

# Nonparametric and Variable-Dimension Bayesian Mixture Models: Analysis, Comparison, and New Methods

by

Jeffrey W. Miller

B.Sc., Georgia Institute of Technology; Atlanta, GA, 2001

M.Sc., Stanford University; Stanford, CA, 2002

M.Sc., Brown University; Providence, RI, 2010

A dissertation submitted in partial fulfillment of the  
requirements for the degree of Doctor of Philosophy  
in The Division of Applied Mathematics at Brown University

PROVIDENCE, RHODE ISLAND

May 2014

© Copyright 2014 by Jeffrey W. Miller

This dissertation by Jeffrey W. Miller is accepted in its present form  
by The Division of Applied Mathematics as satisfying the  
dissertation requirement for the degree of Doctor of Philosophy.

Date \_\_\_\_\_

Matthew T. Harrison, Ph.D., Advisor

Recommended to the Graduate Council

Date \_\_\_\_\_

Stuart Geman, Ph.D., Reader

Date \_\_\_\_\_

Steven N. MacEachern, Ph.D., Reader

Approved by the Graduate Council

Date \_\_\_\_\_

Peter M. Weber, Dean of the Graduate School

## Vita

Jeffrey W. Miller was born on June 20, 1979, in Wilmington, Delaware, and grew up in West Chester, Pennsylvania. He received a Bachelor of Science in Mechanical Engineering from the Georgia Institute of Technology in 2001, a Master of Science in Mechanical Engineering from Stanford University in 2002, and a Master of Science in Mathematics from Brown University in 2010. He served as an officer in the U.S. Air Force during 2002–2005, at the Air Force Research Lab at Tyndall Air Force Base, Florida. During 2005–2008, he worked at Draper Laboratory in Cambridge, Massachusetts.

As a PhD candidate in the Division of Applied Mathematics at Brown University, Jeffrey has been advised and mentored by Matthew T. Harrison and Stuart Geman. In addition to several teaching assistantships, he has been the primary instructor for two courses, Introduction to Machine Learning (summer 2011), and Information Theory (fall 2011), and in 2012 he received the Presidential Award for Excellence in Teaching, awarded to one Brown University graduate student each year. He has also received awards for his research, including the Sigma Xi Outstanding Graduate Student Award in 2013, 1st Prize for Poster by a Young Participant at the 9th Conference on Bayesian Nonparametrics in 2013, and the IBM Thomas J. Watson Research Center Student Research Award at the New England Statistics Symposium in 2011.

## Peer-reviewed publications

J. W. Miller and M. T. Harrison, *Inconsistency of Pitman–Yor process mixtures for the number of components*, 2014+ (Accepted pending minor revisions), arXiv:1309.0024.

J. W. Miller and M. T. Harrison, *A simple example of Dirichlet process mixture inconsistency for the number of components*, Advances in Neural Information Processing Systems (NIPS), Vol. 26, 2013.

J. W. Miller and M. T. Harrison, *Exact sampling and counting for fixed-margin matrices*, The Annals of Statistics, Vol. 41, No. 3, 2013, pp. 1569-1592.

J. W. Miller, *Reduced criteria for degree sequences*, Discrete Mathematics, Vol. 313, Issue 4, 2013, pp. 550-562.

## Preface and Acknowledgments

In the fall of 2011, I decided to try my hand at a problem that Stu Geman had mentioned several times — in his words, a “stone in his shoe” that had been bothering him for some time. He had noticed that Dirichlet process mixtures (DPMs) were often being used to infer the number of components in a finite mixture, a practice which seemed highly questionable, due to the nature of the prior. Still, it was not obvious how the posterior would behave — would it, nonetheless, be consistent for the number of components?

After finding some special cases in which inconsistency did provably occur, I worked for several months with Matt Harrison, trying to determine the limiting behavior of the posterior. Eventually, we found that, although precisely determining the limiting behavior eluded us, we could still prove inconsistency under very general conditions.

Simultaneously, we started exploring the natural alternative of putting a prior on the number of components in a finite mixture, i.e., using a variable-dimension mixture. Naively unaware of the previous work with this model, but being familiar with the DPM, the natural thing to do was to develop an inference algorithm resembling the DPM algorithms. This turned out to be a novel development, quite different from previous inference algorithms for the model, such as reversible jump.

In developing this DPM-style inference algorithm, it became clear that a variable-

dimension mixture could possess many of the same properties as a nonparametric mixture, in particular, representation in terms of a simple restaurant process. This led naturally to the idea of pursuing such a representation for variable-dimension alternatives to other popular nonparametric models, and to my surprise, alternatives to the hierarchical Dirichlet process and the Indian buffet process worked out beautifully in much the same way.

There are many people, without whom, this work would hardly have been possible. I would like to thank my parents for their continual support in all my endeavors, and their interest in understanding and keeping up with my work. My advisor, Matt Harrison, has been as close to ideal as I could imagine — always available and responsive, pushing me to get to the bottom of things and comprehensively address each question, keenly understanding tricky problems and creatively suggesting solutions, and wisely helping to choose projects worth pursuing. It has been an honor and a joy to work with Stu Geman, who, in addition to providing the spark which developed into this thesis, has been an encouraging mentor in many ways; his boundless curiosity is an inspiration. Special thanks also to Steve MacEachern for many insightful comments and suggestions which turned out to be most helpful, even though I often didn't realize it until later!

Thanks to my “lab”-mates, Dan Klein and Dahlia Nadkarni, for many enjoyable group meetings, sharing ideas and working through problems as they arise. Thanks to the regulars of the machine learning reading group, Mike Hughes, Jason Pacheco, Dae Il Kim, Thomas Wiecki, Imri Sofer, and Mark Homer, for many discussions about Bayesian nonparametrics and machine learning. To my best friends, Lauren and Arthur Sugden, thanks for countless evenings of conversations about math, science, and life in general.

Thanks to Tamara Broderick, Dave Duvenaud, and Konstantina Palla for suggesting that a variable-dimension alternative to the IBP might also be interesting to pursue.

The nonparametric Bayesian community has been most welcoming and enjoyable to interact with, and numerous discussions with its members have made this work immeasurably better; in particular, I would like to thank Erik Sudderth, David Dunson, Surya Tokdar, Mike West, Vinayak Rao, Ryan Adams, Annalisa Cerquetti, Peter Müller, Long Nguyen, Levi Boyles, Dan Roy, Zoubin Ghahramani, Yee Whye Teh, Amar Shah, Nick Foti, and Debdeep Pati.



# Contents

|   |           |
|---|-----------|
| <b>Vita</b>   | <b>iv</b> |
| <b>Preface and Acknowledgments</b>  | <b>vi</b> |
| <b>1 Introduction</b>   | <b>1</b>  |
| <b>2 A simple example of Dirichlet process mixture inconsistency for the number of components</b> | <b>4</b>  |
| 2.1 Introduction . . . . .  | 5         |
| 2.2 Discussion . . . . .  | 7         |
| 2.3 Setup . . . . .   | 9         |
| 2.3.1 Dirichlet process mixture model . . . . .   | 10        |
| 2.3.2 Specialization to the standard normal case . . . . .  | 12        |
| 2.4 Simple example of inconsistency . . . . .   | 13        |
| 2.5 Severe inconsistency . . . . .  | 15        |
| <b>3 Inconsistency of Pitman–Yor process mixtures for the number of components</b>                | <b>19</b> |
| 3.1 Introduction . . . . .  | 20        |
| 3.1.1 A motivating example . . . . .  | 20        |
| 3.1.2 Overview of results . . . . .   | 24        |
| 3.1.3 Discussion / related work . . . . .   | 26        |
| 3.1.4 Intuition for the result . . . . .  | 28        |
| 3.1.5 Organization of this chapter . . . . .  | 29        |
| 3.2 Model distribution . . . . .  | 30        |
| 3.2.1 Gibbs partitions . . . . .  | 30        |
| 3.2.2 Gibbs partition mixtures . . . . .  | 33        |
| 3.3 Inconsistency theorem . . . . .   | 35        |
| 3.4 Application to discrete or bounded cases . . . . .  | 39        |
| 3.5 Exponential families and conjugate priors . . . . .   | 42        |
| 3.5.1 Exponential families . . . . .  | 42        |
| 3.5.2 Conjugate priors . . . . .  | 43        |
| 3.6 Application to exponential families . . . . .   | 44        |
| 3.7 Marginal inequalities . . . . .   | 47        |

|          |  |            |
|----------|--|------------|
| 3.8      | Marginal inequality for subsets of the data . . . . .                          | 52         |
| 3.9      | Miscellaneous proofs . . . . .   | 55         |
| 3.10     | Exponential family properties . . . . .  | 58         |
| 3.11     | Bounds on the Laplace approximation . . . . .                                  | 60         |
| 3.12     | Capture lemma . . . . .  | 63         |
| <b>4</b> | <b>The mixture of finite mixtures model</b>                                    | <b>70</b>  |
| 4.1      | Introduction . . . . .   | 71         |
| 4.1.1    | Mixture of finite mixtures (MFM) . . . . .                                     | 73         |
| 4.2      | Properties of MFMs . . . . .   | 76         |
| 4.2.1    | Partition distribution . . . . .   | 76         |
| 4.2.2    | Equivalent models . . . . .  | 78         |
| 4.2.3    | Various basic properties . . . . .   | 80         |
| 4.2.4    | The coefficients $V_n(t)$ . . . . .  | 81         |
| 4.2.5    | Self-consistent marginals . . . . .  | 84         |
| 4.2.6    | Restaurant process / Pólya urn scheme . . . . .                                | 85         |
| 4.2.7    | Random discrete measure formulation . . . . .                                  | 87         |
| 4.2.8    | Density estimates . . . . .  | 89         |
| 4.2.9    | Stick-breaking representation . . . . .  | 91         |
| 4.2.10   | Proofs and details . . . . .   | 93         |
| 4.3      | Asymptotics . . . . .  | 95         |
| 4.3.1    | Posterior asymptotics . . . . .  | 96         |
| 4.3.2    | Relationship between the number of clusters and number of components . . . . . | 104        |
| 4.3.3    | Distribution of the cluster sizes under the prior . . . . .                    | 105        |
| 4.3.4    | Asymptotics of $V_n(t)$ . . . . .  | 107        |
| 4.4      | Inference algorithms . . . . .   | 108        |
| 4.4.1    | Inference with conjugate priors . . . . .                                      | 110        |
| 4.4.2    | Inference with non-conjugate priors . . . . .                                  | 112        |
| 4.4.3    | Justification of the non-conjugate sampler . . . . .                           | 116        |
| <b>5</b> | <b>Experiments with the MFM</b>  | <b>123</b> |
| 5.1      | Univariate normal mixtures . . . . .   | 124        |
| 5.1.1    | Data . . . . .   | 125        |
| 5.1.2    | Model description . . . . .  | 127        |
| 5.1.3    | Approximate inference . . . . .  | 128        |
| 5.1.4    | Density estimation . . . . .   | 129        |
| 5.1.5    | Clustering . . . . .   | 135        |
| 5.1.6    | Number of components and clusters . . . . .                                    | 147        |
| 5.1.7    | Mixing issues . . . . .  | 155        |
| 5.2      | Multivariate skew-normal mixtures . . . . .                                    | 158        |
| 5.2.1    | Data . . . . .   | 159        |
| 5.2.2    | Model description . . . . .  | 162        |
| 5.2.3    | Approximate inference . . . . .  | 162        |
| 5.2.4    | Density estimation . . . . .   | 163        |

|          |   |            |
|----------|---|------------|
| 5.2.5    | Clustering . . . . .  | 164        |
| 5.2.6    | Number of components and clusters . . . . .                                   | 166        |
| <b>6</b> | <b>Combinatorial stochastic processes for other variable-dimension models</b> | <b>169</b> |
| 6.1      | Introduction . . . . .  | 170        |
| 6.2      | Hierarchical mixture of finite mixtures . . . . .                             | 170        |
| 6.2.1    | A hierarchical variable-dimension mixture model . . . . .                     | 172        |
| 6.2.2    | Franchise process . . . . .   | 174        |
| 6.2.3    | Equivalence of the models . . . . .   | 178        |
| 6.2.4    | Inference . . . . .   | 179        |
| 6.3      | Mixture of finite feature models . . . . .                                    | 179        |
| 6.3.1    | A distribution on binary matrices . . . . .                                   | 181        |
| 6.3.2    | A simple urn process . . . . .  | 181        |
| 6.3.3    | Equivalence classes of matrices . . . . .                                     | 183        |
| 6.3.4    | Buffet process . . . . .  | 186        |
| 6.3.5    | Inference . . . . .   | 190        |
| <b>7</b> | <b>Conclusion</b>   | <b>191</b> |

# List of Tables

|     |  |     |
|-----|--|-----|
| 4.1 | Summary of known posterior consistency results . . . . .               | 99  |
| 5.1 | RMS of the differences between pairwise probability matrices . . . . . | 146 |
| 5.2 | MFMM posterior on number of components for the classic galaxy data     | 153 |
| 5.3 | Parameters of the skew-normal mixture used for data simulation . . .   | 160 |

# List of Figures

|      |   |     |
|------|---|-----|
| 2.1  | Posterior on the number of clusters for a univariate normal DPM . . .     | 7   |
| 3.1  | DPM posteriors on the number of clusters . . . . .                        | 22  |
| 3.2  | Partition sampled from the posterior of a bivariate Gaussian DPM . . .    | 23  |
| 3.3  | CDF of the size of the first cluster, given two clusters, in a DPM . . .  | 29  |
| 4.1  | Graphical model for the MFM . . . . .                                     | 74  |
| 4.2  | Alternative graphical models for the MFM . . . . .                        | 79  |
| 4.3  | Graphical models for random discrete measures for the MFM . . . . .       | 87  |
| 5.1  | Histograms of datasets used . . . . .                                     | 126 |
| 5.2  | Estimated densities for the four component distribution . . . . .         | 131 |
| 5.3  | Estimated densities for the Shapley galaxy dataset . . . . .              | 132 |
| 5.4  | Estimated densities for classic galaxy, Shapley galaxy, and SLC . . . .   | 133 |
| 5.5  | Estimated Hellinger distances to the true density . . . . .               | 134 |
| 5.6  | Test set log-likelihood for MFM and DPM . . . . .                         | 136 |
| 5.7  | Sample clusterings from the posterior, on standard normal data . . . .    | 139 |
| 5.8  | Sample clusterings from the posterior, on four component data . . . . .   | 140 |
| 5.9  | Sample clusterings from the posterior, on the classic galaxy data . . . . | 141 |
| 5.10 | Sample clusterings from the posterior, on the Shapley galaxy data . . . . | 142 |
| 5.11 | Sample clusterings from the posterior, on the SLC data . . . . .          | 143 |
| 5.12 | Means of the sorted cluster sizes, given the number of clusters . . . . . | 145 |
| 5.13 | Pairwise probability matrices . . . . .                                   | 148 |
| 5.14 | Posteriors on number of clusters for standard normal data . . . . .       | 150 |
| 5.15 | Posteriors on number of clusters for four component data . . . . .        | 151 |
| 5.16 | Posteriors on number of clusters for the classic galaxy dataset . . . . . | 151 |
| 5.17 | Posteriors on number of clusters for the Shapley galaxy dataset . . . . . | 152 |
| 5.18 | Posteriors on number of clusters for the SLC dataset . . . . .            | 152 |
| 5.19 | Posteriors on the number of components for the MFM . . . . .              | 154 |
| 5.20 | Traceplots of the number of clusters . . . . .                            | 157 |
| 5.21 | Contour plot and scatterplot of a bivariate skew-normal . . . . .         | 160 |
| 5.22 | Skew-normal mixture components used for simulations . . . . .             | 161 |
| 5.23 | Contour plots of density estimates for skew-normal mixtures . . . . .     | 165 |
| 5.24 | Hellinger distances to the true density for skew-normal mixtures . . . .  | 166 |
| 5.25 | Sample clusterings from the posterior for skew-normal mixtures . . . .    | 167 |
| 5.26 | Posteriors of # of components & clusters for skew-normal mixtures . . . . | 168 |
| 6.1  | Graphical models for two representations of the HMFm . . . . .            | 173 |
| 6.2  | Graphical model for combinatorial representation of the HMFm . . . . .    | 178 |
| 6.3  | Sample matrices from the MFFm buffet process . . . . .                    | 188 |

# CHAPTER ONE

---

## Introduction

Nonparametric Bayesian models have found an extraordinarily wide range of applications, including astronomy, meteorology, epidemiology, gene expression profiling, haplotype inference, medical image analysis, survival analysis, econometrics, phylogenetics, species delimitation, computer vision, classification, document modeling, cognitive science, natural language processing, and perhaps more.

Many nonparametric Bayesian models can be viewed as an infinite-dimensional limit of a family of finite-dimensional models. However, another way to construct a flexible Bayesian model is to put a prior on the dimension — that is, to use a variable-dimension model — for example, putting a prior on the number of components in a finite mixture. This thesis (a) analyzes some of the differences and similarities between the nonparametric and variable-dimension approaches, using theory and empirical studies, (b) develops new inference algorithms for these models, and (c) proposes new variable-dimension models and explores their properties.

Primarily, we focus on the Dirichlet process mixture (DPM) and a variable-dimensional alternative that we refer to as the mixture of finite mixtures (MFM) model. One of the main differences between DPMs and MFMs is the behavior of the posterior on the number of clusters. In Chapter 2, we use a simple example to show that, on data from a finite mixture, the DPM posterior on the number of clusters may fail to concentrate at the true number of components. Further, in Chapter 3, we generalize this inconsistency result to a large class of nonparametric mixtures, including DPMs and Pitman–Yor process mixtures over a wide range of families of component distributions. On the other hand, it is known that the MFM posterior on the number of components concentrates at the true number, assuming the data comes from a finite mixture with the same family of component distributions as that used in the model (Nobile, 1994). (Note: The material in Chapters 2 and 3 appears in Miller and Harrison (2013a) and Miller and Harrison (2013b), respectively, so

there is some overlap in their introductions and definitions; on the other hand, this has the benefit of making these two chapters self-contained.)

In Chapter 4, we study the properties of the MFM model, finding that it has many of the same attractive features as the DPM: a simple partition distribution, restaurant process, random discrete measure representation, and in certain special cases, a simple stick-breaking representation, as well as similar exchangeability properties. As a result, many of the same approximate inference algorithms used for nonparametric mixtures can be easily adapted to the MFM. We discuss some of the existing literature on asymptotic results regarding consistency and rates of convergence for the MFM and DPM, for estimating the density, the mixing distribution, and the number of components. We also derive some asymptotic properties of the MFM.

Chapter 5 is an empirical study of the posterior behavior of the MFM model, compared with the DPM. We demonstrate similarities and differences with regard to density estimation, clustering, and the posteriors on the number of clusters and components. These experiments use univariate normal mixtures and bivariate skew-normal mixtures on a variety of real and simulated data sets.

In Chapter 6, we propose two new variable-dimension models, and derive some of their basic properties. For modeling data in separate but related groups, we propose the hierarchical mixture of finite mixtures (HMF<sub>M</sub>) as an alternative to the hierarchical Dirichlet process (HDP) mixture model. For modeling latent binary features, we propose the mixture of finite feature models (MFF<sub>M</sub>) as an alternative to the Indian buffet process (IBP). As with the MFM, these variable-dimension models exhibit some of the same appealing characteristics as their nonparametric counterparts, in particular, simple distributions on discrete structures, exchangeability properties, and representation via combinatorial stochastic processes.



## CHAPTER TWO

---

**A simple example of Dirichlet  
process mixture inconsistency for  
the number of components**

For data assumed to come from a finite mixture with an unknown number of components, it has become common to use Dirichlet process mixtures (DPMs) not only for density estimation, but also for inferences about the number of components. The typical approach is to use the posterior distribution on the number of clusters — that is, the posterior on the number of components represented in the observed data. However, it turns out that this posterior is not consistent — it does not concentrate at the true number of components. In this chapter<sup>1</sup>, we give an elementary proof of this inconsistency in what is perhaps the simplest possible setting: a DPM with normal components of unit variance, applied to data from a “mixture” with one standard normal component. Further, we show that this example exhibits severe inconsistency: instead of going to 1, the posterior probability that there is one cluster converges (in probability) to 0.

## 2.1 Introduction

It is well-known that Dirichlet process mixtures (DPMs) of normals are consistent for the density — that is, given data from a sufficiently regular density  $p_0$  the posterior converges to the point mass at  $p_0$  (see Ghosal (2010) for details and references). However, it is easy to see that this does not necessarily imply consistency for the number of components, since for example, a good estimate of the density might include superfluous components having vanishingly small weight.

Despite the fact that a DPM has infinitely many components with probability 1, it has become common to apply DPMs to data assumed to come from finitely many components or “populations”, and to apply the posterior on the number of clus-

---

<sup>1</sup>The material in this chapter appeared in the proceedings of the NIPS 2013 conference; see Miller and Harrison (2013a).

ters (in other words, the number of components used in the process of generating the observed data) for inferences about the true number of components; see [Huelsenbeck and Andolfatto \(2007\)](#), [Medvedovic and Sivaganesan \(2002\)](#), [Otranto and Gallo \(2002\)](#), [Xing et al. \(2006\)](#), [Fearnhead \(2004\)](#) for a few prominent examples. Of course, if the data-generating process very closely resembles the DPM model, then it makes sense to use this posterior for inferences about the number of clusters (but beware of misspecification; see [Section 2.2](#)). However, in the examples cited, the authors evaluated the performance of their methods on data simulated from a fixed finite number of components or populations, suggesting that they found this to be more realistic than a DPM for their applications.

Therefore, it is important to understand the behavior of this posterior when the data comes from a finite mixture — in particular, does it concentrate at the true number of components? In this chapter, we give a simple example in which a DPM is applied to data from a finite mixture and the posterior distribution on the number of clusters does not concentrate at the true number of components. In fact, DPMs exhibit this type of inconsistency under very general conditions (see [Chapter 3](#)) — however, the aim of this chapter is brevity and clarity. To that end, we focus our attention on a special case that is as simple as possible: a “standard normal DPM”, that is, a DPM using univariate normal components of unit variance, with a standard normal base measure (prior on component means).

The rest of the chapter is organized as follows. In [Section 2.2](#), we address several pertinent questions and consider some suggestive experimental evidence. In [Section 2.3](#), we formally define the DPM model under consideration. In [Section 2.4](#), we give an elementary proof of inconsistency in the case of a standard normal DPM on data from one component, and in [Section 2.5](#), we show that on standard normal data, a standard normal DPM is in fact severely inconsistent.

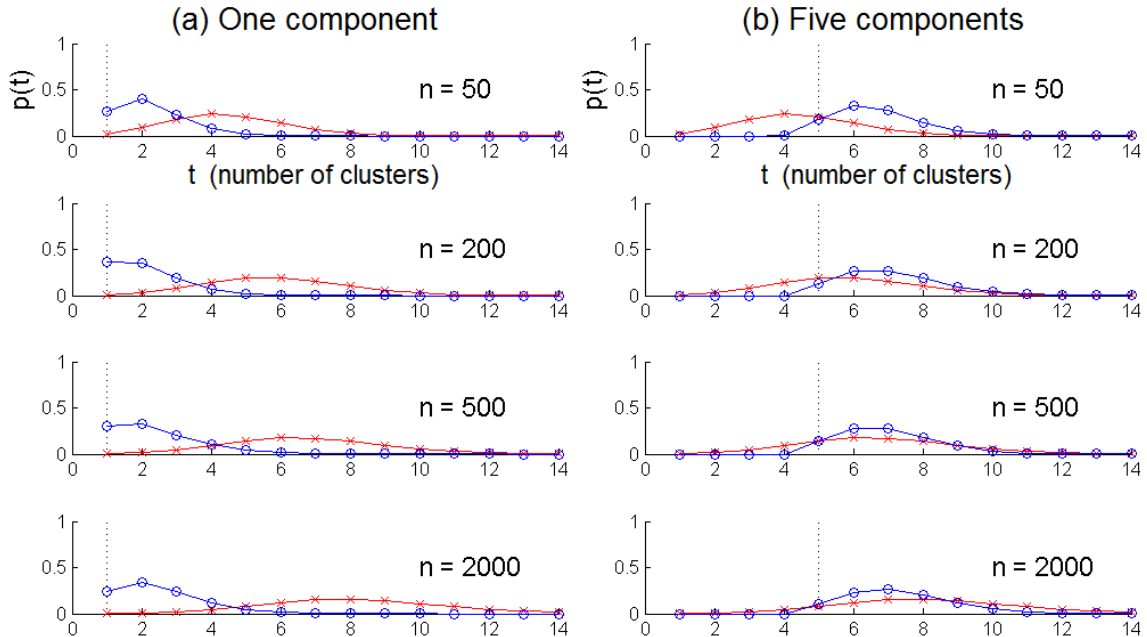


Figure 2.1: Prior (red x) and estimated posterior (blue o) of the number of clusters in the observed data, for a univariate normal DPM on  $n$  i.i.d. samples from (a)  $\mathcal{N}(0, 1)$ , and (b)  $\sum_{k=-2}^2 \frac{1}{5} \mathcal{N}(4k, \frac{1}{2})$ . The DPM had concentration parameter  $\alpha = 1$  and a Normal–Gamma base measure on the mean and precision:  $\mathcal{N}(\mu | 0, 1/c\lambda)\text{Gamma}(\lambda | a, b)$  with  $a = 1$ ,  $b = 0.1$ , and  $c = 0.001$ . Estimates were made using a collapsed Gibbs sampler (MacEachern, 1994), with  $10^4$  burn-in sweeps and  $10^5$  sample sweeps; traceplots and running averages were used as convergence diagnostics. Each plot shown is an average over 5 independent runs.

## 2.2 Discussion

It should be emphasized that these results do not diminish, in any way, the utility of Dirichlet process mixtures as a flexible prior on densities, i.e., for Bayesian density estimation. In addition to their widespread success in empirical studies, DPMs are backed by theoretical guarantees showing that in many cases the posterior on the density concentrates at the true density at the minimax-optimal rate, up to a logarithmic factor (see Ghosal (2010) and references therein).

Many researchers (e.g. West et al. (1994), Onogi et al. (2011), among others) have empirically observed that the DPM posterior on the number of clusters tends to

overestimate the number of components, in the sense that it tends to put its mass on a range of values greater or equal to the true number. Figure 2.1 illustrates this effect for univariate normals, and similar experiments with different families of component distributions yield similar results. Thus, while our theoretical results in Sections 2.4 and 2.5 (and in Chapter 3) are asymptotic in nature, experimental evidence suggests that the issue is present even in small samples.

It is natural to think that this overestimation is due to the fact that the prior on the number of clusters diverges as  $n \rightarrow \infty$ , at a  $\log n$  rate. However, this does not seem to be the main issue — rather, the problem is that DPMS strongly prefer having some tiny clusters and will introduce extra clusters even when they are not needed (see Chapter 3 for an intuitive explanation of why this is the case).

In fact, many researchers have observed the presence of tiny extra clusters (e.g. West et al. (1994), Onogi et al. (2011)), but the reason for this has not previously been well understood, often being incorrectly attributed to the difficulty of detecting components with small weight. These tiny extra clusters are rather inconvenient, especially in clustering applications, and are often dealt with in an *ad hoc* way by simply removing them. It might be possible to consistently estimate the number of components in this way, but this remains an open question.

A more natural solution is the following: if the number of components is unknown, put a prior on the number of components. For example, draw the number of components  $k$  from a probability mass function  $p(k)$  on  $\{1, 2, \dots\}$  with  $p(k) > 0$  for all  $k$ , draw mixing weights  $\pi = (\pi_1, \dots, \pi_k)$  (given  $k$ ), draw component parameters  $\theta_1, \dots, \theta_k$  i.i.d. (given  $k$  and  $\pi$ ) from an appropriate prior, and draw  $X_1, X_2, \dots$  i.i.d. (given  $k$ ,  $\pi$ , and  $\theta_{1:k}$ ) from the resulting mixture. This approach has been widely used (Nobile, 1994, Richardson and Green, 1997, Green and Richardson, 2001, Nobile and

[Fearnside, 2007](#)). Under certain conditions, the posterior on the density has been shown to concentrate at the true density at the minimax-optimal rate, up to a logarithmic factor, for any sufficiently regular true density ([Kruijer et al., 2010](#)). Strictly speaking, as defined, such a model is not identifiable, but it is fairly straightforward to modify it to be identifiable by choosing one representative from each equivalence class. Subject to a modification of this sort, it can be shown (see [Nobile \(1994\)](#)) that under very general conditions, when the data is from a finite mixture of the chosen family, such models are (a.e.) consistent for the number of components, the mixing weights, the component parameters, and the density. Also see [McCullagh and Yang \(2008\)](#) for an interesting discussion about estimating the number of components.

However, as a practical matter, when dealing with real-world data, one would not expect to find data coming exactly from a finite mixture of a known family (except, perhaps, in rare circumstances). Unfortunately, even for a model as in the preceding paragraph, the posterior on the number of components will typically be highly sensitive to misspecification, and it seems likely that in order to obtain robust estimators, the problem itself may need to be reformulated. We urge researchers interested in the number of components to be wary of this robustness issue, and to think carefully about whether they really need to estimate the number of components, or whether some other measure of heterogeneity will suffice.

## 2.3 Setup

In this section, we define the Dirichlet process mixture model under consideration.

### 2.3.1 Dirichlet process mixture model

The DPM model was introduced by [Ferguson \(1983\)](#) and [Lo \(1984\)](#) for the purpose of Bayesian density estimation, and was made practical through the efforts of several authors (see [Escobar and West \(1998\)](#) and references therein); for a more complete list of references, see Chapter 3. In this chapter, we will use  $p(\cdot)$  to denote probabilities under the DPM model (as opposed to other probability distributions that will be considered). The core of the DPM is the so-called Chinese restaurant process (CRP), which defines a certain probability distribution on partitions. Given  $n \in \{1, 2, \dots\}$  and  $t \in \{1, \dots, n\}$ , let  $\mathcal{A}_t(n)$  denote the set of all *ordered* partitions  $(A_1, \dots, A_t)$  of  $\{1, \dots, n\}$  into  $t$  nonempty sets. In other words,

$$\mathcal{A}_t(n) = \left\{ (A_1, \dots, A_t) : A_1, \dots, A_t \text{ are disjoint, } \bigcup_{i=1}^t A_i = \{1, \dots, n\}, |A_i| \geq 1 \forall i \right\}.$$

The CRP with concentration parameter  $\alpha > 0$  defines a probability mass function on  $\mathcal{A}(n) = \bigcup_{t=1}^n \mathcal{A}_t(n)$  by setting

$$p(A) = \frac{\alpha^t}{\alpha^{(n)} t!} \prod_{i=1}^t (|A_i| - 1)!$$

for  $A \in \mathcal{A}_t(n)$ , where  $\alpha^{(n)} = \alpha(\alpha + 1) \cdots (\alpha + n - 1)$ . Note that since  $t$  is a function of  $A$ , we have  $p(A) = p(A, t)$ . (It is more common to see this distribution defined in terms of unordered partitions  $\{A_1, \dots, A_t\}$ , in which case the  $t!$  does not appear in the denominator — however, for our purposes it is more convenient to use the distribution on ordered partitions  $(A_1, \dots, A_t)$  obtained by uniformly permuting the parts. This does not affect the prior or posterior on  $t$ .)

Consider the hierarchical model

$$\begin{aligned}
 p(A, t) &= p(A) = \frac{\alpha^t}{\alpha^{(n)} t!} \prod_{i=1}^t (|A_i| - 1)!, & (2.3.1) \\
 p(\theta_{1:t} \mid A, t) &= \prod_{i=1}^t \pi(\theta_i), \text{ and} \\
 p(x_{1:n} \mid \theta_{1:t}, A, t) &= \prod_{i=1}^t \prod_{j \in A_i} p_{\theta_i}(x_j),
 \end{aligned}$$

where  $\pi(\theta)$  is a prior on component parameters  $\theta \in \Theta$ , and  $\{p_\theta : \theta \in \Theta\}$  is a parametrized family of distributions on  $x \in \mathcal{X}$  for the components. Typically,  $\mathcal{X} \subset \mathbb{R}^d$  and  $\Theta \subset \mathbb{R}^k$  for some  $d$  and  $k$ . Here,  $x_{1:n} = (x_1, \dots, x_n)$  with  $x_i \in \mathcal{X}$ , and  $\theta_{1:t} = (\theta_1, \dots, \theta_t)$  with  $\theta_i \in \Theta$ . This hierarchical model is referred to as a *Dirichlet process mixture (DPM) model*.

The prior on the number of clusters  $t$  under this model is  $p_n(t) = \sum_{A \in \mathcal{A}_t(n)} p(A, t)$ . We use  $T_n$  (rather than  $T$ ) to denote the random variable representing the number of clusters, as a reminder that its distribution depends on  $n$ . Note that we distinguish between the terms “component” and “cluster”: a *component* is part of a mixture distribution (e.g. a mixture  $\sum_{i=1}^{\infty} \pi_i p_{\theta_i}$  has components  $p_{\theta_1}, p_{\theta_2}, \dots$ ), while a *cluster* is the set of indices of data points coming from a given component (e.g. in the DPM model above,  $A_1, \dots, A_t$  are the clusters).

Since we are concerned with the posterior distribution  $p(T_n = t \mid x_{1:n})$  on the number of clusters, we will be especially interested in the marginal distribution on



$(x_{1:n}, t)$ , given by

$$\begin{aligned}
p(x_{1:n}, T_n = t) &= \sum_{A \in \mathcal{A}_t(n)} \int p(x_{1:n}, \theta_{1:t}, A, t) d\theta_{1:t} \\
&= \sum_{A \in \mathcal{A}_t(n)} p(A) \prod_{i=1}^t \int \left( \prod_{j \in A_i} p_{\theta_i}(x_j) \right) \pi(\theta_i) d\theta_i \\
&= \sum_{A \in \mathcal{A}_t(n)} p(A) \prod_{i=1}^t m(x_{A_i}) \tag{2.3.2}
\end{aligned}$$

where for any subset of indices  $S \subset \{1, \dots, n\}$ , we denote  $x_S = (x_j : j \in S)$  and let  $m(x_S)$  denote the single-cluster marginal of  $x_S$ ,

$$m(x_S) = \int \left( \prod_{j \in S} p_{\theta}(x_j) \right) \pi(\theta) d\theta. \tag{2.3.3}$$

### 2.3.2 Specialization to the standard normal case

In this chapter, for brevity and clarity, we focus on the univariate normal case with unit variance, with a standard normal prior on means — that is, for  $x \in \mathbb{R}$  and  $\theta \in \mathbb{R}$ ,

$$\begin{aligned}
p_{\theta}(x) &= \mathcal{N}(x \mid \theta, 1) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(x - \theta)^2\right), \quad \text{and} \\
\pi(\theta) &= \mathcal{N}(\theta \mid 0, 1) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\theta^2\right).
\end{aligned}$$

It is a straightforward calculation to show that the single-cluster marginal is then

$$m(x_{1:n}) = \frac{1}{\sqrt{n+1}} p_0(x_{1:n}) \exp\left(\frac{1}{2} \frac{1}{n+1} \left(\sum_{j=1}^n x_j\right)^2\right), \tag{2.3.4}$$

where  $p_0(x_{1:n}) = p_0(x_1) \cdots p_0(x_n)$  (and  $p_0$  is the  $\mathcal{N}(0, 1)$  density). When  $p_\theta(x)$  and  $\pi(\theta)$  are as above, we refer to the resulting DPM as a *standard normal DPM*.

## 2.4 Simple example of inconsistency

In this section, we prove the following result, exhibiting a simple example in which a DPM is inconsistent for the number of components: even when the true number of components is 1 (e.g.  $\mathcal{N}(\mu, 1)$  data), the posterior probability of  $T_n = 1$  does not converge to 1. Interestingly, the result applies even when  $X_1, X_2, \dots$  are identically equal to a constant  $c \in \mathbb{R}$ . To keep it simple, we set  $\alpha = 1$ ; for more general results, see Chapter 3.

**Theorem 2.4.1.** *If  $X_1, X_2, \dots \in \mathbb{R}$  are i.i.d. from any distribution with  $\mathbb{E}|X_i| < \infty$ , then with probability 1, under the standard normal DPM with  $\alpha = 1$  as defined above,  $p(T_n = 1 \mid X_{1:n})$  does not converge to 1 as  $n \rightarrow \infty$ .*

*Proof.* Let  $n \in \{2, 3, \dots\}$ . Let  $x_1, \dots, x_n \in \mathbb{R}$ ,  $A \in \mathcal{A}_2(n)$ , and  $a_i = |A_i|$  for  $i = 1, 2$ . Define  $s_n = \sum_{j=1}^n x_j$  and  $s_{A_i} = \sum_{j \in A_i} x_j$  for  $i = 1, 2$ . Using Equation 2.3.4 and noting that  $1/(n+1) \leq 1/(n+2) + 1/n^2$ , we have

$$\sqrt{n+1} \frac{m(x_{1:n})}{p_0(x_{1:n})} = \exp\left(\frac{1}{2} \frac{s_n^2}{n+1}\right) \leq \exp\left(\frac{1}{2} \frac{s_n^2}{n+2}\right) \exp\left(\frac{1}{2} \frac{s_n^2}{n^2}\right).$$

The second factor equals  $\exp(\frac{1}{2} \bar{x}_n^2)$ , where  $\bar{x}_n = \frac{1}{n} \sum_{j=1}^n x_j$ . By the convexity of  $x \mapsto x^2$ ,

$$\left(\frac{s_n}{n+2}\right)^2 \leq \frac{a_1+1}{n+2} \left(\frac{s_{A_1}}{a_1+1}\right)^2 + \frac{a_2+1}{n+2} \left(\frac{s_{A_2}}{a_2+1}\right)^2,$$

and thus, the first factor is less or equal to

$$\exp\left(\frac{1}{2}\frac{s_{A_1}^2}{a_1+1} + \frac{1}{2}\frac{s_{A_2}^2}{a_2+1}\right) = \sqrt{a_1+1}\sqrt{a_2+1} \frac{m(x_{A_1})m(x_{A_2})}{p_0(x_{1:n})}.$$

Hence,

$$\frac{m(x_{1:n})}{m(x_{A_1})m(x_{A_2})} \leq \frac{\sqrt{a_1+1}\sqrt{a_2+1}}{\sqrt{n+1}} \exp\left(\frac{1}{2}\bar{x}_n^2\right). \quad (2.4.1)$$

Consequently, we have

$$\begin{aligned} \frac{p(x_{1:n}, T_n = 2)}{p(x_{1:n}, T_n = 1)} &\stackrel{(a)}{=} \sum_{A \in \mathcal{A}_2(n)} n p(A) \frac{m(x_{A_1})m(x_{A_2})}{m(x_{1:n})} \\ &\stackrel{(b)}{\geq} \sum_{A \in \mathcal{A}_2(n)} n p(A) \frac{\sqrt{n+1}}{\sqrt{|A_1|+1}\sqrt{|A_2|+1}} \exp\left(-\frac{1}{2}\bar{x}_n^2\right) \\ &\stackrel{(c)}{\geq} \sum_{\substack{A \in \mathcal{A}_2(n): \\ |A_1|=1}} n \frac{(n-2)!}{n!2!} \frac{\sqrt{n+1}}{\sqrt{2}\sqrt{n}} \exp\left(-\frac{1}{2}\bar{x}_n^2\right) \\ &\stackrel{(d)}{\geq} \frac{1}{2\sqrt{2}} \exp\left(-\frac{1}{2}\bar{x}_n^2\right), \end{aligned}$$

where step (a) follows from applying Equation 2.3.2 to both numerator and denominator, plus using Equation 2.3.1 (with  $\alpha = 1$ ) to see that  $p(A) = 1/n$  when  $A = (\{1, \dots, n\})$ , step (b) follows from Equation 2.4.1 above, step (c) follows since all the terms in the sum are nonnegative and  $p(A) = (n-2)!/n!2!$  when  $|A_1| = 1$  (by Equation 2.3.1, with  $\alpha = 1$ ), and step (d) follows since there are  $n$  partitions  $A \in \mathcal{A}_2(n)$  such that  $|A_1| = 1$ .

If  $X_1, X_2, \dots \in \mathbb{R}$  are i.i.d. with  $\mu = \mathbb{E}X_j$  finite, then by the law of large numbers,

$\bar{X}_n = \frac{1}{n} \sum_{j=1}^n X_j \rightarrow \mu$  almost surely as  $n \rightarrow \infty$ . Therefore,

$$\begin{aligned} p(T_n = 1 \mid X_{1:n}) &= \frac{p(X_{1:n}, T_n = 1)}{\sum_{t=1}^{\infty} p(X_{1:n}, T_n = t)} \leq \frac{p(X_{1:n}, T_n = 1)}{p(X_{1:n}, T_n = 1) + p(X_{1:n}, T_n = 2)} \\ &\leq \frac{1}{1 + \frac{1}{2\sqrt{2}} \exp(-\frac{1}{2}\bar{X}_n^2)} \xrightarrow{\text{a.s.}} \frac{1}{1 + \frac{1}{2\sqrt{2}} \exp(-\frac{1}{2}\mu^2)} < 1. \end{aligned}$$

Hence, almost surely,  $p(T_n = 1 \mid X_{1:n})$  does not converge to 1.  $\square$

## 2.5 Severe inconsistency

In the previous section, we showed that  $p(T_n = 1 \mid X_{1:n})$  does not converge to 1 for a standard normal DPM on any data with finite mean. In this section, we prove that in fact, it converges to 0 on standard normal data. This vividly illustrates that improperly using DPMs in this way can lead to entirely misleading results. The key step in the proof is an application of Hoeffding's strong law of large numbers for U-statistics.

**Theorem 2.5.1.** *If  $X_1, X_2, \dots \sim \mathcal{N}(0, 1)$  i.i.d. then*

$$p(T_n = 1 \mid X_{1:n}) \xrightarrow{\text{Pr}} 0 \quad \text{as } n \rightarrow \infty$$

*under the standard normal DPM with concentration parameter  $\alpha = 1$ .*

*Proof.* For  $t = 1$  and  $t = 2$  define

$$R_t(X_{1:n}) = n^{3/2} \frac{p(X_{1:n}, T_n = t)}{p_0(X_{1:n})}.$$

Our method of proof is as follows. We will show that

$$R_2(X_{1:n}) \xrightarrow[n \rightarrow \infty]{\text{Pr}} \infty$$

(or in other words, for any  $B > 0$  we have  $\mathbb{P}(R_2(X_{1:n}) > B) \rightarrow 1$  as  $n \rightarrow \infty$ ), and we will show that  $R_1(X_{1:n})$  is bounded in probability:

$$R_1(X_{1:n}) = O_P(1)$$

(or in other words, for any  $\varepsilon > 0$  there exists  $B_\varepsilon > 0$  such that  $\mathbb{P}(R_1(X_{1:n}) > B_\varepsilon) \leq \varepsilon$  for all  $n \in \{1, 2, \dots\}$ ). Putting these two together, we will have

$$p(T_n = 1 | X_{1:n}) = \frac{p(X_{1:n}, T_n = 1)}{\sum_{t=1}^{\infty} p(X_{1:n}, T_n = t)} \leq \frac{p(X_{1:n}, T_n = 1)}{p(X_{1:n}, T_n = 2)} = \frac{R_1(X_{1:n})}{R_2(X_{1:n})} \xrightarrow[n \rightarrow \infty]{\text{Pr}} 0.$$

First, let's show that  $R_2(X_{1:n}) \rightarrow \infty$  in probability. For  $S \subset \{1, \dots, n\}$  with  $|S| \geq 1$ , define  $h(x_S)$  by

$$h(x_S) = \frac{m(x_S)}{p_0(x_S)} = \frac{1}{\sqrt{|S|+1}} \exp\left(\frac{1}{2} \frac{1}{|S|+1} \left(\sum_{j \in S} x_j\right)^2\right),$$

where  $m$  is the single-cluster marginal as in Equations 2.3.3 and 2.3.4. Note that when  $1 \leq |S| \leq n-1$ , we have  $\sqrt{n} h(x_S) \geq 1$ . Note also that  $\mathbb{E}h(X_S) = 1$  since

$$\mathbb{E}h(X_S) = \int h(x_S) p_0(x_S) dx_S = \int m(x_S) dx_S = 1,$$

using the fact that  $m(x_S)$  is a density with respect to Lebesgue measure. For  $k \in \{1, \dots, n\}$ , define the U-statistics

$$U_k(X_{1:n}) = \frac{1}{\binom{n}{k}} \sum_{|S|=k} h(X_S)$$

where the sum is over all  $S \subset \{1, \dots, n\}$  such that  $|S| = k$ . By Hoeffding's strong law of large numbers for U-statistics ([Hoeffding, 1961](#)),

$$U_k(X_{1:n}) \xrightarrow[n \rightarrow \infty]{\text{a.s.}} \mathbb{E}h(X_{1:k}) = 1$$

for any  $k \in \{1, 2, \dots\}$ . Therefore, using Equations [2.3.1](#) and [2.3.2](#) we have that for any  $K \in \{1, 2, \dots\}$  and any  $n > K$ ,

$$\begin{aligned} R_2(X_{1:n}) &= n^{3/2} \sum_{A \in \mathcal{A}_2(n)} p(A) \frac{m(X_{A_1}) m(X_{A_2})}{p_0(X_{1:n})} \\ &= n \sum_{A \in \mathcal{A}_2(n)} p(A) \sqrt{n} h(X_{A_1}) h(X_{A_2}) \\ &\geq n \sum_{A \in \mathcal{A}_2(n)} p(A) h(X_{A_1}) \\ &= n \sum_{k=1}^{n-1} \sum_{|S|=k} \frac{(k-1)! (n-k-1)!}{n! 2!} h(X_S) \\ &= \sum_{k=1}^{n-1} \frac{n}{2k(n-k)} \frac{1}{\binom{n}{k}} \sum_{|S|=k} h(X_S) \\ &= \sum_{k=1}^{n-1} \frac{n}{2k(n-k)} U_k(X_{1:n}) \\ &\geq \sum_{k=1}^K \frac{n}{2k(n-k)} U_k(X_{1:n}) \\ &\xrightarrow[n \rightarrow \infty]{\text{a.s.}} \sum_{k=1}^K \frac{1}{2k} = \frac{H_K}{2} > \frac{\log K}{2} \end{aligned}$$

where  $H_K$  is the  $K^{\text{th}}$  harmonic number, and the last inequality follows from the standard bounds ([Graham et al., 1989](#)) on harmonic numbers:  $\log K < H_K \leq \log K + 1$ . Hence, for any  $K$ ,

$$\liminf_{n \rightarrow \infty} R_2(X_{1:n}) > \frac{\log K}{2} \quad \text{almost surely,}$$

and it follows easily that

$$R_2(X_{1:n}) \xrightarrow[n \rightarrow \infty]{\text{a.s.}} \infty.$$

Convergence in probability is implied by almost sure convergence.

Now, let's show that  $R_1(X_{1:n}) = O_P(1)$ . By Equations 2.3.1, 2.3.2, and 2.3.4, we have

$$\begin{aligned} R_1(X_{1:n}) &= n^{3/2} \frac{p(X_{1:n}, T_n = 1)}{p_0(X_{1:n})} = \sqrt{n} \frac{m(X_{1:n})}{p_0(X_{1:n})} \\ &= \frac{\sqrt{n}}{\sqrt{n+1}} \exp\left(\frac{1}{2} \frac{n}{n+1} \left(\frac{1}{\sqrt{n}} \sum_{i=1}^n X_i\right)^2\right) \leq \exp(Z_n^2/2) \end{aligned}$$

where  $Z_n = (1/\sqrt{n}) \sum_{i=1}^n X_i \sim \mathcal{N}(0, 1)$  for each  $n \in \{1, 2, \dots\}$ . Since  $Z_n = O_P(1)$  then we conclude that  $R_1(X_{1:n}) = O_P(1)$ . This completes the proof.  $\square$

# CHAPTER THREE

---

## Inconsistency of Pitman–Yor process mixtures for the number of components



## 3.1 Introduction

In many applications, a finite mixture is a natural model, but it can be difficult to choose an appropriate number of components. To circumvent this choice, investigators are increasingly turning to Dirichlet process mixtures (DPMs), and Pitman–Yor process mixtures (PYMs), more generally. While these models may be well-suited for Bayesian density estimation, some investigators are using them for inferences about the number of components, by considering the posterior on the number of components represented in the observed data. We show that this posterior is not consistent — that is, on data from a finite mixture, it does not concentrate at the true number of components. This result applies to a large class of nonparametric mixtures, including DPMs and PYMs, over a wide variety of families of component distributions, including essentially all discrete families, as well as continuous exponential families satisfying mild regularity conditions (such as multivariate Gaussians).<sup>1</sup>

### 3.1.1 A motivating example

In population genetics, determining the “population structure” is an important step in the analysis of sampled data. As an illustrative example, consider the impala, a species of antelope in southern Africa. Impalas are divided into two subspecies: the common impala occupying much of the eastern half of the region, and the black-faced impala inhabiting a small area in the west. While common impalas are abundant, the number of black-faced impalas has been decimated by drought, poaching, and declining resources due to human and livestock expansion. To assist conservation efforts, [Lorenzen et al. \(2006\)](#) collected samples from 216 impalas, and analyzed the

---

<sup>1</sup>The material in this chapter appears in a manuscript that has been submitted for publication; see [Miller and Harrison \(2013b\)](#).

genetic variation between/within the two subspecies.

