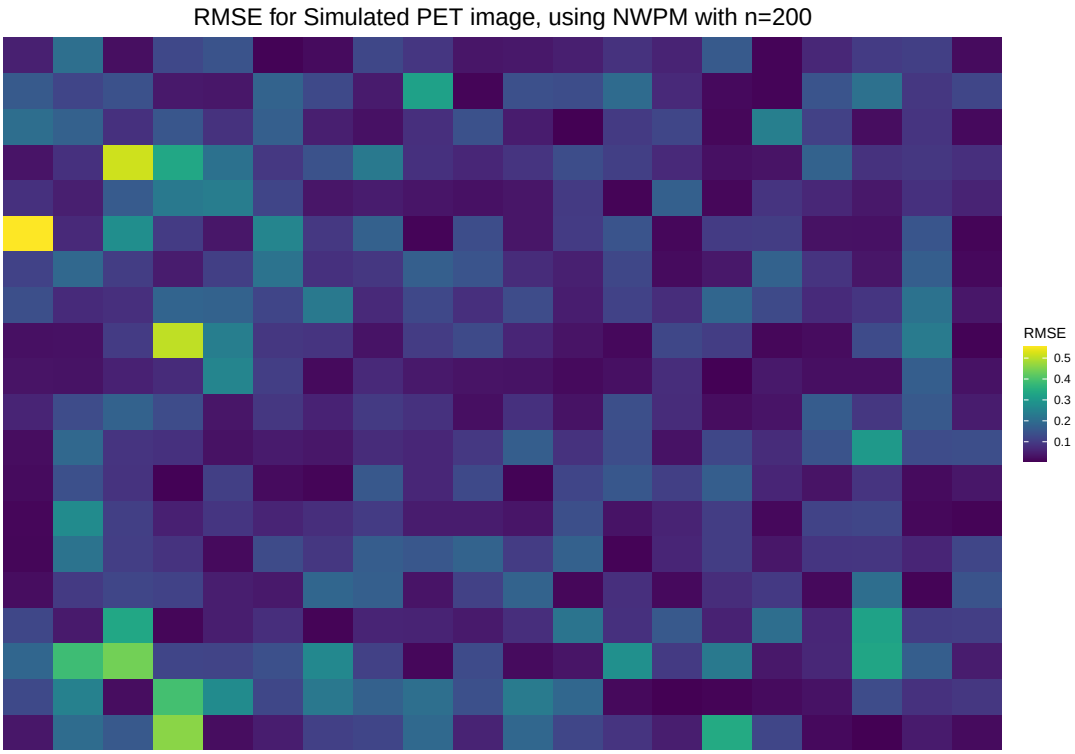


Figure 6.10: Average RMSE at each pixel, for simulated PET data (noise level = 0.5). The NWPM algorithm with $n = 200$ iterations was used.

Looking at these results, we see no notable improvements in performance, for all the studies, when compared to the corresponding Prior 5 scheme studies above. In Figure E.3, we see that NWPM outperforms NWSE, however the NWPM variant here also is more resource intensive. Additionally, the NWPM($N = 200$) with CESS-adaptive scheme does not significantly outperform the corresponding NWPM($N = 200$) with the Prior 5 scheme. Given that this adaptive scheme is more elaborate to implement, therefore more error-prone; in this setting, the minimal benefits do no justify it use.

6.3 Discussion

In this chapter, we evaluated the performance of the NWPM algorithm, together with its two variants , using simulated data. We saw that there was notable, significant improvement in model selection performance when compared to the spatially independent method. This suggest that encoding spatial dependence improves analysis, and the proposed model and method enable this improvement to be seen in these settings.

We looked at simple ways to improve performance further, by studying the trade off between tuning parameters for each of these computational method variants. Additionally we briefly also studied the effect of using an adaptive annealing scheme in the SMC sampler. However, the performance did not seem to be notably affected by these factors. In contrast, we saw that both the approximations and multiple augmentation algorithms gave almost identical performance as the NWPM algorithm. This is very promising, as it allows for greater accessibility and application of this methodology.

Chapter 7

Analysis of Measured PET Data

The human brain starts working the moment you are born and never stops until you stand up to speak in public.

— George Jessel

7.1 Introduction

The simulation studies above have verified the effectiveness of the proposed methodology and allowed for the identification of tuning parameters which lead to promising results in settings close to that of interest. We now turn our attention to measured data sets.

Analysis of the tracer kinetics of dynamic PET images, that use the [^{11}C]-dipronorphine tracer, are generally aimed towards quantification of opioid receptor concentrations in the brain. Such studies involve investigation of neurological disease which often involve change in brain receptor density. For example, PET image studies on normal subjects are used to juxtapose subsequent studies on neurological diseases such as epilepsy.

Investigation of similar PET data sets, using different statistical approaches, have been previously analysed quantitatively by Gunn et al. (2002), Peng et al. (2008) and Jiang et al. (2009); However, these works focused on parameter estimation rather than model selection. Zhou et al. (2013, 2016) are more recent works that investigated model selection; Albeit, in these works, spatial independence is assumed.

7.1.1 PET Meta-Data

We begin by first exploring some of the relevant meta-data details of the PET image that was analysed. This information will help inform the specificities (e.g. cut-off values, discussed below) of analysing this particular PET image. A test-retest pair of measured dynamic PET data of the concentration of the tracer [^{11} C]–diprenorphine in a normal subject, for which an arterial input functions were available, was analysed using the proposed and other methods. The data was initially measured as part of the dynamic PET study in Hammers et al. (2007). In that study, the effects of using lower concentration of tracer was investigated. In addition, the impact of using different slow frequency cutoffs within the analysis of the data was also studied.

PET data acquisition The subject underwent 95 min dynamic [^{11}C]–diprenorphine PET baseline scan. [^{11}C]–diprenorphine is a tracer, of the non-selective antagonist, which binds to neural opioid receptors. The PET scans were acquired in 3D mode on a Siemens/CTI ECAT EXACT3D PET camera. After image reconstruction, the spatial resolution is approximately 5mm. Prior to the scan, a 22-gauge cannula was inserted into a radial artery (after satisfactory Allen's test). Thirty seconds after the start of the scan, [^{11}C]-diprenorphine was injected and flushed through the cannula as a smooth bolus over a total of 30s. The PET data was reconstructed using the re-projection algorithm (Kinahan and Rogers, 1989) with ramp and Colsher filters cutoff at the Nyquist frequency. Reconstructed voxel size were 2.096mm $\times$ 2.096mm $\times$ 2.43mm. Acquisition was performed in listmode (event-by-event) and scans were rebinned into 32 time frames of increasing duration. The lengths of these periods, in seconds, are: (variable length background time, $3\times10, 7\times30, 12\times120, 6\times300, 75, 11\times120, 210, 5\times300, 450$ and 2×600). Frame-by-frame movement correction was performed on the PET images. Overall this resulted in images of dimensions/size $128 \times 128 \times 95$ voxels. This gives a total of $1,556,480$ separate times series respectively to be analysed. Note that this is an upper-bound, since masking over the brain regions will result in (often, orders of magnitude) fewer time series. In this study, we will analyse a cross-section of each of the sagittal, coronal and transversal planes of the brain — the total number of times series studied here is, therefore, roughly 11,000.

Derivation of Input functions Arterial blood was continuously withdrawn at a sampling rate of 5 ml/min through 100 cm of polyethylene tubing. These samples were measure in a BGO detection system, as described in Jones et al. (1994). Continuous sampling was stopped at 15 mins; additionally, discrete blood samples were taken for cross-calibration and determination of the partition of blood radioactivity at 5, 10, 15, 20, 30, 40, 50 and 60 minutes after the start of the scan. Metabolite-correct arterial plasma input functions were then created.

Details of further PET data pre-processing can be found in Hammers et al. (2007, Material and methods)

Decay Correction The PET data is not decay corrected and this should be accounted for within the compartmental model. Gunn et al. (2002) (similarly, Cunningham and Jones (1993) for Spectral Analysis), suggest that the slow frequency cutoff be close to the decay constant of the tracer used. In other words, given that the half life of [^{11}C] is 20.4min, its decay constant is 0.0005663s^{-1}; Thus, the slow frequency cutoff, denoted $\theta_{\min}$, is fixed close to this constant. Hammers et al. (2007) showed that actual parametric values depended heavily on the cutoff slow frequencies(between 0.0008 s^{-1} and 0.00063 s^{-1}). Importantly, the prior ranges of $\theta_{1:m}$ should be based upon $\theta_{\min}$, and subsequently this should be taken into consideration when computing the volume of distribution V_{D}. Following Zhou et al. (2016), in particular see Zhou (2015), we will use the cutoff $0.0007s^{-1}$.

7.2 Data Analysis: [^{11}C]-diprenorphine Data for Opioid Receptor Quantification

Normally-distributed errors have been found to poorly model data of this sort (Zhou et al., 2013). Instead we will use the $t-$distribution to model the additive error variable ϵ. The same gamma prior as used for λ in the simulation studies above, will be used for the scale parameter, τ. A uniform distribution over the interval $[0, 0.5)$ will be used as the prior for $1/\nu$, allowing the likelihood to

vary from having a very heavy tail to being arbitrarily close to normality.

The measured data was analysed using the following methods: Firstly, SMC sampler estimates of the marginal likelihood of each pixel for the three models were computed. This strategy assumes spatial independence, so, as before, we will refer to this method as the independent SMC method. The Prior 5 annealing scheme with $N = 300$ particles and $T = 500$ distributions was used, these values were based on the studies above and numerical results reported by Zhou et al. (2016). Three cross-sectional slices (coronal, transverse and sagittal planes) of the 3-D PET image were analysed, the model order parametric images are shown in Figure 7.1.

Next, the NWPM algorithm was used to analyse the same three cross-sections. An SMC sampler with $N = 300$ particles and $T = 500$ distributions was used as the marginal likelihood estimator. Here, the chain was initialised from the output of the SMC spatially independent model selection above (rather than from the prior distribution, as in the simulation studies above). A pseudo-marginal MH chain of length $n = 75$ was generated. This chain length was based on viability of computation cost for this relatively large data set, and the simulation studies above. In addition, the sagittal cross-sectional image was analysed using a longer $n = 200$ iterations, as discussed below. The results are shown in Figure 7.2.

Finally, Figure 7.3 and 7.4 show the model configuration output from analysing the measured PET data using the NWSE approximation and the NWMA variant, respectively. In this case, an SMC sampler with $N = 400$ particles and $T = 600$ distributions was used to make the marginal likelihood estimates. The pseudo-marginal chain was initialised using the output of these estimates, and these initial estimates where re-used within the NWSE algorithm for $n = 500$ iterations. For the NWMA method, the chain was initialised in the exact same manner, but the marginal estimates where updated every $\kappa = 100$, for a total of $n = 500$ iterations.

CPU runtimes, in hours, are included in the captions of each of the above mentioned figures. They can be found in the square parenthesis.

The volume of distribution V_{D} parametric images of each of these methods can be found in Figures F.1, F.2, F.3 and F.4, Appendix F. The V_{D} images, which we recall depend on the model order, give a clear and detailed visual of the brain.

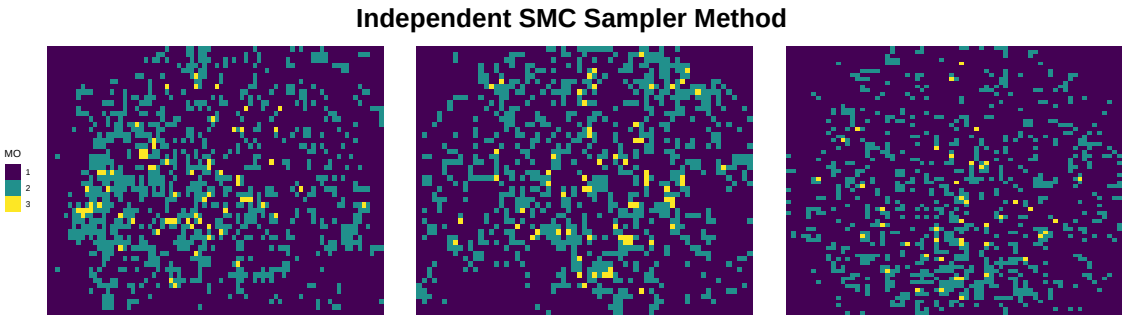


Figure 7.1: Model order parametric image of measured PET data (sagittal(left)[32.27], coronal(middle)[29.07] and transverse(right)[35.21] cross-sections), using spatially independent SMC sampler($N = 300, T = 500$) model selection.

