Advances in Bayesian Modelling and Computation: Spatio-temporal Processes, Model Assessment and Adaptive MCMC

by

Chunlin Ji

Department of Statistical Science
Duke University

Date: _____

Approved:

_____
Mike West, Supervisor

_____
Jerome Reiter

_____
Sayan Mukherjee

_____
Cliburn Chan

_____

Dissertation submitted in partial fulfillment of the
requirements for the degree of Doctor of Philosophy
in the Department of Statistical Science
in the Graduate School of
Duke University

2009

ABSTRACT

Advances in Bayesian Modelling and Computation: Spatio-temporal Processes, Model Assessment and Adaptive MCMC

by

Chunlin Ji

Department of Statistical Science
Duke University

Date: _____

Approved:

_____
Mike West, Supervisor

_____
Jerome Reiter

_____
Sayan Mukherjee

_____
Cliburn Chan

_____

An abstract of a dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Statistical Science in the Graduate School of Duke University

2009

Abstract

The modelling and analysis of complex stochastic systems with increasingly large data sets, state-spaces and parameters provides major stimulus to research in Bayesian nonparametric methods and Bayesian computation. This dissertation presents advances in both nonparametric modelling and statistical computation stimulated by challenging problems of analysis in complex spatio-temporal systems and core computational issues in model fitting and model assessment. The first part of the thesis, represented by chapters 2 to 4, concerns novel, nonparametric Bayesian mixture models for spatial point processes, with advances in modelling, computation and applications in biological contexts. Chapter 2 describes and develops models for spatial point processes in which the point outcomes are latent, where indirect observations related to the point outcomes are available, and in which the underlying spatial intensity functions are typically highly heterogenous. Spatial intensities of inhomogeneous Poisson processes are represented via flexible nonparametric Bayesian mixture models. Computational approaches are presented for this new class of spatial point process mixtures and extended to the context of unobserved point process outcomes. Two examples drawn from a central, motivating context, that of immunofluorescence histology analysis in biological studies generating high-resolution imaging data, demonstrate the modelling approach and computational methodology. Chapters 3 and 4 extend this framework to define a class of flexible Bayesian nonparametric models for inhomogeneous spatio-temporal point processes, adding dynamic models for underlying intensity patterns. Dependent Dirichlet process mixture models are introduced as core components of this new time-varying spatial model. Utilizing such nonparametric mixture models for the spatial process intensity functions allows the introduction of time varia-

tion via dynamic, state-space models for parameters characterizing the intensities. Bayesian inference and model-fitting is addressed via novel particle filtering ideas and methods. Illustrative simulation examples include studies in problems of extended target tracking and substantive data analysis in cell fluorescent microscopic imaging tracking problems.

The second part of the thesis, consisting of chapters 5 and chapter 6, concerns advances in computational methods for some core and generic Bayesian inferential problems. Chapter 5 develops a novel approach to estimation of upper and lower bounds for marginal likelihoods in Bayesian modelling using refinements of existing variational methods. Traditional variational approaches only provide lower bound estimation; this new lower/upper bound analysis is able to provide accurate and tight bounds in many problems, so facilitates more reliable computation for Bayesian model comparison while also providing a way to assess adequacy of variational densities as approximations to exact, intractable posteriors. The advances also include demonstration of the significant improvements that may be achieved in marginal likelihood estimation by marginalizing some parameters in the model. A distinct contribution to Bayesian computation is covered in Chapter 6. This concerns a generic framework for designing adaptive MCMC algorithms, emphasizing the adaptive Metropolized independence sampler and an effective adaptation strategy using a family of mixture distribution proposals. This work is coupled with development of a novel adaptive approach to computation in nonparametric modelling with large data sets; here a sequential learning approach is defined that iteratively utilizes smaller data subsets. Under the general framework of importance sampling based marginal likelihood computation, the proposed adaptive Monte Carlo method and sequential learning approach can facilitate improved accuracy in marginal likelihood computation. The approaches are exemplified in studies of

both synthetic data examples, and in a real data analysis arising in astro-statistics.

Finally, chapter 7 summarizes the dissertation and discusses possible extensions of the specific modelling and computational innovations, as well as potential future work.

Contents

List of Figures

List of Tables

Acknowledgements

The completion of the work for this dissertation represents the opportunity to remember numerous individuals who have influenced my attitude toward PhD study. I would like to acknowledge my debt of gratitude to my supervisors, colleagues, friends and family.

First and foremost, I would like to express my profound gratitude to my supervisor, Professor Mike West, for his kind guidance, support and encouragement throughout my PhD study. I will always remember his help and endless patience. Thanks also to Professor Simon Godsill, for providing kind concern on my research during the SAMSI SMC program. Special thanks are also due to Professor Thomas Kepler, Professor David Dunson, Professor Merlise Clyde, Professor Scott C Schmidler and Professor Cliburn Chan, for their useful discussions in various collaborations.

I would also like to thank several colleagues in the our Statistical Science department for their support and valuable discussions. I thank Dr. Dan Merl, Dr. Ioanna Manolopoulou, Dr. Haige Shen, Dr. Julien Cornebise, Quanli Wang, Hao Wang, Hongxia Yang, and Xiaojing Wang. I want to make special mention of Dr. Ioanna Manolopoulou, Avishek Chakraborty and Scott Schwartz for their kind proof reading.

I am especially grateful to my friend Dr. Yangyang Zhang, Ruopeng Liu and Dr. Bin Liu who have also had an important influence on the development of this dissertation.

I acknowledge supports of the NSF (grant DMS-0342172) and NIH (grant P50-GM081883 and contract HHSN268200500019C). Any opinions, findings and conclusions or recommendations expressed in this work are those of the author and do not necessarily reflect the views of the NSF or NIH.

Finally, but certainly not the least, I would like to acknowledge the support of family. I would like to thank my wife and my parents. Their moral support of my ambition to earn a PhD degree was a major factor in my ability to do so. I will always be grateful for their incredible generosity.

Chapter 1

Introduction

Traditionally, statistical methods for spatial point process assume perfect knowledge of the outcome of the point process. However, in certain situations, the outcome of the spatial point process may not be observed directly, but is measured by some imperfect proxy. For example, in immunological studies of fluorescent intensity images of lymphatic tissue, the observed measurements are fluorescent intensities generated from tagged cell-surface proteins, which indirectly generates noisy observation of cell locations for as many as tens of thousands of cells in the context of background noise. In such situations, inference on the underlying intensity function will depend on accounting for the uncertainty surrounding the outcome of the point process, including the number of realized points as well as their spatial locations. The first part of this dissertation discusses models that address these issues and develops computational Bayesian methods for model fitting and analysis. A flexible Dirichlet process (DP) normal mixture model is used to characterize the highly heterogenous spatial intensity function. Computational approaches are presented for spatial DP mixture and extended to the context of unobserved outcomes. Two examples of immunofluorescence histology data analysis demonstrate the models and computational methodology.

Beyond static spatial point processes, dynamic (or time-varying) spatial point processes have recently gained increasing attention for describing various application problems in areas such as multi-target tracking and cell fluorescent microscopic imaging tracking. In these situations, the spatial point pattern may change dramatically over time, thus it is not straightforward to apply the previous mod-

elling approach. While traditional Dirichlet Process models focus on problems with exchangeable samples from *one* unknown distribution, there is growing interest in extending the Dirichlet Process to accommodate multiple dependent distributions, and this provides an opportunity to extend the static point process models to the dynamic setting. In chapter 3, dependent Dirichlet process mixture models are introduced for complex dynamic systems. By following the previous work on dependent DP mixture modelling (MacEachern, 1999, 2001; Caron et al., 2007), countably infinite mixtures of Gaussian distributions are introduced to represent the unknown density at each time point, while time dependencies are introduced by dynamic linear models for underlying parameters. These methods can be conceived as extensions of the Dirichlet process mixture model (Escobar and West, 1995) to collections of distributions evolving in discrete time. Since dependence is built into the mixing distribution by allowing parameters to evolve dynamically via state-space models, these models can also be regarded as extensions of the Gaussian Dynamic Linear Models (DLMs) (West and Harrison, 1997). For sequential Bayesian inference on these dynamic models, I propose a novel sequential Monte Carlo method, the Rao-Blackwellized particle filter (RBPF). In RBPF, we apply sequential Monte Carlo on the underlying allocation variable for observed data points. Given these allocation variables, we use dynamic models for the parameters of each mixture component, enabling us to process closed form updates for these parameters. Using a simulation study of distribution autoregressive models, we demonstrate the effectiveness of the proposed approach in accommodating multiple dependent distributions.

Chapter 4 builds on the innovations in Chapter 3 to define a flexible Bayesian nonparametric modelling for inhomogeneous spatio-temporal processes. This involves nonparametric spatial process mixture models of intensity functions in which time variation is introduced via dynamic models for underlying parameters. These

models characterize smooth dynamics in time in what may be quite complicated spatial patterns of spatial inhomogeneity in intensity functions. The framework is based on a new time-varying Dirichlet process partition scheme, and physically attractive time propagation models for parameters of nonparametric mixture models for intensities. Bayesian inference and model fitting is addressed, involving novel particle filtering ideas and methods. Illustrative simulation examples in extended target tracking, and substantive data analysis in applications in cell fluorescent microscopic imaging tracking demonstrate analysis with these models.

This dissertation also address some generic computation methods in Bayesian inference, particularly for large data set. Chapter 5 discuss a novel approach of marginal likelihood approximation. As is well known, marginal likelihood is the essential quantity in Bayesian model selection, representing the evidence of a model. However, evaluating marginal likelihoods often involves intractable integration and needs to rely on numerical integration and approximation. Mean-field variational methods has been extensively studied by machine learning and Bayesian learning communities for deterministic approximation of marginal distributions (MacKay, 1995; Jordan et al., 1999; Beal and Ghahramani, 2003; Beal, 2003). Apparently, performing model selection merely based on the lower bounds of log marginal likelihoods can be inappropriate as the approximation error is not quantitatively limited. We provide an upper as well as lower bound for the log marginal likelihood and propose a method based on posterior samples to minimize the upper bound. We also show a quasi-lower bound can be obtained with trivial computation based on the result of optimal upper bound. We demonstrate that by marginalizing some parameters in the model, we can significantly reduce the "discrepancy" between the bounds of log marginal likelihood. However, when some parameters are marginalized, traditional variational method are not feasible. To address this, we present a

method that directly uses a Monte Carlo Stochastic Approximation (MCSA) algorithm to maximize the lower bound, and prove the convergence to the true local maximum lower bound under commonly applicable assumptions.

Chapter 6 consists of three sub-topics which are motivated by posterior sampling for complex distributions and Bayesian inference for large data set problems: 1) adaptive Markov chains have seen renewed interest in recent years due in part to the emergence of certain theoretical guarantees (Haario et al., 2001; Roberts and Rosenthal, 2007). With adaptive MCMC algorithms, the entire sample history of process is used to tune parameters of the proposal density during simulation in order to obtain faster convergence or more efficient estimation. I first present a generic framework to design adaptive MCMC algorithms, emphasizing the adaptive Metropolized independence sampler and effective adaptation strategy using a family of mixture distribution proposals. 2) Motivated by the need for flexible proposal forms in adaptive Monte Carlo methods and effective approach to fitting nonparametric models for large data set, a sequential learning approach for DP mixture model is proposed. This method utilizes only a small subset of the whole data set to update the associated parameters in the mixture distribution iteratively, and gradually approach the optimal DP mixture which minimizing the KL-divergence between the unknown target distribution which generates the data set and the DP mixture distribution. 3) Motivated by the need for effective marginal likelihood computation in complicated Bayesian models (e.g. the velocity-shift model in Bayesian exoplanet searches (Crooks et al., 2007; Bullard, 2009)), the proposed adaptive Monte Carlo method and sequential learning approach are incorporated in the framework of importance sampling, based marginal likelihood computation. More specifically, the adaptive MCMC method is used to draw samples from the target distribution while a truncated DP (TDP) mixture model is tuned by the pro-

posed sequential learning approach utilizing these samples. The well tuned TDP mixture model serves as the importance function for marginal likelihood computation. Both synthetic example and real world application in Bayesian exoplanet searches are presented to demonstrate the performance of proposal methods.

A summary of the dissertation, possible extensions and future work are discussed in Chapter 7.

Chapter 2

Spatial Mixture Modelling for Unobserved Point Processes

We discuss Bayesian modelling and computational methods in analysis of indirectly observed spatial point processes. The context involves noisy measurements on an underlying point process that provide indirect and noisy data on locations of point outcomes. We are interested in problems in which the spatial intensity function may be highly heterogenous, and so is modelled via flexible nonparametric Bayesian mixture models. Analysis aims to estimate the underlying intensity function and the abundance of realized but unobserved points. Our motivating applications involve immunological studies of multiple fluorescent intensity images in sections of lymphatic tissue where the point processes represent geographical configurations of cells. We are interested in estimating intensity functions and cell abundance for each of a series of such data sets to facilitate comparisons of outcomes at different times and with respect to differing experimental conditions. The analysis is heavily computational, utilizing recently introduced MCMC approaches for spatial point process mixtures and extending them to the broader new context here of unobserved outcomes. Further, our example applications are problems in which the individual objects of interest are not simply points, but rather small groups of pixels; this implies a need to work at an aggregate pixel region level and we develop the resulting novel methodology for this. Two examples with immunofluorescence histology data demonstrate the models and computational methodology.

2.1 Introduction

Parametric and nonparametric approaches to spatial point process modelling have been well-studied in recent years (Diggle, 2003; Moller and Waagepetersen, 2004), with increased use of mixtures and convolutional methods for modelling heterogeneity in intensity functions (Wolpert and Ickstadt, 1998). Recently, Kottas and Sanso (2007) proposed the use of the Dirichlet process as a random mixing distribution for mixture-based methods. The full computational machinery of nonparametric Bayesian models has thus been brought to bear on this class of inference problems for point processes. Traditionally, all such methods assume perfect knowledge of the outcome of the point process. However, in situations such as that described below, the outcome of the spatial point process cannot be observed directly but is measured by some imperfect proxy. In such situations, inference on the underlying intensity function will depend on accounting for the uncertainty surrounding the outcome of the point process, the latter including the number of realized points as well as their spatial locations. Our work here defines models that address these issues and develops computational Bayesian methods for model fitting and analysis.

Our motivating applications are immunological studies of multiple fluorescent intensity images of lymphatic tissue. Observed measurements are fluorescent intensities generated from tagged cell-surface proteins; this generates indirect, noisy observation of cell locations for as many as tens of thousands of cells in the context of background noise. The spatial configurations of cells across the 2 or 3-d tissue region is typically hugely heterogenous, so requiring flexible models for underlying intensities. In any one experiment (of many) a series of images may reflect cellular distributions at different times and/or as a response to different interventions and treatments. For each, we aim to characterize the underlying intensity functions

and overall level of abundance of cell types in order to facilitate comparisons across multiple images.

2.2 Latent Spatial Mixture Models

Our general statistical framework jointly models the intensity function of a spatially inhomogeneous Poisson process and the uncertain outcome of the point process. Modelling of the intensity function is similar to that of Kottas and Sanso (2007), but here relying on a Dirichlet process mixture model of multivariate normal densities (rather than beta densities). Incidences of the point process are modelled via a modification of the basic model to represent data on a pixelated grid across image space, and this couples with a generalized linear model for linking noisy measurements (e.g. fluorescence levels, available at the gridded level) to incidences of the point process (e.g. presence of cells).

2.2.1 Basic Spatial Point Process Model

A spatial point process over a finite region $S \subset \mathbb{R}^d$ (here, $d = 2$) generates realizations $x_{1:N} = \{x_1, \ldots, x_N\}$ of $N \geq 0$ points $x_i \in S$. We regard $x_{1:N}$ as the outcome of an inhomogeneous Poisson process with intensity function $\lambda(x) \geq 0$ $(x \in S)$, integrable over S. That is: (a) for any region $s \subseteq S$, the number of points $n(s) = \#\{i = 1 : N | \, x_i \in s\}$ is Poisson with mean $\Lambda(s) = \int_{x \in s} \lambda(x) dx$; and (b) conditional on $\lambda(\cdot)$, $n(s) \perp\!\!\!\perp n(r)$ for any disjoint subsets $s, r \subset S$ (Daley and Vere-Jones, 2003; Diggle, 2003).

Bayesian analysis of *observed* data $x_{1:N}$ arising from a spatial inhomogeneous Poisson process requires first specifying a prior probability model for the intensity function $\lambda(\cdot)$, and then conducting posterior inference on $\lambda(\cdot)$ in light of the realized outcomes $x_{1:N}$. As in Kottas and Sanso (2007), define the overall intensity scale

parameter $\gamma = \int_{x \in S} \lambda(x) dx$ and the probability density (over $x \in S$) $f(x) = \lambda(x)/\gamma$. Then the likelihood function resulting from observed data $x_{1:N}$ can be expressed as

$$p(x_{1:N}|\gamma, f) \propto \exp(-\gamma)\gamma^N \prod_{i=1}^{N} f(x_i) \tag{2.1}$$

as a function of (γ, f). The degree to which underlying spatial heterogeneity can be represented in $\lambda(\cdot)$ is therefore linked to the modelling assumptions surrounding $f(\cdot)$.

2.2.2 Dirichlet Process Mixture Models

To provide flexibility in characterizing spatial heterogeneity in the intensity function we employ the Dirichlet process mixture framework in which the normalized intensity function $f(x)$ is the density of a random mixture of $d-$dimensional normal distributions. This follows Kottas and Sanso (2007) who develop models using mixtures of betas rather than normals. Since we are working on problems with very heterogeneous intensity functions in 2 and 3-d, and with sample sizes N that (though unknown) are large, we very much need the flexibility offered by mixtures of multivariate normals coupled with their relative analytic and computational benefits; we simply truncate and ignore the form of fitted and simulated models outside the finite region S.

A key observation of Kottas and Sanso (2007) was to note that the likelihood function of equation (2.1) depends on $f(\cdot)$ only through the term $\prod_{i=1}^{N} f(x_i)$ and is *precisely* the likelihood that would arise from simple random sampling from $f(\cdot)$ generating data $x_{1:N}$. Thus, for computational purposes, we can then use the standard methods of posterior computation based on any assumed model for $f(\cdot)$. Use of Dirichlet process mixtures is one example.

In brief, $f(\cdot)$ is taken as the density of a distribution arising from the following hierarchical model for independent, $d-$dimensional variates x_i, each with its own parameter $\theta_i = (\mu_i, \Sigma_i)$, a mean vector and variance matrix, respectively. Then the model for $f(\cdot)$ is

$$(x_i|\theta_i) \sim N(x_i|\mu_i, \Sigma_i), \quad (\theta_i|G) \sim G, \quad (G|\alpha, G_0) \sim DP(\alpha, G_0) \tag{2.2}$$

using standard notation. Here $G(\cdot)$ is an uncertain distribution function, $G_0(\cdot)$ is the prior mean of $G(\cdot)$ and $\alpha > 0$ the total mass, or precision of the DP. For conditional conjugacy, it is convenient and common to take the prior as normal-inverse Wishart. The implied distribution corresponding to the density $f(x)$ is a discrete mixture of a countably infinite number of normals. The model notation and structural details are standard and used widely in applied Bayesian inference; key foundational modelling and computational aspects are available in, for example, MacEachern (1994), West et al. (1994), Escobar and West (1995, 1998), MacEachern (1998) and MacEachern and Mueller (1998) and in a broader context in the more recent review paper of Mueller and Quintana (2004). Details that are key to computation are noted below in Section 2.3 and the Appendix A.

2.2.3 Discrete Pixel Region Model

In the immunological application as in other studies in spatial modelling, the data arises in terms of images of the region S within which the individual objects of interest are not simply points, but rather small groups of pixels. This, coupled with the fact that the objects (here, cells) are in any case not directly observed, implies a need to work at an aggregate pixel region level. This can be quite general but, for purposes here, we focus on rectangular pixel regions; in the immunological imaging study, for example, we work at the level of 3×3 pixel regions (in $2-d$) and each region is either occupied by a cell, or not.

Generally, in $d-$dimensions suppose the overall imaged region $S = [-s, s]^d$, for some $s > 0$, and that the level of resolution is a pixels in each dimension. S is then a rectangular grid of a^d pixel regions; label these by interior points x_i, $(i = 1 : a^d)$, and set $\mathcal{X} = \{x_i : \ i = 1 : a^d\}$. Assuming a to be large and with $\Delta = (2s/a)^d$, we have approximate intensity $\Delta\gamma f(x_i)$ for pixel region i and $\sum_{x\in\mathcal{X}} \Delta f(x) \approx \int_{x\in S} f(x)dx = 1$.

Now, for any realization of the point process, each pixel region will be either occupied by an object or not. Define binary variates $y(x) = 1/0$ to represent presence/absence of an object (e.g., a cell) in the pixel region with index point $x \in \mathcal{X}$. Then observing the occurrence of objects at a subset of N regions is equivalent to observing binary data $y(x)$ for *all* $x \in \mathcal{X}$ with $y(x) = 1$ at just the N regions with objects. Suppose the N regions are indexed by $x_{1:N} \in \mathcal{X}$, and write Y for the full set of a^d binaries. It then follows that the likelihood of equation (2.1) is equivalent to

$$p(Y|\gamma, f) \propto \prod_{x\in\mathcal{X}} \{\Delta\gamma f(x)\}^{y(x)} \exp\{-\gamma\Delta f(x)\} \tag{2.3}$$

and

$$p(Y|\gamma, f) \propto \exp(-\gamma)\Delta^N\gamma^N \prod_{i=1}^{N} f(x_i). \tag{2.4}$$

This provides the ability to work at the discretized, pixel region level appropriately, and simply modifies the likelihood with the additional Δ term. Most importantly also, this discrete pixel region version also enables easy development for contexts in which the locations $x_{1:N}$ are not observed directly but are measured with noise, since equation (2.3) delivers a likelihood function for uncertain locations and number of objects $x_{1:N}$ along with (γ, f).

2.2.4 Unobserved Spatial Inhomogeneous Poisson Process

Consider now contexts in which the locations x_i, and their number N, are uncertain. The example of fluorescent intensity images of lymphatic tissue is a key motivation and raises broader modelling questions. There, the specific locations of biological cells are not observed, but reflected in terms of fluorescence generated from labelled cell surface proteins. Under the discrete pixel region formulation, we can incorporate uncertainty about $x_{1:N}$ using equation (2.3), as follows.

Suppose we have observations $z(x)$ at each location $x \in \mathcal{X}$ generated by the measurement process. That is, the measurements represent single pixel region locations with no overlap or interaction. It is practicable to assume that the measurement error distribution depends on x only through presence or absence of objects, i.e., on the $y(x)$ binary indicators, and will usually involve uncertain parameters here denoted by δ. That is, a measurement error model is defined by two density functions $p(z|x,\delta) = p(z|y(x),\delta)$ for $y(x) = 1/0$, where $p(z|y=0,\delta)$ represents background noise in the absence of an object at a specific location, and $p(z|y=1,\delta)$ represents noise in the presence of a signal.

We can now combine $p(Y|\gamma, f)$ of equation (2.3) as the prior for all $y(x)$ with the implied likelihood components $p(z(x)|y(x),\delta)$, based on recorded data $Z = \{z(x) : \ x \in \mathcal{X}\}$. In terms of posterior odds on $y(x) = 1$ versus $y(x) = 0$, this yields conditionally independent posteriors with

$$Odds(y(x) = 1|Z, \gamma, f) = r(z(x))\gamma\Delta f(x) \tag{2.5}$$

where $r(z(x)) = p(z(x)|y(x) = 1,\delta)/p(z(x)|y(x) = 0,\delta)$ for all $x \in \mathcal{X}$.

In the immunological imaging study, appropriate noise models are truncated normals and the posterior odds ratios are trivially evaluated. This is important as we can then embed imputation of Y in the overall MCMC computations.

2.3 Posterior Inference And Sampling Strategies

The overall posterior inference goals are to explore and summarize aspects of the implied joint posterior for all uncertain quantities based on a complete model specification that now includes independent priors on γ, α and any of the hyperparameters δ we may wish to treat as uncertain. In summary, this is the posterior $p(Y, f, \gamma, \alpha, \delta \mid Z)$.

2.3.1 Overall MCMC Framework

The MCMC computational algorithm visits the following components in turn. Each of the imputation steps here draws new variates from the conditional distribution given all other conditioning quantities. In each, only those conditioning quantities that matter are included in the notation.

Sampling the normalized intensity function $f(x)$, its parameters and α

Each MCMC iterate generates a realized density that is a mixture of a finite number of $d-$dimensional normals, $f(x) \equiv f(x|\Theta) = \sum_{j=1}^{k} w_j N(x|\mu_j^*, \Sigma_j^*)$, with parameters $\Theta = \{w_{1:k}, \mu_{1:k}^*, \Sigma_{1:k}^*\}$ changing at each MCMC step, being generated via a two-step process discussed in Section 2.3.2 and the Appendix A. This step also resamples the Dirichlet precision α. Given Θ, the density $f(\cdot|\Theta)$ can be evaluated at the finite set of points $x \in \mathcal{X}$ for further use.

At each iterate, the Θ parameters are drawn from an implicit conditional posterior $p(\Theta|N, x_{1:N})$ where the number of imaged objects, N, and their location indices $x_{1:N}$, are set at current values. As the MCMC progresses these values are resampled as the analysis explores the joint posterior that now also includes uncertainty about $(N, x_{1:N})$.

Sampling spatial object location indicators Y

Equation (2.5) leads to resampling of new values of Y as independent binaries $y(x)$ at each $x \in \mathcal{X}$, based on implied probabilities $Pr(y(x) = 1|z(x), \gamma, f, \delta)$. This generates a complete set of binaries from which those values $y(x) = 1$ identify the new sample size N and pixel region locations $x_{1:N}$. Notice that one by-product is samples from the posterior for N, i.e., the ability to make inferences about the uncertain number of underlying objects as well as their locations. This step explicitly requires the evaluation $f(x) \equiv f(x|\Theta)$, the mixture of normals based on the most recently sampled Θ.

Sampling the overall scale of intensity γ

The form of equation (2.4) makes it clear that a gamma prior is conjugate to the conditional likelihood, leading to a gamma distribution $p(\gamma|N)$.

Sampling hyperparameters δ of the measurement error model

Under a prior $p(\delta)$, these parameters may be generated using some form of Gibbs or Metropolis-Hastings component strategy based on the implied conditional

$$p(\delta|Z, Y) \propto p(\delta) \prod_{x \in \mathcal{X}} p(z(x)|y(x), \delta).$$

For example, normal measurement errors would involve normal mean and variance parameters in δ, one pair for each of the signal and noise error models; in such a case, conditionally conjugate priors would aid in this computational step. In our immunological studies appropriate error models are truncated normals, which introduces a need for Metropolis-Hastings for sampling δ as described in section 2.4.2 and the examples.

2.3.2 Simulation in DP Mixtures

Under the DP model of equation (2.2), G is discrete. This results in any realized set of N parameters $\theta_{1:N} = (\mu_{1:N}, \Sigma_{1:N})$ being configured into some $k \leq N$ *distinct* values $(\mu^*_{1:k}, \Sigma^*_{1:k})$. The DP generates configuration indicators $c_{1:N}$ such that $c_i = j$ indicates $(\mu_i, \Sigma_i) = (\mu^*_j, \Sigma^*_j)$. The original MCMC approaches to generating posterior samples in DP mixtures (MacEachern, 1994; West et al., 1994; Escobar and West, 1995, 1998; MacEachern, 1998; MacEachern and Mueller, 1998) utilize this theory to generate samples from the full joint posterior of k, $(\mu^*_{1:k}, \Sigma^*_{1:k})$ and $c_{1:N}$. Most effective among these approaches are the *collapsed* or *configuration* samplers for DP mixture models originating from MacEachern (1994). More recent approaches are based on the innovative strategy using the *blocked Gibbs sampler* (Ishwaran and James, 2001) that explicitly includes simulation from approximations to the conditional posteriors for the underlying mixing distribution $G(\cdot)$ itself.

In many problems with small or moderate sample sizes N, and when $f(\cdot)$ is well-behaved to the extent that it may be well-approximated by a small mixture of normals, there is little to choose between the configuration and blocked samplers in terms of either computational or statistical efficiencies. However, as N increases, and also with densities $f(\cdot)$ of greater complexity that therefore require larger numbers k of mixture components for adequate representation, the blocked sampler dominates. Configuration sampling iteratively resamples each configuration indicator conditional on the rest; this one-at-a-time update degrades computational efficiency as N increase, and difficulties in moving in configuration space induced by the tight conditioning degrade mixing of the MCMC, and hence statistical efficiency. In contrast, the blocked sampling strategy breaks this configuration conditioning at each iterate, resampling the full set of configuration indicators jointly.

In our immunological applications, N is in the thousands or tens of thousands,

and intensity surfaces can be very heterogeneous, so the blocked sampling strategy is really demanded for efficiency reasons. In fact, the approach is almost mandated in the context of measurement error; as we have seen, values of the normalized intensity $f(x)$ itself are key components of the overall analysis, arising in the conditional posteriors for the latent spatial object location indicators Y in equation (2.5). To evaluate values of the density $f(\cdot)$ requires inference on the underlying mixing distribution $G(\cdot)$ itself, and this is provided by the blocked sampling strategy. Kottas and Sanso (2007) use this strategy, pointing out that it is needed to generate posterior inferences on aspects of $f(\cdot)$ in any case; our new framework with latent spatial process outcomes, large N and heterogeneous intensity patterns very strongly reinforces this choice.

The block sampler involves three linked steps: sampling of the set of configuration indicators $c_{1:N}$, sampling of parameters that define an approximation to the mixing distribution $G(\cdot)$, and sampling of sets of normal model means and variance matrices. A key element is the truncated approximation to the so-called stick-breaking representation of G (Sethuraman, 1994) that effectively defines a finite mixture model with a specified upper bound k on the number of components (Ishwaran and James, 2001). Importantly, then, this approach actually introduces a theoretical approximation to the full DP mixture model through this truncation. The practical relevance of the truncation is limited, however, particularly when dealing with problems with large numbers of components. Moreover, the resulting truncated version can in any case be viewed as a directly specified alternative model in its own right, rather than as an approximation to the DP mixture. The MCMC strategy we use follows that of Kottas and Sanso (2007), with some changes in detail related to the resampling steps for parameters, and is briefly outlined in Appendix A.

2.4 Immunofluorescence Histology Image Analysis

2.4.1 Context

The motivating application for this work arises in immunological studies in mice where multiple images provide data on the spatial configuration of many immune cells in a specific, localized region of lymphatic or spleen tissues. A single experiment views an image as the response to stimulus via injection of a vaccine, the overall context being exploration of responses under candidate vaccine designs. Comparisons involve replicate images from different mice – possibly at different times and under differing treatments – with careful matching and registration of the tissue region across mice. Observed measurements are fluorescent intensities generated from tagged cell-surface proteins that characterize a specific cell type. The pixel region model adopts a very small, 3×3 region of pixels as the level of resolution for modelling; this is small enough to be consistent with each region x being either occupied by a single cell $(y(x) = 1)$ or being unoccupied. Interest lies in characterizing the spatial intensity functions underlying observed data in each image, and feeding the statistical summaries and characterizations into visual and numerical comparisons. For the current paper, we simply explore aspects of the analyses of two example images, focussing on statistical aspects.

2.4.2 Measurement Error Models

Based on exploration of past data and experimentation with different measurement error models, a simple truncated lognormal model for the measurement of the fluorescence intensity appears to be adequate. That is, if $z = z(y(x))$ represents measured fluorescence at pixel region x, then $p(z|y, \delta)$ is defined by

$$(\log(z)|y, \delta) \sim N(m_y, v_y)I(\log(z) < h), \quad (y = 0, 1),$$

where (m_0, v_0) relates to the background noise and (m_1, v_1) to the distribution of signal fluorescence – i.e., the distribution conditional on the pixel region being occupied by a cell. Then $\delta = (m_0, m_1, v_0, v_1)$. Here $h = \log(255)$, the truncation being inherent to the digital fluorescent image generation process; $\log(z)$ is recorded on a scale of $(0, h)$ with values greater than h being truncated.

As described in Section 2.3.1, the overall MCMC can be extended to include a component representing uncertainty about δ. Under the truncated normal model for $(z|\delta)$, this can be done with standard priors, although it raises a need for a Metropolis strategy due to the truncation effects. We adopt independent normal-inverse gamma priors; for each of $y = 0, 1$ independently, $(m_y|v_y) \sim N(m_y|\bar{m}_y, tv_y)$ and $v_y^{-1} \sim Ga(c/2, c\bar{v}_y/2)$ based on specified prior estimates $\bar{m}_y, \bar{v}_y$ of m_y, v_y, respectively. These values are chosen to reflect known scales of (log) fluorescence and background. Specified values of the hyper-parameters t, c are used to define relatively precise priors while allowing for adaptation to the data in each new image data analysis. In the examples here, $\bar{m}_0 = 3.5$, $\bar{m}_1 = 4.5$, $t = 10$, $\bar{v}_0 = 0.1$, $\bar{v}_1 = 0.1$ and $c = 10$. We explore aspects of model fit and adequacy in the following discussion.

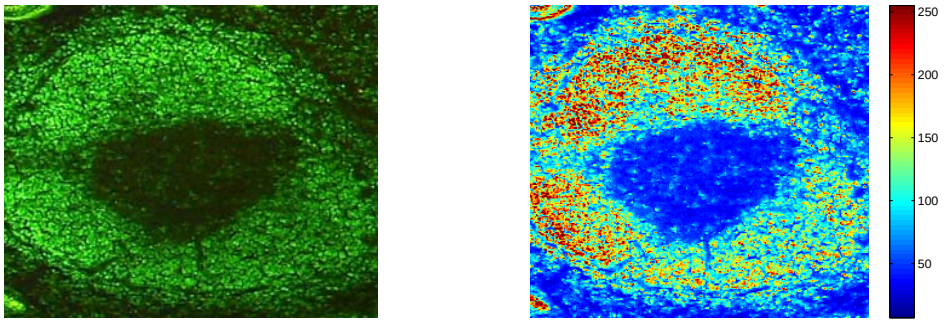
Conditional on the current Y and data Z, the (m_0, v_1) and (m_1, v_1) are conditionally independent. Each of the two conditional posteriors is the product of two terms: updated inverse-gamma based on the full set of $N(\log(z)|m_y, v_y)$ non-truncated terms in the conditional likelihood functions, and the term involving products of cumulative normal distribution function values derived from truncation. The first terms provide suitable Metropolis proposals for the full conditional posteriors, with the contribution to the conditional likelihoods from the truncation normalization providing the terms in the acceptance ratio. Due to the fact that only a relatively small fraction of the data is truncated in these images (almost none,

generally, for background levels, and in the region of up to 15% for signal) with much of the data lying well below the limit, these constructed proposal distributions are typically close to the target conditional posterior and the resulting sampler very effective.

2.4.3 Image Data B220

Figure 2.1a is the original image of the intensity of emission from the fluorescent dye AF350 conjugated to an antibody that binds to B220, a molecule expressed on the surface of B lymphocytes, in a section of a lymph node excised and sectioned 24 hours after subcutaneous alum injection. Figure 2.1b shows the corresponding heat map of intensity levels, which after logging provide the raw numeric data Z; histograms of the fluorescence and logged values Z are shown in Figure 2.2. The image indicates typical heterogeneity of spatial distributions of a very large number of cells. On the resolution analyzed, we have 180×180 pixel regions, each 3×3 pixels representing the locations of individual cells (if present). We take the image region S as $[-5, 5]^2$ so $\Delta = 1/324$.

We use priors as follows. First $\alpha \sim Ga(1, 1)$ and $\gamma \sim Ga(1, 0.001)$ for the two scalar model parameters. The base prior $G_0(\mu, \Sigma)$ is $N(\mu|0, t_0\Sigma)IW(\Sigma|s_0, S_0)$ where $t_0 > 0$, $s_0 > 0$ is the prior degree-of-freedom and $E(\Sigma) = S_0/(s_0 - 2)$ when $s_0 > 2$. Analysis here adopts $t_0 = 50$, $s_0 = 2$ and $S_0 = 0.4I$, based on the specified scale of the image region $S = [-5, 5]^2$ and the expectation of needing a large number of widely dispersed and relatively concentrated normal components to represent a very heterogeneous intensity surface. Further, the truncation of the mixture model uses $k = 250$ as the upper bound on the number of components.



(a) (b)

Figure 2.1: Data image on cell experiment B220 on day 1; (a) shows an image of the original data with fluorescent green tag, and (b) shows the scale of the corresponding intensity data.

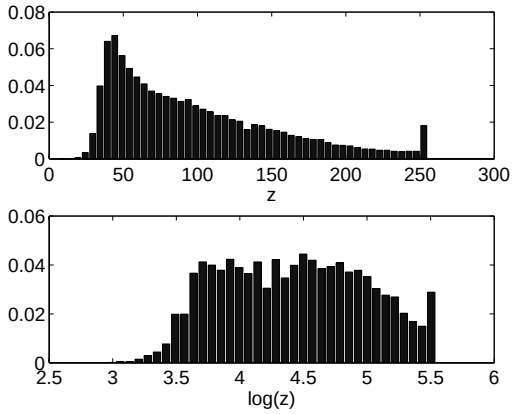
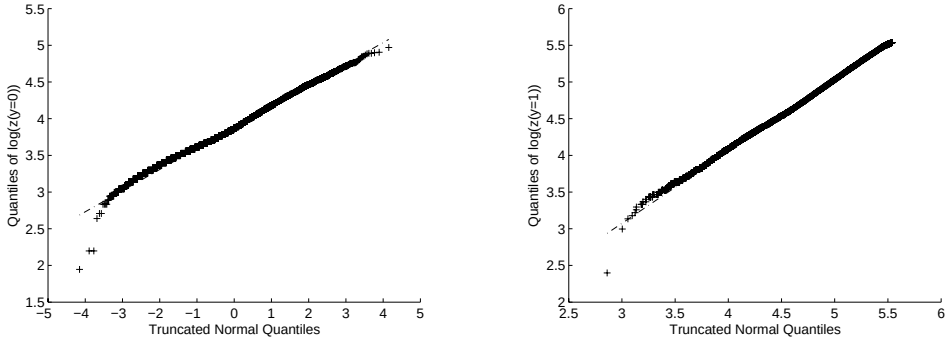


Figure 2.2: Data from B220, day 1: intensities (upper) and log intensities Z (lower).

Some aspects of the analysis with a final 5000 MCMC iterations after burn-in are graphically summarized in the figures. Additional visual confirmation of the relevance of the truncated normal measurement error model is illustrated in Figure 2.3, showing normal qqplots of the Z data partitioned into noise and signal samples based on the current indicators Y at one randomly chosen step on the MCMC analysis. Repeat draws show similar forms. Looking at these graphs for a series of MCMC steps is useful in confirming the stability of the apparent adequacy of the truncated normal model, as reflected in the qqplots across multiple realizations of the signal/noise allocation of the pixel regions. Evidently, the raw data displayed in the lower frame of Figure 2.2 shows the signal/noise structure, but without conditioning on signal/noise assignments it is difficult to develop direct graphical or numerical assessments of the normality assumption; repeat exploration across a series of MCMC samples aids measurably in this exploratory model assessment exercise.



(a) (b)

Figure 2.3: Based on a single, randomly selected draw from the MCMC in analysis of B220, day 1, the data are partitioned into noise $(y = 0)$ and signal $(y = 1)$ and the two corresponding samples of log intensities Z are displayed as normal qqplots: (a) noise, and (b) signal. This provides useful, visual insight into the utility and relevance of the truncated noise measurement error model and represents a nice dissection of the full data in the histogram of Figure 2.2.

Additional snapshots of one of the MCMC iterates are graphed as follows. Figure

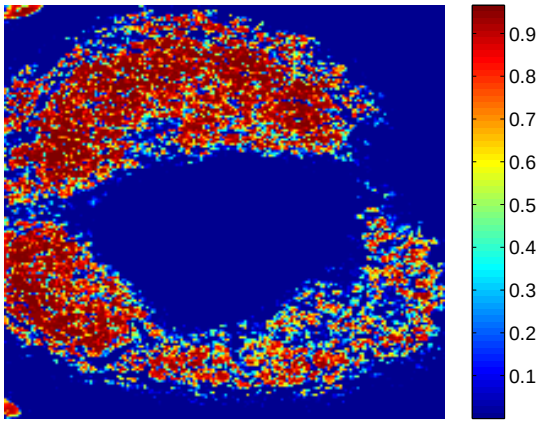


Figure 2.4: Image plot of $Pr(y(x) = 1|Z)$ at one randomly chosen step of the MCMC in analysis of B220, day 1.

2.4 shows an image of one posterior sample of $Pr(y(x) = 1|Z)$ over all $x \in \mathcal{X}$, and Figure 2.5 (upper frame) identifies the corresponding current sampled mixture components overlaid on the data.

The supplemental material on the web site `http://stat.duke.edu/people/theses/JiC.html` contains a series of these "snapshot" figures in a movie, which gives a nice overview of the uncertainty across MCMC steps. Some regions are more stable/less variable than others, and this comes through best by viewing a series of samples through the MCMC. Viewing such figures aids in understanding and illustrating aspects of the model, and comparison of such MCMC snapshots with the real data in Figure 2.1 is illuminating. In terms of posterior estimates, averaging over MCMC iterates produces relevant summaries. For example, Figure 2.5 (lower frame) shows an image of the posterior mean intensity estimate based on averaging Monte Carlo samples $f(x|\Theta)$ over the MCMC steps. In essence and up to a constant, this also represents the Monte Carlo posterior mean of the probabilities in Figure 2.4. Comparisons with Figure 2.1 begin to indicate the ability of the model to reflect the complexity of the data configuration, and with large num-

bers of heterogeneous mixture components adapt to very variable patterns in the underlying spatial intensity.

Monitoring of various parameters aids in assessment of convergence, again at the usual informal level. Rapid stabilization of trajectories of key single parameters, including α, γ, N and δ among others, is typically observed, and that is exemplified in Figures 2.6, 2.7 and 2.8. Good mixing is evident in these and other marginal trajectory plots.

From the applied perspective, the samples for N provide summary approximate posterior inferences on the underlying numbers of occupied pixel regions, i.e. our proxy for the number of cells, as one characteristic of this data set. Figure 2.8b indicates that N very likely lies in the range $8,500 - 8,700$ for B220 on day 1, corresponding to 26-28% coverage of the image.

For a brief visual comparison, a second image of data illustrates the ability of the model analysis to reflect a diversity of patterns of complexity in image intensities. This comes from data captured from tissue in the same experiment, using the same dye-antibody combination, but now after an additional 10 days. Comparisons between selected posterior summaries between day 1 and day 11 clearly indicate that the distribution of the fluorescently labelled cells has changed significantly between day 1 and day 11. For example, the number of the fluorescent labelled cells has apparently also reduced significantly in the later stages; Figure 2.9 suggests N likely lies in the range $8,700 - 9,200$ for B220 on day 11, so that the coverage of the overall image region is slightly increased relative to day 1. Further, there are multiple regions with higher intensities on day 1 that dissipate later on, and the overall intensity becomes fragmented; see the images for day 11 in Figure 2.10 in comparison to those at day 1.