A key part of their analysis consisted of inferring the population structure — that is, partitioning the data into distinct populations, and in particular, determining how many such populations there are. To infer the impala population structure, Lorenzen et al. employed a widely-used tool called STRUCTURE (Pritchard et al., 2000) which, in the simplest version, models the data as a finite mixture, with each component in the mixture corresponding to a distinct population. STRUCTURE uses an *ad hoc* method to choose the number of components, but this comes with no guarantees.

Seeking a more principled approach, Pella and Masuda (2006) proposed using a Dirichlet process mixture (DPM). Now, in a DPM, the number of components is infinite with probability 1, and thus the posterior on the number of components is always, trivially, a point mass at infinity. Consequently, Pella and Masuda instead employed the posterior on the number of clusters (that is, the number of components used in generating the data observed so far) for inferences about the number of components. (The terms “component” and “cluster” are often used interchangeably, but we make the following crucial distinction: a component is part of a mixture distribution, while a cluster is the set of indices of datapoints coming from a given component.) This DPM approach was implemented in a software tool called STRUCTURAMA (Huelsenbeck and Andolfatto, 2007), and demonstrated on the impala data of Lorenzen et al.; see Figure 3.1(a).

STRUCTURAMA has gained acceptance within the population genetics community, and has been used in studies of a variety of organisms, from apples and avocados, to sardines and geckos (Richards et al., 2009, Chen et al., 2009, Gonzalez and Zardoya, 2007, Leaché and Fujita, 2010). Studies such as these can carry significant weight, since they may be used by officials to make informed policy decisions

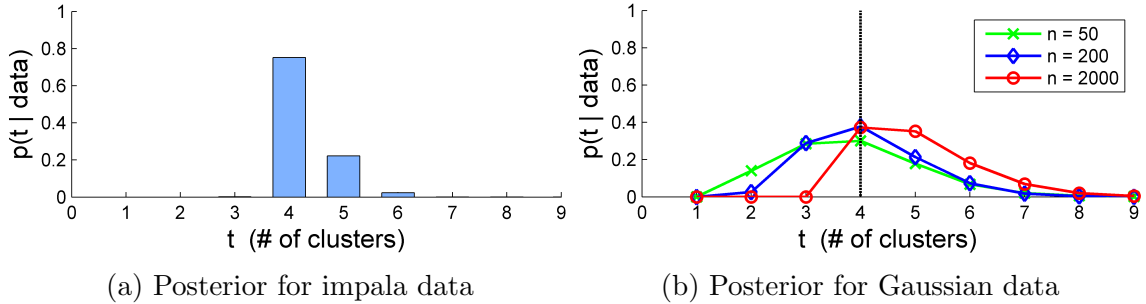


Figure 3.1: Estimated DPM posterior distribution of the number of clusters: (a) For the impala data of Lorenzen et al. ( $n = 216$  datapoints). Our empirical results, shown here, agree with those of Huelsenbeck and Andolfatto. (b) For bivariate Gaussian data from a four-component mixture; see Figure 3.2. Each plot is the average over 10 independently-drawn datasets. (Lines drawn for visualization purposes only.) (For (a) and (b), estimates were made via Gibbs sampling, with  $10^4$  burn-in sweeps and  $10^5$  sample sweeps.)

regarding agriculture, conservation, and public health.

More generally, in a number of applications the same scenario has played out: a finite mixture seems to be a natural model, but requires the user to choose the number of components, while a Dirichlet process mixture offers a convenient way to avoid this choice. For nonparametric Bayesian density estimation, DPMS are indeed attractive, since the posterior on the density exhibits nice convergence properties; see Section 3.1.3. However, in several applications, investigators have drawn inferences from the posterior on the number of clusters — not just the density — on the assumption that this is informative about the number of components. Further examples include gene expression profiling (Medvedovic and Sivaganesan, 2002), haplotype inference (Xing et al., 2006), econometrics (Otranto and Gallo, 2002), and evaluation of inference algorithms (Fearnhead, 2004). Of course, if the data-generating process is well-modeled by a DPM (and in particular, there are infinitely many components), then it is sensible to use this posterior for inference about the number of components represented so far in the data — but that does not seem to be the perspective of these investigators, since they measure performance on simulated data coming from

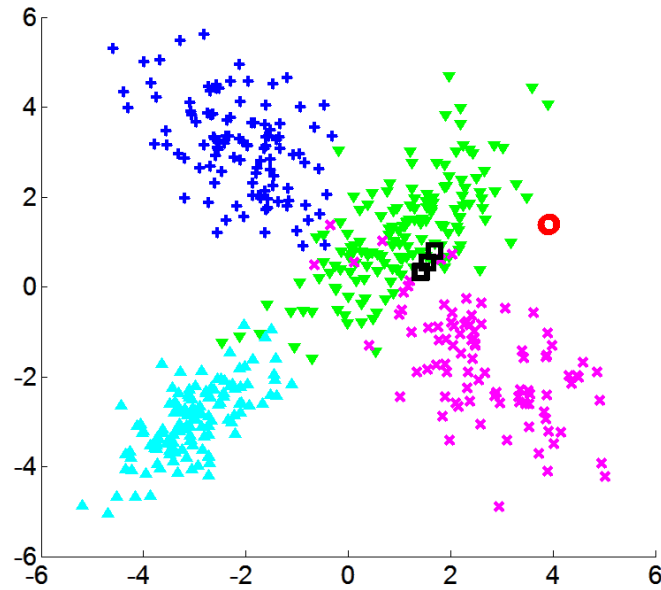


Figure 3.2: A typical partition sampled from the posterior of a Dirichlet process mixture of bivariate Gaussians, on simulated data from a four-component mixture. Different clusters have different marker shapes ( $+$ ,  $\times$ ,  $\nabla$ ,  $\Delta$ ,  $\circ$ ,  $\square$ ) and different colors. Note the tiny “extra” clusters ( $\circ$  and  $\square$ ), in addition to the four dominant clusters.

finitely many components or populations. (To be clear, we should note that although several investigators have used DPMs in this way, most of those who are familiar with nonparametric models are well aware that this usage is questionable.)

Therefore, it is important to understand the properties of this procedure. Simulation results give some cause for concern; for instance, Figures 3.1(b) and 3.2 display results for data from a mixture of two-dimensional Gaussians with four components. Partitions sampled from the posterior often have tiny “extra” clusters, and the posterior on the number of clusters does not appear to be concentrating as the number of datapoints  $n$  increases. This raises a fundamental question that has not been addressed in the literature: With enough data, will this posterior eventually concentrate at the true number of components? In other words, is it consistent?

### 3.1.2 Overview of results

In this chapter, we prove that under fairly general conditions, when using a Dirichlet process mixture, the posterior on the number of clusters will not concentrate at any finite value, and therefore will not be consistent for the number of components in a finite mixture. In fact, our results apply to a large class of nonparametric mixtures including DPMs, and Pitman–Yor process mixtures (PYMs) more generally, over a wide variety of families of component distributions.

Before treating our general results and their prerequisite technicalities, we would like to highlight a few interesting special cases that can be succinctly stated. The terminology and notation used below will be made precise in later sections. To reiterate, our results are considerably more general than the following corollary, which is simply presented for the reader’s convenience.

**Corollary 3.1.1.** *Consider a Pitman–Yor process mixture with component distributions from one of the following families:*

- (a)  $\text{Normal}(\mu, \Sigma)$  (*multivariate Gaussian*),
- (b)  $\text{Exponential}(\theta)$ ,
- (c)  $\text{Gamma}(a, b)$ ,
- (d)  $\text{Log-Normal}(\mu, \sigma^2)$ , *or*
- (e)  $\text{Weibull}(a, b)$  *with fixed shape*  $a > 0$ ,

*along with a base measure that is a conjugate prior of the form in Section 3.5.2, or*

(f) any discrete family  $\{P_\theta\}$  such that  $\bigcap_\theta \{x : P_\theta(x) > 0\} \neq \emptyset$  (e.g., Poisson, Geometric, Negative Binomial, Binomial, Multinomial, etc.),

along with any continuous base measure. Consider any  $t \in \{1, 2, \dots\}$ , except for  $t = N$  in the case of a Pitman–Yor process with parameters  $\sigma < 0$  and  $\vartheta = N|\sigma|$ . If  $X_1, X_2, \dots$  are i.i.d. from a mixture with  $t$  components from the family used in the model, then the posterior on the number of clusters  $T_n$  is not consistent for  $t$ , and in fact,

$$\limsup_{n \rightarrow \infty} p(T_n = t \mid X_1, \dots, X_n) < 1$$

with probability 1.

This is implied by Theorems 3.3.4, 3.4.1, and 3.6.2. These more general theorems apply to a broad class of partition distributions, handling Pitman–Yor processes as a special case, and they apply to many other families of component distributions: Theorem 3.6.2 covers a large class of exponential families, and Theorem 3.4.1 covers families satisfying a certain boundedness condition on the densities (including any case in which the model and data distributions have one or more point masses in common, as well as many location–scale families with scale bounded away from zero). Dirichlet processes are subsumed as a further special case, being Pitman–Yor processes with parameters  $\sigma = 0$  and  $\vartheta > 0$ . Also, the assumption of i.i.d. data from a finite mixture is much stronger than what is required by these results.

Regarding the exception of  $t = N$  when  $\sigma < 0$  in Corollary 3.1.1: posterior consistency at  $t = N$  is possible, however, this could only occur if the chosen parameter  $N$  just happens to be equal to the actual number of components,  $t$ . On the other hand, consistency at any  $t$  can (in principle) be obtained by putting a prior on  $N$ ; see Section 3.1.3 below. In a similar vein, some investigators place a prior on the

concentration parameter  $\vartheta$  in a DPM, or allow  $\vartheta$  to depend on  $n$ ; we conjecture that inconsistency can still occur in these cases, but in this work, we examine only the case of fixed  $\sigma$  and  $\vartheta$ .

### 3.1.3 Discussion / related work

We would like to emphasize that this inconsistency should not be viewed as a deficiency of Dirichlet process mixtures, but is simply due to a misapplication of them. As flexible priors on densities, DPMs are superb, and there are strong results showing that in many cases the posterior on the density converges in  $L_1$  to the true density at the minimax-optimal rate, up to a logarithmic factor (see [Scricciolo \(2012\)](#), [Ghosal \(2010\)](#) and references therein). Further, [Nguyen \(2013\)](#) has recently shown that the posterior on the mixing distribution converges in the Wasserstein metric to the true mixing distribution. However, these results do not necessarily imply consistency for the number of components, since any mixture can be approximated arbitrarily well in these metrics by another mixture with a larger number of components (for instance, by making the weights of the extra components infinitesimally small). There seems to be no prior work on consistency of DPMs (or PYMs) for the number of components in a finite mixture (aside from [Miller and Harrison \(2013a\)](#) — appearing here as Chapter 2 — in which we discuss the very special case of a DPM on data from a univariate Gaussian “mixture” with one component of known variance).

In the context of “species sampling”, several authors have studied the Pitman–Yor process posterior (see [James \(2008\)](#), [Jang et al. \(2010\)](#), [Lijoi et al. \(2007\)](#) and references therein), but this is very different from our situation — in a species sampling model, the observed data is drawn directly from a measure with a Pitman–Yor process prior, while in a PYM model, the observed data is drawn from a mixture

with such a measure as the mixing distribution.

[Rousseau and Mengersen \(2011\)](#) proved an interesting result on “overfitted” mixtures, in which data from a finite mixture is modeled by a finite mixture with too many components. In cases where this approximates a DPM, their result implies that the posterior weight of the extra components goes to zero. In a rough sense, this is complementary to our results, which involve showing that there are always some nonempty (but perhaps small) extra clusters.

Empirically, many investigators have noticed that the DPM posterior tends to overestimate the number of components (e.g. [West et al. \(1994\)](#), [Lartillot and Philippe \(2004\)](#), [Onogi et al. \(2011\)](#)), and such observations are consistent with our theoretical results. This overestimation seems to occur because there are typically a few tiny “extra” clusters. Among researchers using DPMs for clustering, this is an annoyance that is often dealt with by pruning such clusters — that is, by simply ignoring them when calculating statistics such as the number of clusters. It may be possible to obtain consistent estimators in this way, but this remains an open question; [Rousseau and Mengersen’s \(2011\)](#) results may be applicable here.

Under the (strong) assumption that the family of component distributions is correctly specified, one can obtain posterior consistency by simply putting a prior on the number of components in a finite mixture ([Nobile, 1994](#)). (It turns out that putting a prior on  $N$  in a PYM with  $\sigma < 0$ ,  $\vartheta = N|\sigma|$  is a special case of this ([Gnedin and Pitman, 2006](#)).) That said, it seems likely that such estimates will be severely affected by misspecification of the model, which is inevitable in most applications. Robustness to model misspecification seems essential for reliable estimation of the number of components, for real-world data. Other approaches have also been proposed for estimating the number of components ([Henna, 1985](#), [Keribin,](#)



2000, Leroux, 1992, Ishwaran et al., 2001, James et al., 2001, Henna, 2005), and steps toward addressing the issue of robustness have been taken by Woo and Sriram (2006, 2007).

### 3.1.4 Intuition for the result

To illustrate the intuition behind this inconsistency, consider a Dirichlet process with concentration parameter  $\vartheta = 1$ . (Similar reasoning applies for any Pitman–Yor process with  $\sigma \geq 0$ , but the  $\sigma < 0$  case is somewhat different.) It is tempting to think that the prior on the number of clusters is the culprit, since (as is well-known) it diverges as  $n \rightarrow \infty$ . Surprisingly, this does not seem to be the main reason why inconsistency occurs.

Instead, the right intuition comes from examining the prior on partitions, *given* the number of clusters. The prior on ordered partitions  $A = (A_1, \dots, A_t)$  is  $p(A) = (n! t!)^{-1} \prod_{i=1}^t (a_i - 1)!$ , where  $t$  is the number of parts (i.e. clusters) and  $a_i = |A_i|$  is the size of the  $i$ th part. (The  $t!$  comes from uniformly permuting the parts; see Section 3.2.1.) Since there are  $n!/(a_1! \cdots a_t!)$  such partitions with part sizes  $(a_1, \dots, a_t)$ , the conditional distribution of the sizes  $(a_1, \dots, a_t)$  given  $t$  is proportional to  $a_1^{-1} \cdots a_t^{-1}$  (subject to the constraint that  $\sum a_i = n$ ). See Figure 3.3 for the case of  $t = 2$ . The key observation is that, for large  $n$ , this conditional distribution is heavily concentrated in the “corners”, where one or more of the  $a_i$ ’s is small.

By pursuing this line of thought, one can show that the probability of drawing a partition with  $t + 1$  parts and one or more of the  $a_i$ ’s equal to 1 is, at least, the same order of magnitude (with respect to  $n$ ) as the probability of drawing a partition with  $t$  parts. This leads to the basic idea of the proof — if the likelihood of the data is

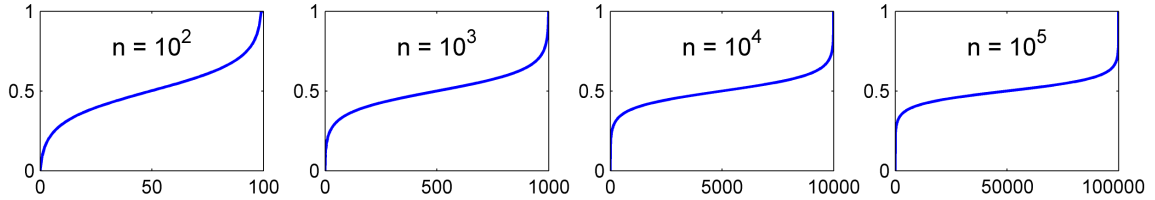


Figure 3.3: The cumulative distribution function of the conditional distribution of  $a_1$  given that  $t = 2$ , for a Dirichlet process with  $\vartheta = 1$ . As  $n$  increases, the distribution becomes concentrated at the extremes.

also the same order of magnitude, then the posterior probability of  $t + 1$  will not be too much smaller than that of  $t$ . Roughly speaking, the posterior will always find it reasonably attractive to split off one element as a singleton. Handling the likelihood is the difficult part, and occupies the bulk of the proof.

### 3.1.5 Organization of this chapter

In Section 3.2, we define Gibbs partition mixture models, which includes Pitman–Yor and Dirichlet process mixtures as special cases. In Section 3.3, we prove a general inconsistency theorem for Gibbs partition mixtures satisfying certain conditions. In Section 3.4, we apply the theorem to cases satisfying a certain boundedness condition on the densities, including discrete families as a special case. In Section 3.5, we introduce notation for exponential families and conjugate priors, and in Section 3.6, we apply the theorem to cases in which the mixture is over an exponential family satisfying some regularity conditions. The remainder of the chapter proves the key lemma used in this application. In Section 3.7, we obtain certain inequalities involving the marginal density under an exponential family with conjugate prior. In Section 3.8, we prove the key lemma of Section 3.6: an inequality involving the marginal density of any sufficiently large subset of the data. Sections 3.9, 3.10, 3.11, and 3.12 contain various supporting results.

## 3.2 Model distribution

Our analysis involves two probability distributions: one which is defined by the model, and another which gives rise to the data. In this section, we describe the model distribution.

Building upon the Dirichlet process (Ferguson, 1973, Blackwell and MacQueen, 1973, Antoniak, 1974), Dirichlet process mixtures were first studied by Antoniak (1974) (who also considered the mixtures of Dirichlet processes arising in the posterior), Berry and Christensen (1979), Ferguson (1983), and Lo (1984), and were later made practical through the efforts of a number of authors (Escobar, 1988, MacEachern, 1994, Escobar and West, 1995, West, 1992, West et al., 1994, Neal, 1992, Liu, 1994, Bush and MacEachern, 1996, MacEachern and Müller, 1998, MacEachern, 1998, Escobar and West, 1998, MacEachern et al., 1999, Neal, 2000). Pitman–Yor process mixtures (Ishwaran and James, 2003) are a generalization of DPMs based on the Pitman–Yor process (Perman et al., 1992, Pitman and Yor, 1997), also known as the two-parameter Poisson–Dirichlet process. The model we consider is, in turn, a generalization of PYMs based on the family of Gibbs partitions (Pitman, 2006, Gnedin and Pitman, 2006).

### 3.2.1 Gibbs partitions

We will use  $p(\cdot)$  to denote probabilities and probability densities under the model. Our model specification begins with a distribution on partitions, or more precisely, on *ordered* partitions. Given  $n \in \{1, 2, \dots\}$  and  $t \in \{1, \dots, n\}$ , let  $\mathcal{A}_t(n)$  denote the set of all ordered partitions  $(A_1, \dots, A_t)$  of  $\{1, \dots, n\}$  into  $t$  nonempty sets (or

“parts”). In other words,

$$\mathcal{A}_t(n) = \left\{ (A_1, \dots, A_t) : A_1, \dots, A_t \text{ are disjoint, } \bigcup_{i=1}^t A_i = \{1, \dots, n\}, |A_i| \geq 1 \ \forall i \right\}.$$

For each  $n \in \{1, 2, \dots\}$ , consider a probability mass function (p.m.f.) on  $\bigcup_{t=1}^n \mathcal{A}_t(n)$  of the form

$$p(A) = v_n(t) \prod_{i=1}^t w_n(|A_i|) \tag{3.2.1}$$

for  $A \in \mathcal{A}_t(n)$ , where  $v_n : \{1, \dots, n\} \rightarrow [0, \infty)$  and  $w_n : \{1, \dots, n\} \rightarrow [0, \infty)$ . This induces a distribution on  $t$  in the natural way, via  $p(t | A) = I(A \in \mathcal{A}_t(n))$ . (Throughout, we use  $I$  to denote the indicator function:  $I(E)$  is 1 if  $E$  is true, and 0 otherwise.) It follows that  $p(A) = p(A, t)$  when  $A \in \mathcal{A}_t(n)$ .

Although it is more common to use a distribution on *unordered* partitions  $\{A_1, \dots, A_t\}$ , for our purposes it is more convenient to work with the corresponding distribution on ordered partitions  $(A_1, \dots, A_t)$  obtained by uniformly permuting the parts. This does not affect the distribution of  $t$ . Under this correspondence, any p.m.f. as in Equation 3.2.1 corresponds to a member of the class of “exchangeable partition probability functions”, or EPPFs (Pitman, 2006). In particular, for any given  $n$  it yields an EPPF in “Gibbs form”, and a random partition from such an EPPF is called a *Gibbs partition* (Pitman, 2006). (Note: We do not assume that, as  $n$  varies, the sequence of p.m.f.s in Equation 3.2.1 necessarily satisfies the marginalization property referred to as “consistency in distribution”.)

For example, to obtain the partition distribution for a Dirichlet process, we can

choose

$$v_n(t) = \frac{\vartheta^t}{\vartheta_{n\uparrow 1} t!} \quad \text{and} \quad w_n(a) = (a-1)! \quad (3.2.2)$$

where  $\vartheta > 0$  and  $x_{n\uparrow\delta} = x(x+\delta)(x+2\delta)\cdots(x+(n-1)\delta)$ , with  $x_{0\uparrow\delta} = 1$  by convention.

(The  $t!$  in the denominator appears since we are working with ordered partitions.)

More generally, to obtain the partition distribution for a Pitman–Yor process, we can choose

$$v_n(t) = \frac{(\vartheta + \sigma)_{t-1\uparrow\sigma}}{(\vartheta + 1)_{n-1\uparrow 1} t!} \quad \text{and} \quad w_n(a) = (1 - \sigma)_{a-1\uparrow 1} \quad (3.2.3)$$

where either  $\sigma \in [0, 1)$  and  $\vartheta \in (-\sigma, \infty)$ , or  $\sigma \in (-\infty, 0)$  and  $\vartheta = N|\sigma|$  for some  $N \in \{1, 2, \dots\}$  (Ishwaran and James, 2003). When  $\sigma = 0$ , this reduces to the partition distribution of a Dirichlet process. When  $\sigma < 0$  and  $\vartheta = N|\sigma|$ , it is the partition distribution obtained by drawing  $q = (q_1, \dots, q_N)$  from a symmetric  $N$ -dimensional Dirichlet with parameters  $|\sigma|, \dots, |\sigma|$ , sampling assignments  $Z_1, \dots, Z_n$  i.i.d. from  $q$ , and removing any empty parts (Gnedin and Pitman, 2006). Thus, in this latter case,  $t$  is always in  $\{1, \dots, N\}$ .

### 3.2.2 Gibbs partition mixtures

Consider the hierarchical model

$$\begin{aligned}
 p(A, t) &= p(A) = v_n(t) \prod_{i=1}^t w_n(|A_i|), \\
 p(\theta_{1:t} \mid A, t) &= \prod_{i=1}^t \pi(\theta_i), \\
 p(x_{1:n} \mid \theta_{1:t}, A, t) &= \prod_{i=1}^t \prod_{j \in A_i} p_{\theta_i}(x_j),
 \end{aligned} \tag{3.2.4}$$

where  $\pi$  is a prior density on component parameters  $\theta \in \Theta \subset \mathbb{R}^k$  for some  $k$ , and  $\{p_\theta : \theta \in \Theta\}$  is a parametrized family of densities on  $x \in \mathcal{X} \subset \mathbb{R}^d$  for some  $d$ . Here,  $x_{1:n} = (x_1, \dots, x_n)$  with  $x_i \in \mathcal{X}$ ,  $\theta_{1:t} = (\theta_1, \dots, \theta_t)$  with  $\theta_i \in \Theta$ , and  $A \in \mathcal{A}_t(n)$ . Assume that  $\pi$  is a density with respect to Lebesgue measure, and that  $\{p_\theta : \theta \in \Theta\}$  are densities with respect to some sigma-finite Borel measure  $\lambda$  on  $\mathcal{X}$ , such that  $(\theta, x) \mapsto p_\theta(x)$  is measurable. (Of course, the distribution of  $x$  under  $p_\theta(x)$  may be discrete, continuous, or neither, depending on the nature of  $\lambda$ .)

For  $x_1, \dots, x_n \in \mathcal{X}$  and  $J \subset \{1, \dots, n\}$ , define the *single-cluster marginal*,

$$m(x_J) = \int_{\Theta} \left( \prod_{j \in J} p_\theta(x_j) \right) \pi(\theta) d\theta, \tag{3.2.5}$$

where  $x_J = (x_j : j \in J)$ , and assume  $m(x_J) < \infty$ . By convention,  $m(x_J) = 1$  when  $J = \emptyset$ . Note that  $m(x_J)$  is a density with respect to the product measure  $\lambda^\ell$  on  $\mathcal{X}^\ell$ , where  $\ell = |J|$ , and that  $m(x_J)$  can (and often will) be positive outside the support of  $\lambda^\ell$ .

**Definition 3.2.1.** We refer to such a hierarchical model as a *Gibbs partition mixture model*.

(Note: This is, perhaps, a slight abuse of the term ‘‘Gibbs partition’’, since we allow  $v_n$  and  $w_n$  to vary arbitrarily with  $n$ .) In particular, it is a *Dirichlet process mixture model* when  $v_n$  and  $w_n$  are as in Equation 3.2.2, or more generally, a *Pitman–Yor process mixture model* when  $v_n$  and  $w_n$  are as in Equation 3.2.3.

We distinguish between the terms ‘‘component’’ and ‘‘cluster’’: a *component* of a mixture is one of the distributions used in it (e.g.  $p_{\theta_i}$ ), while a *cluster* is the set of indices of datapoints coming from a given component (e.g.  $A_i$ ). The prior on the number of clusters under such a model is  $p_n(t) = \sum_{A \in \mathcal{A}_t(n)} p(A)$ . We use  $T_n$ , rather than  $T$ , to denote the random variable representing the number of clusters, as a reminder that its distribution depends on  $n$ .

Since we are concerned with the posterior  $p(T_n = t \mid x_{1:n})$  on the number of clusters, we will be especially interested in the marginal density of  $(x_{1:n}, t)$ , given by

$$\begin{aligned}
 p(x_{1:n}, T_n = t) &= \sum_{A \in \mathcal{A}_t(n)} \int p(x_{1:n}, \theta_{1:t}, A, t) d\theta_{1:t} \\
 &= \sum_{A \in \mathcal{A}_t(n)} p(A) \prod_{i=1}^t \int \left( \prod_{j \in A_i} p_{\theta_i}(x_j) \right) \pi(\theta_i) d\theta_i \\
 &= \sum_{A \in \mathcal{A}_t(n)} p(A) \prod_{i=1}^t m(x_{A_i}). \tag{3.2.6}
 \end{aligned}$$

As usual, the posterior  $p(T_n = t \mid x_{1:n})$  is not uniquely defined, since it can be modified arbitrarily on any subset of  $\mathcal{X}^n$  having probability zero under the model distribution. For definiteness, we will employ the usual version of this posterior,

$$p(T_n = t \mid x_{1:n}) = \frac{p(x_{1:n}, T_n = t)}{p(x_{1:n})} = \frac{p(x_{1:n}, T_n = t)}{\sum_{t'=1}^{\infty} p(x_{1:n}, T_n = t')}$$

whenever the denominator is nonzero, and  $p(T_n = t \mid x_{1:n}) = 0$  otherwise (for

notational convenience).

### 3.3 Inconsistency theorem

The essential ingredients in the main theorem are Conditions 3.3.1 and 3.3.2 below.

For each  $n \in \{1, 2, \dots\}$ , consider a partition distribution as in Equation 3.2.1. For  $n > t \geq 1$ , define

$$c_{w_n} = \max_{a \in \{2, \dots, n\}} \frac{w_n(a)}{a w_n(a-1) w_n(1)} \quad \text{and} \quad c_{v_n}(t) = \frac{v_n(t)}{v_n(t+1)},$$

with the convention that  $0/0 = 0$  and  $y/0 = \infty$  for  $y > 0$ .

**Condition 3.3.1.** *Assume  $\limsup_{n \rightarrow \infty} c_{w_n} < \infty$  and  $\limsup_{n \rightarrow \infty} c_{v_n}(t) < \infty$ , given some particular  $t \in \{1, 2, \dots\}$ .*

For Pitman–Yor processes, Condition 3.3.1 holds for all relevant values of  $t$ , by Proposition 3.3.3 below. Now, consider a collection of single-cluster marginals  $m(\cdot)$  as in Equation 3.2.5. Given  $n \geq t \geq 1$ ,  $x_1, \dots, x_n \in \mathcal{X}$ , and  $c \in [0, \infty)$ , define

$$\varphi_t(x_{1:n}, c) = \min_{A \in \mathcal{A}_t(n)} \frac{1}{n} |S_A(x_{1:n}, c)|$$

where  $S_A(x_{1:n}, c)$  is the set of indices  $j \in \{1, \dots, n\}$  such that the part  $A_\ell$  containing  $j$  satisfies  $m(x_{A_\ell}) \leq c m(x_{A_\ell \setminus j}) m(x_j)$ .

**Condition 3.3.2.** *Given a sequence of random variables  $X_1, X_2, \dots \in \mathcal{X}$ , a collection of single-cluster marginals  $m(\cdot)$ , and  $t \in \{1, 2, \dots\}$ , assume*

$$\sup_{c \in [0, \infty)} \liminf_{n \rightarrow \infty} \varphi_t(X_{1:n}, c) > 0 \quad \text{with probability 1.}$$



Note that Condition 3.3.1 involves only the partition distributions, while Condition 3.3.2 involves only the data distribution and the single-cluster marginals.

**Proposition 3.3.3.** *Consider a Pitman–Yor process. If  $\sigma \in [0, 1)$  and  $\vartheta \in (-\sigma, \infty)$  then Condition 3.3.1 holds for any  $t \in \{1, 2, \dots\}$ . If  $\sigma \in (-\infty, 0)$  and  $\vartheta = N|\sigma|$ , then it holds for any  $t \in \{1, 2, \dots\}$  except  $N$ .*

*Proof.* This is a simple calculation. See Section 3.9. □

**Theorem 3.3.4.** *Let  $X_1, X_2, \dots \in \mathcal{X}$  be a sequence of random variables (not necessarily i.i.d.). Consider a Gibbs partition mixture model. For any  $t \in \{1, 2, \dots\}$ , if Conditions 3.3.1 and 3.3.2 hold, then*

$$\limsup_{n \rightarrow \infty} p(T_n = t \mid X_{1:n}) < 1 \text{ with probability 1.}$$

*If, further, the sequence  $X_1, X_2, \dots$  is i.i.d. from a mixture with  $t$  components, then with probability 1 the posterior of  $T_n$  (under the model) is not consistent for  $t$ .*

*Proof.* This follows easily from Lemma 3.3.5 below. See Section 3.9. □

**Lemma 3.3.5.** *Consider a Gibbs partition mixture model. Let  $n > t \geq 1$ ,  $x_1, \dots, x_n \in \mathcal{X}$ , and  $c \in [0, \infty)$ . If  $\varphi_t(x_{1:n}, c) > t/n$ ,  $c_{w_n} < \infty$ , and  $c_{v_n}(t) < \infty$ , then*

$$p(T_n = t \mid x_{1:n}) \leq \frac{C_t(x_{1:n}, c)}{1 + C_t(x_{1:n}, c)},$$

where  $C_t(x_{1:n}, c) = t c c_{w_n} c_{v_n}(t) / (\varphi_t(x_{1:n}, c) - t/n)$ .

*Proof.* To simplify notation, let us denote  $\varphi = \varphi_t(x_{1:n}, c)$ ,  $C = C_t(x_{1:n}, c)$ , and  $S_A = S_A(x_{1:n}, c)$  for  $A \in \mathcal{A}_t(n)$ . Given  $J \subset \{1, \dots, n\}$  such that  $|J| \geq 1$ , define

$$h_J = w_n(|J|) m(x_J).$$

For  $A \in \mathcal{A}_t(n)$ , let  $R_A = S_A \setminus \left(\bigcup_{i:|A_i|=1} A_i\right)$ , that is,  $R_A$  consists of those  $j \in S_A$  such that the size of the part  $A_\ell$  containing  $j$  is greater than 1. Note that

$$|R_A| \geq |S_A| - t \geq n\varphi - t > 0. \quad (3.3.1)$$

For any  $j \in R_A$ , the part  $A_\ell$  containing  $j$  satisfies

$$\begin{aligned} h_{A_\ell} &= w_n(|A_\ell|) m(x_{A_\ell}) \\ &\leq c_{w_n} |A_\ell| w_n(|A_\ell| - 1) w_n(1) c m(x_{A_\ell \setminus j}) m(x_j) \\ &\leq n c c_{w_n} h_{A_\ell \setminus j} h_j. \end{aligned} \quad (3.3.2)$$

Given  $j \in R_A$ , define  $B(A, j)$  to be the element  $B$  of  $\mathcal{A}_{t+1}(n)$  such that  $B_i = A_i \setminus j$  for  $i = 1, \dots, t$ , and  $B_{t+1} = \{j\}$  (that is, remove  $j$  from whatever part it belongs to, and make  $\{j\}$  the  $(t+1)^{\text{th}}$  part). Define

$$\mathcal{Y}_A = \{B(A, j) : j \in R_A\}.$$

Now, using Equations 3.3.1 and 3.3.2, for any  $A \in \mathcal{A}_t(n)$  we have

$$\begin{aligned} \prod_{i=1}^t h_{A_i} &= \frac{1}{|R_A|} \sum_{\ell=1}^t \sum_{j \in R_A \cap A_\ell} h_{A_\ell} \prod_{i \neq \ell} h_{A_i} \\ &\leq \frac{1}{n\varphi - t} \sum_{\ell=1}^t \sum_{j \in R_A \cap A_\ell} n c c_{w_n} h_{A_\ell \setminus j} h_j \prod_{i \neq \ell} h_{A_i} \\ &= \frac{c c_{w_n}}{\varphi - t/n} \sum_{j \in R_A} \prod_{i=1}^{t+1} h_{B_i(A, j)} \\ &= \frac{c c_{w_n}}{\varphi - t/n} \sum_{B \in \mathcal{A}_{t+1}(n)} \left[ \prod_{i=1}^{t+1} h_{B_i} \right] I(B \in \mathcal{Y}_A). \end{aligned} \quad (3.3.3)$$

For any  $B \in \mathcal{A}_{t+1}(n)$ ,

$$\#\{A \in \mathcal{A}_t(n) : B \in \mathcal{Y}_A\} \leq t, \quad (3.3.4)$$

since there are only  $t$  parts that  $B_{t+1}$  could have come from. Therefore,

$$\begin{aligned} p(x_{1:n}, T_n = t) &\stackrel{(a)}{=} \sum_{A \in \mathcal{A}_t(n)} p(A) \prod_{i=1}^t m(x_{A_i}) \\ &\stackrel{(b)}{=} \sum_{A \in \mathcal{A}_t(n)} v_n(t) \prod_{i=1}^t h_{A_i} \\ &\stackrel{(c)}{\leq} \frac{c c_{w_n}}{\varphi - t/n} v_n(t) \sum_{A \in \mathcal{A}_t(n)} \sum_{B \in \mathcal{A}_{t+1}(n)} \left[ \prod_{i=1}^{t+1} h_{B_i} \right] I(B \in \mathcal{Y}_A) \\ &= \frac{c c_{w_n}}{\varphi - t/n} v_n(t) \sum_{B \in \mathcal{A}_{t+1}(n)} \left[ \prod_{i=1}^{t+1} h_{B_i} \right] \#\{A \in \mathcal{A}_t(n) : B \in \mathcal{Y}_A\} \\ &\stackrel{(d)}{\leq} \frac{c c_{w_n} c_{v_n}(t)}{\varphi - t/n} v_n(t+1) \sum_{B \in \mathcal{A}_{t+1}(n)} \left[ \prod_{i=1}^{t+1} h_{B_i} \right] t \\ &= \frac{t c c_{w_n} c_{v_n}(t)}{\varphi - t/n} \sum_{B \in \mathcal{A}_{t+1}(n)} p(B) \prod_{i=1}^{t+1} m(x_{B_i}) \\ &= C p(x_{1:n}, T_n = t+1), \end{aligned}$$

where (a) is from Equation 3.2.6, (b) is from Equation 3.2.4 and the definition of  $h_j$  above, (c) follows from Equation 3.3.3, and (d) follows from Equation 3.3.4.

If  $p(T_n = t \mid x_{1:n}) = 0$ , then trivially  $p(T_n = t \mid x_{1:n}) \leq C/(C+1)$ . On the other hand, if  $p(T_n = t \mid x_{1:n}) > 0$ , then  $p(x_{1:n}, T_n = t) > 0$ , and therefore

$$\begin{aligned} p(T_n = t \mid x_{1:n}) &= \frac{p(x_{1:n}, T_n = t)}{\sum_{t'=1}^{\infty} p(x_{1:n}, T_n = t')} \\ &\leq \frac{p(x_{1:n}, T_n = t)}{p(x_{1:n}, T_n = t) + p(x_{1:n}, T_n = t+1)} \leq \frac{C}{C+1}. \quad \square \end{aligned}$$

### 3.4 Application to discrete or bounded cases

By Theorem 3.3.4, the following result implies inconsistency in a large class of PYM models, including essentially all discrete cases (or more generally anything with at least one point mass) and a number of continuous cases as well.

**Theorem 3.4.1.** *Consider a family of densities  $\{p_\theta : \theta \in \Theta\}$  on  $\mathcal{X}$  along with a prior  $\pi$  on  $\Theta$  and the resulting collection of single-cluster marginals  $m(\cdot)$  as in Equation 3.2.5. Let  $X_1, X_2, \dots \in \mathcal{X}$  be a sequence of random variables (not necessarily i.i.d.). If there exists  $U \subset \mathcal{X}$  such that*

- (1)  $\liminf_{n \rightarrow \infty} \frac{1}{n} \sum_{j=1}^n I(X_j \in U) > 0$  with probability 1, and
- (2)  $\sup \left\{ \frac{p_\theta(x)}{m(x)} : x \in U, \theta \in \Theta \right\} < \infty$  (where  $0/0 = 0$ ,  $y/0 = \infty$  for  $y > 0$ ),

then Condition 3.3.2 holds for all  $t \in \{1, 2, \dots\}$ .

*Proof.* Suppose  $U \subset \mathcal{X}$  satisfies (1) and (2), and let  $t \in \{1, 2, \dots\}$ . Define  $c = \sup \left\{ \frac{p_\theta(x)}{m(x)} : x \in U, \theta \in \Theta \right\}$ . Let  $n > t$  and  $x_1, \dots, x_n \in \mathcal{X}$ . Now, for any  $x \in U$  and  $\theta \in \Theta$ , we have  $p_\theta(x) \leq c m(x)$ . Hence, for any  $J \subset \{1, \dots, n\}$ , if  $j \in J$  and  $x_j \in U$  then

$$m(x_J) = \int_{\Theta} p_\theta(x_j) \left[ \prod_{i \in J \setminus j} p_\theta(x_i) \right] \pi(\theta) d\theta \leq c m(x_j) m(x_{J \setminus j}). \quad (3.4.1)$$

Thus, letting  $R(x_{1:n}) = \{j \in \{1, \dots, n\} : x_j \in U\}$ , we have  $R(x_{1:n}) \subset S_A(x_{1:n}, c)$  for any  $A \in \mathcal{A}_t(n)$ , and hence,  $\varphi_t(x_{1:n}, c) \geq \frac{1}{n} |R(x_{1:n})|$ .

Therefore, by (1), with probability 1,

$$\liminf_{n \rightarrow \infty} \varphi_t(X_{1:n}, c) \geq \liminf_{n \rightarrow \infty} \frac{1}{n} |R(X_{1:n})| > 0. \quad \square$$

The preceding theorem covers a fairly wide range of cases; here are some examples. Consider a model with  $\{p_\theta\}$ ,  $\pi$ ,  $\lambda$ , and  $m(\cdot)$ , as in Section 3.2.

- (i) **Finite sample space.** Suppose  $\mathcal{X}$  is a finite set,  $\lambda$  is counting measure, and  $m(x) > 0$  for all  $x \in \mathcal{X}$ . Then choosing  $U = \mathcal{X}$ , conditions (1) and (2) of Theorem 3.4.1 are trivially satisfied, regardless of the distribution of  $X_1, X_2, \dots$  (Note that when  $\lambda$  is counting measure,  $p_\theta(x)$  and  $m(x)$  are p.m.f.s on  $\mathcal{X}$ .) It is often easy to check that  $m(x) > 0$  by using the fact that this is true whenever  $\{\theta \in \Theta : p_\theta(x) > 0\}$  has nonzero probability under  $\pi$ . This case covers, for instance, Multinomials (including Binomials), and the population genetics model from Section 3.1.1.

We should mention a subtle point here: when  $\mathcal{X}$  is finite, mixture identifiability might only hold up to a certain maximum number of components (e.g., Teicher (1963), Proposition 4, showed this for Binomials), making consistency impossible in general — however, consistency might still be possible within that identifiable range. Regardless, our result shows that PYMs are not consistent anyway.

Now, suppose  $P$  is a probability measure on  $\mathcal{X}$ , and  $X_1, X_2, \dots \stackrel{\text{iid}}{\sim} P$ . Let us abuse notation and write  $P(x) = P(\{x\})$  and  $\lambda(x) = \lambda(\{x\})$  for  $x \in \mathcal{X}$ .

- (ii) **One or more point masses in common.** If there exists  $x_0 \in \mathcal{X}$  such that  $P(x_0) > 0$ ,  $\lambda(x_0) > 0$ , and  $m(x_0) > 0$ , then it is easy to verify that

conditions (1) and (2) are satisfied with  $U = \{x_0\}$ . (Note that  $\lambda(x_0) > 0$  implies  $p_\theta(x_0) \leq 1/\lambda(x_0)$  for any  $\theta \in \Theta$ .)

(iii) **Discrete families.** Case (ii) essentially covers all discrete families — e.g., Poisson, Geometric, Negative Binomial, or any power-series distribution (see [Sapatinas \(1995\)](#) for mixture identifiability of these) — provided that the data is i.i.d.. For, suppose  $\mathcal{X}$  is a countable set and  $\lambda$  is counting measure. By case (ii), the theorem applies if there is any  $x_0 \in \mathcal{X}$  such that  $m(x_0) > 0$  and  $P(x_0) > 0$ . If this is not so, the model is extremely misspecified, since then the model distribution and the data distribution are mutually singular.

(iv) **Continuous densities bounded on some non-null compact set.** Suppose there exists  $c \in (0, \infty)$  and  $U \subset \mathcal{X}$  compact such that

- (a)  $P(U) > 0$ ,
- (b)  $x \mapsto p_\theta(x)$  is continuous on  $U$  for all  $\theta \in \Theta$ , and
- (c)  $p_\theta(x) \in (0, c]$  for all  $x \in U, \theta \in \Theta$ .

Then condition (1) is satisfied due to item (a), and condition (2) follows easily from (b) and (c) since  $m(x)$  is continuous (by the dominated convergence theorem) and positive on the compact set  $U$ , so  $\inf_{x \in U} m(x) > 0$ . This case covers, for example, the following families (with any  $P$ ):

- (a) Exponential( $\theta$ ),  $\mathcal{X} = (0, \infty)$ ,
- (b) Gamma( $a, b$ ),  $\mathcal{X} = (0, \infty)$ , with variance  $a/b^2$  bounded away from zero,
- (c) Normal( $\mu, \Sigma$ ),  $\mathcal{X} = \mathbb{R}^d$ , (multivariate Gaussian) with  $\det(\Sigma)$  bounded away from zero, and
- (d) many location–scale families with scale bounded away from zero (for instance, Laplace( $\mu, \sigma$ ) or Cauchy( $\mu, \sigma$ ), with  $\sigma \geq \varepsilon > 0$ ).

The examples listed in item (iv) are indicative of a deficiency in Theorem 3.4.1: condition (2) is not satisfied in some important cases, such as multivariate Gaussians with unrestricted covariance. Showing that Condition 3.3.2 still holds, for many exponential families at least, is the objective of the remainder of the chapter.

## 3.5 Exponential families and conjugate priors

### 3.5.1 Exponential families

In this section, we make the usual definitions for exponential families and state the regularity conditions to be assumed. Consider an exponential family of the following form. Fix a sigma-finite Borel measure  $\lambda$  on  $\mathcal{X} \subset \mathbb{R}^d$  such that  $\lambda(\mathcal{X}) \neq 0$ , let  $s : \mathcal{X} \rightarrow \mathbb{R}^k$  be Borel measurable, and for  $\theta \in \Theta \subset \mathbb{R}^k$ , define a density  $p_\theta$  with respect to  $\lambda$  by setting

$$p_\theta(x) = \exp(\theta^\top s(x) - \kappa(\theta))$$

where

$$\kappa(\theta) = \log \int_{\mathcal{X}} \exp(\theta^\top s(x)) d\lambda(x).$$

Let  $P_\theta$  be the probability measure on  $\mathcal{X}$  corresponding to  $p_\theta$ , that is,  $P_\theta(E) = \int_E p_\theta(x) d\lambda(x)$  for  $E \subset \mathcal{X}$  measurable. Any exponential family on  $\mathbb{R}^d$  can be written in the form above by reparametrizing if necessary, and choosing  $\lambda$  appropriately. We will assume the following (very mild) regularity conditions.

**Conditions 3.5.1.** *Assume the family  $\{P_\theta : \theta \in \Theta\}$  is:*

- (1) *full, that is,  $\Theta = \{\theta \in \mathbb{R}^k : \kappa(\theta) < \infty\}$ ,*

- (2) *nonempty, that is,  $\Theta \neq \emptyset$ ,*
- (3) *regular, that is,  $\Theta$  is an open subset of  $\mathbb{R}^k$ , and*
- (4) *identifiable, that is, if  $\theta \neq \theta'$  then  $P_\theta \neq P_{\theta'}$ .*

Most commonly-used exponential families satisfy Conditions 3.5.1, including multivariate Gaussian, Gamma, Poisson, Exponential, Geometric, and others. (A notable exception is the Inverse Gaussian, for which  $\Theta$  is not open.) Let  $\mathcal{M}$  denote the *moment space*, that is,

$$\mathcal{M} = \{\mathbb{E}_\theta s(X) : \theta \in \Theta\}$$

where  $\mathbb{E}_\theta$  denotes expectation under  $P_\theta$ . Finiteness of these expectations is guaranteed, thus  $\mathcal{M} \subset \mathbb{R}^k$ ; see Section 3.10 for this and other well-known properties that we will use.

### 3.5.2 Conjugate priors

Given an exponential family  $\{P_\theta\}$  as above, let

$$\Xi = \left\{ (\xi, \nu) : \xi \in \mathbb{R}^k, \nu > 0 \text{ s.t. } \xi/\nu \in \mathcal{M} \right\},$$

and consider the family  $\{\pi_{\xi, \nu} : (\xi, \nu) \in \Xi\}$  where

$$\pi_{\xi, \nu}(\theta) = \exp(\xi^\top \theta - \nu \kappa(\theta) - \psi(\xi, \nu)) I(\theta \in \Theta)$$

is a density with respect to Lebesgue measure on  $\mathbb{R}^k$ . Here,

$$\psi(\xi, \nu) = \log \int_{\Theta} \exp(\xi^\top \theta - \nu \kappa(\theta)) d\theta.$$



In Section 3.10, we note a few basic properties of this family — in particular, it is a conjugate prior for  $\{P_\theta\}$ .

**Definition 3.5.2.** We will say that an exponential family with conjugate prior is *well-behaved* if it takes the form above, satisfies Conditions 3.5.1, and has  $(\xi, \nu) \in \Xi$ .

## 3.6 Application to exponential families

In this section, we apply Theorem 3.3.4 to prove that in many cases, a PYM model using a well-behaved exponential family with conjugate prior will exhibit inconsistency for the number of components.

**Conditions 3.6.1.** Consider an exponential family with sufficient statistics function  $s : \mathcal{X} \rightarrow \mathbb{R}^k$  and moment space  $\mathcal{M}$ . Given a probability measure  $P$  on  $\mathcal{X}$ , let  $X \sim P$  and assume:

- (1)  $\mathbb{E}|s(X)| < \infty$ ,
- (2)  $\mathbb{P}(s(X) \in \bar{\mathcal{M}}) = 1$ , and
- (3)  $\mathbb{P}(s(X) \in L) = 0$  for any hyperplane  $L$  that does not intersect  $\mathcal{M}$ .

Throughout, we use  $|\cdot|$  to denote the Euclidean norm. Here, a *hyperplane* refers to a set  $L = \{x \in \mathbb{R}^k : x^T y = b\}$  for some  $y \in \mathbb{R}^k \setminus \{0\}$ ,  $b \in \mathbb{R}$ . In Theorem 3.6.2 below, it is assumed that the data comes from a distribution  $P$  satisfying Conditions 3.6.1. In Proposition 3.6.3, we give some simple conditions under which, if  $P$  is a finite mixture from the exponential family under consideration, then Conditions 3.6.1 hold.