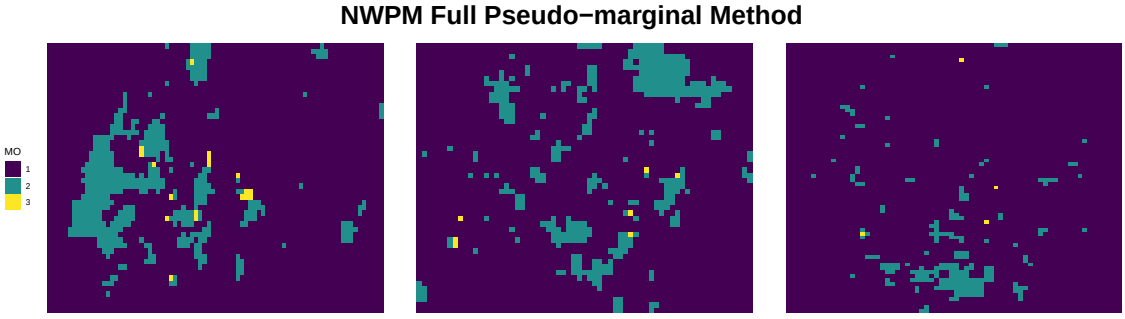


Figure 7.2: Model order parametric image of measured PET data (sagittal(left)[207.75], coronal(middle)[161.55] and transverse(right)[246.89] cross-sections), using the NWPM($n = 75$) algorithm for model selection with spatial dependence. An SMC sampler($N = 300, T = 500$) was used for the marginal likelihood estimator.

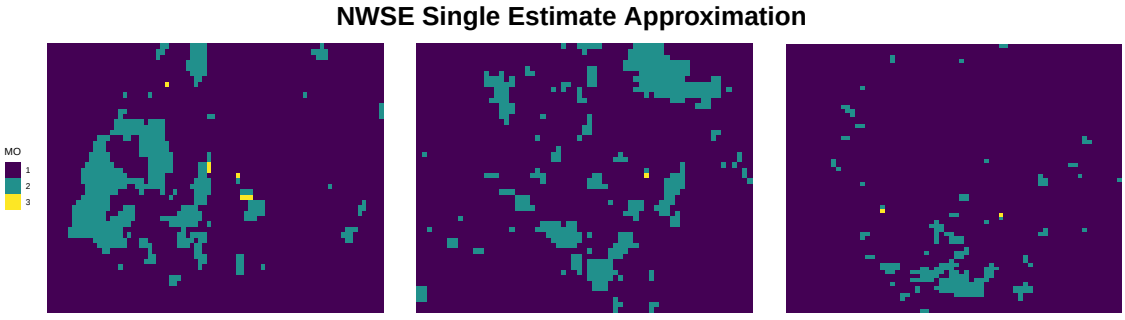


Figure 7.3: Model order parametric image of measured PET data (sagittal(left)[36.51], coronal(middle)[42.63] and transverse(right)[50.94] cross-sections), using the NWSE approximation method($n = 500$) for model selection with spatial dependence. An SMC sampler($N = 400, T = 600$) was used for the marginal likelihood estimator.

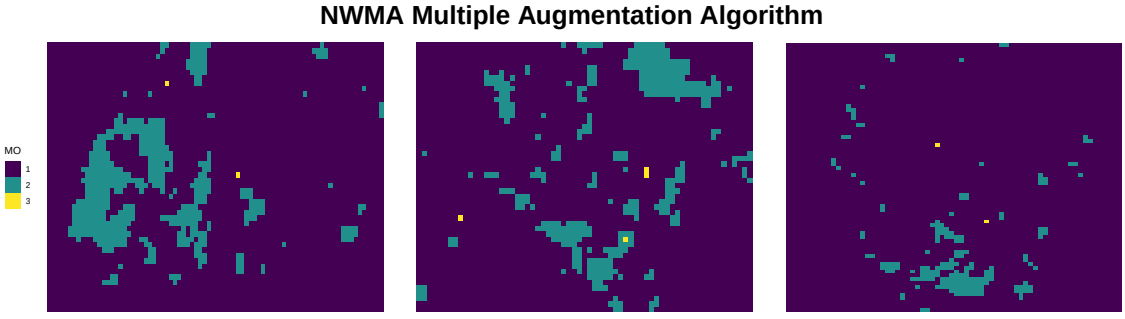


Figure 7.4: Model order parametric image of measured PET data (sagittal(left)[106.49], coronal(middle)[85.07] and transverse(right)[127.86] cross-sections), using the NWMA method($n = 500, \kappa = 100$) for model selection with spatial dependence. An SMC sampler($N = 400, T = 600$) was used for the marginal likelihood estimator.

For further comparison, the sagittal cross-section was also analysed using (corrected)AIC using parameters estimated using the NLS (non-linear least squares; Zhou et al. (2013)) method. This is shown in Figure 7.5. The spatially independent SMC sample model selection and NWPM method(with a longer chain length $n = 200$) model selection are included to show the difference in the final model order parametric image output.

The volume of distribution images in Appendix F, specifically, Figure F.1 and Figure F.2, show very

similar performance for all methods when considering inference of V_D. However, Figure 7.5 shows clear difference in the model order image when using NWPM compared to spatially independent SMC and NLS. It is evident that analysis using NWPM reveals more spatial structure: Albeit, using the SMC independent approach does reveal some of the underlying spatial information when compared to NLS; NWPM has a de-noising effect and improves the clarity of the roughly formed clusters seen in the spatially independent SMC method. In particular, the $m = 1$ and $m = 2$ compartment models are selected to model most of the structural clusters. In contrast to this there is a decrease in the number of $m = 3$ compartment models selected when compared to the non-spatial SMC method.

Note also that, by looking at Figure 7.5 we see that the NWPM algorithm converges to a noisy version of the final output image even at low iterations as $n = 25$. Importantly, there seem to be only small changes at the higher iterations, suggesting that the chain reaches a stable output relatively quickly. Similar behaviour is seen in the simulation studies above, albeit in a smaller, simpler settings.

Furthermore, similar model selection output is seen when using the SE estimate approximation and the multiple augmentation variant, NWMA, of the NWPM algorithm. Comparing both Figure 7.3 and 7.3 with Figure 7.2, we see that the model order outputs are not identical, but qualitatively show many similarities. In particular, we see that the there is a decrease in the total number of 3-compartment models selected when using the SE approximation or the NWMA variant, compared to the full NWPM. A similar effect is seen when comparing the independent SMC model selection with the NWPM - the absolute difference between the volume of distributions, at each voxel, for the two methods is shown in Figure F.5, Appendix F. In particular, as expected the magnitude of the difference is smaller outside the brain region, where the signal is smaller; However, within the regions with higher signal, the differences seem random and evenly spaced out. Finally, we note that the NWSE and NWMA model selection outputs are almost identical, differing only at a very small number of pixels.

7.3 Discussion

The proposed methods reveal new spatial structures in the model order configuration output, of measured PET data, that were otherwise impossible or very difficult to infer using current existing methods. Subsequently, better, more nuanced inferences and conclusions can be made. Albeit, making very precise conclusions about the spatial patterns of these structures, for the images studied here, is difficult. It may be possible there is not enough information in the PET signal to produce outputs of higher precision. However, general inference regarding the model order for regions of the brain can still be stated instead. In brief, the model order configuration output, produced from the proposed framework, is a parametric image that can provide useful, novel information and insights into spatial relations in PET images.

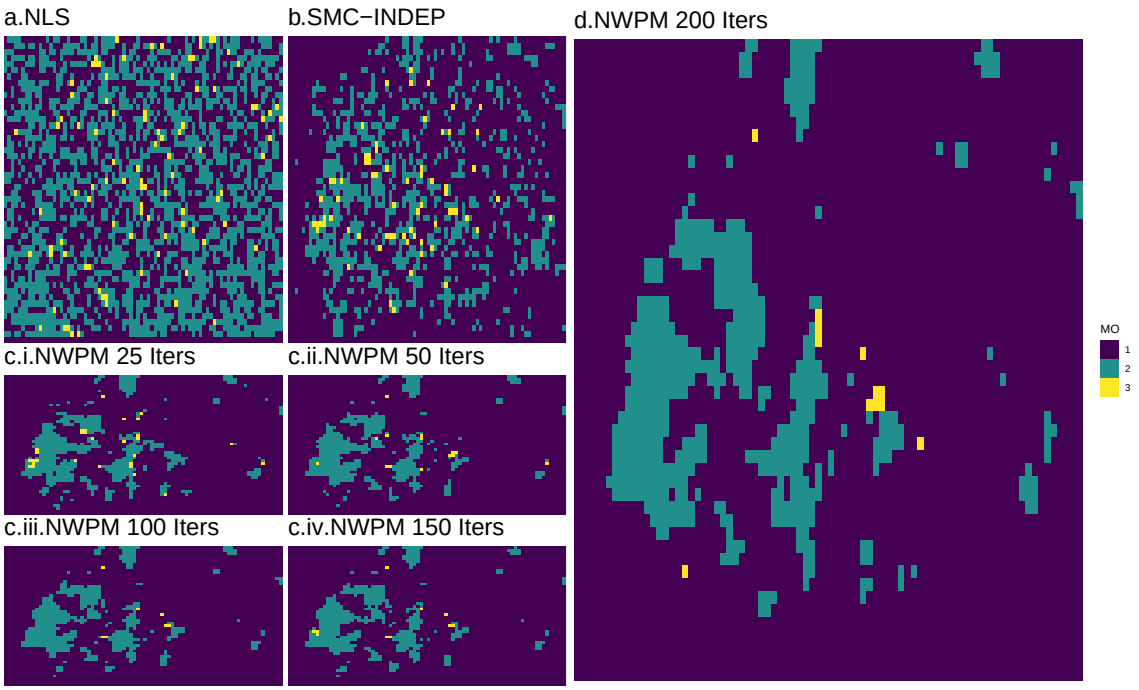


Figure 7.5: Model selection for measured PET images, using: (a.) spatially independent NLS and (b.) SMC; (d.) NWPM with spatial dependence incorporated using Potts model. Part (c.) shows the progress of the NWPM chain.

Chapter 8

An R package for Bayesian Computation using NWPM and SMC

> *Debugging is twice as hard as writing the code in the first place. Therefore, if you write the code as cleverly as possible, you are, by definition, not smart enough to debug it.*
>
> — Brian Kernighan

The `bayespetr` package is software developed during the research in order to implement the NWPM algorithm, and used to illustrate, compare and evaluate performance in the previous chapters. This package contains simple, generic NWPM samplers (together with the NWSE approximation and NWMA extension) and SMC samplers written in the accessible, intuitive and popular R language (R Core Team, 2021). The features of this software can be readily used to analyse spatial data, such as the examples discussed above, using the proposed algorithms.

In addition, another efficient SMC sampler, written in C++ and used for PET analysis, provided by the powerful `vSMC` library (Zhou, 2015), can also be accessed and used through this package; Made possible by modifying and integrating components from this library using the `RCpp` package (Eddelbuettel and François, 2011). In other words, the package also provides direct and easy access to the fast and powerful computational core of SMC samplers by `vSMC`, for the spatial (and non-spatial) analysis of PET images.

In this chapter, the main functionality of this software package is very briefly introduced and discussed. As mentioned before, the source code, documentation and other further relevant information can be found at `https://github.com/dt448/bayespetr`.

8.1 Background

There exists many computational libraries for PET image analyses; for example, the popular and extensive SPM (Statistical parametric mapping; Penny et al. (2011)) library. Other examples, among many others, include: NiftyPET (Markiewicz et al., 2018) and COMKAT (Compartment

Model Kinetic Analysis Tool; Muzic and Cornelius (2001)). These software packages tend to be used for analysis of general neuroimaging data.