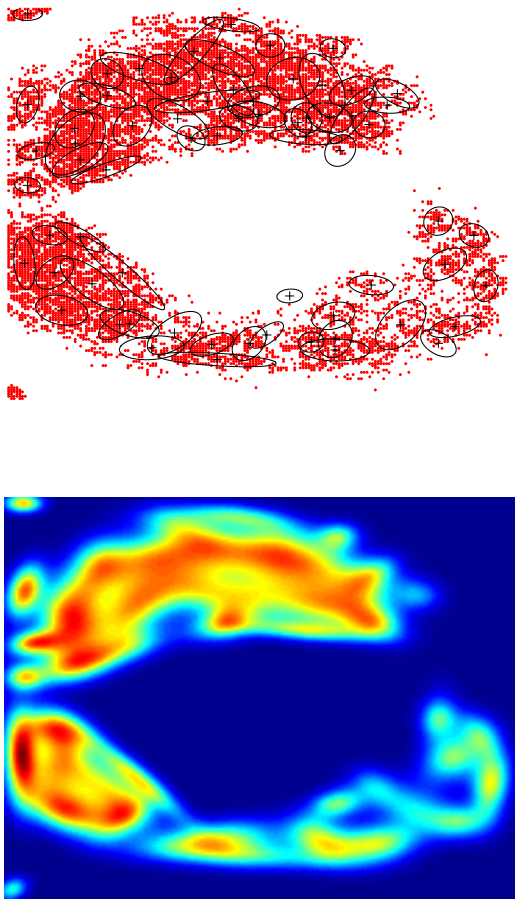
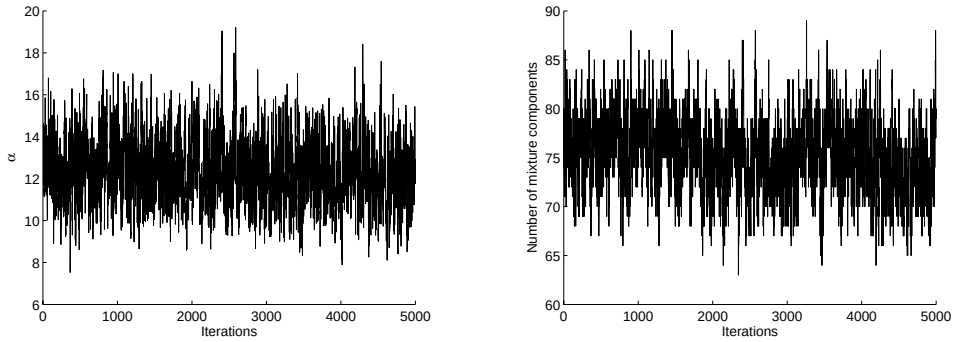


Figure 2.5: *Upper frame:* Scatter plot of the current sampled locations of cells Y at one MCMC step in analysis of B220 on day 1, overlaid with contours representing the location, scale and shape of the corresponding posterior sample of the normal mixture components underlying the intensity function. The contours are drawn at one standard deviation from the means in each of the major and minor axes directions. *Lower frame:* Image plot of the posterior estimate of the normalized intensity function $f(x)$ in analysis of B220 on day 1, based on averages of the sampled surfaces over MCMC steps.



(a) (b)

Figure 2.6: MCMC outputs in analysis of B220, day 1: Trajectories of (a) the Dirichlet process precision parameter α, and (b) the number of realized, non-empty components in the mixture model.

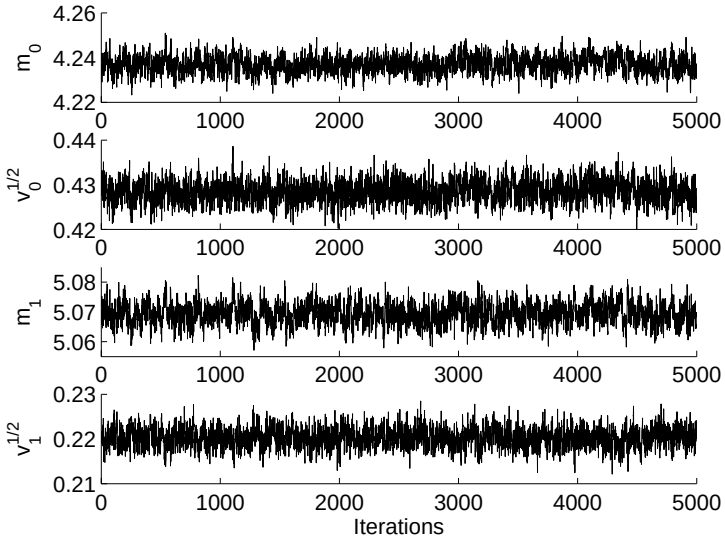
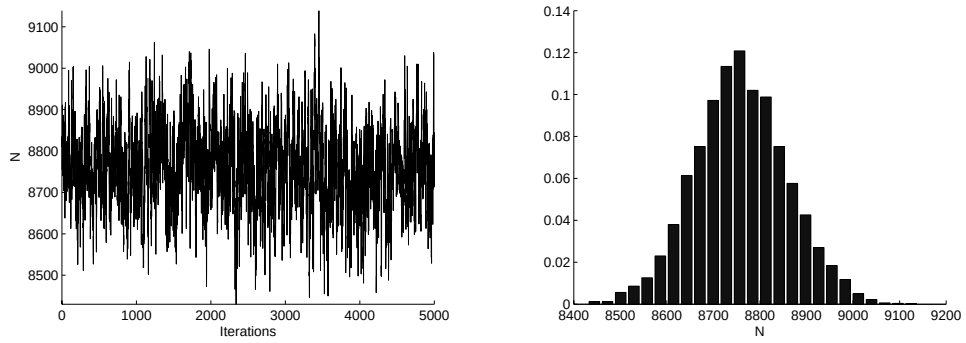


Figure 2.7: Trajectories of MCMC samples of measurement error model parameters in analysis of day 1 data.



(a) (b)

Figure 2.8: Experiment B220, day 1: Plots to show the number of "cells" in the image. (a) Trajectory of sampled N values in the MCMC, (b) the resulting histogram.

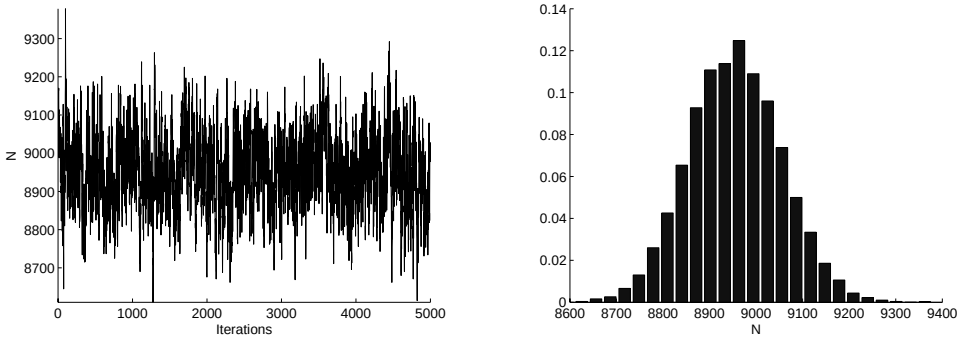


Figure 2.9: Experiment B220, day 11: Plots to show the number of "cells" in the image. (a) Trajectory of sampled N values in the MCMC, (b) the resulting histogram.

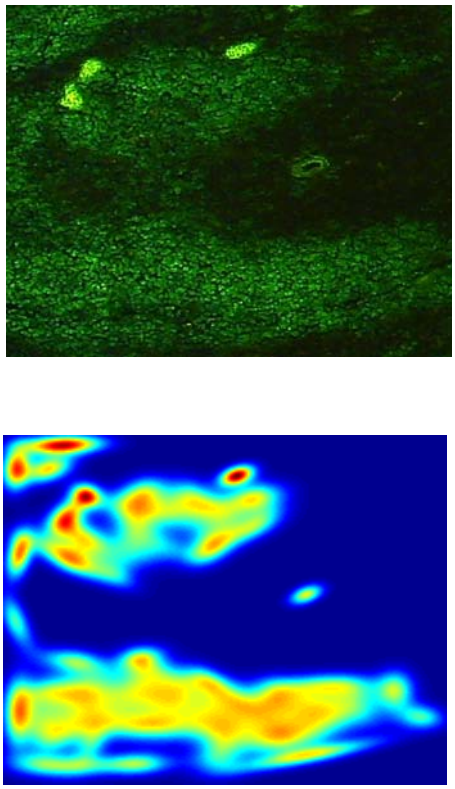


Figure 2.10: B220 on day 11. *Upper frame:* Image of the original data with fluorescent green tag. *Lower frame:* Image plot of the posterior estimate of the normalized intensity function.

2.5 Additional Comments

Our applied studies involve large (though unknown) numbers of point occurrences and intensity mixture models with relatively large numbers of mixture model components to represent potentially complex patterns of variation over the spatial region. Coupled with the need for practically relevant measurement error models to link between observed, noisy data and the underlying latent spatial process of biological relevance, this represents a challenging computational as well as modelling problem context. Our examples shown here, and experiences with other data sets, indicate the relevance and utility of the model developed. The use of flexible,

nonparametric Bayesian mixture models of intensity functions, pioneered by Kottas and Sanso (2007) and extended here, is central and key in engendering adaptability to wildly heterogeneous intensity patterns coupled with robustness and in-built parsimony. The use of effective MCMC samplers is key, and the blocked sampler for Dirichlet process mixture models is attractive from that viewpoint, but also really necessary as our overlaid measurement error structure demands that we have direct, albeit approximate evaluation of the underlying density-intensity function with the MCMC that generates from conditional posteriors of the underlying latent spatial process. In many spatial point process modelling contexts, lack of complete, direct observation on point outcomes is common, and our new methodology provides examples of how the overall analysis framework can be extended to allow for that. Our immunological context generates data sets for which truncated normal models of fluorescence, under both signal and noise at a point location, are adequate, and our experience suggests that we can robustly include learning on measurement error models within the overall analysis. Other contexts may, of course, require alternative measurement error model choices, but the general strategy will apply.

Our use of mixtures of normals for the spatial intensity builds on the well-known framework of normal mixtures and their ability to represent even very highly irregular surfaces. Kottas and Sanso (2007) used mixtures of bivariate betas. The choice of parametric form of the mixands, or "kernels", is to some extent arbitrary, and the use of mixtures of betas is mathematically elegant when the spatial region is a specified rectangle. Normal mixtures do offer advantages, however. In modelling terms, the restricted range of correlations and shapes that bivariate betas are able to represent limits their flexibility. To represent an irregular intensity as accurately as a mixture of normals then requires more beta mixture components. We

have explored this in studies with simulated data and confirmed a need for 4 or 5 times the number of beta than normal components in some examples. This imposes greater computational burden and decreased flexibility. Looking ahead to 3-dimensional, and possibly higher-dimensional extensions, normal mixtures clearly generalize trivially, in both modelling and computational implementation senses. Further, due to the lack of conditional conjugacy for parameters of the bivariate betas within clusters, the MCMC analysis is complicated. Kottas and Sanso (2007) use a traditional Dirichlet process mixture Gibbs sampler with Metropolis-Hasting (MH) steps for sets of beta parameters. Beyond the difficulties in specifying efficient MH proposals, and subsequent inefficiencies, the standard, Polya urn-based Gibbs sampler for these mixtures is inherently slow mixing and this is increasingly problematic with larger sample sizes and numbers of mixture components, such as in our examples. These issues make the MCMC sampler for beta mixtures very slow compared to approaches based on block sampling and that can analytically integrate over parameters, exploiting conditional conjugacy, as in the normal mixture models. In head-to-head comparisons with the data sets here, we find computations in the normal model to be roughly *20 times faster* per iterate than using the beta model algorithm.

Current and potential future areas for consideration include refined computational strategies to increase computational efficiency and enable at least partial parallel implementation to take advantage of both multi-threading and cluster computation. New statistical directions might include consideration of local spatial dependencies in the 0/1 outcomes process, and also potential dependencies at the observational level due to fluorescence scatter across neighboring pixel regions. Potential refinements of prior specifications over the normal variance matrix parameters may also be of interest; for example, mixtures of priors favoring very

different scales of variances in the Σ_j may allow us to more adequately represent very heterogenous images. In the applied context of immunofluorescent images arising in studies of vaccine design, current case studies are focused in part on the context-specific questions of making comparisons between models fitted to two or more images. Further studies are currently exploring extensions of the current approach to deal with problems in which several cells of distinct biological types are marked by fluorescent tags with distinct emission spectra; interest then lies in simultaneously estimating two or more underlying spatial intensity functions for the separate cell types, with a need for dealing with the uncertainties about cell type at any one pixel region location due to frequency interference in recorded intensities. Some of the potential further studies mentioned here has been reported in a recent work (Manolopoulou et al., 2009).

Supplementary material and code

The web page `http://ftp.stat.duke.edu/WorkingPapers/08-25.html` provides freely available Matlab code that implements the method described here. This includes support functions and the examples from this paper as templates for other more general models. The site also provides additional information on aspects of posterior uncertainty and predictive fit in the day-1 example. These include contour and image plots of successive samples of the DP mixture-based intensity surface through a series of MCMC iterations, and associated plots of the changes in the implied MCMC-based posterior mean estimate of the intensity function as it is updated through a series of MCMC iterations. Additional supplementary plots show pixel probabilities representing presence/absence of cell-based fluorescence as they vary over a series of MCMC iterations, accompanied by sampled spatial point patterns – i.e., locations of cells – corresponding to the above probabilities as

the MCMC progresses.

In terms of computational benchmarks (*circa* February 2009), for each one of the examples presented here each iterate of the MCMC algorithm presented takes roughly 2-3 seconds when running on a 2.80 GHz Intel Pentium 4 laptop with 1024 MB memory. Further research will investigate multicore and multiscale implementations that will speed up analyses substantially. The current MCMC is of the order of 20 times faster than alternatives using more traditional Polya urn configuration MCMC samplers.

Chapter 3

Dependent DP Mixture Model and SMC Sampling

In this chapter, dependent Dirichlet process mixture models are introduced for application in complex *dynamic* systems. Following the previous work on dependent DP mixture modelling (MacEachern, 1999, 2001; Muller et al., 2004; Pennell and Dunson, 2006; Dunson et al., 2007; Caron et al., 2007), we utilize a stick-breaking dependent DP mixture model and introduce dependence between mixture distributions on component parameters. MCMC methods for dependent DP mixture models are well studied in the previous literature. However, for application in dynamic systems, sequential Bayesian inference are always required, particularly when interest in prediction for real-time systems. I present a Rao-Blackwellized particle filter (RBPF) for sequential Bayesian inference in dependent DP mixtures. In RBPF, I process sequential Monte Carlo only on mixture component allocation variables for observed data points. Given these allocation variables, I use dynamic models for parameters of the mixture components, enabling closed form updates for these variables. I demonstrate the model and computational approach in a synthetic problem of sequential time-varying density estimation.

3.1 Introduction

Traditional time series analysis is constrained by parametric assumptions for evolution and measurement noise distributions. Even though flexible or non-parametric modelling can be introduced for the evolution process, inferences are often re-

stricted to the moments of the assumed distributions, and thus changes may be overlooked by the model if it can not be captured by those moments. Moreover, in many situations, the measurements are *distributions*, which can present multimodality and other heterogeneous features. For example, in dynamic topic modelling, the number of clusters and the weights/locations of these clusters of topics may change over time (Caron et al., 2007; Srebro and Roweis, 2005); in multiple extended target tracking, observations can be viewed as a spatial point process with a underling highly spatial inhomogeneous intensity function (Gilholm et al., 2005; Singh et al., 2009); in genetic epidemiology studies the interest is the evolution of the distribution of DNA damage over time (Rodriguez and ter Horst, 2008).

This chapter develops flexible modelling for estimation and prediction of densities that evolve in discrete time. Dependent Dirichlet process mixture models in time varying settings are introduced along with computationally efficient algorithms. Following the previous work on dependent DP mixture modelling (MacEachern, 1999, 2001; Muller et al., 2004; Pennell and Dunson, 2006; Dunson et al., 2007; Caron et al., 2007), countably infinite mixtures of Gaussian distributions are employed to represent the unknown density at each time point, and time-to-time dependencies are defined on mixture component weights and locations using dynamic linear models. These methods can be treated as extensions of the Dirichlet process mixture model (Escobar and West, 1995) to collections of distributions evolving in discrete time. Since dependence is built into the mixing distribution by allowing parameters to evolve dynamically, the models we present can also be regarded as extensions of the Gaussian Dynamic Linear Models (DLMs) of West and Harrison (1997).

For application in dynamic systems, sequential Bayesian inference is always required, particularly when interest lies in prediction in real-time systems. As the

analytical update for such complex dynamic system is not feasible, we have to turn to sequential Monte Carlo approaches. However, due to the complexity of dependent DP mixtures, a generic SMC method will suffer from degeneracy problems. To address this, I present a novel sequential Monte Carlo method, Rao-Blackwellized particle filter (RBPF), for the sequential Bayesian inference in dependent DP mixtures. In RBPF, mixture component allocation variables for observed data points follow a nonlinear dynamic system and are updated by sequential Monte Carlo. Then, given these allocation variables, we can obtain dynamic linear models for associated parameters to enable us to process closed-form updates for these parameters. Simulation studies on a synthetic problem of sequential time-varying density estimation demonstrate the performance of the proposed model and computational approach.

3.2 Dependent Dirichlet Process Mixture Models

Most of the applications of Dirichlet Process models focus on problems with exchangeable samples from one unknown distribution. However, in many situations, we cannot assume that the distribution of the observations is fixed; instead, it evolves over time. For example, in a clustering application, the number of clusters and the locations of these clusters may change over time. More specifically, let $t = 1, 2, ..., T$ denote a discrete-time index and assume that we receive N_t observations at each time t, denoted by $\mathbf{x}_t = x_{1:N_t,t}$, which are independent and identically distributed (iid) samples from

$$f_t(\cdot) = \int_{\Theta} p(\cdot|\theta) G_t(d\theta) \tag{3.1}$$

where $p(\cdot|\theta)$ is the mixed pdf and G_t is the mixing distribution which itself is distributed according to a Dirichlet Process

$$G_t \sim DP(\alpha, G_0) \tag{3.2}$$

where G_0 is the base probability measure, and $\alpha > 0$ is the total mass, or precision of the DP.

Developing dependent Dirichlet process mixture models, particularly for time-evolving data, has recently been the focus of significant interest, and researchers have proposed various approaches directed toward specific applications. Most use the stick-breaking representation (Sethuraman, 1994) to introduce dependencies, stimulated by the dependent Dirichlet process framework proposed by MacEachern (1998, 1999, 2001). Under the stick-breaking representation, one can represent a realization of a Dirichlet process by two infinite dimensional vectors of weights and clusters locations, and introduce the dependency either on the weights (Grifin and Steel, 2006) or on the clusters locations in (DeIorio et al., 2004; Gelfand et al., 2005). An early example is the order-based dependent DP (Grifin and Steel, 2006), in which the model is time-reversible but is not Markovian, and it requires one to specify how the mixture weights change over time. Alternatively, convex combinations of independent Dirichlet processes can be used for modelling collections of dependent random measures. The dependency is then introduced through the weight coefficients (Muller et al., 2004; Pennell and Dunson, 2006; Dunson et al., 2007; Dunson and Park, 2008), which leads to an easy way of constructing a MCMC sampling strategy. Alternative approaches are based on the Polya urn-type (Ferguson, 1973) representation of Dirichlet processes, (Walker and Muliere, 2003; Zhu et al., 2005; Caron et al., 2007), implemented by changing the number and locations of clusters over time.

In the following, I present two dependent DP mixtures for time-varying data.

The first one utilizes the Polya Urn scheme, inspired by the work of Caron et al. (2007). The second one is based on the stick-breaking representation, related to (Grifin and Steel, 2006).

3.2.1 Polya Urn Scheme-based Dependent DP Mixture

Under the Polya Urn-type representation of the Dirichlet process (Ferguson, 1973), we introduce a vector of configuration variable $\mathbf{c} = c_{1:N}$ which follows the model,

$$p(c_i = c|c_1, ..., c_{i-1}) = \begin{cases} \frac{n_{i,c}}{i-1+\alpha} & \exists j \in \{1, ..., i-1\}, c_j = c \\ \frac{\alpha}{i-1+\alpha} & \text{for } j = 1, ..., i-1, c_j \neq c \end{cases} \tag{3.3}$$

where $n_{i,c}$ is the number of c_j for $j < i$ that are equal to c. Time dependence on configuration variables $\mathbf{c_t}$ can be introduced by the Generalized Polya Urn random partition model proposed in (Caron et al., 2007). The idea is that at each time step t, first delete randomly a subset of the configuration variables which survived the previous $t-1$ deletion steps and then sample new configuration variables corresponding to the N_t observations $x_{1:N_t,t}$. More specifically, take the same notation from Caron et al. (2007): $\mathbf{c}_{1:t-1}^{t-1}$ (resp. $\mathbf{c}_{1:t-1}^{t}$) denotes the subset of $\mathbf{c}_{1:t-1}$ corresponding to variables having survived in the deletion steps from time 1 to $t-1$ (resp. from time 1 to t); $K_{1:t-1}$ denotes the number of clusters from time 1 to $t-1$; $\mathbf{m}_{t-1}^{t-1}$ (resp. $\mathbf{m}_{t-1}^{t}$) denotes the number of data points in each cluster associated to $\mathbf{c}_{1:t-1}^{t-1}$ (resp. $\mathbf{c}_{1:t-1}^{t}$); $\mathcal{I}(\mathbf{m}_t^t)$ indicates the non-zero entries of $\mathbf{m}_t^t$. In the initialization stage, generalized Polya Urn scheme proceeds the same as traditional Polya Urn scheme (Ferguson, 1973; Escobar and West, 1995). At $t \geq 2$, the generalized Polya Urn scheme proceeds as follows,

- Delete each configuration variable in $\mathbf{c}_{1:t-1}^{t-1}$ with probability $1-\rho$ $(0 \leq \rho \leq 1)$ to obtain $\mathbf{c}_{1:t-1}^{t}$ (hence $\mathbf{m}_{t-1}^{t}$) and set $\mathbf{m}_t^t = \mathbf{m}_{t-1}^t$, $K_t = K_{t-1}$.

- For $i = 1, ..., N_t$

  - with probability $\frac{m_{k,t}^t}{\sum_k m_{k,t}^t + \alpha}$, $k \in \mathcal{I}(\mathbf{m}_t^t)$, set $c_{i,t}^t = k$, $m_{k,t}^t = m_{k,t}^t + 1$,
  - with probability $\frac{\alpha}{\sum_k m_{k,t}^t + \alpha}$, set $K_t = K_t + 1$, $c_{i,t}^t = K_t$, $m_{k,t}^t = 1$.

According to Caron et al. (2007), the sequences $\{\mathbf{c}_{1:t}^t\}$, $\{\mathbf{c}_{1:t-1}^t\}$, $\{\mathbf{m}_{1:t}^t\}$ and $\{\mathbf{m}_{1:t-1}^t\}$ are Markovian. The G_t constructed from this generalized Polya Urn scheme is asymptotically a second order stationary process.

Time Propagation Models for Random Partition

The Generalized Polya Urn scheme reviewed above introduces the time dependence for Dirichlet process through the evolution of variate $\mathbf{m}_{1:t}^t$, but it can not be expressed in explicit formula. To simplify the Generalized Polya Urn scheme, a discount factor δ is utilized for $m_{k,t}$, the number of data points in each cluster, and enables a simple process to introduce dependence on adjacent Dirichlet processes. Specifically, $\mathbf{m}_{t|t-1} = \delta \mathbf{m}_{t-1}$ $(0 < \delta \leq 1)$ and with $m_{k,t|t-1} = 0$ if $m_{k,t|t-1} < 1$ to prevent $m_{k,t|t-1}$ become smaller than 1. By introducing the discount factor, we can reduce the number of data points in each cluster. Thus when using $\mathbf{m}_t$ to build the random partition, we can reduce the correlation between G_t and G_{t-1}. Given the prior information contained in $\mathbf{m}_{t|t-1}$, we assign the allocation variables of data $x_{1:N_t}$ via the generalized Polya Urn: $\mathbf{m}_t = \mathbf{m}_{t|t-1}$, for $i = 1, ..., N_t$,

- with probability $m_{k,t} / \sum_k m_{k,t} + \alpha$, $k \in \mathcal{I}(\mathbf{m}_t)$, set $c_{i,t} = k$, $m_{k,t} = m_{k,t} + 1$,
- with probability $\alpha / \sum_k m_{k,t} + \alpha$, set $K_t = K_t + 1$, $c_{i,t} = K_t$, $m_{k,t} = 1$.

Time Propagation Models for Cluster Means

We assume Dynamic Linear Models (DLMs) for the propagation of the mean $\mu_{k,t}$ of each cluster k over time t and utilize some of the standard notation in West and Harrison (1997). The DLMs for each cluster mean is defined by a DLM quadruple

$$\{H_t, E_t, V_{k,t}, W_{k,t}\}$$

Here, t is the time indicator whereas k indexes each cluster ($k = 1, ..., K_t$). All quantities in the quadruple are assumed to be known. Each cluster mean propagation model can be written as:

$$Observation \quad : \quad \overline{y}_{k,t} = H_t \mu_{k,t} + \varepsilon_{k,t}, \qquad \varepsilon_{k,t} \sim N(0, V_{k,t}\Sigma_{k,t}) \tag{3.4}$$

$$Evolution \quad : \quad \mu_{k,t} = E_t \mu_{k,t-1} + \epsilon_{k,t}, \qquad \epsilon_{k,t} \sim N(0, W_{k,t}\Sigma_{k,t-1}) \tag{3.5}$$

$$Prior \quad : \quad \mu_{k,0} \sim N(0, \tau\Sigma_{k,0}), \qquad \Sigma_{k,0} \sim IW(\cdot|d_0, S_0) \tag{3.6}$$

where observation $\overline{y}_{k,t}$ is the sufficient statistics for cluster mean: given the realization of the allocation variable $\mathbf{c}_t$ for all data points, the data is assigned to groups indexed by k. For a group with index k, denote $y_{k,t} = \{x_{i,t} : c_{i,t} = k\}$, $n_{k,t} = \#\{y_{k,t}\}$, thus $\overline{y}_{k,t} = \sum_{i=1}^{n_{k,t}} y_{k,t}^{(i)}/n_{k,t}$ is the sufficient statistics for cluster mean. For simplicity, the system noise and observation noise are set proportional to cluster variance $\Sigma_{k,t}$ and $\Sigma_{k,t-1}$ with scale $V_{k,t}$ and $W_{k,t}$ respectively. More specially, $V_{k,t} = 1/n_{k,t}$ and $W_{k,t}$ can be set equivalent to τ. Moreover, some standard conditional independence assumptions are necessary: given all parameters the random innovations $v_{k,t}$ and $\epsilon_{k,t}$ are independent across time and mutually independent. Given the previous posterior of $\mu_{k,t}$, $N(\cdot|\nu_{k,t-1}, R_{k,t-1}\Sigma_{k,t-1})$, and assuming we can obtain the observation $\overline{y}_{k,t}$, the sufficient statistics of each cluster mean, we can process sequential inference for the posterior of $\mu_{k,t}$, $N(\cdot|\nu_{k,t}, R_{k,t}\Sigma_{k,t})$, via the following equations:

Predict:

$$\nu_{k,t|t-1} = E_{k,t}\nu_{k,t-1} \tag{3.7a}$$

$$R_{k,t|t-1} = E_{k,t}R_{k,t-1}E_{k,t}^T + W_{k,t} \tag{3.7b}$$

Update:

$$\nu_{k,t} = \begin{cases} \nu_{k,t|t-1} + A_{k,t}e_{k,t}, & \text{if } n_{k,t} > 0 \\ 0, & \text{if } n_{k,t} = 0 \end{cases} \tag{3.8a}$$

$$R_{k,t} = \begin{cases} R_{k,t|t-1} - A_{k,t}Q_{k,t}A_{k,t}^T, & \text{if } n_{k,t} > 0 \\ \tau, & \text{if } n_{k,t} = 0 \end{cases} \tag{3.8b}$$

$$A_{k,t} = R_{k,t|t-1}H_{k,t}Q_{k,t}^{-1} \tag{3.8c}$$

$$e_{k,t} = \overline{y}_{k,t} - \tilde{y}_{k,t} \tag{3.8d}$$

$$\tilde{y}_{k,t} = H_{k,t}\nu_{k,t|t-1} \tag{3.8e}$$

$$Q_{k,t} = H_{k,t}R_{k,t|t-1}H_{k,t}^T + V_{k,t} \tag{3.8f}$$

Time Propagation Models for Covariances

Models of $\Sigma_{k,t}$ varying stochastically over time have been studied thoroughly in time series (West and Harrison, 1997; Quintana and West, 1987). Here we use the "locally smooth", discount factor-based stochastic model (West and Harrison, 1997; Carvalho and West, 2007). The model involves constructing a Markov process in which transition distributions $p(\Sigma_{k,t}|\Sigma_{k,t-1})$ are defined based on matrix-Beta random innovations applied to elements of the Bartlett decomposition of $\Sigma_{k,t-1}$. Here we briefly address the basic ideas and operational results. Based on a specified discount factor ρ, $(0 < \rho \leq 1)$, the matrix Beta-Bartlett stochastic evolution model has the following key implications and features: Beginning at time $t-1$ with the current posterior $\Sigma_{k,t-1} \sim IW(\cdot|d_{k,t-1}, S_{k,t-1})$, the stochastic evolution of $\Sigma_{k,t-1}$ to

$\Sigma_{k,t}$ implies the time t prior

$$\Sigma_{k,t} \sim IW(\cdot|\rho d_{k,t-1}, \rho S_{k,t-1}). \tag{3.9}$$

The time-evolution maintains the inverse-Wishart form for the prior of $\Sigma_{k,t}$, while increasing the spread of the HIW distribution by reducing the degrees-of-freedom and maintaining the location at $S_{k,t-1}/d_{k,t-1}$. Moreover, given the prior distribution for $\Sigma_{k,0}$, namely

$$\Sigma_{k,0} \sim IW(\cdot|d_0, S_0), \tag{3.10}$$

and the data points allocated to each cluster at time step t, then we can infer the posterior for $\Sigma_{k,t}$ via,

$$\Sigma_{k,t} \sim IW(\cdot|d_{k,t}, S_{k,t}) \tag{3.11a}$$

$$d_{k,t} = \begin{cases} \rho d_{k,t-1} + n_{k,t}, & \text{if } n_{k,t} > 0 \\ d_0, & \text{if } n_{k,t} = 0 \end{cases} \tag{3.11b}$$

$$S_{k,t} = \begin{cases} \rho S_{k,t-1} + \Omega, & \text{if } n_{k,t} > 0 \\ S_0, & \text{if } n_{k,t} = 0 \end{cases} \tag{3.11c}$$

where $\Omega = \sum_{i=1}^{n_k} \left(y_{k,t}^{(i)} - \bar{y}_{k,t}\right)\left(y_{k,t}^{(i)} - \bar{y}_{k,t}\right)^T + \frac{R_{k,t|t-1} n_k}{R_{k,t|t-1}+n_k}(\nu_{k,t|t-1} - \bar{y}_{k,t})(\nu_{k,t|t-1} - \bar{y}_{k,t})^T$, with $\nu_{k,t|t-1}$ and $R_{k,t|t-1}$ defined in equations (3.7a) and (3.7b).

3.2.2 Stick Breaking Scheme-based Dependent DP Mixture

The Stick Breaking scheme-based dependent Dirichlet process proposed in MacEachern (1999) is constructed by replacing the base measure underlying Sethuraman's stick-breaking by a stochastic process $\eta(t) : t \in T$; that is,

$$G_t(\cdot) = \sum_{k=1}^{\infty} \pi_k(t) \delta_{\eta_k(t)}(\cdot)$$

where $\eta_k(t)$, for $k = 1, ...,$ are i.i.d. samples from a stochastic process $\eta(t)$ and $\pi_k(t) = V_k(t) \prod_{j=1}^{k-1}(1 - V_j(t))$ with $V_k(t)$ also i.i.d. samples from another stochastic process $V(t)$, $V(t) \sim Beta(1, \alpha(t))$ for all $t \in T$. The resulting DDP defines a distribution on a series of random distributions indexed by $t \in T$, with every G_t being marginally a Dirichlet process.

Here we consider mixture of normal distributions by a discrete-time dependent truncated DP, in which we set the maximum number of components as K. Besides, in our context, we introduce a discount factor δ for parameters in the stochastic process of $V_{k,t}$, and let $V_{k,t} \sim Beta(\delta\kappa_{k,t-1}, \delta\beta_{k,t-1})$. Then $\pi_{k,t}$ can be evaluated by $\pi_{k,t} = V_{k,t} \prod_{j=1}^{k-1}(1 - V_{j,t})$. The allocation variable $\mathbf{c}_t$ can be sampled from its prior, multinomial distribution $Mn(\cdot|\pi_{1,t}, ..., \pi_{K,t})$. Now, given the realization of the allocation variable $\mathbf{c}_t$ for all data points, the sufficient statistics for cluster mean $\overline{y}_{k,t}$ can be figured out easily as addressed in section 3.2.1. In summary, the overall model can be expressed as follows:

$$G_t = \sum_{k=1}^{K} \pi_{k,t}\delta(\mu_{k,t}, \Sigma_{k,t}) \tag{3.12a}$$

$$\pi_{k,t} = V_{k,t} \prod_{j=1}^{k-1}(1 - V_{j,t}) \tag{3.12b}$$

$$V_{k,t} \sim Beta(\cdot|\delta\kappa_{k,t-1}, \delta\beta_{k,t-1}) \tag{3.12c}$$

$$c_{i,t} \sim Mn(\cdot|\pi_{1,t}, ..., \pi_{K,t}) \tag{3.12d}$$

$$\overline{y}_{k,t} = H_t\mu_{k,t} + \varepsilon_{k,t}, \qquad \varepsilon_{k,t} \sim N(0, V_{k,t}\Sigma_{k,t}) \tag{3.12e}$$

$$\mu_{k,t} = E_t\mu_{k,t-1} + \epsilon_{k,t}, \qquad \epsilon_{k,t} \sim N(0, W_{k,t}\Sigma_{k,t-1}) \tag{3.12f}$$

$$\Sigma_{k,t} \sim IW(\cdot|\rho d_{k,t-1}, \rho S_{k,t-1}) \tag{3.12g}$$

with the priors

$$V_{k,0} \sim Beta(1, \alpha) \tag{3.13a}$$

$$\mu_{k,0} \sim N(0, \tau\Sigma_{k,0}) \tag{3.13b}$$

$$\Sigma_{k,0} \sim IW(\cdot|d_0, S_0) \tag{3.13c}$$

3.3 Sampling Methods

In the previous literature (MacEachern, 1999, 2001; Muller et al., 2004; Pennell and Dunson, 2006; Dunson et al., 2007; Caron et al., 2007; Rodriguez and ter Horst, 2008), MCMC methods have been widely used for posterior sampling in dependent DP mixture models. However, for applications in dynamic systems, sequential Bayesian inference is preferable, as analytical updates for such complex dynamic system are not feasible.

SMC for DP mixture models was studied in (Liu, 1996; MacEachern et al., 1999), in which they use collapsed Gibbs sampling method to propose the allocation variable $\mathbf{c}_t$ for each particle and propagate only the allocation variables. Recently Fearnhead (2004) and Fearnhead and Meligkotsidou (2007) gave further explorations of particle filtering technique for mixtures with unknown numbers of components, and provide some efficient resampling strategies. A key point in the above SMC sampling approaches for DP mixtures is that they all marginalized out the associated parameters of each cluster i.e. means and covariances, using SMC to propagate and update only for the allocation variable $\mathbf{c}$.

SMC methods for posterior sampling in dependent DP mixtures were first studied by Caron et al. (2007). However, in their algorithm the cluster means and covariances are proposed and propagated together with the allocation variables. As a result, the SMC algorithm only achieves very low effective sample size (ESS). To

overcome this disadvantage, I present a Rao-Blackwellized particle filter (RBPF) for Bayesian inference of dependent DP mixtures. In RBPF, I apply sequential Monte Carlo only to the allocation variables, while having access to closed-form updates for the mixture component parameters. The Rao-Blackwellisation technique enables the algorithm to achieve high *ESS* and therefore improves the efficiency of posterior estimation.

3.3.1 Rao-Blackwellized Particle Filter

Polya Urn scheme-based dependent DP mixture

The unknown parameters of interest are $(\mathbf{c}_t, \mu_{k,t}, \Sigma_{k,t})$ in the dependent DP mixture model with Polya Urn representation. As in Caron et al. (2007), we can treat all these parameters as the hidden state variable of a particle, and directly apply the sequential importance sampling and resampling process. The drawback of this generic particle filtering is that we can only achieve an extremely low effective sample size. To overcome such problems, we utilize the Rao-Blackwellized technique to facilitate sequential inference. The idea of the Rao-Blackwellized particle filter is to process particle filter for non-linear/non-Gaussian systems while using exact Bayesian inference for linear Gaussian systems via Kalman filter. In our context, the variate $\mathbf{c}_t$ follows a non-linear/non-Gaussian model while $\{\mu_{k,t}, \Sigma_{k,t}\}$ follow linear systems. The RBPF processes particle filter for $\mathbf{c}_t | (\mu_{k,t-1}, \Sigma_{k,t-1})$ and then closed-form updates for $(\mu_{k,t}, \Sigma_{k,t}) | \mathbf{c}_t$.

As discussed in section 3.2.1 , exact Bayesian inference for $\{\mu_{k,t}, \Sigma_{k,t}\}$ consists of two steps: ***Predict*** and ***Update***, which can be expressed as follows.

Predict:

$$\nu_{k,t|t-1} = E_{k,t}\nu_{k,t-1} \tag{3.14a}$$

$$R_{k,t|t-1} = E_{k,t}R_{k,t-1}E_{k,t}^T + W_{k,t} \tag{3.14b}$$

$$d_{k,t|t-1} = \rho d_{k,t-1} \tag{3.14c}$$

$$S_{k,t|t-1} = \rho S_{k,t-1} \tag{3.14d}$$

Update:

$$\nu_{k,t} = \begin{cases} a_{k,t} + A_{k,t}e_{k,t}, & \text{if } n_{k,t} > 0 \\ 0, & \text{if } n_{k,t} = 0 \end{cases} \tag{3.15a}$$

$$R_{k,t} = \begin{cases} R_{k,t|t-1} - A_{k,t}Q_{k,t}A_{k,t}^T, & \text{if } n_{k,t} > 0 \\ \tau, & \text{if } n_{k,t} = 0 \end{cases} \tag{3.15b}$$

$$d_{k,t} = \begin{cases} d_{k,t|t-1} + n_{k,t}, & \text{if } n_{k,t} > 0 \\ d_0, & \text{if } n_{k,t} = 0 \end{cases} \tag{3.15c}$$

$$S_{k,t} = \begin{cases} S_{k,t|t-1} + \Omega, & \text{if } n_{k,t} > 0 \\ S_0, & \text{if } n_{k,t} = 0 \end{cases} \tag{3.15d}$$

$$A_{k,t} = R_{k,t|t-1}H_{k,t}Q_{k,t}^{-1} \tag{3.15e}$$

$$e_{k,t} = \overline{y}_{k,t} - \tilde{y}_{k,t} \tag{3.15f}$$

$$\tilde{y}_{k,t} = H_{k,t}a_{k,t} \tag{3.15g}$$

$$Q_{k,t} = H_{k,t}R_{k,t|t-1}H_{k,t}^T + V_{k,t} \tag{3.15h}$$

$$\Omega = \sum_{i=1}^{n_k} \left(y_{k,t}^{(i)} - \bar{y}_{k,t}\right)\left(y_{k,t}^{(i)} - \bar{y}_{k,t}\right)^T \tag{3.15i}$$

$$+\frac{R_{k,t|t-1}n_k}{R_{k,t|t-1} + n_k}(\nu_{k,t|t-1} - \bar{y}_{k,t})(\nu_{k,t|t-1} - \bar{y}_{k,t})^T \tag{3.15j}$$

The RBPF for Polya Urn scheme-based dependent DP mixture model can now be summarized as follows.

Algorithm 3.1: Rao-Blackwellized particle filter for Polya Urn scheme-based dependent DP mixture

- At $t = 1$, for each particle $j = 1, ..., J$, set a predefined value for $K_0^{(j)}$ and initialize $\left\{\mu_{k,0}^{(j)}, \Sigma_{k,0}^{(j)}, \nu_{k,0}^{(j)}, R_{k,0}^{(j)}, d_{k,0}^{(j)}, S_{k,0}^{(j)}\right\}_{k=1}^{K_0^{(j)}}$. Obtain allocation variables $\mathbf{c}_1^{(j)}$ by generic Collapsed Gibbs sampling (Escobar and West, 1995, and Appendix A). Evaluate $\left\{\nu_{k,1}^{(j)}, R_{k,1}^{(j)}, d_{k,1}^{(j)}, S_{k,1}^{(j)}\right\}_{k=1}^{K_1^{(j)}}$ via (3.14) and (3.15) and draw posterior sample $\left\{\mu_{k,1}^{(j)}, \Sigma_{k,1}^{(j)}\right\}_{k=1}^{K_1^{(j)}}$.

- For $t > 1$:

  - For each particle $j = 1, ..., J$,

    * Update the number of data points in each cluster $m_{k,t|t-1}^{(j)}$ via $\mathbf{m}_{t|t-1}^{(j)} = \delta\mathbf{m}_{t-1}^{(j)}$.

    * Predict: evaluate $\Psi_{t|t-1}^{(j)} = \left\{\nu_{k,t|t-1}^{(j)}, R_{k,t|t-1}^{(j)}, d_{k,t|t-1}^{(j)}, S_{k,t|t-1}^{(j)}\right\}_{k=1}^{K_{t-1}^{(j)}}$ via equation (3.14).

    * Sample $\mathbf{c}_t^{(j)} \sim q(\cdot|\mathbf{m}_{t|t-1}^{(j)}, \Psi_{t|t-1}^{(j)}, \mathbf{x}_t)$ via collapsed Gibbs sampling, and evaluate $\mathbf{m}_t^{(j)}$ and $\pi_{k,t}^{(j)}$.

    * Update: evaluate $\Psi_t^{(j)} = \left\{\nu_{k,t}^{(j)}, R_{k,t}^{(j)}, d_{k,t}^{(j)}, S_{k,t}^{(j)}\right\}_{k=1}^{K_t^{(j)}}$ via equation (3.15).

    * Draw posterior sample $\left\{\mu_{k,t}^{(j)}, \Sigma_{k,t}^{(j)}\right\}_{k=1}^{K_t^{(j)}}$ given $\Psi_t^{(j)}$.

  - Compute the importance weights

$$\widetilde{w}_t^{(j)} \propto w_{t-1}^{(j)} \frac{p(\mathbf{x}_t|\mathbf{c}_t^{(j)})p(\mathbf{c}_t^{(j)}|\mathbf{m}_{t|t-1}^{(j)})}{q(\mathbf{c}_t^{(j)}|\mathbf{m}_{t|t-1}^{(j)}, \Psi_{t|t-1}^{(j)}, \mathbf{x}_t)},$$

with $\sum_{j=1}^{N} \widetilde{w}_t^{(j)} = 1$.

– Resample when ESS is less than a predefined threshold.

Stick Breaking scheme-based dependence DP mixture

The unknown parameters of interest are $(V_{k,t}, \mu_{k,t}, \Sigma_{k,t})$ in the stick breaking. As discussed in section 3.2.1 and 3.2.2, given the allocation $\mathbf{c}_t$ for each data point $\mathbf{x}_t$, exact Bayesian inference for $(V_{k,t}, \mu_{k,t}, \Sigma_{k,t})$ consists of two steps, ***Predict*** and ***Update***, which can be expressed as:

Predict:

$$\kappa_{k,t|t-1} = \delta\kappa_{k,t-1} \tag{3.16a}$$

$$\beta_{k,t|t-1} = \delta\beta_{k,t-1} \tag{3.16b}$$

$$\nu_{k,t|t-1} = E_{k,t}\nu_{k,t-1} \tag{3.16c}$$

$$R_{k,t|t-1} = E_{k,t}R_{k,t-1}E_{k,t}^T + W_{k,t} \tag{3.16d}$$

$$d_{k,t|t-1} = \rho d_{k,t-1} \tag{3.16e}$$

$$S_{k,t|t-1} = \rho S_{k,t-1} \tag{3.16f}$$

Update:

$$\kappa_{k,t} = \begin{cases} \kappa_{k,t|t-1} + n_{k,t}, & \text{if } n_{k,t} > 0 \\ 1, & \text{if } n_{k,t} = 0 \end{cases} \tag{3.17a}$$

$$\beta_{k,t} = \begin{cases} \beta_{k,t|t-1} + \sum_{r=k+1}^{K} n_{r,t}, & \text{if } n_{k,t} > 0 \\ \alpha, & \text{if } n_{k,t} = 0 \end{cases} \tag{3.17b}$$

$$\nu_{k,t} = \begin{cases} \nu_{k,t|t-1} + A_{k,t} e_{k,t}, & \text{if } n_{k,t} > 0 \\ 0, & \text{if } n_{k,t} = 0 \end{cases} \tag{3.17c}$$

$$R_{k,t} = \begin{cases} R_{k,t|t-1} - A_{k,t} Q_{k,t} A_{k,t}^T, & \text{if } n_{k,t} > 0 \\ \tau, & \text{if } n_{k,t} = 0 \end{cases} \tag{3.17d}$$

$$d_{k,t} = \begin{cases} d_{k,t|t-1} + n_{k,t}, & \text{if } n_{k,t} > 0 \\ d_0, & \text{if } n_{k,t} = 0 \end{cases} \tag{3.17e}$$

$$S_{k,t} = \begin{cases} S_{k,t|t-1} + \Omega, & \text{if } n_{k,t} > 0 \\ S_0, & \text{if } n_{k,t} = 0 \end{cases} \tag{3.17f}$$

$$A_{k,t} = R_{k,t|t-1} H_{k,t} Q_{k,t}^{-1} \tag{3.17g}$$

$$e_{k,t} = \overline{y}_{k,t} - \tilde{y}_{k,t} \tag{3.17h}$$

$$\tilde{y}_{k,t} = H_{k,t} a_{k,t} \tag{3.17i}$$

$$Q_{k,t} = H_{k,t} R_{k,t|t-1} H_{k,t}^T + V_{k,t} \tag{3.17j}$$

$$\Omega = \sum_{i=1}^{n_k} \left(y_{k,t}^{(i)} - \bar{y}_{k,t} \right) \left(y_{k,t}^{(i)} - \bar{y}_{k,t} \right)^T \tag{3.17k}$$

$$+ \frac{R_{k,t|t-1} n_k}{R_{k,t|t-1} + n_k} (\nu_{k,t|t-1} - \bar{y}_{k,t})(\nu_{k,t|t-1} - \bar{y}_{k,t})^T \tag{3.17l}$$

Given the previous analysis, the RBPF for Stick Breaking scheme-based dependent DP mixture model can be summarized as follows.