The following theorem follows almost immediately from Lemma 3.8.4, the proof of which will occupy most of the remainder of the chapter.

**Theorem 3.6.2.** *Consider a well-behaved exponential family with conjugate prior (as in Definition 3.5.2), along with the resulting collection of single-cluster marginals  $m(\cdot)$ . Let  $P$  be a probability measure on  $\mathcal{X}$  satisfying Conditions 3.6.1 (for the  $s$  and  $\mathcal{M}$  from the exponential family under consideration), and let  $X_1, X_2, \dots \stackrel{\text{iid}}{\sim} P$ . Then Condition 3.3.2 holds for any  $t \in \{1, 2, \dots\}$ .*

*Proof.* Let  $t \in \{1, 2, \dots\}$  and choose  $c$  according to Lemma 3.8.4 with  $\beta = 1/t$ . We will show that for any  $n > t$ , if the event of Lemma 3.8.4 holds, then  $\varphi_t(X_{1:n}, c) \geq 1/(2t)$ . Since with probability 1, this event holds for all  $n$  sufficiently large, it will follow that with probability 1,  $\liminf_n \varphi_t(X_{1:n}, c) \geq 1/(2t) > 0$ .

So, let  $n > t$  and  $x_1, \dots, x_n \in \mathcal{X}$ , and assume the event of Lemma 3.8.4 holds. Let  $A \in \mathcal{A}_t(n)$ . There is at least one part  $A_\ell$  such that  $|A_\ell| \geq n/t = \beta n$ . Then, by assumption there exists  $R_A \subset A_\ell$  such that  $|R_A| \geq \frac{1}{2}|A_\ell|$  and for any  $j \in R_A$ ,  $m(x_{A_\ell}) \leq c m(x_{A_\ell \setminus j}) m(x_j)$ . Thus,  $R_A \subset S_A(x_{1:n}, c)$ , hence  $|S_A(x_{1:n}, c)| \geq |R_A| \geq \frac{1}{2}|A_\ell| \geq n/(2t)$ . Since  $A \in \mathcal{A}_t(n)$  was arbitrary,  $\varphi_t(x_{1:n}, c) \geq 1/(2t)$ .  $\square$

This theorem implies inconsistency in several important cases. In particular, it can be verified that each of the following is well-behaved (when put in canonical form and given the conjugate prior in Section 3.5.2) and, using Proposition 3.6.3 below, that if  $P$  is a finite mixture from the same family then  $P$  satisfies Conditions 3.6.1:

- (a) Normal( $\mu, \Sigma$ ) (multivariate Gaussian),
- (b) Exponential( $\theta$ ),

- (c) Gamma( $a, b$ ),
- (d) Log-Normal( $\mu, \sigma^2$ ), and
- (e) Weibull( $a, b$ ) with fixed shape  $a > 0$ .

Combined with the cases covered by Theorem 3.4.1, these results are fairly comprehensive.

**Proposition 3.6.3.** *Consider an exponential family  $\{P_\theta : \theta \in \Theta\}$  satisfying Conditions 3.5.1. If  $X \sim P = \sum_{i=1}^t \pi_i P_{\theta(i)}$  for some  $\theta(1), \dots, \theta(t) \in \Theta$  and some  $\pi_1, \dots, \pi_t \geq 0$  such that  $\sum_{i=1}^t \pi_i = 1$ , then*

- (1)  $\mathbb{E}|s(X)| < \infty$ , and
- (2)  $\mathbb{P}(s(X) \in \bar{\mathcal{M}}) = 1$ .

*If, further, the exponential family is continuous (that is, the underlying measure  $\lambda$  is absolutely continuous with respect to Lebesgue measure on  $\mathcal{X}$ ),  $\mathcal{X} \subset \mathbb{R}^d$  is open and connected, and the sufficient statistics function  $s : \mathcal{X} \rightarrow \mathbb{R}^k$  is real analytic (that is, each coordinate function  $s_1, \dots, s_k$  is real analytic), then*

- (3)  $\mathbb{P}(s(X) \in L) = 0$  for any hyperplane  $L \subset \mathbb{R}^k$ .

*Proof.* See Section 3.9. □

Sometimes, Condition 3.6.1(3) will be satisfied even when Proposition 3.6.3 is not applicable. In any particular case, it may be a simple matter to check this condition by using the characterization of  $\mathcal{M}$  as the interior of the closed convex hull of  $\text{support}(\lambda s^{-1})$  (see Proposition 3.10.1(8) in Section 3.10).

### 3.7 Marginal inequalities

Consider a well-behaved exponential family with conjugate prior (as in Definition 3.5.2). In this section, we use some simple bounds on the Laplace approximation (see Section 3.11) to prove certain inequalities involving the marginal density (from Equation 3.2.5),

$$m(x_{1:n}) = \int_{\Theta} \left( \prod_{j=1}^n p_{\theta}(x_j) \right) \pi_{\xi, \nu}(\theta) d\theta$$

of  $x_{1:n} = (x_1, \dots, x_n)$ , where  $x_j \in \mathcal{X}$ . Of course, it is commonplace to apply the Laplace approximation to  $m(X_{1:n})$  when  $X_1, \dots, X_n$  are i.i.d. random variables. In contrast, our application of it is considerably more subtle. For our purposes, it is necessary to show that the approximation is good not only in the i.i.d. case, but in fact whenever the sufficient statistics are not too extreme.

We make extensive use of the exponential family properties in Section 3.10, often without mention. We use  $f'$  to denote the gradient and  $f''$  to denote the Hessian of a (sufficiently smooth) function  $f : \mathbb{R}^k \rightarrow \mathbb{R}$ . For  $\mu \in \mathcal{M}$ , define

$$\begin{aligned} f_{\mu}(\theta) &= \theta^{\top} \mu - \kappa(\theta), \\ \mathcal{L}(\mu) &= \sup_{\theta \in \Theta} (\theta^{\top} \mu - \kappa(\theta)), \\ \theta_{\mu} &= \operatorname{argmax}_{\theta \in \Theta} (\theta^{\top} \mu - \kappa(\theta)), \end{aligned}$$

and note that  $\theta_{\mu} = \kappa'^{-1}(\mu)$  (Proposition 3.10.1).  $\mathcal{L}$  is known as the Legendre transform of  $\kappa$ . Note that  $\mathcal{L}(\mu) = f_{\mu}(\theta_{\mu})$ , and  $\mathcal{L}$  is  $C^{\infty}$  smooth on  $\mathcal{M}$  (since  $\mathcal{L}(\mu) = \theta_{\mu}^{\top} \mu - \kappa(\theta_{\mu})$ ,  $\theta_{\mu} = \kappa'^{-1}(\mu)$ , and both  $\kappa$  and  $\kappa'^{-1}$  are  $C^{\infty}$  smooth). De-

fine

$$\mu_{x_{1:n}} = \frac{\xi + \sum_{j=1}^n s(x_j)}{\nu + n} \quad (3.7.1)$$

(cf. Equation 3.10.1), and given  $x_{1:n}$  such that  $\mu_{x_{1:n}} \in \mathcal{M}$ , define

$$\tilde{m}(x_{1:n}) = (\nu + n)^{-k/2} \exp((\nu + n) \mathcal{L}(\mu_{x_{1:n}})),$$

where  $k$  is the dimension of the sufficient statistics function  $s : \mathcal{X} \rightarrow \mathbb{R}^k$ . The first of the two results of this section provides uniform bounds on  $m(x_{1:n})/\tilde{m}(x_{1:n})$ . Here,  $\tilde{m}(x_{1:n})$  is only intended to approximate  $m(x_{1:n})$  up to a multiplicative constant — a better approximation could always be obtained via the usual asymptotic form of the Laplace approximation.

**Proposition 3.7.1.** *Consider a well-behaved exponential family with conjugate prior. For any  $U \subset \mathcal{M}$  compact, there exist  $C_1, C_2 \in (0, \infty)$  such that for any  $n \in \{1, 2, \dots\}$  and any  $x_1, \dots, x_n \in \mathcal{X}$  satisfying  $\mu_{x_{1:n}} \in U$ , we have*

$$C_1 \leq \frac{m(x_{1:n})}{\tilde{m}(x_{1:n})} \leq C_2.$$

*Proof.* Assume  $U \neq \emptyset$ , since otherwise the result is trivial. Let

$$V = \kappa'^{-1}(U) = \{\theta_\mu : \mu \in U\}.$$

It is straightforward to show that there exists  $\varepsilon \in (0, 1)$  such that  $V_\varepsilon \subset \Theta$  where

$$V_\varepsilon = \{\theta \in \mathbb{R}^k : d(\theta, V) \leq \varepsilon\}.$$

(Here,  $d(\theta, V) = \inf_{\theta' \in V} |\theta - \theta'|$ .) Note that  $V_\varepsilon$  is compact, since  $\kappa'^{-1}$  is continuous.

Given a symmetric matrix  $A$ , define  $\lambda_*(A)$  and  $\lambda^*(A)$  to be the minimal and maximal eigenvalues, respectively, and recall that  $\lambda_*, \lambda^*$  are continuous functions of the entries of  $A$ . Letting

$$\alpha = \min_{\theta \in V_\varepsilon} \lambda_*(\kappa''(\theta)) \quad \text{and} \quad \beta = \max_{\theta \in V_\varepsilon} \lambda^*(\kappa''(\theta)),$$

we have  $0 < \alpha \leq \beta < \infty$  since  $V_\varepsilon$  is compact and  $\lambda_*(\kappa''(\cdot)), \lambda^*(\kappa''(\cdot))$  are continuous and positive on  $\Theta$ . Letting

$$\gamma = \sup_{\mu \in U} e^{-f_\mu(\theta_\mu)} \int_{\Theta} \exp(f_\mu(\theta)) d\theta = \sup_{\mu \in U} e^{-\mathcal{L}(\mu)} e^{\psi(\mu, 1)}$$

we have  $0 < \gamma < \infty$  since  $U$  is compact, and both  $\mathcal{L}$  (as noted above) and  $\psi(\mu, 1)$  (by Proposition 3.10.2) are continuous on  $\mathcal{M}$ . Define

$$h(\mu, \theta) = f_\mu(\theta_\mu) - f_\mu(\theta) = \mathcal{L}(\mu) - \theta^\top \mu + \kappa(\theta)$$

for  $\mu \in \mathcal{M}, \theta \in \Theta$ . For any  $\mu \in \mathcal{M}$ , we have that  $h(\mu, \theta) > 0$  whenever  $\theta \in \Theta \setminus \{\theta_\mu\}$ , and that  $h(\mu, \theta)$  is strictly convex in  $\theta$ . Letting  $B_\varepsilon(\theta_\mu) = \{\theta \in \mathbb{R}^k : |\theta - \theta_\mu| \leq \varepsilon\}$ , it follows that

$$\delta := \inf_{\mu \in U} \inf_{\theta \in \Theta \setminus B_\varepsilon(\theta_\mu)} h(\mu, \theta) = \inf_{\mu \in U} \inf_{u \in \mathbb{R}^k : |u|=1} h(\mu, \theta_\mu + \varepsilon u)$$

is positive, as the minimum of a positive continuous function on a compact set.

Now, applying the Laplace approximation bounds in Corollary 3.11.2 with  $\alpha, \beta, \gamma, \delta, \varepsilon$  as just defined, we obtain  $c_1, c_2 \in (0, \infty)$  such that for any  $\mu \in U$  we have (taking  $E = \Theta, f = -f_\mu, x_0 = \theta_\mu, A = \alpha I_{k \times k}, B = \beta I_{k \times k}$ )

$$c_1 \leq \frac{\int_{\Theta} \exp(tf_\mu(\theta)) d\theta}{t^{-k/2} \exp(tf_\mu(\theta_\mu))} \leq c_2$$

for any  $t \geq 1$ . We prove the result with  $C_i = c_i e^{-\psi(\xi, \nu)}$  for  $i = 1, 2$ .

Let  $n \in \{1, 2, \dots\}$  and  $x_1, \dots, x_n \in \mathcal{X}$  such that  $\mu_{x_{1:n}} \in U$ . Choose  $t = \nu + n$ . By integrating Equation 3.10.1, we have

$$m(x_{1:n}) = e^{-\psi(\xi, \nu)} \int_{\Theta} \exp(t f_{\mu_{x_{1:n}}}(\theta)) d\theta,$$

and meanwhile,

$$\tilde{m}(x_{1:n}) = t^{-k/2} \exp(t f_{\mu_{x_{1:n}}}(\theta_{\mu_{x_{1:n}}})) .$$

Thus, combining the preceding three displayed equations,

$$0 < C_1 = c_1 e^{-\psi(\xi, \nu)} \leq \frac{m(x_{1:n})}{\tilde{m}(x_{1:n})} \leq c_2 e^{-\psi(\xi, \nu)} = C_2 < \infty. \quad \square$$

The second result of this section is an inequality involving a product of marginals.

**Proposition 3.7.2** (Splitting inequality). *Consider a well-behaved exponential family with conjugate prior. For any  $U \subset \mathcal{M}$  compact there exists  $C \in (0, \infty)$  such that we have the following:*

*For any  $n \in \{1, 2, \dots\}$ , if  $A \subset \{1, \dots, n\}$  and  $B = \{1, \dots, n\} \setminus A$  are nonempty, and  $x_1, \dots, x_n \in \mathcal{X}$  satisfy  $\frac{1}{|A|} \sum_{j \in A} s(x_j) \in U$  and  $\mu_{x_B} \in U$ , then*

$$\frac{m(x_{1:n})}{m(x_A)m(x_B)} \leq C \left( \frac{ab}{\nu + n} \right)^{k/2}$$

*where  $a = \nu + |A|$  and  $b = \nu + |B|$ .*

*Proof.* Let  $U'$  be the convex hull of  $U \cup \{\xi/\nu\}$ . Then  $U'$  is compact (as the convex hull of a compact set in  $\mathbb{R}^k$ ) and  $U' \subset \mathcal{M}$  (since  $U \cup \{\xi/\nu\} \subset \mathcal{M}$  and  $\mathcal{M}$  is convex).

We show that the result holds with  $C = C_2 \exp(C_0)/C_1^2$ , where  $C_1, C_2 \in (0, \infty)$  are obtained by applying Proposition 3.7.1 to  $U'$ , and

$$C_0 = \nu \sup_{y \in U'} |(\xi/\nu - y)^\top \mathcal{L}'(y)| + \nu \sup_{y \in U'} |\mathcal{L}(y)| < \infty. \quad (3.7.2)$$

Since  $\mathcal{L}$  is convex (being a Legendre transform) and smooth, then for any  $y, z \in \mathcal{M}$  we have

$$\inf_{\rho \in (0,1)} \frac{1}{\rho} (\mathcal{L}(y + \rho(z - y)) - \mathcal{L}(y)) = (z - y)^\top \mathcal{L}'(y)$$

(by e.g. Rockafellar (1970) 23.1) and therefore for any  $\rho \in (0, 1)$ ,

$$\mathcal{L}(y) \leq \mathcal{L}((1 - \rho)y + \rho z) - \rho(z - y)^\top \mathcal{L}'(y). \quad (3.7.3)$$

Choosing  $y = \mu_{x_{1:n}}$ ,  $z = \xi/\nu$ , and  $\rho = \nu/(n + 2\nu)$ , we have

$$(1 - \rho)y + \rho z = \frac{2\xi + \sum_{j=1}^n s(x_j)}{2\nu + n} = \frac{a\mu_{x_A} + b\mu_{x_B}}{a + b}. \quad (3.7.4)$$

Note that  $\mu_{x_A}, \mu_{x_B}, \mu_{x_{1:n}} \in U'$ , by taking various convex combinations of  $\xi/\nu$ ,  $\frac{1}{|A|} \sum_{j \in A} s(x_j)$ ,  $\mu_{x_B} \in U'$ . Thus,

$$\begin{aligned} (\nu + n)\mathcal{L}(\mu_{x_{1:n}}) &= (a + b)\mathcal{L}(y) - \nu\mathcal{L}(y) \\ &\stackrel{(a)}{\leq} (a + b)\mathcal{L}((1 - \rho)y + \rho z) - (a + b)\rho(z - y)^\top \mathcal{L}'(y) - \nu\mathcal{L}(y) \\ &\stackrel{(b)}{\leq} (a + b)\mathcal{L}\left(\frac{a\mu_{x_A} + b\mu_{x_B}}{a + b}\right) + C_0 \\ &\stackrel{(c)}{\leq} a\mathcal{L}(\mu_{x_A}) + b\mathcal{L}(\mu_{x_B}) + C_0, \end{aligned}$$

where (a) is by Equation 3.7.3, (b) is by Equations 3.7.2 and 3.7.4, and (c) is by the



convexity of  $\mathcal{L}$ . Hence,  $(\nu + n)^{k/2} \tilde{m}(x_{1:n}) \leq (ab)^{k/2} \tilde{m}(x_A) \tilde{m}(x_B) \exp(C_0)$ , so by our choice of  $C_1$  and  $C_2$ ,

$$\frac{m(x_{1:n})}{m(x_A)m(x_B)} \leq \frac{C_2 \tilde{m}(x_{1:n})}{C_1^2 \tilde{m}(x_A) \tilde{m}(x_B)} \leq \frac{C_2 \exp(C_0)}{C_1^2} \left( \frac{ab}{n + \nu} \right)^{k/2}. \quad \square$$

### 3.8 Marginal inequality for subsets of the data

In this section, we prove Lemma 3.8.4, the key lemma used in the proof of Theorem 3.6.2. First, we need a few supporting results.

Given  $y_1, \dots, y_n \in \mathbb{R}^\ell$  (for some  $\ell > 0$ ),  $\beta \in (0, 1]$ , and  $U \subset \mathbb{R}^\ell$ , define

$$\mathcal{I}_\beta(y_{1:n}, U) = \prod_{\substack{A \subset \{1, \dots, n\}: \\ |A| \geq \beta n}} I\left(\frac{1}{|A|} \sum_{j \in A} y_j \in U\right),$$

where as usual,  $I(E)$  is 1 if  $E$  is true, and 0 otherwise.

**Lemma 3.8.1** (Capture lemma). *Let  $V \subset \mathbb{R}^k$  be open and convex. Let  $Q$  be a probability measure on  $\mathbb{R}^k$  such that:*

- (1)  $\mathbb{E}|Y| < \infty$  when  $Y \sim Q$ ,
- (2)  $Q(\bar{V}) = 1$ , and
- (3)  $Q(L) = 0$  for any hyperplane  $L$  that does not intersect  $V$ .

If  $Y_1, Y_2, \dots \stackrel{\text{iid}}{\sim} Q$ , then for any  $\beta \in (0, 1]$  there exists  $U \subset V$  compact such that  $\mathcal{I}_\beta(Y_{1:n}, U) \xrightarrow{\text{a.s.}} 1$  as  $n \rightarrow \infty$ .

*Proof.* The proof is rather long, but not terribly difficult; see Section 3.12.  $\square$

**Proposition 3.8.2.** *Let  $Z_1, Z_2, \dots \in \mathbb{R}^k$  be i.i.d.. If  $\beta \in (0, 1]$  and  $U \subset \mathbb{R}^k$  such that  $\mathbb{P}(Z_j \notin U) < \beta/2$ , then  $\mathcal{I}_\beta(Y_{1:n}, [\frac{1}{2}, 1]) \xrightarrow{\text{a.s.}} 1$  as  $n \rightarrow \infty$ , where  $Y_j = I(Z_j \in U)$ .*

*Proof.* By the law of large numbers,  $\frac{1}{n} \sum_{j=1}^n I(Z_j \notin U) \xrightarrow{\text{a.s.}} \mathbb{P}(Z_j \notin U) < \beta/2$ .

Hence, with probability 1, for all  $n$  sufficiently large,  $\frac{1}{n} \sum_{j=1}^n I(Z_j \notin U) \leq \beta/2$  holds.

When it holds, we have that for any  $A \subset \{1, \dots, n\}$  such that  $|A| \geq \beta n$ ,

$$\frac{1}{|A|} \sum_{j \in A} I(Z_j \in U) = 1 - \frac{1}{|A|} \sum_{j \in A} I(Z_j \notin U) \geq 1 - \frac{1}{\beta n} \sum_{j=1}^n I(Z_j \notin U) \geq 1/2,$$

i.e. when it holds, we have  $\mathcal{I}_\beta(Y_{1:n}, [\frac{1}{2}, 1]) = 1$ . Hence,  $\mathcal{I}_\beta(Y_{1:n}, [\frac{1}{2}, 1]) \xrightarrow{\text{a.s.}} 1$ .  $\square$

In the following,  $\mu_x = (\xi + s(x))/(\nu + 1)$ , as in Equation 3.7.1.

**Proposition 3.8.3.** *Consider a well-behaved exponential family with conjugate prior. Let  $P$  be a probability measure on  $\mathcal{X}$  such that  $\mathbb{P}(s(X) \in \bar{\mathcal{M}}) = 1$  when  $X \sim P$ . Let  $X_1, X_2, \dots \stackrel{\text{iid}}{\sim} P$ . Then for any  $\beta \in (0, 1]$  there exists  $U \subset \mathcal{M}$  compact such that  $\mathcal{I}_\beta(Y_{1:n}, [\frac{1}{2}, 1]) \xrightarrow{\text{a.s.}} 1$  as  $n \rightarrow \infty$ , where  $Y_j = I(\mu_{X_j} \in U)$ .*

*Proof.* Since  $\mathcal{M}$  is open and convex, then for any  $y \in \bar{\mathcal{M}}$ ,  $z \in \mathcal{M}$ , and  $\rho \in (0, 1)$ , we have  $\rho y + (1 - \rho)z \in \mathcal{M}$  (by e.g. Rockafellar (1970) 6.1). Taking  $z = \xi/\nu$  and  $\rho = 1/(\nu + 1)$ , this implies that the set  $U_0 = \{(\xi + y)/(\nu + 1) : y \in \bar{\mathcal{M}}\}$  is contained in  $\mathcal{M}$ . Note that  $U_0$  is closed and  $\mathbb{P}(\mu_X \in U_0) = \mathbb{P}(s(X) \in \bar{\mathcal{M}}) = 1$ . Let  $\beta \in (0, 1]$ , and choose  $r \in (0, \infty)$  such that  $\mathbb{P}(|\mu_X| > r) < \beta/2$ . Letting  $U = \{y \in U_0 : |y| \leq r\}$ , we have that  $U \subset \mathcal{M}$ , and  $U$  is compact. Further,  $\mathbb{P}(\mu_X \notin U) < \beta/2$ , so by applying Proposition 3.8.2 with  $Z_j = \mu_{X_j}$ , we have  $\mathcal{I}_\beta(Y_{1:n}, [\frac{1}{2}, 1]) \xrightarrow{\text{a.s.}} 1$ .  $\square$

**Lemma 3.8.4.** *Consider a well-behaved exponential family with conjugate prior, and the resulting collection of single-cluster marginals  $m(\cdot)$ . Let  $P$  be a probability measure on  $\mathcal{X}$  satisfying Conditions 3.6.1 (for the  $s$  and  $\mathcal{M}$  from the exponential family under consideration), and let  $X_1, X_2, \dots \stackrel{\text{iid}}{\sim} P$ . Then for any  $\beta \in (0, 1]$  there exists  $c \in (0, \infty)$  such that with probability 1, for all  $n$  sufficiently large, the following event holds: for every subset  $J \subset \{1, \dots, n\}$  such that  $|J| \geq \beta n$ , there exists  $K \subset J$  such that  $|K| \geq \frac{1}{2}|J|$  and for any  $j \in K$ ,*

$$m(X_J) \leq c m(X_{J \setminus j}) m(X_j).$$

*Proof.* Let  $\beta \in (0, 1]$ . Since  $\mathcal{M}$  is open and convex, and Conditions 3.6.1 hold by assumption, then by Lemma 3.8.1 (with  $V = \mathcal{M}$ ) there exists  $U_1 \subset \mathcal{M}$  compact such that  $\mathcal{I}_{\beta/2}(s(X_{1:n}), U_1) \xrightarrow{\text{a.s.}} 1$  as  $n \rightarrow \infty$ , where  $s(X_{1:n}) = (s(X_1), \dots, s(X_n))$ . By Proposition 3.8.3 above, there exists  $U_2 \subset \mathcal{M}$  compact such that  $\mathcal{I}_\beta(Y_{1:n}, [\frac{1}{2}, 1]) \xrightarrow{\text{a.s.}} 1$  as  $n \rightarrow \infty$ , where  $Y_j = I(\mu_{X_j} \in U_2)$ . Hence,

$$\mathcal{I}_{\beta/2}(s(X_{1:n}), U_1) \mathcal{I}_\beta(Y_{1:n}, [\frac{1}{2}, 1]) \xrightarrow[n \rightarrow \infty]{\text{a.s.}} 1.$$

Choose  $C \in (0, \infty)$  according to Proposition 3.7.2 applied to  $U := U_1 \cup U_2$ . We will prove the result with  $c = (\nu + 1)^{k/2} C$ . (Recall that  $k$  is the dimension of  $s : \mathcal{X} \rightarrow \mathbb{R}^k$ .)

Let  $n$  large enough that  $\beta n \geq 2$ , and suppose that  $\mathcal{I}_{\beta/2}(s(X_{1:n}), U_1) = 1$  and  $\mathcal{I}_\beta(Y_{1:n}, [\frac{1}{2}, 1]) = 1$ . Let  $J \subset \{1, \dots, n\}$  such that  $|J| \geq \beta n$ . Then for any  $j \in J$ ,

$$\frac{1}{|J \setminus j|} \sum_{i \in J \setminus j} s(X_i) \in U_1 \subset U$$

since  $\mathcal{I}_{\beta/2}(s(X_{1:n}), U_1) = 1$  and  $|J \setminus j| \geq |J|/2 \geq (\beta/2)n$ . Hence, for any  $j \in K$ ,

where  $K = \{j \in J : \mu_{X_j} \in U\}$ , we have

$$\frac{m(X_J)}{m(X_{J \setminus j})m(X_j)} \leq C \left( \frac{(\nu + |J| - 1)(\nu + 1)}{\nu + |J|} \right)^{k/2} \leq C(\nu + 1)^{k/2} = c$$

by our choice of  $C$  above, and

$$\frac{|K|}{|J|} \geq \frac{1}{|J|} \sum_{j \in J} I(\mu_{X_j} \in U_2) = \frac{1}{|J|} \sum_{j \in J} Y_j \geq 1/2$$

since  $\mathcal{I}_\beta(Y_{1:n}, [\frac{1}{2}, 1]) = 1$  and  $|J| \geq \beta n$ . □

### 3.9 Miscellaneous proofs

*Proof of Proposition 3.3.3.* There are two cases: (A)  $\sigma \in [0, 1)$  and  $\vartheta > -\sigma$ , or (B)  $\sigma < 0$  and  $\vartheta = N|\sigma|$ . In either case,  $\sigma < 1$ , so

$$\frac{w_n(a)}{aw_n(a-1)w_n(1)} = \frac{1 - \sigma + a - 2}{a} \leq \frac{1 - \sigma}{2} + 1$$

whenever  $n \geq 2$  and  $a \in \{2, \dots, n\}$ , and hence  $\limsup_n c_{w_n} < \infty$ .

For any  $n > t \geq 1$ , in case (A) we have

$$\frac{v_n(t)}{v_n(t+1)} = \frac{t+1}{\vartheta + t\sigma},$$

and the same holds in case (B) if also  $t < N$ . Meanwhile, whenever  $N < t < n$  in case (B),  $v_n(t)/v_n(t+1) = 0/0 = 0$  by convention. Therefore,  $\limsup_n c_{v_n}(t) < \infty$  in either case, for any  $t \in \{1, 2, \dots\}$  except  $t = N$  in case (B). □

*Proof of Theorem 3.3.4.* Let  $t \in \{1, 2, \dots\}$ , and assume Conditions 3.3.1 and 3.3.2 hold. Let  $x_1, x_2, \dots \in \mathcal{X}$ , and suppose  $\sup_{c \in [0, \infty)} \liminf_n \varphi_t(x_{1:n}, c) > 0$  (which occurs with probability 1). We show that this implies  $\limsup_n p(T_n = t \mid x_{1:n}) < 1$ , proving the theorem.

Let  $\alpha \in (0, \infty)$  such that  $\limsup_n c_{w_n} < \alpha$  and  $\limsup_n c_{v_n}(t) < \alpha$ . Choose  $c \in [0, \infty)$  and  $\varepsilon \in (0, 1)$  such that  $\liminf_n \varphi_t(x_{1:n}, c) > \varepsilon$ . Choose  $N > 2t/\varepsilon$  large enough that for any  $n > N$  we have  $c_{w_n} < \alpha$ ,  $c_{v_n}(t) < \alpha$ , and  $\varphi_t(x_{1:n}, c) > \varepsilon$ . Then by Lemma 3.3.5, for any  $n > N$ ,

$$p(T_n = t \mid x_{1:n}) \leq \frac{C_t(x_{1:n}, c)}{1 + C_t(x_{1:n}, c)} \leq \frac{2t\alpha^2/\varepsilon}{1 + 2t\alpha^2/\varepsilon},$$

since  $\varphi_t(x_{1:n}, c) - t/n > \varepsilon - \varepsilon/2 = \varepsilon/2$  (and  $y \mapsto y/(1+y)$  is monotone increasing on  $[0, \infty)$ ). Since this upper bound does not depend on  $n$  (and is less than 1), then  $\limsup_n p(T_n = t \mid x_{1:n}) < 1$ .  $\square$

*Proof of Proposition 3.6.3.* (1) For any  $\theta \in \Theta$  and any  $j \in \{1, \dots, k\}$ ,

$$\int_{\mathcal{X}} s_j(x)^2 p_\theta(x) d\lambda(x) = \exp(-\kappa(\theta)) \frac{\partial^2}{\partial \theta_j^2} \int_{\mathcal{X}} \exp(\theta^\top s(x)) d\lambda(x) < \infty$$

(Hoffmann-Jørgensen, 1994, 8.36.1). Since  $P$  has density  $f = \sum \pi_i p_{\theta(i)}$  with respect to  $\lambda$ , then

$$\mathbb{E} s_j(X)^2 = \int_{\mathcal{X}} s_j(x)^2 f(x) d\lambda(x) = \sum_{i=1}^t \pi_i \int_{\mathcal{X}} s_j(x)^2 p_{\theta(i)}(x) d\lambda(x) < \infty,$$

and hence

$$(\mathbb{E} |s(X)|)^2 \leq \mathbb{E} |s(X)|^2 = \mathbb{E} s_1(X)^2 + \dots + \mathbb{E} s_k(X)^2 < \infty.$$

(2) Note that  $S_P(s) \subset S_\lambda(s)$  (in fact, they are equal since  $P_\theta$  and  $\lambda$  are mutually absolutely continuous for any  $\theta \in \Theta$ ), and therefore

$$S_P(s) \subset S_\lambda(s) \subset C_\lambda(s) = \bar{\mathcal{M}}$$

by Proposition 3.10.1(8). Hence,

$$\mathbb{P}(s(X) \in \bar{\mathcal{M}}) \geq \mathbb{P}(s(X) \in S_P(s)) = P_{s^{-1}(\text{support}(Ps^{-1}))} = 1.$$

(3) Suppose  $\lambda$  is absolutely continuous with respect to Lebesgue measure,  $\mathcal{X}$  is open and connected, and  $s$  is real analytic. Let  $L \subset \mathbb{R}^k$  be a hyperplane, and write  $L = \{z \in \mathbb{R}^k : z^\top y = b\}$  where  $y \in \mathbb{R}^k \setminus \{0\}$ ,  $b \in \mathbb{R}$ . Define  $g : \mathcal{X} \rightarrow \mathbb{R}$  by  $g(x) = s(x)^\top y - b$ . Then  $g$  is real analytic on  $\mathcal{X}$ , since a finite sum of real analytic functions is real analytic. Since  $\mathcal{X}$  is connected, it follows that either  $g$  is identically zero, or the set  $V = \{x \in \mathcal{X} : g(x) = 0\}$  has Lebesgue measure zero (Krantz, 1992). Now,  $g$  cannot be identically zero, since for any  $\theta \in \Theta$ , letting  $Z \sim P_\theta$ , we have

$$0 < y^\top \kappa''(\theta)y = y^\top (\text{Cov } s(Z))y = \text{Var}(y^\top s(Z)) = \text{Var } g(Z)$$

by Proposition 3.10.1(2) and (3). Consequently,  $V$  must have Lebesgue measure zero. Hence,  $P(V) = 0$ , since  $P$  is absolutely continuous with respect to  $\lambda$ , and thus, with respect to Lebesgue measure. Therefore,

$$\mathbb{P}(s(X) \in L) = \mathbb{P}(g(X) = 0) = P(V) = 0. \quad \square$$

### 3.10 Exponential family properties

We note some well-known properties of exponential families satisfying Conditions 3.5.1. For a general reference on this material, see [Hoffmann-Jørgensen \(1994\)](#). Let  $S_\lambda(s) = \text{support}(\lambda s^{-1})$ , that is,

$$S_\lambda(s) = \{z \in \mathbb{R}^k : \lambda(s^{-1}(U)) \neq 0 \text{ for every neighborhood } U \text{ of } z\}.$$

Let  $C_\lambda(s)$  be the closed convex hull of  $S_\lambda(s)$  (that is, the intersection of all closed convex sets containing it). Given  $U \subset \mathbb{R}^k$ , let  $U^\circ$  denote its interior. Given a (sufficiently smooth) function  $f : \mathbb{R}^k \rightarrow \mathbb{R}$ , we use  $f'$  to denote its gradient, that is,  $f'(x)_i = \frac{\partial f}{\partial x_i}(x)$ , and  $f''(x)$  to denote its Hessian matrix, that is,  $f''(x)_{ij} = \frac{\partial^2 f}{\partial x_i \partial x_j}(x)$ .

**Proposition 3.10.1.** *If Conditions 3.5.1 are satisfied, then:*

- (1)  $\kappa$  is  $C^\infty$  smooth and strictly convex on  $\Theta$ ,
- (2)  $\kappa'(\theta) = \mathbb{E}s(X)$  and  $\kappa''(\theta) = \text{Cov } s(X)$  when  $\theta \in \Theta$  and  $X \sim P_\theta$ ,
- (3)  $\kappa''(\theta)$  is symmetric positive definite for all  $\theta \in \Theta$ ,
- (4)  $\kappa' : \Theta \rightarrow \mathcal{M}$  is a  $C^\infty$  smooth bijection,
- (5)  $\kappa'^{-1} : \mathcal{M} \rightarrow \Theta$  is  $C^\infty$  smooth,
- (6)  $\Theta$  is open and convex,
- (7)  $\mathcal{M}$  is open and convex,
- (8)  $\mathcal{M} = C_\lambda(s)^\circ$  and  $\bar{\mathcal{M}} = C_\lambda(s)$ , and
- (9)  $\kappa'^{-1}(\mu) = \text{argmax}_{\theta \in \Theta}(\theta^\top \mu - \kappa(\theta))$  for all  $\mu \in \mathcal{M}$ . The maximizing  $\theta \in \Theta$  always exists and is unique.

*Proof.* These properties are all well-known. Let us abbreviate [Hoffmann-Jørgensen \(1994\)](#) as HJ. For (1), see HJ 8.36(1) and HJ 12.7.5. For (6),(2),(3), and (4), see HJ 8.36, 8.36.2-3, 12.7(2), and 12.7.11, respectively. Item (5) and openness in (7) follow, using the inverse function theorem ([Knapp, 2005](#), 3.21). Item (8) and convexity in (7) follow, using HJ 8.36.15 and [Rockafellar \(1970\)](#) 6.2-3. Item (9) follows from HJ 8.36.15 and item (4).  $\square$

Given an exponential family with conjugate prior as in Section [3.5.2](#), the joint density of  $x_1, \dots, x_n \in \mathcal{X}$  and  $\theta \in \mathbb{R}^k$  is

$$\begin{aligned} p_\theta(x_1) \cdots p_\theta(x_n) \pi_{\xi, \nu}(\theta) \\ = \exp\left((\nu + n)(\theta^\top \mu_{x_{1:n}} - \kappa(\theta))\right) \exp(-\psi(\xi, \nu)) I(\theta \in \Theta) \end{aligned} \quad (3.10.1)$$

where  $\mu_{x_{1:n}} = (\xi + \sum_{j=1}^n s(x_j))/(\nu + n)$ . The marginal density, defined as in Equation [3.2.5](#), is

$$m(x_{1:n}) = \exp\left(\psi\left(\xi + \sum s(x_j), \nu + n\right) - \psi(\xi, \nu)\right) \quad (3.10.2)$$

when this quantity is well-defined.

**Proposition 3.10.2.** *If Conditions [3.5.1](#) are satisfied, then:*

- (1)  $\psi(\xi, \nu)$  is finite and  $C^\infty$  smooth on  $\Xi$ ,
- (2) if  $s(x_1), \dots, s(x_n) \in S_\lambda(s)$  and  $(\xi, \nu) \in \Xi$ , then  $(\xi + \sum s(x_j), \nu + n) \in \Xi$ ,
- (3)  $\{\pi_{\xi, \nu} : (\xi, \nu) \in \Xi\}$  is a conjugate family for  $\{p_\theta : \theta \in \Theta\}$ , and
- (4) if  $s : \mathcal{X} \rightarrow \mathbb{R}^k$  is continuous,  $(\xi, \nu) \in \Xi$ , and  $\lambda(U) \neq 0$  for any nonempty  $U \subset \mathcal{X}$  that is open in  $\mathcal{X}$ , then  $m(x_{1:n}) < \infty$  for any  $x_1, \dots, x_n \in \mathcal{X}$ .



*Proof.* (1) For finiteness, see [Diaconis and Ylvisaker \(1979\)](#), Theorem 1. Smoothness holds for the same reason that  $\kappa$  is smooth ([Hoffmann-Jørgensen, 1994](#), 8.36(1)). (Note that  $\Xi$  is open in  $\mathbb{R}^{k+1}$ , since  $\mathcal{M}$  is open in  $\mathbb{R}^k$ .)

(2) Since  $C_\lambda(s)$  is convex,  $\frac{1}{n} \sum s(x_j) \in C_\lambda(s)$ . Since  $C_\lambda(s) = \bar{\mathcal{M}}$  and  $\mathcal{M}$  is open and convex ([3.10.1\(7\)](#) and (8)), then  $(\xi + \sum s(x_j))/(\nu + n) \in \mathcal{M}$ , as a (strict) convex combination of  $\frac{1}{n} \sum s(x_j) \in \bar{\mathcal{M}}$  and  $\xi/\nu \in \mathcal{M}$  ([Rockafellar, 1970](#), 6.1).

(3) Let  $(\xi, \nu) \in \Xi$ ,  $\theta \in \Theta$ . If  $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} P_\theta$  then  $s(X_1), \dots, s(X_n) \in S_\lambda(s)$  almost surely, and thus  $(\xi + \sum s(X_j), \nu + n) \in \Xi$  (a.s.) by (2). By [Equations 3.10.1](#) and [3.10.2](#), the posterior is  $\pi_{\xi + \sum s(X_j), \nu + n}$ .

(4) The assumptions imply  $\{s(x) : x \in \mathcal{X}\} \subset S_\lambda(s)$ , and therefore, for any  $x_1, \dots, x_n \in \mathcal{X}$ , we have  $(\xi + \sum s(x_j), \nu + n) \in \Xi$  by (2). Thus, by (1) and [Equation 3.10.2](#),  $m(x_{1:n}) < \infty$ .  $\square$

It is worth mentioning that while  $\Xi \subset \{(\xi, \nu) \in \mathbb{R}^{k+1} : \psi(\xi, \nu) < \infty\}$ , it may be a strict subset — often,  $\Xi$  is not quite the full set of parameters on which  $\pi_{\xi, \nu}$  can be defined.

### 3.11 Bounds on the Laplace approximation

Our proof uses the following simple bounds on the Laplace approximation. These bounds are not fundamentally new, but the precise formulation we require does not seem to appear in the literature, so we have included it for the reader's convenience. [Lemma 3.11.1](#) is simply a multivariate version of the bounds given by [De Bruijn \(1970\)](#), and [Corollary 3.11.2](#) is a straightforward consequence, putting the lemma in

a form most convenient for our purposes.

Given symmetric matrices  $A$  and  $B$ , let us write  $A \preceq B$  to mean that  $B - A$  is positive semidefinite. Given  $A \in \mathbb{R}^{k \times k}$  symmetric positive definite and  $\varepsilon, t \in (0, \infty)$ , define

$$C(t, \varepsilon, A) = \mathbb{P}(|A^{-1/2}Z| \leq \varepsilon\sqrt{t})$$

where  $Z \sim \text{Normal}(0, I_{k \times k})$ . Note that  $C(t, \varepsilon, A) \rightarrow 1$  as  $t \rightarrow \infty$ . Let  $B_\varepsilon(x_0) = \{x \in \mathbb{R}^k : |x - x_0| \leq \varepsilon\}$  denote the closed ball of radius  $\varepsilon > 0$  at  $x_0 \in \mathbb{R}^k$ .

**Lemma 3.11.1.** *Let  $E \subset \mathbb{R}^k$  be open. Let  $f : E \rightarrow \mathbb{R}$  be  $C^2$  smooth with  $f'(x_0) = 0$  for some  $x_0 \in E$ . Define*

$$g(t) = \int_E \exp(-tf(x)) dx$$

for  $t \in (0, \infty)$ . Suppose  $\varepsilon \in (0, \infty)$  such that  $B_\varepsilon(x_0) \subset E$ ,  $0 < \delta \leq \inf\{f(x) - f(x_0) : x \in E \setminus B_\varepsilon(x_0)\}$ , and  $A, B$  are symmetric positive definite matrices such that  $A \preceq f''(x) \preceq B$  for all  $x \in B_\varepsilon(x_0)$ . Then for any  $0 < s \leq t$  we have

$$\frac{C(t, \varepsilon, B)}{|B|^{1/2}} \leq \frac{g(t)}{(2\pi/t)^{k/2} e^{-tf(x_0)}} \leq \frac{C(t, \varepsilon, A)}{|A|^{1/2}} + \left(\frac{t}{2\pi}\right)^{k/2} e^{-(t-s)\delta} e^{sf(x_0)} g(s)$$

where  $|A| = |\det A|$ .

*Remark.* In particular, these assumptions imply  $f$  is strictly convex on  $B_\varepsilon(x_0)$  with unique global minimum at  $x_0$ . Note that the upper bound is trivial unless  $g(s) < \infty$ .

*Proof.* By Taylor's theorem, for any  $x \in B_\varepsilon(x_0)$  there exists  $z_x$  on the line between  $x_0$  and  $x$  such that, letting  $y = x - x_0$ ,

$$f(x) = f(x_0) + y^T f'(x_0) + \frac{1}{2} y^T f''(z_x) y = f(x_0) + \frac{1}{2} y^T f''(z_x) y.$$

Since  $z_x \in B_\varepsilon(x_0)$ , and thus  $A \preceq f''(z_x) \preceq B$ ,

$$\frac{1}{2}y^\top Ay \leq f(x) - f(x_0) \leq \frac{1}{2}y^\top By.$$

Hence,

$$\begin{aligned} e^{tf(x_0)} \int_{B_\varepsilon(x_0)} \exp(-tf(x)) dx &\leq \int_{B_\varepsilon(x_0)} \exp(-\frac{1}{2}(x-x_0)^\top (tA)(x-x_0)) dx \\ &= (2\pi)^{k/2} |(tA)^{-1}|^{1/2} \mathbb{P}(|(tA)^{-1/2}Z| \leq \varepsilon). \end{aligned}$$

Along with a similar argument for the lower bound, this implies

$$\left(\frac{2\pi}{t}\right)^{k/2} \frac{C(t, \varepsilon, B)}{|B|^{1/2}} \leq e^{tf(x_0)} \int_{B_\varepsilon(x_0)} \exp(-tf(x)) dx \leq \left(\frac{2\pi}{t}\right)^{k/2} \frac{C(t, \varepsilon, A)}{|A|^{1/2}}.$$

Considering the rest of the integral, outside of  $B_\varepsilon(x_0)$ , we have

$$0 \leq \int_{E \setminus B_\varepsilon(x_0)} \exp(-tf(x)) dx \leq \exp(-(t-s)(f(x_0) + \delta)) g(s).$$

Combining the preceding four inequalities yields the result.  $\square$

The following corollary tailors the lemma to our purposes. Given a symmetric positive definite matrix  $A \in \mathbb{R}^{k \times k}$ , let  $\lambda_*(A)$  and  $\lambda^*(A)$  be the minimal and maximal eigenvalues, respectively. By diagonalizing  $A$ , it is easy to check that  $\lambda_*(A)I_{k \times k} \preceq A \preceq \lambda^*(A)I_{k \times k}$  and  $\lambda_*(A)^k \leq |A| \leq \lambda^*(A)^k$ .

**Corollary 3.11.2.** *For any  $\alpha, \beta, \gamma, \delta, \varepsilon \in (0, \infty)$  there exist  $c_1 = c_1(\beta, \varepsilon) \in (0, \infty)$  and  $c_2 = c_2(\alpha, \gamma, \delta) \in (0, \infty)$  such that if  $E, f, x_0, A, B$  satisfy all the conditions of Lemma 3.11.1 (for this choice of  $\delta, \varepsilon$ ) and additionally,  $\alpha \leq \lambda_*(A)$ ,  $\beta \geq \lambda^*(B)$ , and  $\gamma \geq e^{f(x_0)}g(1)$ , then*

$$c_1 \leq \frac{\int_E \exp(-tf(x)) dx}{t^{-k/2} \exp(-tf(x_0))} \leq c_2$$

for all  $t \geq 1$ .

*Proof.* The first term in the upper bound of the lemma is  $C(t, \varepsilon, A)/|A|^{1/2} \leq 1/\alpha^{k/2}$ , and with  $s = 1$  the second term is less or equal to  $(t/2\pi)^{k/2} e^{-(t-1)\delta} \gamma$ , which is bounded above for  $t \in [1, \infty)$ . For the lower bound, a straightforward calculation (using  $z^T B z \leq \lambda^*(B) z^T z \leq \beta z^T z$  in the exponent inside the integral) shows that  $C(t, \varepsilon, B)/|B|^{1/2} \geq \mathbb{P}(|Z| \leq \varepsilon\sqrt{\beta})/\beta^{k/2}$  for  $t \geq 1$ .  $\square$

Although we do not need it (and thus, we omit the proof), the following corollary gives the well-known asymptotic form of the Laplace approximation. (As usual,  $g(t) \sim h(t)$  as  $t \rightarrow \infty$  means that  $g(t)/h(t) \rightarrow 1$ .)

**Corollary 3.11.3.** *Let  $E \subset \mathbb{R}^k$  be open. Let  $f : E \rightarrow \mathbb{R}$  be  $C^2$  smooth such that for some  $x_0 \in E$  we have that  $f'(x_0) = 0$ ,  $f''(x_0)$  is positive definite, and  $f(x) > f(x_0)$  for all  $x \in E \setminus \{x_0\}$ . Suppose there exists  $\varepsilon > 0$  such that  $B_\varepsilon(x_0) \subset E$  and  $\inf\{f(x) - f(x_0) : x \in E \setminus B_\varepsilon(x_0)\}$  is positive, and suppose there is some  $s > 0$  such that  $\int_E e^{-sf(x)} dx < \infty$ . Then*

$$\int_E \exp(-tf(x)) dx \sim \left(\frac{2\pi}{t}\right)^{k/2} \frac{\exp(-tf(x_0))}{|f''(x_0)|^{1/2}}$$

as  $t \rightarrow \infty$ .

## 3.12 Capture lemma

In this section, we prove Lemma 3.8.1, which is restated here for the reader's convenience.

The following definitions are standard. Let  $\mathcal{S}$  denote the unit sphere in  $\mathbb{R}^k$ , that is,  $\mathcal{S} = \{x \in \mathbb{R}^k : |x| = 1\}$ . We say that  $H \subset \mathbb{R}^k$  is a *halfspace* if  $H = \{x \in \mathbb{R}^k : x^\top u \prec b\}$ , where  $\prec$  is either  $<$  or  $\leq$ , for some  $u \in \mathcal{S}$ ,  $b \in \mathbb{R}$ . We say that  $L \subset \mathbb{R}^k$  is a *hyperplane* if  $L = \{x \in \mathbb{R}^k : x^\top u = b\}$  for some  $u \in \mathcal{S}$ ,  $b \in \mathbb{R}$ . Given  $U \subset \mathbb{R}^k$ , let  $\partial U$  denote the *boundary of  $U$* , that is,  $\partial U = \bar{U} \setminus U^\circ$ . So, for example, if  $H$  is a halfspace, then  $\partial H$  is a hyperplane. The following notation is also useful: given  $x \in \mathbb{R}^k$ , we call the set  $R_x = \{ax : a > 0\}$  the *ray through  $x$* .

We give the central part of the proof first, postponing some plausible intermediate results for the moment.