In contrast, the recent `kinfitr` (Tjerkaski et al., 2020) is also an R package that is built specifically for studying the kinetics of the PET tracer using approaches such as compartmental models. However, as aforementioned, the majority of statistical methods for PET image analysis has been non-Bayesian; this is also reflected in development, and thus the availability, of software for Bayesian analysis. Still further, given even less exploration of spatial analysis, accessible software implementing Bayesian spatial analysis of PET images are virtually non-existent.

Of course, there are software implementations for general Bayesian computations like the notable `BUGS` (Bayesian inference using Gibbs sampling ;Lunn et al. (2009)); which can also be accessed through R using `rjags` (Plummer, 2016). Indeed, these can also be used to perform PET image analysis. As aforementioned, an important property motivating the use of the SMC method within the NWPM algorithm, is the unbiasedness of the normalising constant estimator. A disadvantage of SMC algorithms is that it can be difficult to implement; Where implementations do exist, they use advanced features of object oriented programming or template meta-programming, in order to be general enough for generic uses. An example of such software is the seminal work by Johansen (2009), who developed the popular C++ library `SMCTC` (SMC Template Class). The fact that this has been written in C++ means that this sampler is very efficient and fast. This is particularly important in complex settings with large data sets such as PET images. However, accessibility maybe reduced due to the use of more advanced programming concepts. Alternatively, `SMCTC` has been made more accessible by `RcppSMC`, which use the Rcpp package for R/C++ integration.

Another example of a generic SMC sampler is the `vSMC` library. This library uses similar templated approach as `SMCTC` but also exploits more modern features of the C++ language. Importantly, it is built towards the capability to do parallel computing; Additionally, it has the functionality to be essentially automatic, with little-to-no manual tuning. Due to its recency, `vSMC` is yet to be made accessible for R users via `Rcpp`. Importantly, earlier versions of this library has in built methods for PET analysis.

Though primarily developed to evaluate the performance of the NWPM algorithm, the `bayespetr` aims to also address many of the above problems. Specifically, it enables accessible software implementation of the computation for Bayesian spatial analysis for complex, large data sets such as PET images.

8.2 The `bayespetr` R package

We begin by briefly introducing the available SMC samplers in the `bayespetr` package, as they play the essential role in the NWPM algorithm. Henceforth, assume that all the SMC samplers presented use the simple Prior 5 annealing scheme, unless specified otherwise. Similarly, the target is a posterior density of a given Bayesian model at a single node/pixel/location.

Firstly, the variadic functional `smc_sampler`, written wholly in R. This sampler takes the following arguments: `logLikelihoodDensity`, `logPriorDensity`, `priorSampler`, `numSamples`, `numDistrbtns`, `dimParameter` and `datum`. The first two arguments are user defined functions which should return the density value of the log-likelihood and the log-prior, respectively. More specifically, `logLikelihoodDensity` and `logPriorDensity` should take the vector argument `parameter` with number of components equivalent to `dimParameter`. Similarly, the `priorSampler` should generate a sample from the user defined initial distribution, usually the prior, of size equivalent to

numSamples. Where possible, density and sampler functions provided by R core functions is recommended. The sampler will then run the SMC sampler algorithm for the annealing scheme with the number of intermediate distributions equivalent to numDistrbtns.

For example, the following functions are used for the toy model studied above (these are provided in the package):

```
toyLogLikelihoodDensity <- function(datum, parameters, likeSigma){
    dnorm(datum, parameters, likeSigma, log = T)
  }

toyLogPriorDensity <- function(parameters, priorSigma, mu){
    dnorm(parameters, mu, priorSigma, log = T)
  }

toyBioPriorSampler <- function(numSamples, mu, likeSigma){
    rnorm(numSamples, mu, priorSigma)
  }
```

To sample from the posterior distribution, call the function with the appropriate variables:

```
smc_sampler(logLikelihoodDensity = toyLogLikelihoodDensity,
            logPriorDensity = toyLogPriorDensity,
            priorSampler = toyBioPriorSampler,
            numSamples = 200,
            numDistributions = 100,
            dimParameter = 1,
            datum = dataSet,
            priorSigma = 5,
            likeSigma = 2,
            mu = 0)
```

The output of the above call is a list variable containing relevant information and quantities; Such as, the weighted sample, normalising constant estimate, volume of distribution, annealing schemes etc.

Similar methods (likelihood density, prior density and prior sampler) for the PET compartmental models are also provided. For high–performance, a C++ clone of this sampler and the associated density functions, for both settings, are also available n this package through Rcpp.

Next, for the PET setting the SMC sampler provided by the vSMC library can also be accessed and used. Importantly, this means that any parallel-computation capability provided by this library, can also be readily exploited if required. Given this advantage, we use this SMC sampler within the NWPM sampler described below.

The vSMC library uses templates of C++ code to implement the SMC method. It is structured through modules that abstract variants of the SMC samplers, see Zhou (2015) and references therein for further discussion. In the original vSMC library, the configuration of the known model parameters and tuning parameters of the SMC sampler is given by an auxiliary file. In the Rcpp form of this package, called via the pet_SMC method, this will be given in the form of lists.

In particular the config_model and config_algo arguments, of the pet_SMC method, both take list variables that configure the compartmental model used and the SMC sampler itself. For

example, the following variable can be used to specify an SMC sampler that uses an AMCMC kernels for local moves, repeated twice at each iteration and uses the CESS-adaptive annealing scheme with threshold $\text{CESS}^* = 0.999$.

```
demo_algo_config = list(mcmc_type = "AMCMC",
                        mcmc_iters = 4,
                        annealing_scheme = "CESS",
                        cess_threshold = 0.999)
```

Similarly, the following examples specifies the time intervals used, the range of the prior (non-informative uniform, see package documentation for specification) distribution, a t-distribution for the error innovations, and $\theta_{\min} = 0.0007$

```
demo_model_config = list(time_intervals = c(45, 10, 10, 10, 30, 30, 30,
                                            30, 30,  30, 30, 120, 120, 120,
                                            120, 120, 120, 120, 120, 120,
                                            120, 120, 120, 300, 300, 300,
                                            300, 300, 300, 600, 600, 600),
                         prior_ranges = c(1e-5, 1e-5, 1e-5,
                                          1e-2, 1e-2, 1e-2,
                                          2e-4, 1e-5, 1e-5,
                                          1e-2, 1e-2, 1e-2,
                                          1e-1,
                                          10,
                                          0,
                                          5e-1),
                         t_distribution = T,
                         decay_lowest_rate = 0.0007)
```

Finally, the SMC sampler method for PET compartment models, can be called; Where, we use the arguments to specify the number of particles, and the dimension of the parameter space. For example, for a 1-compartment model, the `dimParameter` should be set to 3. That is, the parameters are ϕ_1 and θ_1 as well as the variance σ^2 of the error):

```
pet_SMC(numSamples = 200, datum = data1, dimParameter = 7,
        demo_model_config,demo_algo_config,  randSeed = 10)
```

The data (time series) can be given as a `vector` or `data.frame` variable. For 2-D, see blow, an `array` can also be used .

An important modification here is that, the computational pseudo-random number generator used for this sampler are from the standard R library — rather than the generators provided by the vSMC library. In particular, this allows R users to directly set the random seed, enabling reproducible experiments.

Finally, these samplers use a look-up table interpolation optimisation to calculate C_{T}. To generate this table the plasma input function is required, see documentation for further information.

8.2.1 NWPM for PET with compartmental models

We now briefly present functionalities that allow us to analyse dynamical 2-D PET images. Firstly, the SMC sampler, described above, can be used for model selection, assuming no spatial depen-

dence. The relevant method is `pet_smc_indep`, where we simply specify the tuning parameters of the SMC sampler and, as before, pass the relevant configurations.

```
pet_smc_indep(imageData = petImageData,
              smcParameters = list(numSamples = 400,
                                   numDistrbtns = 600),
              config_model = demo_model_config,
              config_algo = demo_algo_config)
```

Next, the NWPM sampler uses similar conventions. Specifically, the length of the pseudo-marginal MH chain is determined by the argument `numberOfIterations`, the coupling constant by `J` and the state space (model order space) at each node $\mathcal{M}$ by `pottsStateSpace`.

```
pet_NWPM(imageData = petImageData,
         numberOfIterations = 50,
         smcParameters = list(numSamples = 200,
                              numDistrbtns = 400),
         J = 0.4,
         pottsStateSpace = c(1,2,3),
         config_model= demo_model_config,
         config_algo = demo_algo_config)
```

The SE variant of this sampler can be called using the method `pet_NWPM_SE` — it uses almost identical arguments. Similarly, the NWMA algorithm can be accessed through the sampler associated to the method `pet_NWMP_MA2`. The extra argument `refreshConst`, here, specifies the "refresh rate" or the frequency of refreshing the marginal estimates, which we recall was denoted κ above.

```
pet_NWPM_MA2(imageData = petImageData,
             numberOfIterations = 500,
             smcParameters = list(numSamples = 200,
                                  numDistrbtns = 400),
             J = 0.4,
             pottsStateSpace = c(1,2,3),
             config_model= demo_model_config,
             config_algo = demo_algo_config,
             refreshConst = 50)
```

An NWPM sampler for the toy model is also provided, this uses the Rcpp SMC sampler (i.e. not the vSMC library). The relevant methods are `toy_SMC`, `toy_NWPM`, `toy_NWPM_MA2` and `toy_NWPM_SE`.

8.3 Summary

Looking at the estimated/recorded CPU runtimes quoted in Section 6.1.4, 6.2.4 and Section 7.2 above, we can see that the quickest method in regards to computational cost is the NWSE method. This is expected as it is essentially an approximation to the very expensive NWPM method. Similarly, the NWMA method acts as an intermediate between these two methods — empirically achieving runtimes close to the NWSE method without the need to do approximations.

In this chapter, we briefly presented and discussed the `bayespetr` package. The primary objective of this package is to provide accessible computational implementation of the NWPM algorithm. This enables researchers and other users to analyse spatial data in the framework proposed in this

thesis. The increasingly popular and widely used R language provides a simple, intuitive programming platform; Thus, it is easy to modify and extend this package. For example, various novel techniques based on the multiple augmentation space can be readily implemented and evaluated with little effort. An auxiliary feature of this package is that it allows easy and direct access to the powerful SMC sampler that can be readily used on parallel hardware. This, allows parallel-computing SMC samplers for researchers and other users who may not be proficient in C++ or advanced programming concepts. In summary, this software package allows for non-spatial and spatial Bayesian model selection, in a simple, intuitive format.

Chapter 9

Conclusions

If you are not having fun, you are not learning. There's a pleasure in finding things out.

— Richard Feynman

In this work, we have illustrated a novel computational method for effectively incorporating spatial dependence when performing model selection at each location within a graph. This approach extends the pseudo-marginal method, in a number of directions, within the context that it has been developed and evaluated in. By exploiting the structure of the problem at hand, the developed methodology allows for considerable flexibility in the updating of state variables and marginal likelihood estimates relative to generic pseudo-marginal algorithms. The empirical study suggests that this approach can yield better inference than methods which impose assumptions of full spatial independence and, in particular, can reveal clearer spatial structure in the image of underlying model orders. As with pseudo-marginal algorithms, essentially any unbiased marginal likelihood estimators used in non-spatial analysis can be used within this algorithm.

This methodology was motivated by considering the problem of model selection in PET images. In such a setting, both the amount of data and the complexity of the models lead to substantial computational requirements. The images of model order provided by this approach supplement the information provided by volume of distribution and similar macro-parameters in PET images. For instance, Figure 7.5 could be used to make the interpretation that the signal in certain regions of the brain (posterior) requires a higher model order when compared to other (anterior) regions.

In order to reduce the computational cost, a multiple augmentation variant of the pseudo-marginal algorithm is introduced, in addition to a more extreme approximate form in which in which each marginal likelihood is estimated only once. The performance of these methods in the examples explored here is very encouraging: it suggests that where non-spatial analysis has been performed, spatial dependence can be incorporated with the existing estimates very easily and with little computational costs. Doing so carefully can result in similar performance to the full NWPM algorithm.

In summary, the NWPM algorithm, together with its approximations and extensions, enables a intuitive, natural approach to incorporating spatial dependence in realistic, complex and challenging settings; such as analysis of PET images. Further, the methods extensions and approximations means that the subsequent performance and results of incorporating spatial information is very

accessible and readily applicable.