Algorithm 3.2: Rao-Blackwellized particle filter for stick breaking scheme-based dependent DP mixture

- At $t = 1$, for each particle $j = 1, ..., J$, set the truncated values of K for TDP mixture and initialize $\left\{V_{k,0}^{(j)}, \mu_{k,0}^{(j)}, \Sigma_{k,0}^{(j)}, \kappa_{k,0}^{(j)}, \beta_{k,0}^{(j)}\nu_{k,0}^{(j)}, R_{k,0}^{(j)}, d_{k,0}^{(j)}, S_{k,0}^{(j)}\right\}_{k=1}^{K}$. Obtain allocation variable $\mathbf{c}_1^{(j)}$ by generic blocked Gibbs sampling (Ishwaran and James, 2001, and Appendix A). Evaluate $\left\{\kappa_{k,1}^{(j)}, \beta_{k,1}^{(j)}, \nu_{k,1}^{(j)}, R_{k,1}^{(j)}, d_{k,1}^{(j)}, S_{k,1}^{(j)}\right\}_{k=1}^{K}$ via (3.16) and (3.17), draw posterior sample $\left\{V_{k,1}^{(j)}, \mu_{k,1}^{(j)}, \Sigma_{k,1}^{(j)}\right\}_{k=1}^{K}$, and evaluate $\left\{\pi_{k,1}^{(j)}\right\}_{k=1}^{K}$.

- For $t > 1$:

  - For each particle $j = 1, ..., J$,

    * Predict: evaluate $\Psi_{t|t-1}^{(j)} = \{\kappa_{k,t|t-1}^{(j)}, \beta_{k,t|t-1}^{(j)}, \nu_{k,t|t-1}^{(j)}, R_{k,t|t-1}^{(j)}, d_{k,t|t-1}^{(j)}, S_{k,t|t-1}^{(j)}\}_{k=1}^{K}$ via equation (3.16), and evaluate $\left\{\pi_{k,t|t-1}^{(j)}\right\}_{k=1}^{K}$.

    * Sample $\mathbf{c}_t^{(j)} \sim q(\cdot|\pi_{t|t-1}^{(j)}, \Psi_{t|t-1}^{(j)}, \mathbf{x}_t)$ via blocked Gibbs sampling.

    * Update: evaluate $\Psi_t^{(j)} = \left\{\kappa_{k,t}^{(j)}, \beta_{k,t}^{(j)}, \nu_{k,t}^{(j)}, R_{k,t}^{(j)}, d_{k,t}^{(j)}, S_{k,t}^{(j)}\right\}_{k=1}^{K}$ via equation (3.17).

    * Draw posterior sample $\left\{V_{k,t}^{(j)}, \mu_{k,t}^{(j)}, \Sigma_{k,t}^{(j)}\right\}_{k=1}^{K}$ given $\Psi_t^{(j)}$, and evaluate $\left\{\pi_{k,t}^{(j)}\right\}_{k=1}^{K}$.

  - Compute the importance weights

$$\widetilde{w}_t^{(j)} \propto w_{t-1}^{(j)} \frac{p(\mathbf{x}_t|\mathbf{c}_t^{(j)})p(\mathbf{c}_t^{(j)}|\pi_{t|t-1}^{(j)})}{q(\mathbf{c}_t^{(j)}|\pi_{t|t-1}^{(j)}, \Psi_{t|t-1}^{(j)}, \mathbf{x}_t)},$$

    with $\sum_{j=1}^{N} \widetilde{w}_t^{(j)} = 1$.

  - Resample when ESS is less than a predefined threshold.

3.3.2 Density Estimation

This section discuss the smoothing and predictive density estimation related to the above particle filtering for dependent DP mixtures. One primary goal of our analysis is to obtain density estimates which facilitate borrowing information across time and predict the shape of the density at future periods.

After the RBPF has been performed through the entire dataset, we get an approximate representation of $f_t(\cdot|\mathbf{x}_{1:t})$ for each time step $t \in 1, ..., T$, consisting of weighted particles $\{(\boldsymbol{\pi}_t, \boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)^{(j)}, \widetilde{w}_t^{(j)}; j = 1, 2, ..., J\}$. Filtered density estimation is given by

$$\hat{f}_t(\cdot|\mathbf{x}_{1:t}) = E\left[\int N(\cdot|\mu_t, \Sigma_t) G_t(d\mu_t, d\Sigma_t)|\mathbf{x}_{1:t}\right] \tag{3.18}$$

$$= \int N(\cdot|\mu_t, \Sigma_t) E\left[G_t(d\mu_t, d\Sigma_t)|\mathbf{x}_{1:t}\right] \tag{3.19}$$

Hence, given the weighted samples $\{(\boldsymbol{\pi}_t, \boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)^{(j)}, \widetilde{w}_t^{(j)}; j = 1, 2, ..., J\}$, the posterior estimate of $f(\cdot)$ is

$$\hat{f}(\cdot) \approx \sum_{j=1}^{J} \widetilde{w}_t^{(j)} \sum_{k=1}^{K_t^{(j)}} \pi_{k,t}^{(j)} N(\cdot|\mu_{k,t}^{(j)}, \Sigma_{k,t}^{(j)}) \tag{3.20}$$

Also, k-step ahead density prediction, $\hat{f}_{t+k}(\cdot|\mathbf{x}_{1:t})$, can be obtained in a similar way.

3.4 Simulation study

This section presents a simulation study where we compare the performance of our dynamic density estimation model against some regular methods: kernel density

estimation and finite mixture via EM algorithm. The example is similar to examples studied in Caron et al. (2007) and Rodriguez and ter Horst (2008). The true model used to simulate the data corresponds to a sequence of 500 distributions, $\{f_t(\cdot)\}_{t=1}^{500}$, specified as

$$f_t(x) = \epsilon_t N(x| -2, .25) + (1 - \epsilon_t) N(x|\phi_t, .25), \tag{3.21}$$

where $\epsilon_t = 0.2 + 0.001t$ for $t = 1, ..., 500$ and $\phi_t = 0.99\phi_{t-1}$ $(t = 1, ..., 500)$ with $\phi_1 = 2$. Figure 3.1a shows the time-varying distributions $f_t(\cdot)$. For each simulation, we generate 20 observations from $f_t(\cdot)$ illustrated in Figure 3.1b. The small sample sizes introduce a complication to the density estimation process, and allow us to demonstrate 1) the advantages of borrowing information across time, and 2) that traditional estimation such as kernel density estimation and finite mixture via EM algorithm can be highly unreliable for small sample sizes.

In this study, we examine the stick breaking scheme-based dependent DP mixture and the RBPF algorithm for this model. Hyperparameters in the dependent DP mixture as well as control parameters in the RBPF algorithm are set as shown in Table 3.1. After the RBPF is applied on the synthetic data, we obtain the weighted samples $\{(\boldsymbol{\pi}_t, \boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)^{(j)}, w_t^{(j)}; j = 1, 2, ..., J\}$. The effective sample size shown in Figure 3.3 implies the efficiency of the proposed particle filtering. As discussed in Section 3.3.2, the $\{f_t(\cdot)\}_{t=1}^{500}$ can be estimated via

$$\hat{f}_t(\cdot) = \sum_{j=1}^{N} \tilde{w}_t^{(j)} \sum_{k=1}^{K_t^{(j)}} \pi_{k,t}^{(j)} J(\cdot|\mu_{k,t}^{(j)}, \Sigma_{k,t}^{(j)}) \tag{3.22}$$

This sequence of posterior estimates is shown in Figure 3.2. Comparing with the shape of the true distribution $f_t(\cdot)$, our method can rebuild the distribution well. To quantify the estimation error, we record the KL-divergence between true density functions and their estimates over each iteration; see Figure 3.2. For comparison, we show the KL-divergence between $f_t(\cdot)$ and two other regular methods: (a)

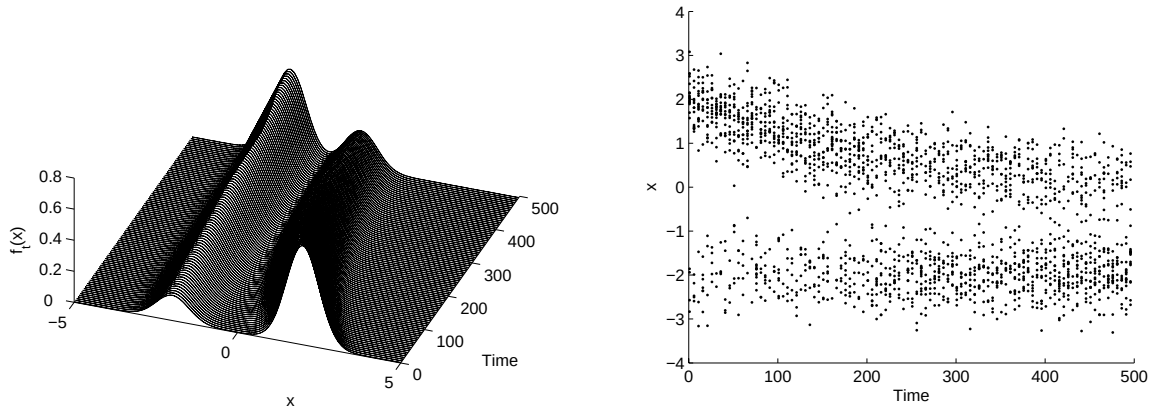
Parameter	Value
K	10
τ	10
d_0	1
S_0	**I**
J	100

Table 3.1: Parameter setting in dependent DP mixture model and RBPF filtering algorithm.

kernel density estimation shown in Figure 3.5a, (b) finite mixture (with two components) estimated by EM algorithm shown in Figure 3.5b. These estimates are highly unreliable compared to our method.

Figure 3.6 displays the posterior medians of the numbers of non-zero weights in the sequence of mixture components. This approximates the true values well and implies our proposed approach is reliable in rebuilding key aspects of the unknown distribution. Moreover, the combined parameter and state filtering approach proposed in Liu and West (2001) can be easily incorporated into our algorithm for parameter learning for the DP total mass parameter α, or parameters of the DLMs such as E_t, H_t, etc. In this study, we fixed the associated parameters in DLMs, and applied parameter learning only for α. Figure 3.7 shows the result of the parameter learning for α, which approaches a reasonable value 0.8 (experimental number of DP components $k \approx \alpha \log(N_t)$, Escobar and West, 1995).

To further demonstrate performance of our proposed method, a movie is created to show $\hat{f}_t(\cdot)$ over time, $f_t(\cdot)$ and histograms of the observed data. The movie is available at `http://stat.duke.edu/people/theses/JiC.html`. The last frame of the movie is shown in Figure 3.8.



(a) (b)

Figure 3.1: Synthetic data: (a) the true densities used to generate the data; (b) plots of synthetic data per iteration.

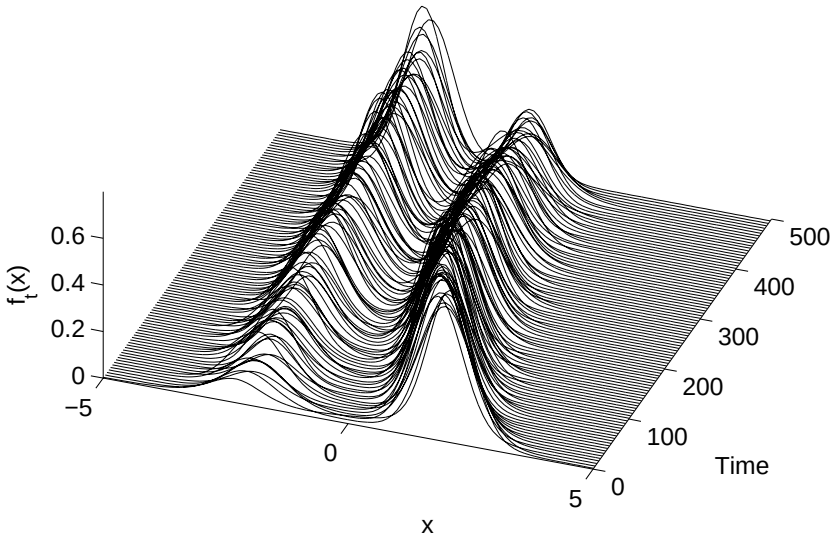


Figure 3.2: Posterior density estimates using RBPF.

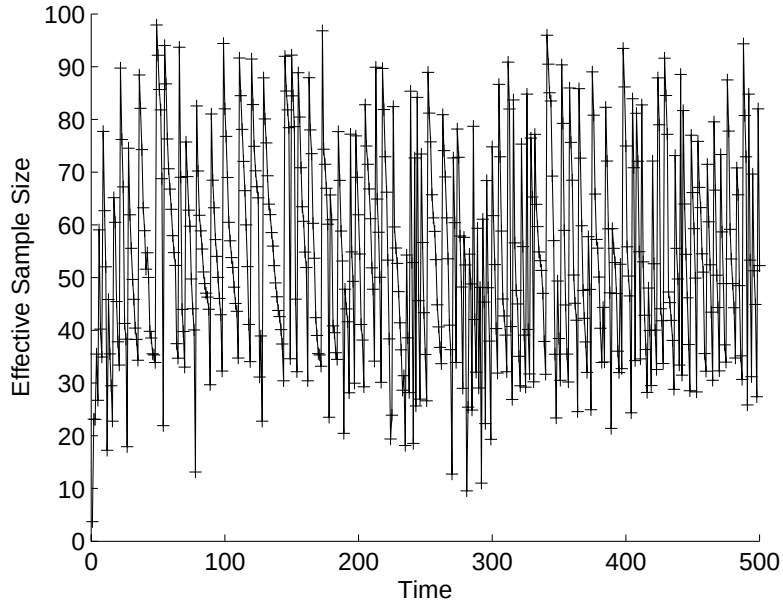


Figure 3.3: Effective sample size of Rao-Blackwellized particle filter (mean: 54.5, std: 18.4).

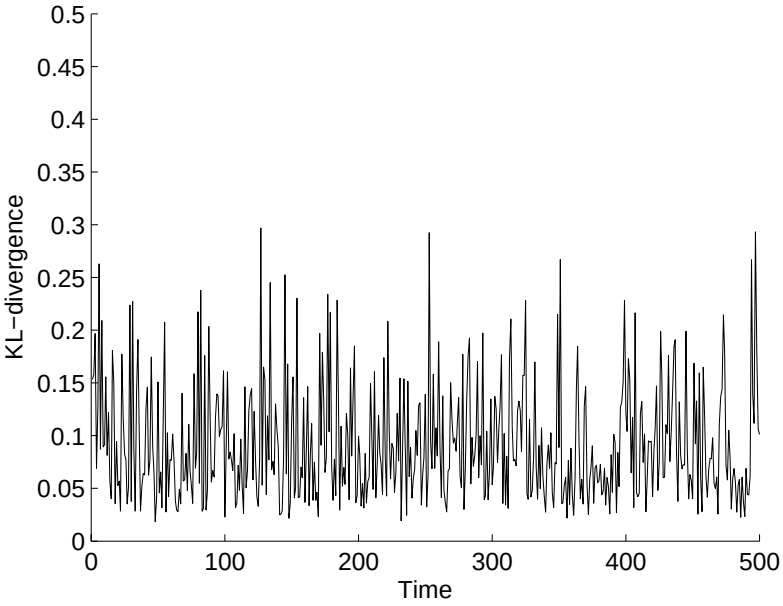
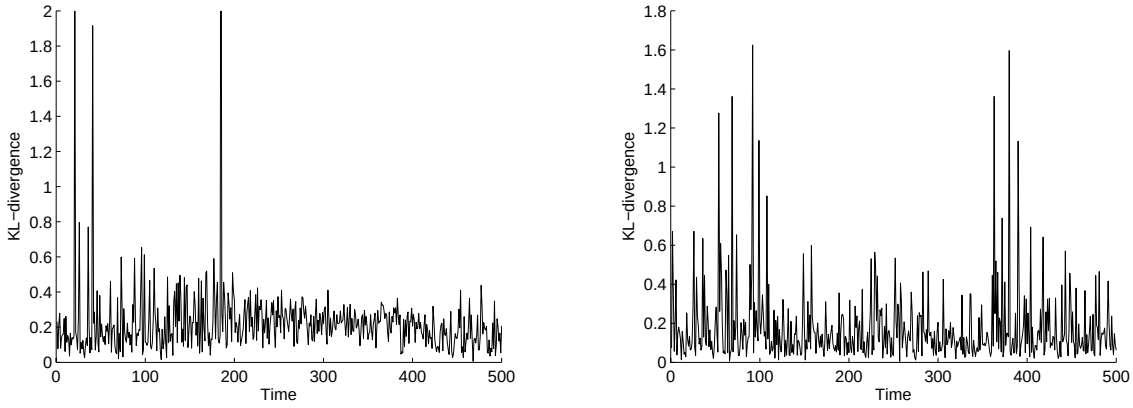


Figure 3.4: Plot of the KL-divergence between true density functions and RBPF-based posterior estimates.



(a) (b)

Figure 3.5: Plot of the KL-divergence between true density functions and two regular estimates: (a) kernel density estimation, (b) finite mixture estimated by EM algorithm.

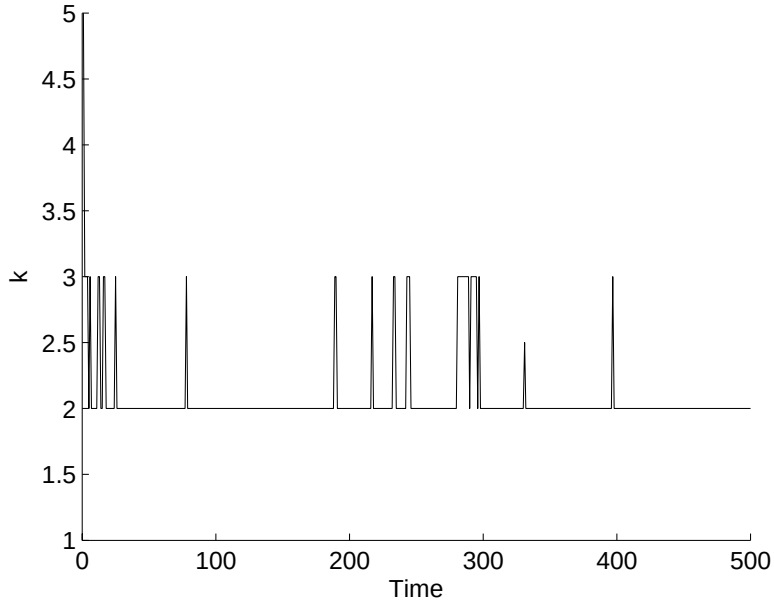


Figure 3.6: Plot of posterior median of the number of non-zero weights of mixture components.

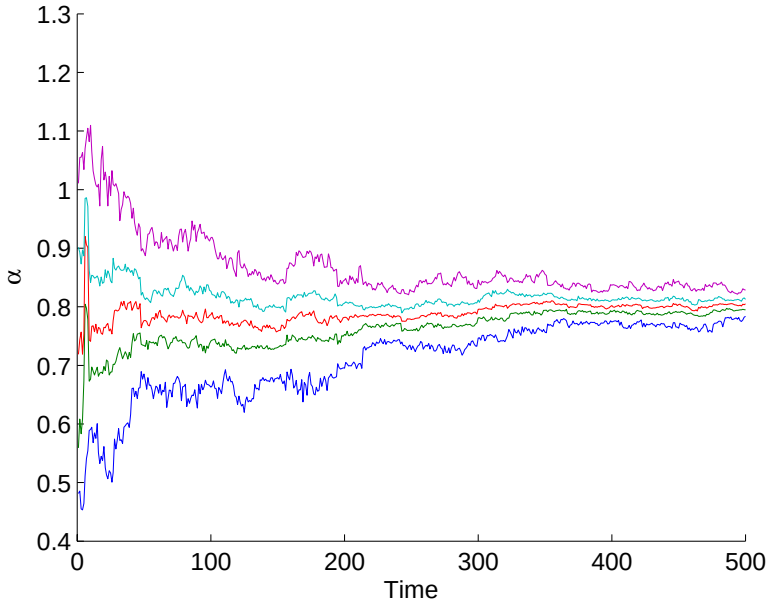


Figure 3.7: Trajectories of posterior quantiles (2.5%, 25%, 50%, 75%, 97.5%) of the posteriors for α.

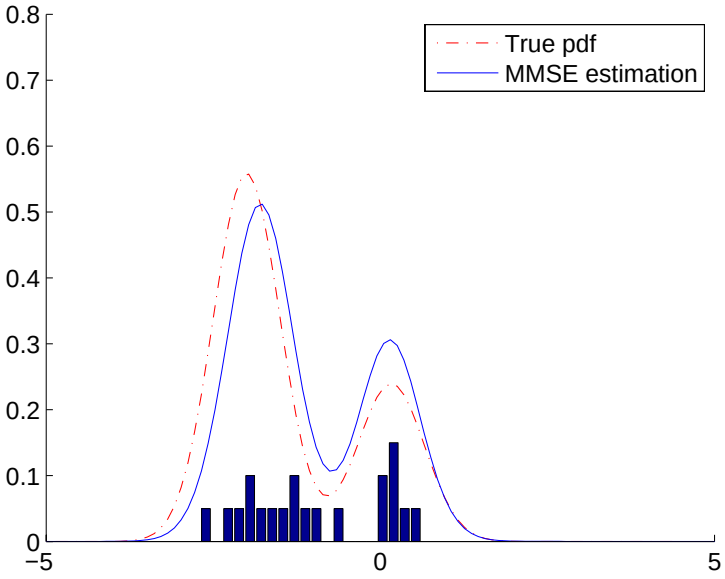


Figure 3.8: Last frame of the movie of estimates of densities $f_t(\cdot)$: true distribution shown by red dashed curve, posterior estimate shown by blue curve, observed data shown by blue histogram.

Chapter 4

Bayesian Nonparametric Modelling for Time-varying Spatial Point Processes

Using the models and methods of Chapter 3 as a basis, we now discuss flexible Bayesian nonparametric modelling for inhomogeneous spatio-temporal processes. This involves nonparametric spatial process mixture models of intensity functions, in which time variation is introduced via dynamic models for underlying parameters. These models characterize smooth dynamics in time in what may be quite complicated spatial patterns of spatial inhomogeneity in intensity functions. The framework is based on a time-varying dependent Dirichlet process, and physically attractive time propagation models for parameters of nonparametric mixture models for intensities. Bayesian inference and model fitting are addressed using novel particle filtering methods based on Chapter 3. Illustrative simulation examples in extended target tracking, and substantive data analysis in applications in cell fluorescent microscopic imaging tracking demonstrate analysis with these models.

4.1 Spatial Mixture Modelling for Dynamic Point Process

Chapter 2 introduces the mixture modelling for spatial point process, which enable us to characterize the intensity of a static spatial point process. In order to model the intensity of dynamic spatial points process, we use the model of Chapter 2 at each time point t, adding models of Chapter 3 to relate over time.

The model theory and notation of Chapter 2 is now used for each of a sequence

of time point $t = 1, 2, ..., T$. All parameters and data and processes are now additionally subscribed by t. At a particular time frame t in a dynamic spatial point process, we first specify a prior model for the intensity function $\lambda_t(\cdot)$, and then conduct posterior inference on $\lambda_t(\cdot)$ in light of the realized outcomes $x_{1:N,t}$. As in Chapter 2, we define the overall intensity scale parameter $\gamma_t = \int_{x_t \in S} \lambda_t(x)dx$ and the probability density (over $x_t \in S$) $f_t(x_t) = \lambda_t(x_t)/\gamma_t$. Given observed data $x_{1:N_t}$, the likelihood function can be expressed as

$$p(x_{1:N_t}|\gamma_t, f_t) \propto \exp(-\gamma_t)\gamma_t^{N_t} \prod_{i=1}^{N_t} f_t(x_{i,t}) \tag{4.1}$$

as a function of (γ_t, f_t).

We employ the Dirichlet process mixture framework in which the normalized intensity function $f_t(\cdot)$ is the density of a random mixture of normals, with mixing distribution distributed according to a Dirichlet Process $G \sim DP(\alpha, G_0)$ where $G_0(\cdot)$ is the prior mean of $G(\cdot)$ and $\alpha > 0$ the precision of the DP. We use the same (α, G_0) for all times t.

4.1.1 Dynamic Dirichlet Process Mixture Modelling

In many situations, like multi-target tracking (Gilholm et al., 2005) and cell fluorescent microscopic imaging tracking (Sigal et al., 2006; Gordon et al., 2007; Wang et al., 2009), the spatial point process is not static but evolves over time. We regards such process as inhomogeneous spatio temporal point processes (or dynamic/time-varying spatial point process). An intuitive way for modelling such dynamic spatial point processes is to introduce time dependence in the intensity function of succussive spatial point processes. However, directly modelling for a series of intensity functions with high spatial heterogeneity is generically infeasible. The alternative used here is to build a flexible nonparametric mixture model for the intensity and

introduce time dependence for underlying parameters in the mixture model, which therefore enables us to model the dynamics of the intensity function and characterize the spatial point process. This directly builds on our work in Chapter 3.

4.1.2 Dependent Dirichlet Process Mixture

We have the following nonparametric mixture model for $f_t(\cdot)$ with $\phi_{i,t} = (\mu_{i,t}), \Sigma_{i,t}$,

$$(x_{i,t}|\phi_{i,t}) \sim p(x_{i,t}|\phi_{i,t}), \quad (\phi_{i,t}|G_t) \sim G_t, \quad (G_t|\alpha, G_0) \sim DP(\alpha, G_0) \tag{4.2}$$

where G_t is a discrete-time dependent Dirichlet process as discussed in Chapter 3. We employ the stick breaking scheme-based dependent DP mixture because it involves the computationally effective approach, blocked Gibbs sampling, which is significantly faster than the collapsed Gibbs sampling used in the Polya Urn scheme-based model. The dependent Dirichlet process with stick breaking representation can be expressed as follows:

$$G_t = \sum_{k=1}^{K} \pi_{k,t}\delta(\mu_{k,t}, \Sigma_{k,t}) \tag{4.3a}$$

$$\pi_{k,t} = V_{k,t}\prod_{j=1}^{k-1}(1 - V_{j,t}) \tag{4.3b}$$

$$V_{k,t} \sim Beta(\cdot|\delta\kappa_{k,t-1}, \delta\beta_{k,t-1}) \tag{4.3c}$$

$$c_{i,t} \sim Mn(\cdot|\pi_{1,t}, ..., \pi_{K,t}) \tag{4.3d}$$

$$\overline{y}_{k,t} = H_t\mu_{k,t} + \varepsilon_{k,t}, \qquad \varepsilon_{k,t} \sim N(0, V_{k,t}\Sigma_{k,t}) \tag{4.3e}$$

$$\mu_{k,t} = E_t\mu_{k,t-1} + \epsilon_{k,t}, \qquad \epsilon_{k,t} \sim N(0, W_{k,t}\Sigma_{k,t-1}) \tag{4.3f}$$

$$\Sigma_{k,t} \sim IW(\cdot|\rho d_{k,t-1}, \rho S_{k,t-1}) \tag{4.3g}$$

with the priors

$$V_{k,0} \sim Beta(1,\alpha) \tag{4.4a}$$

$$\mu_{k,0} \sim N(0,\varrho\Sigma_{k,0}) \tag{4.4b}$$

$$\Sigma_{k,0} \sim IW(\cdot|d_0,S_0) \tag{4.4c}$$

The discount factor-based stochastic model for $V_{k,t}$ and cluster covariances $\Sigma_{k,t}$ remain the same as discussed in Chapter 3. However, a physically attractive time propagation model for cluster means is proposed instead of previous naive model discussed in Chapter 3.

Dynamic Model for Cluster Means

Previous studies (Caron et al., 2007) typically assumed stationary models for $\mu_{k,t}$ to obtain a first-order stationary DPM process. In Chapter 3, we used random walk models for $\mu_{k,t}$. However, such models are not suitable to represent non-stationary stochastic process. For example, in multi-target tracking and cell fluorescent microscopic imaging tracking, $\mu_{k,t}$ are utilized to represent the position of maneuvering targets or cells which may be highly non-linear non-stationary stochastic process.

Here we assume the dynamics of 'target' position and its first derivation (aka 'velocity'), $\boldsymbol{\mu}_{k,t} = [\mu_{k,t}, \dot{\mu}_{k,t}]^T$, evolve according to a near constant velocity model (Bar-Shalom and Fortmann, 1988), a physically attractive dynamic model, as follows:

$$\boldsymbol{\mu}_{k,t} = E_{k,t}\boldsymbol{\mu}_{k,t-1} + B\epsilon_{k,t} \tag{4.5}$$

$$= \begin{bmatrix} \mathbf{I}_d & \tau\mathbf{I}_d \\ \mathbf{0}_d & \mathbf{I}_d \end{bmatrix} \begin{bmatrix} \mu_{k,t} \\ \dot{\mu}_{k,t} \end{bmatrix} + \begin{bmatrix} \tau^2/2 \\ \tau \end{bmatrix} \epsilon_{k,t} \tag{4.6}$$

where d is the dimension of $\mu_{k,t}$, τ is the time interval between t and $t-1$ and $\epsilon_{k,t}$ is a zero-mean Gaussian distributed random vector with a covariance matrix

$W_{k,t}\Sigma_{k,t-1}$. Here $W_{k,t}$ is set equivalent to ϱ.

Given the realization of the allocation variable in the DP mixture model, we can then obtain the observation $\overline{y}_{k,t}$, the sufficient statistics for cluster mean k, i.e. the mean of all data points allocated to cluster k at time t. Then the measurement model for $\boldsymbol{\mu}_{k,t}$ is

$$\begin{aligned} \overline{y}_{k,t} &= H_t\boldsymbol{\mu}_{k,t} + \varepsilon_{k,t} && (4.7) \\ &= \begin{bmatrix} \mathbf{I}_d & \mathbf{0}_d \\ \mathbf{0}_d & \mathbf{0}_d \end{bmatrix} \begin{bmatrix} \mu_{k,t} \\ \dot{\mu}_{k,t} \end{bmatrix} + \varepsilon_{k,t} && (4.8) \end{aligned}$$

where $\varepsilon_{k,t} \sim N(0, V_{k,t}\Sigma_{k,t})$ with $V_{k,t} = 1/n_{k,t}$.

Assume the posterior of $\boldsymbol{\mu}_{k,t}$ at time $t-1$ is $N(\cdot|\nu_{k,t-1}, R_{k,t-1}\begin{bmatrix} \Sigma_{k,t-1} & \mathbf{0}_d \\ \mathbf{0}_d & \Sigma_{k,t-1} \end{bmatrix})$, then, we can process sequential inference for the posterior of $\boldsymbol{\mu}_{k,t}$,

$$N(\cdot|\nu_{k,t}, R_{k,t}\begin{bmatrix} \Sigma_{k,t-1} & \mathbf{0}_d \\ \mathbf{0}_d & \Sigma_{k,t-1} \end{bmatrix}),$$

via the following equations,

Predict:

$$\begin{aligned} \nu_{k,t|t-1} &= E_{k,t}\nu_{k,t-1} && (4.9a) \\ R_{k,t|t-1} &= E_{k,t}R_{k,t-1}E_{k,t}^T + \begin{bmatrix} \frac{\tau^4}{4}W_{k,t} & \frac{\tau^2}{2}W_{k,t} \\ \frac{\tau^2}{2}W_{k,t} & \tau W_{k,t} \end{bmatrix} && (4.9b) \end{aligned}$$

Update:

$$\nu_{k,t} = \begin{cases} \nu_{k,t|t-1} + A_{k,t}e_{k,t}, & \text{if } n_{k,t} > 0 \\ 0_{2d}, & \text{if } n_{k,t} = 0 \end{cases} \tag{4.10a}$$

$$R_{k,t} = \begin{cases} R_{k,t|t-1} - A_{k,t}Q_{k,t}A_{k,t}^T, & \text{if } n_{k,t} > 0 \\ \varrho\mathbf{I}_{2d}, & \text{if } n_{k,t} = 0 \end{cases} \tag{4.10b}$$

$$A_{k,t} = R_{k,t|t-1}H_{k,t}Q_{k,t}^{-1} \tag{4.10c}$$

$$e_{k,t} = \overline{y}_{k,t} - \tilde{y}_{k,t} \tag{4.10d}$$

$$\tilde{y}_{k,t} = H_{k,t}\nu_{k,t|t-1} \tag{4.10e}$$

$$Q_{k,t} = H_{k,t}R_{k,t|t-1}H_{k,t}^T + V_{k,t} \tag{4.10f}$$

4.1.3 Likelihood Function for Inhomogeneous Poisson Process

At each time step t a set or frame of measurements $\mathbf{x}_t = \{x_{1,t}, ..., x_{N_t,t}\}$ becomes available. Each of these N_t measurements originates from one entity (i.e. 'target' or cell). The totality of measurements $\mathbf{x}_t$ received over the observation region at a particular time step t can be treated as an inhomogeneous (or nonhomogeneous) Poisson point process with the likelihood function

$$p(\mathbf{x}_t|\gamma, f_t) \propto \exp(-\gamma_t)\gamma^{N_t}\prod_{i=1}^{N_t} f_t(x_{i,t}) \tag{4.11}$$

where γ_t is relevant to the expected total number of measurements received in current time frame. The density function $f_t(\cdot)$ is modelled by the dependent DP mixture model as discussed above. Moreover, due to the factor that γ_t can be cancelled in calculating the sample weights in sequential Monte Carlo implementations, thus, likelihood of the observation is *precisely* the likelihood that would arise from simple random sampling from $f_t(\cdot)$ generating data $x_{1:N,t}$.

4.2 Sequential Monte Carlo Implementation

Bayesian inference (filtering) for the unknown state parameter $\Phi = \{\mathbf{c}_t, \boldsymbol{V}_t, \boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t\}$ in the dependent DP mixture model can be expressed as follows:

$$p(\boldsymbol{\Phi}_t|\mathbf{x}_{t-1}) = \int p(\boldsymbol{\Phi}_t|\boldsymbol{\Phi}_{t-1})p(\boldsymbol{\Phi}_{t-1}|\mathbf{x}_{t-1})d\boldsymbol{\Phi}_{t-1} \tag{4.12}$$

$$p(\boldsymbol{\Phi}_t|\mathbf{x}_t) \propto p(\mathbf{x}_t|\boldsymbol{\Phi}_t)p(\boldsymbol{\Phi}_t|\mathbf{x}_{t-1}) \tag{4.13}$$

In general, due to the complexity of the system equation $p(\boldsymbol{\Phi}_t|\boldsymbol{\Phi}_{t-1})$, closed-form sequential inference for $p(\boldsymbol{\Phi}_t|\mathbf{x}_t)$ is analytically intractable. We therefore turn to sequential Monte Carlo (SMC) methods.

4.2.1 Rao-Blackwellized Particle Filter

As discussed in Chapter 3, to overcome the problem of extremely low effective sample size in sequential Monte carlo for dependent DP mixture, we utilize the Rao-Blackwellized technique via the two steps as follows:

Predict:

$$\kappa_{k,t|t-1} = \delta\kappa_{k,t-1} \tag{4.14a}$$

$$\beta_{k,t|t-1} = \delta\beta_{k,t-1} \tag{4.14b}$$

$$\nu_{k,t|t-1} = E_{k,t}\nu_{k,t-1} \tag{4.14c}$$

$$R_{k,t|t-1} = E_{k,t}R_{k,t-1}E_{k,t}^T + \begin{bmatrix} \frac{\tau^4}{4}W_{k,t} & \frac{\tau^2}{2}W_{k,t} \\ \frac{\tau^2}{2}W_{k,t} & \tau W_{k,t} \end{bmatrix} \tag{4.14d}$$

$$d_{k,t|t-1} = \rho d_{k,t-1} \tag{4.14e}$$

$$S_{k,t|t-1} = \rho S_{k,t-1} \tag{4.14f}$$

Update:

$$\kappa_{k,t} = \begin{cases} \kappa_{k,t|t-1} + n_{k,t}, & \text{if } n_{k,t} > 0 \\ 1, & \text{if } n_{k,t} = 0 \end{cases} \tag{4.15a}$$

$$\beta_{k,t} = \begin{cases} \beta_{k,t|t-1} + \sum_{r=k+1}^{K} n_{r,t}, & \text{if } n_{k,t} > 0 \\ \alpha, & \text{if } n_{k,t} = 0 \end{cases} \tag{4.15b}$$

$$\nu_{k,t} = \begin{cases} \nu_{k,t|t-1} + A_{k,t}e_{k,t}, & \text{if } n_{k,t} > 0 \\ 0_{2d}, & \text{if } n_{k,t} = 0 \end{cases} \tag{4.15c}$$

$$R_{k,t} = \begin{cases} R_{k,t|t-1} - A_{k,t}Q_{k,t}A_{k,t}^T, & \text{if } n_{k,t} > 0 \\ \varrho\mathbf{I}_{2d}, & \text{if } n_{k,t} = 0 \end{cases} \tag{4.15d}$$

$$d_{k,t} = \begin{cases} d_{k,t|t-1} + n_{k,t}, & \text{if } n_{k,t} > 0 \\ d_0, & \text{if } n_{k,t} = 0 \end{cases} \tag{4.15e}$$

$$S_{k,t} = \begin{cases} S_{k,t|t-1} + \Omega, & \text{if } n_{k,t} > 0 \\ S_0, & \text{if } n_{k,t} = 0 \end{cases} \tag{4.15f}$$

$$A_{k,t} = R_{k,t|t-1}H_{k,t}Q_{k,t}^{-1} \tag{4.15g}$$

$$e_{k,t} = \overline{y}_{k,t} - \tilde{y}_{k,t} \tag{4.15h}$$

$$\tilde{y}_{k,t} = H_{k,t}\nu_{k,t|t-1} \tag{4.15i}$$

$$Q_{k,t} = H_{k,t}R_{k,t|t-1}H_{k,t}^T + V_{k,t} \tag{4.15j}$$

$$\Omega = \sum_{i=1}^{n_k} \left(y_{k,t}^{(i)} - \bar{y}_{k,t}\right)\left(y_{k,t}^{(i)} - \bar{y}_{k,t}\right)^T \tag{4.15k}$$

$$+\frac{R_{k,t|t-1}n_k}{R_{k,t|t-1} + n_k}(\nu_{k,t|t-1} - \bar{y}_{k,t})(\nu_{k,t|t-1} - \bar{y}_{k,t})^T \tag{4.15l}$$

Given the previous analysis, the RBPF for dependent DP mixture model for dynamic spatial point process application can be summarized as follows.

Algorithm 4.1: Rao-Blackwellized particle filter for dynamic spatial point process

- At $t = 1$, for each particle $j = 1, ..., J$, set the truncated values K for TDP mixture and initialize $\left\{V_{k,0}^{(j)}, \mu_{k,0}^{(j)}, \Sigma_{k,0}^{(j)}, \kappa_{k,0}^{(j)}, \beta_{k,0}^{(j)}\nu_{k,0}^{(j)}, R_{k,0}^{(j)}, d_{k,0}^{(j)}, S_{k,0}^{(j)}\right\}_{k=1}^{K}$. Obtain allocation variable $\mathbf{c}_1^{(j)}$ by blocked Gibbs sampling. Evaluate $\{\kappa_{k,1}^{(j)}, \beta_{k,1}^{(j)}, \nu_{k,1}^{(j)}, R_{k,1}^{(j)}, d_{k,1}^{(j)}, S_{k,1}^{(j)}\}_{k=1}^{K}$ via (4.14 and 4.15). Draw posterior sample $\left\{V_{k,1}^{(j)}, \mu_{k,1}^{(j)}, \Sigma_{k,1}^{(j)}\right\}_{k=1}^{K}$, and evaluate $\left\{\pi_{k,1}^{(j)}\right\}_{k=1}^{K}$.

- For $t > 1$

  - For each particle $j = 1, ..., J$

    * Predict: evaluate $\Psi_{t|t-1}^{(j)} = \{\kappa_{k,t|t-1}^{(j)}, \beta_{k,t|t-1}^{(j)}, \nu_{k,t|t-1}^{(j)}, R_{k,t|t-1}^{(j)}, d_{k,t|t-1}^{(j)}, S_{k,t|t-1}^{(j)}\}_{k=1}^{K}$ via equation (4.14), and evaluate $\left\{\pi_{k,t|t-1}^{(j)}\right\}_{k=1}^{K}$.

    * Sample $\mathbf{c}_t^{(j)} \sim q(\cdot|\pi_{t|t-1}^{(j)}, \Psi_{t|t-1}^{(j)}, \mathbf{x}_t)$ via blocked Gibbs sampling.

    * Update: evaluate $\Psi_t^{(j)} = \left\{\kappa_{k,t}^{(j)}, \beta_{k,t}^{(j)}, \nu_{k,t}^{(j)}, R_{k,t}^{(j)}, d_{k,t}^{(j)}, S_{k,t}^{(j)}\right\}_{k=1}^{K}$ via equation (4.15).

    * Draw posterior sample $\left\{V_{k,t}^{(j)}, \mu_{k,t}^{(j)}, \Sigma_{k,t}^{(j)}\right\}_{k=1}^{K}$ given $\Psi_t^{(j)}$, and evaluate $\left\{\pi_{k,t}^{(j)}\right\}_{k=1}^{K}$.

  - Compute the importance weights

$$\widetilde{w}_t^{(j)} \propto w_{t-1}^{(j)} \frac{p(\mathbf{x}_t|\mathbf{c}_t^{(j)})p(\mathbf{c}_t^{(j)}|\pi_{t|t-1}^{(j)})}{q(\mathbf{c}_t^{(j)}|\pi_{t|t-1}^{(j)}, \Psi_{t|t-1}^{(j)}, \mathbf{x}_t)},$$

    with $\sum_{j=1}^{N} \widetilde{w}_t^{(j)} = 1$.

  - Resample when ESS is less than a predefined threshold.

4.2.2 Presentation of Estimation Results

We present two approaches to utilize the SMC samples. The first, approximate posterior mean estimates of the functions $f_t(\cdot)$, are of interest in most applications. The second, MAP sequence estimates may be preferable in tracking applications when inference on trajectories.