**Lemma 3.12.1** (Capture lemma). *Let  $V \subset \mathbb{R}^k$  be open and convex. Let  $P$  be a probability measure on  $\mathbb{R}^k$  such that:*

- (1)  $\mathbb{E}|X| < \infty$  when  $X \sim P$ ,
- (2)  $P(\bar{V}) = 1$ , and
- (3)  $P(L) = 0$  for any hyperplane  $L$  that does not intersect  $V$ .

If  $X_1, X_2, \dots \stackrel{\text{iid}}{\sim} P$ , then for any  $\beta \in (0, 1]$  there exists  $U \subset V$  compact such that  $\mathcal{I}_\beta(X_{1:n}, U) \xrightarrow{\text{a.s.}} 1$  as  $n \rightarrow \infty$ .

*Proof.* Without loss of generality, we may assume  $0 \in V$  (since otherwise we can translate to make it so, obtain  $U$ , and translate back). Let  $\beta \in (0, 1]$ . By Proposition 3.12.3 below, for each  $u \in \mathcal{S}$  there is a closed halfspace  $H_u$  such that  $0 \in H_u^\circ$ ,  $R_u$  intersects  $V \cap \partial H_u$ , and  $\mathcal{I}_\beta(X_{1:n}, H_u) \xrightarrow{\text{a.s.}} 1$  as  $n \rightarrow \infty$ . By Proposition 3.12.5 below, there exist  $u_1, \dots, u_r \in \mathcal{S}$  (for some  $r > 0$ ) such that the set  $U = \bigcap_{i=1}^r H_{u_i}$  is compact

and  $U \subset V$ . Finally,

$$\mathcal{I}_\beta(X_{1:n}, U) = \prod_{i=1}^r \mathcal{I}_\beta(X_{1:n}, H_{u_i}) \xrightarrow[n \rightarrow \infty]{\text{a.s.}} 1. \quad \square$$

The main idea of the lemma is exhibited in the following simpler case, which we will use to prove Proposition 3.12.3.

**Proposition 3.12.2.** *Let  $V = (-\infty, c)$ , where  $-\infty < c \leq \infty$ . Let  $P$  be a probability measure on  $\mathbb{R}$  such that:*

- (1)  $\mathbb{E}|X| < \infty$  when  $X \sim P$ , and
- (2)  $P(V) = 1$ .

*If  $X_1, X_2, \dots \stackrel{\text{iid}}{\sim} P$ , then for any  $\beta \in (0, 1]$  there exists  $b < c$  such that  $\mathcal{I}_\beta(X_{1:n}, (-\infty, b]) \xrightarrow[n \rightarrow \infty]{\text{a.s.}} 1$  as  $n \rightarrow \infty$ .*

*Proof.* Let  $\beta \in (0, 1]$ . By continuity from above, there exists  $a < c$  such that  $\mathbb{P}(X > a) < \beta$ . If  $\mathbb{P}(X > a) = 0$  then the result is trivial, taking  $b = a$ . Suppose  $\mathbb{P}(X > a) > 0$ . Let  $b$  such that  $\mathbb{E}(X \mid X > a) < b < c$ , which is always possible, by a straightforward argument (using  $\mathbb{E}|X| < \infty$  in the  $c = \infty$  case). Let  $B_n = B_n(X_1, \dots, X_n) = \{i \in \{1, \dots, n\} : X_i > a\}$ . Then

$$\begin{aligned} \frac{1}{|B_n|} \sum_{i \in B_n} X_i &= \frac{1}{\frac{1}{n}|B_n|} \frac{1}{n} \sum_{i=1}^n X_i I(X_i > a) \\ &\xrightarrow[n \rightarrow \infty]{\text{a.s.}} \frac{\mathbb{E}(X I(X > a))}{\mathbb{P}(X > a)} = \mathbb{E}(X \mid X > a) < b. \end{aligned}$$

Now, fix  $n \in \{1, 2, \dots\}$ , and suppose  $0 < |B_n| < \beta n$  and  $\frac{1}{|B_n|} \sum_{i \in B_n} X_i < b$ , noting that with probability 1, this happens for all  $n$  sufficiently large. We show that this

implies  $\mathcal{I}_\beta(X_{1:n}, (-\infty, b]) = 1$ . This will prove the result.

Let  $A \subset \{1, \dots, n\}$  such that  $|A| \geq \beta n$ . Let  $M = \{\pi_1, \dots, \pi_{|A|}\}$  where  $\pi$  is a permutation of  $\{1, \dots, n\}$  such that  $X_{\pi_1} \geq \dots \geq X_{\pi_n}$  (that is,  $M \subset \{1, \dots, n\}$  consists of the indices of  $|A|$  of the largest entries of  $(X_1, \dots, X_n)$ ). Then  $|M| = |A| \geq \beta n \geq |B_n|$ , and it follows that  $B_n \subset M$ . Therefore,

$$\frac{1}{|A|} \sum_{i \in A} X_i \leq \frac{1}{|M|} \sum_{i \in M} X_i \leq \frac{1}{|B_n|} \sum_{i \in B_n} X_i \leq b,$$

as desired.  $\square$

The first of the two propositions used in Lemma 3.12.1 is the following.

**Proposition 3.12.3.** *Let  $V$  and  $P$  satisfy the conditions of Lemma 3.12.1, and also assume  $0 \in V$ . If  $X_1, X_2, \dots \stackrel{\text{iid}}{\sim} P$  then for any  $\beta \in (0, 1]$  and any  $u \in \mathcal{S}$  there is a closed halfspace  $H \subset \mathbb{R}^k$  such that*

- (1)  $0 \in H^\circ$ ,
- (2)  $R_u$  intersects  $V \cap \partial H$ , and
- (3)  $\mathcal{I}_\beta(X_{1:n}, H) \xrightarrow{\text{a.s.}} 1$  as  $n \rightarrow \infty$ .

*Proof.* Let  $\beta \in (0, 1]$  and  $u \in \mathcal{S}$ . Either (a)  $R_u \subset V$ , or (b)  $R_u$  intersects  $\partial V$ .

(Case (a)) Suppose  $R_u \subset V$ . Let  $Y_i = X_i^\top u$  for  $i = 1, 2, \dots$ . Then  $\mathbb{E}|Y_i| \leq \mathbb{E}|X_i||u| = \mathbb{E}|X_i| < \infty$ , and thus, by Proposition 3.12.2 (with  $c = \infty$ ) there exists  $b \in \mathbb{R}$  such that  $\mathcal{I}_\beta(Y_{1:n}, (-\infty, b]) \xrightarrow{\text{a.s.}} 1$ . Let us choose this  $b$  to be positive, which is always possible since  $\mathcal{I}_\beta(Y_{1:n}, (-\infty, b])$  is nondecreasing as a function of  $b$ . Let  $H = \{x \in \mathbb{R}^k : x^\top u \leq b\}$ . Then  $0 \in H^\circ$ , since  $b > 0$ , and  $R_u$  intersects  $V \cap \partial H$  at  $bu$ ,

since  $R_u \subset V$  and  $bu^T u = b$ . And since  $\frac{1}{|A|} \sum_{i \in A} Y_i \leq b$  if and only if  $\frac{1}{|A|} \sum_{i \in A} X_i \in H$ , we have  $\mathcal{I}_\beta(X_{1:n}, H) \xrightarrow{\text{a.s.}} 1$ .

(Case (b)) Suppose  $R_u$  intersects  $\partial V$  at some point  $z \in \mathbb{R}^k$ . Note that  $z \neq 0$  since  $0 \notin R_u$ . Since  $\bar{V}$  is convex, it has a supporting hyperplane at  $z$ , and thus, there exist  $v \in \mathcal{S}$  and  $c \in \mathbb{R}$  such that  $G = \{x \in \mathbb{R}^k : x^T v \leq c\}$  satisfies  $\bar{V} \subset G$  and  $z \in \partial G$  (Rockafellar, 1970, 11.2). Note that  $c > 0$  and  $V \cap \partial G = \emptyset$  since  $0 \in V$  and  $V$  is open. Letting  $Y_i = X_i^T v$  for  $i = 1, 2, \dots$ , we have

$$\mathbb{P}(Y_i \leq c) = \mathbb{P}(X_i^T v \leq c) = \mathbb{P}(X_i \in G) \geq \mathbb{P}(X_i \in \bar{V}) = P(\bar{V}) = 1,$$

and hence,

$$\mathbb{P}(Y_i \geq c) = \mathbb{P}(Y_i = c) = \mathbb{P}(X_i^T v = c) = \mathbb{P}(X_i \in \partial G) = P(\partial G) = 0,$$

by our assumptions on  $P$ , since  $\partial G$  is a hyperplane that does not intersect  $V$ . Consequently,  $\mathbb{P}(Y_i < c) = 1$ . Also, as before,  $\mathbb{E}|Y_i| < \infty$ . Thus, by Proposition 3.12.2, there exists  $b < c$  such that  $\mathcal{I}_\beta(Y_{1:n}, (-\infty, b]) \xrightarrow{\text{a.s.}} 1$ . Since  $c > 0$ , we may choose this  $b$  to be positive (as before). Letting  $H = \{x \in \mathbb{R}^k : x^T v \leq b\}$ , we have  $\mathcal{I}_\beta(X_{1:n}, H) \xrightarrow{\text{a.s.}} 1$ . Also,  $0 \in H^\circ$  since  $b > 0$ .

Now, we must show that  $R_u$  intersects  $V \cap \partial H$ . First, since  $z \in R_u$  means  $z = au$  for some  $a > 0$ , and since  $z \in \partial G$  means  $z^T v = c > 0$ , we find that  $u^T v > 0$  and  $z = cu/u^T v$ . Therefore, letting  $y = bu/u^T v$ , we have  $y \in R_u \cap V \cap \partial H$ , since

- (i)  $b/u^T v > 0$ , and thus  $y \in R_u$ ,
- (ii)  $y^T v = b$ , and thus  $y \in \partial H$ ,
- (iii)  $0 < b/u^T v < c/u^T v$ , and thus  $y$  is a (strict) convex combination of  $0 \in V$  and



$z \in \bar{V}$ , hence  $y \in V$  (Rockafellar, 1970, 6.1).  $\square$

To prove Proposition 3.12.5, we need the following geometrically intuitive facts.

**Proposition 3.12.4.** *Let  $V \subset \mathbb{R}^k$  be open and convex, with  $0 \in V$ . Let  $H$  be a closed halfspace such that  $0 \in H^\circ$ . Let  $T = \{x/|x| : x \in V \cap \partial H\}$ . Then*

- (1)  $T$  is open in  $\mathcal{S}$ ,
- (2)  $T = \{u \in \mathcal{S} : R_u \text{ intersects } V \cap \partial H\}$ , and
- (3) if  $x \in H$ ,  $x \neq 0$ , and  $x/|x| \in T$ , then  $x \in V$ .

*Proof.* Write  $H = \{x \in \mathbb{R}^k : x^\top v \leq b\}$ , with  $v \in \mathcal{S}$ ,  $b > 0$ . Let  $\mathcal{S}_+ = \{u \in \mathcal{S} : u^\top v > 0\}$ . (1) Define  $f : \partial H \rightarrow \mathcal{S}_+$  by  $f(x) = x/|x|$ , noting that  $0 \notin \partial H$ . It is easy to see that  $f$  is a homeomorphism. Since  $V$  is open in  $\mathbb{R}^k$ , then  $V \cap \partial H$  is open in  $\partial H$ . Hence,  $T = f(V \cap \partial H)$  is open in  $\mathcal{S}_+$ , and since  $\mathcal{S}_+$  is open in  $\mathcal{S}$ , then  $T$  is also open in  $\mathcal{S}$ . Items (2) and (3) are easily checked.  $\square$

**Proposition 3.12.5.** *Let  $V \subset \mathbb{R}^k$  be open and convex, with  $0 \in V$ . If  $(H_u : u \in \mathcal{S})$  is a collection of closed halfspaces such that for all  $u \in \mathcal{S}$ ,*

- (1)  $0 \in H_u^\circ$  and
- (2)  $R_u$  intersects  $V \cap \partial H_u$ ,

*then there exist  $u_1, \dots, u_r \in \mathcal{S}$  (for some  $r > 0$ ) such that the set  $U = \bigcap_{i=1}^r H_{u_i}$  is compact and  $U \subset V$ .*

*Proof.* For  $u \in \mathcal{S}$ , define  $T_u = \{x/|x| : x \in V \cap \partial H_u\}$ . By part (1) of Proposition 3.12.4,  $T_u$  is open in  $\mathcal{S}$ , and by part (2),  $u \in T_u$ , since  $R_u$  intersects  $V \cap \partial H_u$ . Thus,  $(T_u : u \in \mathcal{S})$  is an open cover of  $\mathcal{S}$ . Since  $\mathcal{S}$  is compact, there is a finite subcover: there exist  $u_1, \dots, u_r \in \mathcal{S}$  (for some  $r > 0$ ) such that  $\bigcup_{i=1}^r T_{u_i} \supset \mathcal{S}$ , and in fact,  $\bigcup_{i=1}^r T_{u_i} = \mathcal{S}$ . Let  $U = \bigcap_{i=1}^r H_{u_i}$ . Then  $U$  is closed and convex (as an intersection of closed, convex sets). Further,  $U \subset V$  since for any  $x \in U$ , if  $x = 0$  then  $x \in V$  by assumption, while if  $x \neq 0$  then  $x/|x| \in T_{u_i}$  for some  $i \in \{1, \dots, r\}$  and  $x \in U \subset H_{u_i}$ , so  $x \in V$  by Proposition 3.12.4(3).

In order to show that  $U$  is compact, we just need to show it is bounded, since we already know it is closed. Suppose not, and let  $x_1, x_2, \dots \in U \setminus \{0\}$  such that  $|x_n| \rightarrow \infty$  as  $n \rightarrow \infty$ . Let  $v_n = x_n/|x_n|$ . Since  $\mathcal{S}$  is compact, then  $(v_n)$  has a convergent subsequence such that  $v_{n_i} \rightarrow u$  for some  $u \in \mathcal{S}$ . Then for any  $a > 0$ , we have  $av_{n_i} \in U$  for all  $i$  sufficiently large (since  $av_{n_i}$  is a convex combination of  $0 \in U$  and  $|x_{n_i}|v_{n_i} = x_{n_i} \in U$  whenever  $|x_{n_i}| \geq a$ ). Since  $av_{n_i} \rightarrow au$ , and  $U$  is closed, then  $au \in U$ . Thus,  $au \in U$  for all  $a > 0$ , i.e.  $R_u \subset U$ . But  $u \in T_{u_j}$  for some  $j \in \{1, \dots, r\}$ , so  $R_u$  intersects  $\partial H_{u_j}$  (by Proposition 3.12.4(2)), and thus  $au \notin H_{u_j} \supset U$  for all  $a > 0$  sufficiently large. This is a contradiction. Therefore,  $U$  is bounded.  $\square$

# CHAPTER FOUR

---

## The mixture of finite mixtures model

## 4.1 Introduction

In this chapter, we explore the properties of a class of variable-dimension mixture models that we refer to as mixtures of finite mixtures (MFMs), finding that in fact they share many of the same attractive properties as Dirichlet process mixtures (DPMs) — a simple partition distribution, restaurant process, random discrete measure interpretation, stick-breaking representation in some cases, and exchangeability properties — and that many of the inference algorithms developed for DPMs can be directly applied to MFMs as well. Also, as with DPMs, in some cases there are theoretical guarantees on the posterior rate of convergence for density estimation with MFMs.

Further, in Chapter 5, we empirically observe that the posterior properties of MFMs and DPMs are remarkably similar in many ways. For instance, density estimates under the two models are usually nearly indistinguishable (which we might expect since both models are consistent for the density).

Despite these similarities, there are at least two major differences between MFMs and the usual nonparametric mixture models (e.g., DPMs and Pitman–Yor process mixtures (PYMs)):

- (1) *The prior distribution of the number of clusters is very different.* In an MFM, one has complete control over the distribution of the number of components — leading to control over the distribution of the number of clusters — and (under the prior) as the sample size grows, the number of clusters converges to a finite value with probability 1 (but has no *a priori* bound). In a DPM or PYM (with discount parameter  $\sigma \in [0, 1)$ ), the distribution of the number of

clusters follows a particular parametric form, and (under the prior) the number of clusters grows to infinity with probability 1.

- (2) *Given the number of clusters, the prior distribution of the cluster sizes is very different.* In particular (under the prior), in an MFM the sizes of the clusters are all the same order of magnitude (asymptotically with respect to the number of samples), while in a DPM or PYM (with  $\sigma \in [0, 1)$ ) many of the clusters are negligible in size relative to the largest clusters.

These differences in the priors carry over to differences in certain aspects of the posteriors. Partitions sampled from the DPM posterior tend to have multiple small transient clusters, which are not seen in samples from the MFM posterior; consequently, the DPM posterior on the number of clusters tends to put more probability mass on higher numbers of clusters than the MFM. Further, given the number of clusters, the entropy of a partition sampled from the DPM posterior tends to be lower than one sampled from the MFM posterior.

One could argue that, compared to the DPM, the MFM is a more natural Bayesian approach for a data distribution of unknown complexity (if something is unknown, put a prior on it), and consequently, that it enjoys greater interpretability and conceptual simplicity. Further, in an MFM, the parameter space is a countable union of finite-dimensional spaces, while in a DPM or PYM, the parameter space is infinite-dimensional.

There are also some disadvantages to using MFMs. A minor inconvenience is that the coefficients of the partition distribution need to be precomputed when doing inference. A more significant issue is that the mixing time of incremental Markov chain Monte Carlo (MCMC) samplers can be worse than for the DPM, since the MFM

dislikes small clusters, making it difficult to create or destroy substantial clusters by moving one element at a time. A natural solution to this mixing issue would be to use a split-merge sampler.

In general, we would not say that one model is uniformly better than the other — the choice of whether to use an MFM or a DPM or PYM (or something else) should be made based on the appropriateness of the model to the application at hand.

#### 4.1.1 Mixture of finite mixtures (MFM)

Consider the following variable-dimension mixture model:

$$\begin{aligned}
 K &\sim p(k), \text{ where } p(k) \text{ is a p.m.f. on } \{1, 2, \dots\} \\
 (\pi_1, \dots, \pi_k) &\sim \text{Dirichlet}_k(\gamma, \dots, \gamma), \text{ given } K = k \\
 Z_1, \dots, Z_n &\stackrel{\text{iid}}{\sim} \pi, \text{ given } \pi \\
 \theta_1, \dots, \theta_k &\stackrel{\text{iid}}{\sim} H, \text{ given } K = k \\
 X_j &\sim f_{\theta_{Z_j}} \text{ independently for } j = 1, \dots, n, \text{ given } \theta_{1:K}, Z_{1:n}.
 \end{aligned} \tag{4.1.1}$$

Here,  $H$  is a prior or “base measure” on  $\Theta \subset \mathbb{R}^\ell$ , and  $\{f_\theta : \theta \in \Theta\}$  is a family of probability densities with respect to a sigma-finite measure  $\lambda$  on  $\mathcal{X} \subset \mathbb{R}^d$ . (As usual, we give  $\Theta$  and  $\mathcal{X}$  the Borel sigma-algebra.) In a typical application, the values  $X_1, \dots, X_n$  would be observed, and all other variables would be hidden/latent. See Figure 4.1 for a graphical model representation.

We refer to this as a *mixture of finite mixtures* (MFM) model, since there does not seem to be a well-established, unambiguous name. This type of model has been studied by many authors (Nobile, 1994, Phillips and Smith, 1996, Richardson

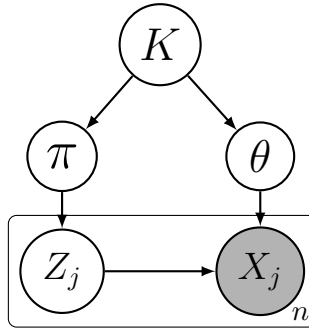


Figure 4.1: Graphical model for the MFM.

and Green, 1997, Green and Richardson, 2001, Stephens, 2000, Nobile and Fearnside, 2007) and has found many applications, including neuron spike classification (Nguyen et al., 2003), survival analysis (Marin et al., 2005), DNA restriction mapping (Lee et al., 1998), and gene expression profiling (Vogl et al., 2005).

The most common approach to inference in such a model is reversible jump Markov chain Monte Carlo (RJMCMC) (Richardson and Green, 1997, Green and Richardson, 2001), although other strategies have been proposed as well (Nobile, 1994, Phillips and Smith, 1996, Stephens, 2000, Nobile and Fearnside, 2007). Reversible jump is a very general and widely-applicable method, but it is not a “black box”. In contrast, a nice aspect of many DPM samplers is that they are fairly generic. In their paper developing a reversible jump sampler for the DPM, Green and Richardson (2001) pointed out that it would be interesting if, conversely, DPM-style samplers could be used for the MFM model:

*In view of the intimate correspondence between DP and DMA models discussed above, it is interesting to examine the possibilities of using either class of MCMC methods for the other model class. We have been unsuccessful in our search for incremental Gibbs samplers for the DMA models, but it turns out to be reasonably straightforward to implement reversible*

*jump split/merge methods for DP models.* (Green and Richardson, 2001)

Note: They seem to use the term Dirichlet-multinomial allocation (DMA) to refer both to the MFM model above with  $k$  random, and to the corresponding model with  $k$  fixed (the latter of which is the more standard usage, (Chen, 1980)); to avoid confusion, we say MFM instead of DMA for the  $k$  random case.

The key to many DPM samplers is that the model can be characterized by a nice distribution on combinatorial structures — namely, the Chinese restaurant process. It turns out, as we observe below, that MFMs have a similar characterization, enabling the application of the same sampling algorithms to them.

It is important to note that in the MFM, we assume a symmetric Dirichlet with a single parameter  $\gamma > 0$  not depending on  $k$ . This assumption is key to deriving a simple form for the partition probability distribution and the resulting restaurant process. Assuming symmetry in the distribution of  $\pi$  is quite natural, since the distribution of  $X_1, \dots, X_n$  under any asymmetric distribution on  $\pi$  would be the same as if this were replaced by its symmetrized version (e.g., if the entries of  $\pi$  were uniformly permuted). Assuming the same  $\gamma$  for all  $k$  is a genuine restriction, albeit a fairly natural one, often made in such models even when not strictly necessary (Nobile, 1994, Richardson and Green, 1997, Green and Richardson, 2001, Stephens, 2000, Nobile and Fearnside, 2007). Note that prior information about the relative sizes of the mixing weights  $\pi_1, \dots, \pi_k$  can be introduced through  $\gamma$  — roughly speaking, small  $\gamma$  favors lower entropy  $\pi$ 's, while large  $\gamma$  favors higher entropy  $\pi$ 's.

On the other hand, we will make very few assumptions on  $p(k)$ , the distribution of the number of components. For practical purposes, we need the infinite series  $\sum_{k=1}^{\infty} p(k)$  to converge to 1 reasonably quickly, although any choice of  $p(k)$  arising in



practice should not be a problem. For certain theoretical purposes — in particular, consistency for the number of components — it is desirable to have  $p(k) > 0$  for all  $\{1, 2, \dots\}$ .

This chapter is organized as follows. In Section 4.2, we derive a number of properties of the MFM model, including the partition distribution, restaurant process, random discrete measure formulation and other equivalent models, stick-breaking representation in a special case, density estimates, and various basic properties. In Section 4.3, we discuss existing results on the posterior asymptotics of the density, the mixing distribution, and the number of components, and we derive some asymptotic properties of the MFM model. In Section 4.4, we show how the properties derived in Section 4.2 facilitate inference with Markov chain Monte Carlo (MCMC) algorithms; in particular, we show how partition-based samplers for the DPM — in the case of both conjugate and non-conjugate priors — can be directly adapted to the MFM.

## 4.2 Properties of MFMs

### 4.2.1 Partition distribution

The primary observation on which our development relies is that the distribution on partitions induced by an MFM takes a form which is simple enough that it can be easily computed. Consider the MFM model defined above. Let  $\mathcal{C}$  denote the unordered partition of  $[n] := \{1, \dots, n\}$  induced by  $Z_1, \dots, Z_n$ ; in other words,  $\mathcal{C} = \{E_i : |E_i| > 0\}$  where  $E_i = \{j : Z_j = i\}$  for  $i = 1, 2, \dots$ . Then, as we show

below,

$$p(\mathcal{C}) = V_n(t) \prod_{c \in \mathcal{C}} \gamma^{(|c|)} \quad (4.2.1)$$

where  $t = |\mathcal{C}|$  is the number of parts in the partition,

$$V_n(t) = \sum_{k=1}^{\infty} \frac{k^{(t)}}{(\gamma k)^{(n)}} p(k), \quad (4.2.2)$$

$x^{(m)} = x(x+1) \cdots (x+m-1)$ , and  $x_{(m)} = x(x-1) \cdots (x-m+1)$ . (By convention,  $x^{(0)} = 1$  and  $x_{(0)} = 1$ .)

Viewed as a function of the part sizes ( $|c| : c \in \mathcal{C}$ ), Equation 4.2.1 is an “exchangeable partition probability function” (EPPF) in the terminology of Pitman (1995, 2006) — that is, it is a symmetric function of the part sizes (and depends on  $\mathcal{C}$  only through  $t = |\mathcal{C}|$  and the part sizes). Consequently,  $\mathcal{C}$  is an *exchangeable random partition* of  $[n]$ ; that is, its distribution is invariant under permutations of  $[n]$  (alternatively, this can be seen directly from the definition of the model, since  $Z_1, \dots, Z_n$  are exchangeable). More specifically, Equation 4.2.1 is a member of the family of Gibbs partition distributions (Pitman, 2006). This can be viewed as a special case of the general results of Gnedin and Pitman (2006) characterizing the extremal points of the space of Gibbs partition distributions; also see Lijoi et al. (2008), Ho et al. (2007), Cerquetti (2008, 2011), Gnedin (2010), and Lijoi and Prünster (2010) for further results on Gibbs partitions. However, the utility of this representation for inference in a wide variety of variable-dimension mixture models does not seem to have been previously explored in the literature.

The derivation of Equation 4.2.1 is straightforward, and we provide it here. Letting  $E_i = \{j : z_j = i\}$  as above, and writing  $\mathcal{C}(z)$  for the partition induced by

$z = (z_1, \dots, z_n)$ , by Dirichlet-multinomial conjugacy we have

$$p(z|k) = \int p(z|\pi)p(\pi|k)d\pi = \frac{\Gamma(k\gamma) \prod_{i=1}^k \Gamma(|E_i| + \gamma)}{\Gamma(\gamma)^k \Gamma(n + k\gamma)} = \frac{1}{(k\gamma)^{(n)}} \prod_{c \in \mathcal{C}(z)} \gamma^{(|c|)},$$

provided that  $p(k) > 0$ . It follows that for any partition  $\mathcal{C}$  of  $[n]$ ,

$$\begin{aligned} p(\mathcal{C}|k) &= \sum_{z \in [k]^n : \mathcal{C}(z) = \mathcal{C}} p(z|k) \\ &= \#\{z \in [k]^n : \mathcal{C}(z) = \mathcal{C}\} \frac{1}{(\gamma k)^{(n)}} \prod_{c \in \mathcal{C}} \gamma^{(|c|)} \\ &= \frac{k^{(t)}}{(\gamma k)^{(n)}} \prod_{c \in \mathcal{C}} \gamma^{(|c|)}, \end{aligned} \tag{4.2.3}$$

where  $t = |\mathcal{C}|$ , since  $\#\{z \in [k]^n : \mathcal{C}(z) = \mathcal{C}\} = \binom{k}{t} t! = k^{(t)}$ . Finally,

$$p(\mathcal{C}) = \sum_{k=1}^{\infty} p(\mathcal{C}|k)p(k) = \left( \prod_{c \in \mathcal{C}} \gamma^{(|c|)} \right) \sum_{k=1}^{\infty} \frac{k^{(t)}}{(\gamma k)^{(n)}} p(k) = V_n(t) \prod_{c \in \mathcal{C}} \gamma^{(|c|)},$$

with  $V_n(t)$  as in Equation 4.2.2 above.

## 4.2.2 Equivalent models

For various purposes, it useful to write the model (Equation 4.1.1) in various equivalent ways. First, note that instead of sampling only  $\theta_1, \dots, \theta_k \stackrel{\text{iid}}{\sim} H$  given  $K = k$ , we could simply sample  $\theta_1, \theta_2, \dots \stackrel{\text{iid}}{\sim} H$  independently of  $K$ , and the distribution of  $X_{1:n}$  would be the same; the graphical model for this version is in Figure 4.2(a).

Now,  $Z_{1:n}$  determines which subset of the i.i.d. variables  $\theta_1, \theta_2, \dots$  will actually be used, and the indices of this subset are independent of  $\theta_{1:\infty}$ ; hence, denoting these random indices  $I_1 < \dots < I_T$ , we have that  $\theta_{I_1}, \dots, \theta_{I_T} | Z_{1:n}$  are i.i.d. from  $H$ .

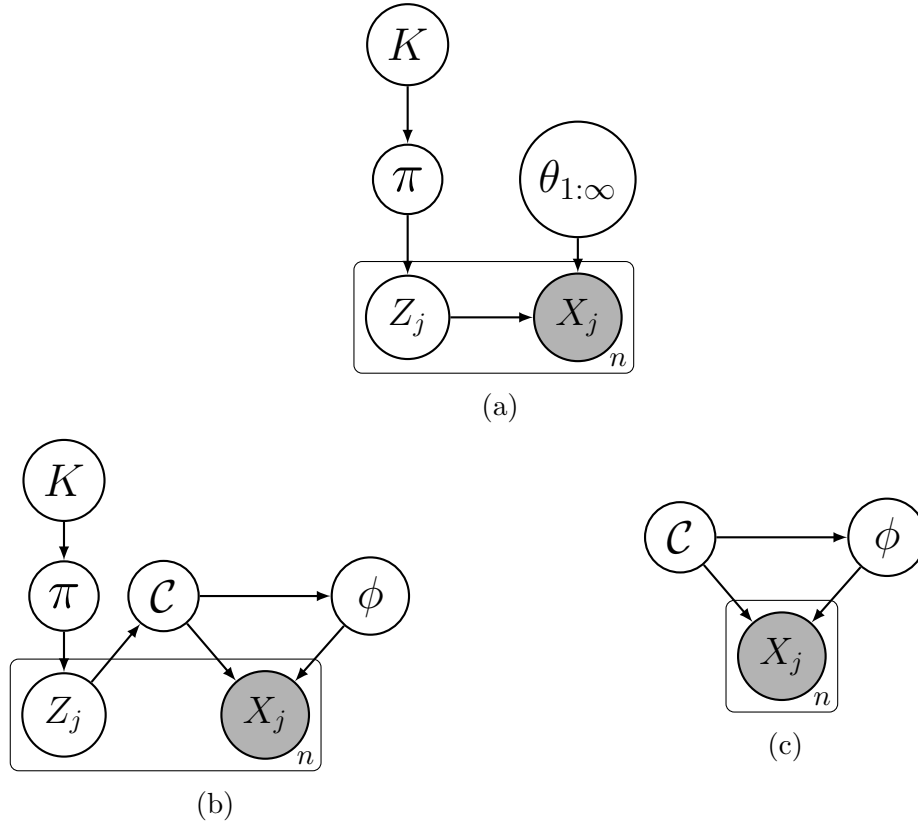


Figure 4.2: Alternative graphical models for the MFM.

From this we see that we could equivalently write the model in terms of  $\mathcal{C}$  as follows, resulting in the same distribution for  $\mathcal{C}$  and  $X_{1:n}$  (see Figure 4.2(b,c)):

$$\begin{aligned}
 \mathcal{C} &\sim p(\mathcal{C}), \text{ with } p(\mathcal{C}) \text{ as in Equation 4.2.1} \\
 \phi_c &\stackrel{\text{iid}}{\sim} H \text{ for } c \in \mathcal{C}, \text{ given } \mathcal{C} \\
 X_j &\sim f_{\phi_c} \text{ independently for } j \in c, c \in \mathcal{C}, \text{ given } \phi, \mathcal{C}.
 \end{aligned}
 \tag{4.2.4}$$

Note that for notational convenience, here,  $\phi = (\phi_c : c \in \mathcal{C})$  is a tuple of  $t = |\mathcal{C}|$  parameters  $\phi_c \in \Theta$ , one for each part  $c \in \mathcal{C}$ .

This representation of the model is particularly useful for doing inference, since one does not have to deal with cluster labels or empty components. The formulation of models starting from a partition distribution has been a fruitful approach, exem-

plified by the development of product partition models (Quintana and Iglesias, 2003, Quintana, 2006, Dahl, 2009, Park and Dunson, 2010, Müller et al., 2011, Barry and Hartigan, 1992); also see Müller and Quintana (2010) for a review.

### 4.2.3 Various basic properties

We list here some basic properties of the MFM model above. The derivations are all straightforward, but for completeness we provide them in Section 4.2.10.1. Denoting  $x_c = (x_j : j \in c)$  and  $m(x_c) = \int_{\Theta} [\prod_{j \in c} f_{\theta}(x_j)] H(d\theta)$  (with the convention that  $m(x_{\emptyset}) = 1$ ), we have

$$p(x_{1:n}|\mathcal{C}) = \prod_{c \in \mathcal{C}} m(x_c). \quad (4.2.5)$$

The number of components  $K$  and the number of clusters  $T = |\mathcal{C}|$  are related by

$$p(t|k) = \frac{k^{(t)}}{(\gamma k)^{(n)}} \sum_{\mathcal{C}:|\mathcal{C}|=t} \prod_{c \in \mathcal{C}} \gamma^{(|c|)}, \quad (4.2.6)$$

$$p(k|t) = \frac{1}{V_n(t)} \frac{k^{(t)}}{(\gamma k)^{(n)}} p(k), \quad (4.2.7)$$

where in  $p(t|k)$ , the sum is over partitions  $\mathcal{C}$  of  $[n]$  such that  $|\mathcal{C}| = t$ . The formula for  $p(k|t)$  is required for doing inference about the number of components  $K$  based on posterior samples of  $\mathcal{C}$ ; fortunately, it is easy to compute. We have the conditional independence relations

$$\mathcal{C} \perp K \mid T, \quad (4.2.8)$$

$$X_{1:n} \perp K \mid T. \quad (4.2.9)$$

#### 4.2.4 The coefficients $V_n(t)$

Recall that  $p(\mathcal{C}) = V_n(t) \prod_{c \in \mathcal{C}} \gamma^{(|c|)}$ , where  $t = |\mathcal{C}|$  and

$$V_n(t) = \sum_{k=1}^{\infty} \frac{k_{(t)}}{(\gamma k)^{(n)}} p(k).$$

The numbers  $V_n(t)$  satisfy the recursion

$$V_{n+1}(t+1) = V_n(t)/\gamma - (n/\gamma + t)V_{n+1}(t) \quad (4.2.10)$$

for any  $0 \leq t \leq n$  and  $\gamma > 0$ ; this is easily seen by plugging the identity

$$k_{(t+1)} = (\gamma k + n)k_{(t)}/\gamma - (n/\gamma + t)k_{(t)}$$

into the expression for  $V_{n+1}(t+1)$ . This recursion is a special case of a more general recursion for a large class of partition distributions ([Gnedin and Pitman, 2006](#)). In the case of  $\gamma = 1$ , [Gnedin \(2010\)](#) has given a nice example of a distribution on  $K$  for which both  $p(k)$  and  $V_n(t)$  have closed-form expressions.

In previous work on the MFM model, it has been common for  $p(k)$  to be chosen to be proportional to a Poisson distribution restricted to a subset of the positive integers ([Phillips and Smith, 1996](#), [Stephens, 2000](#), [Nobile and Fearnside, 2007](#)), and [Nobile \(2005\)](#) has proposed a theoretical justification for this choice. Interestingly, the model has some nice mathematical properties if one instead chooses  $K - 1$  to be given a Poisson distribution, that is,  $p(k) = \text{Poisson}(k - 1|\lambda)$  for some  $\lambda > 0$ . One example of this arises here (for another example, see [Section 4.2.9](#)): it turns out that

if  $p(k) = \text{Poisson}(k-1|\lambda)$  and  $\gamma = 1$  then (see Section 4.2.10.2 for details)

$$V_n(0) = \frac{1}{\lambda^n} \left( 1 - \sum_{k=1}^n p(k) \right). \quad (4.2.11)$$

However, to do inference, it is not necessary to choose  $p(k)$  to have any particular form. To do inference, we just need to be able to compute  $p(\mathcal{C})$ , and in turn, we need to be able to compute  $V_n(t)$ . To this end, note that  $k_{(t)}/(\gamma k)^n \leq k^t/(\gamma k)^n$ , and thus the infinite series for  $V_n(t)$  converges rapidly when  $t \ll n$ . It always converges to a finite value when  $1 \leq t \leq n$ ; this is clear from the fact that  $p(\mathcal{C}) \in [0, 1]$ . This finiteness can also be seen directly from the series since  $k^t/(\gamma k)^n \leq 1/\gamma^n$ , and in fact, this shows that the series for  $V_n(t)$  converges at least as rapidly (up to a constant) as the series  $\sum_{k=1}^{\infty} p(k)$  converges to 1.

Hence, for any reasonable choice of  $p(k)$  (i.e., not having an extraordinarily heavy tail),  $V_n(t)$  can easily be numerically approximated to a high level of precision. For instance, in Chapter 5, we apply the model using

$$p(k) \propto \begin{cases} 1 & \text{if } k \in \{1, \dots, 30\} \\ 1/(k-30)^2 & \text{if } k > 30, \end{cases}$$

and computing the coefficients  $V_n(t)$  presents no problems, even though this has a fairly heavy tail (the mean is infinite) and does not take an analytic form. Note that  $V_n(t)$  is often exceedingly small, making it necessary to use a standard technique such as the “log-sum-exp trick” in order to avoid numerical underflow (that is, represent each term in log space and use  $\log(\exp(a)+\exp(b)) = \log(\exp(a-m)+\exp(b-m))+m$ , where  $m = \max\{a, b\}$ , to compute the log of the sum).

Although we find it easier to use this numerical approach for all  $V_n(t)$  values, it

would also be possible to use the recursion in Equation 4.2.10 (taking care, again, regarding underflow) and numerically approximate only the required values of  $V_n(0)$  (or use the expression for  $V_n(0)$  in Equation 4.2.11 if  $K - 1 \sim \text{Poisson}(\lambda)$  and  $\gamma = 1$ ).

In the course of doing inference,  $n$  is usually fixed and the MCMC sampling methods we describe below typically only require one to compute  $V_n(t)$  for  $t = 1, \dots, t_{\max}$  for some relatively small  $t_{\max}$ , since higher values of  $t$  have such low probability that in practice they are never visited by the sampler. Consequently, it suffices to precompute  $V_n(1), \dots, V_n(t_{\max})$  for some suitably chosen  $t_{\max}$ , and this takes a negligible amount of time compared to running the sampler.

Also, on a practical note, the samplers described below involve moving one element at a time between clusters, so in order to achieve irreducibility of the Markov chain, it is necessary to have  $\{t \in \{1, 2, \dots\} : V_n(t) > 0\}$  be a block of consecutive integers. In fact, it turns out that this is always the case (and this block includes  $t = 1$ ), since for any  $k$  such that  $p(k) > 0$ , we have  $V_n(t) > 0$  for all  $t = 1, \dots, k$ .

When we come to the restaurant process in Section 4.2.6 below, it will become clear that precomputing the  $V_n(t)$ 's effectively amounts to precomputing the probabilities for “choosing a new table”. The idea of precomputing these probabilities has previously been applied to DPMS with a prior on the concentration parameter  $\alpha$  (MacEachern, 1998), in which the pre-computation consists of numerically integrating  $\alpha$  out.

In the MFM model as described above (Equation 4.1.1), the Dirichlet parameter  $\gamma$  is fixed, but in some cases it might be interesting to put a prior on  $\gamma$ , to allow greater flexibility regarding the relative sizes of the components. (Thanks to Vinayak Rao for this suggestion.) Although we have not tried this, it should be straightforward to



periodically update  $\gamma$  with a Metropolis-Hastings step and recompute the required  $V_n(t)$  coefficients, possibly using a simple caching scheme.

### 4.2.5 Self-consistent marginals

For each  $n = 1, 2, \dots$ , let  $q_n(\mathcal{C})$  denote the distribution on partitions of  $[n]$  as defined above (Equation 4.2.1). This family of partition distributions has the property that  $q_m$  coincides with the “marginal” distribution on partitions of  $[m]$  induced by  $q_n$  when  $n \geq m$ ; in other words, drawing a sample from  $q_n$  and removing elements  $m + 1, \dots, n$  from it yields a sample from  $q_m$ . This can be seen directly from the model definition (Equation 4.1.1), since  $\mathcal{C}$  is the partition induced by the  $Z$ ’s, and the distribution of  $Z_{1:m}$  is the same when the model is defined with any  $n \geq m$ . This property is sometimes referred to as *consistency in distribution* (Pitman, 2006).

Using Kolmogorov’s extension theorem (e.g., Durrett, 1996), one can show that this implies the existence of a unique probability distribution on partitions of the positive integers  $\mathbb{Z}_{>0} = \{1, 2, \dots\}$  such that the marginal distribution on partitions of  $[n]$  is  $q_n$  for all  $n = 1, 2, \dots$ . A random partition of  $\mathbb{Z}_{>0}$  from such a distribution is a *combinatorial stochastic process*; for background, see Pitman (2006).

It is instructive to also verify this self-consistency property for MFMs by the more tedious approach of directly marginalizing. Suppose  $\mathcal{C}$  is a partition of  $[n]$ , denote by  $\mathcal{C}_*$  the partition of  $[n + 1]$  obtained from  $\mathcal{C}$  by making  $n + 1$  a singleton, and for  $c \in \mathcal{C}$  denote by  $\mathcal{C}_c$  the partition of  $[n + 1]$  obtained from  $\mathcal{C}$  by adding  $n + 1$  to  $c$ . We would like to show that  $q_{n+1}(\mathcal{C}_*) + \sum_{c \in \mathcal{C}} q_{n+1}(\mathcal{C}_c) = q_n(\mathcal{C})$ , and indeed, letting

$$t = |\mathcal{C}|,$$

$$\begin{aligned} & V_{n+1}(t+1)\gamma \prod_{c' \in \mathcal{C}} \gamma^{(|c'|)} + \sum_{c \in \mathcal{C}} V_{n+1}(t)(\gamma + |c|) \prod_{c' \in \mathcal{C}} \gamma^{(|c'|)} \\ &= \left( V_{n+1}(t+1)\gamma + V_{n+1}(t)(t\gamma + n) \right) \left( \prod_{c \in \mathcal{C}} \gamma^{(|c|)} \right) = V_n(t) \prod_{c \in \mathcal{C}} \gamma^{(|c|)}, \end{aligned}$$

the last step following from the recursion, Equation 4.2.10. The property follows by induction.

## 4.2.6 Restaurant process / Pólya urn scheme

Pitman (1996) considered a general class of “restaurant processes”, or Pólya urn schemes, corresponding to exchangeable partition probability functions (EPPFs). Following this general rule, we can derive a restaurant process for the MFM.

For any vector of random variables  $(Y_1, \dots, Y_n)$ , we can obtain a sample by first drawing  $Y_1$ , then  $Y_2|Y_1$ , and so on, up to  $Y_n|Y_{n-1}, \dots, Y_1$ . In the same way, if we write  $\mathcal{C}_1, \dots, \mathcal{C}_n$  for the partitions of  $[1], \dots, [n]$ , respectively, induced by a partition  $\mathcal{C}_n$  of  $[n]$ , then for any distribution on partitions of  $[n]$ , we can obtain a sample  $\mathcal{C}_n$  by sampling  $\mathcal{C}_1, \mathcal{C}_2|\mathcal{C}_1$ , and so on. So, in a crude sense, any distribution on partitions has a “restaurant process” describing this sequence of conditional distributions, but it may depend on  $n$  and it may be quite complicated.

However, since we are in the fortunate circumstance that our partition distribution has the self-consistency property described in the previous section, when  $\mathcal{C}_n \sim q_n$ , we have  $\mathcal{C}_m \sim q_m$  for  $m = 1, \dots, n$  (where  $q_n$  denotes the distribution on partitions as in Equation 4.2.1). Consequently, there is a single restaurant process

that works for all  $n$ , and it takes a simple form involving  $q_m$  at step  $m$ , since

$$p(\mathcal{C}_m | \mathcal{C}_{m-1}, \dots, \mathcal{C}_1) = p(\mathcal{C}_m | \mathcal{C}_{m-1}) \propto q_m(\mathcal{C}_m) I(\mathcal{C}_m \setminus m = \mathcal{C}_{m-1}),$$

where  $\mathcal{C} \setminus m$  denotes  $\mathcal{C}$  with element  $m$  removed.

Recalling that  $q_m(\mathcal{C}_m) = V_m(|\mathcal{C}_m|) \prod_{c \in \mathcal{C}_m} \gamma^{(|c|)}$ , we have, letting  $t = |\mathcal{C}_{m-1}|$ ,

$$p(\mathcal{C}_m | \mathcal{C}_{m-1}) \propto \begin{cases} V_m(t+1)\gamma & \text{if } m \text{ is a singleton in } \mathcal{C}_m, \text{ i.e., } \{m\} \in \mathcal{C}_m \\ V_m(t)(\gamma + |c|) & \text{if } c \in \mathcal{C}_{m-1} \text{ and } c \cup \{m\} \in \mathcal{C}_m, \end{cases}$$

for  $\mathcal{C}_m$  such that  $\mathcal{C}_m \setminus m = \mathcal{C}_{m-1}$  (and  $p(\mathcal{C}_m | \mathcal{C}_{m-1}) = 0$  otherwise).

In other words, we have the following restaurant process for the MFM:

- The first customer sits at a table:  $\mathcal{C}_1 = \{\{1\}\}$ .
- For  $n = 2, 3, \dots$ , the  $n$ th customer sits ...

at an existing table  $c \in \mathcal{C}_{n-1}$  with probability  $\propto |c| + \gamma$

at a new table with probability  $\propto \frac{V_n(t+1)}{V_n(t)}\gamma$

where  $t = |\mathcal{C}_{n-1}|$ .

Clearly, this bears a resemblance to the Chinese restaurant process, in which the  $n$ th customer sits at an existing table  $c$  with probability  $\propto |c|$  or at a new table with probability  $\propto \alpha$  (the concentration parameter) ([Blackwell and MacQueen, 1973](#)).



Figure 4.3: Graphical models for (a) the random discrete measure formulation, and (b) the equivalent construction of the distribution of  $\beta_1, \dots, \beta_n$ .

### 4.2.7 Random discrete measure formulation

The MFM can also be formulated starting from a distribution on discrete measures that is analogous to the Dirichlet process. Let

$$K \sim p_K(k)$$

$$(\pi_1, \dots, \pi_k) \sim \text{Dirichlet}_k(\gamma, \dots, \gamma), \text{ given } K = k$$

$$\theta_1, \dots, \theta_k \stackrel{\text{iid}}{\sim} H, \text{ given } K = k$$

$$G = \sum_{i=1}^K \pi_i \delta_{\theta_i}$$

and denote the distribution of  $G$  by  $\mathcal{M}(p_K, \gamma, H)$ . Then the distribution of  $X_{1:n}$  is the same as before if we take  $X_1, \dots, X_n | G$  i.i.d. from the resulting mixture, namely,

$$f_G(x) := \int f_\theta(x) G(d\theta) = \sum_{i=1}^K \pi_i f_{\theta_i}(x).$$

So, in this notation, the model is (Figure 4.3(a)):

$$G \sim \mathcal{M}(p_K, \gamma, H)$$

$$X_1, \dots, X_n \sim f_G, \text{ given } G.$$

When  $H$  is continuous, the random discrete measure  $G$  belongs to the class of “species sampling” models studied by Pitman (1996), who proved many properties regarding this type of distribution. For completeness, we derive the prediction rule for  $G$ , along with some other properties. We refer to Pitman (1996), Hansen and Pitman (2000), Ishwaran and James (2003, 2001), Lijoi et al. (2005a, 2007), and Lijoi et al. (2008) for more background on species sampling models and further examples.

Let  $G \sim \mathcal{M}(p_K, \gamma, H)$  and let  $\beta_1, \dots, \beta_n \stackrel{\text{iid}}{\sim} G$ , given  $G$ . Then, by construction, the joint distribution of  $(\beta_1, \dots, \beta_n)$  (with  $G$  marginalized out) is the same as  $(\theta_{Z_1}, \dots, \theta_{Z_n})$  in the original model (Equation 4.1.1). Further, letting  $\mathcal{C}$  denote the partition induced by  $Z_1, \dots, Z_n$  as usual, we have  $(\theta_{Z_1}, \dots, \theta_{Z_n}) = (\phi_{c_1}, \dots, \phi_{c_n})$  where  $c_j$  is defined to be the  $c \in \mathcal{C}$  such that  $j \in c$ , and  $\phi_c$  is defined to be equal to the  $\theta_i$  such that  $Z_j = i$  for all  $j \in c$ . By the same argument as in Section 4.2.2,  $(\phi_c : c \in \mathcal{C})$  are i.i.d. from  $H$  given  $\mathcal{C}$ .