9.1 Contributions

A generic hierarchical model, that enables incorporation of spatial dependence for image and image-like data sets in a natural and intuitive manner was presented in Chapter 5. The associated novel method for inference and model selection from this model, was also introduced in said chapter. This efficient, readily applicable algorithm carefully exploits the structure of the objective; through utilising popular and well–studied Monte Carlo methods in prudent conjunction with existing, proven models which assume full spatial independence. The result is an accessible framework which uses unbiased marginal likelihood estimators, that are often used in existing non-spatial analysis, to allow for the inference based on spatial structures in data set. Flexible extensions and approximations of the NWPM algorithm, that greatly reduce computational requirements, were also presented. These variants of the algorithm, expand the accessibility of this proposed approach further; Allowing for significant reduction in the computation overhead. The multiple space augmentation gives rise to a framework that has the potential for many future algorithmic innovations. These methods were empirically studied extensively in the setting of PET image analysis, revealing spatial structures in the model order image outputs.

Software implementation of these algorithms was briefly introduced in Chapter 8. The presented R package, `bayespetr`, provides direct access to the NWPM algorithm, as well as the two variants studied in the numerical studies above. The intuitive but specialised R language, means that this package extends the accessibility of these methods to more researchers and other users; Additionally, access to the integrated computational core of C++ SMC sampler(s) means that the high performance of the C++ language, is also maintained.

9.2 Future Directions

An important factor in the application of almost all MCMC methods, including pseudo-marginal algorithms, such as the NWPM here, is the time and effort required for manual tuning. More specifically, in this case, heuristically determining the coupling constant; and the tuning of the SMC sampler parameters N and T, to allow for optimal mixing. Roughly speaking, tuning N to allow for optimal efficiency could be made adaptive using SMC variance estimators, for example using the estimator as proposed by Lee and Whiteley (2018) or Du and Guyader (2021). Similarly, future work could be dedicated to exploring approaches to incorporate methods for inferring the coupling constant, such as Møller et al. (2006); Moores et al. (2020), within the NWPM algorithm.

For example, looking at Figure 6.10, in general we see that the RMSE is higher at "harder" nodes. That is, nodes of the ground truth configuration that have neighbourhoods with higher number of states that are different to said node. This motivates the consideration that N and similarly T need not be fixed. Instead, generalise these parameters at each node and, even to, graphical iterations. In other words, denote this $N_v^{(i)}$ and $T_v^{(i)}$ — these values could be determined in an adaptive manner, by criterion involving the SMC variance estimators alluded to above. This means that more computational resource is allocated to nodes that require it. Doing this in a careful, principled manner will improve both model selection and inference in an efficient manner. This extension could also be used within the multiple augmentation algorithms; efficiently, changing the tuning parameters at each re-estimation of the marginals.

Next, we have seen that empirical evidence suggest that it is possible to attain similar performance using approximations and multiple augmentation variants. This is useful, as for most cases of realistic application a lower computational requirement is highly appealing. However, ideally further theoretical results and properties could be explored to formally understand the performance seen empirically. For example, the theoretical exploration of the effect of the SE approximation by Proposition 5.6.1 can be extended to the whole graph. Similar exploration can also be undertaken for the multiple augmentation approach, particularly in comparison to the NWPM algorithm.

Part III

Appendices and Bibliography

Bibliography

H. Akaike. *Information Theory and an Extension of the Maximum Likelihood Principle*, pages 199–213. Springer New York, New York, NY, 1973.

C. Andrieu and G. O. Roberts. The pseudo-marginal approach for efficient Monte Carlo computations. *The Annals of Statistics*, 37(2):697–725, 2009.

C. Andrieu and M. Vihola. Convergence properties of pseudo-marginal Markov chain Monte Carlo algorithms. *Annals of Applied Probability*, 25(2):1030–1077, 2015.

C. Andrieu and M. Vihola. Establishing some order amongst exact approximations of MCMCs. *Annals of Applied Probability*, 26(5), 2016.

C. Andrieu, A. Doucet, and R. Holenstein. Particle Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 72(3):269–342, 2010.

J. A. Aston and A. M. Johansen. Bayesian inference on the brain : Bayesian solutions to selected problems in neuroimaging. In *Current Trends in Bayesian Methodology with Applications*, pages 1–36. CRC Press, 2015.

S. Banerjee, B. P. Carlin, and A. E. Gelfand. *Hierarchical Modeling and Analysis for Spatial Data*. Chapman and Hall CRC, 2014.

M. A. Beaumont. Estimation of Population Growth or Decline in Genetically Monitored Populations. *Genetics*, 164(3):1139–1160, 2003.

J. Besag. Spatial Interaction and the Statistical Analysis of Lattice Systems. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(2), 1974.

J. Besag. Statistical analysis of dirty pictures. *Journal of Applied Statistics*, 20(5-6), 1993.

M. Bezener, J. Hughes, and G. Jones. Bayesian spatiotemporal modeling using hierarchical spatial priors, with applications to functional magnetic resonance imaging (with discussion). *Bayesian Analysis*, 13(4), 2018.

S. Brooks, A. Gelman, G. Jones, and X.-L. Meng. *Handbook of Markov Chain Monte Carlo*. CRC press, 2011.

A. Buck, G. Westera, G. von Schulthess, and C. Burger. Modeling Alternatives for Cerebral Carbon-11-Iomazenil Kinetics. *Journal of nuclear medicine, Society of Nuclear Medicine*, 37: 699–705, 1996.

Y. Chen. Another look at rejection sampling through importance sampling. *Statistics and Probability Letters*, 72(4), 2005.

N. Chopin. A Sequential Particle Filter Method for Static Models. *Biometrika*, 89:539–552, 2001.

N. Chopin. Central limit theorem for sequential Monte Carlo methods and its application to Bayesian inference. *Annals of Statistics*, 32(6), 2004.

N. Chopin and O. Papaspiliopoulos. An Introduction to Sequential Monte Carlo. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 52(4), 2020.

G. Claeskens and N. L. Hjort. *Model Selection and Model Averaging*. Cambridge University Press, 2010.

A. H. Compton. A Quantum Theory of the Scattering of X-rays by Light Elements. *Physical Review*, 21(5), 1923.

P. Congdon. *Bayesian Models for Categorical Data*. Wiley and Sons, 2006.

N. Cressie. *Statistics For Spatial Data*. John Wiley & Sons, 1992.

V. J. Cunningham and T. Jones. Spectral analysis of dynamic PET studies. *Journal of Cerebral Blood Flow and Metabolism*, 13(1):15–23, 1993.

P. Del Moral. Nonlinear Filtering Using Random Particles. *Theory of Probability & Its Applications*, 40(4):690–701, 1996.

P. Del Moral. *Feynman-Kac Formulae: Genealogical and Interacting Particle Systems With Applications*, volume 100. New York: Springer, 2004.

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

R. Douc, O. Cappé, and E. Moulines. Comparison of resampling schemes for particle filtering. In *Image and Signal Processing and Analysis, 2005. ISPA 2005. Proceedings of the 4th International Symposium*, volume 2005, 2005.

A. Doucet and A. M. Johansen. A Tutorial on Particle Filtering and Smoothing : Fifteen years later. *Handbook of Nonlinear Filtering*, 2011.

A. Doucet, M. K. Pitt, G. Deligiannidis, and R. Kohn. Efficient implementation of Markov chain Monte Carlo when using an unbiased likelihood estimator. *Biometrika*, 102(2):295–313, 2015.

Q. Du and A. Guyader. Variance estimation in adaptive sequential Monte Carlo. *The Annals of Applied Probability*, 31(3):1021–1060, 2021.

S. Duane, A. D. Kennedy, B. J. Pendleton, and D. Roweth. Hybrid monte carlo. *Physics letters B*, 195(2):216–222, 1987.

D. Eddelbuettel and R. François. Rcpp: Seamless R and C++ integration. *Journal of Statistical Software*, 40(8), 2011.

B. Efron. Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, 1979.

R. G. Everitt. Bayesian parameter estimation for latent markov random fields and social networks. *Journal of Computational and Graphical Statistics*, 2012.

Y. Fan, G. Emvalomenos, C. Grazian, and S. R. Meikle. Pet-abc: fully bayesian likelihood-free inference for kinetic models. *Physics in Medicine & Biology*, 66(11):115002, 2021.

S. H. Friedberg, A. J. Insel, and L. E. Spence. *Linear Algebra*. Featured Titles for Linear Algebra (Advanced) Series. Pearson Education, 2003.

S. Geisser. The predictive sample reuse method with applications. *Journal of the American Statistical Association*, 70(350), 1975.

A. E. Gelfand and D. K. Dey. Bayesian Model Choice: Asymptotics and Exact Calculations. *Journal of the Royal Statistical Society. Series B (Methodological)*, 56, 1994.

A. E. Gelfand, P. Diggle, P. Guttorp, and M. Fuentes. *Handbook of Spatial Statistics*. CRC Press, 2010.

D. Geman. Random fields and inverse problems in imaging. pages 115–193. Springer Berlin Heidelberg, 1990. ISBN 978-3-540-46718-2. doi: 10.1007/bfb0103042.

S. Geman and D. Geman. Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-6(6): 721–741, 1984.

M. Gerber, N. Chopin, and N. Whiteley. Negative association, ordering and convergence of resampling methods. *Annals of Statistics*, 47(4), 2019.

R. J. Glauber. Time-dependent statistics of the Ising model. *Journal of Mathematical Physics*, 4 (2), 1963.

N. J. Gordon, D. J. Salmond, and A. F. M. Smith. Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEE Proceedings F - Radar and Signal Processing*, 140(2):107–113, apr 1993.

P. J. Green. Reversible jump Markov chain monte carlo computation and Bayesian model determination. *Biometrika*, 82(4), 1995.

P. J. Green and J. Heikkinen. Trans-dimensional Markov chain Monte Carlo. In *Highly Structured Stochastic Systems*. University Press, 2003.

R. N. Gunn, S. R. Gunn, and V. J. Cunningham. Positron emission tomography compartmental models. *Journal of Cerebral Blood Flow and Metabolism*, 21(6):635–652, 2001.

R. N. Gunn, S. R. Gunn, F. E. Turkheimer, J. A. Aston, and V. J. Cunningham. Positron emission tomography compartmental models: A basis pursuit strategy for kinetic modeling. *Journal of Cerebral Blood Flow and Metabolism*, 22(12):1425–1439, 2002.

M. Hairer and J. C. Mattingly. Yet Another Look at Harris' Ergodic Theorem for Markov Chains. In *Seminar on Stochastic Analysis, Random Fields and Applications VI*. 2011.

A. Hammers, M. C. Asselin, F. E. Turkheimer, R. Hinz, S. Osman, G. Hotton, D. J. Brooks, J. S. Duncan, and M. J. Koepp. Balancing bias, reliability, noise properties and the need for parametric maps in quantitative ligand PET: [11C]diprenorphine test-retest data. *NeuroImage*, 38(1):82–94, 2007.

T. E. Harris. The Existence of Stationary Measures for Certain {M}arkov Processes. *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, 1954–1955, vol. II*, 1956.

W. A. Hastings. Monte Carlo Sampling Methods Using Markov Chains and Their Applications. *Biometrika*, 57(1):97–109, 1970.

D. M. Higdon. Auxiliary variable methods for Markov chain Monte Carlo with applications. *Journal of the American Statistical Association*, 93(442), 1998.

M. Hinne, Q. F. Gronau, D. van den Bergh, and E.-J. Wagenmakers. A conceptual introduction to bayesian model averaging. *Advances in Methods and Practices in Psychological Science*, 3(2): 200–215, 2020.

M. A. Hurn, O. K. Husby, and H. Rue. *A Tutorial on Image Analysis*. Number April. Springer New York, New York, NY, 2003.

E. Ising. Beitrag zur Theorie des Ferromagnetismus. *Zeitschrift für Physik*, 31(1):253–258, 1925.

A. Jasra, D. A. Stephens, A. Doucet, and T. Tsagaris. Inference for Lévy-Driven Stochastic Volatility Models via Adaptive Sequential Monte Carlo. *Scandinavian Journal of Statistics*, 38 (1):1–22, 2011.

H. Jeffreys. Some Tests of Significance, Treated by the Theory of Probability. *Mathematical Proceedings of the Cambridge Philosophical Society*, 31:203–222, 1935.