Posterior Mean Estimation

The RBPF gives the output of weighted particles $\{(\boldsymbol{\pi}_t, \boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)^{(j)}, \widetilde{w}_t^{(j)}; j = 1, 2, ..., J\}$ to approximate $p(\boldsymbol{\pi}_t, \boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t|\mathbf{x}_{1:t})$ for each time step $t \in 1, ..., T$. Filtered density estimation is given by

$$\hat{f}_t(\cdot|\mathbf{x}_{1:t}) = E\left[\int N(\cdot|\mu_t, \Sigma_t) G_t(d\mu_t, d\Sigma_t)|\mathbf{x}_{1:t}\right] \tag{4.16}$$

$$= \int N(\cdot|\mu_t, \Sigma_t) E\left[G_t(d\mu_t, d\Sigma_t)|\mathbf{x}_{1:t}\right]. \tag{4.17}$$

Given $\{(\boldsymbol{\pi}_t, \boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)^{(j)}, \widetilde{w}_t^{(j)}; j = 1, 2, ..., J\}$, the mean estimate for $f(\cdot)$ can be approximated via

$$\hat{f}(\cdot) \approx \sum_{j=1}^{J} \widetilde{w}_t^{(j)} \sum_{k=1}^{K_t^{(j)}} \pi_{k,t}^{(j)} N(\cdot|\mu_{k,t}^{(j)}, \Sigma_{k,t}^{(j)}). \tag{4.18}$$

Also, k-step ahead density predictions, $\hat{f}_{t+k}(\cdot|\mathbf{x}_{1:t})$, can be obtained in a similar way.

Maximum a Posteriori Sequence Estimation

As is well known, the choice of mean, MAP or other estimates is dependent on the demands of the application and the inherent choice of loss functions, whether made explicit or not. In tracking applications, the trajectory is always of interest. If we select only one sample in each time step and link them together then we can

evaluate a posterior draw from the trajectory. Here we appeal to the MAP Sequence Estimation method proposed in Godsill et al. (2001) which can efficiently computes the optimal trajectory over all combinations of the filtered states. The methods rely on a particle cloud representation of the filtering distribution which evolves through time using an SMC method. MAP sequence estimation is then performed using a classical dynamic programming technique, the Viterbi algorithm (Viterbi, 1967), applied to the discretised version of the state space (Godsill et al., 2001). Details of the MAP sequential estimation method is presented in Appendix B.

4.3 Applications

This section discusses some applications of the dynamic spatial point point process and demonstrates the modelling and algorithm.

4.3.1 Multiple Extended Target Tracking

In classical target tracking problems, it is often assumed that at most a single measurement is received from a point target at each time step. However, in many cases, high resolution sensors are able to resolve individual features on an extended object. A straightforward way to address this problem is to model the target as a set of point sources, each of which may be the origin of a sensor measurement. In a previous study of Gilholm et al. (2005), the authors represent the measurements over the sensor observation region as a spatial non-homogeneous Poisson point process, where multiple measurements/points may originate from an entity (target). It also shown that this leads to an exact expression for the overall measurement likelihood that does not involve explicit assignments between the measurements and the entities. In this work, we treat the multiple extended target tracking problem as a dynamic spatial point process, and model it by the rigorous Bayesian model

discussed in the previous section. Generally speaking, the components in the mixture model are used to represent the target. Dynamics in the mixture components can capture the dynamics of each target, while the nonparametric setting can capture scenarios like 'birth', 'death', 'split','merge' which are difficult but common in multi-target tracking.

Illustrative simulation examples

The synthetic data set is generated as follows. At initiation stage, 5 extended targets are shown in the observation region; each of these targets generates 20 random points around its location. At time steps 10, 20, 50 one more target emerges in the observation region which also generates 20 points in consequential time steps. The overall observations are shown on both x and y coordinates over time in Figure 4.1.

Hyperparameters in the dynamic DP mixture model as well as control parameters in the RBPF algorithm are set as shown in Table 4.1. After applying the filtering algorithm, we obtain the weighted samples $\{(\boldsymbol{\pi}_t, \boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)^{(j)}, w_t^{(j)}; j = 1, 2, ..., J\}$. Using the MAP sequence estimation discussed in Section 4.2.2, we can obtain a sequence of mixtures $\{(\boldsymbol{\pi}_t, \boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)^{(j^*)}\}_{t=1}^{100}$, from which we can reconstructed the trajectory of each target by linking the mixture component with the same label in adjacent time steps. The reconstructed trajectories of the central locations of each extended target are shown in Figure 4.1. The underling intensity functions are estimated via

$$\hat{f}_t(\cdot) = \sum_{j=1}^{N} \tilde{w}_t^{(j)} \sum_{k=1}^{K_t^{(j)}} \pi_{k,t}^{(j)} N(\cdot|\mu_{k,t}^{(j)}, \Sigma_{k,t}^{(j)}). \tag{4.19}$$

The spatial intensity functions (true values and estimates) in each coordinate are shown in Figure 4.3. The posterior medians of the numbers of non-zero weights of mixture components is shown in Figure 4.2. It can be observed that our pro-

Parameter	Value
K	20
ρ	100
d_0	2
S_0	$\mathbf{I}$
τ	1
J	100

Table 4.1: Parameter setting in dependent DP mixture model and RBPF filtering algorithm.

posed approach efficiently solves the problem of tracking multiple extended targets simultaneously and modelling the dynamics of the underling intensity function.

To further demonstrate performance of our proposed method, three movies are created to show $\hat{f}_t(\cdot)$ over time: 1) plot of MAP sequence estimation of target location and trajectory; 2) image plot of posterior means of spatial intensities; 3) 3D plot of posterior means of spatial intensities with MAP sequence estimates of target trajectories. These movies are available at `http://stat.duke.edu/people/theses/JiC.html`. The last frame of these movies are shown in Figure 4.4 and 4.5.

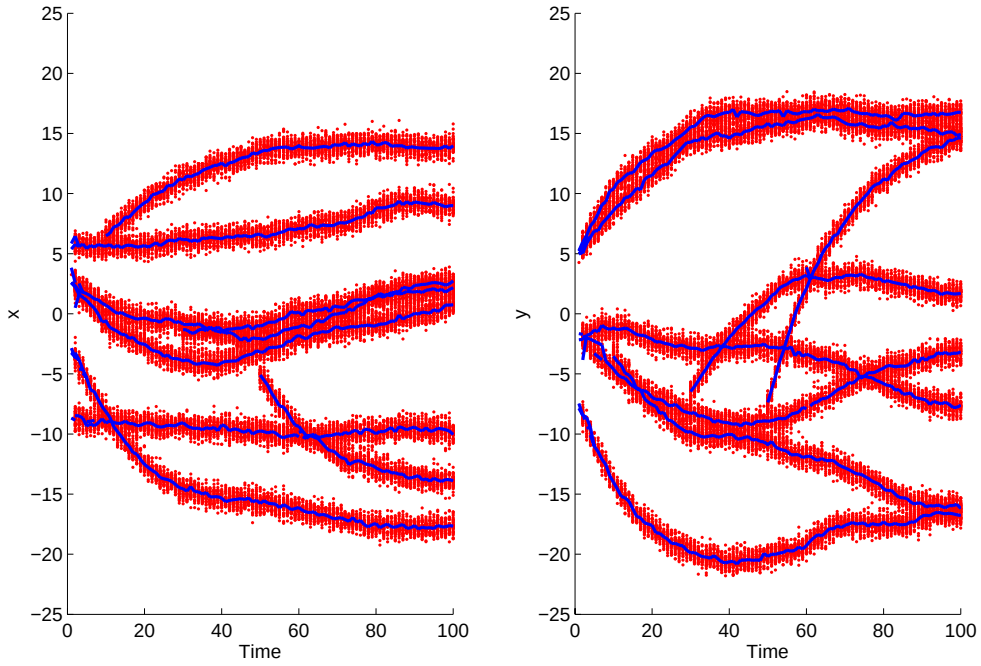


Figure 4.1: Synthetic dynamic spatial point process shown in both x and y coordinate, and reconstructed trajectories of each 'extended target': red dots represent observations and blue curves represent target trajectories.

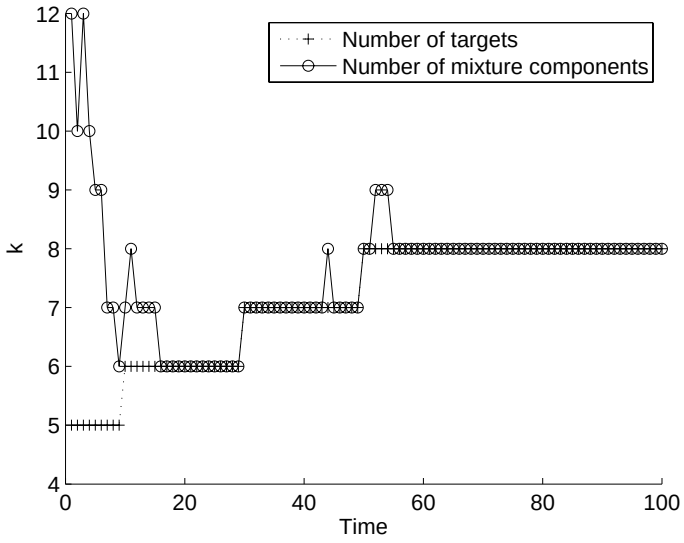
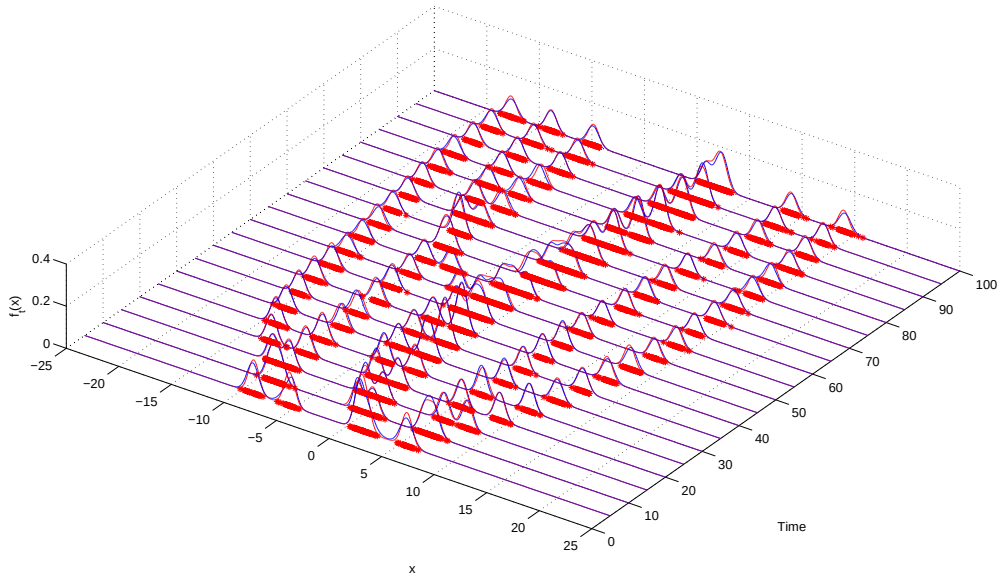
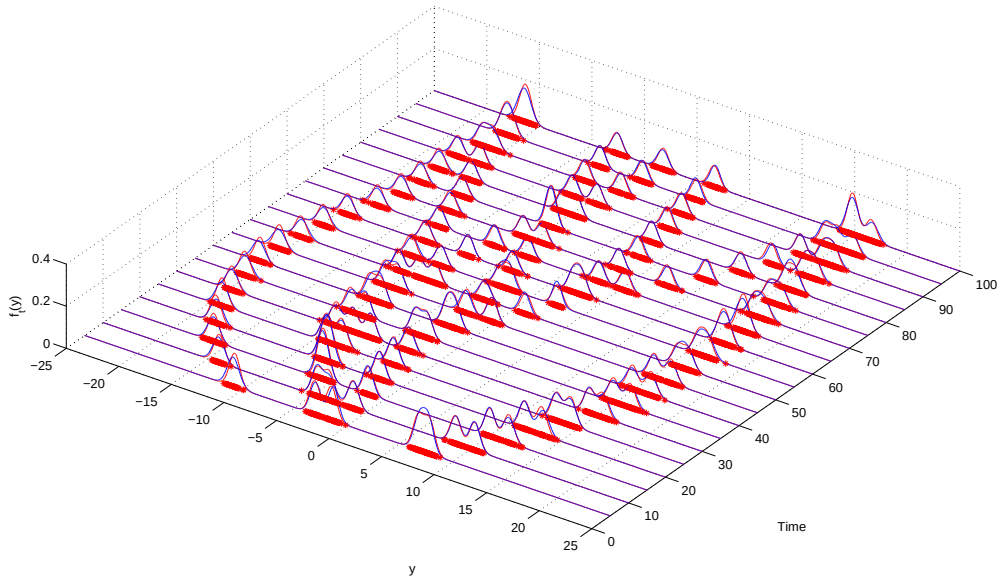


Figure 4.2: Plot of exact number of targets and posterior median of the number of non-zero mixture component weights.



(a)



(b)

Figure 4.3: Plots of spatial intensity functions in each coordinate: the true spatial intensity is shown by the red curve, the posterior means of the spatial intensity functions are shown by blue curves, and the spatial point pattern is shown by red dots. (a) plots in coordinate x, (b) plots in coordinate y.

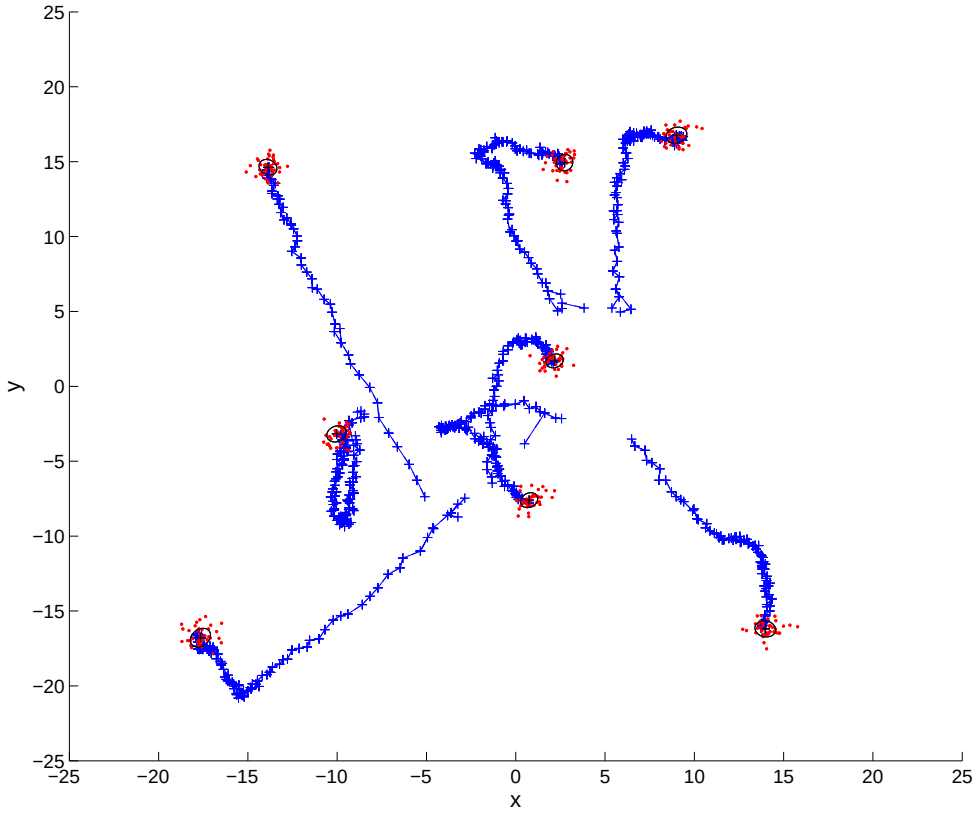
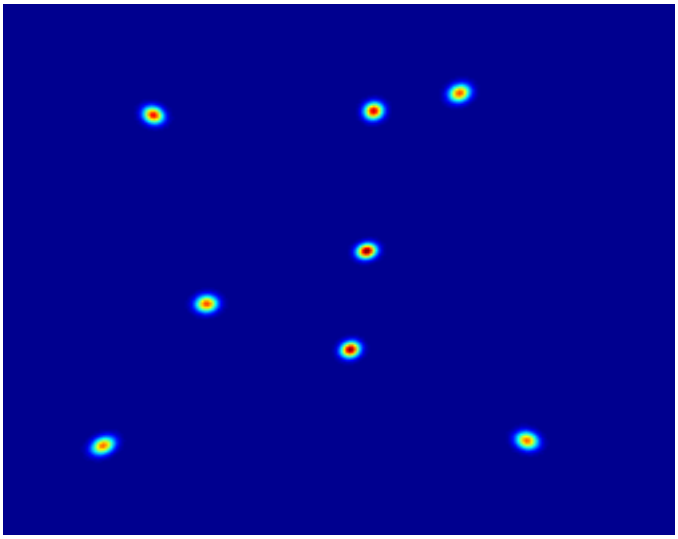
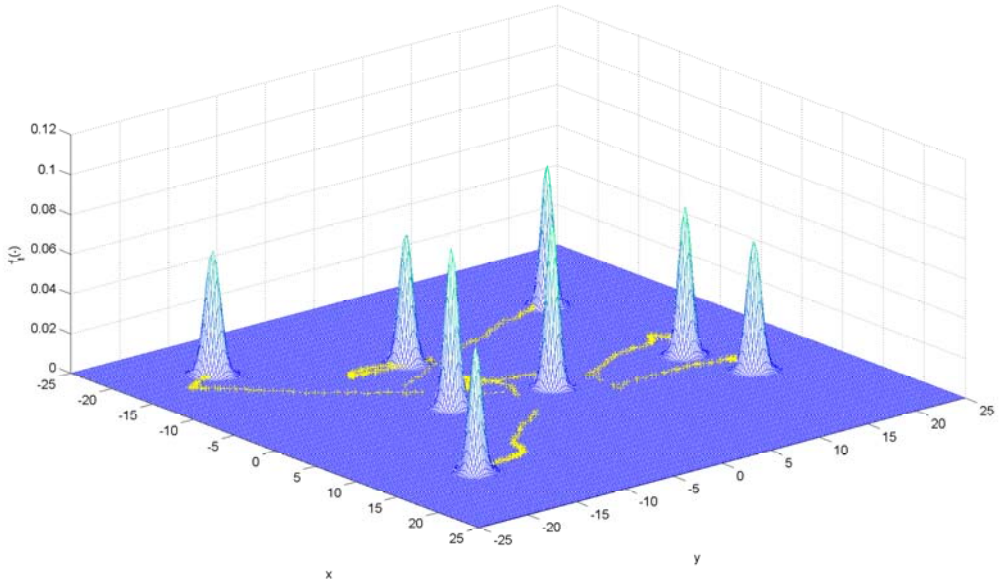


Figure 4.4: Last frame of the movie of target trajectory: target observations are shown by red dots, MAP sequence estimation of target position is represented by + (mean) and ellipse (standard deviation), and target trajectories are shown by blue curves.



(a)



(b)

Figure 4.5: Last frame of the movie of spatial intensity: (a) image plot of the posterior mean of spatial intensity; (b) 3D plot of the posterior mean of spatial intensity as well as MAP sequence estimates of target trajectories shown by yellow curve.

4.3.2 Cell Tracking

Recently, emerging time-lapse microscopy technologies allow detailed data generation on dynamic cellular processes at the single cell level (Megason and Fraser, 2007; Longo and Hasty, 2006). This technology has been broadly applied to investigate various biological problems such as biological noise in the dynamics of gene regulation, competence pathways, cell growth and proliferation (Elowitz et al., 2002; Raser and O' Shea, 2004; Rosenfeld et al., 2005; Levine and Davidson, 2005).

In using time-lapse microscopy technologies, we extract sequences of image frames from the time-lapse movies which can produce sufficiently high temporal and spatial resolution allowing time courses of gene expression dynamics at the single-cell level. These time courses are of critical importance in investigating the dynamics of cellular networks. However, extracting cells as objects in images, and tracking them in sequential images, is one of central technical challenges for single-cell fluorescent microscopy studies in systems biology (Megason and Fraser, 2007). In previous study, several mathematical and statistical models have been proposed to model such dynamic systems (Kask et al., 1999; Golding et al., 2005; Rosenfeld et al., 2006), but no rigorous statistical model has been proposed for modelling of the dynamics of multiple cells from tracking perspective.

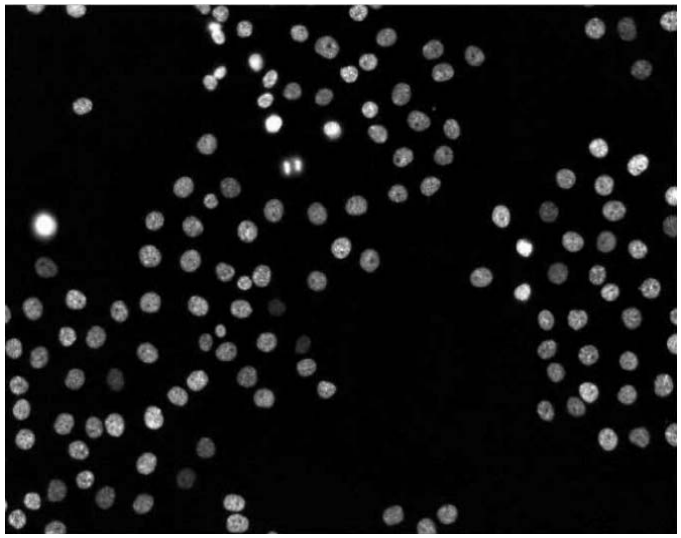
In this work, we provide the dynamic spatial DP mixture model for tracking the cells in fluorescence microscopic image, where we extracted the cells from the sequence image as dynamic spatial point processes.

Simulation study

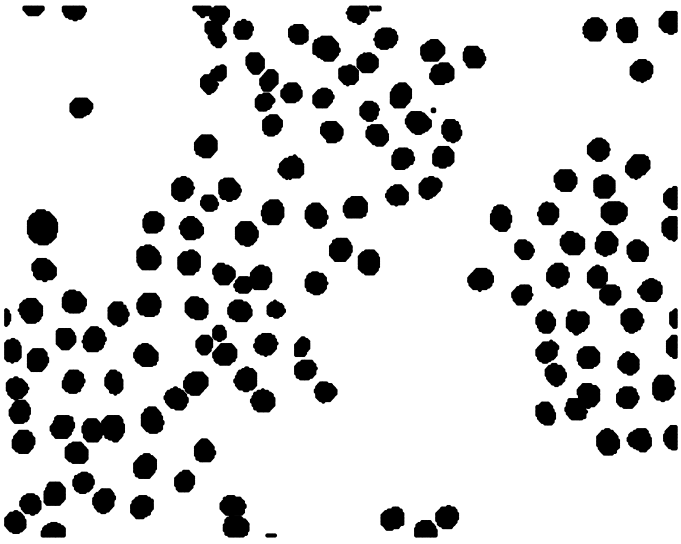
Our experiment data is a movie consisting 497 frames of cell fluorescence microscopic images. Since our proposed model has to deal with spatial point process, we need to generated the spatial point process corresponding to the location of

cells from each individual image. In previous study (Wang et al., 2009), image segmentation technique has been proposed, which can be applied to extract the pixels representing the location of cells. However, based on exploration of the experimentation data, we find that a simple likelihood-ratio test on the fluorescence intensity of image can generate the spatial location of points representing cells, which can serve as the observation of the spatial point process. Figure 4.6a and 4.6b show original data of a cell fluorescence microscopic image at last time frame of the movie and its corresponding spatial point pattern generated by the likelihood-ratio test (Wang et al., 2008, 2009).

Given the observations, it is straightforward to apply our proposed modelling and computation approach on the image sequence, to obtain the posterior of dynamic spatial intensities. Figure 4.7 show the MAP sequence estimate of the spatial intensity function, while posterior means are shown in Figure 4.8. As can be seen, the intensity function has high spatial inhomogeneity and can be well captured by our proposed model. The quantitative analysis of overall intensities of spatial locations of cells may therefore facilitate understanding of its system dynamics. Bedsides, tracking of individual cell is also of interest, particularly if we want to rebuild its lineage tree. To demonstrate the ability of our proposed method in dealing with individual cell tracking, we generate a movie by linking the MAP sequence estimation of the intensity function. This movie is available at `http://stat.duke.edu/people/theses/JiC.html`, the last frame of which is shown in Figure 4.9. As can be seen in the movie, each individual cell is captured by one or a few mixture components. Wit labels of the mixture components, we can identify and track each cell in adjacent time frame, and link them to reconstruct the trajectory of each cell as shown in Figure 4.10.



(a)



(b)

Figure 4.6: Human cell imaging data: (a) original data of cell fluorescence microscopic image at last time step, (b) spatial point pattern generated by the image segmentation at last time step.

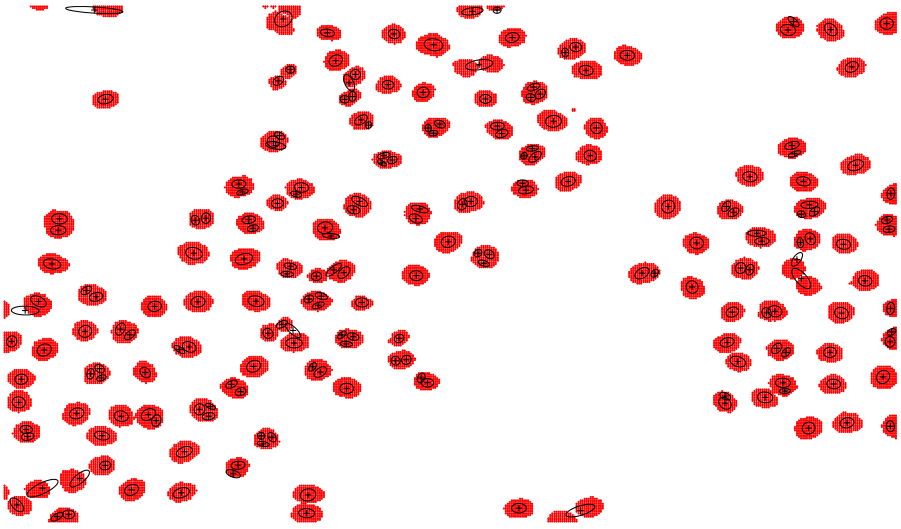
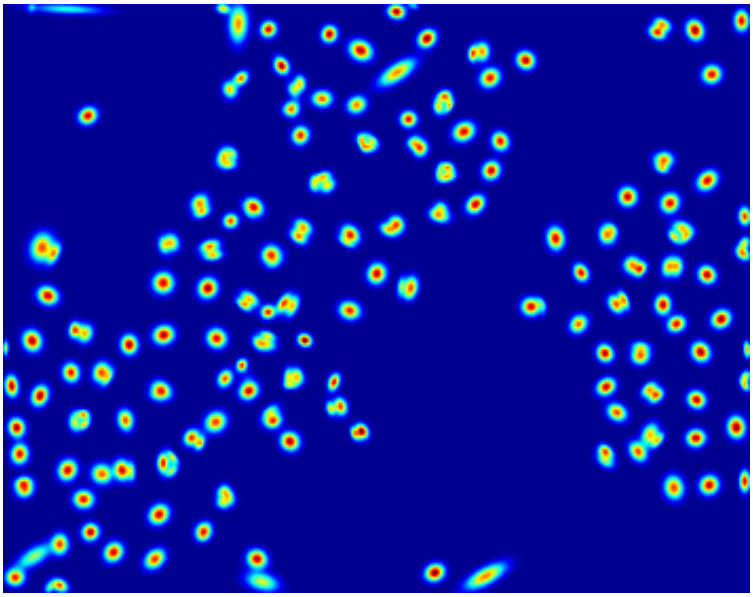
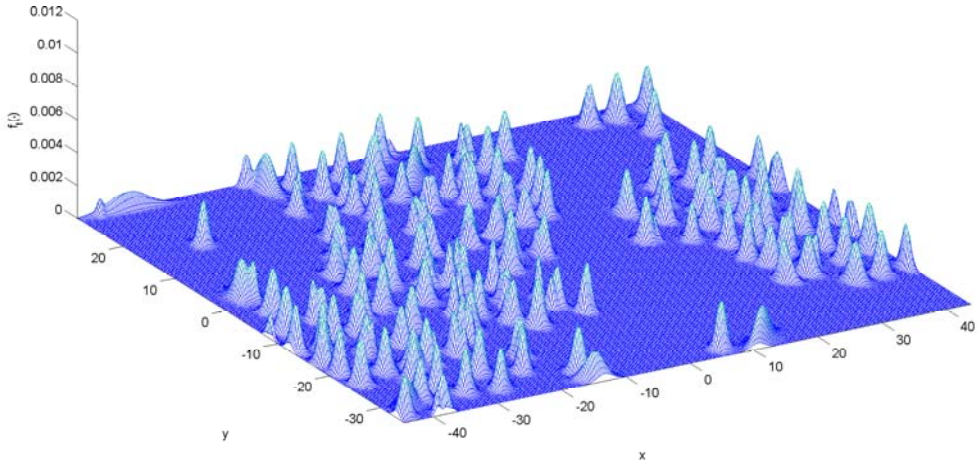


Figure 4.7: MAP sequence estimation of the spatial intensity function: red dots are the realization of spatial point pattern, + and ellipse represent the mean and standard deviation of mixture components.



(a)



(b)

Figure 4.8: Posterior mean of the spatial intensity function at last time step: (a) image plot, (b) 3D plot.

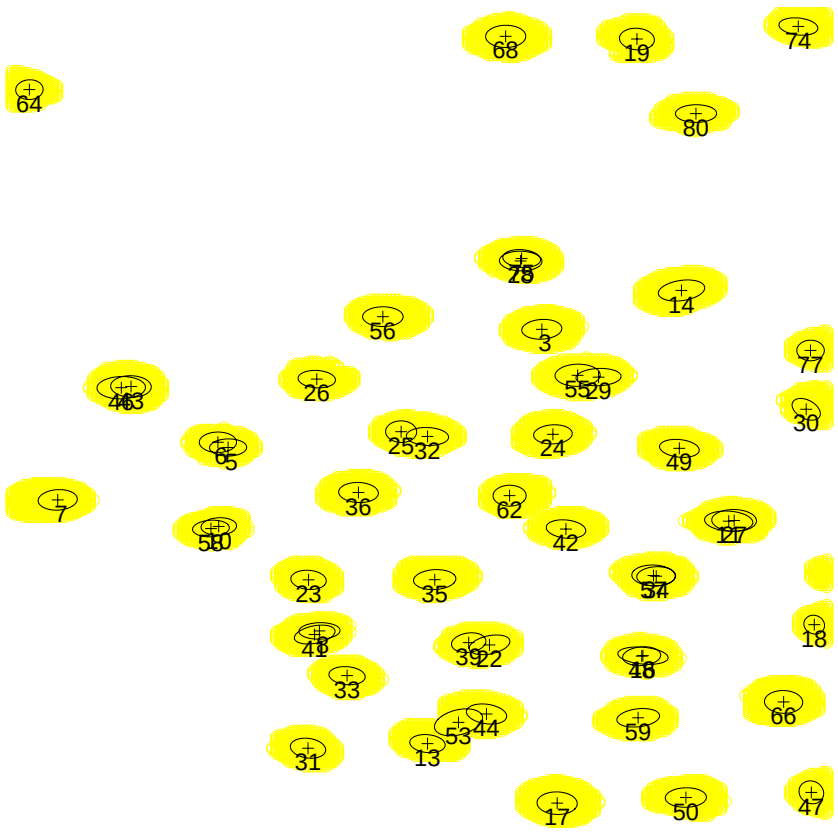
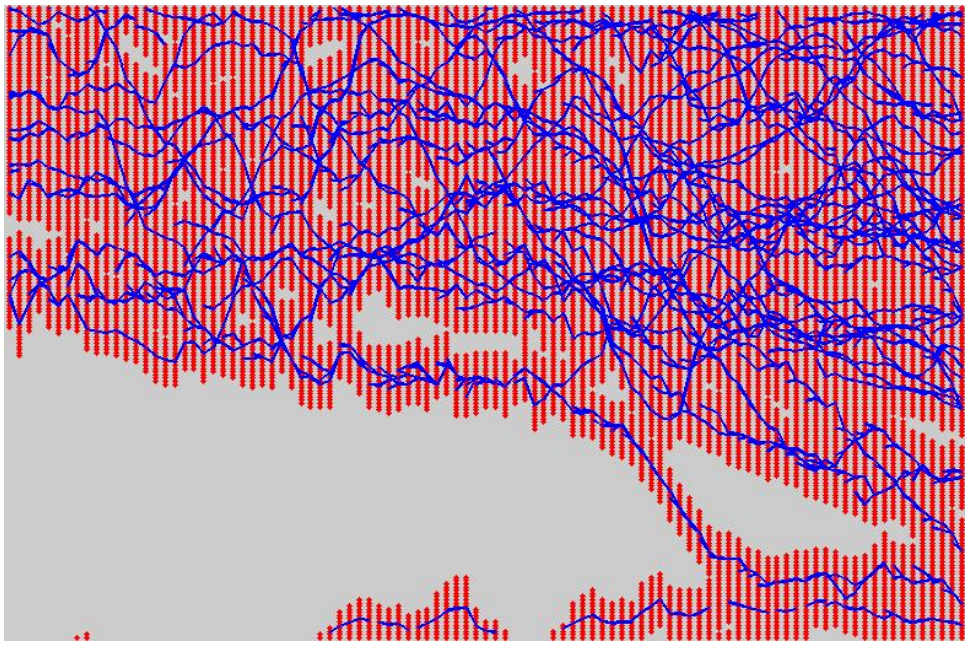
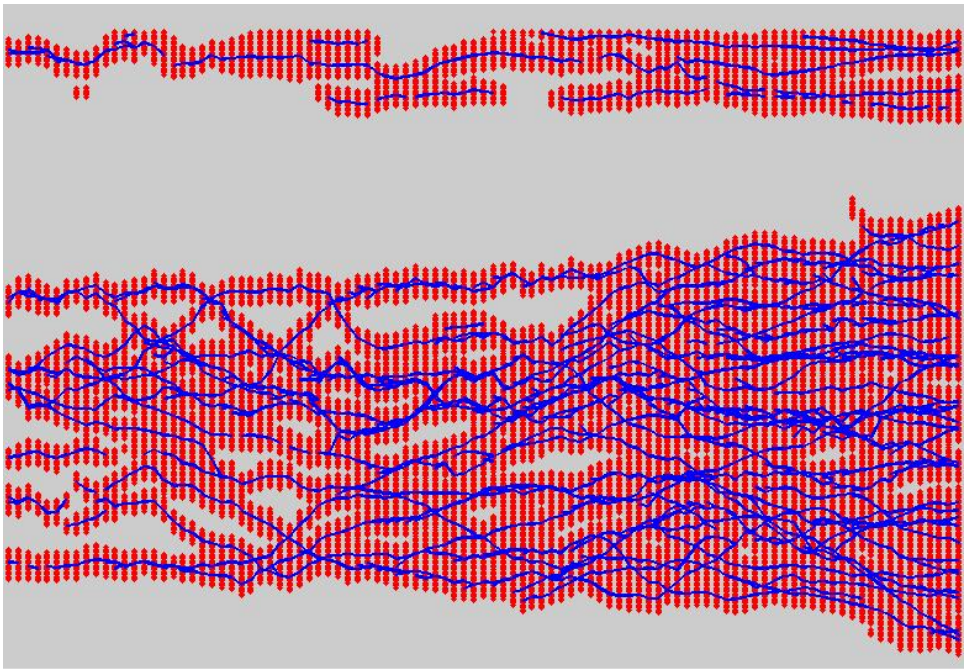


Figure 4.9: Last frame of the movie of cell tracking in a zoom in area: observed spatial point process is shown by yellow dots, the MAP sequence estimates of the spatial intensity function is represented by + (mean) and ellipse (standard deviation). Moreover, to identify the mixture components, each component is labeled by a number.



(a)



(b)

Figure 4.10: Reconstructed trajectories of each cell shown in both x and y coordinate: red dots represent observations and blue curves represent the reconstructed trajectories.

Chapter 5

Marginal Likelihood Approximation

We discuss some novel approaches for estimation of the upper and lower bounds of the log marginal likelihood in certain Bayesian models. Apparently, performing model selection merely based on only lower bounds of log marginal likelihoods can be inappropriate as the approximation error is not quantitatively limited. We provide an upper bound for the log marginal likelihood to couple with a lower bound and propose a method based on the posterior samples to minimize this upper bound. We provide a quasi-lower bound that can be obtained with trivial computation based on the result of optimal upper bound. We also demonstrate that by marginalizing some parameters in the Bayesian model, we can significantly reduce the error between the bounds of log marginal likelihood. However, when some parameters are marginalized, the optimal lower bound cannot be obtained using the traditional variational methods. To address this, we present a method that directly uses a Monte Carlo Stochastic Approximation (MCSA) algorithm to maximize the lower bound, and prove the convergence to the true local maximum lower bound under commonly satisfied assumptions.

5.1 Introduction

The marginal likelihood is the essential quantity in Bayesian model selection, representing the evidence of a model. However, evaluating marginal likelihoods often involves intractable integration and relies on numerical integration and approximation. Mean-field variational methods, initially developed in statistical physics and extensively studied by machine learning and Bayesian learning communities

for deterministic approximation of marginal distributions (MacKay, 1995; Jordan et al., 1999; Jaakkola and Jordan, 2000; Humphreys and Titterington, 2000; Ueda and Ghahramani, 2002; Jordan, 2004; Wang and Titterington, 2004), have been implemented in the model selection context (Corduneanu and Bishop, 2001; Beal, 2003).

For model M and data D, the marginal likelihood denoted in a general form is

$$p(D|M) = \int_{\boldsymbol{\Theta}} p(\boldsymbol{\theta}, D|M) \mathrm{d}\boldsymbol{\theta}, \tag{5.1}$$

with $\boldsymbol{\theta} = \{\theta_1, \ldots, \theta_K\} \in \boldsymbol{\Theta}$ representing all the model parameters. In certain cases, some of the parameters can be analytically integrated out, reducing the dimension of the integral, and the integration over remaining parameters was the same form.

For any density function $q(\boldsymbol{\theta}; \boldsymbol{\gamma})$ that is parameterized with $\boldsymbol{\gamma} = \{\gamma_1, \ldots, \gamma_J\} \in \boldsymbol{\Gamma}$ and has the same support as the posterior density function $p(\boldsymbol{\theta}|D, M)$, Jensen's inequality

$$\log p(D|M) \geq \int_{\boldsymbol{\Theta}} q(\boldsymbol{\theta}; \boldsymbol{\gamma}) \log \frac{p(\boldsymbol{\theta}, D|M)}{q(\boldsymbol{\theta}; \boldsymbol{\gamma})} \mathrm{d}\boldsymbol{\theta} \tag{5.2}$$

provides a lower bound for the log marginal likelihood, which can be maximized with respect to $\boldsymbol{\gamma}$ and serve as an approximation to the log marginal likelihood. This lower bound optimization corresponds to minimization of the Kullback-Leibler divergence between the variational density $q(\boldsymbol{\theta}; \boldsymbol{\gamma})$ and model parameter posterior density $p(\boldsymbol{\theta}|D, M)$.

The current mean-field variational methods use variational density form factorized over hidden variables and model parameters (or construct such settings by treating certain model parameters as hidden variables), and rely on EM algorithms to provide solutions to the lower bound optimization (Beal, 2003). Similar to its application in variational MLE with missing data (Celeux and Diebolt, 1992; Delyon et al., 1999), the Stochastic Approximation algorithm based on an iterative

Monte Carlo procedure can be employed in cases where the expectation step in the EM algorithm cannot be performed in closed form. Wang and Titterington (2004) have shown that, for the mean-field variational densities of exponential family, this optimization converges to the true local maximum lower bound with probability one.

Apparently, performing model selection merely based on the lower bounds of log marginal likelihoods can be inappropriate as the approximation error is not quantitatively limited. Here, we show an upper bound for the log marginal likelihood and propose a method based on Markov chain Monte Carlo methods to minimize this upper bound. This optimization is equivalent to minimization of the Kullback-Leibler divergence between $p(\boldsymbol{\theta}|D, M)$ and $q(\boldsymbol{\theta}; \boldsymbol{\gamma})$, and, for $q(\boldsymbol{\theta}; \boldsymbol{\gamma})$ of exponential family, converges to the true global minimum upper bound almost surely or with probability one.

We also discuss a quasi-lower bound that can be obtained with trivial computation based on the result of optimal upper bound. We demonstrate that by marginalizing some parameter in the Bayesian model, we can significantly reduce the error between the bounds of log marginal likelihood. We present a method that directly uses Monte Carlo Stochastic Approximation (MCSA) algorithm to maximize the lower bound, and prove the convergence to the true local maximum lower bound with probability one when $q(\boldsymbol{\theta}; \boldsymbol{\gamma})$ takes an exponential family form.

In the following sections, the methodology regarding the upper bound and lower bound optimization derivation and optimization is described in detail. The performance of this new method is demonstrated with two examples. The first example is a Bayesian linear regression model, in which the analytical form of marginal likelihood is available. In the second example, we investigate our proposed approach on finite Gaussian mixture models. Both simulation examples show

that our approach can give reasonable bounds for the log marginal likelihood.

5.2 Upper Bound Computation with MCMC

When $q(\boldsymbol{\theta};\boldsymbol{\gamma}) = p(\boldsymbol{\theta}|D,M)$, the inequality in (5.2) turns into equality

$$\log p(D|M) = \int_{\boldsymbol{\Theta}} p(\boldsymbol{\theta}|D,M)\log p(D,\boldsymbol{\theta}|M)\mathrm{d}\boldsymbol{\theta} - \int_{\boldsymbol{\Theta}} p(\boldsymbol{\theta}|D,M)\log p(\boldsymbol{\theta}|D,M)\mathrm{d}\boldsymbol{\theta}. \tag{5.3}$$

The second term $-\int_{\boldsymbol{\Theta}} p(\boldsymbol{\theta}|D,M)\log p(\boldsymbol{\theta}|D,M)\mathrm{d}\boldsymbol{\theta}$ is the mathematical entropy of $p(\boldsymbol{\theta}|D,M)$. According to Gibbs' inequality, for any $q(\boldsymbol{\theta};\boldsymbol{\gamma})$,

$$-\int_{\boldsymbol{\Theta}} p(\boldsymbol{\theta}|D,M)\log p(\boldsymbol{\theta}|D,M)\mathrm{d}\boldsymbol{\theta} \leq -\int_{\boldsymbol{\Theta}} p(\boldsymbol{\theta}|D,M)\log q(\boldsymbol{\theta};\boldsymbol{\gamma})\mathrm{d}\boldsymbol{\theta}. \tag{5.4}$$

Inserting this into (5.3) leads to an upper bound of log marginal likelihood

$$U(\boldsymbol{\gamma}) = \int_{\boldsymbol{\Theta}} p(\boldsymbol{\theta}|D,M)\log\frac{p(\boldsymbol{\theta},D|M)}{q(\boldsymbol{\theta};\boldsymbol{\gamma})}\mathrm{d}\boldsymbol{\theta} \geq \log p(D|M). \tag{5.5}$$

Given a variational proposal distribution $q(\boldsymbol{\theta};\boldsymbol{\gamma})$, this upper bound is the expectation of $\log\frac{p(\boldsymbol{\theta},D|M)}{q(\boldsymbol{\theta};\boldsymbol{\gamma})}$ with respect to the posterior distribution $p(\boldsymbol{\theta}|D,M)$. Since $p(\boldsymbol{\theta}|D,M)$ does not have an explicit form, the evaluation of this upper bound becomes a Monte Carlo integration problem that depends on samples from $p(\boldsymbol{\theta}|D,M)$. As is well known, these posterior samples of model parameters can be obtained through an MCMC sampler with target the desired posterior distribution.

It is straightforward to show that to approach the optimal $U(\boldsymbol{\gamma})$ is equivalent to minimizing the Kullback-Leibler divergence between $q(\boldsymbol{\theta};\boldsymbol{\gamma})$ and $p(\boldsymbol{\theta}|D,M)$,

$$\mathcal{D}\left[p||q\right] \equiv \int_{\boldsymbol{\Theta}} p(\boldsymbol{\theta}|D,M)\log\frac{p(\boldsymbol{\theta}|D,M)}{q(\boldsymbol{\theta};\boldsymbol{\gamma})}\mathrm{d}\boldsymbol{\theta}, \tag{5.6}$$

which is actually the error term (or say discrepancy) between $U(\boldsymbol{\gamma})$ and $\log p(D|M)$. Since we assume that $q(\boldsymbol{\theta};\boldsymbol{\gamma})$ comes from the exponential family, which is well known to be log-concave, $\mathcal{D}\left[p||q\right]$, as a linear functional of $\log q(\boldsymbol{\theta};\boldsymbol{\gamma})$ must be convex with respect to $\boldsymbol{\gamma}$. Hence, the global minimum can be found by solving the equations

$$\frac{\partial}{\partial\gamma_j}\mathcal{D}\left[p||q\right] = -\int_{\boldsymbol{\Theta}} p(\boldsymbol{\theta}|D,M)\left[\frac{\partial}{\partial\gamma_j}\log q(\boldsymbol{\theta};\boldsymbol{\gamma})\right]\mathrm{d}\boldsymbol{\theta} = 0, \tag{5.7}$$

$j = 1,\ldots,J$.