Therefore, we have the following equivalent construction for  $(\beta_1, \dots, \beta_n)$ :

$$\mathcal{C}_n \sim q_n, \text{ with } q_n \text{ as in Section 4.2.5}$$

$$\phi_c \stackrel{\text{iid}}{\sim} H \text{ for } c \in \mathcal{C}_n, \text{ given } \mathcal{C}_n$$

$$\beta_j = \phi_c \text{ for } j \in c, c \in \mathcal{C}_n, \text{ given } \mathcal{C}_n, \phi.$$

See Figure 4.3(b). Due to the self-consistency property of  $q_1, q_2, \dots$ , we can sample  $\mathcal{C}_n, (\phi_c : c \in \mathcal{C}_n), \beta_{1:n}$  sequentially for  $n = 1, 2, \dots$  by sampling from the restaurant process for  $\mathcal{C}_n | \mathcal{C}_{n-1}$ , sampling  $\phi_{\{n\}}$  from  $H$  if  $n$  is seated at a new table (or setting  $\phi_{c \cup \{n\}} = \phi_c$  if  $n$  is seated at  $c \in \mathcal{C}_{n-1}$ ), and setting  $\beta_n$  accordingly.

In particular, if the base measure  $H$  is continuous (that is, if  $H(\{\theta\}) = 0$  for all  $\theta \in \Theta$ ), then conditioning on  $\beta_{1:n-1}$  is the same as conditioning on  $\mathcal{C}_{n-1}, (\phi_c :$

$c \in \mathcal{C}_{n-1}$ ),  $\beta_{1:n-1}$ , so we can sample  $\beta_n | \beta_{1:n-1}$  in the same way as was just described. Therefore, when  $H$  is continuous, the distribution of  $\beta_n$  given  $\beta_1, \dots, \beta_{n-1}$  is proportional to

$$\gamma \frac{V_n(t+1)}{V_n(t)} H + \sum_{c \in \mathcal{C}_{n-1}} (|c| + \gamma) \delta_{\phi_c}, \quad (4.2.12)$$

where  $t = |\mathcal{C}_{n-1}|$ , with the  $\mathcal{C}_{n-1}$  and  $(\phi_c : c \in \mathcal{C}_{n-1})$  induced by  $\beta_1, \dots, \beta_{n-1}$ , or equivalently,

$$\gamma \frac{V_n(t+1)}{V_n(t)} H + \sum_{j=1}^{n-1} \delta_{\beta_j} + \gamma \sum_{j=1}^t \delta_{\beta_j^*},$$

where  $\beta_1^*, \dots, \beta_t^*$  are the distinct values taken by  $\beta_1, \dots, \beta_{n-1}$ . For comparison, if  $G \sim \text{DP}(\alpha, H)$  instead, the distribution of  $\beta_n$  given  $\beta_1, \dots, \beta_{n-1}$  is proportional to

$$\alpha H + \sum_{j=1}^{n-1} \delta_{\beta_j}.$$

This is the classic Pólya urn scheme for the Dirichlet process, described by [Blackwell and MacQueen \(1973\)](#).

### 4.2.8 Density estimates

In this section, we derive formulas for density estimation with the MFM using samples of  $\mathcal{C}, \phi | x_{1:n}$ , or, if the single-cluster marginals can be computed, using samples of  $\mathcal{C} | x_{1:n}$ . Using the restaurant process (at the end of Section 4.2.6 above), it is straightforward to show that, if  $\mathcal{C}$  is a partition of  $[n]$  and  $\phi = (\phi_c : c \in \mathcal{C})$  then

$$p(x_{n+1} | \mathcal{C}, \phi, x_{1:n}) \propto \frac{V_{n+1}(t+1)}{V_{n+1}(t)} \gamma m(x_{n+1}) + \sum_{c \in \mathcal{C}} (|c| + \gamma) f_{\phi_c}(x_{n+1}) \quad (4.2.13)$$

where  $t = |\mathcal{C}|$ , and, using the recursion for  $V_n(t)$  (Equation 4.2.10), this is normalized when multiplied by  $V_{n+1}(t)/V_n(t)$ . Further,

$$p(x_{n+1} \mid \mathcal{C}, x_{1:n}) \propto \frac{V_{n+1}(t+1)}{V_{n+1}(t)} \gamma m(x_{n+1}) + \sum_{c \in \mathcal{C}} (|c| + \gamma) \frac{m(x_{c \cup \{n+1\}})}{m(x_c)}, \quad (4.2.14)$$

with the same normalization constant. Therefore, when the single-cluster marginals  $m(x_c)$  can be easily computed, Equation 4.2.14 can be used to estimate the posterior predictive density  $p(x_{n+1} \mid x_{1:n})$  based on samples from  $\mathcal{C} \mid x_{1:n}$ . When  $m(x_c)$  cannot be easily computed, Equation 4.2.13 can be used to estimate  $p(x_{n+1} \mid x_{1:n})$  based on samples from  $\mathcal{C}, \phi \mid x_{1:n}$ , along with samples  $\theta_1, \dots, \theta_N \stackrel{\text{iid}}{\sim} H$  to approximate  $m(x_{n+1}) \approx \frac{1}{N} \sum_{i=1}^N f_{\theta_i}(x_{n+1})$ .

The posterior predictive density is, perhaps, the most natural estimate of the density. However, following [Green and Richardson \(2001\)](#), another way to obtain a fairly natural estimate is by assuming that element  $n+1$  is added to an existing cluster; this will be very similar to the posterior predictive density when  $n$  is sufficiently large. To this end, we define  $p_*(x_{n+1} \mid \mathcal{C}, \phi, x_{1:n}) = p(x_{n+1} \mid \mathcal{C}, \phi, x_{1:n}, |\mathcal{C}_{n+1}| = |\mathcal{C}|)$ , where  $\mathcal{C}_{n+1}$  is the partition of  $[n+1]$ , and observe that

$$p_*(x_{n+1} \mid \mathcal{C}, \phi, x_{1:n}) = \sum_{c \in \mathcal{C}} \frac{|c| + \gamma}{n + \gamma t} f_{\phi_c}(x_{n+1})$$

where  $t = |\mathcal{C}|$  ([Green and Richardson, 2001](#)). Using this, we can estimate the density by

$$\frac{1}{N} \sum_{i=1}^N p_*(x_{n+1} \mid \mathcal{C}^{(i)}, \phi^{(i)}, x_{1:n}), \quad (4.2.15)$$

where  $(\mathcal{C}^{(1)}, \phi^{(1)}), \dots, (\mathcal{C}^{(N)}, \phi^{(N)})$  are samples from  $\mathcal{C}, \phi \mid x_{1:n}$ .

The corresponding expressions for the DPM are all very similar, with the obvious changes, using its restaurant process instead.

These formulas are conditional on additional parameters such as  $\gamma$  for the MFM, and  $\alpha$  for the DPM. If priors are placed on such parameters and they are sampled along with  $\mathcal{C}$  and  $\phi$  given  $x_{1:n}$ , then the posterior predictive density can be estimated using the same formulas as above, but also using the posterior samples of these additional parameters.

### 4.2.9 Stick-breaking representation

The Dirichlet process has an elegant stick-breaking representation for the mixture weights  $\pi_1, \pi_2, \dots$  (Sethuraman, 1994, Sethuraman and Tiwari, 1981). This extraordinarily clarifying perspective has inspired a number of other nonparametric models (MacEachern et al., 1999, MacEachern, 2000, Hjort, 2000, Ishwaran and Zarepour, 2000, Ishwaran and James, 2001, Griffin and Steel, 2006, Dunson and Park, 2008, Chung and Dunson, 2009, Rodriguez and Dunson, 2011, Broderick et al., 2012), has provided insight into the properties of other models (Favaro et al., 2012, Teh et al., 2007, Thibaux and Jordan, 2007, Paisley et al., 2010), and has been used to develop efficient inference algorithms (Ishwaran and James, 2001, Blei and Jordan, 2006, Papaspiliopoulos and Roberts, 2008, Walker et al., 2007, Kalli et al., 2011).

In a certain special case — namely, when  $p(k) = \text{Poisson}(k - 1|\lambda)$  and  $\gamma = 1$  — we have noticed that the MFM also has a nice stick-breaking representation. (There may also be other cases of which we are not yet aware.) This is another example of the nice mathematical properties resulting from this choice of  $p(k)$  and  $\gamma$ . Consider the following “stick-breaking” procedure:



*Take a unit-length stick, and break off i.i.d. Exponential( $\lambda$ ) pieces until you run out of stick.*

In other words, let  $\varepsilon_1, \varepsilon_2, \dots \stackrel{\text{iid}}{\sim} \text{Exponential}(\lambda)$ , define  $\tilde{K} = \min\{j : \sum_{i=1}^j \varepsilon_i > 1\}$ , and set  $\tilde{\pi}_i = \varepsilon_i$  for  $i = 1, \dots, \tilde{K} - 1$  and  $\tilde{\pi}_{\tilde{K}} = 1 - \sum_{i=1}^{\tilde{K}-1} \tilde{\pi}_i$ . Then the resulting stick lengths  $\tilde{\pi}$  have the same distribution as the mixture weights  $\pi$  in the MFM model when  $p(k) = \text{Poisson}(k - 1 | \lambda)$  and  $\gamma = 1$ . This is a consequence of a standard construction of a Poisson process; for details, see Section 4.2.10.3.

This suggests an alternative way of constructing a variable-dimension mixture: take any sequence of nonnegative random variables  $(\varepsilon_1, \varepsilon_2, \dots)$  (not necessarily independent or identically distributed) such that  $\sum_{i=1}^{\infty} \varepsilon_i > 1$  with probability 1, and define  $\tilde{K}$  and  $\tilde{\pi}$  as above. Although the distribution of  $\tilde{K}$  and  $\tilde{\pi}$  may be complicated, in some cases it might still be possible to do inference based on the stick-breaking representation. This might be an interesting way to introduce different kinds of prior information on the mixture weights, however, we have not explored this possibility.

## 4.2.10 Proofs and details

### 4.2.10.1 Derivation of various basic properties

Here, we derive the properties listed in Section 4.2.3. Abbreviate  $x = x_{1:n}$ ,  $z = z_{1:n}$ , and  $\theta = \theta_{1:k}$ , and assume  $p(z, k) > 0$ . Then  $p(x|\theta, z, k) = \prod_{i=1}^k \prod_{j \in E_i} f_{\theta_i}(x_j)$  and

$$\begin{aligned} p(x|z, k) &= \int_{\Theta^k} p(x|\theta, z, k) p(d\theta|k) = \prod_{i=1}^k \int_{\Theta} \left[ \prod_{j \in E_i} f_{\theta_i}(x_j) \right] H(d\theta_i) \\ &= \prod_{i=1}^k m(x_{E_i}) = \prod_{c \in \mathcal{C}(z)} m(x_c). \end{aligned}$$

Since this last expression depends only on  $z, k$  through  $\mathcal{C} = \mathcal{C}(z)$ , we have

$$\prod_{c \in \mathcal{C}} m(x_c) = p(x|z, k) = p(x|z) = p(x|\mathcal{C}),$$

establishing Equation 4.2.5. Next, recall that  $p(\mathcal{C}|k) = \frac{k^{(t)}}{(\gamma k)^{(n)}} \prod_{c \in \mathcal{C}} \gamma^{(|c|)}$  (where  $t = |\mathcal{C}|$ ) from Equation 4.2.3, and thus

$$p(t|k) = \sum_{\mathcal{C}: |\mathcal{C}|=t} p(\mathcal{C}|k) = \frac{k^{(t)}}{(\gamma k)^{(n)}} \sum_{\mathcal{C}: |\mathcal{C}|=t} \prod_{c \in \mathcal{C}} \gamma^{(|c|)},$$

(where the sum is over partitions  $\mathcal{C}$  of  $[n]$  such that  $|\mathcal{C}| = t$ ) establishing Equation 4.2.6. Equation 4.2.7 follows, since

$$p(k|t) \propto p(t|k)p(k) \propto \frac{k^{(t)}}{(\gamma k)^{(n)}} p(k),$$

(provided  $p(t) > 0$ ) and the normalizing constant is precisely  $V_n(t)$ . To see that  $\mathcal{C} \perp K \mid T$  (Equation 4.2.8), note that if  $t = |\mathcal{C}|$  then

$$p(\mathcal{C}|t, k) = \frac{p(\mathcal{C}, t|k)}{p(t|k)} = \frac{p(\mathcal{C}|k)}{p(t|k)},$$

(provided  $p(t, k) > 0$ ) and due to the form of  $p(\mathcal{C}|k)$  and  $p(t|k)$  just above, this quantity does not depend on  $k$ ; hence,  $p(\mathcal{C}|t, k) = p(\mathcal{C}|t)$ . To see that  $X \perp K \mid T$  (Equation 4.2.9), note that from the graphical model in Figure 4.2(b),  $X \perp K \mid \mathcal{C}$ ; using this in addition to  $\mathcal{C} \perp K \mid T$ , we have

$$p(x|t, k) = \sum_{\mathcal{C}:|\mathcal{C}|=t} p(x|\mathcal{C}, t, k)p(\mathcal{C}|t, k) = \sum_{\mathcal{C}:|\mathcal{C}|=t} p(x|\mathcal{C}, t)p(\mathcal{C}|t) = p(x|t).$$

#### 4.2.10.2 $V_n(0)$ in the Poisson case with $\gamma = 1$

To see Equation 4.2.11, observe that when  $p(k) = \text{Poisson}(k-1|\lambda)$  and  $\gamma = 1$ , we have

$$V_n(0) = \sum_{k=1}^{\infty} \frac{1}{k^{(n)}} \frac{e^{-\lambda} \lambda^{k-1}}{(k-1)!} = \lambda^{-n} \sum_{k=1}^{\infty} \frac{e^{-\lambda} \lambda^{k+n-1}}{(k+n-1)!} = \lambda^{-n} \sum_{k=n+1}^{\infty} p(k),$$

and  $\sum_{k=n+1}^{\infty} p(k) = 1 - \sum_{k=1}^n p(k)$ .

#### 4.2.10.3 Proof of the stick-breaking representation

Let  $\varepsilon_1, \varepsilon_2, \dots \stackrel{\text{iid}}{\sim} \text{Exponential}(\lambda)$ ,  $\tilde{K} = \min\{j : \sum_{i=1}^j \varepsilon_i > 1\}$ ,  $\tilde{\pi}_i = \varepsilon_i$  for  $i = 1, \dots, \tilde{K} - 1$  and  $\tilde{\pi}_{\tilde{K}} = 1 - \sum_{i=1}^{\tilde{K}-1} \tilde{\pi}_i$ .

Let  $S_j = \sum_{i=1}^j \varepsilon_i$  and define  $N(\tau) = \#\{j : S_j \leq \tau\}$ . Then  $N(\tau)$  is a Poisson pro-

cess with rate  $\lambda$  (e.g., [Durrett, 1996](#)). Hence,  $N(\tau) \sim \text{Poisson}(\lambda\tau)$ , and in particular, since  $\tilde{K} = N(1) + 1$  a.s., we have  $\tilde{K} - 1 \sim \text{Poisson}(\lambda)$ .

It is a standard result that  $(S_1, \dots, S_m) | N(\tau) = m$  has the same joint distribution as the order statistics  $(U_{(1)}, \dots, U_{(m)})$  of i.i.d. uniform random variables  $U_1, \dots, U_m \sim \text{Uniform}[0, \tau]$ . Considering the case of  $\tau = 1$ , the conditional density of  $(S_1, \dots, S_m) | N(1) = m$  is therefore  $m! I(0 < s_1 < \dots < s_m < 1)$ , and since the change of variables from  $(S_1, \dots, S_m)$  to  $(\varepsilon_1, \dots, \varepsilon_m)$  has Jacobian equal to 1, it follows that  $(\varepsilon_1, \dots, \varepsilon_m) | N(1) = m$  has density  $m! I(\sum_{i=1}^m \varepsilon < 1, \varepsilon_i > 0 \forall i)$ . Therefore, taking  $k = m + 1$ , we see that  $\tilde{\pi} | \tilde{K} = k$  has the uniform  $k$ -dimensional Dirichlet distribution.

### 4.3 Asymptotics

In this section, we first summarize and compare asymptotic results for posterior inference in the mixture of finite mixtures (MFM) and Dirichlet process mixture (DPM) models with respect to three objects of possible interest: the density, the mixing distribution, and the number of components. We then provide some basic asymptotic results for various aspects of the MFM: the asymptotic relationship between the number of components and the number of clusters, the asymptotic behavior of  $V_n(t)$ , and the form of the conditional distribution on part sizes given the number of parts.

### 4.3.1 Posterior asymptotics

To set things up, we first pose the questions of interest from a general perspective. Suppose  $\{f_\theta : \theta \in \Theta\}$  is a parametric family of densities with respect to a measure on  $\mathcal{X} \subset \mathbb{R}^d$ . Given a discrete probability measure  $G = \sum_{i=1}^{\infty} \pi_i \delta_{\theta_i}$  (where  $\theta_1, \theta_2, \dots \in \Theta$ ,  $\delta_\theta$  is the unit point mass at  $\theta \in \Theta$ , and  $\pi_1, \pi_2, \dots \geq 0$  such that  $\sum_i \pi_i = 1$ ), let  $f_G$  denote the density of the resulting mixture, that is,

$$f_G(x) = \int_{\Theta} f_\theta(x) G(d\theta) = \sum_{i=1}^{\infty} \pi_i f_{\theta_i}(x),$$

and let  $s(G)$  denote the number of elements in the support of  $G$ , that is,  $s(G) = |\text{support}(G)| \in \{1, 2, \dots\} \cup \{\infty\}$ .

Suppose we observe data  $X_1, \dots, X_n \sim f_*$  for some true density  $f_*$ . The Bayesian approach is to put a prior on  $f_*$ , and given  $X_1, \dots, X_n$ , make posterior inferences about  $f_*$ . Mixtures of the form  $f_G$  above tend to generate a very flexible class of densities, and considerable success has been obtained by constructing Bayesian models following this form, Dirichlet process mixtures being the prime example.

In the density estimation setting, the true density  $f_*$  does not have to take the form  $f_G$  in order to obtain posterior guarantees — only general regularity conditions are required; meanwhile, when estimating the mixing distribution and the number of clusters, it is typically necessary to assume  $f_*$  is of the form  $f_G$ .

Further, although it is irrelevant for density estimation, to have any hope of consistency for the mixing distribution, it is necessary to assume that the family  $\{f_\theta\}$  is *mixture identifiable* in the following sense: if  $s(G), s(G') < \infty$  and  $f_G = f_{G'}$  a.e., then  $G = G'$ . Note that we are not constraining the  $\theta$ 's to be distinct and

ordered (nor constraining the  $\pi$ 's to be positive), so the  $\pi$ 's and  $\theta$ 's will be non-identifiable; however, the assumption here is identifiability of the measure  $G$ , and this is satisfied for many commonly-used families  $\{f_\theta\}$ , such as multivariate Gaussian, Gamma, Poisson, Exponential, Cauchy (Teicher, 1963, Yakowitz and Spragins, 1968) in the continuous case, and Poisson, Geometric, Negative Binomial, or any power-series distribution (Sapatinas, 1995) in the discrete case.

It should be said, perhaps, at this point, that inferences about the mixing distribution or the number of components will typically be sensitive to misspecification of the family  $\{f_\theta\}$ ; for instance, any sufficiently regular density can be approximated arbitrarily well by a mixture of Gaussians, so even if the true density is very close to but not exactly a finite mixture Gaussians, the model will introduce “extra” components in order to fit the data distribution. On the other hand, when making inferences about the density, there is no assumption that the true density  $f_*$  has any particular form (other than general regularity conditions), so misspecification is not a serious issue.

Nonetheless, it is desirable to obtain consistency results for the mixing distribution and the number of components when the model is well-specified, since lack of consistency under ideal conditions would be a red flag.

Given a prior on  $G$ , then, consider the following questions.

- (1) *Density estimation.* Does the posterior on the density concentrate at the true density, and if so, at what rate of convergence? That is, for an appropriate metric  $\text{dist}(\cdot, \cdot)$ , does

$$\mathbb{P}_{\text{model}}(\text{dist}(f_G, f_*) < \varepsilon \mid X_{1:n}) \xrightarrow[n \rightarrow \infty]{\mathbb{P}_{\text{data}}} 1$$

for all  $\varepsilon > 0$ , and if so, how rapidly can we let  $\varepsilon_n$  go to 0 and still have

$$\mathbb{P}_{\text{model}}(\text{dist}(f_G, f_*) < C\varepsilon_n \mid X_{1:n}) \xrightarrow[n \rightarrow \infty]{\mathbb{P}_{\text{data}}} 1$$

for some  $C > 0$ ? The standard choices of metric here are  $L_1$  and Hellinger.

- (2) *Mixing distribution.* Does the posterior on the mixing distribution concentrate at the true mixing distribution (assuming there is one)? That is, for an appropriate metric  $\text{dist}(\cdot, \cdot)$ , does

$$\mathbb{P}_{\text{model}}(\text{dist}(G, G_*) < \varepsilon \mid X_{1:n}) \xrightarrow[n \rightarrow \infty]{\mathbb{P}_{\text{data}}} 1$$

for all  $\varepsilon > 0$ ? The Wasserstein metric has been used in this context ([Nguyen, 2013](#)).

- (3) *Number of components.* Does the posterior on the number of components concentrate at the true number of components, for data from a finite mixture? That is, does

$$\mathbb{P}_{\text{model}}(s(G) = s(G_*) \mid X_{1:n}) \xrightarrow[n \rightarrow \infty]{\mathbb{P}_{\text{data}}} 1?$$

Also of interest, due to its (albeit, questionable) use in practice for inference about the number of components, is the number of clusters  $T$ , defined — when the prior on  $G$  is introduced via priors on  $\pi$  and  $\theta$  — as the number of distinct values of the latent allocation variables  $Z_1, \dots, Z_n \sim \pi$ , where  $X_i \sim f_{\theta_{Z_i}}$ . Does the posterior on the number of clusters  $T$  concentrate at the true number of *components*? That is, does

$$\mathbb{P}_{\text{model}}(T = s(G_*) \mid X_{1:n}) \xrightarrow[n \rightarrow \infty]{\mathbb{P}_{\text{data}}} 1?$$

In varying degrees of generality, answers to these questions for both MFMs and

DPMs have been provided in the literature, as summarized in Table 4.1. We provide references for these results in the sections below. It should be emphasized that most of these results — especially regarding rates of convergence — have only been shown for certain types of models; the table is simply meant to give a rough idea of the general picture.

Table 4.1: Summary of known posterior consistency results (Yes = consistent, No = not consistent); see text for explanation.

|                      | MFMs               | DPMs               |
|----------------------|--------------------|--------------------|
| Density estimation   | Yes (optimal rate) | Yes (optimal rate) |
| Mixing distribution  | Yes                | Yes                |
| Number of components | Yes                | No                 |

Here, by “optimal rate”, we mean optimal up to a logarithmic factor in  $n$ . Also, note that if the true number of components is finite, then the DPM posterior on the number of components is trivially inconsistent, since the number of components is infinite with probability 1; the inconsistency in the table refers to inconsistency of the number of clusters  $T$ .

#### 4.3.1.1 Density estimation

Density estimation is perhaps the primary type of application for a flexible Bayesian mixture model, such as a DPM or MFM. In this setting, the true density  $f_*$  is assumed only to satisfy certain regularity conditions; it does not have to be of the form assumed in the model. Posterior inferences for  $f_*$  enable, for example, predictive inference regarding future data points  $X_m$  for  $m > n$ , inference about functionals of  $f_*$ , or decision-theoretic inferences to minimize a loss of interest. Since posterior asymptotics of the density are not our main focus, here we simply reference the primary results from the literature.



For DPMs, there has been great progress in establishing posterior consistency and rates of concentration for the density, in a number of cases. For a general overview, see Ghosal (2010), and for background, see Ghosh and Ramamoorthi (2003). For some of the many contributions to this area, see Ghosal et al. (1999), Barron et al. (1999), Ghosal and Van der Vaart (2001), Lijoi et al. (2005b), Tokdar (2006), Ghosh and Ghosal (2006), Tang and Ghosal (2007), Ghosal and Van der Vaart (2007), Walker et al. (2007), James (2008), Wu and Ghosal (2010), Bhattacharya and Dunson (2010), Khazaei et al. (2012), Scricciolo (2012), Pati et al. (2013), and references therein.

More recently, these efforts have been extended to variable-dimension mixtures, and similar results have been obtained. Under fairly general conditions, Kruijer (2008) and Kruijer et al. (2010) have shown that for a large class of variable-dimension location-scale mixtures, the posterior on the density concentrates at the true density at the minimax-optimal rate (up to a logarithmic factor). It should be noted that these optimality proofs apply to a model that is not exactly an MFM as defined in Equation 4.1.1, since there is a single scale parameter shared among all mixture components, rather than a separate scale parameter for each component; the same assumption is made in the DPM optimality proofs. However, this can be viewed as extending the hierarchical model by first sampling the common scale from some prior, and given this, defining the MFM as before.

These results suggest that for density estimation, variable-dimension mixtures may perform as well as Dirichlet process mixtures.

### 4.3.1.2 Mixing distribution

Now, suppose the true density  $f_*$  is in fact a finite mixture from the assumed family  $\{f_\theta\}$ , that is,  $f_* = f_{G_*}$  for some  $G_*$  with  $s(G_*) < \infty$ .

Under this (strong) assumption, MFMs can be shown to be consistent for the mixing distribution (and the density, and the number of components), under very general conditions, by appealing to Doob’s theorem (Nobile, 1994). Roughly speaking, Doob’s theorem says that for a correctly-specified, identifiable model  $P_\theta$ , with probability 1, if the parameter  $\theta_0$  is drawn from the prior, and the data is i.i.d. from  $P_{\theta_0}$  (given  $\theta_0$ ), then the posterior concentrates at  $\theta_0$  (Doob, 1949, Ghosh and Ramamoorthi, 2003). The theorem applies under very general conditions, however, a slightly subtle point is that the condition “with probability 1, if the parameter  $\theta_0$  is drawn from the prior” allows for an exceptional set of prior probability zero, and in an infinite-dimensional space, this exceptional set can be quite large even for seemingly reasonable priors (Freedman, 1963, Diaconis and Freedman, 1986).

Fortunately, in an MFM, the parameter space is a countable union of finite-dimensional spaces, since for each  $k = 1, 2, \dots$ , a mixture with  $k$  components has a finite-dimensional parameter space. Consequently, one can choose a prior such that any set of prior probability zero also has Lebesgue measure zero (where Lebesgue measure is extended in the natural way to this disjoint union of spaces), and indeed, an MFM will have this property whenever  $p(k) > 0$  for all  $k = 1, 2, \dots$  and Lebesgue measure is absolutely continuous with respect to the base measure  $H$  (for instance, if  $H$  has a density with respect to Lebesgue measure that is strictly positive on  $\Theta$ ).

Consequently, Doob’s theorem can be applied to prove consistency of MFMs for Lebesgue almost-all true mixing distributions (Nobile, 1994). (A minor techni-

cal issue is that Doob's theorem also requires identifiability, and mixtures are not identifiable in the usual sense; however, as long as the family  $\{f_\theta\}$  is mixture identifiable, the theorem can be applied by defining a new parameter space containing one representative from each equivalence class.)

Regarding DPM consistency for the mixing distribution, [Nguyen \(2013\)](#) has shown that the DPM posterior on the mixing measure  $G$  concentrates at the true mixing measure  $G_*$  (in the Wasserstein metric), and has provided a rate of convergence.

#### 4.3.1.3 Number of components

As mentioned in the preceding section, when the true density  $f_*$  is a finite mixture from the assumed family  $\{f_\theta\}$ , under very general conditions the MFM posterior on the number of components will concentrate at the true number of components, by Doob's theorem ([Nobile, 1994](#)). (Other approaches have also been proposed for estimating the number of components ([Henna, 1985](#), [Keribin, 2000](#), [Leroux, 1992](#), [Ishwaran et al., 2001](#), [James et al., 2001](#), [Henna, 2005](#)), including methods with claims of robustness ([Woo and Sriram, 2006, 2007](#)).) On the other hand, DPMs are not well-suited to estimating the number of components in a finite mixture, as discussed at length in Chapters [2](#) and [3](#).

In Chapter [3](#), we gave conditions under which the number of clusters in a Gibbs partition mixture model is inconsistent for the number of components, and observed that DPMs satisfy the conditions on the partition distribution. Since MFMs are also Gibbs partition mixtures, the fact that they are consistent for the number of components implies that these conditions are not satisfied for MFMs. This is easy to

verify directly, using the following asymptotic expression for  $V_n(t)$ , which is proved in Section 4.3.4 below: for any  $t \in \{1, 2, \dots\}$ , if  $p_K(t) > 0$  then

$$V_n(t) \sim \frac{t!}{(\gamma t)^{(n)}} p_K(t) \sim \frac{t!}{n!} \frac{\Gamma(\gamma t)}{n^{\gamma t - 1}} p_K(t) \quad (4.3.1)$$

as  $n \rightarrow \infty$ . (Here,  $p_K$  is the distribution of the number of components  $K$ , and  $\gamma > 0$  is the Dirichlet parameter.) Let  $\mathcal{C}$  be the random partition of  $[n]$  as in Section 4.2.1, and let  $A = (A_1, \dots, A_T)$  be the ordered partition of  $[n]$  obtained by randomly ordering the parts of  $\mathcal{C}$ , uniformly among the  $T!$  possible choices, where  $T = |\mathcal{C}|$ . Note that given  $\mathcal{C}$ , the  $T!$  such choices all yield distinct  $A$ 's. Therefore,

$$p(A) = \frac{p(\mathcal{C})}{|\mathcal{C}|!} = \frac{1}{t!} V_n(t) \prod_{c \in \mathcal{C}} \gamma^{(|c|)} = \frac{1}{t!} V_n(t) \prod_{i=1}^t \gamma^{(|A_i|)},$$

where  $t = |\mathcal{C}|$ . The inconsistency conditions are in terms of a distribution on  $A$  of the form  $p(A) = v_n(t) \prod_{i=1}^t w_n(|A_i|)$ , and we can put the MFM distribution on  $A$  in this form by choosing  $v_n(t) = V_n(t)/t!$  and  $w_n(a) = \gamma^{(a)}$ . One of the conditions for inconsistency is that  $\limsup_n v_n(t)/v_n(t+1) < \infty$  for some  $t$ . This condition is not satisfied for MFMs — provided that  $p_K(k) > 0$  for all  $k = 1, 2, \dots$  — since for any  $t \in \{1, 2, \dots\}$ , by Equation 4.3.1,

$$\frac{v_n(t)}{v_n(t+1)} = (t+1) \frac{V_n(t)}{V_n(t+1)} \sim \frac{\Gamma(\gamma t)}{\Gamma(\gamma(t+1))} \frac{p_K(t)}{p_K(t+1)} n^\gamma \rightarrow \infty$$

as  $n \rightarrow \infty$ .

### 4.3.2 Relationship between the number of clusters and number of components

In the MFM, it is perhaps intuitively clear that, under the prior at least, the number of clusters  $T = |\mathcal{C}|$  should behave very similarly to the number of components  $K$  when  $n$  is large. It turns out that under the posterior they also behave very similarly for large  $n$ . Assume  $p_K(k) > 0$  for all  $k$ . We show that for any  $x_1, x_2, \dots \in \mathcal{X}$ ,  $k \in \{1, 2, \dots\}$ , if  $\liminf_n p(K = k | x_{1:n}) > 0$  then

$$p(T = k | x_{1:n}) \sim p(K = k | x_{1:n})$$

as  $n \rightarrow \infty$ . See [Nobile \(2005\)](#) for related results.

To show this, first we observe that for any  $t \in \{1, 2, \dots\}$ ,

$$p_n(K = t | T = t) = \frac{1}{V_n(t)} \frac{t^{(t)}}{(\gamma t)^{(n)}} p_K(t) \xrightarrow[n \rightarrow \infty]{} 1 \quad (4.3.2)$$

(where  $p_n$  denotes the MFM distribution with  $n$  samples), by Equations [4.2.7](#) and [4.3.1](#). Next, write

$$p(K = k | x_{1:n}) = \sum_{t=1}^k p(K = k | T = t, x_{1:n}) p(T = t | x_{1:n}), \quad (4.3.3)$$

and note that by Equations [4.2.9](#) and [4.3.2](#),  $p(K = k | T = t, x_{1:n}) = p_n(K = k | T = t) \rightarrow I(k = t)$  (also see [Nobile \(2005\)](#)). If  $\liminf_n p(K = k | x_{1:n}) > 0$ , then dividing Equation [4.3.3](#) by  $p(K = k | x_{1:n})$  and taking  $n \rightarrow \infty$ , the terms for  $t = 1, \dots, k - 1$  go to 0, and we are left with  $1 = \lim_n p(T = k | x_{1:n}) / p(K = k | x_{1:n})$ .

### 4.3.3 Distribution of the cluster sizes under the prior

Here, we illustrate one of the major differences between the MFM and DPM priors, namely, the distribution on partitions given the number of clusters/parts. (See [Green and Richardson \(2001\)](#) for similar observations along these lines.) Under the prior, the MFM prefers all parts to be roughly the same order of magnitude, while the DPM prefers having some parts very small. Interestingly, these prior influences remain visible in certain aspects of the posterior, even in the limit as  $n$  goes to infinity, as shown by our results on inconsistency of DPMs for the number of components in a finite mixture; see Chapters 2 and 3. In Chapter 5 below, we also empirically observe differences in the posteriors.

Let  $\mathcal{C}$  be the random partition of  $[n]$  as in Section 4.2.1, and let  $A = (A_1, \dots, A_T)$  be the ordered partition of  $[n]$  obtained by randomly ordering the parts of  $\mathcal{C}$ , uniformly among the  $T!$  such choices, where  $T = |\mathcal{C}|$ . Recall that

$$p(A) = \frac{p(\mathcal{C})}{|\mathcal{C}|!} = \frac{1}{t!} V_n(t) \prod_{i=1}^t \gamma^{(|A_i|)},$$

where  $t = |\mathcal{C}|$ . Now, let  $S = (S_1, \dots, S_T)$  be the vector of part sizes of  $A$ , that is,  $S_i = |A_i|$ . Then

$$p(S = s) = \sum_{A: S(A)=s} p(A) = \frac{n!}{s_1! \cdots s_t!} \frac{1}{t!} V_n(t) \prod_{i=1}^t \gamma^{(s_i)} = V_n(t) \frac{n!}{t!} \prod_{i=1}^t \frac{\gamma^{(s_i)}}{s_i!}$$

for  $s \in \Delta_t$ , where  $\Delta_t = \{s \in \mathbb{Z}^t : \sum_i s_i = n, s_i \geq 1 \forall i\}$  (i.e., the  $t$ -part compositions of  $n$ ). For any  $x > 0$ , writing  $x^{(m)}/m! = \Gamma(x + m)/(m! \Gamma(x))$  and using Stirling's approximation, we have

$$\frac{x^{(m)}}{m!} \sim \frac{m^{x-1}}{\Gamma(x)}$$

as  $m \rightarrow \infty$ . This yields the approximations

$$p(S = s) \approx \frac{V_n(t)}{\Gamma(\gamma)^t} \frac{n!}{t!} \prod_{i=1}^t s_i^{\gamma-1}$$

and

$$p(S = s \mid T = t) \approx \kappa \prod_{i=1}^t s_i^{\gamma-1}$$

for  $s \in \Delta_t$ , where  $\kappa$  is a normalization constant. Note that  $p(s|t)$ , although a discrete distribution, has approximately the same shape as a symmetric  $t$ -dimensional Dirichlet distribution. This would be obvious if we were conditioning on the number of components  $K$ , and it makes intuitive sense when conditioning on  $T$ , since  $K$  and  $T$  are essentially the same for large  $n$ .

It is interesting to compare this to the corresponding distributions for the Chinese restaurant process. In the CRP, we have  $p_{\text{CRP}}(\mathcal{C}) = \frac{\alpha^t}{\alpha^{(n)}} \prod_{c \in \mathcal{C}} (|c| - 1)!$ , and  $p_{\text{CRP}}(A) = p_{\text{CRP}}(\mathcal{C})/|\mathcal{C}|!$  as before, so for  $s \in \Delta_t$ ,

$$p_{\text{CRP}}(S = s) = \frac{n!}{s_1! \cdots s_t!} \frac{1}{t!} \frac{\alpha^t}{\alpha^{(n)}} \prod_{i=1}^t (s_i - 1)! = \frac{n!}{\alpha^{(n)}} \frac{\alpha^t}{t!} s_1^{-1} \cdots s_t^{-1}$$

and

$$p_{\text{CRP}}(S = s \mid T = t) \propto s_1^{-1} \cdots s_t^{-1}, \tag{4.3.4}$$

which has the same shape as a  $t$ -dimensional Dirichlet distribution with all the parameters taken to 0 (noting that this is normalizable since  $\Delta_t$  is finite). Asymptotically in  $n$ , the distribution  $p_{\text{CRP}}(s|t)$  puts all of its mass in the “corners” of the discrete simplex  $\Delta_t$ , while under the MFM,  $p(s|t)$  remains more evenly dispersed.

### 4.3.4 Asymptotics of $V_n(t)$

Recall that (Equation 4.2.2)

$$V_n(t) = \sum_{k=1}^{\infty} \frac{k^{(t)}}{(\gamma k)^{(n)}} p_K(k),$$

for  $1 \leq t \leq n$ , with  $\gamma > 0$  and  $p_K$  a p.m.f. on  $\{1, 2, \dots\}$ . Here, we consider the asymptotics of  $V_n(t)$ : we show that for any  $t \in \{1, 2, \dots\}$ , if  $p_K(t) > 0$  then

$$V_n(t) \sim \frac{t!}{(\gamma t)^{(n)}} p_K(t) \sim \frac{t!}{n!} \frac{\Gamma(\gamma t)}{n^{\gamma t - 1}} p_K(t) \quad (4.3.5)$$

as  $n \rightarrow \infty$ . Hence, asymptotically,  $V_n(t)$  has a simple interpretation — it behaves like the  $k = t$  term in the series. To show this, we use the following elementary result; this is a special case of the dominated convergence theorem, but since the proof is so simple we provide it here.

**Proposition 4.3.1.** *For  $j = 1, 2, \dots$ , let  $a_{1j} \geq a_{2j} \geq \dots \geq 0$  such that  $a_{ij} \rightarrow 0$  as  $i \rightarrow \infty$ . If  $\sum_{j=1}^{\infty} a_{1j} < \infty$  then  $\sum_{j=1}^{\infty} a_{ij} \rightarrow 0$  as  $i \rightarrow \infty$ .*

*Proof.* Letting  $\varepsilon_R = \sum_{j>R} a_{1j}$ , we have  $\varepsilon_R \rightarrow 0$  as  $R \rightarrow \infty$ . For any  $R > 0$ ,  $\limsup_i \sum_{j=1}^{\infty} a_{ij} \leq \limsup_i \sum_{j=1}^R a_{ij} + \varepsilon_R = \varepsilon_R$ . Taking  $R \rightarrow \infty$  gives the result.  $\square$

For any  $x > 0$ , writing  $x^{(n)}/n! = \Gamma(x+n)/(n! \Gamma(x))$  and using Stirling's approximation, we have

$$\frac{x^{(n)}}{n!} \sim \frac{n^{x-1}}{\Gamma(x)}$$

as  $n \rightarrow \infty$ . Therefore, the  $k = t$  term of  $V_n(t)$  is

$$\frac{t^{(t)}}{(\gamma t)^{(n)}} p_K(t) \sim \frac{t!}{n!} \frac{\Gamma(\gamma t)}{n^{\gamma t - 1}} p_K(t).$$



The first  $t - 1$  terms of  $V_n(t)$  are 0, so to establish Equation 4.3.5, we need to show that the rest of the series, divided by the  $k = t$  term, goes to 0. (Recall that we have assumed  $p_K(t) > 0$ .) To this end, let

$$b_{nk} = (\gamma t)^{(n)} \frac{k^{(t)}}{(\gamma k)^{(n)}} p_K(k).$$

We must show that  $\sum_{k=t+1}^{\infty} b_{nk} \rightarrow 0$  as  $n \rightarrow \infty$ . We apply Proposition 4.3.1 with  $a_{ij} = b_{t+i,t+j}$ . Clearly, for any  $k > t$ ,  $b_{1k} \geq b_{2k} \geq \dots \geq 0$ . Further, for any  $k > t$ ,

$$\frac{(\gamma t)^{(n)}}{(\gamma k)^{(n)}} \sim \frac{n^{\gamma t-1} \Gamma(\gamma k)}{\Gamma(\gamma t) n^{\gamma k-1}} \rightarrow 0$$

as  $n \rightarrow \infty$ , hence, for any  $k > t$ ,  $b_{nk} \rightarrow 0$  as  $n \rightarrow \infty$ . Finally, observe that  $\sum_{k=t+1}^{\infty} b_{nk} \leq (\gamma t)^{(n)} V_n(t) < \infty$  for any  $n \geq t$ . Therefore, by Proposition 4.3.1,  $\sum_{k=t+1}^{\infty} b_{nk} \rightarrow 0$  as  $n \rightarrow \infty$ . This proves Equation 4.3.5.

## 4.4 Inference algorithms

The usual approach to inference in variable-dimension mixture models is reversible jump Markov chain Monte Carlo (Richardson and Green, 1997). However, now that we have established that MFMs (as defined in Equation 4.1.1) have many of the same attractive properties as DPMs, much of the extensive body of work on Markov chain Monte Carlo (MCMC) samplers for the DPM can be directly applied to MFMs. One advantage of this is that these samplers tend to be applicable to a wide range of component distribution families  $\{f_{\theta}\}$ , without requiring one to design specialized moves. We show how this works for two well-known MCMC sampler algorithms:

- (1) “Algorithm 3”, for conjugate priors (MacEachern, 1994, Neal, 1992, 2000), and

- (2) “Algorithm 8”, for non-conjugate priors (Neal, 2000, MacEachern and Müller, 1998).

The names Algorithm 3 and Algorithm 8 come from Neal (2000); it has become fairly common to refer to these algorithms by the numbers he gave them. Algorithm 3 is a collapsed sampler (i.e., the component parameters have been integrated out) involving the partition  $\mathcal{C}$ , while Algorithm 8 is an auxiliary variable algorithm involving  $\mathcal{C}$  and the component parameters.

In fact, we describe a generalization of Algorithm 8, allowing a distribution other than the prior to be used for the auxiliary variables. This allows the algorithm to be used even when one cannot sample from the prior, as well as allowing for the possibility of improved mixing time with a better auxiliary variable distribution.

A well-known issue with incremental samplers such as these, when applied to DPMs, is that the mixing time can be somewhat slow, since it may take a long time to create or destroy substantial clusters by moving one element at a time. With MFMs, this issue seems to be exacerbated, since MFMs tend to put small probability (compared with DPMs) on partitions with tiny clusters (see Section 4.3.3), making it difficult for the sampler to move through these regions of the sample space; see Section 5.1.7 for examples of this in experiments.

To circumvent this issue, split-merge samplers for DPMs have been proposed in which a large number of elements can be reassigned in a single move (Dahl, 2003, 2005, Jain and Neal, 2004, 2007). In the same way as the incremental samplers, it should be possible to directly apply such split-merge samplers to MFMs, using the properties of the partition distribution described in Section 4.2. More generally, it seems likely that any partition-based MCMC sampler for DPMs could be applied to

MFMs as well. (We should point out that split-merge samplers are often used with reversible jump MCMC also (Richardson and Green, 1997, Green and Richardson, 2001); however, as before, there might be an advantage to using the DPM-style split-merge samplers in terms of the ease of application to new models.)

#### 4.4.1 Inference with conjugate priors

When  $H$  is a conjugate prior for the family of component distributions  $\{f_\theta : \theta \in \Theta\}$ , often there is an easy-to-compute expression for the “single-cluster marginal likelihood”,

$$m(x_c) = \int \prod_{j \in c} f_\theta(x_j) H(d\theta),$$

of a subset  $c \subset [n]$  of the data. To do posterior inference, we would like to draw samples from the posterior on the partition  $\mathcal{C}$ ,

$$p(\mathcal{C}|x_{1:n}) \propto p(x_{1:n}|\mathcal{C})p(\mathcal{C}) = \left( \prod_{c \in \mathcal{C}} m(x_c) \right) p(\mathcal{C}),$$

(using Equation 4.2.5). The following MCMC algorithm provides an easy way to sample (approximately) from  $p(\mathcal{C}|x_{1:n})$ . The algorithm was described by MacEachern (1994), Neal (1992), and Neal (2000) (Algorithm 3) for DPMs; here, we show how it is immediately adapted to MFMs. For ease of comparison, we describe both the MFM and DPM versions, side-by-side. The DPM concentration parameter is denoted by  $\alpha$ .

Given a partition  $\mathcal{C}$ , denote by  $\mathcal{C} \setminus j$  the partition obtained by removing element  $j$  from  $\mathcal{C}$ . We use restaurant terminology, by way of analogy with the restaurant process described in Section 4.2.6.

**Algorithm 4.4.1** (Collapsed sampler for MFM and DPM).

Initialize  $\mathcal{C} = \{[n]\}$  (i.e., a single cluster).

Repeat the following  $N$  times, to obtain  $N$  samples:

For  $j = 1, \dots, n$ :

(1) Choose to reseat customer  $j$  ...

|   |   |                                      |
|---|---|--------------------------------------|
|   | <u>MFM</u>                                      | <u>DPM</u>                           |
| at table $c \in \mathcal{C} \setminus j$ with probability $\propto$ | $( c  + \gamma) \frac{m(x_{c \cup j})}{m(x_c)}$ | $ c  \frac{m(x_{c \cup j})}{m(x_c)}$ |
| at a new table with probability $\propto$                           | $\gamma \frac{V_n(t+1)}{V_n(t)} m(x_j)$         | $\alpha m(x_j)$                      |

where  $t = |\mathcal{C} \setminus j|$ .

(2) Update  $\mathcal{C}$  to reflect this reassignment.

It is reasonable to assume that  $m(x_{1:n}) > 0$  (since otherwise something is probably wrong with the model); this ensures that  $m(x_c) > 0$  for all  $c \subset [n]$ . Further,  $V_n(t) > 0$  if and only if  $t < t_* + 1$  where  $t_* = \sup\{k : p_K(k) > 0\}$  (allowing, possibly,  $t_* = \infty$ ); see Section 4.2.4. Hence, the transition probabilities are all well-defined and positive up to  $t_*$  (and there is zero probability of going beyond  $t_*$ ).

This algorithm is usually described as a Gibbs sampler in terms of the allocation variables  $Z_1, \dots, Z_n$ , however, in that formulation one has to accommodate the fact that these variables must take particular values. Our description of the algorithm is directly in terms of the partition  $\mathcal{C}$ . This can be seen as a blocked Gibbs sampler in terms of the  $n \times n$  binary matrix in which entry  $(i, j)$  is 1 when  $i$  and  $j$  belong to the same cluster, taking the  $l$ th block ( $l = 1, \dots, n$ ) to be all the entries  $(i, j)$  such that  $i = l$  or  $j = l$ . Another easy way to see that it has the correct invariant

distribution (the posterior,  $p(\mathcal{C}|x_{1:n})$ ) is as a special case of the auxiliary variable algorithm described in Section 4.4.3.2 below; see that section for this argument.

Irreducibility is satisfied, since, for example, from any state  $\mathcal{C}^0$  with  $p(\mathcal{C}^0|x_{1:n}) > 0$ , there is a positive probability of reaching the state  $\mathcal{C} = \{[n]\}$  (i.e., one cluster) in a single sweep, by moving each element to the cluster containing element 1. (Note that this requires  $V_n(t) > 0$  for all  $t = 1, \dots, |\mathcal{C}^0|$ , and this is indeed the case; see the discussion above and observe that  $p(\mathcal{C}^0|x_{1:n}) > 0$  implies  $V_n(|\mathcal{C}^0|) > 0$ .) The chain is aperiodic, since there is positive probability of remaining in the same state.

#### 4.4.2 Inference with non-conjugate priors

Often, the need arises to use a non-conjugate prior — for instance, we may wish to use a family of component distributions  $\{f_\theta : \theta \in \Theta\}$  for which there does not exist a conjugate prior. Neal (2000) described a clever auxiliary variable method (Algorithm 8), based on a very similar algorithm of MacEachern and Müller (1998), for inference in DPMS with a non-conjugate prior. The method is particularly attractive since it does not require approximation of the marginal likelihoods  $m(x_c)$  to compute the transition probabilities, and consequently, is exact in the sense that its invariant distribution is exactly equal to the posterior, and not an approximation thereof.

We show how this algorithm can be adapted to MFMs as well, in the same manner as the previous section. In fact, we describe a generalization of the algorithm allowing for a distribution other than the prior to be used for the auxiliary variables. This permits the algorithm to be used when it is not possible to sample exactly from the prior, and also introduces the possibility of improved mixing time via a better auxiliary variable distribution. We show that when the “single-cluster posterior”

(defined below) is used instead of the prior, this reduces — with one minor exception — to the sampler referred to as Algorithm 2 in Neal (2000) (for earlier descriptions of the algorithm, see MacEachern (1994), West et al. (1994), MacEachern (1998)).