C. R. Jiang, J. A. Aston, and J. L. Wang. Smoothing dynamic positron emission tomography time courses using functional principal components. *NeuroImage*, 47(1):184–193, 2009.

A. M. Johansen. SMCTC: Sequential Monte Carlo in C++. *Journal of Statistical Software*, 30(6), 2009.

A. K. Jones, V. J. Cunningham, S. K. Ha-Kawa, T. Fujiwara, Q. Liyii, S. K. Luthra, J. Ashburner, S. Osman, and T. Jones. Quantitation of [11C]diprenorphine cerebral kinetics in man acquired by PET using presaturation, pulse-chase and tracer-only protocols. *Journal of Neuroscience Methods*, 1994.

R. L. Kashyap. Inconsistency of the AIC Rule for Estimating the Order of Autoregressive Models. *IEEE Transactions on Automatic Control*, 25(5), 1980.

R. E. Kass and A. E. Raftery. Bayes factors. *Journal of the American Statistical Association*, 90 (430):773–795, 1995.

P. E. Kinahan and J. G. Rogers. Analytic 3D image reconstruction using all detected events. *IEEE Transactions on Nuclear Science*, 36(1):964–968, 1989.

A. Kong. A Note on Importance Sampling using Standardized Weights. Technical Report 348, University of Chicago, Department of Statistics, 1992.

K. Kubota, T. Matsuzawa, M. Ito, K. Ito, T. Fujiwara, Y. Abe, S. Yoshioka, H. Fukuda, J. Hatazawa, and R. Iwata. Lung Tumor Imaging by Positron Emission Tomography Using C-11 L-Methionine. *Journal of nuclear medicine : official publication, Society of Nuclear Medicine*, 26:37–42, 1985.

S. Kullback and R. A. Leibler. On Information and Sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951.

A. Lee and N. Whiteley. Variance estimation in the particle filter. *Biometrika*, 105(3):609–625, 2018.

S. Lee and A. Karagrigoriou. An Asymptotically Optimal Selection of the Order of a Linear Process. *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002)*, 63(1):93–106, 2001.

K. Lehtio, V. Oikonen, S. Nyman, T. Graonroos, A. Roivainen, O. Eskola, and H. Minn. Quantifying tumour hypoxia with fluorine-18 fluoroerythronitroimidazole([F-18]FETNIM) and PET

using the tumour to plasma ratio. *European journal of nuclear medicine and molecular imaging*, 30:101–108, 2003.

F. Lindsten, A. M. Johansen, C. A. Naesseth, B. Kirkpatrick, T. B. Schön, J. A. Aston, and A. Bouchard-Côté. Divide-and-Conquer With Sequential Monte Carlo. *Journal of Computational and Graphical Statistics*, 26(2):445–458, 2017.

J. S. Liu. Metropolized independent sampling with comparisons to rejection sampling and importance sampling. *Statistics and Computing*, 6(2):113–119, 1996.

J. S. Liu. *Monte Carlo strategies in Scientific Computing*, volume 10. Springer, 2001.

S. Livingstone and G. Zanella. The Barker proposal: combining robustness and efficiency in gradient-based MCMC. *arXiv preprint arXiv:1908.11812*, 2020.

J. Logan, J. S. Fowler, N. D. Volkow, A. P. Wolf, S. L. Dewey, D. J. Schlyer, R. R. MacGregor, R. Hitzemann, B. Bendriem, S. J. Gatley, and D. R. Christman. Graphical Analysis of Reversible Radioligand Binding from Time-Activity Measurements Applied to [N-11C-Methyl]-Cocaine PET Studies in Human Subjects. *Journal of Cerebral Blood Flow & Metabolism*, 10(5): 740–747, 1990.

D. Lunn, D. J. Spiegelhalter, A. Thomas, and N. Best. The BUGS project: Evolution, critique and future directions. *Statistics in Medicine*, 28(25), 2009.

D. A. Mankoff, A. F. Shields, M. M. Graham, J. M. Link, J. F. Eary, and K. A. Krohn. Kinetic analysis of 2-[carbon-11]thymidine PET imaging studies: Compartmental model and mathematical analysis. *Journal of Nuclear Medicine*, 39(6):1043–1055, 1998.

P. J. Markiewicz, M. J. Ehrhardt, K. Erlandsson, P. J. Noonan, A. Barnes, J. M. Schott, D. Atkinson, S. R. Arridge, B. F. Hutton, and S. Ourselin. NiftyPET: a High-throughput Software Platform for High Quantitative Accuracy and Precision PET Imaging and Analysis. *Neuroinformatics*, 16(1), 2018.

V. Matveev and R. Shrock. Complex-temperature singularities in Potts models on the square lattice. *Physical Review E - Statistical Physics, Plasmas, Fluids, and Related Interdisciplinary Topics*, 54(6), 1996.

N. S. Metropolis. The beginning of the Monte Carlo method. *Los Alamos Science (1987 Special Issue dedicated to Stanis?aw Ulam)*, 1987.

N. S. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. J. Teller. Equation of State Calculations by Fast Computing Machines. *Journal of Chemical Physics*, 21:1087–1092, 1953.

S. Meyn and R. L. Tweedie. *Markov chains and stochastic stability, second edition.* 2009.

J. Møller, A. N. Pettitt, R. Reeves, and K. K. Berthelsen. An efficient Markov chain Monte Carlo method for distributions with intractable normalising constants. *Biometrika*, 93(2), 2006.

M. Moores, G. K. Nicholls, A. N. Pettitt, and K. Mengersen. Scalable bayesian inference for the inverse temperature of a hidden potts model. *Bayesian Analysis*, 15(1), 2020.

E. D. Morris, C. J. Endres, K. C. Schmidt, B. T. Christian, R. F. Muzic, and R. E. Fisher. Kinetic Modeling in Positron Emission Tomography. In *Emission Tomography: The Fundamentals of PET and SPECT*, pages 499–540. Elsevier Inc., 2004.

J. Muzic and S. Cornelius. COMKAT: Compartment model kinetic analysis tool. *Journal of Nuclear Medicine*, 42(4), 2001.

C. A. Naesseth, F. Lindsten, and T. B. Schön. Elements of sequential Monte Carlo, 2019.

R. M. Neal. Annealed importance sampling. *Statistics and Computing*, 11(2):125–139, 2001.

M. Newton and A. E. Raftery. Approximate Bayesian Inference by the Weighted Likelihood Bootstrap. *Journal of the Royal Statistical Society Series B-Methodological*, 56:3–48, 1994.

M. S. Oh and J. O. Berger. Adaptive importance sampling in monte carlo integration. *Journal of Statistical Computation and Simulation*, 41(3-4), 1992.

L. Onsager. Crystal statistics. I. A two-dimensional model with an order-disorder transition. *Physical Review*, 65(3-4):117–149, 1944.

C. Patlak S. and R. Blasberg. Graphical Evaluation of Blood-to-Brain Transfer Constants from Multiple-Time Uptake Data. Generalizations. *Journal of cerebral blood flow and metabolism : official journal of the International Society of Cerebral Blood Flow and Metabolism*, 5:584–590, 1986.

J. Y. Peng, J. A. Aston, R. N. Gunn, C. Y. Liou, and J. Ashburner. Dynamic positron emission tomography data-driven analysis using sparse Bayesian learning. *IEEE Transactions on Medical Imaging*, 27(9):1356–1369, 2008.

W. D. Penny, K. J. Friston, J. T. Ashburner, S. J. Kiebel, and T. E. Nichols. *Statistical parametric mapping: the analysis of functional brain images*. Elsevier, 2011.

M. E. Phelps, S. R. Cherry, and M. Dahlbom. *PET: Physics, instrumentation, and scanners*. Springer New York, 2006.

M. Plummer. rjags: Bayesian graphical models using MCMC, 2016.

R. B. Potts. Some generalized order-disorder transformations. *Mathematical Proceedings of the Cambridge Philosophical Society*, 48(1):106–109, 1952.

R Core Team. R: A language and environment for statistical computing., 2021. URL https://www.r-project.org/.

A. E. Raftery, D. Madigan, and J. A. Hoeting. Bayesian Model Averaging for Linear Regression Models. *Journal of the American Statistical Association*, 92(437), 1997.

C. P. Robert. *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation*, volume 91. Springer Texts in Statistics, 2007.

C. P. Robert and G. Casella. *Monte Carlo Statistical Methods*. Springer-Verlag, 2005.

G. O. Roberts and J. S. Rosenthal. Optimal scaling of discrete approximations to Langevin diffusions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60:255–268, 1998.

G. O. Roberts and J. S. Rosenthal. Optimal Scaling for Various Metropolis-Hastings Algorithms. *Statistical Science*, 16(4), 2001.

G. O. Roberts and R. L. Tweedie. Exponential Convergence of Langevin Diffusions and Their Discrete Approximations. *Bernoulli*, 2, 1995.

G. O. Roberts, A. Gelman, and W. R. Gilks. Weak Convergence And Optimal Scaling Of Random Walk Metropolis Algorithms. *The Annals of Applied Probability*, 7, 1997.

D. B. Rubin. Bayesianly justifiable and relevant frequency calculations for the applied statistician. *The Annals of Statistics*, pages 1151–1172, 1984.

I. Sarikaya. PET studies in epilepsy. *Am J Nucl Med Mol Imaging*, 5(5), 2015.

K. C. Schmidt. Which linear compartmental systems can be analyzed by spectral analysis of PET output data summed over all compartments? *Journal of Cerebral Blood Flow and Metabolism*, 19(5):560–569, 1999.

G. Schwarz. Estimating the Dimension of a Model. *The Annals of Statistics*, 6:461–464, 1978.

G. A. Seber and C. Wild. *Nonlinear Regression*. Wiley, 2003.

J. Shao. Linear model selection by cross-validation. *Journal of the American Statistical Association*, 88(422), 1993.

C. Sherlock. Optimal Scaling for the Pseudo-Marginal Random Walk Metropolis: Insensitivity to the Noise Generating Mechanism. *Methodology and Computing in Applied Probability*, 18(3), 2016.

C. Sherlock, A. H. Thiery, G. O. Roberts, and J. S. Rosenthal. On the efficiency of pseudo-marginal random walk metropolis algorithms. *Annals of Statistics*, 43(1), 2015.

B. W. Silverman. *Density Estimation for Statistics and Data Analysis*. 1986.

C. Y. Sin and H. White. Information criteria for selecting possibly misspecified parametric models. *Journal of Econometrics*, 71(1-2), 1996.

D. J. Spiegelhalter, N. G. Best, B. P. Carlin, and A. Van Der Linde. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 64(4), 2002.

T. Spinks, T. Jones, P. Bloomfield, D. Bailey, M. Miller, D. Hogg, W. Jones, K. Vaigneur, J. Reed, J. Young, D. Newport, C. Moyers, M. Casey, and R. Nutt. Physical characteristics of the ECAT EXACT3D positron tomograph. *Physics in Medicine & Biology*, 45(9):2601, 2000.

M. Stone. Cross-Validatory Choice and Assessment of Statistical Predictions (With Discussion). *Journal of the Royal Statistical Society: Series B (Methodological)*, 38(1), 1976.

M. Stone. An Asymptotic Equivalence of Choice of Model by Cross-Validation and Akaike's Criterion. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1), 1977.

R. H. Swendsen and J. S. Wang. Nonuniversal critical dynamics in Monte Carlo simulations. *Physical Review Letters*, 58(2), 1987.

K. Takeuchi. Distribution of informational statistics and a criterion of model fitting. *Mathematical Sciences*, 1976.

D. Thesingarajah and A. M. Johansen. The Node-wise Pseudo-marginal Method. *arXiv preprint arXiv:2109.08573*, 2021. URL http://arxiv.org/abs/2109.08573.

L. Tierney. A note on Metropolis-Hastings kernels for general state spaces. *Annals of Applied Probability*, 8(1), 1998.

H. Tjelmeland and J. Besag. Markov random fields with higher-order interactions. *Scandinavian Journal of Statistics*, 25(3), 1998.

J. Tjerkaski, S. Cervenka, L. Farde, and G. J. Matheson. Kinfitr — an open-source tool for reproducible PET modelling: validation and evaluation of test-retest reliability. *EJNMMI Research*, 10(1), 2020.