For $q(\boldsymbol{\theta};\boldsymbol{\gamma})$ of exponential family form, these equations can be solved analytically, given the Monte Carlo samples of $p(\boldsymbol{\theta}|D,M)$, $\{\boldsymbol{\theta}^{(i)} : i = 1,\ldots,N\}$ generated through MCMC. It is trivial to prove that the estimated solution $\hat{\boldsymbol{\gamma}}_U$ almost surely converges to the true solution $\boldsymbol{\gamma}_U$ if $\hat{\boldsymbol{\gamma}}_U$ can be analytically expressed. Then with the estimated optimum $\hat{\boldsymbol{\gamma}}_U$, the global minimum upper bound of log marginal likelihood U_o can be estimated by

$$\hat{U}_o = \frac{1}{N}\sum_{i=1}^{N}\log\frac{p(\boldsymbol{\theta}^{(i)},D|M)}{q(\boldsymbol{\theta}^{(i)};\hat{\boldsymbol{\gamma}}_U)}. \tag{5.8}$$

When $N\to\infty$, we can prove the almost sure convergence of $\hat{U}_o$ to U_o if $\hat{\boldsymbol{\gamma}}_U$ can be analytically expressed.

5.3 Lower Bound Computation

This section shows that optimal lower bound of log marginal likelihood can be approximated in several ways. Variational Bayesian methods are standard approaches (Jordan et al., 1999; Ghahramani and Beal, 2001; Beal, 2003; Beal and Ghahramani, 2003; Xing et al., 2003; Blei and Jordan, 2004). Here we present two alternative methods for evaluating lower bounds: quasi-lower bounds and MCSA lower bounds.

5.3.1 Variational Methods

We briefly introduce variational Bayesian methods, which have been well presented (Jordan et al., 1999; Ghahramani and Beal, 2001; Blei and Jordan, 2004). Recall that we are considering a model with parameters $\boldsymbol{\theta}$ and observation D. The posterior can be written as

$$p(\boldsymbol{\theta}|D, M) = \exp\{\log p(\boldsymbol{\theta}, D|M) - \log p(D|M)\}. \tag{5.9}$$

The variational lower bound on the log marginal probability is

$$L(\boldsymbol{\gamma}) = E_q[\log(p(D, \boldsymbol{\theta}|M)] - E_q[\log q(\boldsymbol{\theta})] \leq \log p(D|M). \tag{5.10}$$

Note that this bound holds for any distribution $q(\boldsymbol{\theta})$.

For the optimization of this bound to be computationally tractable, we assume a fully-factorized form $q_{\boldsymbol{\gamma}}(\boldsymbol{\theta}) = \prod_{j=1}^{J} q_{\gamma_j}(\theta_j)$, where $\boldsymbol{\gamma} = \{\gamma_1, ..., \gamma_J\}$ are variational parameters, and each distribution is in the exponential family, i.e. $q_{\gamma_j}(\theta_j) = h(\theta_j) \exp\{\gamma_j^T \theta_j - a(\gamma_j)\}$. By taking the derivative of $L(\boldsymbol{\gamma})$ with respect to each γ_i, the optimal γ_j maximizing $L(\boldsymbol{\gamma})$ satisfies

$$\gamma_j = [a''(\gamma_j)]^{-1} \left(\frac{\partial}{\partial \nu_j} E_q[\log p(\theta_j|D, \theta_{-j}, M)] - \frac{\partial}{\partial \gamma_j} E_q[\log h(\theta_j)] \right). \tag{5.11}$$

A further simplification can be achieved, if $p(\theta_j|D, \theta_{-j}, M)$ is also an exponential family distribution,

$$p(\theta_j|D, \theta_{-j}, M) = h(\theta_j) \exp\{g(\theta_{-j}, D, M)^T \theta_j - a(g(\theta_{-j}, D, M))\},$$

where $g(\theta_{-j}, D, M)$ is the natural parameter for θ_j conditioning on the remaining other variational parameters and the observations. The maximum of $L(\boldsymbol{\gamma})$ is attained at

$$\gamma_j = E_q[g(\theta_{-j}, D, M)]. \tag{5.12}$$

Then a coordinate ascent algorithm can be derived, in which we iteratively maximize the bound with respect to each γ_j, holding the other variational parameters fixed (see Ghahramani and Beal (2001); Blei and Jordan (2004) for details).

5.3.2 Quasi-Lower Bound

As observed, the optimization of upper and lower bounds is equivalent to minimization of $\mathcal{D}[p||q]$ and $\mathcal{D}[q||p]$. Since the KL-divergence is not a symmetric distance measure, generally $\text{argmin}\mathcal{D}[p||q] \neq \text{argmin}\mathcal{D}[q||p]$. However, if the distribution p is unimodal, then $\boldsymbol{\gamma}_U = \text{argmin}\mathcal{D}[p||q]$ allows $\mathcal{D}[q||p]$ to approximate its minimum. As analyzed in Section 5.2, given samples drawn from posterior $p(\theta|D, M)$, it is generally straightforward to estimate $\hat{\boldsymbol{\gamma}}_U$ which minimizes the KL-divergence $\mathcal{D}[p||q]$ and thus enables $\mathcal{D}[q||p]$ to approximate its minimum. So we can use $\hat{\boldsymbol{\gamma}}_U$ to estimate a *quasi lower bound* as follows: draw the Monte Carlo samples $\{\boldsymbol{\theta}^{(i)} : i = 1, \ldots, N\}$ from $q(\boldsymbol{\theta}; \hat{\boldsymbol{\gamma}}_U)$, then obtain the estimate of the quasi-lower bound using

$$\hat{L}_o = \frac{1}{N} \sum_{i=1}^{N} \log \frac{p(\boldsymbol{\theta}^{(i)}, D|M)}{q(\boldsymbol{\theta}^{(i)}; \hat{\boldsymbol{\gamma}}_U)}. \tag{5.13}$$

5.3.3 Lower Bound Optimization by MCSA

As mentioned in Blei and Jordan (2004), if the variational distribution $q_{\boldsymbol{\gamma}}(\boldsymbol{\theta})$ is not a fully factorized distribution, then the analytical iterative update equation for variational parameters derived in the variational algorithm may not be applicable. However, such a scenario is common. For example, a model with only one parameter remaining after marginalizing other model parameters. We will show later in the simulation study that by marginalizing some model parameters, the 'discrepancy' between the bounds of log marginal likelihood can be reduced significantly.

However in such scenarios, the optimal lower bound cannot be obtained using traditional variational methods. We present an alternative method which uses a Monte Carlo Stochastic Approximation (MCSA) algorithm to maximize the lower bound directly. We also prove that the algorithm converges to the true local maximum lower bound with probability one when $q(\boldsymbol{\theta};\boldsymbol{\gamma})$ takes an exponential family form.

Given a parameterized variational density $q(\boldsymbol{\theta};\boldsymbol{\gamma})$, the lower bound of the log marginal likelihood as a function of $\boldsymbol{\gamma}$ can be written as

$$L(\boldsymbol{\gamma}) = \int_{\Theta} q(\boldsymbol{\theta};\boldsymbol{\gamma}) \log \frac{p(\boldsymbol{\theta}, D|M)}{q(\boldsymbol{\theta};\boldsymbol{\gamma})} \mathrm{d}\boldsymbol{\theta}. \tag{5.14}$$

By taking the derivative of $L(\boldsymbol{\gamma})$, the optimum $\boldsymbol{\gamma}_L$ that maximizes $L(\boldsymbol{\gamma})$ is the solution of the system of equations,

$$\frac{dL(\boldsymbol{\gamma})}{d\boldsymbol{\gamma}} = -\int_{\boldsymbol{\Theta}} q(\boldsymbol{\theta};\boldsymbol{\gamma}) \left[1 + \log \frac{q(\boldsymbol{\theta};\boldsymbol{\gamma})}{p(\boldsymbol{\theta}, D|M)}\right] \frac{d}{d\gamma} \log q(\boldsymbol{\theta};\boldsymbol{\gamma}) d\boldsymbol{\theta} = 0. \tag{5.15}$$

Define $h(\boldsymbol{\theta};\boldsymbol{\gamma}) = \left[1 + \log \frac{q(\boldsymbol{\theta};\boldsymbol{\gamma})}{p(\boldsymbol{\theta}, D|M)}\right] \frac{d}{d\gamma} \log q(\boldsymbol{\theta};\boldsymbol{\gamma})$. As can be seen, these equations are non-linear and cannot be solved analytically. To numerically find the solution, we present a Monte Carlo Stochastic Approximation (MCSA) algorithm.

Stochastic Approximation (SA) (Robbins and Monro, 1951; Kushner and Yin, 1997) is a class of algorithms for finding the roots of possibly non-linear equation $f(x) = 0$, in the situation where only noisy measurements of $f(x)$ are available. Robbins-Monro algorithm (Robbins and Monro, 1951), the simplest form of SA, is a recursive process as

$$x^{(t+1)} = x^{(t)} + s^{(t+1)} \zeta^{(t+1)} \tag{5.16}$$

with some initial $x^{(0)}$. Here $\left\{s^{(t)}, t \geq 1\right\}$ is a sequence of stepsizes which satisfies standard conditions: $\sum_{t=1}^{\infty} s^{(t)} = \infty$ and $\sum_{t=1}^{\infty} \left[s^{(t)}\right]^2 < \infty$. For any $t \geq 1$. $\zeta^{(t)}$ is a

noisy measurement of $f(x)$, i.e.

$$\zeta^{(t)} = f(x) + \xi^{(t)}, \tag{5.17}$$

where $\left\{\xi^{(t)}, t \geq 1\right\}$ is the noise sequence.

In our case, x is $\boldsymbol{\gamma}$ and function $f(\boldsymbol{\gamma}) = \int_{\boldsymbol{\Theta}} h(\boldsymbol{\theta}; \boldsymbol{\gamma}) d\boldsymbol{\theta}$. Assume we have Monte Carlo samples $\{\boldsymbol{\theta}^{(i)} : i = 1, \ldots, N\}$ from the distribution $q(\boldsymbol{\theta}; \boldsymbol{\gamma})$. Then $f(\boldsymbol{\gamma})$ can be evaluated by its Monte Carlo estimate

$$\zeta(\boldsymbol{\gamma}) = -\frac{1}{N} \sum_{i=1}^{N} \left\{ \left[1 + \log \frac{q(\boldsymbol{\theta}^{(i)}; \boldsymbol{\gamma})}{p(\boldsymbol{\theta}^{(i)}, D|M)} \right] \frac{d}{d\boldsymbol{\gamma}} \log q(\boldsymbol{\theta}^{(i)}; \boldsymbol{\gamma}) \right\}. \tag{5.18}$$

The Central Limit Theorem,

$$\xi(\boldsymbol{\gamma}) = [\zeta(\boldsymbol{\gamma}) - f(\boldsymbol{\gamma})] \rightarrow N\left(0, \frac{\sigma^2}{N}\right), \quad \text{as} \quad n \rightarrow \infty \tag{5.19}$$

implies that $\xi(\boldsymbol{\gamma})$ is Gaussian noise, with mean zero and variance $\frac{\sigma^2}{N}$ with $\sigma^2 = \frac{1}{N-1} \sum_{i=1}^{N} \left(h(\boldsymbol{\theta}^{(i)}; \boldsymbol{\gamma}) - \zeta(\boldsymbol{\gamma})\right)$ (Robert and Casella, 2004).

Using Robbins-Monro algorithm, we can estimate $\boldsymbol{\gamma}$ iteratively via

$$\boldsymbol{\gamma}^{(t+1)} = \boldsymbol{\gamma}^{(t)} + s^{(t+1)} \zeta\left(\boldsymbol{\gamma}^{(t)}\right). \tag{5.20}$$

Then using the estimate $\hat{\boldsymbol{\gamma}}_L$ produced through above iterative procedure and the Monte Carlo samples $\{\boldsymbol{\theta}^{(i)} : i = 1, \ldots, N\}$ from $q(\boldsymbol{\theta}; \hat{\boldsymbol{\gamma}}_L)$, we obtain the estimate of the optimal lower bound conditional on the kernel form of the variational density function,

$$\hat{L}_o = \frac{1}{N} \sum_{i=1}^{N} \log \frac{p(\boldsymbol{\theta}^{(i)}, D|M)}{q(\boldsymbol{\theta}^{(i)}; \hat{\boldsymbol{\gamma}}_L)}. \tag{5.21}$$

When the number of iterative steps in the stochastic approximation tends to infinity, this estimated lower bound converges to the true maximum lower bound L_o with probability one. The proof of this conclusion is presented in Appendix D.

5.4 Applications

The proposed log marginal likelihood approximation approaches can be applied to a wide range of Bayesian models. In this section, we present two examples to demonstrate the performance of our proposed methods. The first model is a generic Bayesian linear regression; by using a conjugate prior setting, we can derive the analytical form of log marginal likelihood, which can help us understanding the performance of its upper and lower bounds. The second one is a mixture model for with a large number of data point, which does not have a tractable analytical log marginal likelihood. We estimated the log marginal likelihood by the standard approach proposed in Chib (1995) and compared with various upper and lower bounds.

5.4.1 Bayesian Linear Regression

Assume we have the predictors $X = (\mathbf{x}_1, ..., \mathbf{x}_n)^T$, and response variables $y = (y_1, ..., y_n)^T$. The linear regression model can be written as

$$y_i \sim N\left(\beta x_i, \sigma^2\right) \tag{5.22a}$$

$$\beta \sim N\left(0, \tau\sigma^2\mathbf{I}\right) \tag{5.22b}$$

$$\sigma^2 \sim IG\left(h_0, k_0\right) \tag{5.22c}$$

where hyperparameters τ, h_0, and k_0 are assumed to be fixed and known. Due to the conjugate setting, the marginal likelihood has a closed form,

$$p(y|X, \tau, h_0, k_0) = \int\int p(y|X, \beta, \sigma^2)p(\beta|\tau\sigma^2 I)p(\sigma^2|h_0, k_0)d\beta d\sigma^2 \tag{5.23a}$$

$$= (2\pi)^{-\frac{n}{2}}|C|^{-\frac{1}{2}}k_0^{h_0}\frac{\Gamma\left(\frac{n}{2}+h_0\right)}{\Gamma(h_0)}\left(\frac{y^T C^{-1}y}{2}+k_0\right)^{-\frac{n}{2}+h_0} \tag{5.23b}$$

where $C = I_{nn} + \tau XX^T$.

Moreover, if we marginalize out β in the model, we have

$$p(y|X,\sigma^2,\tau) = \int p(y|X,\beta,\sigma^2)p(\beta|\tau\sigma^2 I)d\beta \quad (5.24a)$$

$$= (2\pi)^{-\frac{n}{2}}|B|^{-\frac{1}{2}}\exp(-\frac{1}{2}y^T B^{-1}y) \quad (5.24b)$$

where $B = \sigma^2 I_{nn} + \tau\sigma^2 XX^T$.

In order to draw posterior samples from this Bayesian linear model, we turn to Gibbs sampling, in which the conditional posteriors are,

$$\beta \sim N\left((\tau^{-1}\mathbf{I} + X^T X)^{-1}X^T y, \sigma^2(\tau^{-1}I + X^T X)^{-1}\right) \quad (5.25)$$

$$\sigma^2 \sim IG\left(\frac{n+p}{2} + h_0, \frac{(y-X\beta)^T(y-X\beta)}{2} + \frac{\beta^T\beta}{2}\tau^{-1} + k_0\right). \quad (5.26)$$

We study the order selection in a polynomial regression, in which for order q, the predictors are $\mathbf{x} = [1, x, x^2, ..., x^q]^T$. The data is synthetic with $q = 3$, $n = 20$ and know variance $\sigma^2 = 10$. The hyperparameters are set as $\tau = 0.1$, $h_0 = 1$, and $k_0 = 1$.

Seven quantities will be studied for the log marginal likelihood estimation for the Bayesian polynomial regression model:

ML**:** ML is the exact log marginal likelihood, which can be calculated by equation (5.23) given the data (X, y) and the model hyperparameters.

U_1**:** U_1 is the estimate of the upper bound of log marginal likelihood by marginalizing out coefficients β in the model. Thus we set the variational distribution as $q(\boldsymbol{\theta};\boldsymbol{\gamma}) = IG(\sigma^2|h,k)$. After obtaining posterior samples $\{(\beta,\sigma^2)^{(i)} : i = 1,\ldots,N\}$ generated through Gibbs sampling, optimal variational parameters h_U and k_U can be estimated by (5.7). Thus, the upper bound can be evaluated using equation (5.8) with $p(\boldsymbol{\theta}^{(i)}, D|M) = p(y|X,\sigma^{2(i)},\tau)IG(\sigma^{2(i)}|h_0,k_0)$.

U_2: U_2 is the estimate of the upper bound of log marginal likelihood with respect to the full parameterized model defined as (5.22), in which we set the variational distribution as $q(\boldsymbol{\theta};\boldsymbol{\gamma}) = N(\beta|\mu,\Omega)IG(\sigma^2|h,k)$. After obtaining posterior samples $\{(\beta,\sigma^2)^{(i)} : i = 1,\ldots,N\}$ generated through Gibbs sampling, optimal variational parameters μ_U, Ω_U, h_U , and k_U can be estimated using (5.7). Thus, the upper bound can be evaluated by equation (5.8), in which $p(\boldsymbol{\theta}^{(i)}, D|M) = p(y|X,\beta^{(i)},\sigma^{2(i)})N(\beta^{(i)}|0,\tau\sigma^2\mathbf{I})IG(\sigma^{2(i)}|h_0,k_0)$.

L_1: L_1 is the estimate of the lower bound of log marginal likelihood for the model marginalizing out coefficients β by the MCSA method proposed in Section 5.3.3. The variational distribution is $q(\boldsymbol{\theta};\boldsymbol{\gamma}) = IG(\sigma^2|h,k)$, where parameters h_L and h_L are estimated by the MCSA algorithm. Samples $\{\sigma^{2(i)} : i = 1,\ldots,N\}$ are drawn from $IG(\sigma^2|h_L,k_L)$, and then used to estimate the quasi-lower bound using equation (5.21), in which $p(\boldsymbol{\theta}^{(i)}, D|M) = p(y|X,\sigma^{2(i)},\tau)$ $IG(\sigma^{2(i)}|h_0,k_0)$.

L_2: L_2 is the estimate of the quasi-lower bound of log marginal likelihood by marginalizing out coefficients β in the model. The variational distribution is $q(\boldsymbol{\theta};\boldsymbol{\gamma}) = IG(\sigma^2|h,k)$, where the parameter h_L and h_L are set the same as h_U and k_U. Samples $\{\sigma^{2(i)} : i = 1,\ldots,N\}$ are drawn from $IG(\sigma^2|h_L,k_L)$, and then used to estimate the quasi lower bound using equation (5.21), in which $p(\boldsymbol{\theta}^{(i)}, D|M) = p(y|X,\sigma^{2(i)},\tau)IG(\sigma^{2(i)}|h_0,k_0)$.

L_3: L_3 is the estimate of the quasi-lower bound of log marginal likelihood with respect to the full parameterized model defined as (5.22), in which we set the variational distribution as $q(\boldsymbol{\theta};\boldsymbol{\gamma}) = N(\beta|\mu,\Omega)IG(\sigma^2|h,k)$, where the parameter μ_L, σ_L^2, h_L and h_L are set the same as μ_U, σ_U^2, h_U and k_U. Samples $\{(\beta,\sigma^2)^{(i)} : i = 1,\ldots,N\}$ are drawn from $N(\beta|\mu_L,\Omega_L)IG(\sigma^2|h_L,k_L)$, and

then used to estimate the quasi-lower bound using equation (5.21), in which $p(\boldsymbol{\theta}^{(i)}, D|M) = p(y|X, \beta^{(i)}, \sigma^{2(i)})N(\beta^{(i)}|0, \tau\sigma^2\mathbf{I})IG(\sigma^{2(i)}|h_0, k_0)$.

L_4**:** L_4 is the estimate of the lower bound of log marginal likelihood of using variational method. The variational distribution is $q(\boldsymbol{\theta}; \boldsymbol{\gamma}) = N(\beta|\mu, \Omega)IG(\sigma^2|h, k)$, where the optimal parameter μ_L, ${\sigma^2}_L$, h_L and h_L are estimated by the variational method, i.e. the coordinate ascent algorithm presented in Appendix C. Samples $\{(\beta, \sigma^2)^{(i)} : i = 1, \ldots, N\}$ are drawn from $N(\beta|\mu_L, \Omega_L)IG(\sigma^2|h_L, k_L)$, and then used to estimate the lower bound using equation (5.21), in which $p(\boldsymbol{\theta}^{(i)}, D|M) = p(y|X, \beta^{(i)}, \sigma^{2(i)})N(\beta^{(i)}|0, \tau\sigma^2\mathbf{I})IG(\sigma^{2(i)}|h_0, k_0)$.

We show the synthetic data is approximated by polynomials of varying orders in Figure 5.1. We run the process of computing the values of these seven quantities 100 time, and show their means and standard deviations in Table 5.1. From the comparison, we find U_1 has better performance than U_2, and L_1, L_2 have better performance than L_3 L_4, which means that by marginalizing out some parameter in the Bayesian linear model, we can significantly reduce the errors between the exactly value of log marginal likelihood and its upper/lower bounds. Moreover, due to the restrictive assumption in using variational method, it cannot deal with models with only one parameter (i.e. the case we marginalize out β in the Bayesian linear regression). Moreover, the curves of the means of U_1 and L_1 in Figure 5.2 show that the preferable order is $q = 3$, which is the 'true' model in our context.

5.4.2 Mixture Model

Mixture models are widely used conventional tools for modelling complex probability distributions, and consist of a linear combination of some number K of simpler, component distributions. We focus on the marginal likelihood approximation in

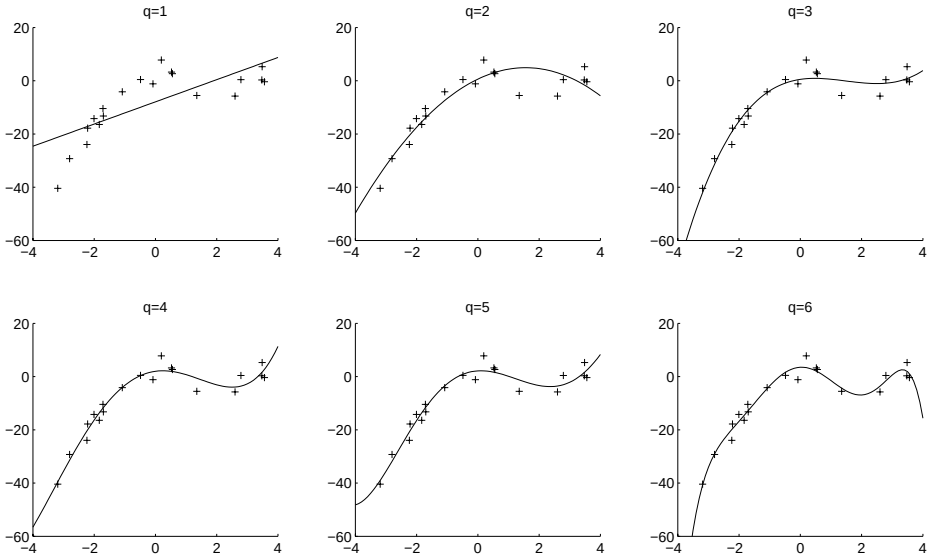


Figure 5.1: Synthetic data approximated by polynomials of varying orders.

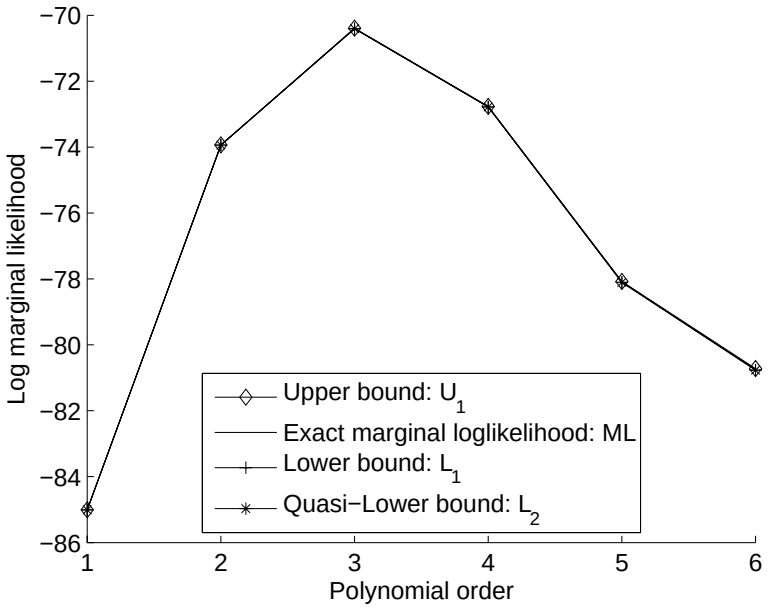


Figure 5.2: Plot of the analytic value of the log marginal likelihood of the Bayesian linear model with varying number of order q, and means of upper bound (U_1), lower bound (L_1) and quasi-lower bound (L_2) of the log marginal likelihood for 100 Monte Carlo runs.

	U_2	U_1	ML	L_1	L_2	L_3	L_4
1	-84.9591 (± 0.0045)	-85.0046 (± 0.0012)	-85.0068	-85.0074 (± 0.0006)	-85.0089 (± 0.0008)	-85.0521 (± 0.0049)	-85.0475 (± 0.0028)
2	-73.8644 (± 0.0067)	-73.9329 (± 0.0015)	-73.9374	-73.9395 (± 0.0010)	-73.9414 (± 0.0011)	-74.0038 (± 0.0052)	-73.9974 (± 0.0032)
3	-70.3021 (± 0.0079)	-70.3933 (± 0.0021)	-70.4009	-70.4054 (± 0.0012)	-70.4077 (± 0.0013)	-70.4871 (± 0.0060)	-70.4791 (± 0.0039)
4	-72.6467 (± 0.0093)	-72.7603 (± 0.0027)	-72.772	-72.7792 (± 0.0016)	-72.7825 (± 0.0020)	-72.8781 (± 0.0065)	-72.8676 (± 0.0040)
5	-77.9364 (± 0.00105)	-78.0741 (± 0.0029)	-78.1043	-78.1013 (± 0.0016)	-78.1045 (± 0.0018)	-78.2152 (± 0.0078)	-78.2040 (± 0.0042)
6	-80.5663 (± 0.00127)	-80.7266 (± 0.0039)	-80.7477	-80.7622 (± 0.0020)	-80.7656 (± 0.0020)	-80.8920 (± 0.0085)	-80.8781 (± 0.0048)

Table 5.1: Analytic values of the log marginal likelihood of the Bayesian linear model and Monte Carlo estimation of various lower and upper bounds: mean and standard deviation.

multivariate normal mixture distributions. By introducing a latent variable, also known as allocation variable z, we can interpret the normal mixture as follows:

$$x_i \sim N\left(\mu_{z_i}, \Sigma_{z_i}\right) \tag{5.27}$$

$$z_i \sim Mn\left(1, q_{0,1}, ..., q_{0,K}\right) \tag{5.28}$$

$$\mu_k \sim N\left(0, \tau_0^{-1}\Sigma_k\right) \tag{5.29}$$

$$\Sigma_k \sim IW\left(d_0, S_0\right) \tag{5.30}$$

Given the realization of the allocation variables z, the data is assigned across the K components. For a component with index k, define $y_k = \{x_i : z_i = k\}$ and $n_k = \#\{y_k\}$. Then the likelihood can be expressed as $p(y_k|\mu_k, \Sigma_k) = \prod_{l=1}^{n_k} N(y_{l,k}|\mu_k, \Sigma_k)$ where l is used to index the data points in component k. By marginalizing out the cluster mean μ_k of component k, the likelihood becomes

$$p(y_k|\Sigma_k, \tau_0) = \int p(y_k|\mu_k, \Sigma_k)p(\mu_k|0, \tau_0^{-1}\Sigma_k)d\mu_k \tag{5.31}$$

$$= \frac{\left(\frac{\tau_0}{\tau_0+n_k}\right)^{\frac{p}{2}}}{(2\pi)^{\frac{n_k p}{2}}|\Sigma_k|^{\frac{n_k}{2}}} \exp\left[-\frac{\text{Tr}(Q_0\Sigma_k^{-1}) + \frac{\tau_0 n_k}{\tau_0+n_k}\bar{y}^T\Sigma_k^{-1}\bar{y}}{2}\right] \tag{5.32}$$

where $Q_0 = \sum_{i=1}^{n_k}(y_k^{(i)} - \bar{y}_k)(y_k^{(i)} - \bar{y}_k)^T$.

After marginalizing out both cluster mean μ_k and variance Σ_k, the likelihood becomes

$$p(y_k|\tau_0, d_0, S_0) = \int\int p(y_k|\mu_k, \Sigma_k)p(\mu_k|0, \tau_0^{-1}\Sigma_k)p(\Sigma_k|d_0, S_0)d\mu_k d\Sigma_k \quad (5.33)$$

$$= \frac{\left(\frac{\tau_0}{\tau_0+n_k}\right)^{\frac{p}{2}}}{(\pi)^{\frac{n_k p}{2}}} \frac{\Gamma_p\left(\frac{d_0+n_k}{2}\right)}{\Gamma_p\left(\frac{d_0}{2}\right)} \frac{|S_0|^{\frac{d_0}{2}}}{\left(S_0 + Q_0 + \frac{\tau_0 n_k}{\tau_0+n_k}\bar{y}_k\bar{y}_k{}^T\right)^{\frac{d_0+n_k}{2}}} \quad (5.34)$$

Under a conditionally conjugate set up, Gibbs sampling can be used to drawn posterior samples using the following conditional posteriors:

$$z_i = Mn(1, w_{i,1}, ...w_{i,K}) \quad (5.35a)$$

$$\mu_k \sim N(\cdot|m_k, \tau_k^{-1}\Sigma_k) \quad (5.35b)$$

$$\Sigma_k \sim IW(\cdot|d_k, S_k) \quad (5.35c)$$

where

$$w_{i,k} \propto q_{0,k}N(x_i|\mu_k, \Sigma_k) \quad (5.36a)$$

$$m_k = \begin{cases} (\tau_0 + n_k)^{-1}(m_0\tau_0 + n_k\bar{y}_k), & \text{if } n_k > 0 \\ m_0, & \text{if } n_k = 0 \end{cases} \quad (5.36b)$$

$$\tau_k = \begin{cases} \tau_0 + n_k, & \text{if } n_k > 0 \\ \tau_0, & \text{if } n_k = 0 \end{cases} \quad (5.36c)$$

$$d_k = \begin{cases} d_0 + n_k, & \text{if } n_k > 0 \\ d_0, & \text{if } n_k = 0 \end{cases} \quad (5.36d)$$

$$S_k = \begin{cases} S_0 + Q_0 + \frac{n_k\tau_0}{(n_k+\tau_0)}(m_0 - \bar{y}_k)(m_0 - \bar{y}_k)^T, & \text{if } n_k > 0 \\ S_0, & \text{if } n_k = 0 \end{cases} \quad (5.36e)$$

Seven quantities will be studied for the log marginal likelihood estimation for the normal mixture model,

U_1: U_1 is the estimate of the upper bound of the log marginal likelihood for normal mixture model with μ_k and Σ_k being marginalized out. Only the allocation variable z remains and the variational distribution is set as $q(\boldsymbol{\theta};\boldsymbol{\gamma}) = \prod_{i=1}^{n} Mn(z_i|1, w_{i,1}, ...w_{i,K})$. After obtaining posterior samples $\{z_{1:n}^{(j)}\}_{j=1}^{M}$ using Gibbs sampling addressed in (5.35), we estimate the optimal parameters $w_{1:n,1:K}^{U}$ in the variational distribution using equation (5.7). Thus, the upper bound can be evaluated using equation (5.8), in which $p(\boldsymbol{\theta}^{(j)}, D|M) = \prod_{k=1}^{K} p(y_k^{(j)}|\tau_0, d_0, S_0) \prod_{i=1}^{n} Mn(z_i^{(j)}|1, q_{0,1}, ..., q_{0,K})$.

U_2: U_2 is the estimate of the upper bound of the log marginal likelihood for normal mixture model with μ_k being marginalized out. The variational distribution is set as $q(\boldsymbol{\theta};\boldsymbol{\gamma}) = \prod_{k=1}^{K} IW(\Sigma_k|\kappa_k, \Psi_k) \prod_{i=1}^{n} Mn(z_i|1, w_{i,1}, ...w_{i,K})$. After obtaining posterior samples $\{z_{1:n}^{(j)}, \Sigma_{1:K}^{(j)}\}_{j=1}^{M}$ using Gibbs sampling addressed in (5.35), we estimate the parameters $w_{1:n,1:K}^{U}$, $\kappa_{1:K}^{U}$ and $\Psi_{1:K}^{U}$ in the variational distribution using equation (5.7). Thus, the upper bound can be evaluated using equation (5.8), in which $p(\boldsymbol{\theta}^{(j)}, D|M) = \prod_{k=1}^{K} p(y_k^{(j)}|\Sigma_k^{(j)}, \tau_0)$ $\prod_{k=1}^{K} IW(\Sigma_k^{(j)}|d_0, S_0) \prod_{i=1}^{n} Mn(z_i^{(j)}|1, q_{0,1}, ..., q_{0,K})$.

U_3: U_3 is the estimate of the upper bound of the log marginal likelihood for the full parameterized model. The variational distribution is set as $q(\boldsymbol{\theta};\boldsymbol{\gamma}) = \prod_{k=1}^{K} N(\mu_k|\nu_k, \Omega_k) \prod_{k=1}^{K} IW(\Sigma_k|\kappa_k, \Psi_k) \prod_{i=1}^{n} Mn(z_i|1, w_{i,1}, ...w_{i,K})$. After obtaining posterior samples $\{z_{1:n}^{(j)}, \mu_{1:K}^{(j)}, \Sigma_{1:K}^{(j)}\}_{j=1}^{M}$ using Gibbs sampling addressed in (5.35), we estimate the parameters $w_{1:n,1:K}^{U}$, $\nu_{1:K}^{U}$, $\Omega_{1:K}^{U}$, $\kappa_{1:K}^{U}$ and $\Psi_{1:K}^{U}$ in the variational distribution by using equation (5.7). Thus, the upper bound can be evaluated by equation (5.8), in which $p(\boldsymbol{\theta}^{(j)}, D|M) = \prod_{k=1}^{K} p(y_k^{(j)}|\mu_k^{(j)}, \Sigma_k^{(j)})$ $\prod_{k=1}^{K} N(\mu_k^{(j)}|0, \tau_0^{-1}\Sigma_k^{(j)}) \prod_{k=1}^{K} IW(\Sigma_k^{(j)}|d_0, S_0) \prod_{i=1}^{n} Mn(z_i^{(j)}|1, q_{0,1}, ..., q_{0,K})$.

L_1: L_1 is the estimate of the quasi-lower bound of the log marginal likelihood

with μ_k and Σ_k being marginalized out. The variational distribution is set as $q(\boldsymbol{\theta};\boldsymbol{\gamma}) = \prod_{i=1}^{n} Mn(z_i|1, w_{i,1}, ...w_{i,K})$. Parameters $w^L_{1:n,1:K}$ in the variational distribution take the same values as $w^U_{1:n,1:K}$. Samples $\{z^{(j)}_{1:n}\}_{j=1}^{M}$ are drawn from $\prod_{i=1}^{n} Mn(z_i|1, w^L_{i,1}, ..., w^L_{1,K})$ and then the quasi-lower bound can be estimated by equation (5.21) in which $p(\boldsymbol{\theta}^{(j)}, D|M) = \prod_{k=1}^{K} p(y^{(j)}_k|\tau_0, d_0, S_0)$ $\prod_{i=1}^{n} Mn(z^{(j)}_i|1, q_{0,1}, ..., q_{0,K})$.

L_2: L_2 is the estimate of the quasi-lower bound of the log marginal likelihood with μ_k being marginalized out. The variational distribution is set as $q(\boldsymbol{\theta};\boldsymbol{\gamma}) = \prod_{k=1}^{K} IW(\Sigma_k|\kappa_k, \Psi_k) \prod_{i=1}^{n} Mn(z_i|1, w_{i,1}, ...w_{i,K})$. Parameters $w^L_{1:n,1:K}$, $\kappa^L_{1:K}$ and $\Psi^L_{1:K}$ in the variational distribution take the same values as $w^U_{1:n,1:K}$, $\kappa^U_{1:K}$ and $\Psi^U_{1:K}$ respectively. Samples $\{z^{(j)}_{1:n}, \Sigma^{(j)}_{1:K}\}_{j=1}^{M}$ are drawn from $\prod_{k=1}^{K} IW(\Sigma_k|\kappa^L_k, \Psi^L_k)$ $\prod_{i=1}^{n} Mn(z_i|1, w^L_{i,1}, ..., w^L_{i,K})$, and then the quasi-lower bound can be estimated by equation (5.21), with $p(\boldsymbol{\theta}^{(j)}, D|M) = \prod_{k=1}^{K} p(y^{(j)}_k|\Sigma^{(j)}_k, \tau_0) \prod_{k=1}^{K} IW(\Sigma^{(j)}_k|d_0, S_0)$ $\prod_{i=1}^{n} Mn(z^{(j)}_i|1, q_{0,1}, ..., q_{0,K})$.

L_3: L_3 is the estimate of quasi-lower bound of log marginal likelihood for the fully parameterized model. The variational distribution is set as $q(\boldsymbol{\theta};\boldsymbol{\gamma}) = \prod_{k=1}^{K} N(\mu_k|\nu_k, \Omega_k) \prod_{k=1}^{K} IW(\Sigma_k|\kappa_k, \Psi_k) \prod_{i=1}^{n} Mn(z_i|1, w_{i,1}, ...w_{i,K})$. Parameters $w^L_{1:n,1:K}$, $\nu^L_{1:K}$, $\Omega^L_{1:K}$, $\kappa^L_{1:K}$ and $\Psi^L_{1:K}$ in the variational distribution take the same values as $w^U_{1:n,1:K}$, $\nu^U_{1:K}$, $\Omega^U_{1:K}$, $\kappa^U_{1:K}$ and $\Psi^U_{1:K}$ respectively. Samples $\{z^{(j)}_{1:n}, \mu^{(j)}_{1:K}, \Sigma^{(j)}_{1:K}\}_{j=1}^{M}$ are drawn from $\prod_{k=1}^{K} N(\mu_k|\nu^L_k, \Omega^L_k) \prod_{k=1}^{K} IW(\Sigma_k|\kappa^L_k, \Psi^L_k) \prod_{i=1}^{n} Mn(z_i|1, w^L_{i,1:K})$, and then the quasi-lower bound can be estimated by equation (5.21), with $p(\boldsymbol{\theta}^{(j)}, D|M) = \prod_{k=1}^{K} p(y^{(j)}_k|\mu^{(j)}_k, \Sigma^{(j)}_k) \prod_{k=1}^{K} N(\mu^{(j)}_k|0, \tau_0^{-1}\Sigma^{(j)}_k) \prod_{k=1}^{K} IW(\Sigma^{(j)}_k|d_0, S_0)$ $\prod_{i=1}^{n} Mn(z^{(j)}_i|1, q_{0,1}, ..., q_{0,K})$.

L_4: L_4 is the estimate of lower bound of log marginal likelihood using the vari-

ational method. Variational methods for normal mixture models have been studied in Corduneanu and Bishop (2001) and Wang and Titterington (2004). The variational distribution is $q(\boldsymbol{\theta};\boldsymbol{\gamma}) = \prod_{k=1}^{K} N(\mu_k|\nu_k,\Omega_k)\prod_{k=1}^{K} IW(\Sigma_k|\kappa_k,\Psi_k)$ $\prod_{i=1}^{n} Mn(z_i|1,w_{i,1},...w_{i,K})$. Optimal parameters $w^L_{1:n,1:K}$, $\nu^L_{1:K}$, $\Omega^L_{1:K}$, $\kappa^L_{1:K}$ and $\Psi^L_{1:K}$ can be estimated by variational method i.e. the coordinate ascent algorithm presented in Appendix C. Samples $\{z^{(j)}_{1:n},\mu^{(j)}_{1:K},\Sigma^{(j)}_{1:K}\}_{j=1}^{M}$ are drawn from $\prod_{k=1}^{K} N(\mu_k|\nu^L_k,\Omega^L_k)\prod_{k=1}^{K} IW(\Sigma_k|\kappa^L_k,\Psi^L_k)\prod_{i=1}^{n} Mn(z_i|1,w^L_{i,1},...,w^L_{i,K})$, and then the lower bound can be estimated by equation (5.21), with $p(\boldsymbol{\theta}^{(j)},D|M)$ the same as in L_3.

We investigate our proposed approach on a synthetic data set which has been studied in Corduneanu and Bishop (2001): 600 data points generated from a mixture of five bivariate normals with means: $[0,0]$, $[3,-3]$, $[3,3]$, $[-3,3]$, $[-3,-3]$ and covariances: $[1,0;0,1]$, $[1,0.5;0.5,1]$, $[1,-0.5;-0.5,1]$, $[1,0.5;0.5,1]$, $[1,-0.5;-0.5,1]$. The synthetic data are shown in Figure 5.3. We run the process of computing the values of these seven quantities 20 time, and show their means and standard deviations in Table 5.2. From the comparison, we find U_1 (L_1) has better performance than U_2 (L_2) while U_2 (L_2) performs better than U_3(L_3), which means that by marginalizing out some parameter in the model, can significantly reduce the spread between the upper/lower bounds. Moreover, as the variational method can only be applied to the full parameterized mixture model, it cannot achieve performance as good as either L_1 or L_2; as shown in Figure 5.4 it has a big 'gap' to L_1, which implies that the lower bound given by the variational method could be far from the exact value of log likelihood. Moreover, the curves of the means of U_1, L_1 in Figure 5.4 show that the preferable order is $K = 5$, which is the 'true' model in our context.

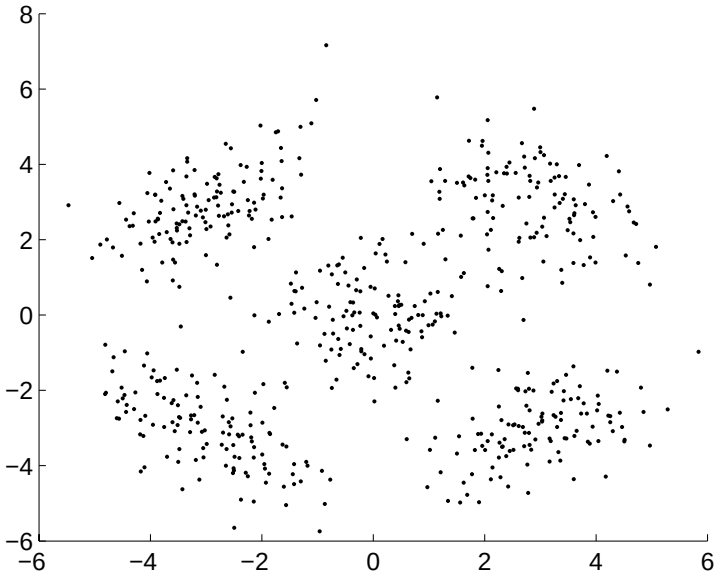


Figure 5.3: 600 data points sampled from the mixture of 5 bivariate Gaussians.