As before, we present the algorithm in terms of the partition  $\mathcal{C}$  (instead of the allocation variables  $Z_1, \dots, Z_n$  as is usually done). Recall from Section 4.2.2 that the model can be equivalently formulated as:

$$\begin{aligned} \mathcal{C} &\sim p(\mathcal{C}) \\ \phi_c &\stackrel{\text{iid}}{\sim} H \text{ for } c \in \mathcal{C}, \text{ given } \mathcal{C} \\ X_j &\sim f_{\phi_c} \text{ independently for } j \in c, c \in \mathcal{C}, \text{ given } \phi, \mathcal{C}. \end{aligned}$$

The algorithm constructs a Markov chain with state  $(\mathcal{C}, \phi)$ , where  $\phi = (\phi_c : c \in \mathcal{C})$ , producing samples from the posterior  $\mathcal{C}, \phi | x_{1:n}$ . (Note that for mathematical convenience, the parameters  $\phi_c$  are indexed by subsets  $c$ , rather than numbers, but of course any indexing scheme could be used in a software implementation.)

The algorithm has two “parameters”:  $s$  is the number of auxiliary variables to be used, and  $(R_x : x \in \mathcal{X})$  is a family of probability distributions on  $\Theta$ , parametrized by points  $x \in \mathcal{X}$  in the data space. (On first reading, one can think of the special case in which  $R_x = H$  for all  $x$ .) Assume  $H$  and  $R_x$ , for all  $x \in \mathcal{X}$ , have densities  $h$  and  $r_x$  with respect to a common measure  $\nu$ . Further, assume  $f_\theta(x)$ ,  $h(\theta)$ , and  $r_x(\theta)$  are strictly positive for all  $x \in \mathcal{X}$ ,  $\theta \in \Theta$ .

We use  $\pi(\theta | x_c)$  to denote the density of the *single-cluster posterior* with respect to  $\nu$ ,

$$\pi(\theta | x_c) = \left( \prod_{j \in c} f_\theta(x_j) \right) h(\theta) / m(x_c).$$

Also, we denote  $\phi \setminus j = (\phi'_{c'} : c' \in \mathcal{C} \setminus j)$  where  $\phi'_{c'} = \phi_c$  if  $c' \subset c$ .

**Algorithm 4.4.2** (Auxiliary variable sampler for MFM and DPM).

*Initialize  $\mathcal{C} = \{[n]\}$ , and initialize  $\phi_{[n]} \in \Theta$ .*

*Repeat the following  $N$  times, to obtain  $N$  samples:*

- *For each  $c \in \mathcal{C}$ , sample  $\phi_c \sim \pi(\phi_c | x_c)$ , or move  $\phi_c$  according to a Markov chain guaranteed to converge to this distribution.*

- *For  $j = 1, \dots, n$ :*

(1) *If  $j$  is seated alone, set  $\eta_1 \leftarrow \phi_{\{j\}}$  and sample  $\eta_2, \dots, \eta_s \stackrel{\text{iid}}{\sim} R_{x_j}$ ; otherwise, sample  $\eta_1, \dots, \eta_s \stackrel{\text{iid}}{\sim} R_{x_j}$ .*

(2) *Choose to reseate customer  $j \dots$*

*at table  $c \in \mathcal{C} \setminus j$  with probability  $\propto$*

$$\begin{array}{cc} \underline{MFM} & \underline{DPM} \\ (|c| + \gamma) f_{\phi'_c}(x_j) & |c| f_{\phi'_c}(x_j) \end{array}$$

*at a new table with parameter  $\eta_i$  ( $i = 1, \dots, s$ ) with probability  $\propto$*

$$\begin{array}{cc} \underline{MFM} & \underline{DPM} \\ \frac{\gamma}{s} \frac{V_n(t+1)}{V_n(t)} \frac{f_{\eta_i}(x_j) h(\eta_i)}{r_{x_j}(\eta_i)} & \frac{\alpha}{s} \frac{f_{\eta_i}(x_j) h(\eta_i)}{r_{x_j}(\eta_i)} \end{array}$$

*where  $t = |\mathcal{C} \setminus j|$  and  $\phi' = \phi \setminus j$ .*

(3) *Update  $\mathcal{C}$  and  $\phi$  to reflect this reassignment.*

To be clear, when reseating customer  $j$ , we sample from among  $|\mathcal{C} \setminus j| + s$  choices:  $|\mathcal{C} \setminus j|$  existing tables or a new table with one of the  $s$  parameters  $\eta_1, \dots, \eta_s$ .

Note that for the DPM, the original Algorithm 8 of Neal (2000) is the special

case obtained when  $r_x = h$  for all  $x \in \mathcal{X}$ . Further, we recover a slight variation of Algorithm 2 (Neal, 2000, MacEachern, 1994, West et al., 1994, MacEachern, 1998) (also see Ishwaran and James (2001) for a general formulation) when  $s = 1$  and  $r_x$  is the single-cluster posterior,  $r_x(\theta) = \pi(\theta|x)$ . For, in this case, in step (2),

$$\frac{f_{\eta_1}(x_j)h(\eta_1)}{r_{x_j}(\eta_1)} = m(x_j),$$

and thus, for the DPM, the probability of choosing a new table with parameter  $\eta_1$  is proportional to  $\alpha m(x_j)$ . If  $j$  was not seated alone in step (1), then  $\eta_1$  was sampled from  $\pi(\theta|x_j)$ , and this is equivalent to Algorithm 2; the only difference is that here, if  $j$  was seated alone then  $\eta_1$  was set equal to  $\phi_{\{j\}}$ , rather than being sampled from  $\pi(\theta|x_j)$ .

This suggests that mixing time could be improved by choosing  $r_x(\theta)$  to approximate  $\pi(\theta|x)$  (rather than using  $r_x = h$ ). Note that if  $r_x(\theta)$  can be chosen to be exactly  $\pi(\theta|x)$ , then there seems to be (essentially) no point in choosing  $s > 1$ , since it ends up being equivalent to  $s = 1$ , except on the occasion that  $j$  is seated alone.

On the other hand, if  $r_x(\theta)$  is not equal to  $\pi(\theta|x)$ , then choosing  $s$  large can be seen as approximating this choice (as noted by Neal (2000) in the case of  $r_x = h$ ), since then

$$\frac{1}{s} \sum_{i=1}^s \frac{f_{\eta_i}(x_j)h(\eta_i)}{r_{x_j}(\eta_i)} \approx \int f_{\eta}(x_j)H(d\eta) = m(x_j),$$

making the probability of choosing a new table approximately the same as in Algorithm 2, and given that we choose a new table, choosing among  $\eta_1, \dots, \eta_s$  proportionally to  $f_{\eta_i}(x_j)h(\eta_i)/r_{x_j}(\eta_i)$  amounts to an approximate sample from  $\pi(\theta|x_j)$ . It should be emphasized, however, that the algorithm has exactly the correct invariant distribution (see Section 4.4.3), and the preceding remarks are simply intended to



aid the intuition.

From this perspective, the original Algorithm 8 can be interpreted as employing a Monte Carlo approximation to  $\pi(\theta|x_j)$  (based on samples from the prior), while the generalization above can be interpreted as using importance sampling. Consequently, if there are many points  $x_j$  for which  $\pi(\theta|x_j)$  is very different than  $h(\theta)$ , and we have a decent approximation to  $\pi(\theta|x)$  that is easy to evaluate and sample from, then we would expect the algorithm above to mix significantly better than the original Algorithm 8.

In Section 4.4.3, we show the correctness of the algorithm above. In Chapter 5, we apply this algorithm in experiments.

### 4.4.3 Justification of the non-conjugate sampler

In this section, we show that the non-conjugate sampler in Section 4.4.2 produces a Markov chain converging to the posterior distribution  $\mathcal{C}, \phi | x_{1:n}$ .

For a Markov chain  $(U_i) = (U_0, U_1, U_2, \dots)$  in a measurable space  $(\mathcal{U}, \mathcal{A})$  with a countably-generated sigma-algebra  $\mathcal{A}$ , it is sufficient to have an invariant distribution  $\mu$ , irreducibility, and aperiodicity to guarantee convergence to the invariant distribution  $\mu$ , for  $\mu$ -almost all initial points  $u_0$  (Tierney, 1994). If  $U_0$  is drawn from a distribution that is absolutely continuous with respect to  $\mu$ , this guarantees convergence with probability 1.

Before establishing these properties, we recall some standard definitions; for general background on Markov chains, we refer to Tierney (1994), Robert and

Casella (2004), and Meyn and Tweedie (2009). A *transition kernel* is a function  $Q : \mathcal{U} \times \mathcal{A} \rightarrow [0, 1]$  such that  $u \mapsto Q(u, A)$  is measurable for each fixed  $A \in \mathcal{A}$ , and  $Q(u, \cdot)$  is a probability measure on  $(\mathcal{U}, \mathcal{A})$  for each fixed  $u \in \mathcal{U}$ . The Markov chain  $(U_i)$  is said to have transition kernel  $Q$  if

$$\mathbb{P}(U_i \in A \mid U_{i-1} = u) = Q(u, A).$$

A probability measure  $\mu$  on  $(\mathcal{U}, \mathcal{A})$  is *invariant* with respect to  $Q$  if

$$\mu(A) = \int_{\mathcal{U}} \mu(du)Q(u, A) \tag{4.4.1}$$

for all  $A \in \mathcal{A}$ ; this is conveniently abbreviated as  $\mu = \mu Q$ . For  $A \in \mathcal{A}$ , let  $T(A) = \inf\{i > 0 : U_i \in A\}$  denote the first time the chain hits  $A$  after time 0. Given a sigma-finite measure  $\varphi$  on  $(\mathcal{U}, \mathcal{A})$ , the chain  $(U_i)$  is  *$\varphi$ -irreducible* if, for any  $A \in \mathcal{A}$  such that  $\varphi(A) > 0$ , we have  $\mathbb{P}(T(A) < \infty \mid U_0 = u) > 0$  for all  $u \in \mathcal{U}$ . The chain is *irreducible* if it is  $\varphi$ -irreducible for some sigma-finite measure  $\varphi$  such that  $\varphi(\mathcal{U}) > 0$ . Assuming  $Q$  has invariant distribution  $\mu$ , it is *periodic* if there exist disjoint sets  $A_1, \dots, A_d \in \mathcal{A}$  for some  $d \geq 2$ , such that (a)  $\mu(A_1) > 0$ , (b)  $Q(u, A_{i+1}) = 1$  for all  $u \in A_i$ , for  $i = 1, \dots, d - 1$ , and (c)  $Q(u, A_1) = 1$  for all  $u \in A_d$ . Otherwise, it is *aperiodic*.

#### 4.4.3.1 Irreducibility and aperiodicity

Let  $\mathcal{U} = \bigcup \{\mathcal{C}\} \times \Theta^{|\mathcal{C}|}$  where the union is over partitions  $\mathcal{C}$  of  $[n]$  such that  $p(\mathcal{C}) > 0$ . Note that the sigma-algebra  $\mathcal{A}$  associated to  $\mathcal{U}$  is countably generated, since  $\mathcal{U}$  is the disjoint union of finitely many spaces of the form  $\Theta^t$ , and  $\Theta \subset \mathbb{R}^\ell$  has been given the Borel sigma-algebra (which is countably generated).

Consider the Markov chain with state  $U = (\mathcal{C}, \phi) \in \mathcal{U}$  resulting from the non-conjugate sampler; for definiteness, let  $U_0$  be the state after the first  $\phi$  move, and each successive  $U_i$  be the state after a complete sweep through the reassignments of  $j = 1, \dots, n$  and the following  $\phi$  move. Let  $Q$  be the corresponding transition kernel, and take  $\mu$  to be the posterior distribution of  $\mathcal{C}, \phi \mid x_{1:n}$ . Here, we show that the chain is irreducible and aperiodic; in Section 4.4.3.2, we verify that  $\mu$  is an invariant distribution.

To show irreducibility, choose  $\varphi = \delta_{\mathcal{C}^*} \times \Pi(\cdot \mid x_{[n]})$ , where  $\mathcal{C}^* = \{[n]\}$  (i.e., one cluster) and  $\Pi(\cdot \mid x_{[n]})$  is the single-cluster posterior, that is,  $\Pi(B \mid x_{[n]}) = \int_B \pi(\theta \mid x_{[n]}) \nu(d\theta)$ . Let  $A \in \mathcal{A}$  such that  $\varphi(A) > 0$ , let  $u \in \mathcal{U}$ , and write  $u = (\mathcal{C}^0, \phi^0)$ . First, observe that similarly to the case of the conjugate prior algorithm, since  $p(\mathcal{C}^0) > 0$ , we have  $V_n(|\mathcal{C}^0|) > 0$ , and there is positive probability of reaching  $\mathcal{C}^*$  in a single pass through the reassignments of  $j = 1, \dots, n$ , since  $f_\theta(x) > 0$  for all  $x \in \mathcal{X}$ ,  $\theta \in \Theta$  by assumption. Define  $B = \{\theta \in \Theta : (\mathcal{C}^*, \theta) \in A\}$ . When  $\mathcal{C} = \mathcal{C}^*$ , the  $\phi$  move consists of moving  $\phi_{[n]}$  according to a Markov chain guaranteed to converge to  $\Pi(\cdot \mid x_{[n]})$ , so since  $\Pi(B \mid x_{[n]}) = \varphi(A) > 0$ , it is guaranteed that  $\phi_{[n]}$  will hit  $B$  in finitely many steps, with probability 1. Since there is positive probability of staying at  $\mathcal{C}^*$  during successive reassignments of  $j = 1, \dots, n$ , we have  $\mathbb{P}(T(A) < \infty \mid U_0 = u) > 0$ .

To show aperiodicity, suppose  $d \geq 2$  and  $A_1, \dots, A_d \in \mathcal{A}$  are disjoint sets such that  $\mu(A_1) > 0$ ,  $Q(u, A_{i+1}) = 1$  for all  $u \in A_i$ , for  $i = 1, \dots, d-1$ , and  $Q(u, A_1) = 1$  for all  $u \in A_d$ . Pick  $\mathcal{C}^*$  such that the set  $B_1 = \{\phi : (\mathcal{C}^*, \phi) \in A_1\}$  satisfies  $\mu(\{\mathcal{C}^*\} \times B_1) > 0$ , i.e.,  $\mathbb{P}(\mathcal{C} = \mathcal{C}^*, \phi \in B_1 \mid X_{1:n} = x_{1:n}) > 0$ . Then  $\mathbb{P}(\phi \in B_1 \mid \mathcal{C} = \mathcal{C}^*, X_{1:n} = x_{1:n}) > 0$  also. Let  $B_i = \{\phi : (\mathcal{C}^*, \phi) \in A_i\}$  for  $i = 2, \dots, d$ . Then  $B_1, \dots, B_d$  are disjoint sets, and since there is a positive probability of staying at  $\mathcal{C}^*$  (with no modifications to  $\phi$ ) during the reassignments of  $j = 1, \dots, n$ , the transition kernel  $K$  for the  $\phi$  move must satisfy  $K(\phi, B_{i+1}) = 1$  for all  $\phi \in B_i$ ,

for  $i = 1, \dots, d - 1$ , and  $K(\phi, B_1) = 1$  for all  $\phi \in B_d$ . However, this cannot be, since by assumption,  $K$  is a transition kernel for a Markov chain that converges to the posterior distribution  $\phi \mid \mathcal{C}^*, x_{1:n}$ , and therefore  $K$  must be aperiodic (and have invariant distribution  $\phi \mid \mathcal{C}^*, x_{1:n}$ ) (Tierney, 1994). Hence, a contradiction is obtained, and the chain  $(U_i)$  must be aperiodic.

#### 4.4.3.2 Invariant distribution

A surprisingly useful way of constructing a transition kernel with a particular invariant distribution  $\mu$  is via auxiliary variables (Besag and Green, 1993, Edwards and Sokal, 1988): let  $U \sim \mu$  and introduce a random variable  $V$  that is dependent on  $U$ ; if the current state of the chain is  $u_i$ , then sample  $V = v \mid U = u_i$ , and make a move from  $u_i$  to  $u_{i+1}$  according to a transition kernel  $R_v$  for which the conditional distribution  $U \mid V = v$  is invariant. A transition kernel  $Q$  constructed in this way (for the overall move from  $u_i$  to  $u_{i+1}$ ) has  $\mu$  as an invariant distribution. Often, such a  $Q$  will not yield an irreducible chain, but the composition  $Q_1 Q_2 \cdots Q_m$  of a well-chosen sequence of finitely-many such kernels will yield irreducibility (and will always still have the same invariant distribution).

An important special case arises when  $V = f(U)$  for some function  $f$  (Tierney, 1994); this can be thought of as partitioning  $u$ -space according to the level sets of  $f$ , and making a move according to the conditional distribution of the level set containing the current state. The Gibbs sampler can be thought of as a further special case in which  $U$  consists of multiple variables and  $f$  is a projection onto a subset of variables (typically, projecting onto all but one).

The approach of the preceding paragraph can be used to see that the conjugate

prior algorithm in Section 4.4.1 has the correct invariant distribution: let  $U = \mathcal{C}$ , take  $\mu$  to be the posterior distribution  $p(\mathcal{C} | x_{1:n})$ , define  $f_j(\mathcal{C}) = \mathcal{C} \setminus j$  for  $j = 1, \dots, n$ , and define the transition matrices

$$Q_j(\mathcal{C}^0, \mathcal{C}) = p(\mathcal{C} | f_j(\mathcal{C}) = f_j(\mathcal{C}^0), X_{1:n} = x_{1:n})$$

for  $j = 1, \dots, n$ . Then, the posterior  $p(\mathcal{C} | x_{1:n})$  is an invariant distribution for each  $Q_j$ . Applying  $Q_1 Q_2 \cdots Q_n$  is equivalent to the conjugate prior algorithm described in Section 4.4.1.

To show that the non-conjugate prior algorithm in Section 4.4.2 has the correct invariant distribution, we actually use the auxiliary variable technique in two ways: first in a way resembling the preceding paragraph, and again to sample from the conditional distribution. We focus on the MFM case; the DPM case is essentially the same. Let  $U = (\mathcal{C}, \phi)$ , take  $\mu$  to be the posterior distribution of  $\mathcal{C}, \phi | x_{1:n}$ , define  $f_j(\mathcal{C}, \phi) = (\mathcal{C} \setminus j, \phi \setminus j)$  for  $j = 1, \dots, n$ , and define the transition kernels

$$Q_j((\mathcal{C}^0, \phi^0), A) = \mathbb{P}((\mathcal{C}, \phi) \in A | f_j(\mathcal{C}, \phi) = f_j(\mathcal{C}^0, \phi^0), X_{1:n} = x_{1:n}).$$

Clearly, this conditional distribution has a discrete part and a “continuous” part:  $j$  can be placed in one of the clusters  $c \in \mathcal{C} \setminus j$ , in which case  $\phi$  is completely determined, or  $j$  can be placed in a separate cluster, in which case  $\phi_{\{j\}}$  is free. (Strictly speaking, it is continuous only if  $H$  is continuous.) Recall that  $H$  has density  $h$  with respect to  $\nu$ . With respect to a measure on  $\mathcal{U}$  (that is, on the space of  $(\mathcal{C}, \phi)$  pairs) consisting of a unit point mass at these discrete points, and  $\nu$  on the “continuous” part, the

density of the conditional distribution is

$$p(\mathcal{C}, \phi \mid \mathcal{C} \setminus j, \phi \setminus j, x_{1:n}) \propto \begin{cases} f_{\phi'_c}(x_j)(|c| + \gamma) & \text{if } c \in \mathcal{C} \setminus j \text{ and } c \cup j \in \mathcal{C} \\ f_\eta(x_j) \gamma \frac{V_n(t+1)}{V_n(t)} h(\eta) & \text{if } \{j\} \in \mathcal{C} \text{ and } \phi_{\{j\}} = \eta \in \Theta \\ 0 & \text{otherwise,} \end{cases}$$

where  $t = |\mathcal{C} \setminus j|$  and  $\phi' = \phi \setminus j$ . (Note that sampling from this distribution would be easy if we could compute  $m(x_j) = \int f_\eta(x_j) h(\eta) \nu(d\eta)$  and sample from  $\pi(\eta|x_j) \propto f_\eta(x_j) h(\eta)$ , however, since we are in the non-conjugate case, this is not possible.) In order to construct a move for this conditional distribution, the algorithm uses another auxiliary variable technique, resulting in the probabilities as given in the algorithm; see the next section for details.

#### 4.4.3.3 A simple variable-dimension move

Here, we consider — in an abstract setting — a particular auxiliary variable technique for a simple variable-dimension move; then we apply it to the situation at hand.

Let  $a_1, \dots, a_t \geq 0$  and  $g : \mathcal{Y} \rightarrow [0, \infty)$ ,  $\mathcal{Y} \subset \mathbb{R}^\ell$ , such that  $\kappa = \int_{\mathcal{Y}} g(y) \nu(dy) < \infty$ , where  $\nu$  is a sigma-finite Borel measure on  $\mathcal{Y}$ , and define  $c = \sum_i a_i + \kappa$ . Define  $X \in \{0, 1, \dots, t\}$  and  $Y \in \mathcal{Y} \cup \{*\}$ , where  $*$  is an arbitrary element not in  $\mathcal{Y}$ , such that

$$\begin{aligned} \mathbb{P}(X = i) &= a_i/c \text{ for } i = 1, \dots, t, & \mathbb{P}(X = 0) &= \kappa/c, \\ \mathbb{P}(Y = * \mid X > 0) &= 1, & Y \mid X = 0 & \text{has density } g(y)/\kappa. \end{aligned}$$

Suppose we know  $a_1, \dots, a_t$  and can evaluate  $g(y)$  for any  $y \in \mathcal{Y}$ , and we would like to sample  $(X, Y)$ , but we cannot easily evaluate  $\kappa$  or  $c$ . A transition kernel for which

the distribution of  $(X, Y)$  is invariant can be constructed using auxiliary variables, as follows.

Choose a distribution  $R$  on  $\mathcal{Y}$  that we can easily sample from, and that has a density  $r$  with respect to  $\nu$  that is strictly positive and can be evaluated at any  $y \in \mathcal{Y}$ . Fix  $s \in \{1, 2, \dots\}$ . Let  $M \sim \text{Uniform}\{1, \dots, s\}$  independently of  $X$  and  $Y$ . Given  $X, Y, M$ : if  $X > 0$  then let  $Z_1, \dots, Z_s \stackrel{\text{iid}}{\sim} R$ , and if  $X = 0$  then let  $Z_M = Y$  and  $Z_i \stackrel{\text{iid}}{\sim} R$  for  $i \in \{1, \dots, s\} \setminus \{M\}$ . Denote  $Z = (Z_1, \dots, Z_s)$ .

The following move preserves the distribution of  $(X, Y)$ : from state  $(x, y)$ , sample  $Z = z \mid X = x, Y = y$ , then sample  $X, Y \mid Z = z$ , and discard  $z$ . To sample  $Z \mid X, Y$ , we can just sample  $M, Z \mid X, Y$  as described above, and discard  $M$ . It turns out that sampling  $X, Y \mid Z$  is also easy, since it can be shown that this is a discrete distribution placing mass  $a_x/C$  at  $(x, *)$  for  $x = 1, \dots, t$ , and mass  $\frac{1}{Cs} \frac{g(z_i)}{r(z_i)}$  at  $(0, z_i)$  for  $i = 1, \dots, s$ , where  $C = \sum_{x=1}^t a_x + \frac{1}{s} \sum_{i=1}^s \frac{g(z_i)}{r(z_i)}$  is the normalizing constant. (Note: if  $H$  is not continuous, then multiple  $z_i$ 's may coincide, and in this case their mass contributions accumulate.) In fact, from this we can see that, *algorithmically*, there is actually no need to sample  $M$ : by symmetry, the same transition kernel results from always choosing  $M = 1$ .

We apply this to finish showing that the non-conjugate algorithm has the correct invariant distribution, continuing the discussion from the end of Section 4.4.3.2. Replace  $a_1, \dots, a_t$  by  $(a_c : c \in \mathcal{C} \setminus j)$  where  $a_c = f_{\phi'_c}(x_j)(|c| + \gamma)$ . Take  $\mathcal{Y} = \Theta$  and  $g(\eta) = f_\eta(x_j) \gamma \frac{V_n(t+1)}{V_n(t)} h(\eta)$  for  $\eta \in \Theta$ , and take  $R = R_{x_j}$ . Then, the move above is precisely the reassignment move made for each  $j = 1, \dots, n$  in Algorithm 4.4.2.

# CHAPTER FIVE

---

## Experiments with the MFM



In this chapter, we apply the MFM model in various ways, and compare it to the corresponding DPM. Our objectives are to empirically demonstrate:

- (1) posterior consistency properties,
- (2) similarities and differences between the MFM and DPM,
- (3) correctness of the inference algorithms and agreement with previously published results, and
- (4) broad applicability and ease of application.

To this end, we consider two families of mixtures: univariate normal and bivariate skew-normal. We apply these mixture models to a variety of datasets — simulated and real — and consider the posterior behavior of the models with respect to density estimation, clustering, and the number of components and clusters. We also compare the mixing time of the MCMC samplers for the MFM and DPM.

For a number of other experiments with the MFM, see [Nobile \(1994\)](#), [Phillips and Smith \(1996\)](#), [Richardson and Green \(1997\)](#), [Stephens \(2000\)](#), [Green and Richardson \(2001\)](#), and [Nobile and Fearnside \(2007\)](#).

## 5.1 Univariate normal mixtures

Here, we apply the MFM and DPM models in the most standard setting: mixtures of univariate normals. Partly in order to demonstrate agreement with previously published results, some of the experiments are modeled after those of [Richardson and Green \(1997\)](#) and [Green and Richardson \(2001\)](#), who developed reversible jump

MCMC samplers for both the MFM and DPM models (in the univariate normal setting) and compared the two models empirically, by considering density estimates, deviance, mean posterior cluster sizes, the entropy of posterior partitions, and the posteriors on the number of components and clusters.

Rather than using reversible jump for inference, we use the (non-conjugate) incremental Gibbs sampler from Section 4.4.2. In addition to comparing density estimates, mean posterior cluster sizes, and the posteriors on the number of components and clusters, we also consider Hellinger distance to the true density (if known), likelihood on a held-out test set, sample clusterings, and pairwise probability matrices.

The characteristics we observe here appear to be true for other families of component distributions as well; for instance, in Section 5.2, we observe similar properties with mixtures of bivariate skew-normal distributions.

### 5.1.1 Data

We consider the following data distributions and datasets.

- *Standard normal.* The data is  $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} \mathcal{N}(0, 1)$ . This can be interpreted as a mixture with a single component.
- *Four components.* The data is  $X_1, \dots, X_n$  i.i.d. from  $\sum_{i=1}^4 \pi_i \mathcal{N}(x|\mu_i, \sigma_i^2)$  where  $\pi = (0.44, 0.25, 0.3, 0.01)$ ,  $\mu = (5, 5, 8, 10)$ , and  $\sigma = (1.2, 0.2, 0.6, 0.2)$ . See Figure 5.1. Note that two of the components have the same mean, and one of the components has fairly small weight: 0.01.
- *Classic galaxy dataset.* The “galaxy dataset” (Roeder, 1990) is a standard

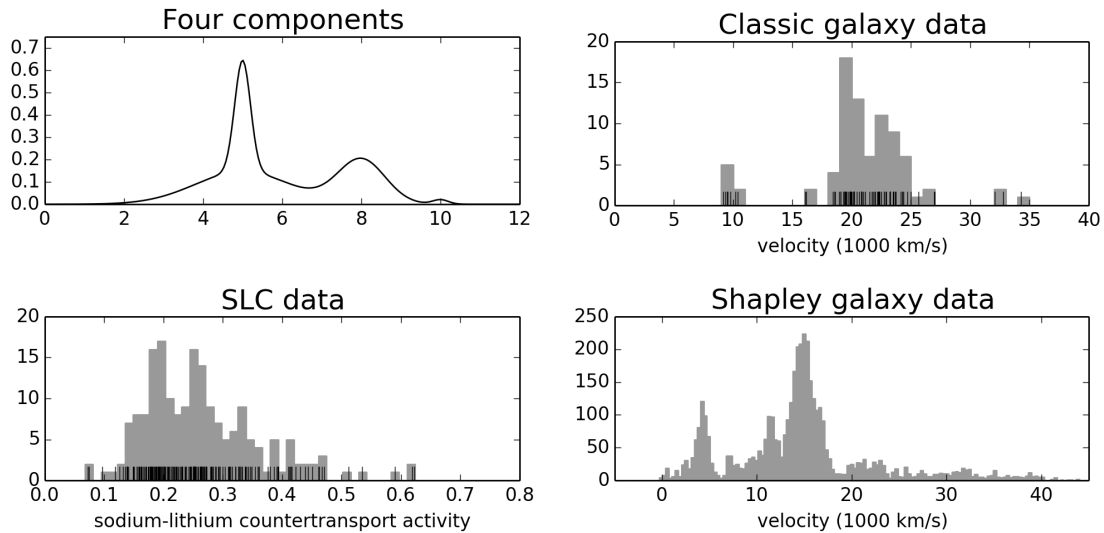


Figure 5.1: Datasets used, in addition to standard normal. Histograms are shown for the classic galaxy, Shapley galaxy, and SLC datasets, accompanied by rug plots for classic galaxy and SLC. Note: The Shapley dataset also has a small amount of additional data above 45,000 km/s, extending in a long tail up to nearly 80,000 km/s.

benchmark dataset for mixture models. It consists of measurements of the velocities of 82 galaxies in the Corona Borealis region. See Figure 5.1.

- *Shapley galaxy dataset*. This is a larger dataset of the same type as the classic galaxy dataset, containing measurements of the velocity of 4215 galaxies in the Shapley supercluster, a large concentration of gravitationally-interacting galaxies (Drinkwater et al., 2004). See Figure 5.1. The clustering tendency of galaxies continues to be a subject of interest in astronomy. Since the assumption of normal components appears to be incorrect, this dataset provides the opportunity to see how the MFM and DPM behave on a relatively large amount of real data with a potentially misspecified model.
- *Sodium-lithium countertransport (SLC) dataset*. The “SLC dataset” (Roeder, 1994) consists of the SLC activity level of 190 individuals; it is another benchmark dataset for mixture models, particularly for estimating the number of

components. See Figure 5.1.

### 5.1.2 Model description

To enable comparison, we use the exact same model as in Richardson and Green (1997) and Green and Richardson (2001). The component densities are univariate normal,

$$f_{\theta}(x) = f_{\mu,\lambda}(x) = \mathcal{N}(x|\mu, \lambda^{-1}) = \sqrt{\frac{\lambda}{2\pi}} \exp\left(-\frac{\lambda}{2}(x - \mu)^2\right),$$

and the base measure (prior)  $H$  on  $\theta = (\mu, \lambda)$  is

$$\mu \sim \mathcal{N}(\mu_0, \sigma_0^2), \quad \lambda \sim \Gamma(a, b),$$

independently (where the gamma distribution  $\Gamma(a, b)$  is parametrized to have density  $\frac{b^a}{\Gamma(a)} x^{a-1} e^{-bx}$ ). Further, a hyperprior is placed on  $b$ , by taking  $b \sim \Gamma(a_0, b_0)$ . The remaining parameters are set to  $\mu_0 = (\max\{x_i\} + \min\{x_i\})/2$ ,  $\sigma_0 = \max\{x_i\} - \min\{x_i\}$ ,  $a = 2$ ,  $a_0 = 0.2$ , and  $b_0 = 10/\sigma_0^2$ . Note that the parameters  $\mu_0$ ,  $\sigma_0$ , and  $b_0$  are functions of the observed data  $x_1, \dots, x_n$ . See Richardson and Green (1997) for the rationale behind these parameter choices. (Note: This choice of  $\sigma_0$  may be a bit too large, affecting the posteriors on the number of clusters and components — see Section 5.1.6 — however, we stick with it to enable comparisons to Richardson and Green (1997).)

For the MFM, following Richardson and Green (1997), we take  $K \sim \text{Uniform}\{1, \dots, 30\}$  and  $\gamma = 1$  for the finite-dimensional Dirichlet parameters, as a default. The Shapley galaxy data, however, is highly heterogeneous and  $K \leq 30$  is too restrictive,

so on it we use

$$p(k) = \begin{cases} c & \text{if } k \in \{1, \dots, 30\} \\ c/(k-30)^2 & \text{if } k > 30 \end{cases}$$

where  $c = 1/(30 + \pi^2/6)$ , and  $\gamma = 1$ . For the DPM, we consider two versions: a fixed concentration parameter of  $\alpha = 1$  (DPM-fixed), and a random concentration parameter with  $\alpha \sim \text{Exponential}(1)$  (DPM-random).

Note that taking  $\mu$  and  $\lambda$  to be independent results in a non-conjugate prior. Such a prior is appropriate when the location of the data is not informative about the scale (and vice versa).

### 5.1.3 Approximate inference

The hyperprior on  $b$  could be handled in at least two different ways. First, integrating  $b$  out would give rise to a distribution on  $(\mu, \lambda)$  that we could have called  $H$ , and in fact, in the case above  $b$  could be analytically integrated out since it has been given a conjugate prior. Alternatively, we could use Gibbs sampling (i.e., adding  $b$  to the state of the Markov chain, running the sampler given  $b$  as usual, and periodically sampling  $b$  given everything else). We use the latter approach in order to illustrate that there is no difficulty in handling more general hyperprior structures.

Since the prior is non-conjugate, we use the non-conjugate sampler described in Section 4.4.2. For simplicity, we use a single auxiliary variable ( $s = 1$ ) and use the prior  $H$  (given the current value of the hyperparameter  $b$ ) for the auxiliary variable distributions  $R_x$ . For the  $\phi$  move (that is, moving  $\phi_c = (\mu_c, \lambda_c)$  according to the single-cluster posterior  $\pi(\phi_c|x_c)$ , for  $c \in \mathcal{C}$ ), we use partial conjugacy to sample from  $\pi(\mu_c|\lambda_c, x_c)$  and then  $\pi(\lambda_c|\mu_c, x_c)$ .

In contrast to [Richardson and Green \(1997\)](#), we do not restrict the parameter space in any way (e.g., forcing the component means to be ordered to obtain identifiability). All of the quantities we consider are invariant to the labeling of the clusters. See [Jasra et al. \(2005\)](#) for an interesting discussion on this point.

For the DPM with random  $\alpha$  (DPM-random), we numerically integrate out the prior on  $\alpha$  in a pre-computation step, as described by [MacEachern \(1998\)](#).

For some of the datasets, multiple runs were performed using different sets of samples; for others, a single run was performed with the whole dataset. In each run, the sampler executed 100,000 burn-in sweeps, and 200,000 sample sweeps (for a total of 300,000). Judging by traceplots and running averages of various statistics, this appeared to be sufficient for mixing. The number  $t = |\mathcal{C}|$  and sizes ( $|c| : c \in \mathcal{C}$ ) of the clusters were recorded after each sweep. To reduce memory storage requirements, the full state  $(\mathcal{C}, \phi)$  of the chain was recorded once every 10 sweeps. For each run, the seed of the random number generator was initialized to the same value for both the MFM and DPM.

For a dataset of size  $n$ , the sampler used for these experiments took approximately  $2.5 \times 10^{-6} n$  seconds per sweep, using a 2.80 GHz processor with 6 GB of RAM.

#### 5.1.4 Density estimation

Following [Green and Richardson \(2001\)](#), we use Equation [4.2.15](#) to estimate the density based on samples from the posterior of  $\mathcal{C}, \phi \mid x_{1:n}$ . The 20,000 recorded samples (out of the 200,000 sample sweeps) are used.

Overall, the MFM and DPM yield highly similar density estimates. For large  $n$ , this makes sense, since both models are consistent for the density; the degree of similarity for smaller  $n$  as well is interesting. The estimated densities using the DPM with fixed  $\alpha$  and random  $\alpha$  were very similar, so the density estimation results displayed here are for fixed  $\alpha$ , unless mentioned otherwise.

#### 5.1.4.1 Estimated densities

Figures 5.2 and 5.3 show the estimated densities for increasing amounts of data ( $n \in \{50, 200, 500, 2000\}$ ) from the four component distribution and Shapley dataset, respectively; for the Shapley dataset, random subsets of size  $n$  were used. Figure 5.4 shows the estimated densities for the classic galaxy, (full) Shapley galaxy, and SLC datasets.

#### 5.1.4.2 Hellinger distance

For the simulated data sets (i.e., standard normal and four component), the true density is known, so we can evaluate the performance of a given density estimate by considering its distance from the true density. Hellinger distance is one of the standard metrics used in theoretical studies of the convergence of density estimates, so it makes for a natural choice in this context. The Hellinger distance between densities  $f$  and  $g$  is

$$H(f, g) = \left( \frac{1}{2} \int \left| \sqrt{f(x)} - \sqrt{g(x)} \right|^2 dx \right)^{1/2}.$$

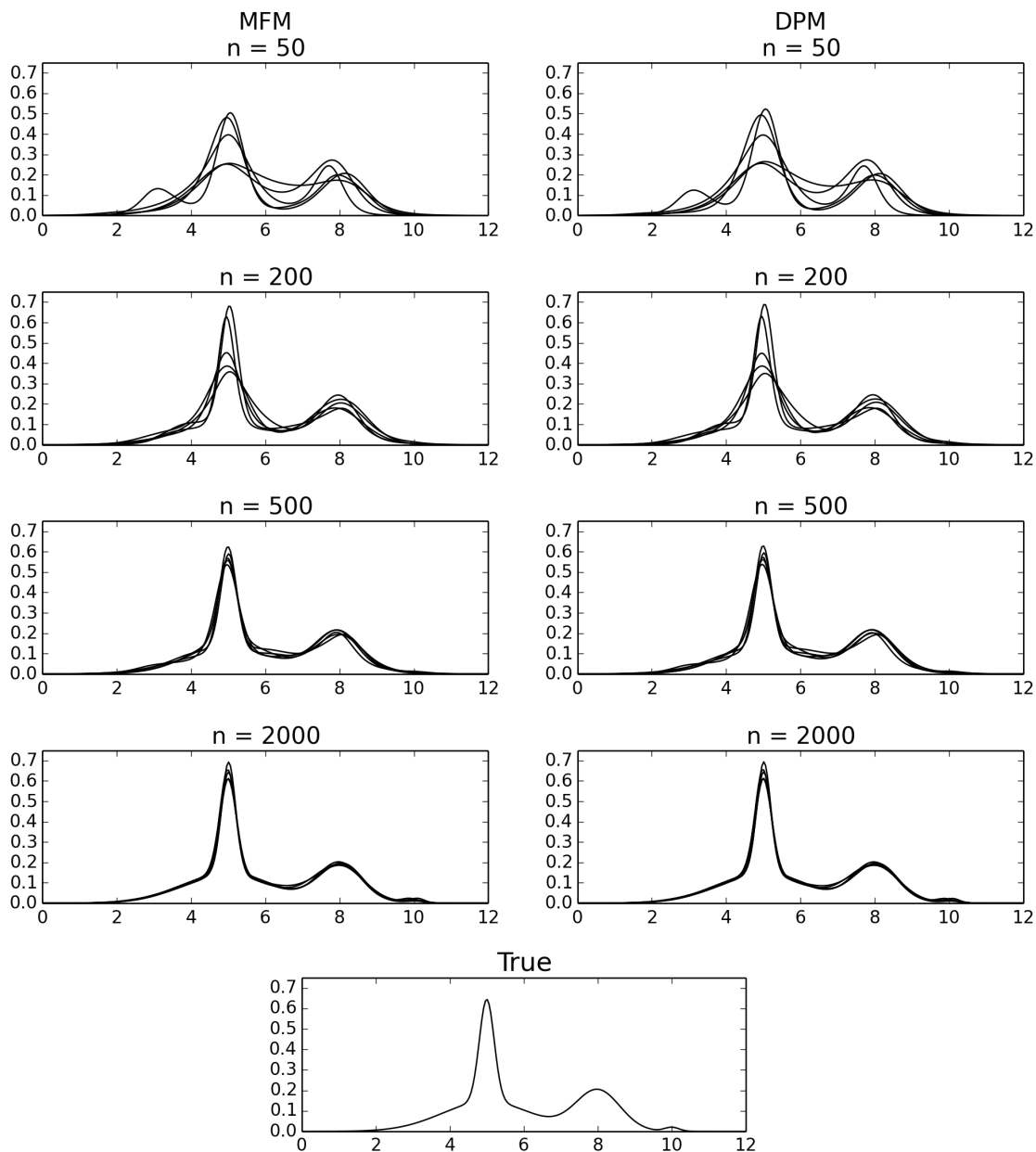


Figure 5.2: Estimated densities for MFM (left) and DPM (right) on data from the four component distribution (bottom plot). Results for 5 datasets are displayed for each  $n \in \{50, 200, 500, 2000\}$ . For each dataset, the MFM and DPM estimated densities are nearly indistinguishable. As  $n$  increases, the estimates appear to converge to the true density, as expected.



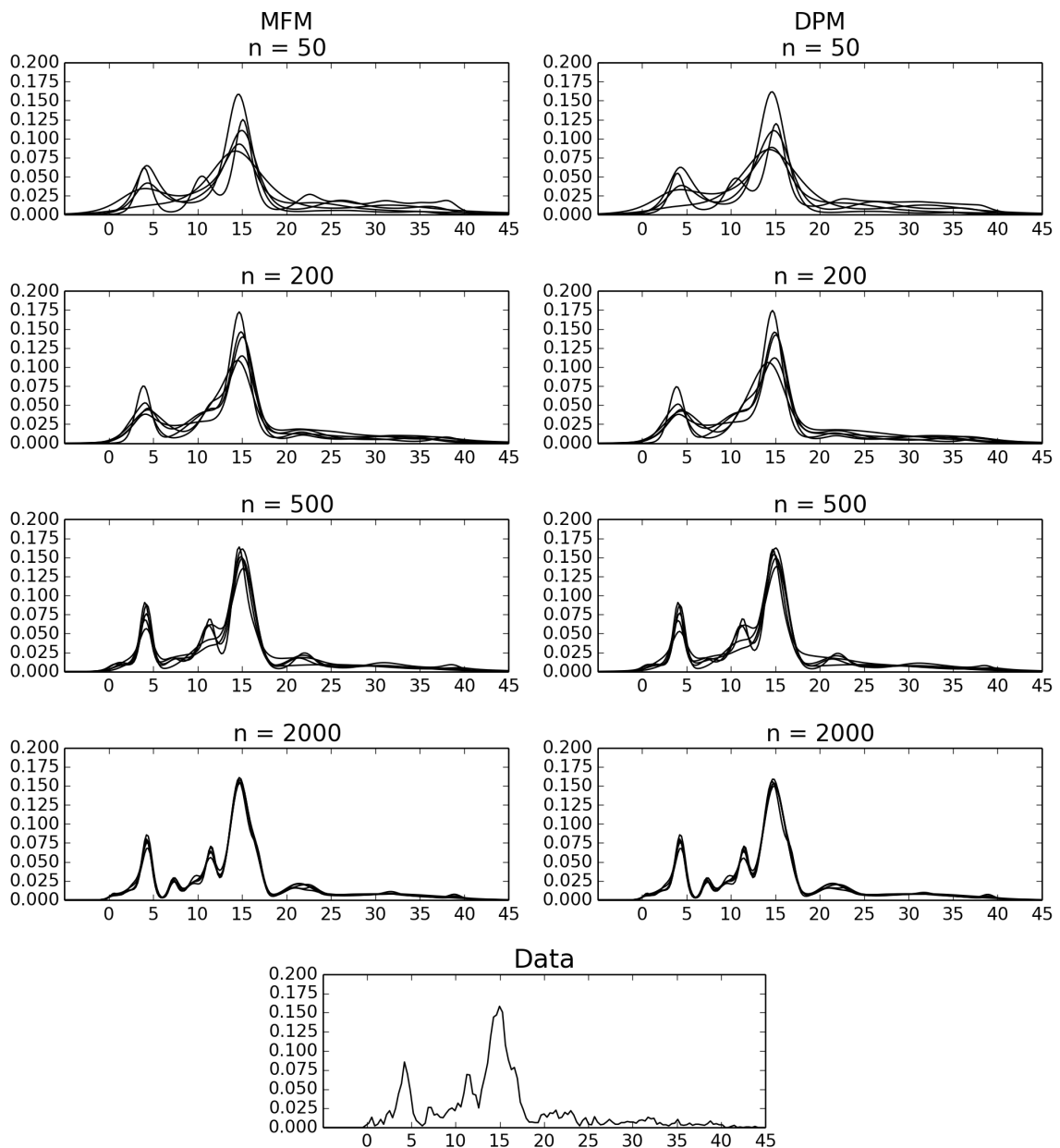


Figure 5.3: Estimated densities for MFM (left) and DPM (right) on velocity data (units: 1000 km/s) from the Shapley galaxy dataset (bottom plot, histogram normalized to a density). For each  $n \in \{50, 200, 500, 2000\}$ , five random subsets of size  $n$  were used. Again, for each set of data, the MFM and DPM estimated densities are nearly indistinguishable.

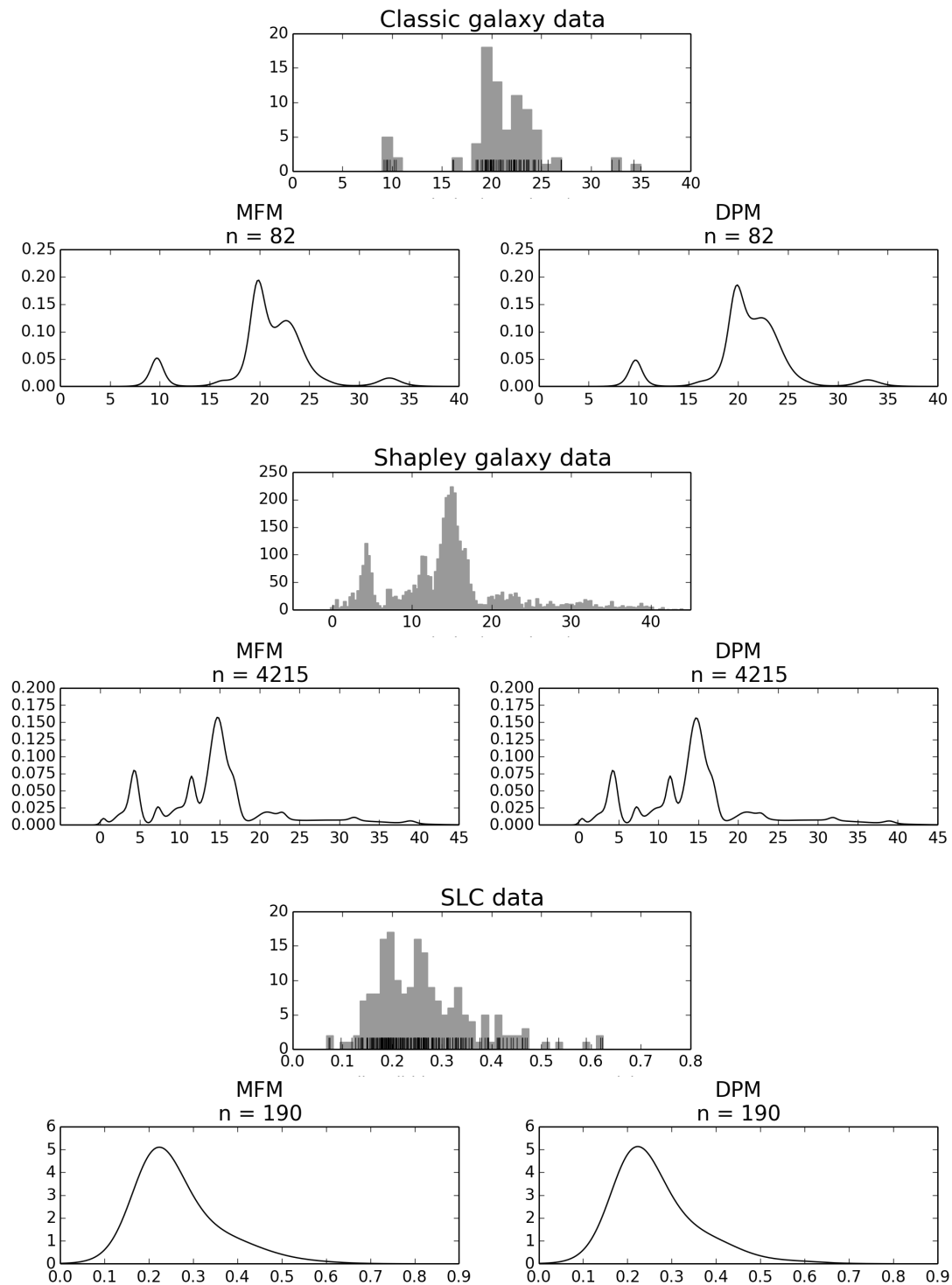


Figure 5.4: Estimated densities using MFM and DPM on the classic galaxy, Shapley galaxy, and SLC datasets. Once again, the MFM and DPM results are visually very similar. For the classic galaxy dataset, we can compare the MFM estimated density to the results of [Richardson and Green \(1997\)](#), and they appear to be similar.