F. E. Turkheimer, R. Hinz, and V. J. Cunningham. On the Undecidability Among Kinetic Models: From Model Selection to Model Averaging. *Journal of cerebral blood flow and metabolism : official journal of the International Society of Cerebral Blood Flow and Metabolism*, 23:490–498, 2003.

L. Wasserman. Bayesian model selection and model averaging. *Journal of mathematical psychology*, 44(1):92–107, 2000.

G. Winkler. *Image Analysis, Random Fields and Dynamic Monte Carlo Methods*. Springer Berlin Heidelberg, 1995.

F. Y. Wu. The Potts model. *Reviews of Modern Physics*, 54(1), 1982.

Y. Zhou. vSMC: Parallel Sequential Monte Carlo in C++. *Journal of Statistical Software*, 2015.

Y. Zhou, S. C. Huang, M. Bergsneider, and D. F. Wong. Improved parametric image generation using spatial-temporal analysis of dynamic PET studies. *NeuroImage*, 15(3):697–707, 2002.

Y. Zhou, J. A. Aston, and A. M. Johansen. Bayesian model comparison for compartmental models with applications in positron emission tomography. *Journal of Applied Statistics*, 40(5):993–1016, 2013.

Y. Zhou, A. M. Johansen, and J. A. Aston. Toward Automatic Model Comparison: An Adaptive Sequential Monte Carlo Approach. *Journal of Computational and Graphical Statistics*, 25(3): 701–726, 2016.

Appendix A

Derivation of the ESS Statistic

Following on from the argument of Section 2.5.2, using the asymptotic delta method approximation gives us that,

$$\begin{aligned}\mathrm{Var}_\nu[\widehat{I}^{(\mathrm{IS})}] &\approx \frac{1}{n}\left(\frac{\mathrm{Var}_\nu[HW]}{(\mathbb{E}_\nu W)^2} - 2\frac{\mathbb{E}_\nu HW}{(\mathbb{E}_\nu W)^3}\mathrm{Cov}(HW, W) + \frac{(\mathbb{E}_\nu HW)^2}{(\mathbb{E}_\nu W)^4}\mathrm{Var}_\nu[W]\right) \\ &= \frac{1}{n}\left(\underbrace{\mathrm{Var}_\nu[HW]}_{\text{(I)}} - 2\mathbb{E}_\mu[H]\underbrace{\mathrm{Cov}_\nu(HW, W)}_{\text{(II)}} + \mathbb{E}_\mu[H]^2\mathrm{Var}_\nu[W]\right) \qquad \text{via (2.10) and (2.11).}\end{aligned}$$

Next, using the definition of the co-variance, $\mathrm{Cov}(X, Y) = \mathbb{E}[XY] - (\mathbb{E}X)(\mathbb{E}Y)$, we have that firstly

$$\begin{aligned}\text{(II)} = \mathrm{Cov}_\nu(HW, W) &= \mathbb{E}_\nu[WH \cdot W] - (\mathbb{E}_\nu[HW])(\mathbb{E}_\nu W) \\ &= \mathbb{E}_\nu[HW^2] - \mathbb{E}_\mu[H] \qquad \text{via (2.10) and (2.11) again} \\ &= \mathbb{E}_\mu[HW] - \mathbb{E}_\mu[H] \\ &= \mathrm{Cov}_\mu(H, W) + \mathbb{E}_\mu[H]\mathbb{E}_\mu W - \mathbb{E}_\mu[H] \qquad \text{def of } \mathrm{Cov}_\mu(H, W).\end{aligned}$$

For (I), note that firstly

$$\mathrm{Var}_\nu[HW] = \mathbb{E}_\nu[H^2W^2] - (\mathbb{E}_\nu[WH])^2 = \mathbb{E}_\mu[H^2W] - \mathbb{E}_\mu[H]^2.$$

To make $\mathbb{E}_\mu(H^2W)$ more digestible, we use the second-order delta method approximation to give us,

$$\mathbb{E}_\mu(H^2W) \approx \mathbb{E}_\mu[W]\mathbb{E}_\mu[H]^2 + 2\mathbb{E}_\mu[H]\mathrm{Cov}_\mu(W, H) + \mathbb{E}_\mu[W]\mathrm{Var}_\mu[H] \qquad .$$

Finally, returning to the variance of the IS estimator, and substituting (I) and (II),

$$
\begin{aligned}
\mathrm{Var}_\nu[\widehat{I}^{(\mathrm{IS})}] &= \frac{1}{n}\left(\underbrace{\mathrm{Var}_\nu[HW]}_{\text{(I)}} - 2\mathbb{E}_\mu[H]\underbrace{\mathrm{Cov}_\nu(HW,W)}_{\text{(II)}} + \mathbb{E}_\mu[H]^2\mathrm{Var}_\nu[W]\right) \\
&\approx \frac{1}{n}\Big(\mathbb{E}_\mu[W]\mathbb{E}_\mu[H]^2 + 2\mathbb{E}_\mu[H]\mathrm{Cov}_\mu(W,H) + \mathbb{E}_\mu[W]\mathrm{Var}_\mu[H] - \mathbb{E}_\mu[H]^2 \\
&\qquad - 2\mathbb{E}_\mu[H]\Big\{\mathrm{Cov}_\mu(H,W) + \mathbb{E}_\mu[H]\mathbb{E}_\mu W - \mathbb{E}_\mu[H]\Big\} \\
&\qquad + \mathbb{E}_\mu[H]^2\mathrm{Var}_\nu[W]\Big) \\
&= \frac{1}{n}\Big(\mathbb{E}_\mu[H]^2\{1 + \mathrm{Var}_\nu[W] - \mathbb{E}_\mu W\} + \mathbb{E}_\mu[W]\mathrm{Var}_\mu[H]\Big) \\
&= \frac{1}{n}\Big(\mathbb{E}_\mu[H]^2\{1 + \mathrm{Var}_\nu(W) - \mathrm{Var}_\nu(W) - 1\} + \{\mathrm{Var}_\nu(W) + 1)\}\mathrm{Var}_\mu(H)\Big) \qquad \text{using (2.12) above} \\
&= \mathrm{Var}_\mu[\widehat{I}^{(\mathrm{MC})}](1 + \mathrm{Var}_\nu(W)) \qquad \text{using (2.13) above.}
\end{aligned}
$$

Appendix B

Compartmental Models Forms

B.1 One Tissue Compartmental Model

Figure B.1 below represents the one tissue compartmental model:

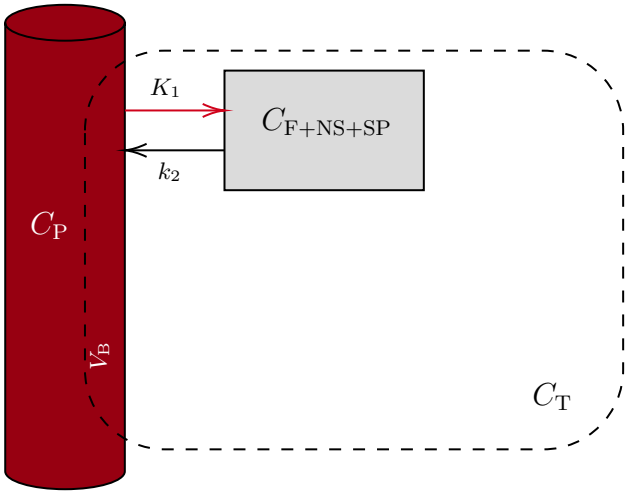


Figure B.1: A diagram of the 1–compartment model; Here, the flow between compartments C_i are represented by arrows and the constants K_i are the rate of flow.

The transition matrix and inflow vectors are thus:

$$A = [-k_2], \mathbf{b} = [K_2].$$

The IRF is then given by,

$$H(t) = \phi_1 e^{-\vartheta_1 t},$$

where,

$$\phi_1 = K_1,$$
$$\vartheta_1 = k_2.$$

From Theorem 2.2 of Gunn et al. (2001) it follows that

$$\begin{aligned} V_{\mathrm{D}} &= \frac{\phi_1}{\vartheta_1}, \\ &= \frac{K_1}{k_2}. \end{aligned}$$

B.2 Two Tissue Compartmental Model

Figure B.2below represents the one tissue compartmental model:

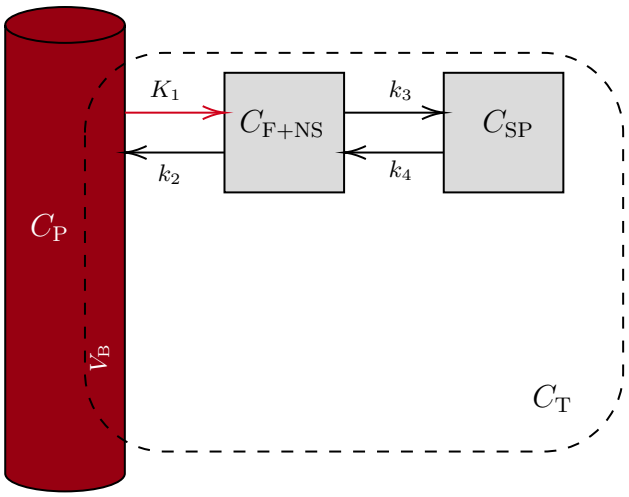


Figure B.2: A diagram of the 2–compartment model; Here, the flow between compartments C_i are represented by arrows and the constants K_i are the rate of flow.

The transition matrix and inflow vectors are thus:

$$A = \begin{bmatrix} -k_2 - k_3 & k_4 \\ k_3 & -k_4 \end{bmatrix}, \mathbf{b} = \begin{bmatrix} K_1 \\ 0 \end{bmatrix}.$$

The IRF is then given by,

$$H_{TP}(t) = \phi_1 e^{-\vartheta_1 t} + \phi_2 e^{-\vartheta_2 t},$$

where,

$$\begin{aligned} \phi_1 &= \frac{K_1(\vartheta_1 - k_3 - k_4)}{\Delta}, \\ \phi_2 &= \frac{K_1(\vartheta_2 - k_3 - k_4)}{-\Delta}, \\ \vartheta_1 &= \frac{k_2 + k_3 + k_4 + \Delta}{2}, \\ \vartheta_2 &= \frac{k_2 + k_3 + k_4 - \Delta}{2}, \\ \Delta &= \sqrt[+]{(k_2 + k_3 + k_4)^2 + 4k_2k_4}. \end{aligned}$$

From Theorem 2.2 of Gunn et al. (2001) it follows that

$$\begin{aligned} V_{\mathrm{D}} &= \frac{\phi_1}{\vartheta_1} + \frac{\phi_2}{\vartheta_2}, \\ &= \frac{K_1}{k_2}(1 + \frac{k_3}{k_4}). \end{aligned}$$

B.3 Three Tissue Compartmental Model

Figure 4.4 above represents the one tissue compartmental model. The transition matrix and inflow vectors are thus:

$$A = \begin{bmatrix} -k_2 - k_3 & k_4 & k_6 \\ k_3 & -k_4 & 0 \\ k_5 & 0 & -k_6 \end{bmatrix}, \mathbf{b} = \begin{bmatrix} K_1 \\ 0 \\ 0 \end{bmatrix}.$$