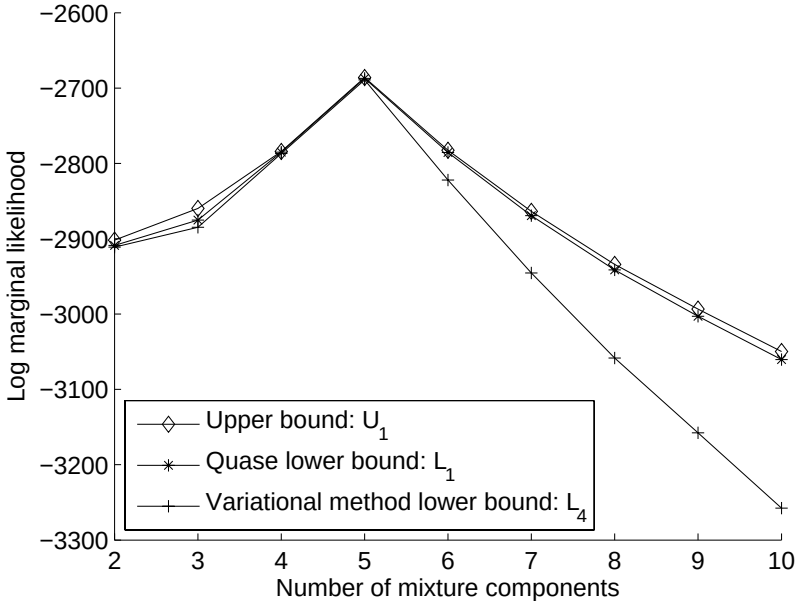


Figure 5.4: Plot of the lower bound L_1, and upper bound L_2 of the log marginal likelihood of the mixture model with varying number of components. For comparison, the lower bound L_4 estimated by the variational method is also shown in the plot.

	U_3	U_2	U_1	L_1	L_2	L_3	L_4
k=2	-2895.4 (± 22.6)	-2895.7 (± 21.6)	-2901.9 (± 25.5)	-2908.9 (± 27.5)	-2909.4 (± 27.7)	-2911.2 (± 27.5)	-2910.2 (± 27.3)
k=3	-2836.7 (± 80.6)	-2848.3 (± 38.8)	-2859.8 (± 21.1)	-2863.4 (± 14.3)	-2877.0 (± 15.4)	-2884.8 (± 26.9)	-2884.2 (± 24.8)
k=4	-2781.6 (± 21.8)	-2782.3 (± 21.9)	-2784.1 (± 21.1)	-2785.1 (± 21.4)	-2786.3 (± 22.5)	-2786.9 (± 22.6)	-2787.2 (± 22.6)
k=5	-2683.0 (± 0.15)	-2683.8 (± 0.23)	-2686.0 (± 0.09)	-2687.2 (± 0.06)	-2688.6 (± 0.08)	-2689.2 (± 0.07)	-2689.2 (± 0.08)
k=6	-2777.1 (± 1.73)	-2779.5 (± 1.93)	-2782.5 (± 1.57)	-2785.4 (± 1.20)	-2788.2 (± 1.12)	-2821.6 (± 4.05)	-2822.2 (± 4.58)
k=7	-2857.4 (± 2.44)	-2860.5 (± 2.39)	-2863.7 (± 2.31)	-2869.4 (± 1.77)	-2873.4 (± 1.60)	-2946.0 (± 9.48)	-2945.7 (± 8.68)
k=8	-2925.7 (± 2.88)	-2931.3 (± 3.42)	-2934.4 (± 3.31)	-2941.2 (± 2.12)	-2946.5 (± 1.95)	-3059.7 (± 8.18)	-3058.5 (± 8.41)
k=9	-2982.6 (± 8.01)	-2990.9 (± 8.92)	-2993.7 (± 7.23)	-3002.7 (± 4.23)	-3010.9 (± 3.98)	-3153.5 (± 17.8)	-3151.9 (± 17.1)
k=10	-3038.9 (± 3.10)	-3049.0 (± 4.14)	-3052.3 (± 3.98)	-3060.5 (± 2.19)	-3067.6 (± 1.89)	-3246.1 (± 10.5)	-3243.8 (± 10.4)

Table 5.2: Monte Carlo estimation of various lower and upper bounds of the log marginal likelihood of mixture model: mean and standard deviation over repeat simulations.

5.5 Discussion

Our variational method provides both lower and upper bounds of log marginal likelihoods that are optimized under a certain variational density form. These bounds not only facilitate more reliable model selection but also provide a way to see the advantage of the variational density as an approximation to the posterior density of model parameters. It is also worth noting that our proposed method is more general in terms of the variational density form, since in the MCSA optimization for the lower bound, the variational density does not need to take a factorized form. This is important when either only one parameter remains after marginalized out other model parameters or the model parameters left in the integration are so dependent on each other that the factorized variational density is a poor approximation to the posterior density of these parameters.

Additionally, the coexistence of the upper and lower bounds can relax the re-

quirement to find exact *optimal* bounds. Apparently, a single bound strongly relies on the optimization because its distance to the true value of log marginal likelihood itself is not bounded. The requirement of reducing this distance is imposed on the bound optimization algorithms and makes marginal likelihood computation difficult when it involves in tremendously large number of model parameters. In particular, the lower bound optimization that depends on EM or MCSA is more stressed by dimensionality, and convergence can be unbearably slow.

Such stress can be reduced when the bounds on both sides are available. On one hand, if the distance between these two bounds is small enough to distinguish different models, their optima are not necessary. One the other hand, since we use the same variational density form $q(\boldsymbol{\theta};\boldsymbol{\gamma})$ (though this is not necessary otherwise) for both bounds, if one of the bounds is easier to be optimized, the corresponding optimum variational density can be applied to compute the other bound, which may not be optimized but may be good enough for the purpose of model selection. This is due to the fact that the tightness of the bounds is essentially determined by how good an approximation $q(\boldsymbol{\theta};\boldsymbol{\gamma})$ is to $p(\boldsymbol{\theta}|D,M)$. In such a context, the upper bound shows peculiar advantage because of the better convergence property of MCMC method compared to MCSA. Moreover, simulation studies also verify that the upper bound and the quasi-lower bound whose variational parameters take the same optimal values are close to each other and facilitate more reliable model selection.

Furthermore, simulation studies show that by marginalizing out some parameters in the model, the 'discrepancy' between bounds of the log marginal likelihood and its exact value can be significantly reduced. The marginalization not only reduces the dimension of the parameter space, which enables us to approximate the posterior more easily, but can also reduce the correlation between model parame-

ters. Since we always use separated variational distributions for each of the model parameters to approximate the posterior of correlated model parameters, the KL-divergence between the posterior and variational distributions cannot approach zero no matter how we tune the associated parameters in the variational distributions, resulting in the 'discrepancy' between bounds of log marginal likelihood and its exact value. Due to the above factors, the marginalization technique used in our upper/lower bounds approximation method enable us to reduce this 'discrepancy'.

Chapter 6

Adaptive Monte Carlo Methods, Sequential Learning and Marginal Likelihood Computation

In this chapter, adaptive Monte Carlo sampling methods are introduced aimed at obtaining faster convergence rates and more efficient estimation. Adaptive Markov chains have seen renewed interest in recent years due in part to the emergence of certain theoretical guarantees (Haario et al., 2001; Roberts and Rosenthal, 2007). With adaptive MCMC algorithms, the entire sample history of a process is used to tune parameters of the proposal density during simulation. We present a general framework to design adaptive MCMC algorithms, emphasizing the adaptive Metropolized independence sampler (AMIS). To handle multimodality, we develop a simple but effective adaptation strategy using a family of mixture distribution proposals.

Motivated by the demand for flexible proposal forms in adaptive Monte Carlo methods and the need for effective approaches to fitting nonparametric model for large data sets, we propose a sequential learning method for truncated DP mixture (TDP) models, which utilizes only a small subset of the whole data set to update the associated parameters in the TDP mixture distribution iteratively, and gradually approach the optimal TDP mixture in the sense of minimizing the KL-divergence between the unknown target distribution generating the data and the TDP mixture distribution. This sequential learning approach can be incorporated into the adaptive SMC sampler to enhance the flexibility of the proposal distribution. Simulation studies are provided to demonstrate the efficiency of our proposed methods.

One of our primary goal in this study is to use adaptive Monte Carlo methods to improve the estimation of the marginal likelihood in Bayesian inference. We present the adaptive Monte Carlo based marginal likelihood computation method. The adaptive MCMC method is used to draw samples from the target distribution, while a TDP mixture model is tuned by the proposed sequential learning approach utilizing these samples. Finally, the well tuned TDP mixture model serves as the importance function for marginal likelihood computation. Both synthetic example and a real world application in Bayesian Exoplanet Searches are presented to demonstrate the performance of the proposed method.

6.1 Adaptive Markov Chain Monte Carlo

Markov chain Monte Carlo methods are widely used to sample from analytically intractable probability distributions arising in statistics (Gilks et al., 1996; Robert and Casella, 2004). The efficiency of MCMC methods is of significant practical importance, and loosely speaking is determined by the convergence rate of the chain and asymptotic variance of ergodic averages, both of which are controlled by the spectral gap of the Markov kernel. Thus the efficiency of MCMC algorithms can depend significantly on the design of the Markov transition kernel; see e.g. (Hastings, 1970; Gelman et al., 1996; Mira, 2001; Roberts and Rosenthal, 2001). However choice of effective kernels and their associated tuning parameters is often difficult in precisely those problems where MCMC is most needed: high dimensional problems where we know little *a priori* about the shape of the (potentially multimodal) target posterior distribution.

Due to the difficulty of obtaining rapidly mixing Markov chains for simulating complicated target distributions, *adaptive* MCMC algorithms have been proposed which use the previous history of the chain to automatically tune or "learn" the pro-

posal distribution parameters during simulation, with the goal of obtaining faster convergence or more efficient estimation (Gelfand and Sahu, 1994; Gilks et al., 1998). In adaptive MCMC, the proposal distribution is continually or periodically modified with the aim of improving efficiency. Although this idea is intuitively appealing, such algorithms generally fail to yield Markov chains, making design of adaptive MC schemes with theoretical convergence guarantees more challenging. Gilks et al. (1998) and Brockwell and Kadane (2005) approach this via regeneration times, at which the kernel may be modified while producing independent tours each generating correct ergodic averages. More recently, Haario et al. (2001) give an ergodic theorem for an adaptive Metropolis scheme based on the Robbins-Munro stochastic approximation algorithm (Robbins and Monro, 1951), and this result has led to significant renewed interest in adaptive algorithms and theory (Andrieu et al., 2005; Andrieu and Moulines, 2006; Erland, 2003; Roberts and Rosenthal, 2007). Recently, Roberts and Rosenthal (2007) provided a simple elegant proof and concise set of conditions under which ergodic theorems can be obtained. One such condition requires that the magnitude of adaptation is continually decreasing in such a way that convergence of the chain to the target distribution in the limit is guaranteed; this kind of algorithm is referred as *diminishing adaptation* by Erland (2003). The other is a *bounded convergence* condition, which essentially guarantees that all transition kernels considered have bounded convergence time.

In this work, we describe a general approach to the design of adaptive MCMC algorithms which utilizes a mixture distribution for the proposal kernel, and adapts the parameters of this proposal distribution to minimize Kullback-Leibler divergence from the target distribution. We illustrate our approach using a Metropolized independence sampler (MIS) (Hastings, 1970; Tierney, 1994), a special case of the Metropolis algorithm where the proposal is independent of the current state. (The

method described here utilizes the stochastic approximation approach of Ji (2006). Andrieu and Moulines (2006) have proposed a closely related method for adapting MIS mixtures using KL divergence, although to our knowledge it has not been applied to the variable selection problems studied here.)

6.1.1 Adaptive Metropolized Independence Sampler

Performance of MIS samplers is strongly dependent on the proposal distribution selected. Our adaptation strategy tunes the parametrized proposal distribution to approximate the target distribution in the sense of minimizing Kullback-Leibler (KL)-divergence. Thus for independence proposal density $q(x;\psi)$ with parameters ψ, and target distribution $\pi(x)$, we wish to find the optimal parameters ψ^* which minimize $\mathcal{D}\left[\pi(x) \parallel q(x;\psi)\right] = \mathbb{E}_\pi\left[\log \frac{\pi(x)}{q(x;\psi)}\right]$. Then ψ^* is obtained as a root of the derivative of $\mathcal{D}[\pi(x) \parallel q(x;\psi)]$:

$$h(\psi) = -\int \frac{\pi(x)}{q(x;\psi)} \frac{\partial}{\partial \psi} q(x;\psi) = 0 \tag{6.1}$$

where we assume the integrand is continuous. Exact solution of the integral equation (6.7) is generally intractable, as $h(\psi)$ involves an integral with respect to the target distribution $\pi(x)$ which cannot be calculated directly. However, denoting $f(x,\psi) = \frac{\partial}{\partial \psi}[\log \frac{\pi(x)}{q(x;\psi)}]$ and assuming $f(x,\psi) \in L_2(\pi)$, we can approximate $h(\psi)$ by Monte Carlo integration:

$$h(\psi) \approx \frac{1}{K} \sum_{k=1}^{K} f(X^{(k)}, \psi),$$

where $X^{(k)} \sim \pi(x)$.

When $q(x;\psi)$ is in the exponential family, so $q(x;\psi) = c(x)\exp\left(t(x)'\psi - A(\psi)\right)$ in canonical form with natural parameter ψ, we obtain $\int \pi(x)t(x) = \frac{\partial}{\partial \psi} A(\psi)$, which

says that we should match the expected sufficient statistics under π to the moments of q. However $E_\pi(t(x))$ is an integral of the general form we are constructing the MCMC algorithm to calculate in the first place, and thus assumed to be analytically intractable. Instead, we adaptively match the moments of q to a Monte Carlo approximation of $E_\pi(t(x))$ based on the current sample history.

Let $\hat{h}(X^{(1:K)};\psi)$ denote the estimate of $h(\psi)$ based on the previous sample path $X^{(1:k)}$ from $\pi(x)$, which can be therefore viewed as a noisy 'observation' of $h(\psi)$. A common approach to obtaining roots of $h(\psi) = 0$ when only noisy evaluations of $h(\psi)$ is the Stochastic Approximation (SA) algorithm (Robbins and Monro, 1951; Kushner and Yin, 1997). Stochastic approximation is an iterative algorithm expressed as

$$\begin{aligned}\psi_{n+1} &= \psi_n + r_{n+1}(h(\psi_n) + \xi_{n+1}) \\ &= \psi_n + r_{n+1}\,\hat{h}(X_n^{(1:K)};\psi_n),\end{aligned} \tag{6.2}$$

where $X_n^{(1:K)} \sim \pi(x)$ are samples generated by Metropolis MCMC with proposal distribution $q(x;\psi_n)$, $\{\xi_n\}$ is a sequence of 'noise' (where the Monte Carlo estimate $\hat{h}(X_n^{(1:K)};\psi_n)$ can be interpreted as $h(\psi_n)+\xi_n$), and $\{r_n\}$ is a sequence of decreasing step-sizes satisfying $\sum_n r_n = \infty$ and $\sum_n r_n^2 < \infty$.

In our case, SA can be viewed as performing an iterative gradient descent, with Monte Carlo approximation of the gradient at each iteration. It is easily verified that when q is an exponential family distribution $\mathcal{D}[\pi(x) \parallel q(x;\psi)]$ is convex. Then the sequence $\{\psi_n\}$ defined by equation (6.8) converges to the unique root of equation (6.7) under mild conditions on $\{\xi_n\}$ and $\{r_n\}$ (Andrieu et al., 2005). However Andrieu and Moulines (2006) also show that an adaptive proposal $q(x;\psi)$ for MIS with ψ unrestricted does not guarantee convergence of the algorithm. A straightforward solution due to Haario et al. (2001) is to use an additional fixed mixture component $q(x;\zeta)$ which is not modified during the adaptive updating; in what follows

we take $q(x;\zeta) = N(x;\tilde{\mu},\tilde{\Sigma})$ for some fixed $(\tilde{\mu},\tilde{\Sigma})$. As a simple illustrative example, choosing the adaptive proposal distributions $q(x;\psi)$ to be normal $N(x;\mu,\Sigma)$ with parameters $\psi = (\mu,\Sigma)$ yields the following algorithm:

Algorithm 6.1: Adaptive Metropolized Independence Sampler

- Initialization: Choose $\psi_0 = (\mu_0,\Sigma_0)$ and set $n = 0$.
- Iteration $n+1$:

  1. Simulate K samples $X_{n+1}^{(1:K)}$ by MIS wrt $\pi(x)$ with proposal distribution

  $$q_n = \lambda N(x;\tilde{\mu},\tilde{\Sigma}) + (1-\lambda)N(x;\mu_n,\Sigma_n)$$

  2. Update the parameters of the adaptive proposal by

  $$\mu_{n+1} = \mu_n + r_{n+1}\left[\frac{1}{K}\sum_{k=1}^{K}\left(X_{n+1}^{(k)} - \mu_n\right)\right]$$

  $$\Sigma_{n+1} = \Sigma_n + r_{n+1}\left[\frac{1}{K}\sum_{k=1}^{K}\left(X_{n+1}^{(k)} - \mu_n\right)\left(X_{n+1}^{(k)} - \mu_n\right)^T - \Sigma_n\right]$$

  where r_{n+1} is the step-size of the SA algorithm.

The covariance update of Step 2 is similar to that of (Haario et al., 2001), but as we use an independence proposal rather than a random walk proposal the mean is also approximated. The above adaptive MCMC algorithm satisfies the *diminishing adaptation* condition of Roberts and Rosenthal (2007) as long as the step-size sequence $r_n \to 0$, and the *bounded convergence* is satisfied for $\lambda > 0$, ensuring asymptotic convergence and a weak law of large numbers for this algorithm (Roberts and Rosenthal, 2007).

6.1.2 Adaptive MIS with Mixture Proposal Distribution

When the adaptive Metropolized independence sampler given above is applied to sample from a multimodal target distribution, it will generally perform poorly due to the difficulty in approximating the posterior with a unimodal q distribution. An alternative is to take q to be a mixture distribution, and adapt the mixture component parameters to approximate the multimodal target distribution by minimizing KL-divergence. This results in an adaptive proposal of the form $q(x) = \lambda N(x; \tilde{\mu}, \tilde{\Sigma}) + (1-\lambda)\sum_{m=1}^{M} w_m N(x; \mu_m, \Sigma_m)$, where $\psi = (w_{1:M}, \mu_{1:M}, \Sigma_{1:M})$ are the parameters to be adapted, and M is the number of mixture components. M can be set to a relatively large number to ensure that the proposal can adequately cover multiple modes. Then Ji (2006) derived the adaptation strategy:

Algorithm 6.2: Adaptive MCMC with Mixture Distribution Proposal

- Initialization: Choose $(w_0, \mu_0, \mathbf{\Sigma}_0) = \{w_{i,0}, \mu_{i,0}, \Sigma_{i,0}; i = 1, 2, \ldots, M\}$, and set $n \leftarrow 0$.
- Iteration $n+1$:
  1. Simulate a new state X_{n+1} by MIS wrt $\pi(x)$ with proposal distribution

$$q_n(x) \;=\; \lambda\, q_0(x; \tilde{\mu}, \tilde{\Sigma}) \;+\; (1-\lambda)\, q(x; w_n, \mu_n, \mathbf{\Sigma}_n)$$

  2. Update the parameters w_{n+1} (or $\{w_{i,n+1}\}$) by

$$\widetilde{w}_{i,n+1} = w_{i,n} + r_{i,n+1}(O_i(X_{n+1}) - 1) \tag{6.3}$$

$$w_{i,n+1} = \widetilde{w}_{i,n+1} / \sum_{i=1}^{M} \widetilde{w}_{i,n+1} \tag{6.4}$$

3. Update the parameters $(\mu_{n+1}, \mathbf{\Sigma}_{n+1})$ (or $\{\mu_{i,n+1}, \Sigma_{i,n+1}\}$) by

$$\mu_{i,n+1} = \mu_{i,n} + \alpha_{i,n+1}\left(X_{n+1} - \mu_{i,n}\right) \tag{6.5}$$

$$\Sigma_{i,n+1} = \Sigma_{i,n} + \alpha_{i,n+1}\left[\left(X_{n+1} - \mu_{i,n}\right)\left(X_{n+1} - \mu_{i,n}\right)^T - \Sigma_{i,n}\right] \tag{6.6}$$

where

$$O_i(X_{n+1}) = \frac{N(X_{n+1}; \mu_{i,n}, \Sigma_{i,n})}{\sum_{m=1}^{M} w_{m,n} N(X_{n+1}; \mu_{m,n}, \Sigma_{m,n})}$$

and $\alpha_{i,n+1} = r_{i,n+1} w_{m,n} O_i(X_{n+1})$.

For notational simplicity (6.3-6.6) show using a single sample $X^{(n+1)}$ for updating parameters, although as discussed previously using $K > 1$ samples samples $X_{n+1}^{(1:K)}$ will enable the SA algorithm converge more smoothly.

6.1.3 Extensions

For the Bayesian variable selection problem, we use a family of proposals containing both a point mass component and a Gaussian mixture family. Under our adaptation strategy, the mixing weight of the point mass component adapts to approximate the posterior inclusion probability of its associated variable, while the Gaussian mixture distribution approximates the non-zero component of the coefficient's posterior distribution. This mixture proposal enables the efficient mixing between models with and without the variable included, and the resulting sampling scheme performs parameter estimation and variable selection simultaneously (Ji and Schmidler, 2008).

When performing Bayesian variable selection using priors of the form $(1 - p)\delta_0(\cdot) + p\mathcal{N}(\cdot|0, \sigma)$, the resulting conditional posterior is a mixture of point mass

and an normal-likelihood product. When this conditional distribution is not available in closed form (e.g. due to nonlinearity or non-conjugacy), sampling from the posterior via MCMC can be difficult. In particular, random-walk Metropolis can converge very slowly due to multimodality, and an MIS sampler will perform poorly unless the proposal distribution can be chosen in advance to closely approximates the target distribution. However, the adaptive mixture MIS algorithm described in the previous section can successfully handle both of these difficulties. We need simply to modify the family of proposal mixture distributions to include both point mass and normal components:

$$q(x) = \lambda N(x; \tilde{\mu}, \tilde{\Sigma}) + (1 - \lambda)\Big[w_0\delta(x) + \sum_{m=1}^{M} w_m N(\mu_m, \Sigma_m)\Big],$$

where the parameters $\psi = (w_{0:M}, \mu_{1:M}, \Sigma_{1:M})$ can be tuned using an adaptive scheme similar to that of the previous section.

6.1.4 Example

We present a simple concrete example to illustrate the performance of our adaptive MIS with point mass mixture proposal. Suppose we consider inclusion or exclusion of a single parameter, with posterior posterior distribution given by point mass mixture

$$\pi(x) = 0.3\delta(x) + 0.7N(x; 5, 1).$$

We apply the adaptive MIS Algorithm 3 to sample from this target distribution $\pi(x)$. We set $M = 1$, making the proposal distribution of the form $w_1\delta + w_2 N(\mu, \sigma)$. Therefore for this illustrative example, q and π are of the same parametric (w_1, μ, σ) family, and it is expected that q will converge to π.

Figures 6.1- 6.2 show results on this simple example, using initial parameter

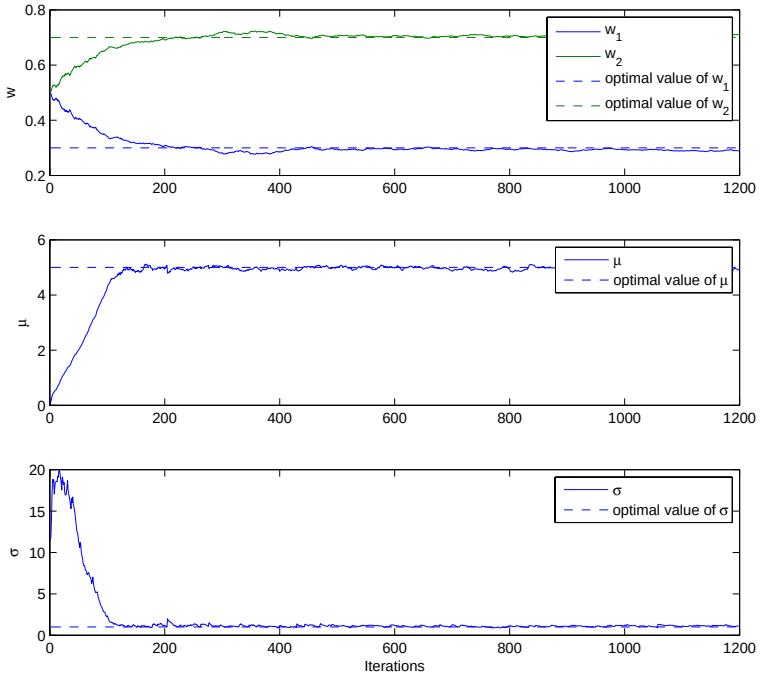


Figure 6.1: Trace plots of proposal distribution parameters. Under the adaptive MIS algorithm described in text, proposal parameters converge under to their optimal values: $w = [0.3, 0.7]$, $\mu^* = 5$, $\sigma^* = 1$.

values $w_1 = 0.5$, $w_2 = 0.5$, $\mu = 0$, and $\sigma = 10$, and SA step-size $r_n = 0.1/n$. Figures 6.1 shows traces of proposal distribution parameters (w_1, w_2, μ, σ), where it can be seen that all parameters converge to their respective optimal values $w_1^* = 0.3, w_2^* = 0.7, \mu^* = 5$, and $\sigma^* = 1$, and therefore the proposal distribution converges to the target distribution. Figure 6.2 compares the performance of this adaptive scheme with a non-adaptive Metropolized independence sampler using fixed proposal $0.5\delta + 0.5N(0, 100)$, via posterior histograms and autocorrelation plots. The adaptive algorithm is seen to perform significantly better.

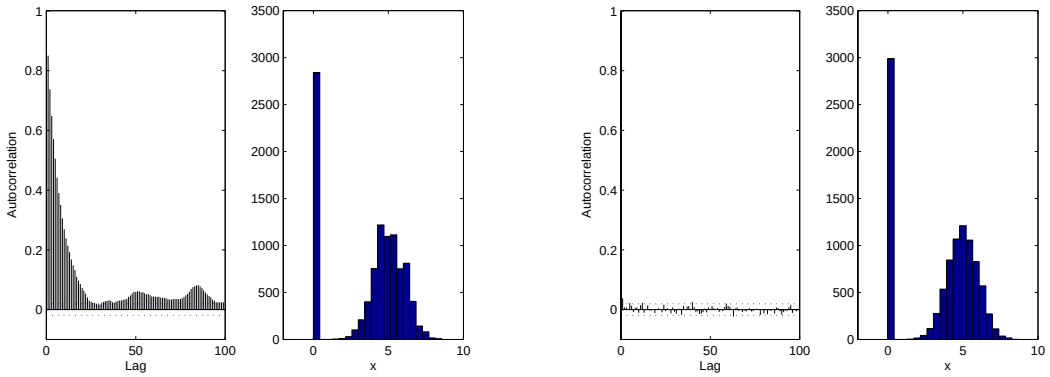


Figure 6.2: Autocorrelation and posterior histogram for toy example obtained from (a) MIS algorithm with fixed proposal distribution, versus (b) adaptive MIS algorithm with point mass mixture proposal.

6.2 Sequential Learning for DP Mixtures

In this section, we discuss some sequential learning methods for DP mixture. One motivating example arises from the adaptive Monte Carlo methods discussed in last section. When the adaptive Metropolized independence sampler given above is applied, in order to sample from a multimodal target distribution, it will generally perform poorly due to the difficulty of approximating the posterior with a unimodal q distribution. An intuitive idea is to take q to be a mixture distribution, and adapt the mixture component parameters to approximate the multimodal target distribution by minimizing the KL-divergence (Ji, 2006). This results in an adaptive proposal of the form $q(x) = \lambda N(x; \tilde{\mu}, \tilde{\Sigma}) + (1-\lambda)\sum_{m=1}^{M} w_m N(x; \mu_m, \Sigma_m)$, where $\psi = (w_{1:M}, \mu_{1:M}, \Sigma_{1:M})$ are the parameters to be adapted, and M is the number of mixture components. Although M can be set to a relatively large number to ensure that the proposal can adequately cover multiple modes, a nonparametric proposal is more elegant since it can provide more flexibility. As a result, we require methods to sequentially tune the nonparametric proposal distribution by learning from the history information of samples.

Another motivating application arises from fitting a nonparametric model like DP mixture model for large data set, for example in flow cytometry study (Chan et al., 2008), where the number of data points may range from thousands to millions. Therefore, using conventional approaches such as MCMC, or variational methods on such data set requires an extremely long time. However, by using a subset of the data and sequentially learning the nonparametric model to fit the distribution of the full data set, we can significantly reduce the computation time.

Specifically, assume a set of data $X = [x_1, ..., x_N]$ subjected to an unknown distribution $\pi(\cdot)$. Our goal is to fit a nonparametric model for the full data set X or tune the nonparametric distribution to approximate $\pi(\cdot)$ in a sequential fashion: at each iteration we choose X_t, a subset of X selected either randomly or by design, then update the nonparametric distribution by learning from X_t.

Inspired by the adaptive MCMC with mixture proposal (Ji, 2006; Ji and Schmidler, 2008) as well as stochastic approximation version of EM (Celeux and Diebolt, 1992; Delyon et al., 1999), a sequential learning approach for fitting a DP mixture model is proposed in this work. The idea is to minimize the KL-divergence between $\pi(\cdot)$ and a nonparametric DP mixture distribution. Owing to the advantage of truncated DP mixtures, we can obtain a closed form update formula to iteratively update the parameters of the truncated DP mixtures.

6.2.1 Stochastic Approximation for DP Mixtures

The idea of the stochastic approximation for sequential learning for DP mixtures comes from the adaptation strategy that tends to minimize the KL-divergence between the distribution of interest and the proposal in mixture fashion in adaptive MCMC methods (Ji, 2006; Ji and Schmidler, 2008). More specifically, the adaptive proposal is in the form of $q(x) = \lambda N(x; \tilde{\mu}, \tilde{\Sigma}) + (1 - \lambda) \sum_{m=1}^{M} w_m N(x; \mu_m, \Sigma_m)$,

where $\psi = (w_{1:M}, \mu_{1:M}, \Sigma_{1:M})$ are the parameters to be adapted, and M is the predefined number of mixture components. We propose a stochastic approximation based sequential leaning approach to tune parameters $(w_{1:M}, \mu_{1:M}, \Sigma_{1:M})$ by gradually learning from samples of the unknown distribution $\pi(\cdot)$, enabling $q(\cdot)$ to approximate $\pi(\cdot)$ in the sense of minimizing the KL-divergence $\mathcal{D}[\pi(\cdot)|q(\cdot)]$.

The idea of our sequential learning approach is to tune the nonparametric proposal distribution, expressed as a truncated DP mixture, in order to approximate the target distribution in the sense of minimizing KL-divergence. Thus for proposal density $q(x;\psi)$ with parameters ψ, and target distribution $\pi(x)$, we wish to find the optimal parameters ψ^* which minimize $\mathcal{D}\left[\pi(x) \parallel q(x;\psi)\right] = \mathbb{E}_\pi\left[\log \frac{\pi(x)}{q(x;\psi)}\right]$. Then ψ^* is obtained as a root of the derivative of $\mathcal{D}[\pi(x) \parallel q(x;\psi)]$:

$$h(\psi) = -\int \frac{\pi(x)}{q(x;\psi)} \frac{\partial}{\partial \psi} q(x;\psi) = 0. \tag{6.7}$$

Exact solution of the integral equation (6.7) is generally intractable, as $h(\psi)$ involves an integral with respect to the target distribution $\pi(x)$ which is usually a complex distribution and cannot be calculated directly. However, since we can obtain samples $X_{1:T}$ from $\pi(x)$ at each time index t, denoting $f(x,\psi) = \frac{\partial}{\partial \psi}[\log \frac{\pi(x)}{q(x;\psi)}]$ and assuming $f(x,\psi) \in L_2(\pi)$, we can approximate $h(\psi)$ by Monte Carlo integration:

$$h(\psi) \approx \frac{1}{M} \sum_{k=1}^{M} f(x_t^{(k)}, \psi)$$

where $X_t = \{x_t^{(k)} : x_t^{(k)} \sim \pi(x)\}$.

Let $\hat{h}(x^{(1:K)};\psi)$ denote the estimate of $h(\psi)$ based on $x_t^{(1:k)}$ sampled from $\pi(x)$, which can be therefore viewed as a noisy 'observation' of $h(\psi)$. One available approach to obtaining roots of $h(\psi) = 0$ when only noisy evaluations of $h(\psi)$ is the Stochastic Approximation (SA) algorithm (Robbins and Monro, 1951; Kushner and

Yin, 1997). The SA algorithm iteratively updates ψ to approximate its optimal values by the following formula,

$$\begin{aligned}\psi_{t+1} &= \psi_t + r_{t+1}(h(\psi_t) + \xi_{t+1}) \\ &= \psi_t + r_{t+1}\,\hat{h}(x_t^{(1:K)}; \psi_t)\end{aligned} \tag{6.8}$$

where $x_t^{(1:K)} \sim \pi(x)$ is our observed data, $\{\xi_t\}$ is a sequence of 'noise' (thus the Monte Carlo estimate $\hat{h}(x_t^{(1:K)}; \psi_t)$ can be interpreted as $h(\psi_t) + \xi_t$), and $\{r_t\}$ is a sequence of decreasing step-sizes satisfying $\sum_t r_t = \infty$ and $\sum_t r_t^2 < \infty$.

We assumed that $q(x; \psi)$ is a truncated DP normal mixture, which can be expressed as, $q(x; \psi) = \sum_{k=1}^{K} w_k N(\cdot|\mu_k^*, \Sigma_k^*)$, where $w_k = V_k \prod_{j=1}^{k-1}(1 - V_j)$. Let ψ denote the set $(V_k, \mu_k^*, \Sigma_k^*)$. The the partial derivative of $\mathcal{D}[\pi(x) \parallel q(x; \psi)]$ with respect to V_k, μ_k^* and Σ_k^* can be derived as follows (refer to Appendix E for details of derivation),

$$h_{V_k}(x; \psi) = \int \pi(x) \frac{-\sum_{l=k+1}^{K} V_l \prod_{j \le l-1, j \neq k}(1 - V_j) q(x|\mu_l, \Sigma_l) + \prod_{j=1}^{k-1}(1 - V_j) q(x|\mu_k, \Sigma_k)}{\sum_{m=1}^{K} w_m q(x|\mu_m, \Sigma_m)} dx \tag{6.9}$$

$$H_{\mu_k}(x; \psi) \propto \int \pi(x) \frac{w_k q(x; \mu_k, \Sigma_k)}{\sum_{m=1}^{K} w_m q(x; \mu_m, \Sigma_m)} \times (x - \mu_k)\, dx \tag{6.10}$$

$$H_{\Sigma_k}(x; \psi) \propto \int \pi(x) \frac{w_k q(x; \mu_k, \Sigma_k)}{\sum_{m=1}^{K} w_m q(x; \mu_m, \Sigma_m)} \times \Big((x - \mu_k)\,(x - \mu_k)^T - \Sigma_k \Big)\, d\Sigma \tag{6.11}$$

Given the observations $X_t = \{x_t^{(i)}\}_{i=1}^{N_t}$ from $\pi(x)$, the Monte Carlo approximation of these partial derivatives are

$$H_{V_k}(X_t; \psi) = \frac{1}{N_t} \sum_{i=1}^{N_t} \frac{-\sum_{l=k+1}^{K} V_l \prod_{j \le l-1, j \neq k}(1 - V_j) q(x_t^{(i)}|\mu_l, \Sigma_l) + \prod_{j=1}^{k-1}(1 - V_j) q(x_t^{(i)}|\mu_k, \Sigma_k)}{\sum_{m=1}^{K} w_m q(x_t^{(i)}|\mu_m, \Sigma_m)} \tag{6.12}$$

$$H_{\mu_k}(x;\psi) \propto \frac{1}{N_t}\sum_{i=1}^{N_t} \frac{w_k q(x_t^{(i)};\mu_k,\Sigma_k)}{\sum_{m=1}^{K} w_m q(x_t^{(i)};\mu_m,\Sigma_m)} \times \left(x_t^{(i)} - \mu_k\right) \tag{6.13}$$

$$H_{\Sigma_k}(x;\psi) \propto \frac{1}{N_t}\sum_{i=1}^{N_t} \frac{w_k q(x_t^{(i)};\mu_k,\Sigma_k)}{\sum_{m=1}^{K} w_m q(x_t^{(i)};\mu_m,\Sigma_m)} \times \left(\left(x_t^{(i)} - \mu_k\right)\left(x_t^{(i)} - \mu_k\right)^T - \Sigma_k\right) \tag{6.14}$$

Therefore, we can iteratively update V_k, μ_k and Σ_k using the Stochastic Approximation approach, and yielding the following algorithm:

Algorithm 6.3: Sequential Learning for DP Mixtures

- Initialization: Choose $\psi_0 = (V_0, \mu_0, \Sigma_0)$ and set $t = 1$.
- For $t > 1$, update the $V_{k,t}$, $w_{k,t}$, $\mu_{k,t}$ and $\Sigma_{k,t}$ (for $k = 1, ..., K$) as follows,

$$V_{k,t+1} = V_{k,t} + r_{k,n+1} H_{V_k}(X_t;\psi_t) \tag{6.15a}$$

$$w_{k,t+1} = V_{k,t+1}\prod_{j=1}^{k-1}(1 - V_{j,t+1}) \tag{6.15b}$$

$$\mu_{k,t+1} = \mu_{k,t} + \frac{r'_{k,t+1}}{N_t}\sum_{i=1}^{N_t} \alpha_{k,t+1}^{(i)}\left(x_t^{(i)} - \mu_{k,t}\right) \tag{6.15c}$$

$$\Sigma_{k,t+1} = \Sigma_{k,t} + \frac{r'_{k,t+1}}{N_t}\sum_{i=1}^{N_t} \alpha_{k,t+1}^{(i)}\left((x_t^{(i)} - \mu_{k,t})(x_t^{(i)} - \mu_{k,t})^T - \Sigma_{k,t}\right) \tag{6.15d}$$

where

$$H_{V_k}(X_t;\psi_t) = \frac{1}{N_t}\sum_{i=1}^{N_t} \frac{-\sum_{l=k+1}^{K} V_l \prod_{j\le l-1, j\neq k}(1-V_j) q(x_t^{(i)}|\mu_{l,t},\Sigma_{l,t}) + \prod_{j=1}^{k-1}(1-V_j) q(x_t^{(i)}|\mu_{k,t},\Sigma_{k,t})}{\sum_{m=1}^{K} w_{m,t} q(x_t^{(i)}|\mu_{m,t},\Sigma_{m,t})},$$

$\alpha_{k,t+1}^{(i)} = \frac{w_{k,t} q(x_t^{(i)};\mu_{k,t},\Sigma_{k,t})}{\sum_{m=1}^{K} w_{m,t} q(x_t^{(i)};\mu_{m,t},\Sigma_{m,t})}$, $r_{k,n+1}$ and $r'_{k,t+1}$ are the step-size in the stochastic approximation algorithm.

As the iterative steps in the stochastic approximation go to infinity, the estimated V_t, w_t, μ_t and Σ_t converge to the optimal set minimizing the $\mathcal{D}\left[\pi(x) \parallel q(x;\psi)\right]$ with probability one. The proof of this result is straightforward using the same derivation shown in Appendix D.

Example

We demonstrate the behavior of the proposed sequential learning algorithm by applying it to a synthetic data set (shown in Figure 6.3): 5000 data points generated from a mixture of four bivariate normals with weights: $[0.3, 0.4, 0.29, 0.01]$, means: $[-1.75, 0]$, $[0, 0]$, $[2, 1]$, $[5, 5]$ and covariances: $[0.6, 0.5; 0.5, 0.6]$, $[0.4, -0.25; -0.25, 0.4]$, $[0.25, 0.15; 0.15, 2]$, $[0.3, 0.2; 0.2, 0.25]$. In our sequential learning context, we assume that at each iteration the observation is a subset of 500 data points uniformly selected from these 5000 data points.

We wish to fit a TDP mixture model for this data set. The SA based sequential learning algorithm for a TDP mixture model is initialized as follows: the maximum number of components is set as $K = 20$, means of normal components μ_k (for $k = 1, ..., K$) are randomly initialized in range $[-10, 10] \times [-10, 10]$, the covariance of normal components are set as $\Sigma_k = 2\mathbf{I}$ (for $k = 1, ..., K$), and set $V_k = 1/(K - k + 1)$. The step-size in the SA algorithm is set as $r_{k,t} = 1/(t + 50)$ and $r'_{k,t} = 10/(t + 50)$. The algorithm runs as described in Algorithm 6.3 until satisfies some termination condition such as a prespecified total number of iterations, or monitoring iterative changes of the log likelihood. In our simulation study, we simply set the termination condition as a total number of 200 iterations. The output of the sequential learning algorithm is w_k, μ_k, Σ_k for ($k = 1, ..., K'$), where K' is the number of components with non-zero weight. We shows the fitted mixture

model in the final iteration with the data in Figure 6.3. The KL-divergence between the 'true' target distribution and our fitted TDP mixture is monitored and shown in Figure 6.4. Moreover, in Figure 6.5 we show the log likelihood of TDP mixture model at each iteration comparing with the log likelihood of finite mixture model computed by EM algorithm. Finally in Figure 6.6 we show the number of mixture components with non-zero weights per iterations.

The simulation study shows that the proposed sequential learning algorithm can iteratively tune a TDP mixture model to fit the data. The proposed algorithm affords at least two advantages as we expected: 1) it can deal with TDP mixture model, which can provide more flexibility in modelling than a mixture model with a fixed component number; 2) instead of learning from the whole data set, at each iteration, it requires only a subset of the data, reducing the computation costs.

Moreover, as can be observed, the data set contains a small group of approximately 50 data points, which may lead to a difficulty in fitting a mixture for it. For example, when we use the EM algorithm to tune a normal mixture with fixed number of components for this data set, if the prespecified number of mixture components is small, the small group may easily be neglected by the fitting algorithm due to poor initialization; while if the number of mixture components is very large, the mixture model tends to over fit the data. However, in our proposed sequential learning algorithm, we tune the parameters in a TDP mixture based on previous estimation of these parameters and the current observation, therefore even when the number of data points in its group is small we can still 'remember' the mixture component corresponding to it. This property may be helpful in exploring rare subtypes in a large data set, which is of interest in flow cytometry data study (Chan et al., 2008).

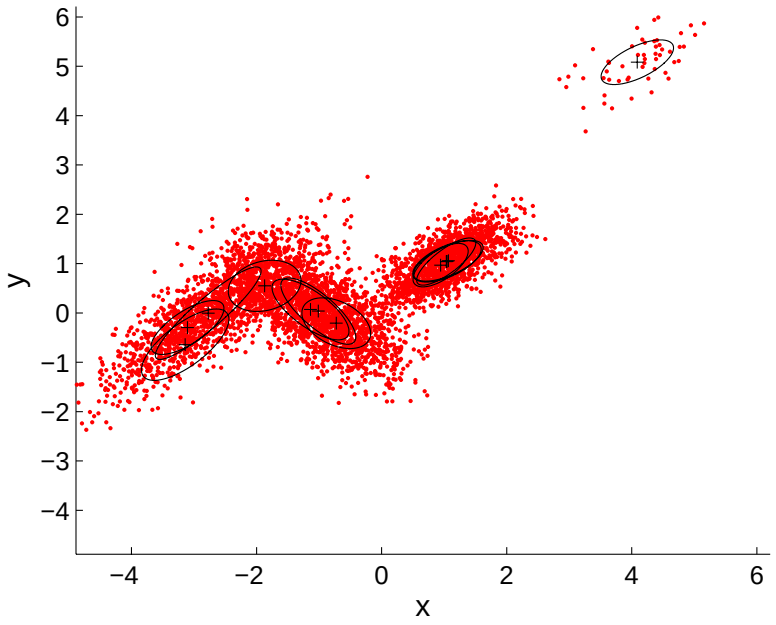


Figure 6.3: Synthetic data points are shown in red dots. The final fitted TDP mixture are presented with + representing the mean of normal component and ellipse representing one standard deviation.