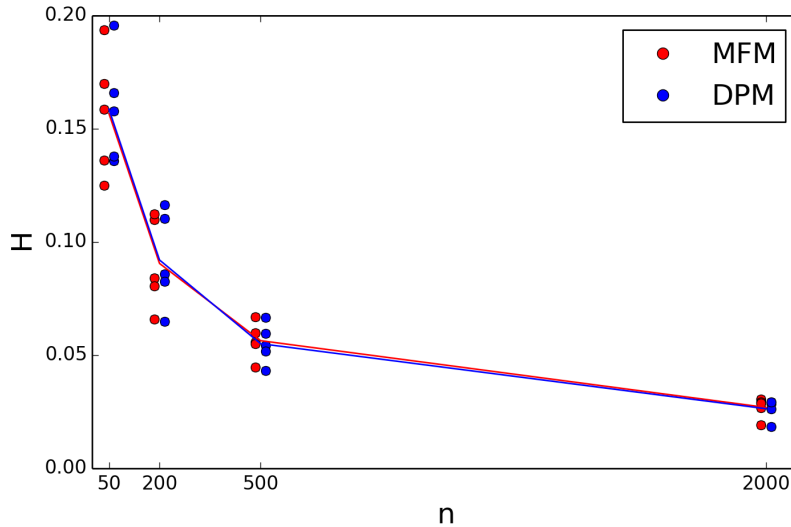


Figure 5.5: Estimated Hellinger distances for MFM (red, left) and DPM (blue, right) density estimates of the four component distribution. For each  $n \in \{50, 200, 500, 2000\}$ , five independent datasets of size  $n$  were used, and the lines connect the averages of the estimated distances.

If we can sample from  $g$ , and  $g(x) > 0$  whenever  $f(x) > 0$ , then we can estimate the squared Hellinger distance by using a simple Monte Carlo approximation,

$$H(f, g)^2 = \frac{1}{2} \int \left| \sqrt{f(x)} - \sqrt{g(x)} \right|^2 dx \approx \frac{1}{2N} \sum_{i=1}^N \left( \sqrt{\frac{f(Y_i)}{g(Y_i)}} - 1 \right)^2$$

where  $Y_1, \dots, Y_N \stackrel{\text{iid}}{\sim} g$ . For the simulated data sets, the true density is simply a mixture of normals, so it is easy to sample from (and it is strictly positive).

Figure 5.5 shows the estimated Hellinger distance between the true density and the estimated density of the four component distribution, for increasing amounts of data ( $n \in \{50, 200, 500, 2000\}$ ). Each distance was estimated with  $N = 10^3$  independent samples from the true density; the same samples were used to evaluate both MFM and DPM. The MFM and DPM results are very similar.

### 5.1.4.3 Log-likelihood on a test set

For a real dataset, of course, we cannot compute the Hellinger distance. In this situation, it is common to evaluate performance by measuring how well the estimated density fits a held-out subset of the data. A standard measure of fit is the log-likelihood on the held-out data.

To this end, we compute the density estimate  $\hat{f}_n$  as before using a “training set” of  $n$  points from a dataset, and compute the log-likelihood

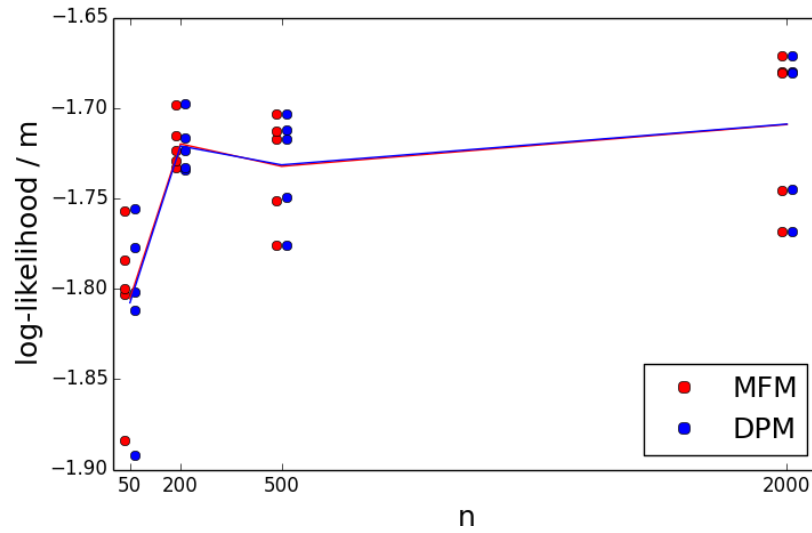
$$\ell(\hat{f}_n; y_{1:m}) = \sum_{j=1}^m \log \hat{f}(y_j)$$

where  $y_1, \dots, y_m$  is a “test set” of  $m$  points from the dataset, disjoint from the training set.

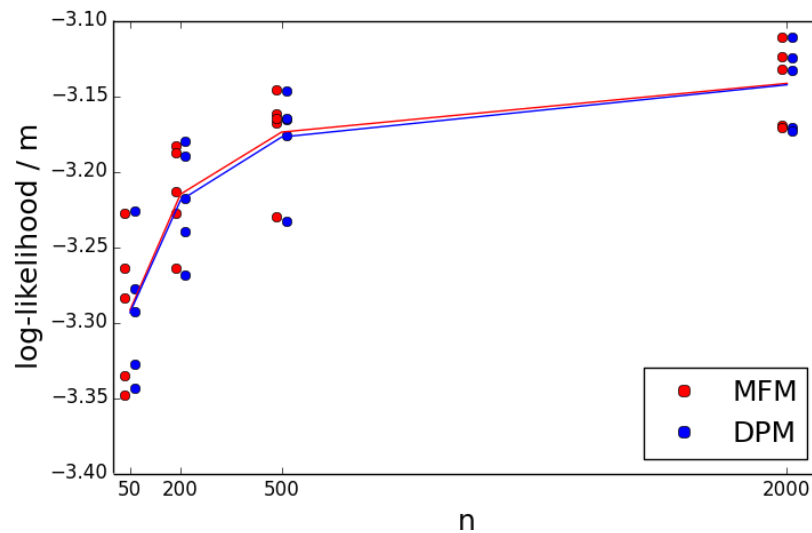
For this experiment, we used the Shapley galaxy dataset, and also, for comparison, the four component distribution. Figure 5.6 shows  $\frac{1}{m}\ell(\hat{f}_n; y_{1:m})$  for the MFM and DPM estimated densities, using training sets of increasing size ( $n \in \{50, 200, 500, 2000\}$ ), and test sets of size  $m = 1000$ . For each  $n$ , five different training sets (each with a different test set) were used; the same training and test sets were used for the MFM and DPM. The MFM and DPM results are very similar.

### 5.1.5 Clustering

In addition to density estimation, mixture models are very frequently used for clustering. In this section, we examine various properties of the MFM and DPM posteriors that are relevant to clustering. In contrast with density estimation, some significant



(a) Four component distribution



(b) Shapley galaxy dataset

Figure 5.6: Test set log-likelihood (divided by  $m$ ) for MFM (red, left) and DPM (blue, right) density estimates. For each  $n \in \{50, 200, 500, 2000\}$ , five training sets of size  $n$  were used, each with a test set of size  $m = 1000$ .

differences emerge between the MFM and DPM with respect to clustering.

It is well-known that partitions, or “clusterings”,  $\mathcal{C}$  sampled from the DPM posterior tend to have small transient clusters (e.g., [West et al. \(1994\)](#), [Lartillot and Philippe \(2004\)](#), [Onogi et al. \(2011\)](#), and others). Here, we empirically observe that MFMs do not exhibit this behavior. When using DPMs for clustering, these small extra clusters can be a nuisance. This makes MFMs a potentially attractive alternative to DPMs for model-based clustering.

The differences we observe are what one might expect, based on the rather extreme differences in the conditional distribution of the partition *a priori*, given the number of clusters (as discussed in [Section 4.3.3](#)). Roughly speaking, for a given number of clusters, the DPM has a strong preference for lower entropy partitions.

In addition to comparing the MFM’s clustering behavior to the DPM with fixed  $\alpha$ , we also compare to the DPM with random  $\alpha$ . When the data is well-modeled by a finite mixture, it appears that putting a prior on  $\alpha$  enables the DPM to adapt somewhat and reduce the tendency to make extra clusters, however, empirically it does seem to still have this tendency to some degree.

#### 5.1.5.1 Sample clusterings

First, we look at some samples of clusterings from the posterior. To visualize the clusters, we make a scatter plot in which the x-values are the data points and the y-values are the cluster assignments; the y-values are then slightly vertically jittered in order to distinguish nearby points.

[Figure 5.7](#) shows three representative samples for each of the three different mod-

els — the MFM, the DPM with random  $\alpha$  (DPM-random), and the DPM with fixed  $\alpha$  (DPM-fixed) — on  $n = 500$  points from the standard normal distribution (i.e., a mixture with one component). The MFM samples typically have a single cluster. The DPM-random samples usually have a single cluster, but occasionally have one or more small extra clusters. The DPM-fixed samples almost always have multiple small extra clusters. It should be noted that the extra clusters are “transient”, in the sense that they do not consistently contain the same points.

Figure 5.8 shows samples from the three models, on  $n = 2000$  points from the four component distribution. All three models fairly consistently have 4 dominant clusters corresponding to the 4 components, accounting for the majority of the data. With respect to extra clusters, we observe the same trend as before: the MFM usually has 4 or 5 clusters, DPM-random usually has 4 to 6, and DPM-fixed usually has 4 to 8. When the MFM does have extra clusters, they are typically substantial in size, whereas the DPM’s extra clusters range in size and usually some are very small. Also, even though one of the true components has small weight (0.01, so it accounts for only around 20 data points out of 2000), all three models clearly recognize it by creating a separate cluster.

For the classic galaxy dataset (Figure 5.9), all the models tended to have small clusters. This seems attributable to fact that the data set is fairly heterogeneous and rather small ( $n = 82$ ).

For the Shapley galaxy dataset (Figure 5.10), all the models have a fairly large number of clusters (10 to 20). Nonetheless, as before, the DPM models usually have a few tiny clusters in addition. Interestingly, the behavior seems quite different on the long flat tail at the upper range of values — the MFM often puts all of these values together in one cluster, while the DPMs often break it up into several clusters.

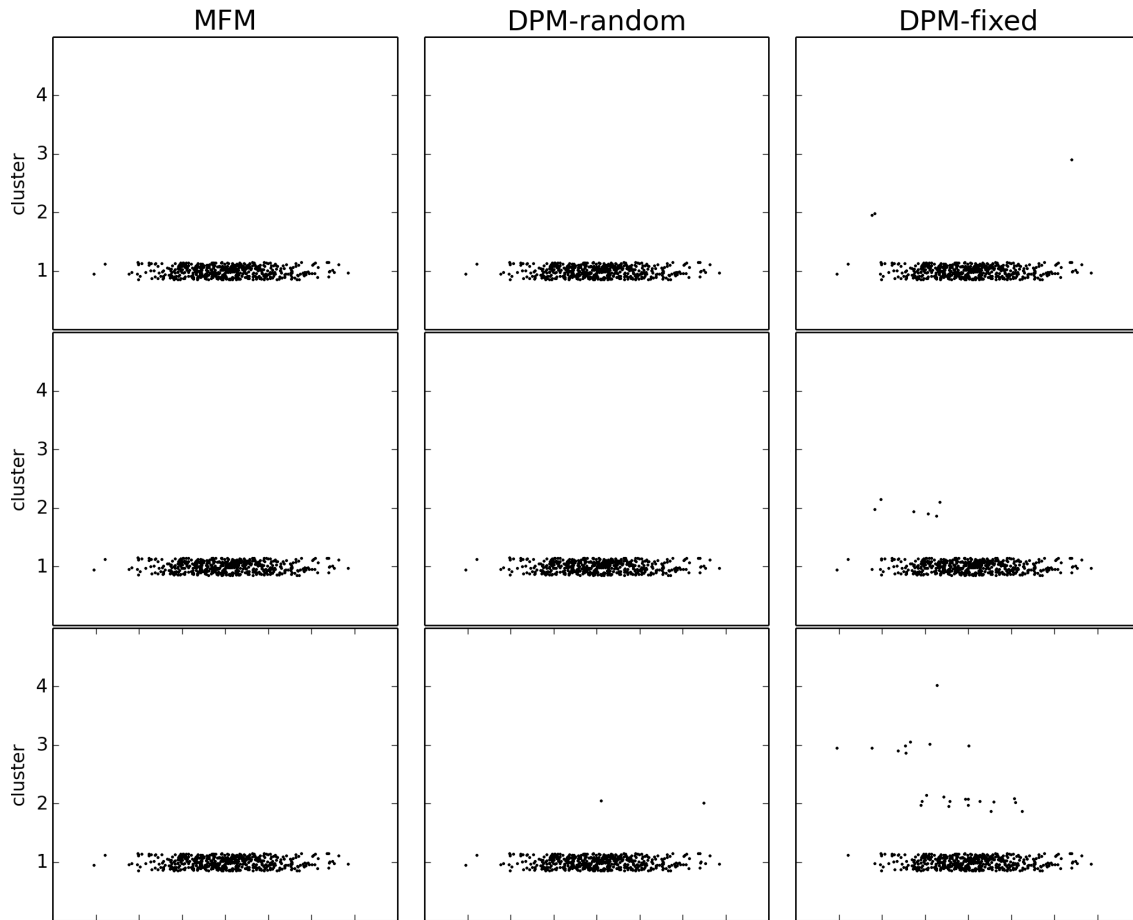


Figure 5.7: Typical sample clusterings from the posterior, on standard normal data. See the text for discussion.

Also, in this case, DPM-fixed tends to have *fewer* clusters than DPM-random (see also Section 5.1.6), apparently due to the prior on  $t$ .

Figure 5.11 shows samples from the three models on the SLC dataset. We observe similar trends as on the one component and four component data.

The preceding heuristic observations are based on looking at individual samples. In the subsequent sections, we consider more principled estimates of various posterior properties.



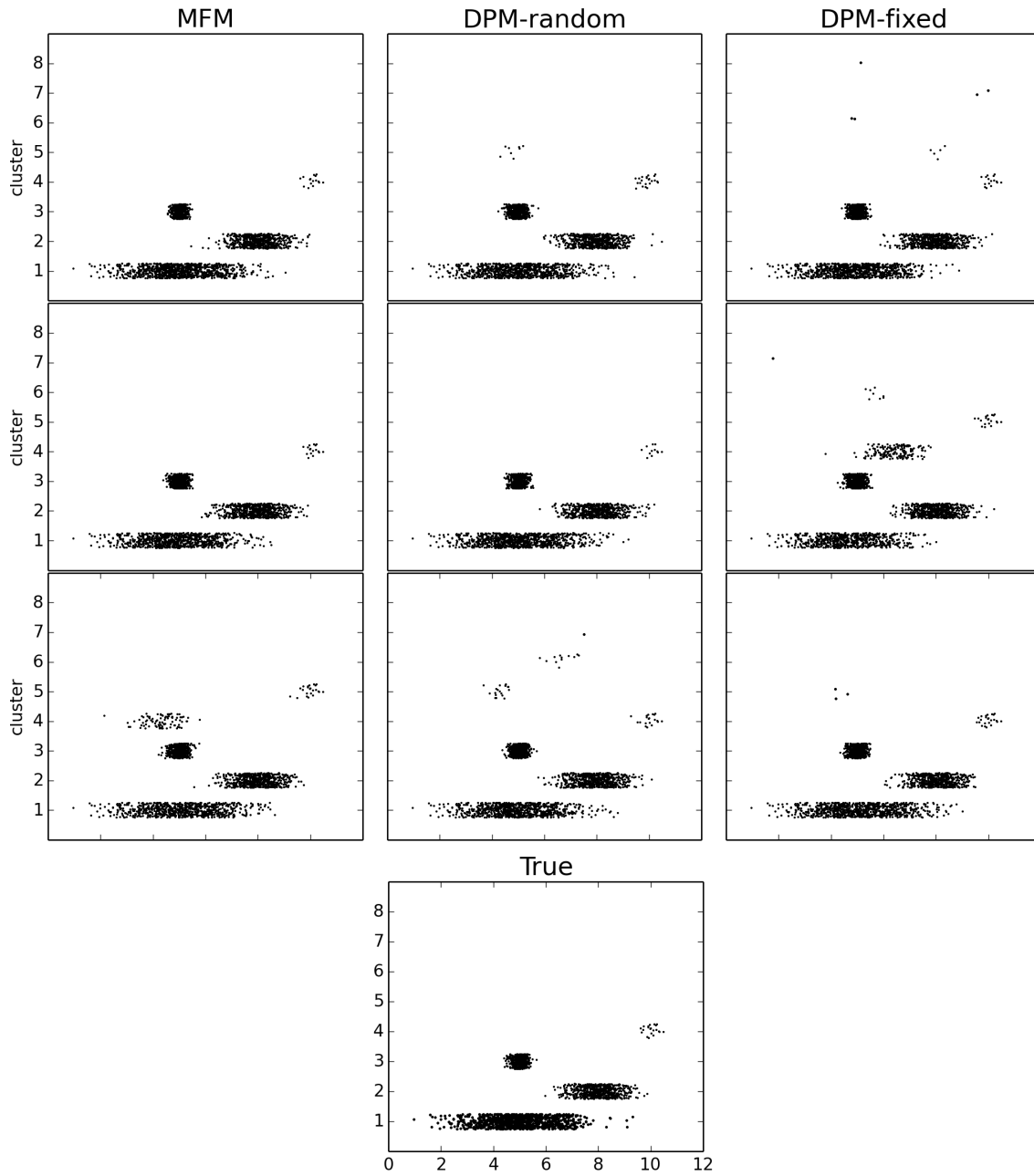


Figure 5.8: Typical sample clusterings from the posterior, on four component data. The true component assignments are shown at bottom. See the text for discussion.

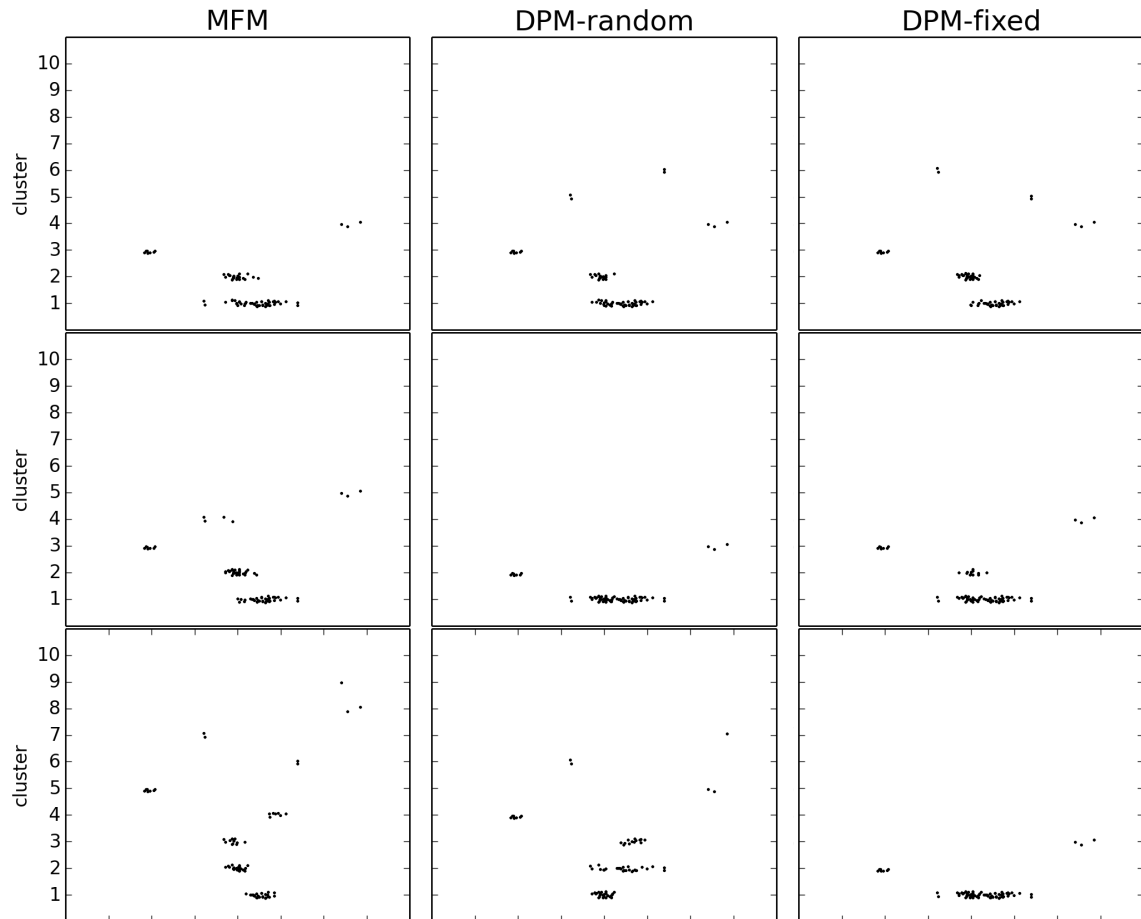


Figure 5.9: Typical sample clusterings from the posterior, on the classic galaxy dataset. See the text for discussion.

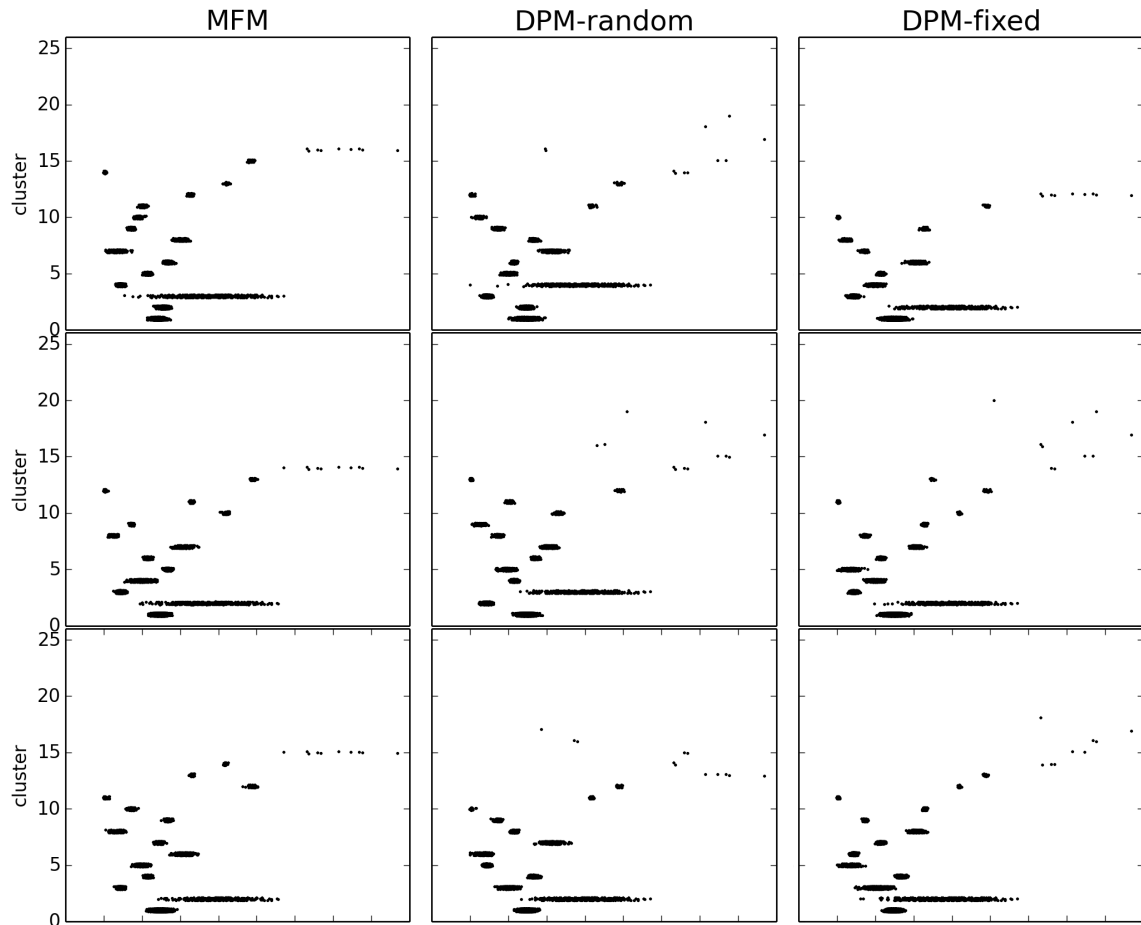


Figure 5.10: Typical sample clusterings from the posterior, on the Shapley galaxy dataset. See the text for discussion.

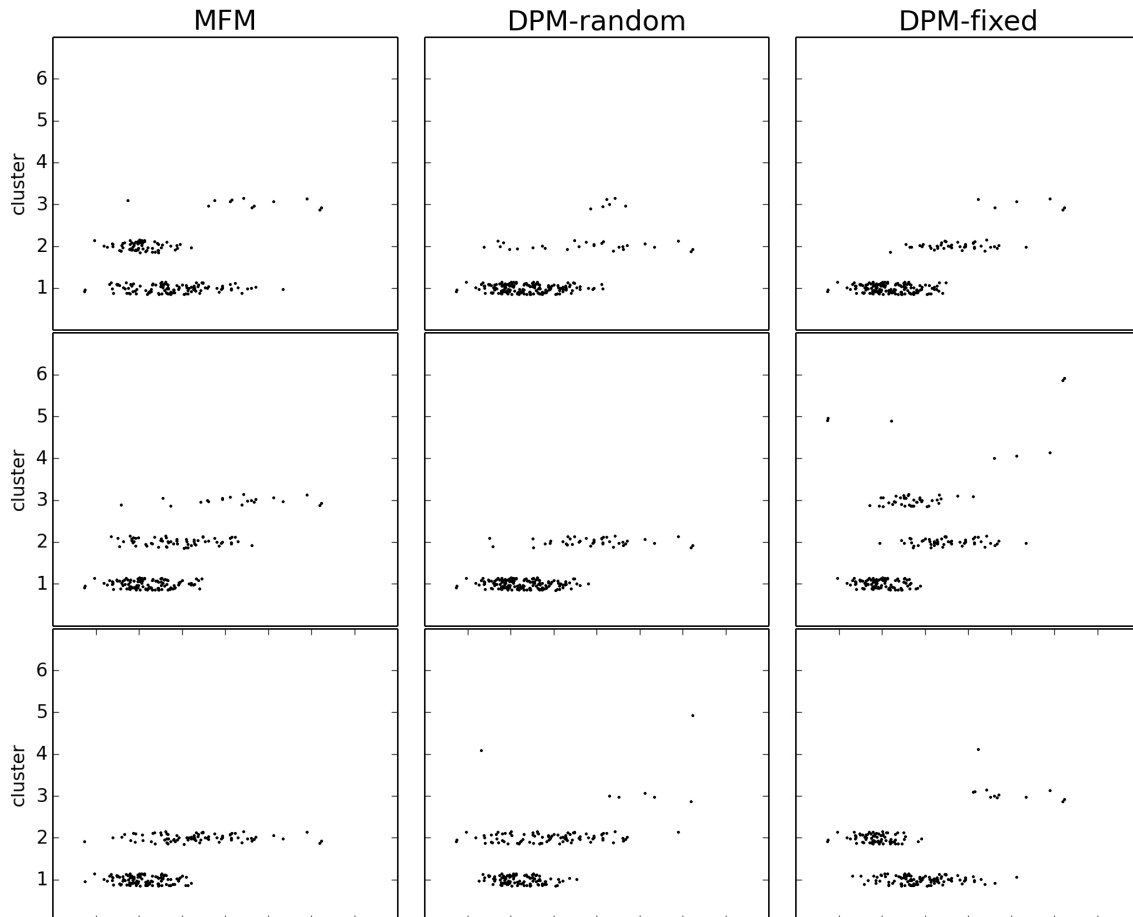


Figure 5.11: Typical sample clusterings from the posterior, on the SLC dataset. See the text for discussion.

### 5.1.5.2 Cluster sizes

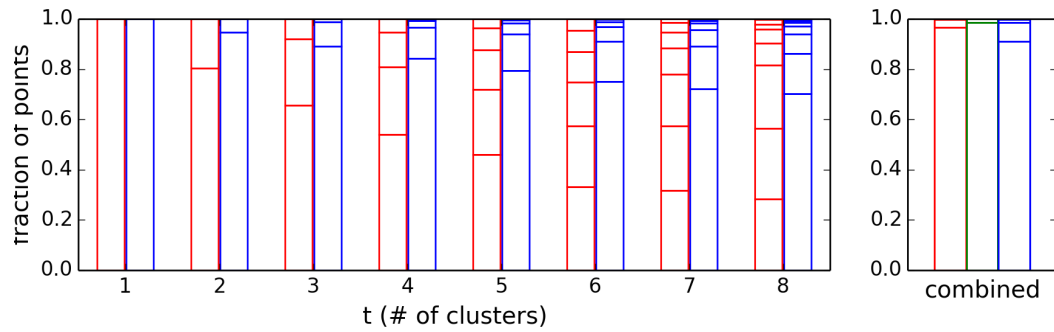
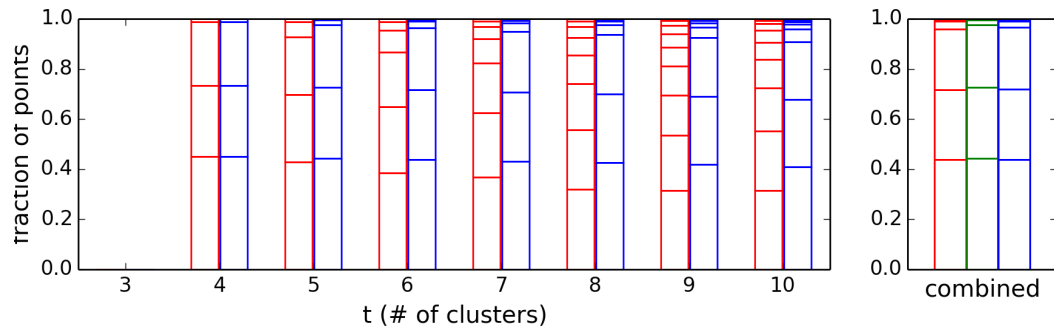
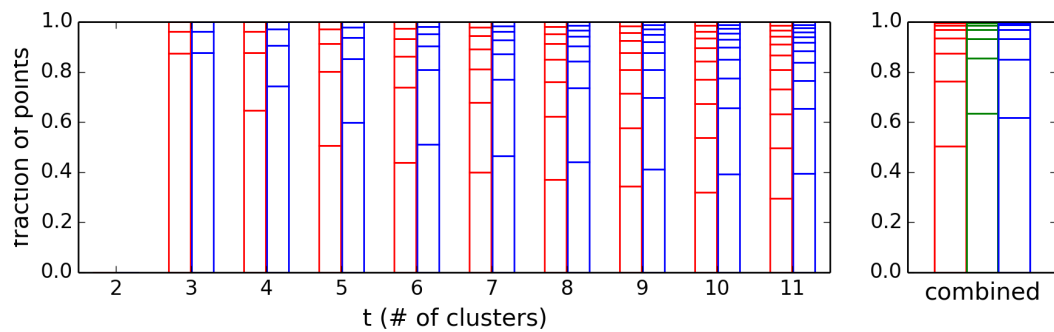
To summarize the relative sizes of the clusters  $c \in \mathcal{C}$  for clusterings  $\mathcal{C}$  sampled from the posterior, we consider the posterior means of the sorted cluster sizes. In other words, let  $N_1 \geq N_2 \geq \dots$  such that  $(N_i : i \in \{1, 2, \dots\}, N_i > 0)$  is a permutation of  $(|c| : c \in \mathcal{C})$ , and consider the posterior means  $\mathbb{E}(N_i \mid x_{1:n})$ . Following [Green and Richardson \(2001\)](#), we also consider the mean sizes given the number of clusters  $t = |\mathcal{C}|$ , that is,  $\mathbb{E}(N_i \mid t, x_{1:n})$ .

Figure 5.12 shows estimates of these conditional and unconditional mean sizes for the standard normal distribution with  $n = 500$ , the four component distribution with  $n = 2000$ , the classic galaxy dataset, and the (full) Shapley galaxy dataset. To visualize these quantities, we draw a box for each  $i = 1, 2, \dots$ , stacked up from largest to smallest. The estimates shown are for values of  $t$  with sufficiently large posterior probability that we have enough samples to compute the estimates. Note that conditional on  $t$ , the DPM with fixed  $\alpha$  (DPM-fixed) and with random  $\alpha$  (DPM-random) are formally equivalent.

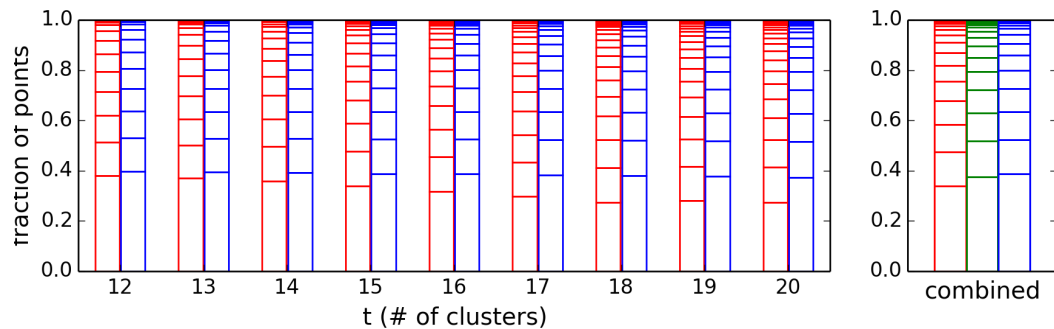
Similarly to [Green and Richardson \(2001\)](#), we very consistently observe that under the MFM, the mean cluster sizes given  $t$  are “more equally-sized” than under the DPM; more precisely, the discrete distribution corresponding to the sizes (i.e., with probabilities  $p_i = N_i/n$ ) has higher entropy under the MFM.

### 5.1.5.3 Pairwise probability matrix

A useful summary of the posterior distribution of the cluster assignments  $\mathcal{C}$  is the matrix containing, for each pair of points, the posterior probability that they belong

(a) Standard normal distribution,  $n = 500$ (b) Four component distribution,  $n = 2000$ 

(c) Classic galaxy dataset



(d) Shapley galaxy dataset

Figure 5.12: Left: Means of the sorted cluster sizes, given  $t$ , for MF (red, left) and DM (blue, right). Right: Means of the sorted cluster sizes, unconditionally, for MF (red, left), DM-random (green, middle), and DM-fixed (blue, right).

to the same cluster. In other words, entry  $(i, j)$  is  $p_{ij} = \mathbb{P}(\exists c \in \mathcal{C} \text{ s.t. } i, j \in c \mid x_{1:n})$ . This matrix is often used in clustering applications (Medvedovic and Sivaganesan, 2002, Kim et al., 2006, Dahl, 2006, Murino et al., 2010).

For each dataset and each model, we estimate this matrix using 5000 samples of  $\mathcal{C}$ . Figure 5.13 displays estimates of these matrices for the four component distribution ( $n = 500$ ), the classic galaxy dataset ( $n = 82$ ), and a random subset of the Shapley galaxy dataset ( $n = 500$ ). The matrices shown are for the DPM with fixed  $\alpha$ ; the random  $\alpha$  case is nearly identical. The indices of the matrix for the four component distribution are ordered by true component origin, then by the value of the data point; the other matrices are ordered by the value of the data point. The MFM and DPM matrices are visually similar in all three cases, however we observe — particularly for the galaxy dataset — that some of the blocks are darker in the DPM matrices (i.e., these pairs are more likely to be together); this is a real effect, not simply a plotting artifact. This seems to be attributable to the higher entropy of MFM clusterings compared to DPM clusterings.

Table 5.1: RMS of the differences between estimated matrices.

|                          | $k = 1$ | $k = 4$ | galaxy | Shapley | SLC  |
|--------------------------|---------|---------|--------|---------|------|
| MFM vs. DPM-fixed        | 0.09    | 0.03    | 0.14   | 0.04    | 0.04 |
| MFM vs. DPM-random       | 0.03    | 0.03    | 0.16   | 0.03    | 0.06 |
| DPM-fixed vs. DPM-random | 0.12    | 0.02    | 0.03   | 0.03    | 0.04 |

Table 5.1 displays the root-mean-square (RMS) of the differences between the estimated matrices for the different models, i.e., for matrices  $(p_{ij})$  and  $(q_{ij})$ , we compute  $(\frac{1}{n^2} \sum_{i,j} (p_{ij} - q_{ij})^2)^{1/2}$ . For the one component data we use  $n = 500$ , and for SLC,  $n = 190$ . To get a sense of the accuracy of these estimated matrices, the RMS distance between two estimates of the MFM matrix based on independent sets of 5000 samples was computed for each dataset, and in each case it was approximately 0.01. For the most part, the differences between the models are small, but there are

some noticeable disparities: the MFM differs from DPM-fixed and DPM-random on the galaxy dataset, meanwhile, DPM-fixed differs from the others on the standard normal data.

### 5.1.6 Number of components and clusters

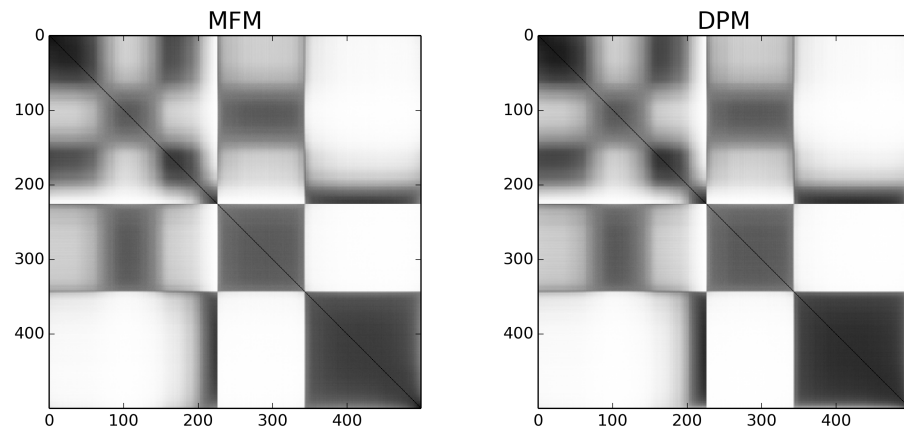
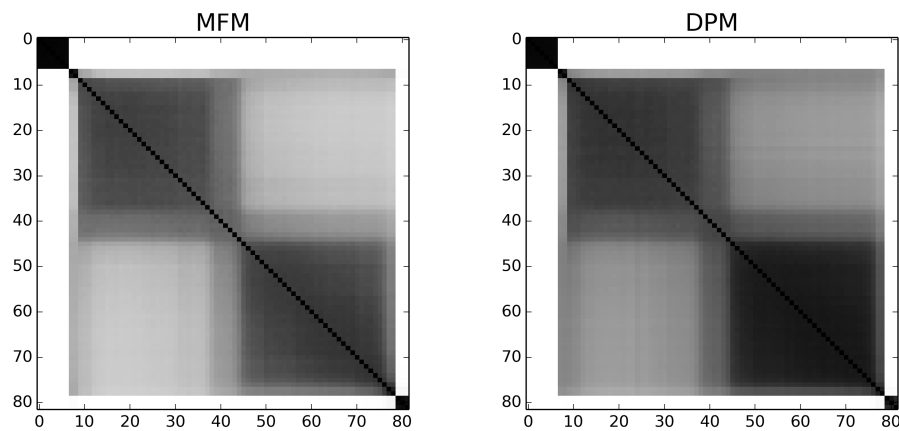
It is tempting to use the posterior of the number of components  $k$  or of the number of clusters  $t$  as a measure of the heterogeneity of the data, however, there are some subtle issues with this that one should bear in mind. (See [Aitkin \(2001\)](#) for a detailed discussion in the context of the galaxy dataset.) In particular,

- (1) these posteriors can be strongly affected by the base measure  $H$ , and
- (2) these posteriors can be highly sensitive to misspecification of the family of component distributions  $\{f_\theta\}$ .

Issue (1) can be seen, for instance, in the case of normal mixtures: it might seem desirable to choose the prior on the component means to have large variance in order to be less informative, however, this causes the posterior of  $t$  to favor smaller values of  $t$ . In the MFM, the same is true for the posterior of  $k$  ([Richardson and Green, 1997](#), [Stephens, 2000](#), [Jasra et al., 2005](#)). With some care, this issue can be dealt with by varying the base measure  $H$  and observing the effect on the posterior; for instance, see [Richardson and Green \(1997\)](#).

Issue (2) is more serious; in practice, we typically cannot expect our choice of  $\{f_\theta : \theta \in \Theta\}$  to contain the true component densities (assuming the data is even from a mixture). When the model is misspecified in this way, the posteriors of  $k$



(a) Four component distribution,  $n = 500$ 

(b) Classic galaxy dataset

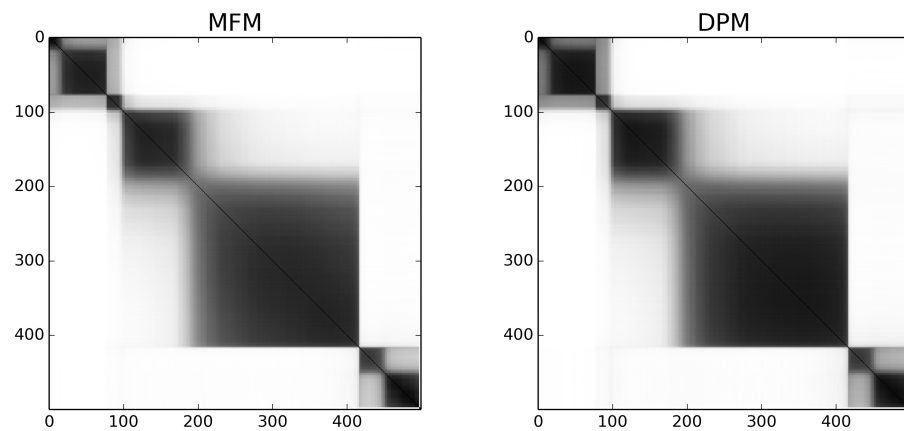
(c) Shapley galaxy dataset (random subset of size  $n = 500$ )

Figure 5.13: Matrices showing, for each pair of data points, the posterior probability of belonging to the same cluster (white is probability 0, black is 1).

and  $t$  will typically be severely affected and depend strongly on  $n$ . For instance, if the model uses mixtures of normals, and the true data distribution is not a finite mixture of normals, then these posteriors will diverge as  $n$  increases. Consequently, the effects of misspecification need to be carefully considered if these posteriors are to be used as measures of heterogeneity.

Here, we consider the posteriors of  $k$  and  $t$ , not for the purpose of measuring heterogeneity, but simply to empirically demonstrate:

- (1) posterior consistency (and inconsistency), assuming correct specification,
- (2) differences between the MFM and DPM posteriors, and
- (3) correctness of the inference algorithms.

The posterior of  $t = |\mathcal{C}|$  is easily estimated from posterior samples of  $\mathcal{C}$ . Figures 5.14 (standard normal data), 5.15 (four component data), 5.16 (classic galaxy data), 5.17 (Shapley galaxy data), and 5.18 (SLC) show estimates of  $p(t|x_{1:n})$  for the MFM, DPM-fixed, and DPM-random. For the standard normal, four component, and Shapley galaxy data, for each  $n \in \{50, 200, 500, 2000\}$  (as well as  $n = 5000$  for the four component data) we show the average over 5 datasets (using random subsets for the Shapley galaxy data).

Under the DPM, the prior on  $t$  diverges as  $n$  grows, and this influences the posterior. Following Green and Richardson (2001), we can eliminate this influence by “tilting” the posterior in postprocessing; that is, since  $p(t|x_{1:n}) \propto p(x_{1:n}|t)p(t)$ , we can switch to any other prior  $q(t)$  on  $t$  simply by multiplying the estimated posterior

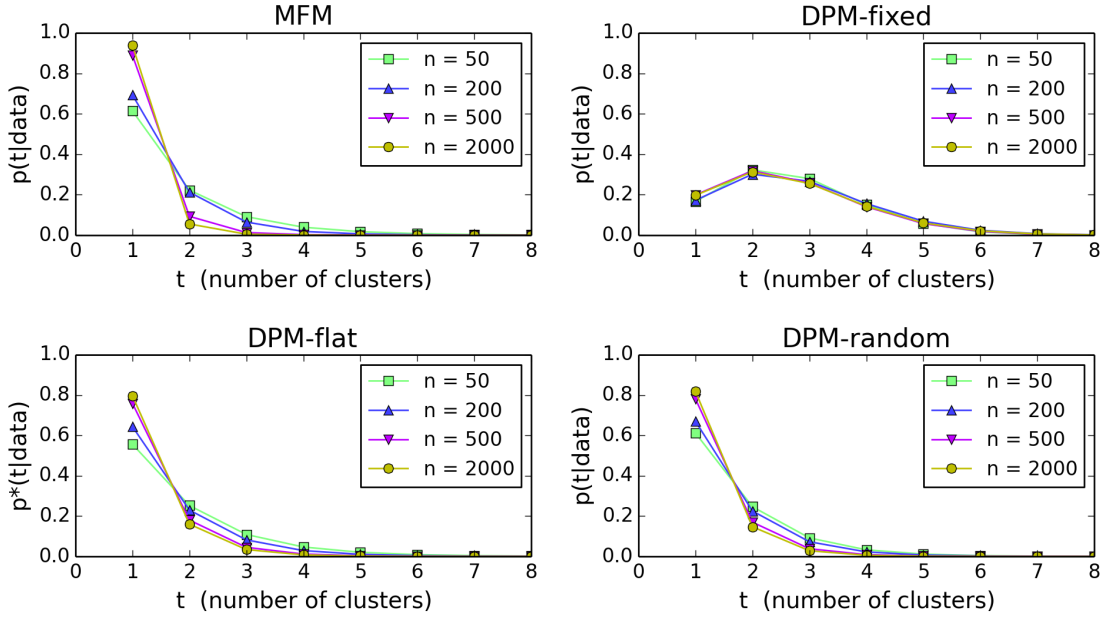


Figure 5.14: Posterior distribution of  $t$  on standard normal data.

of  $t$  by  $q(t)/p(t)$  and renormalizing. To this end, define

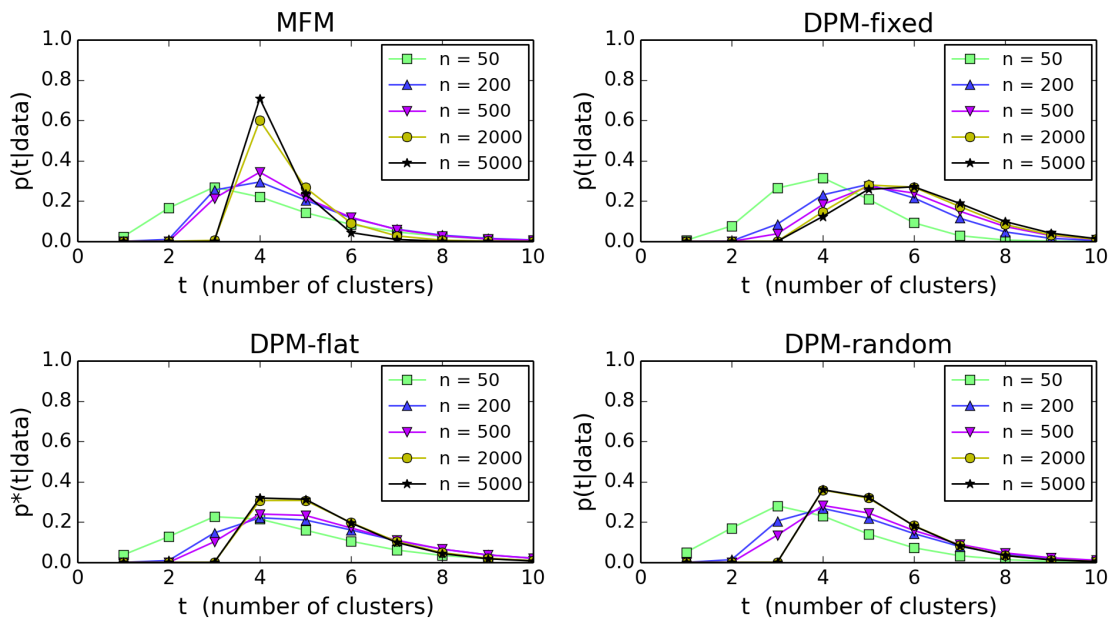
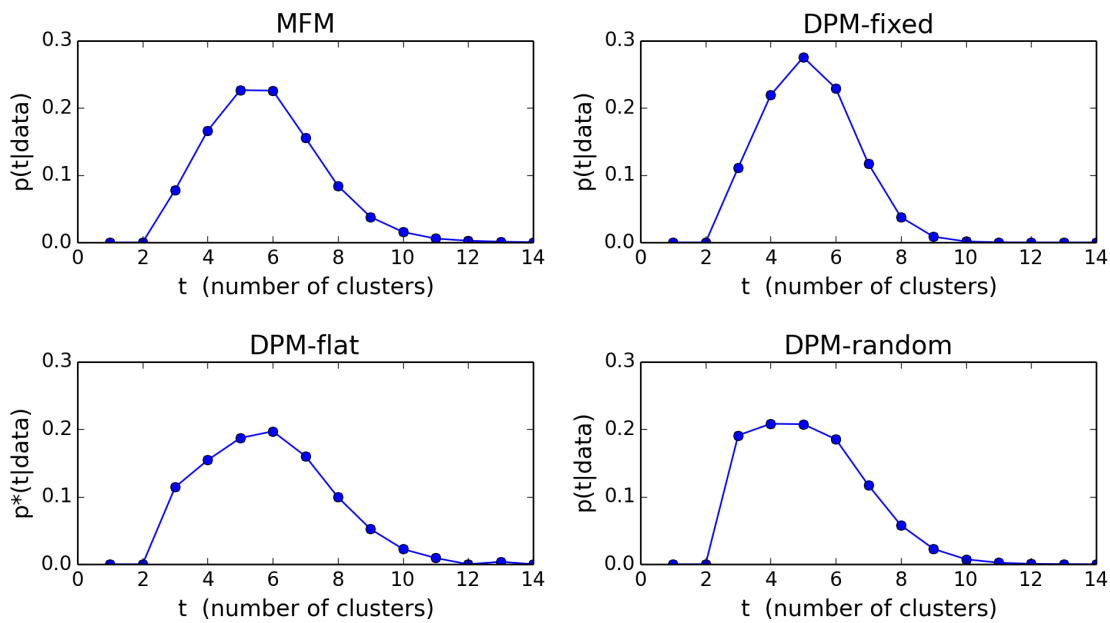
$$p^*(t | x_{1:n}) \propto \frac{p(t | x_{1:n})}{p_n(t)}$$

where  $p_n(t)$  is the prior on  $t$  when there are  $n$  data points; we refer to this as “DPM-flat”. This can be interpreted as corresponding to an (improper) uniform prior on  $t$ ; in practice it is numerically equivalent to a uniform prior on the range of  $t$  values sampled. In the figures mentioned above, we also display  $p^*(t|x_{1:n})$  (DPM-flat).