The IRF is then given by,

$$H_{TP}(t) = \phi_1 e^{-\vartheta_1 t} + \phi_2 e^{-\vartheta_2 t} + \phi_3 e^{-\vartheta_3 t},$$

where,

$$\begin{aligned} \phi_1 &= \frac{K_1(k_3(k_6 - \vartheta_1) + (k_4 - \vartheta_1)(k_5 + k_6 - \vartheta_1))}{(\vartheta_1 - \vartheta_2)(\vartheta_1 - \vartheta_3)}, \\ \phi_2 &= \frac{K_1(k_3(k_6 - \vartheta_2) + (k_4 - \vartheta_2)(k_5 + k_6 - \vartheta_2))}{(\vartheta_2 - \vartheta_1)(\vartheta_2 - \vartheta_3)}, \\ \phi_3 &= \frac{K_1(k_3(k_6 - \vartheta_3) + (k_4 - \vartheta_3)(k_5 + k_6 - \vartheta_3))}{(\vartheta_3 - \vartheta_1)(\vartheta_3 - \vartheta_2)}, \\ \vartheta_1 &= \frac{\Gamma_1}{3} - 2\sqrt{\Delta_1}\cos\left(\frac{\Upsilon}{3}\right), \\ \vartheta_2 &= \frac{\Gamma_1}{3} - 2\sqrt{\Delta_1}\cos\left(\frac{\Upsilon + 2\pi}{3}\right), \\ \vartheta_3 &= \frac{\Gamma_1}{3} - 2\sqrt{\Delta_1}\cos\left(\frac{\Upsilon + 4\pi}{3}\right), \\ \Upsilon &= \begin{cases} \cos^{-1}\left(\sqrt[+]{\frac{\Delta_2^2}{\Delta_1^3}}\right) & : \quad \Delta_2 < 0 \\ \cos^{-1}\left(\sqrt[-]{\frac{\Delta_2^2}{\Delta_1^3}}\right) & : \quad \Delta_2 > 0 \end{cases}, \\ \Delta_1 &= -\frac{1}{9}(3\Gamma_2 - \Gamma_1^2), \\ \Delta_2 &= \frac{1}{54}(2\Gamma_1^3 - 9\Gamma_2\Gamma_2 + 27\Gamma_3), \\ \Gamma_1 &= k_2 + k_3 + k_4 + k_5 + k_6, \\ \Gamma_2 &= k_2k_4 + k_2k_6 + k_3k_6 + k_4k_5 + k_4k_6, \\ \Gamma_3 &= k_2k_4k_6. \end{aligned}$$

From Theorem 2.2 of Gunn et al. (2001) it follows that

$$\begin{aligned} V_{\mathrm{D}} &= \frac{\phi_1}{\vartheta_1} + \frac{\phi_2}{\vartheta_2} + \frac{\phi_3}{\vartheta_3} \\ &= \frac{K_1}{k_2}(1 + \frac{k_3}{k_4} + \frac{k_5}{k_6}). \end{aligned}$$

Appendix C

PET Model Equations

The posterior distribution for the PET compartmental models with normally-distributed errors is as follows: Let

$$\iota_i(\phi, \vartheta) := \frac{C_{\mathrm{T}}(t_i; \phi, \vartheta)}{t_i - t_{i-1}},$$

the likelihood for data $y = (y_1, \ldots, y_k)$ can be written:

$$\prod_{i=1}^{k} \sqrt{\frac{\lambda}{2\pi\iota_i(\phi, \vartheta)}} \exp\left\{-\frac{\lambda}{2\iota_i(\phi, \vartheta)}(y_i - C_{\mathrm{T}}(t_i; \phi, \vartheta))^2\right\}.$$

The prior (joint) distributions over ϕ, ϑ and the precision parameter $\lambda = \frac{1}{\sigma_2}$ is given by:

$$\lambda^{\alpha-1} e^{-\beta\lambda} \prod_{j=1}^{m} I_{[\phi_j^a, \phi_j^b]} I_{[\vartheta_j^a, \vartheta_j^b]}$$

Here $\alpha = \beta = 10^{-3}$, the parameters of the prior distribution of λ. And ϕ_j^a and ϕ_j^a are the lower and upper bounds of the truncation interval of parameters ϕ_j and corresponding notation is used for θ_j.

For the $t-$distributed errors, the observation y_i has a t distribution with location $C_{\mathrm{T}}(t_i)$, scale $\frac{t_i - t_{i-1}}{C_{\mathrm{T}}(t_i)}\tau$, and degrees of freedom ν. The likelihood is,

$$\prod_{i=1}^{k} \left\{ \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma(\frac{\nu}{2})} \left(\frac{\tau}{\iota_i(\phi, \vartheta)\pi\nu}\right)^{\frac{1}{2}} \times \left(1 + \frac{\tau}{\nu\iota_i(\phi, \vartheta)}(y_i - C_{\mathrm{T}}(t_i; \phi, \vartheta))^2\right)^{-\frac{\nu+1}{2}} \right\}.$$

The prior density is given by:

$$\tau^{\alpha-1} e^{-\beta\tau} \times \frac{1}{\nu^2} \times I_{[a,b]}\left(\frac{1}{\nu}\right) \prod_{i=1}^{m} I_{[\phi_i^a, \phi_i^b]}(\phi_i) I_{[\vartheta_i^a, \vartheta_i^b]}(\vartheta_i),$$

where $\alpha = \beta = 10^{-3}$, the parameters of prior distribution of τ; $a = 0$ and $b = 0.5$.

Appendix D

MCMC Traces for Experiments and Long-run chains

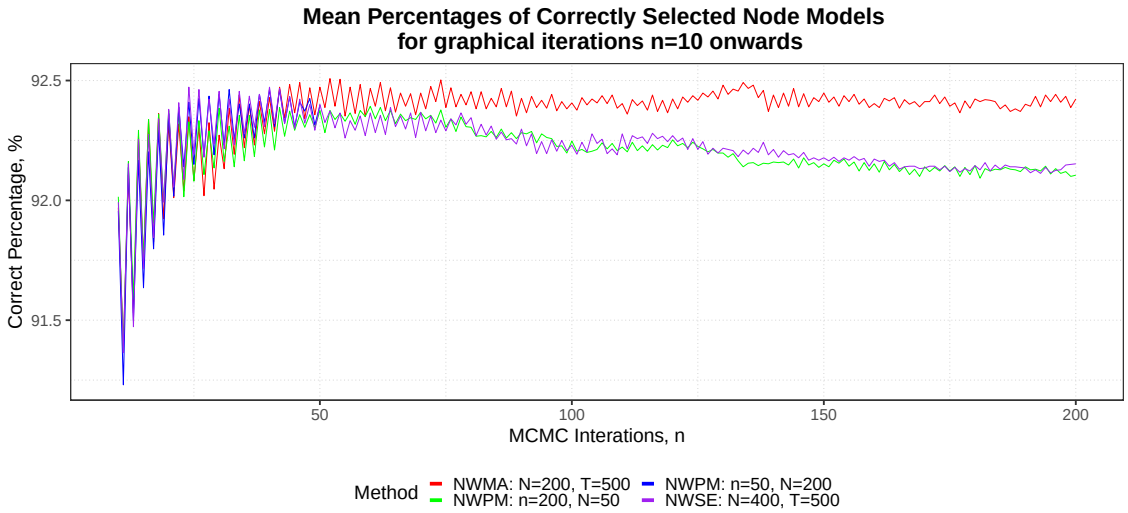


Figure D.1: Average percentages(%) of the whole toy model image ($20 \times 20 = 400$ pixels) where the correct model order was selected at each iteration of the MCMC chain. MCMC traces, of each method, for graphical iterations $n = 10$ on wards are shown.

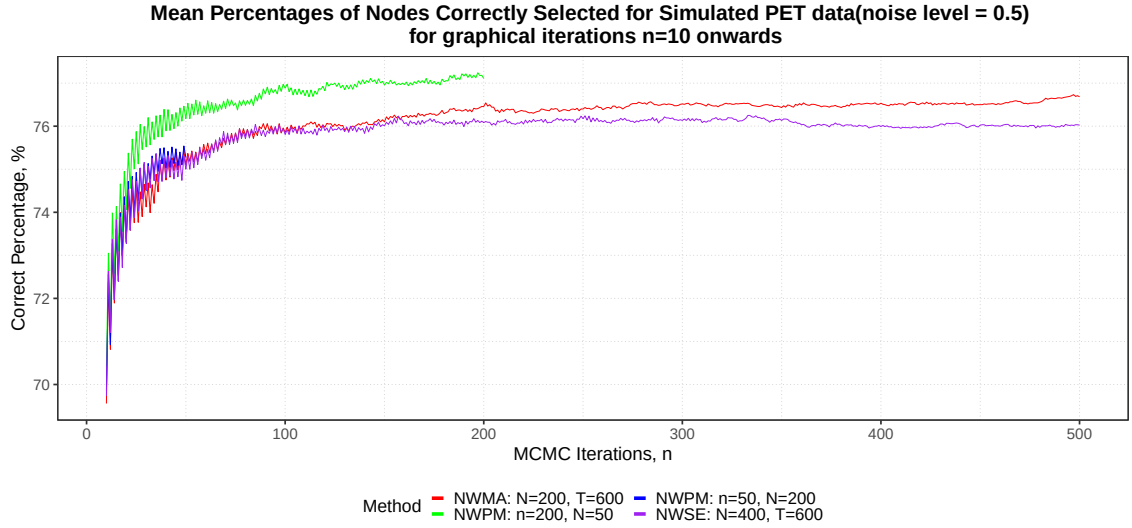


Figure D.2: average percentages(%) of the whole image ($20 \times 20 = 400$ pixels) where the correct model order was selected at each iteration of the pseudo-marginal MH chain. MCMC traces, of each method, for graphical iterations $n = 10$ and on wards are shown.

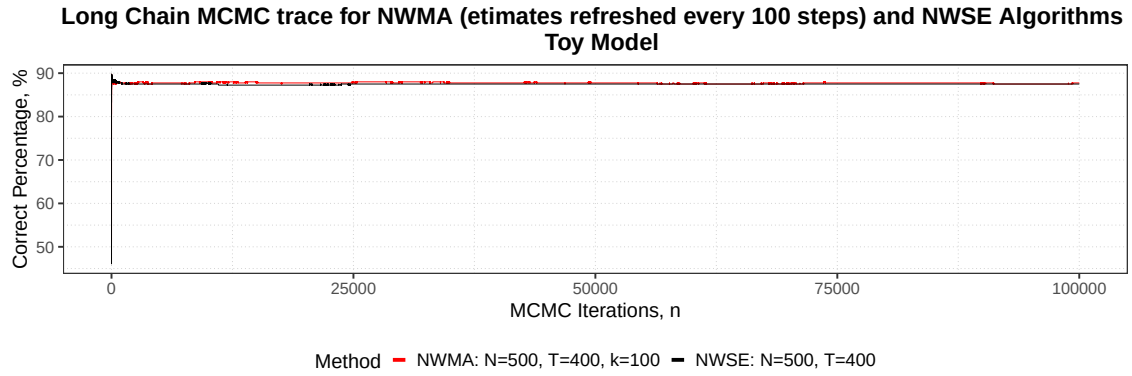


Figure D.3: MCMC traces of the percentage of correctly selected nodes, for the toy model. The NWSE and NWMA (refreshing marginal estimates every $\kappa = 100$ graphical iterations). The SMC sampler with $N = 400, T = 500$ was used as the marginal likelihood estimator for both methods.

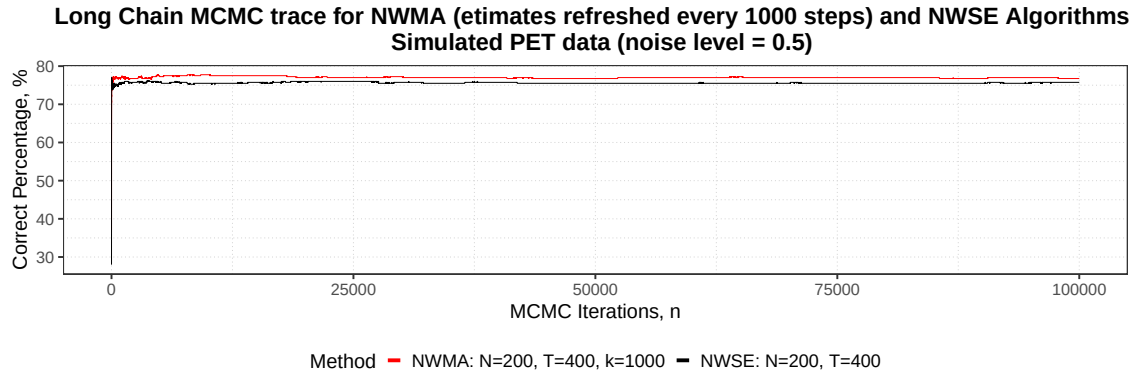


Figure D.4: MCMC traces of the percentage of correctly selected nodes, for simulated 2-D PET image, using the NWMA (refreshing marginal estimates every $\kappa = 1000$ graphical iterations) and NWSE methods. For both methods, the SMC sampler, with $N = 200, T = 400$, was used.