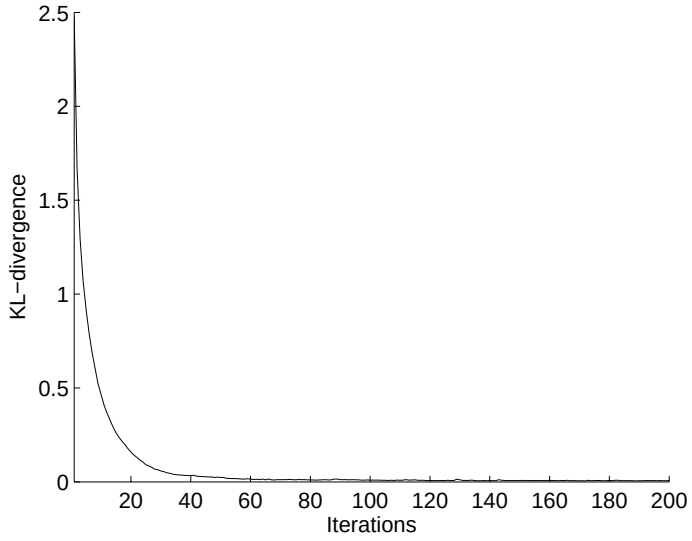


Figure 6.4: The KL-divergence between the true target distribution and the estimated TDP normal mixture model per iteration.

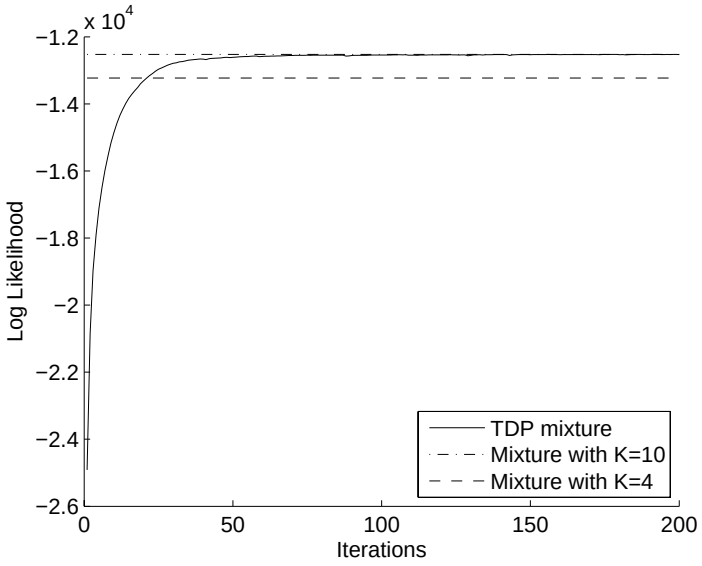


Figure 6.5: Log likelihood of the TDP normal mixture model per iteration is shown by the solid curve. For comparison, the log likelihood of finite normal mixture model with components number $K = 4$ and $K = 10$ computed through the EM algorithm are also shown by the dashed line and the dash-dot line respectively.

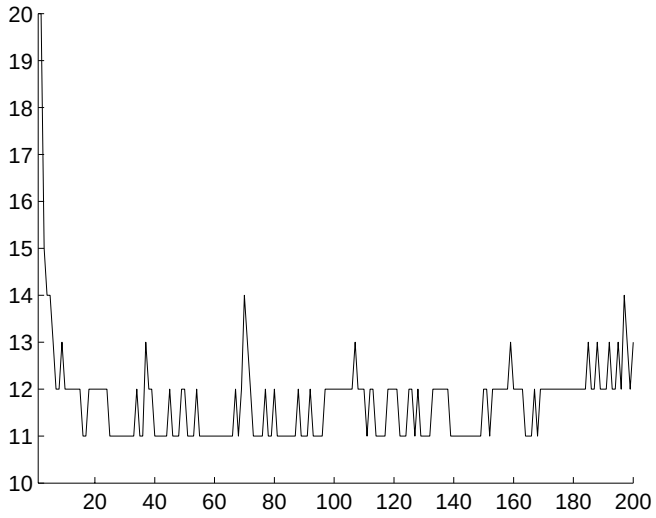


Figure 6.6: Plot of the number of mixture components with non-zero weights per iteration.

6.3 Marginal Likelihood Computation

As discussed in Chapter 5, marginal likelihood, also known as 'Evidence', is a key quantity for Bayesian model evaluation, comparison and selection. Computing marginal likelihood is an integration problem which is generally prohibitive in most applications because of the intractability of the likelihood function. Numerous methods haven been proposed based on various Monte Carlo integration approaches (Newton and Raftery, 1994; Gelfand and Dey, 1994; Chib, 1995; Meng and Wong, 1996; Robert and Casella, 2004; Crooks et al., 2007; Lefebvre et al., 2009), among which importance sampling is one of the most straightforward approaches. In importance sampling, an easy-to-use probability density is designed, called the importance function $g(\theta)$, such that given a set of random samples from it $\{\theta^{(i)}\}_{i=1}^{N}$, the marginal likelihood can be approximated by

$$p(D|M) \simeq \frac{1}{N}\sum_{i=1}^{N}\frac{L(D|\theta^{(i)},M)p(\theta^{(i)}|M)}{g(\theta^{(i)})}. \tag{6.16}$$

The terms $\frac{L(D|\theta^{(i)},M)p(\theta^{(i)}|M)}{g(\theta^{(i)})}$ in (6.16) are often referred to as the importance "weights", denoted $w_i(\theta^{(i)})$.

6.3.1 Marginal Likelihood Computation by Adaptive Importance Sampling

It is a well known fact that the efficiency of importance samplers largely depends on how closely the importance function resembles the shape of the target distribution, $\pi(\theta|D,M) \propto L(D|\theta,M)p(\theta|M)$. According to the analysis in Crooks et al. (2007), if the importance function has thinner tails than the target distribution, the estimate of (6.16) will tend to be unstable since the weights can be arbitrarily large; meanwhile if the importance function has fatter tails, then lots of samples

$\theta^{(i)}$ will be wasted on the area where the target distribution has very small values, resulting to a biased estimate. The best case is an importance function with slightly heavier tails than the target. In a previous study (Crooks et al., 2007), several improved importance sampling approaches were proposed, base on different criteria to tune the importance function to mimic the target distribution. It has been verified that the importance function which minimizes the KL-divergence between the posterior distribution and importance function shows the best performance in marginal likelihood computation (Crooks et al., 2007).

Here, a nonparametric importance function is proposed instead of a traditional parametric importance function to mimic the target distribution, which is supported to provide more flexibility. The process consist of two parts: first, use certain Monte Carlo approach to sample from the posterior efficiently, meanwhile using a sequential learning approach for the nonparametric importance function by iteratively learning from the posterior samples; after the importance function has been well tuned, we draw a large set of samples and evaluate the marginal likelihood through equation (6.16). The approach is named as adaptive importance sampling for marginal likelihood computation, which is summarized as follows,

- Sampling and adaptation:
  - Sampling: sample $\{\theta^{(i)}\}_{i=1}^{N}$ from the posterior distribution $p(\theta|D, M)$.
  - Sequential learning: use a sequential learning approach proposed in 6.2 to adapt the importance function $g(\theta|\psi)$ to approximate $p(\theta|D, M)$.
- ML computation: sampling $\{\theta^{(j)}\}_{j=1}^{N}$ from $g(\theta|\psi^*)$ and evaluate

$$p(D|M) \simeq \frac{1}{N}\sum_{j=1}^{N}\frac{L(D|\theta^{(i)}, M)p(\theta^{(j)}|M)}{g(\theta^{(j)}|\psi^*)} \tag{6.17}$$

where ψ^* denotes the result from sequential learning, i.e. ψ_T in the stochastic approximation for the TDP mixture when T is sufficient large.

Note that if the adaptive MCMC discussed in Section 6.1 is used to sample from $p(\theta|D, M)$, then the sequential learning process can be embedded into the sampling process: after we obtain a block of samples from the posterior using adaptive MCMC, we iteratively update the nonparametric importance function by learning from this data set. Moreover, for sure, we may also use the nonparametric importance function as the adaptive proposal in adaptive MCMC. However, as shown in Andrieu and Moulines (2006), an adaptive proposal $q(x; \psi)$ (even in nonparametric form) for MIS with ψ unrestricted does not guarantee convergence of the algorithm. A straightforward solution is to use an additional fixed mixture component $q(x; \zeta)$ which is not modified during the adaptive updating, as discussed in Section 6.1.

6.3.2 Simulation Study

We demonstrate the proposed approach on a target function $\pi(\cdot)$, which is a outer product of seven univariate distributions, with the marginal likelihood exactly equal to 1. These seven distributions are:

1. $\frac{3}{5}Ga(10 + x|2, 3) + \frac{2}{5}Ga(10 - x|2, 5)$
2. $\frac{3}{4}skN(x|3, 1, 5) + \frac{1}{4}skN(x| - 3, 3, -6)$
3. $T(x|0, 9, 4)$
4. $\frac{1}{2}Be(x + 3|3, 3) + \frac{1}{2}N(x|0, 1)$
5. $\frac{1}{2}\epsilon(x|1) + \frac{1}{2}\epsilon(-x|1)$
6. $skN(x|0, 8, -3)$

7. $\frac{1}{8}N(x|-10,0.1)+\frac{1}{4}N(x|0,0.15)+\frac{5}{8}N(x|7,0.2)$

where $Ga(\cdot|\alpha,\beta)$ denotes the gamma distribution, $N(\cdot|\mu,\sigma)$ denotes the normal distribution, $skN(\cdot|\mu,\sigma,\alpha)$ denotes the skew-normal distribution, $T(\cdot|\mu,\sigma,df)$ denotes the student-T distribution, $Be(\cdot|\alpha,\beta)$ denotes the beta distribution, and $\epsilon(\cdot|\lambda)$ denotes the exponential distribution. From the definition, we can see that dimension 2 has two modes bracketing a deep ravine, dimension 4 has one low, broad mode that overlaps a second sharper mode, and dimension 7 has three distinct, well-separated modes. Only dimension 5 is symmetric.

For the posterior sampling, adaptive MCMC with a mixture distribution proposal discussed in Section 6.1.2 is applied, in which we use a mixture of 5 univariate normal distributions for each of the 7 dimensions. At each iteration, we sample 200 samples from the posterior and use these samples to update the proposal distribution as well as the nonparametric importance function which is a truncated DP mixture of 7 dimensional multivariate normal distribution with truncated components number $K = 40$, each with a mean uniformly chosen from the parameter space and a covariance matrix of $4\mathbf{I}$.

After about 250 iterations, we observed that the changes in the importance function becomes tiny, through monitoring the KL-divergence between the true target distribution $\pi(\cdot)$ and the estimated importance function shown in Figure 6.8. To show the efficiency of our proposed algorithm, we compare the true target distribution, the kernel density estimation of samples drawn by adaptive MCMC algorithm and the fitted TDP normal mixture importance function in univariate style in Figure 6.7. As can be observed, the drawn samples can effectively represent the target distribution, while the fitted TDP normal mixture importance function matches the true target distribution well. Both of these two factors enable us to achieve a good performance in IS based marginal likelihood estimation. Moreover, the poste-

rior samples obtained by using the adaptive Metropolis independence sampler are shown in Figure 6.9, as well as the fitted TDP normal mixture importance function. After the importance function has been well tuned, we draw $100,000$ samples from the importance function and evaluate the marginal likelihood via equation (6.16).

For comparison, we also implement the adaptive importance sampling (AIS) studied in Cappé et al. (2008), in which the importance function is adapted through a EM algorithm by utilizing samples form importance sampling with previous tuned importance function. The initial IS density q_0 is chosen similarly as the setting in sequential learning algorithm: a mixture of 40 Gaussian components, each with a mean uniformly chosen from the parameter space and a covariance matrix of $4 \times \mathbf{I}$. The number of samples for each IS iteration is $N = 100,000$, while the total iteration times are set equally to be 50.

We summarize the comparison simulation results as follows: the *ESS* for the IS marginal likelihood estimator is 56723 in our proposal algorithm, compared with 1265.5 in the AIS algorithm; scatter plots shown in Figure 6.9 (resp. Figure 6.10) are used to present the bivariate posterior of the model parameters. The marginal likelihood evaluated by our algorithm is 0.9988 ± 0.0036 compared with 0.6257 ± 0.024 by AIS algorithm. Moreover, our algorithm require drawing $600,000$ samples compared with $5000,000$ for the AIS algorithm, showing an advantage in computational cost.

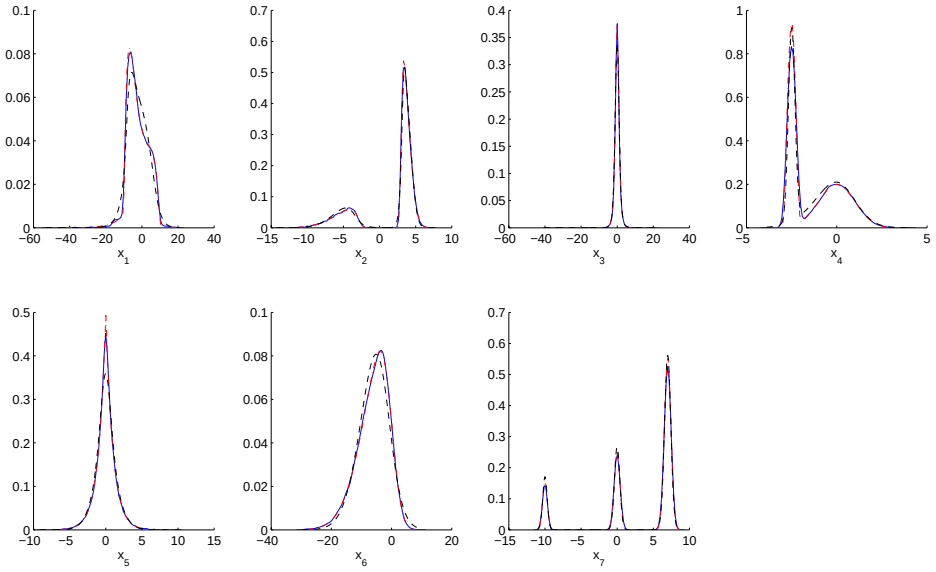


Figure 6.7: Plots of three density functions for comparison: the true univariate density of each dimension is shown by the red dashed dot curve; kernel density estimation of samples drawn by the adaptive MCMC algorithm is shown by the solid blue curve; a well tuned truncated DP mixture nonparametric importance function is shown by black dashed curve.

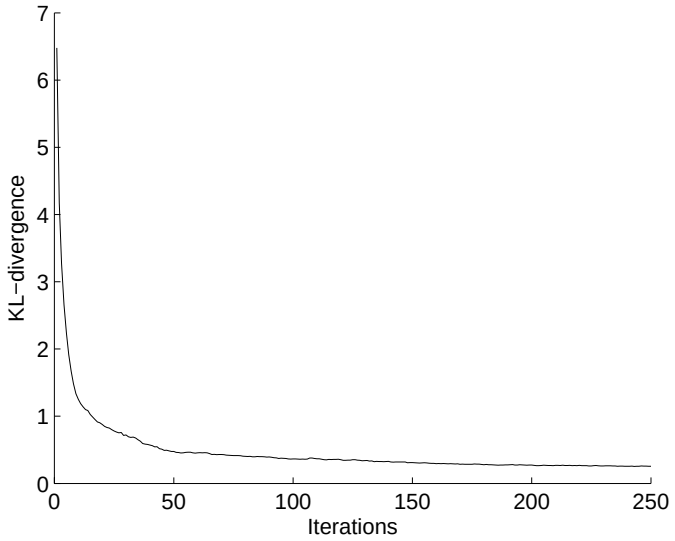


Figure 6.8: The KL-divergence between the true target distribution and the estimated TDP normal mixture model per iteration.

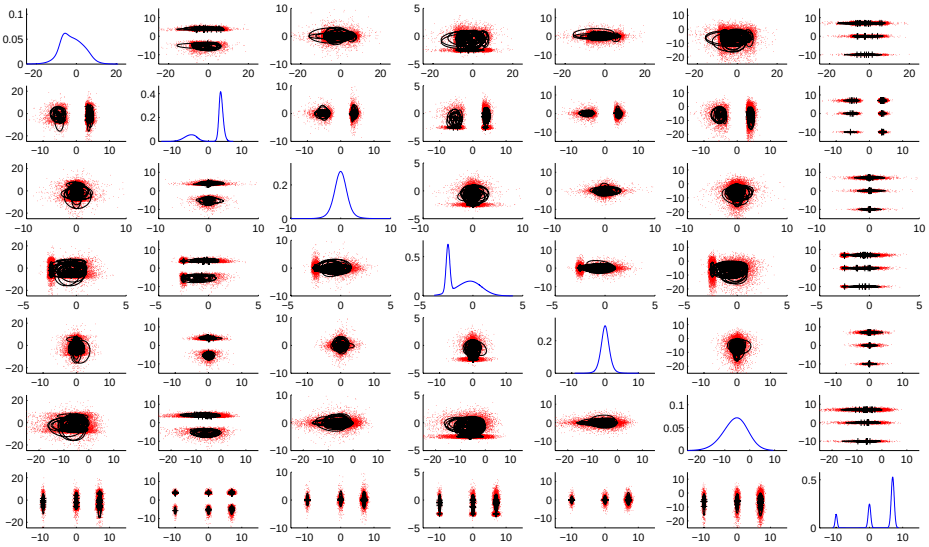


Figure 6.9: Scatter plot of samples obtained using the adaptive Metropolis independence sampling. The final fitted TDP mixture is presented in a bivariate style, with + representing the mean of normal component and ellipse representing the standard deviation, as well as in univariate style by the blue curve.

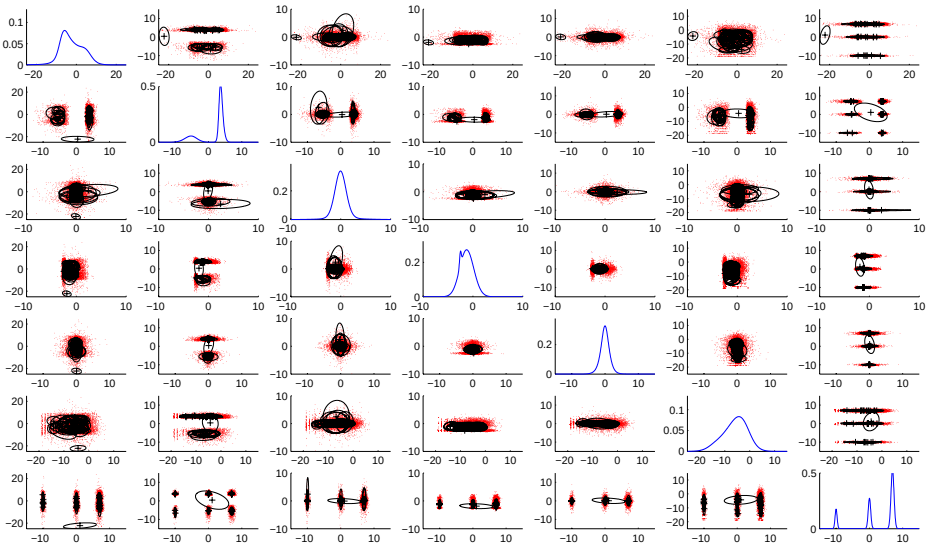


Figure 6.10: Scatter plot of samples obtained using the AIS algorithm. The fitted mixture by EM algorithm is presented in a bivariate style, with + representing the mean of normal component and ellipse representing the standard deviation, as well as in univariate style by the blue curve.

6.4 Application in Bayesian Exoplanet Searches

Extra-solar planet (Exoplanet) science was motivated by the famous Fermis paradox "where is everybody? where are they?" Webb (2002) and is of great importance to help us understand the origin and evolution of the Solar System. In this section, we use the advanced Bayesian computational method proposed in Section 6.3 to draw inferences on the number of planet, if any exist, based on the radial velocity (RV) data set observed by astronomers.

6.4.1 The Velocity-shift Model

The velocity-shift model has been described in detail in Crooks et al. (2007) and Bullard (2009). Here we give a brief introduction. The astronomers data are $y_i \triangleq \{t_i, v_i, \sigma_i\}$, $i \in 1, \ldots, o$, where t_i indexes the time of the ith observation, v_i is the observed velocity at time t_i, and σ_i is estimated error of the velocity observation. Moreover, we assume a source of (presumably) Gaussian error from "stellar jitter", the random fluctuations in the luminosity of a star. The observed velocities will be related to the orbital parameters through

$$v_i \sim \mathcal{N}\left(C + \triangle V(t_i|\phi), \sigma_i^2 + s^2\right), \tag{6.18}$$

where C is the constant center-of-mass velocity of a star relative to Earth, $\triangle V(t_i|\phi)$ is a function to be defined below, ϕ represents a set of orbital parameters, and s^2 is the "stellar jitter" variance. The set of parameters $\{C, s^2, \phi\}$ is denoted by θ.

The No-planets Model: $\mathbf{M}_0$

If no planets orbiting a star, then the function $\triangle V$ is zero for all values of t:

$$V_i(t) = 0, \forall t \in \mathbb{R} \tag{6.19}$$

In this case, only two elements in θ need to be estimated: C and s^2. So the no-planet model is

$$v_i \sim \mathcal{N}\left(C, \sigma_i^2 + s^2\right). \tag{6.20}$$

The Single Planet Model: $\mathbf{M}_1$

If a single planet orbiting a star, then ϕ will contain five components: $\phi = (K, P, e, \omega, M_0)$, where $K \geq 0$ is the velocity semi-amplitude, $P > 0$ is the orbital period, $0 \leq e \leq 1$ denotes eccentricity, $0 \leq \omega \leq 2\pi$ argument of periastron and $0 \leq M_0 \leq 2\pi$ denotes the mean anomaly at time $t = 0$. Given these parameters, the velocity shift at time t is then just

$$\triangle V(t|\phi) = K[\cos(\omega + T(t)) + e\cos(\omega)], \tag{6.21}$$

where $T(t)$ is the "true anomaly at time t" and can be given by

$$T(t) = 2\arctan\left[\tan(\frac{E(t)}{2})\sqrt{\frac{1+e}{1-e}}\right], \tag{6.22}$$

where $E(t)$, called the "eccentric anomaly at time t", is the solution of the transcendental equation

$$E(t) - e\sin(E(t)) = \text{mod}\left(\frac{2\pi}{P}t + M_0, 2\pi\right). \tag{6.23}$$

Priors on Parameters

We describe prior distributions of parameters in the velocity-shift model. These priors were recommended by Ford and Gregory (2006) based on both mathematical convenience and their approximate realism.

A: No-planet model

For the no-planet model shown in (6.20), we choose independent prior for C and s,

$$\pi_C(C) = \begin{cases} \dfrac{1}{C_{max} - C_{min}}, & \text{for} \quad C_{min} \leq C \leq C_{max} \\ 0, & \text{otherwise,} \end{cases} \tag{6.24}$$

and

$$\pi_s(s) = \begin{cases} \dfrac{1}{\log\left(1 + \frac{s_{max}}{s_0}\right)} \cdot \dfrac{1}{s_0 + s}, & \text{for} \quad 0 < s \leq s_{max} \\ 0, & \text{otherwise.} \end{cases} \tag{6.25}$$

Thus, the complete prior is given by:

$$\pi_0(C, s) = \frac{1}{C_{max} - C_{min}} \cdot \frac{1}{\log\left(1 + \frac{s_{max}}{s_0}\right)} \cdot \frac{1}{s_0 + s}, \tag{6.26}$$

and where $C_{min} \leq C \leq C_{max}, 0 < s \leq s_{max}$.

B: Single planet model

For analysis of model parameters, such as posterior sampling, it is convenient to work in a parameter space where the parameters are not so highly correlated as they are in the space defined by $(C, K, P, e, \omega, M_0, s)$ in the single planet model. It is also convenient to have the posterior in a shape like Gaussian distribution. For these reasons, some useful transformation of variables are given (Bullard, 2009),

- Translate t_i (essentially reparameterizing M_0) so that $t = 0$ occurs in the middle of all observations: we place it at the weighted mean of the observation times, with the weights being inversely proportional to the measurement errors. This transformation reduces correlations between M_0 and P.

- Use Poincar é variables $x \equiv e\cos\omega$ and $y \equiv e\sin\omega$ instead of e and ω to reduce correlations between M_0 and ω. This transformation is particularly important for low eccentricity orbits.

- Use $z \equiv (\omega + M_0)\text{mod}2\pi$ instead of M_0 to reduce correlations between these two parameters when $e \ll 1$.

Besides, it is useful to work in $\dot{P} \equiv \log P$ rather than in P, in $\dot{K} \equiv \log K$ rather than in K, and in $\dot{s} \equiv \log s$ rather than in s.

In this transformed space, the prior given for the one-planet model with a full Keplerian orbit can be expressed as,

$$\pi_1(C, \dot{K}, \dot{P}, x, y, z, \dot{s}) = \kappa_1 \cdot \exp\dot{K} \cdot \frac{1}{1 + \frac{\exp\dot{K}}{K_0}} \cdot \frac{1}{\sqrt{x^2 + y^2}} \cdot \exp\dot{s} \cdot \frac{1}{1 + \frac{\exp\dot{s}}{s_0}}, \quad (6.27)$$

where

$$\kappa_1 = \frac{1}{C_{max} - C_{min}} \cdot \frac{1}{\log\left(1 + \frac{K_{max}}{K_0}\right)} \cdot \frac{1}{K_0} \cdot \frac{1}{\log\left(\frac{P_{max}}{P_{min}}\right)} \cdot \left(\frac{1}{2\pi}\right)^2 \cdot \frac{1}{\log\left(1 + \frac{s_{max}}{s_0}\right)} \cdot \frac{1}{s_0}, \quad (6.28)$$

where $C_{min} \leq C \leq C_{max}$, $\dot{K} \leq \log(K_{max})$, $\log(P_{min}) \leq \dot{P} \leq \log(P_{max})$, $x^2 + y^2 < 1$, $0 \leq z \leq 2\pi$, and $\dot{s} \leq \log(s_{max})$.

C: Values of the constants

The choices of constants are partially based on physical realities, (e.g. an orbit which is too small will engender the planet getting consumed by the star),

$$P_{\min} \equiv 1 \text{ day} \tag{6.29a}$$

$$P_{max} \equiv 1000 \text{ years} \tag{6.29b}$$

$$K_0 \equiv 1 \text{ ms}^{-1} \tag{6.29c}$$

$$K_{max} \equiv 2128 \text{ ms}^{-1} \tag{6.29d}$$

$$C_{min} \equiv -K_{max} \tag{6.29e}$$

$$C_{max} \equiv K_{max} \tag{6.29f}$$

$$s_0 \equiv 1 \text{ ms}^{-1} \tag{6.29g}$$

$$s_{max} \equiv K_{max} \tag{6.29h}$$

6.4.2 Marginal Likelihoods

It is of great interest for astronomers to schedule telescope time in order to maximize the probability of a significant observation such as determining the total number of planets in the system or detecting at least one planet. From a statistician viewpoint, finding the number of planets in a system is a model selection problem. Under Bayesian model selection, we need to calculate of marginal likelihoods of two or more models in order to calculate Bayes factors such as

$$BF(M_{p_1}, M_{p_1}) = \frac{m(x|M_{p_1})}{m(x|M_{p_2})}, \tag{6.30}$$

where $m(x|M_p)$ is the marginal likelihood of the model with p planets, which can be evaluated by integrating over the parameter space

$$p(x|M_p) = \int L(x|\theta_p, M_p) p(\theta_p|M_p) d\theta_p, \tag{6.31}$$

and where θ_p is a vector describing the parameters for the model M_p. Here $p(\theta_p|M_p)$ is the prior density of the parameters and $L(x|\theta_p, M_p)$ is the likelihood function.

Consequentially, exoplanet detection requires calculation of marginal likelihoods for each possible number of planets, which is not trivial for non-zero planet models $M_i, i > 0$. Even for one-planet model M_1, there are three factors that are intertwined to make the computation of $p(x|M_1)$ particularly challenging (Bullard, 2009): (a) multimodality in the likelihood function $L(x|\theta_1, M_1)$; (b) multidimensionality in the M_1 parameter space; (c) high nonlinearity in the single planet model M_1.

Estimating Marginal Likelihoods by Adaptive Importance Sampling

The marginal likelihood computation in Bayesian exoplanet search discussed in the section above can be well solved by the adaptive marginal likelihood computation method proposed in Section 6.1.1. In the sampling and adaptation part, adaptive MCMC with mixture distribution proposal is applied to sample from the posterior $p(\theta|x, D_p) \propto L(x|\theta_p, M_p)p(\theta_p|M_p)$. At each adaption iteration, the posterior samples are used to update the proposal distribution as well as the nonparametric importance function which is a truncated DP mixture of $2 + 5p$ dimensional multivariate normal distribution. After the algorithm runs sufficient long and the changes of nonparametric importance function become negligible, we draw a large set of samples from the tuned importance function and evaluate the marginal likelihood via equation (6.16).

Adaptive MCMC Sampling

We now focus on the adaptive MCMC for posterior sampling in the Bayesian exoplanet search model. We can use a proposal distribution in the form of a mixture

of univariate normal distributions for each of the $2+5p$ dimensions. However, by a simple exploration, we find that the orbital period, variate P, has extremely small variance and has strong correlation to other variants. Supposing that the distribution of P has two well separated modes, which is a quite general assumption, it is straightforward to verify that the two modes of P yield very different set of values for the remaining variants. Thus, if we update the $2+5p$ variants one by one by a Metropolis-within-Gibbs algorithm (no matter how well we can tune the proposal), when P jumps from one mode to another, it will result in extremely low acceptance probabilities since the other $2+5p-1$ variants do not support this new P to fit our model. Therefore it is challenging to explore the space of P, particular when P is multi-modes. On the other hand, we can use a proposal distribution in the form of mixture of $2+5p$ dimensional multivariate normal distributions. In that case, however, a large number of mixture components is required in order to cover the high dimensional space, resulting a heavy computation burden and difficulty in adapting the proposal.

Here we proposed a novel adaptive MCMC framework for posterior sampling in case where variants may be highly correlated. The adaptive proposal for variate P is a mixture of univariate normal distribution, while for other variants, denoted by θ_d, the proposal is a mixture of bivariate normal distributions to model P and θ_d jointly. A new sample is drawn as follows: first draw a new sample from the proposal distribution of P, then given the new value of $P^{(new)}$ sample a new value for each θ_d from the conditional distribution $p(\theta_d|P^{(new)})$, which can be obtained easily by the following scheme: given that we have obtained a mixture for joint distribution of (y,x), means that

$$\begin{pmatrix} y \\ x \end{pmatrix} \sim \sum_{j=1}^{k} \pi_j N \left(\begin{pmatrix} \mu_{jy} \\ \mu_{jx} \end{pmatrix}, \begin{pmatrix} \Sigma_{jx} & R_j' \\ R_j & \Sigma_{jy} \end{pmatrix} \right).$$

Then the conditional distribution $p(y|x)$ has the following closed form expression

$$p(y|x) = \sum_{j=1}^{k} \pi_j^*(x) N(y|\mu_{jy} + R_j \Sigma_{jx}^{-1} (x - \mu_{jx}), \Sigma_{jx} - R_j \Sigma_{jx}^{-1} R_j'), \tag{6.32}$$

where $\pi_j^*(x) = \pi_j p_j(x) / \sum_{r=1}^{k} \pi_r p_r(x)$ is the non-linear weights.

After draw a set of samples from the posterior distribution, these samples are then used to update the mixture distribution proposal for both P and θ_d using the adaptation strategies introduced in Section 6.1.2.

6.4.3 Simulation Studies

We present a example to demonstrate our proposed adaptive marginal likelihood computation method on model assessment on exoplanet searches. The example involves the metal-rich G5IV star HD88133 with radial velocities given in Fischer et al. (2005). Here we focus on model assessment of two potential models, no-planet model M_0 and one-planet model M_1.

As presented in Section 6.3.2, we apply the same method for marginal likelihoods computation for both models. Figure 6.11 shows autocorrelation plots for the sampled parameters in one-planet model M_1, which implies the effectiveness of our adaptive MIS sampler. Given the posterior samples, we can obtain the estimation of parameters in the radial velocity model. A fitted radial velocity function is plotted in Figure 6.12, which shows the model fitting observations well. $50,000$ samples are drawn from the tuned TDP mixture for evaluating of marginal likelihood. The effective sample size $12,810$ implies the efficiency of the estimation. We carry out 50 runs for marginal likelihoods of both models, resulting the following estimations: No-planet model M_0: $1.6815e-37(\pm 2.7012e-39)$; One-planet Model M_1: $2.8732e-35(\pm 8.1360e-37)$. The Bayes factor $BF(M_1 : M_0) \approx 171$ supports for

the hypothesis of one-planet model M_1, which is consistent with the underground true-a planet orbiting HD88133 with a period of 3.41 days (Fischer et al., 2005).

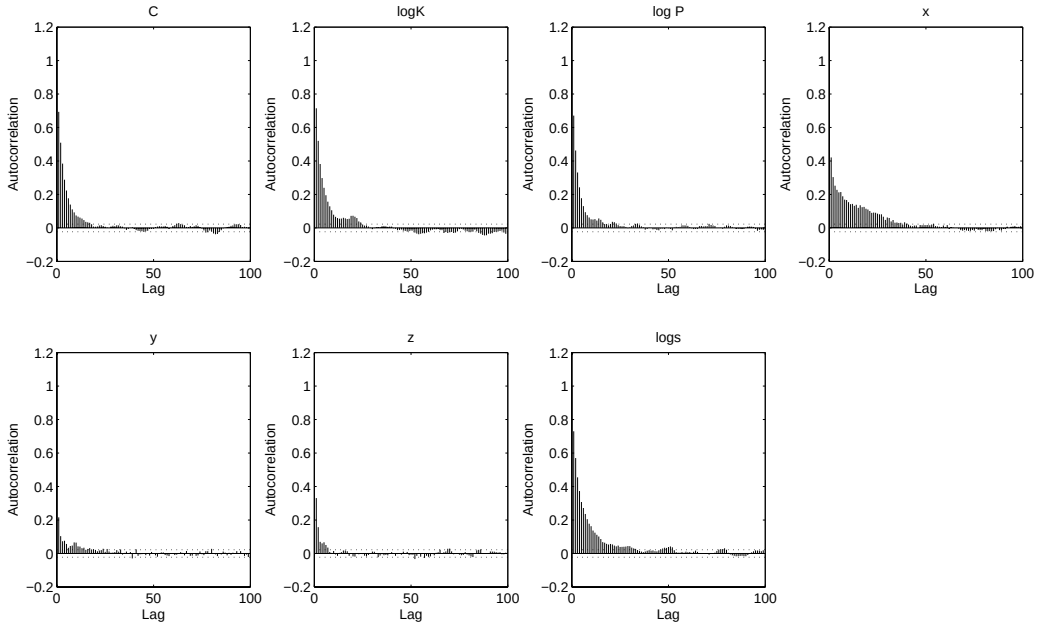


Figure 6.11: Autocorrelation plots for posterior samples of transformed parameters in one-planet model M_1.

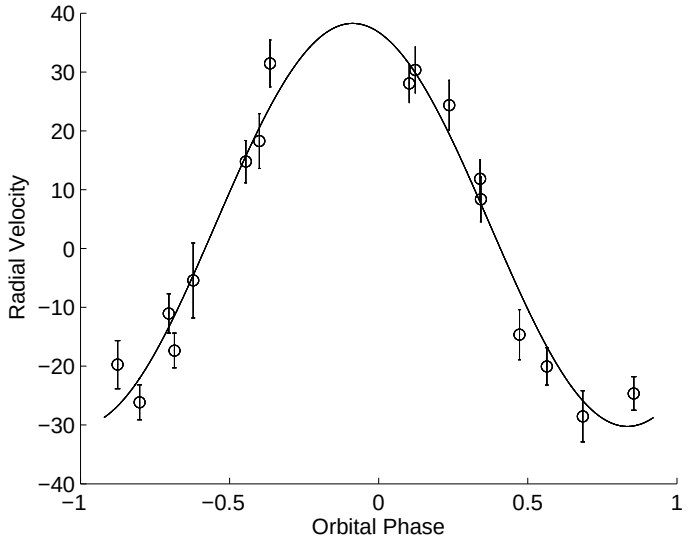


Figure 6.12: Phased radial velocities for HD88133 with an orbital period of 3.41 days: circles and error bars representing the observations of HD88133; the curve representing the fitted velocity-shift model.

Chapter 7

Conclusion and Further Study

7.1 Summary

This dissertation presented the development of Bayesian nonparametric modelling and associated computational methods in analysis of spatial point processes. The first scenario studied in this work is indirectly observed spatial point processes, which involves noisy measurements on an underlying point process that provide indirect and noisy data on locations of point outcomes. The motivation for this research arises from the analysis of fluorescent intensity images in sections of lymphatic tissue where the point processes represent geographical configurations of cells. Analysis of fluorescent intensity images has gained increasing interests in recent years. Our applied studies involve large (though unknown) numbers of point occurrences and intensity mixture models with relatively large numbers of mixture model components representing potentially complex patterns of variation over the spatial region. The use of flexible, nonparametric Bayesian mixture models of intensity functions is central and key in engendering adaptability to wildly heterogeneous intensity patterns coupled with robustness and in-built parsimony. The use of effective MCMC samplers is key, and the blocked sampler for Dirichlet process mixture models is attractive from that viewpoint, but also really necessary as our overlaid measurement error structure demands that we have direct, albeit approximate evaluation of the underlying density-intensity function with the MCMC that generates from conditional posteriors of the underlying latent spatial process. In many spatial point process modelling contexts, lack of complete, direct observation

on point outcomes is common, and our new methodology provides examples of how the overall analysis framework can be extended to allow for that.

In the second scenario, we deal with dynamic spatial point processes, which are motivated by multiple extended target tracking and cell fluorescent microscopic imaging tracking problems. No satisfactory statistical methodology has been particularly available for inference in these types of problems. In this work, we develop the dependent DP mixture model for the time varying setting and provide a novel computational method for Bayesian inference and model-fitting. Our proposed approach can be applied to deal with problems where the measurement of objects is a distribution, which can present multimodality as its features. As a straightforward extension, the dependent DP mixture modelling is further developed to deal with dynamic spatial point processes. Utilizing such nonparametric mixture models for the spatial process intensity functions, we introduce time variation via dynamic models for underlying parameters. The filtering method developed for dependent DP mixture can be easily tuned to solve the sequential Bayesian inference in such scenario. We demonstrated that our proposed dependent DP mixture modelling outperformed naive modelling without introducing dependence between adjacent time frames in rebuilding the underling distribution of interest. In extended target tracking, our proposed approach can give the estimation of not only the target trajectories but also the intensity driving the point processes, and thus facilitate a rigorous Bayesian analysis of such tracking problems. Moreover, the proposed model can be scaled to deal with substantive data analysis in cell fluorescent microscopic imaging tracking, where the number of targets is significant larger than cases studied in traditional multi-target tracking.

The second part of this dissertation discusses computation methods in Bayesian inference. We provides approaches to estimate the upper and lower bounds for log

marginal likelihood in a Bayesian model. While traditional variational approaches only provide lower bound estimation, our lower/upper bounds not only facilitate more reliable model selection but also give a way to show the goodness of the variational density as an approximation to the posterior density of model parameters. Moreover, by marginalizing some parameters in the model, the 'discrepancy' between bounds of log marginal likelihood can be significantly reduced. Extensive simulation studies show the efficiency of our proposed marginal likelihood approximation approach.

Finally, in the last chapter we first presents a generic framework to design adaptive MCMC algorithms, emphasizing the adaptive Metropolized independence sampler and effective adaptation strategy using a family of mixture distribution proposals. To fit a nonparametric model for large data sets, a sequential learning approach for DP mixture model is also proposed, which utilizes only small subsets of the whole data set to update the associated parameters in mixture distribution iteratively. Under the general framework of importance sampling based marginal likelihood computation, the proposed adaptive Monte Carlo method and sequential learning approach can facilitate the marginal likelihood computation and improve its performance. The performance of the proposal method is demonstrated on synthetic examples and a real world application in Bayesian Exoplanet Searches.

7.2 Extensions and Further Study

Nonparametric Spatial Mixture Modelling

In current studies, we only discuss the analysis of 2D fluorescent intensity images. One straightforward extension of the current approach is to deal with 3D fluorescent intensity images or even 4D fluorescent intensity images where time is treated

as a dimension. Using spatial mixture modelling for 4D fluorescence images with high temporal resolution,

$$\begin{pmatrix} y \\ t \end{pmatrix} \sim \sum_{j=1}^{K} \pi_j N \left(\begin{pmatrix} \mu_{jy} \\ \mu_{jt} \end{pmatrix}, \begin{pmatrix} \Sigma_{jt} & R_j' \\ R_j & \Sigma_{jy} \end{pmatrix} \right),$$

then the conditional distribution $p(y|t)$ is in closed-form as follows:

$$p(y|t) = \sum_{j=1}^{K} \pi_j^*(t) N(y|\mu_{jy} + R_j \Sigma_{jt}^{-1} (x - \mu_{jt}), \Sigma_{jt} - R_j \Sigma_{jt}^{-1} R_j'), \tag{7.1}$$

where $\pi_j^*(t) = \pi_j p_j(t) / \sum_{r=1}^{K} \pi_r p_r(t)$ are the non-linear weights. Through such a model, we can therefore infer the time courses of the directional drift of tagged cell types in the fluorescence images, which is of great interest of fluorescence imaging analysis.

This proposed nonparametric spatial mixture modelling suggests a rigorous statistical treatment for time-varying spatial point processes. The proposed modelling and inference approach can be developed and investigated on more realistic examples in multiple extended targets tracking, e.g. observation consisting of signals and clutters, and real-world problems like high resolution radar tracking and image tracking.

Marginal Likelihood Approximation

In the current work, the marginal likelihood approximation by lower and upper bounds method only examined in parametric models. It is of interest to use such methods for Bayesian model comparison of nonparametric models. For further study, we are interested in specification of the criterion of nonparametric models comparison, using our proposed computational tools for nonparametric model comparison and addressing these in realistic data studies.

Adaptive Monte Carlo Methods

Population based Monte Carlo methods, e.g. Sequential Monte Carlo (SMC), have been proposed and widely studied for several decades. While adaptation in single chain MCMC algorithm has been studied in this work, the possibility exists to combine the population based methods with adaptive strategies. For example, running a number of parallel chains which can exchange state information with others, the proposal distribution of each chain can be adaptively tuned in order to explore the space more efficiently. The information exchange mechanism between chains and learning strategy of the adaptive proposal should be carefully chosen in order to guarantee the ergodicity of the chains.

For applications, further study will focus on some non- traditional areas, particularly optimization problems. Adaptive Monte Carlo approach has demonstrated great succuss in dealing with complex system optimization in the emerging area of metamaterials (Liu et al., 2009). In further work, we will extend adaptive Monte Carlo as standard tools for general optimization problems, particularly utilizing its ability to deal with uncertainties in many real world optimization problems.

Appendix A

Gibbs Sampling for Dirichlet Process Mixture Model

Collapsed Gibbs sampling

The collapsed Gibbs sampling for Dirichlet process mixtures involves the following steps:

- For $i \in \{1, ..., N\}$ draw a new value for c_i from the posteriors as defined by,

$$p(c_i = j | x_{1:N}, \mathbf{c}, K, \theta^*, \alpha, G_0) = q_{i,j} \propto \begin{cases} n_{-i,j} \cdot p(x_i|\theta_j^*), & \text{if } j \in \{1, ..., K\} \\ \alpha \cdot h(x_i), & \text{if } j = 0 \end{cases} \tag{A.1}$$

  where $n_{-i,j}$ is the number of occurrences of j in all indicator variables $\mathbf{c}$ except c_i; $p(x_i|\theta_j^*)$ is the likelihood; $h(x_i)$ is a weight obtained via $h(x_i) = \int p(x_i|\theta) dG_0(\theta)$.

- For $k \in \{1, ..., K\}$, independently sample θ_k^* from the relevant component posterior

$$p(\theta_k^* | x_{1:N}, \mathbf{c}) \propto \prod_{i \in \{i: c_i = k\}} \mathcal{N}(x_i|\theta_k^*) G_0(\theta_k^*) \tag{A.2}$$

  .

Blocked Gibbs sampling

The block Gibbs sampling for Dirichlet process mixtures involves the following steps:

- Resample configuration indicators $c_{1:N}$ from $1 : K$ with probabilities

$$Pr(c_i = j) \propto w_j N(x_i|\mu_j^*, \Sigma_j^*), \qquad (j = 1 : K),$$

independently over $i = 1 : N$. This reconfigures the N points independently among the K components, and delivers counts $n_j = \#\{c_i = j,\ i = 1 : N\}$ for $j = 1 : K$. Note that some components may be empty, with $n_j = 0$ for some j.