Appendix E

CESS-adaptive Annealing Scheme Study

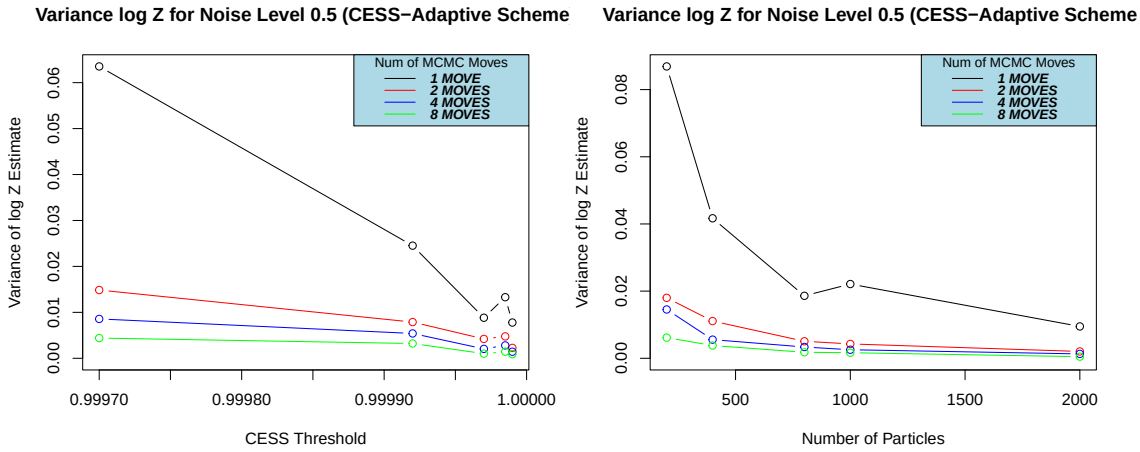


Figure E.1: Variance log of SMC normalising constant estimator for different number of particles and CESS thresholds (chosen to be roughly comparable to the design used for Prior 5 scheme), for noise level = 0.5.

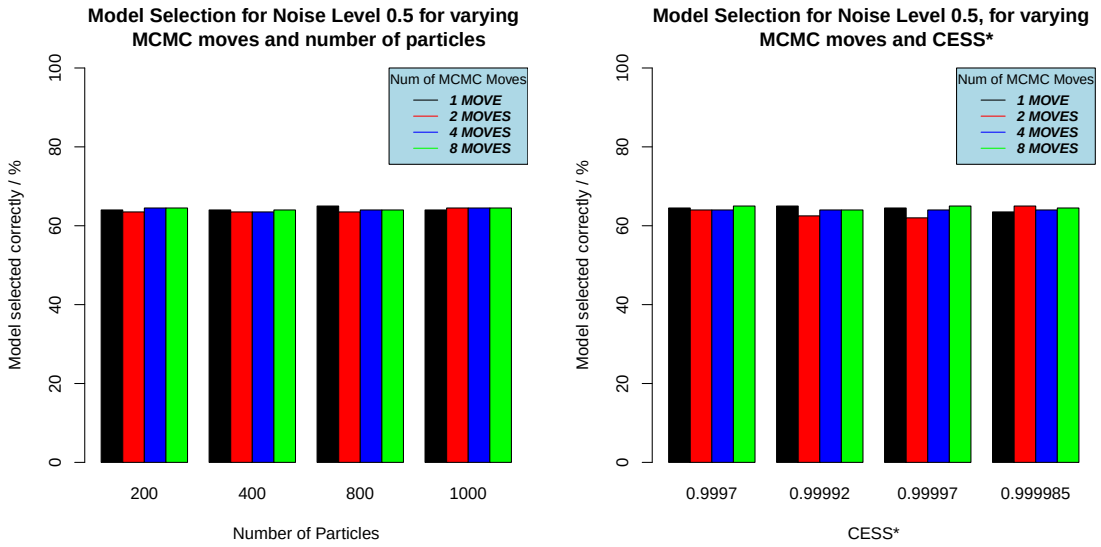


Figure E.2: Model selection for simulated PET data (with noise level 0.5) for varying number of MCMC moves and CESS*, using SMC sampler.

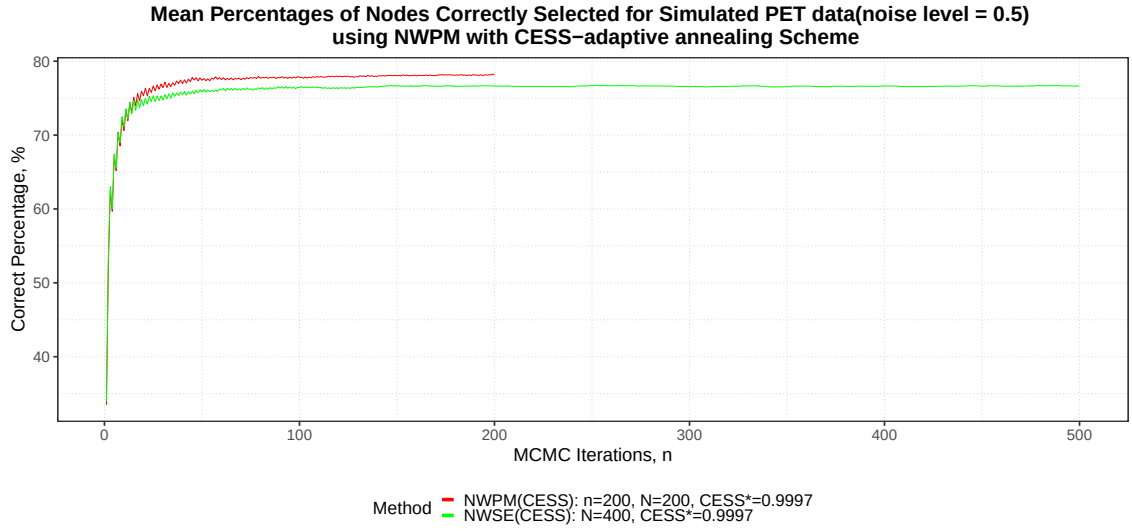


Figure E.3: Average percentages(%) of the whole image where the correct model order was selected at each iteration of the pseudo-marginal MH chain, using: NWPM($n = 200, N = 200$) and NWSE($n = 500, N = 400$). A CESS-adaptive annealing scheme SMC sampler with CESS$^* = 0.9997$ was used in both cases.

Appendix F

Volume of Distribution for Measured Images

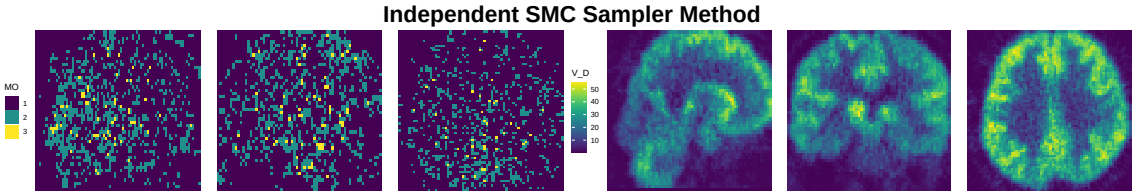


Figure F.1: Model order and volume of distribution parametric image of measured PET data, using spatially independent SMC sampler model selection.

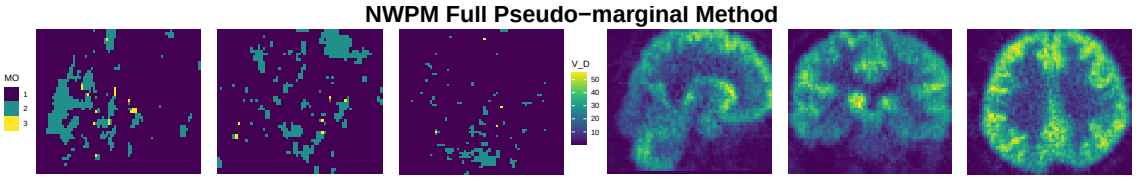


Figure F.2: Model order and volume of distribution parametric image of measured PET data, using NWPM method for model selection with spatial dependence.

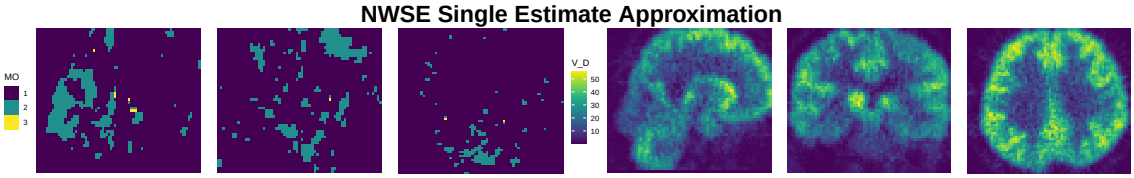


Figure F.3: Model order and volume of distribution parametric image of measured PET data, using NWSE approximation for model selection with spatial dependence.

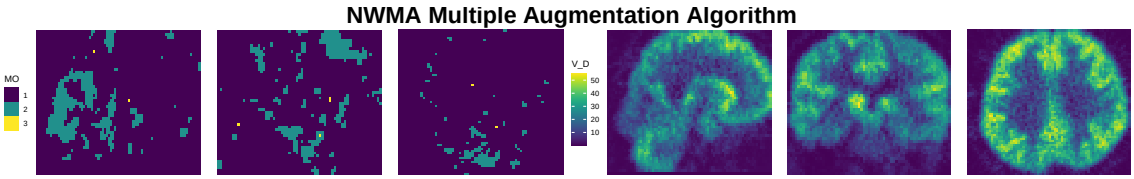


Figure F.4: Model order and volume of distribution parametric image of measured PET data, using the NWMA (multiple augmentation) variant algorithm.

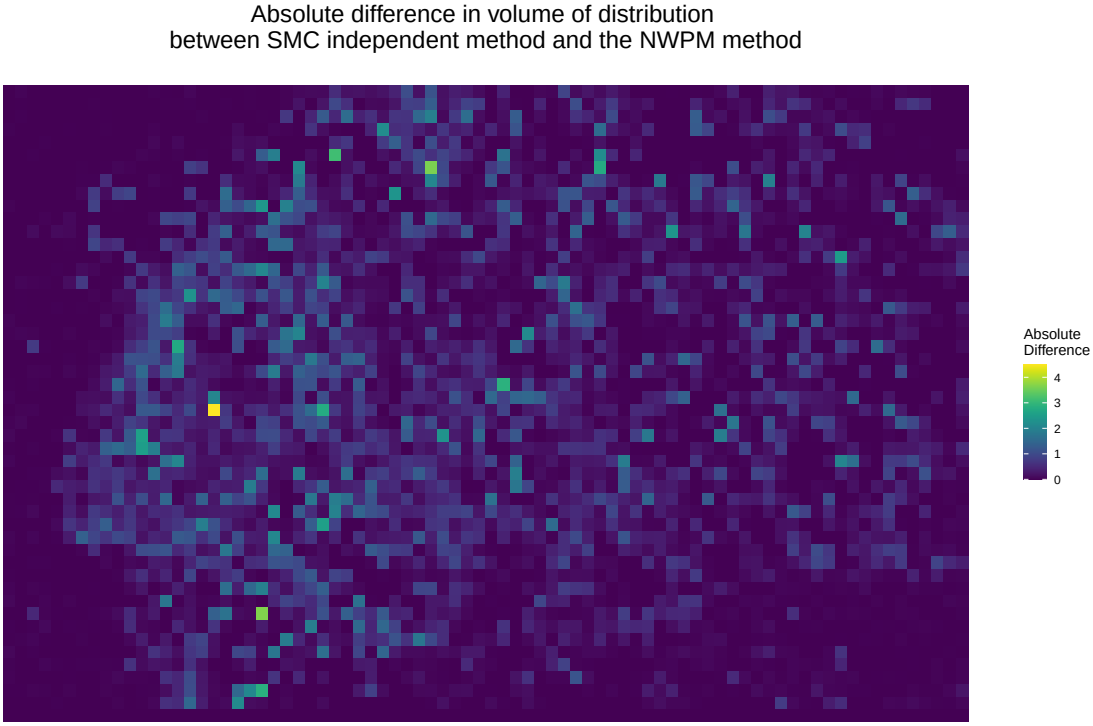


Figure F.5: Absolute difference in volume of distribution between the SMC independent method (Figure F.1) and NWPM method (Figure F.2), for the sagittal cross-section.