- For $j \in \{1, ..., K\}$, independently sample new θ_j^* from $p(\theta_j^*|x_{1:N}, c_{1:N})$. The model has $G_0(\mu, \Sigma) = N(\mu|0, t_0\Sigma)IW(\Sigma|s_0, S_0)$ where $t_0 > 0$, $s_0 > 0$ is the prior degree-of-freedom, and $E(\Sigma) = S_0/(s_0 - 2)$ when $s_0 > 2$. This leads to conditional normal-inverse Wishart distributions for each of the k parameters. This straightforward step samples a new set of k parameters, including new draws from $G_0(\cdot)$ for cases with $n_j = 0$.

- For each component $j = 1 : (K - 1)$, compute $\alpha_j = 1 + n_j$ and $\beta_j = \alpha + \sum_{r=j+1}^{K} n_r$ and then sample independent beta variates $v_j \sim Be(\alpha_j, \beta_j)$; set $v_K = 1$. Compute new values of the component probabilities via $\pi_1 = v_1$ and $\pi_j = v_j \prod_{r=1}^{j-1}(1 - v_r)$ for $j = 2 : K$.

Appendix B

MAP sequence estimation

Viterbi Algorithm for MAP sequence estimation (Godsill et al., 2001) can be summarized as follows:

- Initialization. For $1 \leq j \leq J$,

$$\varphi_1(j) = \log p(\boldsymbol{\theta}_1^{(j)}, \mathbf{c}_1^{(j)}) + \log p(\mathbf{x}_1 | \boldsymbol{\theta}_1^{(j)}, \mathbf{c}_1^{(j)}). \tag{B.1}$$

- Recursion. For $2 \leq t \leq T$ and $1 \leq m \leq J$,

$$\varphi_t(m) = \log p(\mathbf{x}_t | \boldsymbol{\theta}_t^{(m)}, \mathbf{c}_t^{(m)}) + \max_j \left[\varphi_{t-1}(j) + \log p(\boldsymbol{\theta}_t^{(m)}, \mathbf{c}_t^{(m)} | \boldsymbol{\theta}_{t-1}^{(j)}, \mathbf{c}_{t-1}^{(j)}) \right]. \tag{B.2}$$

$$\psi_t(m) = \arg\max_j \left[\varphi_{t-1}(j) + \log p(\boldsymbol{\theta}_t^{(m)}, \mathbf{c}_t^{(m)} | \boldsymbol{\theta}_{t-1}^{(j)}, \mathbf{c}_{t-1}^{(j)}) \right]. \tag{B.3}$$

- Termination. $j_T = \arg\max_j \varphi_T(j)$, $\hat{\boldsymbol{\theta}}_T^{MAP}(T) = \boldsymbol{\theta}_T^{(j_T)}$.

- Backtracking. For $t = T-1, T-2, ..., 1$, $j_t = \psi_{t+1}(j_{t+1})$ and $\hat{\boldsymbol{\theta}}_t^{MAP}(t) = \boldsymbol{\theta}_t^{(j_t)}$.

Note that $\boldsymbol{\theta}$ denotes $(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ in the DP mixture model.

Appendix C

Variational Inference in Exponential Families

This appendix reviews the coordinate ascent algorithm discussed in (Blei and Jordan, 2004). Recall that we are considering a model index M, with parameters $\boldsymbol{\theta}$ and observed variables D. The lower bound on the log marginal likelihood is

$$\log p(D|M) \geq E_q[\log(p(D, \boldsymbol{\theta}|M)] - E_q[\log q(\boldsymbol{\theta})]. \tag{C.1}$$

This bound holds for any distribution $q(\boldsymbol{\theta})$. To apply the variational method, we assume a fully-factorized form $q_\nu(\boldsymbol{\theta}) = \prod_{i=1}^N q_{\nu_i}(\theta_i)$ where $\nu = \{\nu_1, ..., \nu_N\}$ are variational parameters and each distribution is in the exponential family (Ghahramani and Beal, 2001). Here we show a coordinate ascent algorithm in which the bound is maximized we iteratively with respect to each ν_i, holding the other variational parameters fixed.

Using the chain rule, the low bound $E_q[\log(p(D, \boldsymbol{\theta}|M)] - E_q[\log q_\nu(\boldsymbol{\theta})]$ depending on ν_i is

$$\ell_i = E_q[\log p(\theta_i|D, \theta_{-i}, M)] - E_q[\log q_{\nu_i}(\theta_i)]. \tag{C.2}$$

Given the variational distribution $q_{\nu_i}(\theta_i)$ is in the exponential family, with form $q_{\nu_i}(\theta_i) = h(\theta_i) \exp\{\nu_i^T \theta_i - a(\nu_i)\}$. To optimize ℓ_i, we take the derivative with respect to ν_i, namely

$$\frac{\partial}{\partial \nu_i} \ell_i = \frac{\partial}{\partial \nu_i} \left(E_q[\log p(\theta_i|D, \theta_{-i}, M)] - E_q[\log h(\theta_i)] \right) - \nu_i^T a''(\nu_i). \tag{C.3}$$

The optimal ν_i satisfies

$$\nu_i = [a''(\nu_i)]^{-1} \left(\frac{\partial}{\partial \nu_i} E_q[\log p(\theta_i|D, \theta_{-i}, M)] - \frac{\partial}{\partial \nu_i} E_q[\log h(\theta_i)] \right). \tag{C.4}$$

The result in Equation (C.4) is general. In many applications of mean field methods, a further simplification can be achieved if the conditional distribution $\log p(\theta_i|D, \theta_{-i}, M)$ is an exponential family distribution,

$$p(\theta_i|D, \theta_{-i}, M) = h(\theta_i) \exp\{g(\theta_{-i}, D, M)^T \theta_i - a(g(\theta_{-i}, D, M))\},$$

where $g(\theta_{-i}, D, M)$ denotes the natural parameter for θ_i . A simplified expression of the first derivative of $E_q[\log p(\theta_i|D, \theta_{-i}, M)]$ can be obtained:

$$\frac{\partial}{\partial \nu_i} E_q[\log p(\theta_i|D, \theta_{-i}, M)] \quad = \quad \frac{\partial}{\partial \nu_i} E_q[\log h(\theta_i)] + \frac{\partial}{\partial \nu_i} E_q[g(\theta_{-i}, D, M)]^T a''(\nu_i).$$

Using the first derivative in Equation (C.4), the maximum is attained at

$$\nu_i = E_q[g(\theta_{-i}, D, M)]. \tag{C.5}$$

A coordinate ascent algorithm based on Equation (C.5) can be defined by iteratively updating ν_i for $i \in \{1, ..., N\}$. Such an algorithm can find a local maximum of Equation (C.1), under the condition that the right-hand side of Equation (C.2) is strictly convex (Bertsekas, 1999).

Appendix D

Proofs of Convergence of MCSA Algorithm

Stochastic Approximation (SA) is a class of algorithms to finding the roots of possibly non-linear equation $f(\boldsymbol{\gamma}) = 0$, in the situation where only noisy measurements of $f(\boldsymbol{\gamma})$ are available. In its simplest form, the Robbins-Monro algorithm is a recursive process as follows,

$$\boldsymbol{\gamma}^{(t+1)} = \boldsymbol{\gamma}^{(t)} + r^{(t+1)}\zeta^{(t+1)} \tag{D.1}$$

where $\left\{r^{(t)}, t \geq 1\right\}$ is a sequence of stepsizes which satisfies standard conditions: $\sum_{t=1}^{\infty} r^{(t)} = \infty$ and $\sum_{t=1}^{\infty} \left[r^{(t)}\right]^2 < \infty$ and for any $t \geq 1$, ζ is a noisy measurement of $f(\boldsymbol{\gamma})$:

$$\zeta^{(t+1)} = f(\boldsymbol{\gamma}) + \xi^{(t+1)} \tag{D.2}$$

where $\left\{\xi^{(t)}, t \geq 1\right\}$ is the so called noise sequence.

In our case, we denote $h(\boldsymbol{\theta};\boldsymbol{\gamma}) = \left[1 + \log \frac{q(\boldsymbol{\theta}^{(i)};\boldsymbol{\gamma})}{p(\boldsymbol{\theta}^{(i)}, D|M)}\right] \frac{d}{d\boldsymbol{\gamma}} \log q(\boldsymbol{\theta}^{(i)};\boldsymbol{\gamma})$. Assume we have Monte Carlo samples $\{\boldsymbol{\theta}^{(i)} : i = 1, \ldots, N\}$ from the distribution $q(\boldsymbol{\theta};\boldsymbol{\gamma})$, then $f(\boldsymbol{\gamma})$ in equation (5.15) can be evaluated by its Monte Carlo estimate,

$$\zeta(\boldsymbol{\gamma}) = -\frac{1}{N}\sum_{i=1}^{N} h(\boldsymbol{\theta}^{(i)};\boldsymbol{\gamma}). \tag{D.3}$$

The Central Limit Theorem,

$$\xi(\boldsymbol{\gamma}) = [\zeta(\boldsymbol{\gamma}) - f(\boldsymbol{\gamma})] \to N\left(0, \frac{\sigma^2}{N}\right), \quad \text{as} \quad n \to \infty \tag{D.4}$$

implies that $\xi(\boldsymbol{\gamma})$ is Gaussian noise, with mean zero and variance $\frac{\sigma^2}{N}$ with $\sigma^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(h(\boldsymbol{\theta}^{(i)};\boldsymbol{\gamma}) - \zeta(\boldsymbol{\gamma})\right)$ (Robert and Casella, 2004).

By using the iterative stochastic approximation method, we can estimate $\boldsymbol{\gamma}$ iteratively via

$$\boldsymbol{\gamma}^{(t+1)} = \boldsymbol{\gamma}^{(t)} + r^{(t+1)}\zeta\left(\boldsymbol{\gamma}^{(t)}\right). \tag{D.5}$$

Here we present a theorem to show that $\boldsymbol{\gamma}^{(t)} \to \boldsymbol{\gamma}_L$ in probability one as $t \to \infty$.

Theorem 1. *Consider the following conditions:*

(A1) *By the central limit theorem,* $\xi(\boldsymbol{\gamma}) \to N\left(0, \frac{\sigma^2}{N}\right)$ *in distribution and*

$$\int_{\Theta} h(\boldsymbol{\theta};\boldsymbol{\gamma})^2 q(\boldsymbol{\theta};\boldsymbol{\gamma}) d\boldsymbol{\theta} < \infty.$$

(A2) Γ *is an open subset of* R^{n_γ}*. The mean field* $f : \Gamma \to R^{n_\gamma}$ *is continuous and there exists a continuously differentiable function* $w : \Gamma \to [0, \infty)$ *(with the convention* $w(\boldsymbol{\gamma}) = \infty$ *when* $\boldsymbol{\gamma} \notin \Gamma$ *) such that:*

1. *For any* $M > 0$*, the level set* $W_M \equiv \{\boldsymbol{\gamma} \in \Gamma, w(\boldsymbol{\gamma}) \leq M\} \subset \Gamma$ *is compact,*
2. *The set of stationary point(s)* $\mathcal{L} \equiv \{\boldsymbol{\gamma} \in \Gamma, \langle \nabla w(\boldsymbol{\gamma}), f(\boldsymbol{\gamma})\rangle = 0\}$ *belongs to the interior of* Γ*,*
3. *For any* $\boldsymbol{\gamma} \in \Gamma$ *,* $\langle \nabla w(\boldsymbol{\gamma}), f(\boldsymbol{\gamma})\rangle \leq 0$ *and the closure of* $w(\mathcal{L})$ *has an empty interior.*

(A3) *The sequence* $\{r^{(t)}, t \geq 1\}$ *is non-increasing, positive and*

$$\sum_{t=1}^{\infty} r^{(t)} = \infty \quad \textit{and} \quad \sum_{t=1}^{\infty} \left[r^{(t)}\right]^2 < \infty. \tag{D.6}$$

Assume (A1-3). Then,

$$P\left[\lim_{t\to\infty} d\left(\boldsymbol{\gamma}^{(t)}, \mathcal{L}\right) = 0\right] = 1. \tag{D.7}$$

Proof: the recursion is expressed as follows

$$\boldsymbol{\gamma}^{(t+1)} = \boldsymbol{\gamma}^{(t)} + r^{(t+1)}\zeta^{(t+1)} = \boldsymbol{\gamma}^{(t)} + r^{(t+1)} f\left(\boldsymbol{\gamma}\right) + r^{(t+1)}\xi^{(t+1)}, \tag{D.8}$$

where $f\left(\boldsymbol{\gamma}\right)$ is the function of interest and $\xi^{(t+1)}$ is a random perturbation

$$\begin{aligned} \xi^{(t+1)} &= f\left(\boldsymbol{\gamma}\right) - \zeta^{(t+1)} && \text{(D.9)} \\ &= \int_{\Theta} h(\boldsymbol{\theta};\boldsymbol{\gamma}) d\boldsymbol{\theta} - \frac{1}{N}\sum_{i=1}^{N} h(\boldsymbol{\theta}^{(i)};\boldsymbol{\gamma}). && \text{(D.10)} \end{aligned}$$

Define $M_n = \sum_{t=1}^{n} r^{(t)}\xi^{(t)}$. Then

$$\begin{aligned} E\left[M_{n+1}|M_k, k \leq n\right] &= E\left[\sum_{t=1}^{n+1} r^{(t)}\xi^{(t)}|M_k, k \leq n\right] && \text{(D.11)} \\ &= E\left[\sum_{t=1}^{n} r^{(t)}\xi^{(t)}|M_k, k \leq n\right] + E\left[r^{(t+1)}\xi^{(t+1)}|M_k, k \leq n\right] && \text{(D.12)} \\ &= M_n + r^{(t+1)} E\left[\xi^{(t+1)}\right]. && \text{(D.13)} \end{aligned}$$

Since $\xi^{(t+1)} = \int_{\Theta} h(\boldsymbol{\theta};\boldsymbol{\gamma}) d\boldsymbol{\theta} - \frac{1}{N}\sum_{i=1}^{N} h(\boldsymbol{\theta}^{(i)};\boldsymbol{\gamma}) \to 0$, almost sure (a.s.), as $N \to \infty$, $E\left[M_{n+1}|M_k, k \leq n\right] \to M_n$ a.s. or with probability one. Therefore $\{M_n, n \geq 1\}$ is a F-martingale.

Then by the martingale inequality,

$$P\left\{\sup_{n \geq j \geq m} |M_j - M_m| \geq \mu\right\} \leq \frac{E\left|\sum_{i=m}^{n-1} r^{(i)}\xi^{(i)}\right|}{\mu} \tag{D.14}$$

which implies

$$\lim_{m\to\infty} P\left\{\sup_{j \geq m} |M_j - M_m| \geq \mu\right\} = 0 \tag{D.15}$$

so that

$$\lim_{m\to\infty}\left(\sup_{m \leq j \leq m(n,T)} \left\|\sum_{i=n}^{j} r^{(i)}\xi^{(i)}\right\|\right) = 0 \quad \text{for all} \quad T > 0 \tag{D.16}$$

where $m(n,T) \equiv \max\left\{k : r^{(n)} + ... + r^{(k)} \leq T\right\}$. This condition is called Kushner and Clark's condition, which is an important sufficient condition on the noise sequence for the convergence of stochastic approximation algorithms. By the theorem in Kushner and Yin (1997), we can obtain

$$P\left[\lim_{t\to\infty} d\left(\boldsymbol{\gamma}^{(t)}, \mathcal{L}\right) = 0\right] = 1. \tag{D.17}$$

In case the noise sequence $\left\{\xi^{(t)}\right\}$ is Markov state dependent noise,

$$P\left\{\xi^{(t+1)} \in \cdot | \xi^{(i+1)}, \boldsymbol{\gamma}^{(i+1)}, i \leq t\right\} = P\left\{\xi^{(t+1)} \in \cdot | \xi^{(t)}, \boldsymbol{\gamma}^{(t)}\right\},$$

we can obtain a similar but more complex proof for the convergence,

$$P\left[\lim_{t\to\infty} d\left(\boldsymbol{\gamma}^{(t)}, \mathcal{L}\right) = 0\right] = 1. \tag{D.18}$$

Using the estimate $\hat{\boldsymbol{\gamma}}_L$ produced through the above iterative procedure and the Monte Carlo samples $\{\boldsymbol{\theta}^{(i)} : i = 1, \ldots, N\}$ from $q(\boldsymbol{\theta}; \hat{\boldsymbol{\gamma}}_L)$, we obtain the estimate of the optimal lower bound conditional on the kernel form of the variational density function, namely

$$\hat{L}(\hat{\boldsymbol{\gamma}}_L) = \frac{1}{N}\sum_{i=1}^{N} \log \frac{p(\boldsymbol{\theta}^{(i)}, D|M)}{q(\boldsymbol{\theta}^{(i)}; \hat{\boldsymbol{\gamma}}_U)}, \tag{D.19}$$

When the iterative steps in the stochastic approximation go to infinity, this estimated lower bound converges to the true maximum lower bound $L(\boldsymbol{\gamma}_L)$ with probability 1.

Appendix E

Derivation of Sequential Learning for DP Mixture Models

The truncated DP mixture proposal distribution is in the form of

$$q(x;\psi) = \sum_{k=1}^{K} w_k q(x|\theta_k), \tag{E.1}$$

where $w_k = V_k \prod_{j=1}^{k-1}(1-V_j)$ and (V_k, θ_k) (also denoted by ψ) are the parameters to be optimized. The KL-divergence between the target distribution $\pi(x)$ and the candidate proposal distribution $q(x;\psi)$ is then

$$\mathcal{D}[\pi(x)||q(x;\psi)] = \int \pi(x) \log \frac{\pi(x)}{q(x;\psi)} dx. \tag{E.2}$$

To find the optimal parameter (ψ^*) that minimizes the KL-divergence $\mathcal{D}[\pi(x)||q(x;\psi)]$ or equivalently maximize $\widehat{\mathcal{D}}(\psi) = E_\pi[\log q(x;\psi)]$, we firstly obtain the first-order partial derivative of $\widehat{\mathcal{D}}(\psi)$ with respect to each V_k (for $k = 1, ..., K-1$) as follows:

$$\begin{aligned}
h_{V_k}(x;\psi) &= \frac{\partial}{\partial V_k}\left[\int \pi(x) \log q(x;\psi) dx\right] && \text{(E.3)}\\
&= \int \pi(x) \frac{\partial}{\partial V_k}\left[\log \sum_{m=1}^{K} w_m q(x|\theta_m)\right] dx && \text{(E.4)}\\
&= \int \pi(x) \frac{-\sum_{l=k+1}^{K} V_l \prod_{j\le l-1, j\neq k}(1-V_j) q(x|\theta_l) + \prod_{j=1}^{k-1}(1-V_j) q(x|\theta_k)}{\sum_{m=1}^{K} w_m q(x|\theta_m)} dx. && \text{(E.5)}
\end{aligned}$$

Note that due to the truncation of the DP mixture, V_K is always set equal to 1.

As we have stated, this partial derivative involves an intractable integration w.r.t. complex $\pi(x)$. Instead, we can evaluate estimate based on the sample $X_t = \{x_t^{(i)}\}_{i=1}^{N_t}$ from $\pi(x)$, that is,

$$H_{V_k}(X_t;\psi) = \frac{1}{N_t}\sum_{i=1}^{N_t} \frac{-\sum_{l=k+1}^{K} V_l \prod\limits_{j\le l-1, j\neq k}(1-V_j)q(x_t^{(i)}|\theta_l) + \prod\limits_{j=1}^{k-1}(1-V_j)q(x_t^{(i)}|\theta_k)}{\sum\limits_{m=1}^{K} w_m q(x_t^{(i)}|\theta_m)}. \tag{E.6}$$

Therefore, by applying the stochastic approximation technique, we can get the recursive update equation for V_k as follows:

$$V_{k,t+1} = V_{k,t} + r_{k,n+1} H_{V_k}(X_t;\psi_t), \tag{E.7}$$

where

$$H_{V_k}(X_t;\psi_t) = \frac{1}{N_t}\sum_{i=1}^{N_t} \frac{-\sum_{l=k+1}^{K} V_l \prod\limits_{j\le l-1, j\neq k}(1-V_j)q(x_t^{(i)}|\theta_{l,t}) + \prod\limits_{j=1}^{k-1}(1-V_j)q(x_t^{(i)}|\theta_{k,t})}{\sum\limits_{m=1}^{K} w_{m,t} q(x_t^{(i)}|\theta_{m,t})}, \tag{E.8}$$

and $r_{k,n+1}$ is the step-size in the stochastic approximation algorithm. Give the updated $V_{k,t+1}$ (for $k = 1, ..., K$), we can evaluate the $w_{k,t+1}$ trivially by $w_{k,t+1} = V_{k,t+1}\prod\limits_{j=1}^{k-1}(1-V_{j,t+1})$.

In order to estimate the parameter θ in each of the mixture component, the first-order partial derivative of $\widehat{\mathcal{D}}(\psi)$ with respect to θ_k (for $k = 1, ..., K$) is also needed.

We see that

$$h_{\theta_k}(x;\psi) = \frac{\partial}{\partial\theta_k}\left[\int \pi(x)\log q(x;\psi)dx\right] \tag{E.9}$$

$$= \int \pi(x)\frac{\partial}{\partial\theta_k}\left[\log\sum_{m=1}^{K} w_m q(x|\theta_m)\right]dx \tag{E.10}$$

$$= \int \pi(x)\frac{w_k\frac{\partial}{\partial\theta_k}q(x;\theta_k)}{\sum_{m=1}^{K} w_m q(x;\theta_m)}dx \tag{E.11}$$

$$= \int \pi(x)\frac{w_k q(x;\theta_k)\frac{1}{q(x;\theta_k)}\frac{\partial}{\partial\theta_k}q(x;\theta_k)}{\sum_{m=1}^{K} w_m q(x;\theta_m)} \tag{E.12}$$

$$= \frac{w_k q(x;\theta_k)}{\sum_{m=1}^{K} w_m q(x;\theta_m)}\frac{\partial}{\partial\theta_k}\log q(x;\theta_k). \tag{E.13}$$

Estimation based on sample $X_t = \{x_t^{(i)}\}_{i=1}^{N_t}$ is of form

$$H_{\theta_k}(x;\psi) = \frac{1}{N_t}\sum_{i=1}^{N_t}\frac{w_k q(x_t^{(i)};\theta_k)}{\sum_{m=1}^{K} w_m q(x_t^{(i)};\theta_m)}\frac{\partial}{\partial\theta_k}\log q(x_t^{(i)};\theta_k).$$

In case that the mixture component $q(x;\theta_k)$ is a Gaussian distribution $\mathcal{N}(x;\mu_k,\Sigma_k)$, where $\theta_k = (\mu_k,\Sigma_k)$, $H_{\theta_k}(x;\psi)$ becomes

$$H_{\mu_k}(x;\psi) = \frac{1}{N_t}\sum_{i=1}^{N_t}\frac{w_k q(x_t^{(i)};\theta_k)}{\sum_{m=1}^{K} w_m q(x_t^{(i)};\theta_m)}\frac{\partial}{\partial\mu_k}\log\mathcal{N}(x_t^{(i)};\mu_k,\Sigma_k) \tag{E.14}$$

$$\propto \frac{1}{N_t}\sum_{i=1}^{N_t}\frac{w_k q(x_t^{(i)};\theta_k)}{\sum_{m=1}^{K} w_m q(x_t^{(i)};\theta_m)}\left(x_t^{(i)}-\mu_k\right), \tag{E.15}$$

$$H_{\Sigma_k}(x;\psi) = \frac{1}{N_t}\sum_{i=1}^{N_t}\frac{w_k q(x_t^{(i)};\theta_k)}{\sum_{m=1}^{K} w_m q(x_t^{(i)};\theta_m)}\frac{\partial}{\partial\Sigma_k}\log\mathcal{N}(x_t^{(i)};\mu_k,\Sigma_k) \tag{E.16}$$

$$\propto \frac{1}{N_t}\sum_{i=1}^{N_t}\frac{w_k q(x_t^{(i)};\theta_k)}{\sum_{m=1}^{K} w_m q(x_t^{(i)};\theta_m)}\left((x_t^{(i)}-\mu_k)(x_t^{(i)}-\mu_k)^T-\Sigma_k\right) \tag{E.17}$$

Hence, we derive the following recursive update equations for $\theta_k = (\mu_k, \Sigma_k)$

$$\mu_{k,t+1} = \mu_{k,t} + r'_{k,t+1} \frac{1}{N_t} \sum_{i=1}^{N_t} \alpha_{k,t+1}^{(i)} \left(x_t^{(i)} - \mu_{k,t} \right) \tag{E.18}$$

$$\Sigma_{k,t+1} = \Sigma_{k,t} + r'_{k,t+1} \frac{1}{N_t} \sum_{i=1}^{N_t} \alpha_{k,t+1}^{(i)} \left((x_t^{(i)} - \mu_{k,t})(x_t^{(i)} - \mu_{k,t})^T - \Sigma_{k,t} \right) \tag{E.19}$$

where

$$\alpha_{k,t+1}^{(i)} = \frac{w_{k,t} q(x_t^{(i)}; \theta_{k,t})}{\sum_{m=1}^{K} w_{m,t} q(x_t^{(i)}; \theta_{m,t})}. \tag{E.20}$$

$r'_{k,t+1}$ is also the step-size in the stochastic approximation algorithm.

Bibliography

Andrieu, C. and Moulines, E. (2006), "On the ergodicity properties of some adaptive MCMC algorithms," *Annals of Applied Probability*, 16, 1462–1505.

Andrieu, C., Moulines, E., and Priouret, P. (2005), "Stability of stochastic approximation under verifiable conditions," *SIAM Journal on Control and Optimization*, 44, 283–312.

Bar-Shalom, Y. and Fortmann, T. (1988), *Tracking and Data Association*, Academic Press Professional, Inc. San Diego, CA, USA.

Beal, M. (2003), "Variational algorithms for approximate Bayesian inference," Ph.D. thesis, Gatsby Computational Neuroscience Unit, University College London.

Beal, M. and Ghahramani, Z. (2003), "The variational Bayesian EM algorithm for incomplete data: with application to scoring graphical model structures," in *Bayesian Statistics 7*, eds. J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith, and M. West, pp. 453–464, Oxford University Press.

Bertsekas, D. (1999), *Nonlinear Programming*, Athena Scientific, Belmont, MA.

Blei, D. and Jordan, M. (2004), "Variational inference for Dirichlet process mixtures," *Bayesian Analysis*, 1, 121–144.

Brockwell, A. E. and Kadane, J. B. (2005), "Identification of regeneration times in MCMC, with application to adaptive schemes," *Journal of Computational and Graphical Statistics*, 14, 436–458.

Bullard, F. (2009), "Exoplanet Detection: a Comparison of Three Statistics or How Long Should It Take to Find a Small Planet?" Ph.D. thesis, Duke University.

Cappé, O., Douc, R., Guillin, A., Marin, J., and Robert, C. (2008), "Adaptive importance sampling in general mixture classes," *Statistics and Computing*, 18, 447–459.

Caron, F., Davy, M., and Doucet, A. (2007), "Generalized polya urn for time-varying Dirichlet process mixtures," in *Proceedings of the International Conference on Uncertainty in Artificial Intelligence(UAI)*, Vancouver, Canada.

Carvalho, C. and West, M. (2007), "Dynamic Matrix-Variate Graphical Models," *Bayesian Analysis*, 2, 69–98.

Celeux, G. and Diebolt, J. (1992), "A stochastic approximation type EM algorithm for the mixture problem," *Stochastics and Stochastics Reports*, 41, 127–146.

Chan, C., Feng, F., Ottinger, J., Foster, D., West, M., and Kepler, T. (2008), "Statistical mixture modeling for cell subtype identification in flow cytometry," *Cytometry-Part A*, 73, 693–701.

Chib, S. (1995), "Marginal Likelihood from the Gibbs Output," *Journal of the American Statistical Association*, 90, 1313–1321.

Corduneanu, A. and Bishop, C. (2001), "Variational Bayesian model selection for mixture distributions," in *Proceedings Eighth International Conference on Artificial Intelligence and Statistics*, eds. T. Richardson and T. Jaakkola, pp. 27–34, Morgan Kaufmann.

Crooks, J. L., Berger, J. O., and Loredo, T. J. (2007), "Posterior-guided importance sampling for calculating marginal likelihoods with application to bayesian exo-

planet searches," Duke University, Department of Statistical Science, Discussion Paper 2007-26.

Daley, D. and Vere-Jones, D. (2003), *An Introduction to the Theory of Point Processes (2nd edn.)*, New York: Springer Verlag.

DeIorio, M., Mueller, P., Rosner, G. L., and MacEachern, S. (2004), "An ANOVA model for dependent random measures," *Journal of the America Statistical Association*, 99, 205–215.

Delyon, B., Lavielle, M., and Moulines, E. (1999), "Convergence of a stochastic approximation version of the EM algorithm," *The Annals of Statistics*, 27, 94–128.

Diggle, P. (2003), *Statistical Analysis of Spatial Point Patterns*, London: Arnold.

Dunson, D. B. and Park, J.-H. (2008), "Kernel stick-breaking processes," *Biometrika*, 95, 307–323.

Dunson, D. B., Pillai, N., and Park., J.-H. (2007), "Bayesian density regression," *Journal of the Royal Statistical Society: Series B*, 69, 163 – 183.

Elowitz, M. B., Levine, A. J., Siggia, E. D., and Swain, P. S. (2002), "Stochastic gene expression on a single cell," *Science*, 297, 1183–1186.

Erland, S. (2003), "On Eigen-Decompositions and Adaptivity of Markov Chains," Ph.D. thesis, Norwegian University of Science and Technology.

Escobar, M. and West, M. (1995), "Bayesian density estimation and inference using mixtures," *Journal of the American Statistical Association*, 90, 577–588.

Escobar, M. and West, M. (1998), "Computing nonparametric hierarchical models," in *Practical Nonparametric and Semiparametric Bayesian Statistics*, eds. D. Dey, P. Mueller, and D. Sinha, pp. 1–22, New York: Springer Verlag.

Fearnhead, P. (2004), "Particle filters for mixture models with an unknown number of components," *Statistics and Computing*, 14, 11–21.

Fearnhead, P. and Meligkotsidou, L. (2007), "Filtering methods for mixture models," *Journal of Computational and Graphical Statistics*, 16, 586–607.

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

Fischer, D., Laughlin, G., Butler, P., Marcy, G., Johnson, J., Henry, G., Valenti, J., Vogt, S., Ammons, M., Robinson, S., et al. (2005), "A hot Saturn planet orbiting HD 88133, from the N2K consortium," *the Astrophysical Journal*, 620, 481–486.

Ford, E. and Gregory, P. (2006), "Bayesian model selection and extrasolar planet detection," *Arxiv preprint astro-ph/0608328*.

Gelfand, A. and Dey, D. (1994), "Bayesian model choice: asymptotics and exact calculations," *Journal of the Royal Statistical Society: Series B*, 56, 501–514.

Gelfand, A. E. and Sahu, S. K. (1994), "On Markov chain Monte Carlo acceleration." *Journal of Computational and Graphical Statistics*, 3, 261–276.

Gelfand, A. E., Kottas, A., and MacEachern, S. N. (2005), "Bayesian nonparametric spatial modeling with Dirichlet process mixing," *Journal of the American Statistical Association*, 100, 1021–1035.

Gelman, A. G., Roberts, G. O., and Gilks, W. R. (1996), "Efficient Metropolis jumping rules," in *Bayesian Statistics V*, eds. A. D. J.M. Bernardo, J.O. Berger and A. Smith, pp. 599–608, Oxford University Press, Oxford.

Ghahramani, Z. and Beal, M. (2001), "Propagation algorithms for variational Bayesian learning," in *Advances in Neural Information Processing Systems*, eds. T. Leen, T. Dietterich, and V. Tresp, vol. 13, pp. 507–513, MIT Press.

Gilholm, K., Godsill, S., Maskell, S., and Salmond, D. (2005), "Poisson models for extended target and group tracking," in *Proceedings of SPIE: Signal and Data Processing of Small Targets*, vol. 5913, pp. 230–241.

Gilks, W. R., Richardson, S., and Spiegelhalter, D. J. (1996), *Markov Chain Monte Carlo in Practice*, Chapman and Hall, London.

Gilks, W. R., Roberts, G. O., and Sahu, S. K. (1998), "Adaptive Markov chain Monte Carlo through regeneration," *Journal of the American Statistical Association*, 93, 1045–1054.

Godsill, S., Doucet, A., and West, M. (2001), "Maximum *a posteriori* sequence estimation using Monte Carlo particle filters," *Ann. Inst. Stat. Math.*, 53, 82–96.

Golding, I., Paulsson, J., Zawilski, S., and Cox, E. (2005), "Real-time kinetics of gene activity in individual bacteria," *Cell*, 123, 1025–1036.

Gordon, A., Colman-Lerner, A., Chin, T. E., Benjamin, K. R., Yu, R. C., and Brent, R. (2007), "Single-cell quantification of molecules and rates using open-source microscope-based cytometry," *Nature Methods*, 4, 175–181.

Grifin, J. E. and Steel, M. F. J. (2006), "Order-based dependent Dirichlet processes," *Journal of the American Statistical Association*, 101, 179–194.

Haario, H., Saksman, E., and Tamminen, J. (2001), "An adaptive Metropolis algorithm," *Bernoulli*, 7, 223–242.

Hastings, W. K. (1970), "Monte Carlo sampling methods using Markov chains and their applications," *Biometrika*, 57, 97–109.

Humphreys, K. and Titterington, D. (2000), "Approximate Bayesian inference for simple mixtures," in *Proceedings in Computational Statistics*, COMPSTAT'2000, Springer-Verlag.

Ishwaran, H. and James, L. (2001), "Gibbs sampling methods for stick-breaking priors," *Journal of the American Statistical Association*, 96, 161–173.

Jaakkola, T. and Jordan, M. (2000), "Bayesian parameter estimation via variational methods," *Statistics and Computing*, 10, 25–37.

Ji, C. (2006), "Adaptive Monte Carlo Methods for Bayesian Inference," Master's thesis, University of Cambridge, UK.

Ji, C. and Schmidler, S. C. (2008), "Adaptive Markov chain Monte Carlo for Bayesian variable selection," *Discussion Papaer*.

Jordan, M. (2004), "Graphical models," *Statistical Science*, 19, 140–15.

Jordan, M., Ghahramani, Z., Jaakkola, T., and Saul, K. (1999), "An introduction to variational methods for graphical models," *Machine Learning*, 37, 183–233.

Kask, P., Palo, K., Ullmann, D., and Gall, K. (1999), "Fluorescence-intensity distribution analysis and its application in biomolecular detection technology," in *Proceedings of the National Academy of Sciences, USA*, vol. 96, pp. 13756–13761.

Kottas, A. and Sanso, B. (2007), "Bayesian mixture modeling for spatial Poisson process intensities, with applications to extreme value analysis," *Journal of Statistical Planning and Inference (Special Issue on Bayesian Inference for Stochastic Processes)*, 137, 3151–3163.

Kushner, H. J. and Yin, G. G. (1997), *Stochastic Approximation Algorithms and Applications*, Springer-Verlag, New York.

Lefebvre, G., Steele, R., Vandal, A., Narayanan, S., and Arnold, D. (2009), "Path sampling to compute integrated likelihoods: an adaptive approach," *Journal of Computational and Graphical Statistics*, 18, 415–437.

Levine, M. and Davidson, E. (2005), "Gene regulatory networks for development," in *Proceedings of the National Academy of Sciences, USA, 102*, pp. 4936–4942.

Liu, J. (1996), "Nonparametric hierarchical Bayes via sequential imputations," *The Annals of Statistics*, 24, 911–930.

Liu, J. and West, M. (2001), "Combined parameter and state estimation in simulation-based filtering," in *Sequential Monte Carlo Methods in Practice*, eds. A. Doucet, N. de Freitas, and N. Gordon, pp. 197–224, Springer-Verlag, New York.

Liu, R., Ji, C., Mock, J. J., Chin, J. Y., Cui, T. J., and Smith, D. R. (2009), "Broadband ground-plane cloak," *Science*, 323, 366–369.

Longo, D. and Hasty, J. (2006), "Dynamics of single-cell gene expression," *Molecular Systems Biology*, 2, 64.

MacEachern, S. (1994), "Estimating normal means with a conjugate style Dirichlet process prior," *Communications in Statistics: Simulation and Computation*, 23, 727–741.

MacEachern, S. (1999), "Dependent nonparametric processes," in *ASA Proceedings of the Section on Bayesian Statistical Science*, pp. 50–55.

MacEachern, S. and Mueller, P. (1998), "Estimating mixture of Dirichlet process models," *Journal of Computational and Graphical Statistics*, 7, 223–238.

MacEachern, S., Clyde, M., and Liu, J. (1999), "Sequential importance sampling for nonparametric Bayes models: The next generation," *The Canadian Journal of Statistics*, 27, 251–267.

MacEachern, S. N. (1998), "Computational methods for mixture of Dirichlet process models," in *Practical Nonparametric and Semiparametric Bayesian Statistics*, eds. D. Dey, P. Mueller, and D. Sinha, pp. 23–43, New York: Springer Verlag.

MacEachern, S. N. (2001), "Decision theoretic aspects of dependent nonparametric processes," in *Bayesian Methods with Applications to Science, Policy and Official Statistics*, ed. E. George, pp. 551–560, Creta: International Society for Bayesian Analysis.

MacKay, D. (1995), "Developments in probabilistic modelling with neural networks ensemble learning," in *Neural Networks: Artificial Intelligence and Industrial Applications. Proceedings of the 3rd Annual Symposium on Neural Networks*, pp. 191–198, Nijmegen, Netherlands.

Manolopoulou, I., Wang, X., Ji, C., Lynch, H. E., Stewart, S., Sempowski, G. D., Alam, S. M., West, M., and Kepler, T. B. (2009), "Statistical analysis of cellular aggregates in immunofluorescence histology," *BMC bioinformatics (submitted)*.

Megason, S. G. and Fraser, S. (2007), "Imaging in systems biology," *Cell*, 130, 784–795.

Meng, X. and Wong, W. (1996), "Simulating ratios of normalizing constants via a simple identity: A theoretical exploration," *Statistica Sinica*, 6, 831–860.

Mira, A. (2001), "Ordering and improving the performance of Monte Carlo Markov chains," *Statistical Science*, 16, 340–350.

Moller, J. and Waagepetersen, R. (2004), *Statistical Inference and Simulation for Spatial Point Processes*, London: Chapman and Hall.

Mueller, P. and Quintana, F. (2004), "Nonparametric Bayesian data analysis," *Statistical Science*, 19, 95–110.

Muller, P., Quintana, F., and Rosner., G. (2004), "A method for combining inference across related nonparametric Bayesian models." *Journal of the Royal Statistical Society: Series B*, 66, 735–749.

Newton, M. A. and Raftery, A. E. (1994), "Approximate Bayesian inference with the weighted likelihood bootstrap," *Journal of the Royal Statistical Society: Series B*, 56, 3–48.

Pennell, M. L. and Dunson, D. B. (2006), "Bayesian semiparametric dynamic frailty models for multiple event time data," *Biometrics*, 62, 1044–52.

Quintana, J. M. and West, M. (1987), "An analysis of international exchange rates using multivariate DLMs," *The Statistician*, 36, 275–281.

Raser, J. M. and O' Shea, E. K. (2004), "Control of stochasticity in eukaryotic gene expression," *Science*, 304, 1811–1814.

Robbins, H. and Monro, S. (1951), "A stochastic approximation method," *Annals of Mathematical Statstics*, 22, 400–407.

Robert, C. P. and Casella, G. (2004), *Monte Carlo Statistical Methods*, New York: Springer-Verlag, 2nd edn.

Roberts, G. O. and Rosenthal, J. S. (2001), "Optimal scaling for various Metropolis-Hastings algorithms," *Statistical Science*, 16, 351–367.

Roberts, G. O. and Rosenthal, J. S. (2007), "Coupling and ergodicity of adaptive Markov chain Monte Carlo algorithms," *Journal of Applied Probability*, 44, 458–475.

Rodriguez, A. and ter Horst, E. (2008), "Bayesian dynamic density estimation," *Bayesian Analysis*, 3, 339–366.

Rosenfeld, N., Young, J. W., A., U., Swain, P. S., and Elowitz, M. (2005), "Gene regulation at the single-cell level," *Science*, 307, 1962–1965.

Rosenfeld, N., Perkins, T., Alon, U., Elowitz, M., and Swain, P. (2006), "A fluctuation method to quantify In vovo fluorescence data," *Biophysical Journal*, 91, 759–766.

Sethuraman, J. (1994), "A constructive definition of Dirichlet priors," *Statistica Sinica*, 4, 639–650.

Sigal, A., Milo, R., Cohen, A., Geva-Zatorsky, N., Klein, Y., Alaluf, I., Swerdlin, N., Perzov, N., Danon, T., Liron, Y., Raveh, T., Carpenter, A. E., Lahav, G., and Alon, U. (2006), "Dynamic proteomics in individual human cells uncovers widespread cell-cycle dependence of nuclear proteins," *Nature Methods*, 3, 525–531.

Singh, S. S., Vo, B. N., Baddeley, A., and Zuyev, S. (2009), "Filters for spatial point processes," *SIAM Journal on Control and Optimization*, 48, 2275–2295.

Srebro, N. and Roweis, S. (2005), "Time-varying topic models using dependent Dirichlet processes," Tech. rep., Department of Computer Science, University of Toronto.

Tierney, L. (1994), "Markov chains for exploring posterior distributions," *Annals of Statistics*, 22, 1701–1728.

Ueda, N. and Ghahramani, Z. (2002), "Bayesian model search for mixture models based on optimizing variational bounds," *Neural Networks*, 15, 1223–1241.

Viterbi, A. J. (1967), "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm," *IEEE Trans. Info. Theory*, IT-13, 260–269.

Walker, S. and Muliere, P. (2003), "A bivariate Dirichlet process," *Statistics and Probability Letters*, 64, 1–7.

Wang, B. and Titterington, D. (2004), "Convergence and asymptotic normality of variational Bayesian approximations for exponential family models with missing values," in *Proceedings of the 20th Conference in Uncertainty in Artificial Intelligence*, eds. M. Chickering and J. Halpern, pp. 577–584.

Wang, Q., You, L., and West, M. (2008), "CellTracer: Software for automated image segmentation and lineage mapping for single-cell studies," Discussion Paper 08-22, Department of Statistical Science, Duke University.

Wang, Q., Niemi, J., Tan, C., You, L., and West., M. (2009), "Image segmenation and dynamic lineage analysis in single-cell flourescence microscopy," *Cytometry A (to appear)*.

Webb, S. (2002), "Where is everybody," *Fifty Solutions to the Fermis Paradox, Copernicus, New York*.

West, M. and Harrison, P. (1997), *Bayesian Forecasting and Dynamic Models*, Springer-Verlag, New York, 2nd edn.

West, M., Mueller, P., and Escobar, M. (1994), "Hierarchical priors and mixture models, with application in regression and density estimation," in *Aspects of Uncertainty: A Tribute to D.V. Lindley*, eds. P. Freeman and A. Smith, pp. 363–386, London: Wiley.

Wolpert, R. and Ickstadt, K. (1998), "Poisson/Gamma random field models for spatial statistics," *Biometrika*, 85, 251–267.

Xing, E., Jordan, M., and Russell, S. (2003), "A generalized mean field algorithm for variational inference in exponential families," in *Proceedings of the 19th Annual Conference on Uncertainty in AI*, eds. C. Meek and U. Kjerulff, pp. 583–591, Morgan Kaufmann Publishers.

Zhu, X., Ghahramani, Z., and Lafferty, J. (2005), "Time-sensitive Dirichlet process mixture models," Tech. rep., Carnegie Mellon University, Pittsburgh, PA.

Biography

Chunlin Ji was born in Hebei province, China on March 13, 1981. He received the B.Eng. Degree from the Northeastern University, Shenyang, China, in 2003, MPhil Degree in Engineering from the Department of Engineering, the University of Cambridge, United Kingdom, in 2006, M.Sc. Degree in Statistical Science from the Department of Statistical Science, Duke University, United States, in 2008. He was a junior research assistant in the Department of Electronic Engineering, Chinese University of Hong Kong, Hong Kong, in 2004, and a research assistant in the Department of Electronic Engineering, City University of Hong Kong, Hong Kong, in 2006. From August 2006, he started to pursue the Ph.D. Degree in Statistics, in the Department of Statistical Science, Duke University, United States.