For the DPM, the posterior of the number of components  $k$  is always trivially a point mass at infinity. For the MFM, to compute the posterior of  $k$ , note that

$$p(k|x_{1:n}) = \sum_{t=1}^{\infty} p(k|t, x_{1:n})p(t|x_{1:n}) = \sum_{t=1}^{\infty} p(k|t)p(t|x_{1:n}),$$

by Equation 4.2.9, and the formula for  $p(k|t)$  is given by Equation 4.2.7; therefore, it is easy to transform our estimate of the posterior of  $t$  into an estimate of the posterior of  $k$ . Figure 5.19 shows the posterior of  $k$  alongside that of  $t$  for the

Figure 5.15: Posterior distribution of  $t$  on four component data.Figure 5.16: Posterior distribution of  $t$  on the classic galaxy dataset.

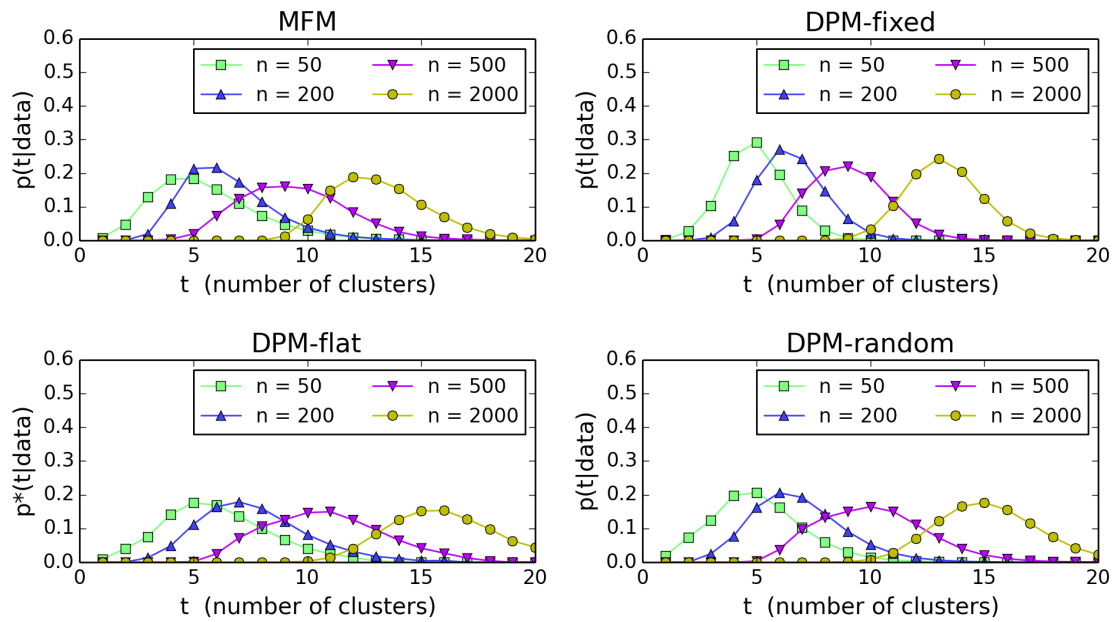


Figure 5.17: Posterior distribution of  $t$  on the Shapley galaxy dataset.

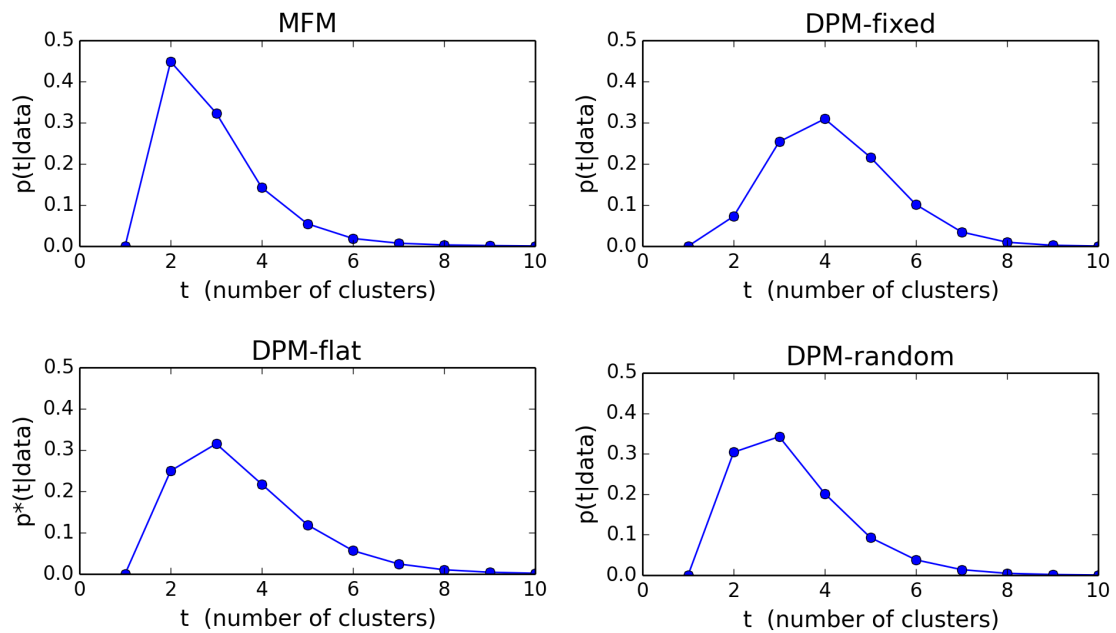


Figure 5.18: Posterior distribution of  $t$  on the SLC dataset.

MFM, for each dataset. We know that the difference between  $p(t|x_{1:n})$  and  $p(k|x_{1:n})$  becomes negligible as  $n$  grows (see Section 4.3.2); empirically, in these experiments the difference is modest for  $n = 50$  and becomes indistinguishable for larger values of  $n$ . Note that the mass of  $p(k|x_{1:n})$  is always shifted “to the right” (i.e. to higher values), compared to  $p(t|x_{1:n})$ .

For the classic galaxy dataset, we can compare the MFM posterior estimated here with the estimate of Richardson and Green (1997); since the exact same model is used, the results should be very similar. Indeed, as Table 5.2 shows, they are.

Table 5.2: MFM posterior on  $k$  for the classic galaxy dataset.

| $k$  | 1     | 2     | 3     | 4     | 5     | 6     | 7     |
|------|-------|-------|-------|-------|-------|-------|-------|
| Here | 0.000 | 0.000 | 0.067 | 0.139 | 0.188 | 0.194 | 0.156 |
| R&G  | 0.000 | 0.000 | 0.061 | 0.128 | 0.182 | 0.199 | 0.160 |

| 8     | 9     | 10    | 11    | 12    | 13    | 14    | 15    |
|-------|-------|-------|-------|-------|-------|-------|-------|
| 0.107 | 0.066 | 0.038 | 0.021 | 0.011 | 0.006 | 0.003 | 0.002 |
| 0.109 | 0.071 | 0.040 | 0.023 | 0.013 | 0.006 | 0.003 | 0.002 |

On the finite mixture data, the MFM posteriors on  $k$  and  $t$  appear to be concentrating at the true number of components (although perhaps slowly on the four component data), and DPM-fixed certainly does not appear to be concentrating. DPM-random and DPM-flat are somewhere in-between — it is not apparent whether they are concentrating at the true number of components; we would conjecture that they will not concentrate.

These empirical observations are consistent with what we know from theory: the MFM is consistent for the number of components (see Section 4.3.1.3) and DPM-fixed is not (see Chapters 2 and 3).

Note that the four component data is rather challenging for this task, since one of the components has a weight of only 0.01; this may contribute to the slow

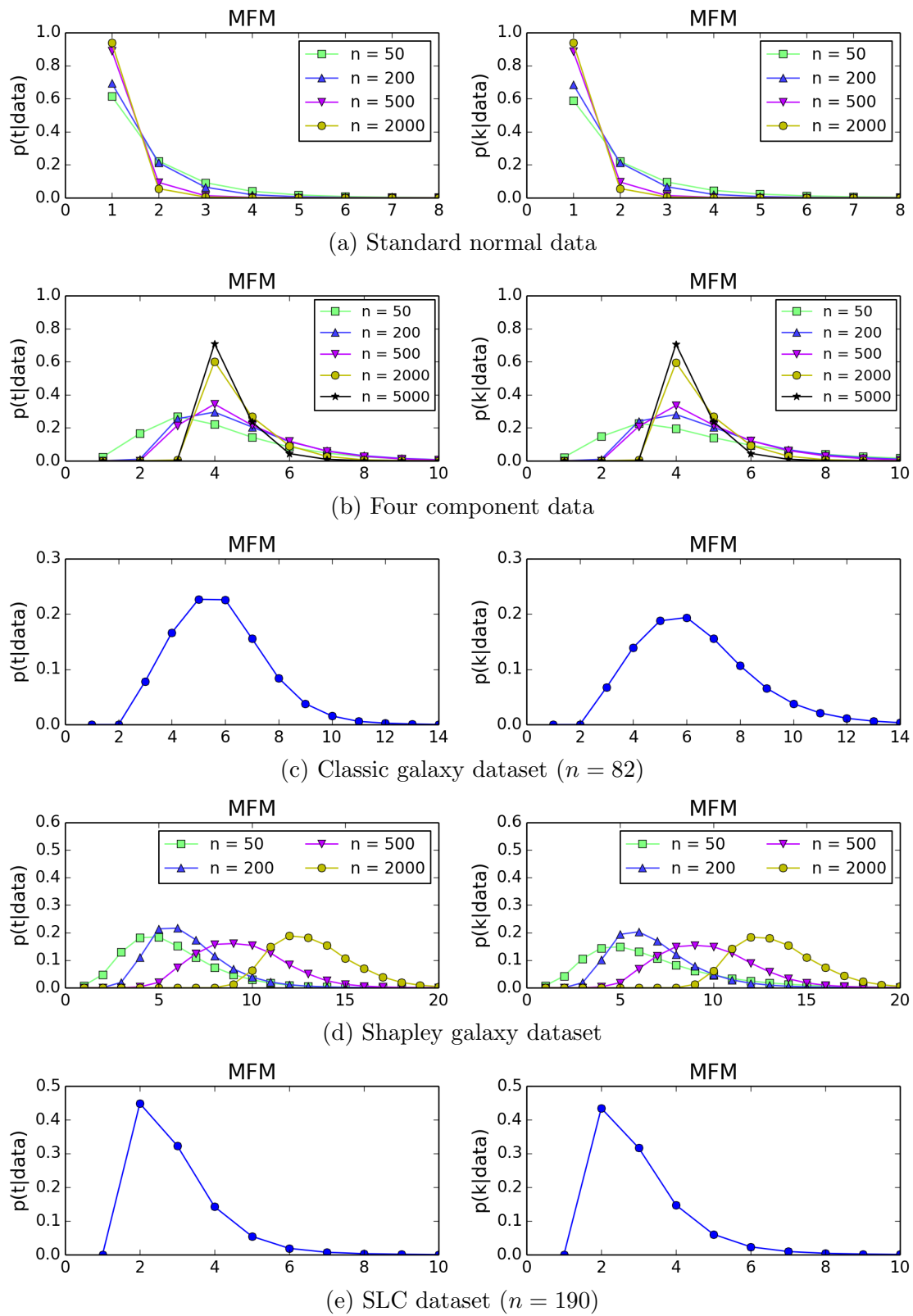


Figure 5.19: Posterior distributions of  $t$  (left) and  $k$  (right) for the MFM.

concentration of the MFM posterior at  $k = 4$ . For comparison, in Section 5.2.6, the MFM posterior concentrates much more quickly for data from a three-component mixture of bivariate skew-normals.

The SLC dataset is often used as a benchmark for methods estimating the number of components in a finite mixture. The sodium-lithium countertransport (SLC) activity of red blood cells is associated with hypertension, and is thought to have a genetic basis. Assuming a normal model, the objective is to determine whether a “simple dominance model” (two components, with weights  $p^2 + 2pq$  and  $q^2$ ) or an “additive model” (three components, with weights  $p^2$ ,  $2pq$ , and  $q^2$ ) is more appropriate; see (Roeder, 1994) for details. Some methods find more evidence for two components, while others indicate three components — there is no clear consensus (Woo and Sriram, 2006, Roeder, 1994, Ishwaran et al., 2001, Chen et al., 2012). The MFM results shown here (Figures 5.18 and 5.19) are consistent with either two or three components (as are the DPM-random and DPM-flat results), however, as expected, using the DPM-fixed posterior on  $t$  would appear to be rather misleading.

It is interesting to note that on the Shapley galaxy dataset, DPM-fixed favors a smaller number of clusters than the others; by comparing it with DPM-flat, we can see that this is due to the prior on the number of clusters.

### 5.1.7 Mixing issues

As described in Section 4.4 (and shown empirically in this chapter), it is straightforward to adapt partition-based DPM samplers to the MFM. An issue that arises with the incremental MCMC samplers is that when  $n$  is large, the mixing time for the MFM can be significantly worse than for the DPM.



To illustrate, in Figure 5.20, we display examples of traceplots of the number of clusters  $t$  for posterior samples using the MFM and DPM, on  $n = 50$  and  $n = 2000$  data points from the four component distribution. For visualization purposes, we uniformly jitter the  $t$  values within  $t \pm 0.25$ , and we display only every 10th sample. Only the burn-in period ( $10^5$  sweeps) is displayed. DPM-fixed is shown; DPM-random is similar.

When  $n = 50$ , the mixing is fine for both MFM and DPM, however, when  $n = 2000$  the MFM mixing is substantially worse. Out of the 5 runs with  $n = 2000$ , the example shown was a particularly bad run in this respect, but this behavior is not too uncommon, based on previous experiments.

The explanation for this seems to be that — as we know from Section 4.3.3 — the DPM likes having small clusters, while the MFM does not. Consequently, the MFM takes a longer time to create or destroy substantial clusters by reassigning one element at a time, since in order to do so it must move through the regions of low probability where there are small clusters. We have observed similar behavior for the conjugate prior algorithm also (Section 4.4.1).

While traceplots of  $t$  are particularly useful for illustrating this issue, it is worth mentioning that they are a bit misleading with regard to DPM mixing. Since the DPM eagerly creates and destroys small transient clusters,  $t$  is significantly more variable than it would be if these tiny clusters were ignored, and consequently, the trace of  $t$  indicates a level of mixing that is artificially high. (Adding independent noise to any trace will appear to improve mixing, in an artificial way.) It seems that mixing would be more appropriately assessed by using a statistic that is not strongly affected by tiny clusters, for instance, the entropy of the clustering,  $-\sum_{c \in \mathcal{C}} \frac{|c|}{n} \log \frac{|c|}{n}$ , or the sum of the squares of the sizes,  $\sum_{c \in \mathcal{C}} |c|^2$ . Nonetheless, by considering the

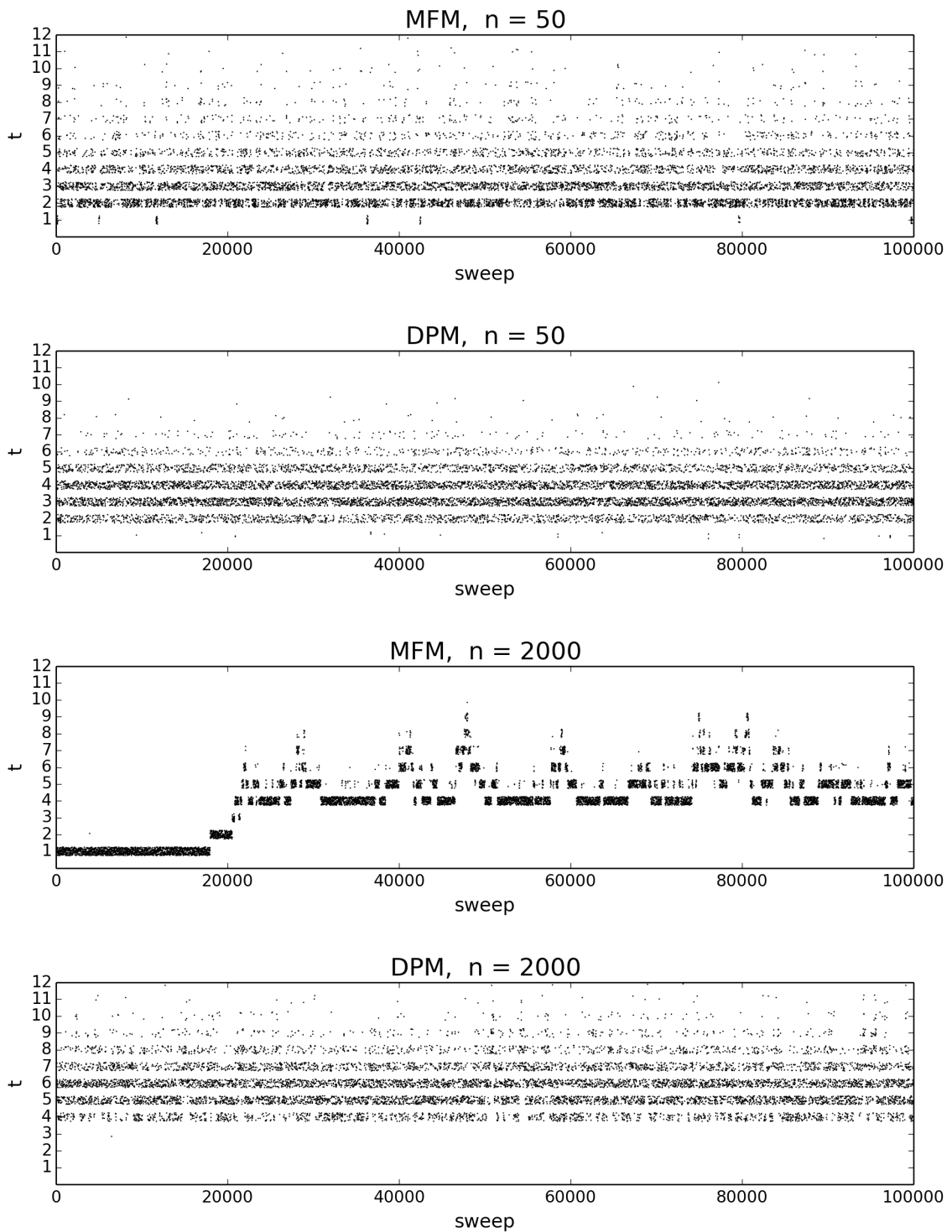


Figure 5.20: Examples of traceplots of the number of clusters  $t$  for the first  $10^5$  sweeps (the burn-in period), on four component data.

traces of various other statistics, it seems that the DPM does indeed mix significantly better than the MFM when  $n$  is large.

A natural solution would be to use a split-merge sampler (Dahl, 2003, 2005, Jain and Neal, 2004, 2007); although we have not explored this, we expect it to be straightforward to adapt existing DPM split-merge samplers to the MFM.

## 5.2 Multivariate skew-normal mixtures

In order to illustrate the flexibility with which the MFM model can be used, in this section we apply it to mixtures of multivariate skew-normals. This example is fairly general, in the following respects.

- *Fully non-conjugate.* For the univariate normal mixtures, we used a non-conjugate prior, but one in which each parameter separately had a conjugate prior; here, we use a prior with no such conjugacy properties.
- *Unbounded  $K$ .* We use a prior on the number of components  $K$  under which  $K$  has no *a priori* upper bound (and indeed,  $p_K(k) > 0$  for all  $k = 1, 2, \dots$ ).
- *Multivariate.* We use a multivariate family of component distributions. Going from univariate to multivariate is routine, due to the convenient form of the samplers. For visualization purposes, we show results for the bivariate case.

Overall, the results are quite similar to the case of univariate normal mixtures, so we discuss only a subset of the posterior properties considered in the previous section.

Introduced by [Azzalini and Dalla Valle \(1996\)](#) and [Azzalini and Capitanio \(1999\)](#), the multivariate skew-normal distribution  $\mathcal{SN}(\xi, Q, w)$  with parameters  $\xi \in \mathbb{R}^d$  (location),  $Q \in \mathbb{R}^{d \times d}$  positive definite (scale and correlation), and  $w \in \mathbb{R}^d$  (skew), has density

$$\mathcal{SN}(x \mid \xi, Q, w) = 2\mathcal{N}(x \mid \xi, Q) \Phi(w^T S^{-1}(x - \xi))$$

for  $x \in \mathbb{R}^d$ , where  $S \in \mathbb{R}^{d \times d}$  is the diagonal matrix with  $S_{ii} = \sqrt{Q_{ii}}$ , and  $\Phi$  is the univariate standard normal CDF. This is a generalization of the multivariate normal family (which can be recovered by setting  $w = 0$ ) which retains some of its nice properties. Intuitively, it is a multivariate normal which has been skewed, according to the skew parameter  $w \in \mathbb{R}^d$ , in the direction of  $w/|w|$  by a magnitude of  $|w|$ .

Figure 5.21 shows a contour plot and a scatterplot of  $n = 2000$  points from  $\mathcal{SN}(\xi, Q, w)$  with  $\xi = (-1.5, -1)^T$ ,  $Q = \begin{pmatrix} 6.25 & 1.5 \\ 1.5 & 1 \end{pmatrix}$ , and  $w = (1, 3)^T$ . A convenient parametrization of  $Q$  in the bivariate case ( $d = 2$ ), is as  $Q = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$ , where  $\sigma_1, \sigma_2 > 0$  and  $\rho \in (-1, 1)$  is the correlation coefficient; for instance, in the preceding example,  $\sigma_1 = 2.5$ ,  $\sigma_2 = 1$ , and  $\rho = 0.6$ .

Due to its increased flexibility over the multivariate normal, the skew-normal may be useful for applications involving mixture models, particularly clustering.

### 5.2.1 Data

We consider data simulated from a three-component mixture of skew-normals  $\sum_{i=1}^3 \pi_i \mathcal{SN}(x \mid \xi_i, Q_i, w_i)$ , with parameters as shown in Table 5.3. Figure 5.22 displays contour plots and samples from each of the three mixture components.

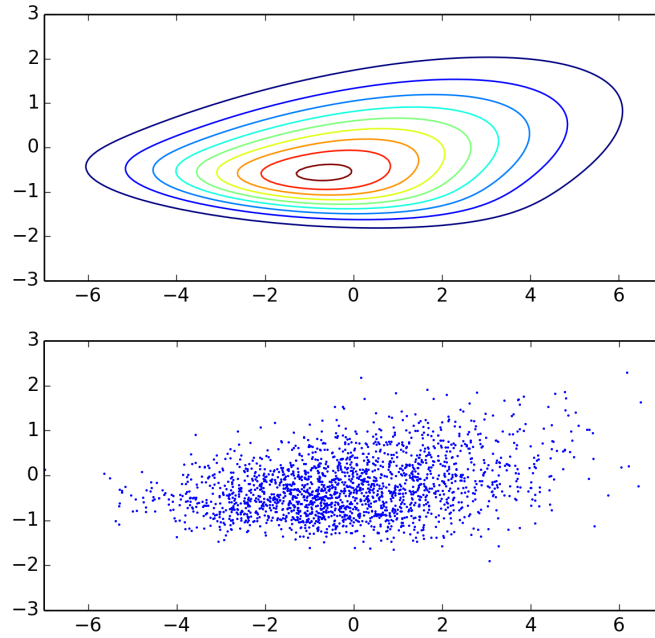


Figure 5.21: Contour plot and scatterplot of a bivariate skew-normal distribution.

Table 5.3: Parameters of the skew-normal mixture used for data simulation

| $i$ | $\pi_i$ | $\xi_{i1}$ | $\xi_{i2}$ | $\sigma_{i1}$ | $\sigma_{i2}$ | $\rho_i$ | $w_{i1}$ | $w_{i2}$ |
|-----|---------|------------|------------|---------------|---------------|----------|----------|----------|
| 1   | 0.45    | 2          | -2         | 2.5           | 1             | -0.6     | -3       | 6        |
| 2   | 0.3     | 0          | 0          | 2             | 1             | 0.95     | 6        | 0        |
| 3   | 0.25    | -1         | 2          | 1             | 1             | 0        | 0        | -2       |

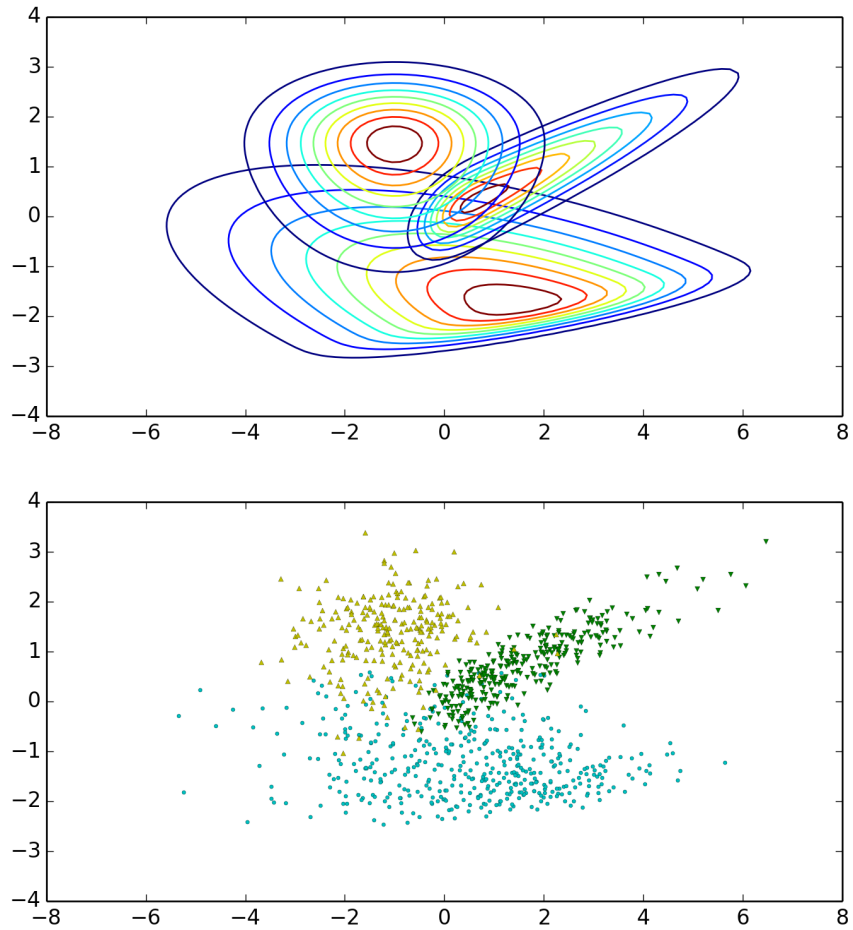


Figure 5.22: Mixture components used for simulations (top: contour plots, bottom: scatterplots). The scatterplots shows 450, 300, and 250 samples from the three components, respectively.

### 5.2.2 Model description

We take the family of component densities to be bivariate skew-normal,

$$f_{\theta}(x) = \mathcal{SN}(x \mid \xi, Q, w),$$

with the parametrization  $\theta = (\xi_1, \xi_2, \log \sigma_1, \log \sigma_2, \text{logit } \frac{p+1}{2}, w_1, w_2)$  (where  $\text{logit } p = \log \frac{p}{1-p}$ ). For the base measure  $H$ , for convenience we simply choose  $\theta \sim \mathcal{N}(0, C)$  where  $C^{1/2} = \text{diag}(5, 5, 2, 2, 2, 4, 4)$ .

Note that this parametrization covers all of  $\mathbb{R}^7$ . This parametrization and choice of  $H$  was chosen rather arbitrarily, without any particular regard to the structure of the skew-normal distribution; a more principled choice of  $H$  is almost certainly possible. The location and scale of  $H$  was chosen to be roughly appropriate for the data at hand.

For the MFM parameters, we use  $\gamma = 1$  and

$$p(k) = \begin{cases} c & \text{if } k \in \{1, \dots, 30\} \\ c/(k-30)^2 & \text{if } k > 30 \end{cases}$$

where  $c = 1/(30 + \pi^2/6)$ . For the DPM, we use only a fixed concentration parameter,  $\alpha = 1$ .

### 5.2.3 Approximate inference

Since  $H$  is not a conjugate prior for  $\{f_{\theta}\}$ , we again use the non-conjugate sampler (Section 4.4.2). For the auxiliary variable distribution, we again use  $H$ . For

the  $\phi$  move (that is, sampling the component parameters), we use the Metropolis algorithm with a normal proposal distribution. Specifically, for  $c \in \mathcal{C}$ , we propose  $\phi'_c = \phi_c + 0.1 \varepsilon$  where  $\varepsilon_1, \dots, \varepsilon_7 \stackrel{\text{iid}}{\sim} \mathcal{N}(0, 1)$  (and accept  $\phi'_c$  with probability  $\min\{1, \pi(\phi'_c|x_c)/\pi(\phi_c|x_c)\}$ ); this is repeated 20 times between each pass through the reassignments of  $j = 1, \dots, n$ . Note that the acceptance probability is

$$\min \left\{ 1, \frac{h(\phi'_c) \prod_{j \in c} f_{\phi'_c}(x_j)}{h(\phi_c) \prod_{j \in c} f_{\phi_c}(x_j)} \right\},$$

which does not actually require computing the posterior density. Although a small amount of experimentation was done with these parameters, this move could almost certainly be improved upon (e.g., perhaps by scaling the proposal distribution separately for each dimension).

For each  $n \in \{50, 200, 500, 2000\}$ , five independent sets of data from the true distribution were considered. As before, for each such set, the sampler was run for 100,000 burn-in sweeps and 200,000 samples sweeps (for a total of 300,000); this appeared to be sufficient for mixing.

For a dataset of size  $n$ , the sampler used for these experiments took approximately  $1.5 \times 10^{-5} n$  seconds per sweep, using a 2.80 GHz processor with 6 GB of RAM.

### 5.2.4 Density estimation

As before, Equation 4.2.15 was employed to estimate the density based on samples from the posterior of  $\mathcal{C}, \phi \mid x_{1:n}$ . The 20,000 recorded samples (out of the 200,000 sample sweeps) were used. Figure 5.23 shows representative examples of the estimated densities for increasing amounts of data ( $n \in \{50, 200, 500\}$ ). Although slight



differences between the MFM and DPM can be detected, overall, the densities appear to be quite similar. The wiggleness of the contour lines for the DPM is probably due to the presence of tiny clusters in the posterior samples; see below.

Since the true density is known, we can assess the performance of the density estimates using Hellinger distance (see Section 5.1.4). Figure 5.24 shows the estimated Hellinger distance between the true density and the estimated densities, for increasing amounts of data. Each distance was estimated with  $N = 10^3$  independent samples from the true density; the same samples were used to evaluate both MFM and DPM. As usual, the MFM and DPM results are very similar.

### 5.2.5 Clustering

We display some representative clusterings sampled from the posterior. Figure 5.25 shows (Top) three samples from the MFM and the DPM, and (Bottom) the true component assignments of the points, for  $n = 500$  data points. The markers for points in small clusters have been enlarged for visualization purposes. All of the data points are visible within the window shown.

Similarly to before, we observe that both models consistently have three dominant clusters corresponding to the three true components, and that the DPM tends to have small transient extra clusters, while the MFM does not.

The behavior of the means of the sorted cluster sizes is similar to the univariate normal case. Also, the pairwise probability matrices for the MFM and DPM are very similar to one another.

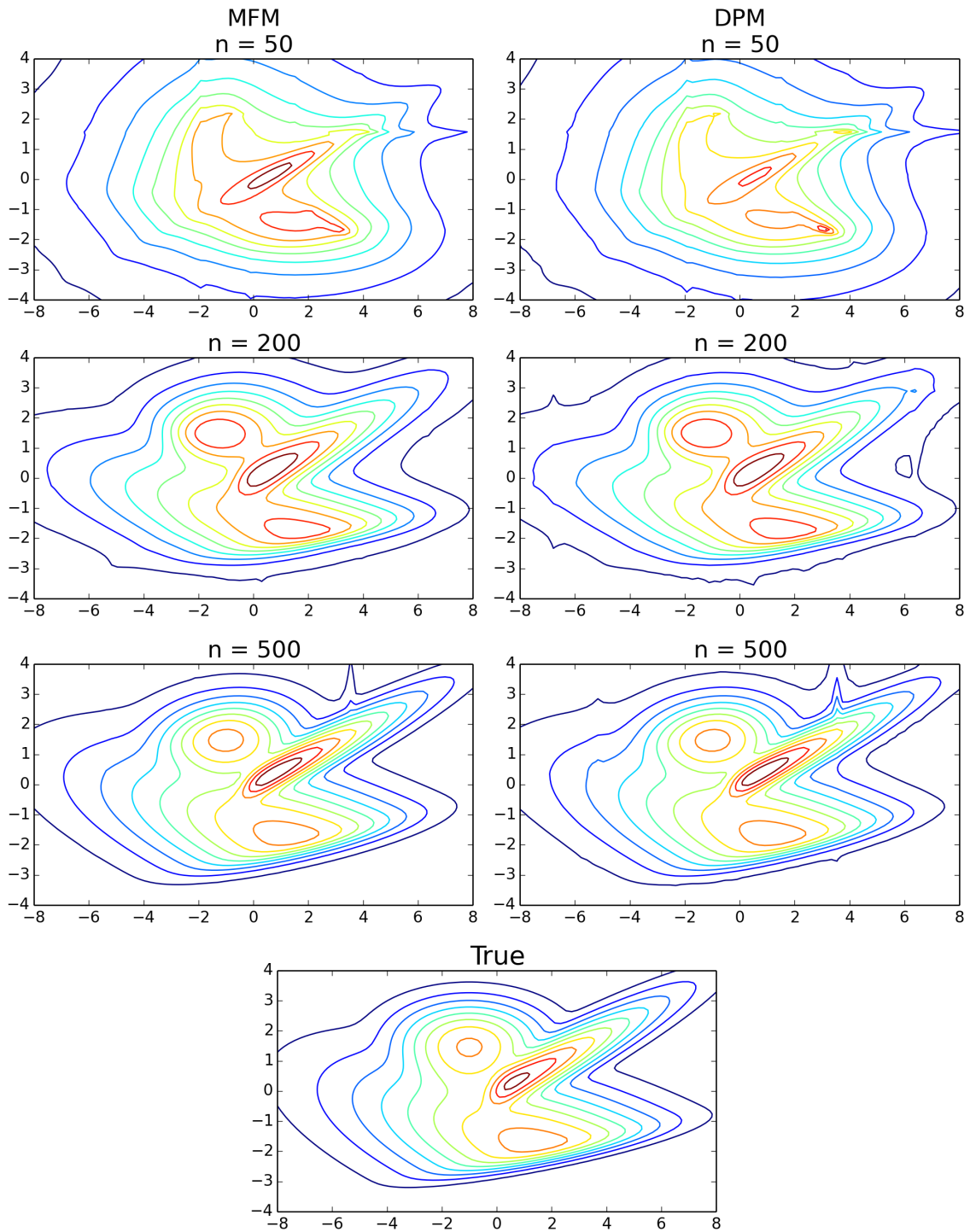


Figure 5.23: Contour plots of typical density estimates for the MFM (left) and DPM (right); compare with the true density (bottom).

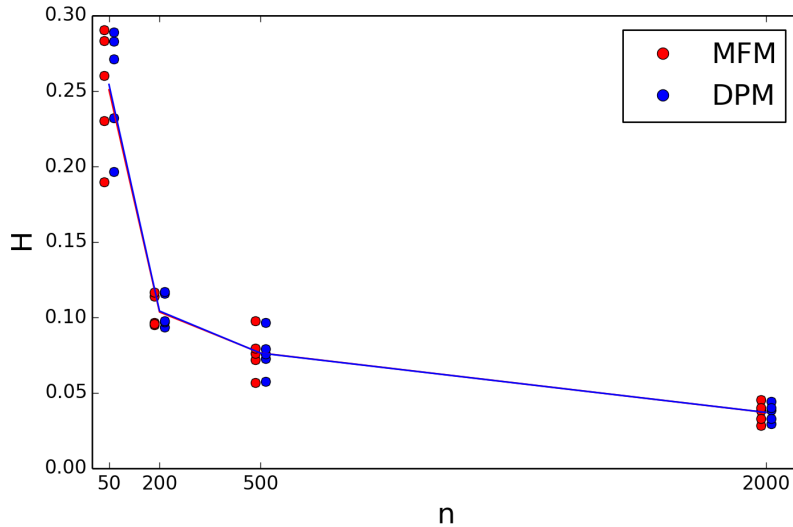


Figure 5.24: Estimated Hellinger distances for MFM (red, left) and DPM (blue, right) density estimates. For each  $n \in \{50, 200, 500, 2000\}$ , five independent datasets of size  $n$  were used, and the lines connect the averages of the estimated distances.

## 5.2.6 Number of components and clusters

We estimate the posteriors on the number of components and clusters, as described in Section 5.1.6. As explained there, we do not consider these for the purpose of measuring heterogeneity, but simply to demonstrate posterior consistency properties and differences between the MFM and DPM.

Figure 5.26 shows estimates of  $p(t|x_{1:n})$  and  $p(k|x_{1:n})$  for the MFM,  $p(t|x_{1:n})$  for the DPM (DPM-fixed), and  $p^*(t|x_{1:n})$  for the DPM (DPM-flat). For each  $n \in \{50, 200, 500, 2000\}$ , we show the average over 5 datasets. As before, the MFM appears to be concentrating at the true number of components, the DPM does not appear to be, and for DPM-flat it is not clear (although it does seem to be “stalled”, suggesting that it might not concentrate).

This empirical observation for the MFM is actually rather interesting, since skew-normal mixtures are not mixture identifiable in general, and therefore we have no

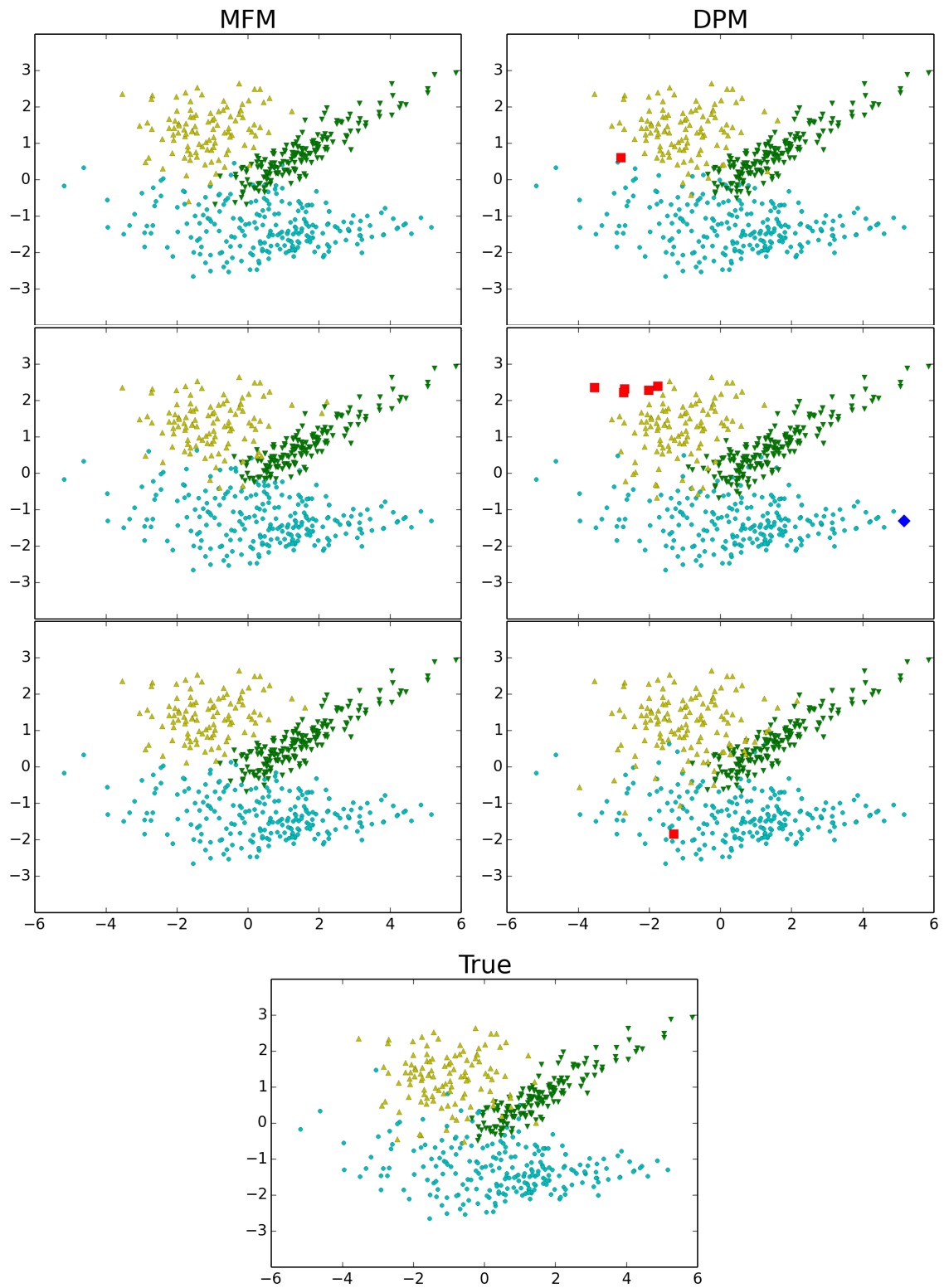


Figure 5.25: Typical sample clusterings from the posterior; see the text.

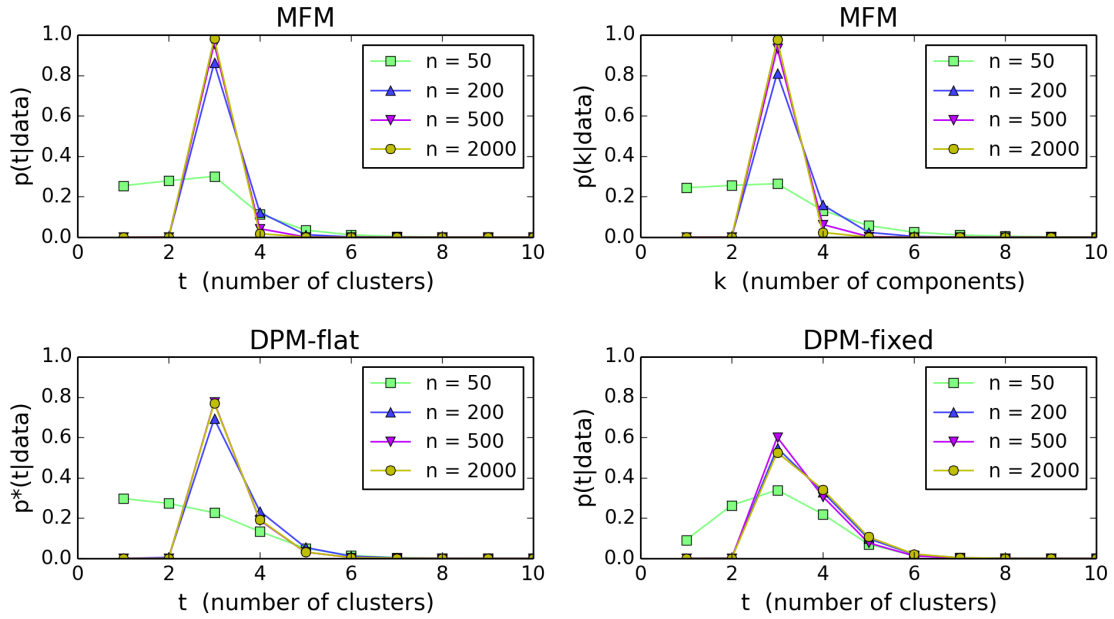


Figure 5.26: Posterior distributions of  $k$  and  $t$ .

guarantee that the posterior will concentrate at the true number of components. This lack of identifiability can be easily exhibited (as pointed out to me by Dan Klein) by observing that

$$\frac{1}{2}\mathcal{N}(x \mid \xi, Q, w) + \frac{1}{2}\mathcal{N}(x \mid \xi, Q, -w) = \mathcal{N}(x \mid \xi, Q, 0)$$

for any  $x, \xi, Q, w$ , since  $\Phi(z) + \Phi(-z) = 2\Phi(0)$  for any  $z \in \mathbb{R}$ .

# CHAPTER SIX

---

## Combinatorial stochastic processes for other variable-dimension models

## 6.1 Introduction

Like the Dirichlet process mixture (DPM), a number of other nonparametric models are derived as a certain infinite-dimensional limit of a family of finite-dimensional models. For each such model, we could alternatively construct a variable-dimension model by putting a prior on the dimension, in the same way that we constructed the MFM.

As we have seen with the MFM, a variable-dimension model can have many of the same characteristics as an infinite-dimensional model, as well as having certain advantages. Two key observations underlying this development were that the MFM has a nice partition distribution, and that this distribution can be generated by a simple combinatorial stochastic process (a restaurant process, in this case).

In this chapter, we show that for two other popular nonparametric models — the hierarchical Dirichlet process (HDP) and the Indian buffet process (IBP) — the natural variable-dimensional counterparts also have many of the same attractive properties as the nonparametric, infinite-dimensional models. In particular, they give rise to nice distributions on discrete structures that can be generated by simple combinatorial stochastic processes (a “franchise process” and a “buffet process”, respectively).

## 6.2 Hierarchical mixture of finite mixtures

In many applications, the data comes in groups that may be related in some ways but may also have significant differences. For instance, in modeling a set of documents,

the words in each document may be considered as a group of data points, and documents may be related by sharing various topics. One way to model such data would be to combine the data from all groups and use a single mixture model, however, this does not allow for possible differences between the groups. At the other extreme, one could use a separate mixture model for each group, but this does not take advantage of possible relationships between the groups. A third way — the one we pursue here — is to use a mixture model for each group and allow mixture components to be shared among groups.

The hierarchical Dirichlet process (HDP) ([Teh et al., 2004, 2006](#)) is a model of this type, in which each group is modeled using a Dirichlet process mixture for which the base measure is itself a draw from a Dirichlet process, and is shared among all groups. Due to the discreteness of random measures drawn from a DP, this allows subsets of the component parameters of these different mixtures to be shared. The HDP is a special case of the Analysis of Densities model of [Tomlinson and Escobar \(1999\)](#), and can also be viewed as an example of a dependent Dirichlet process ([MacEachern et al., 1999, MacEachern, 2000, MacEachern et al., 2001](#)). The HDP has seen a wide range of applications, including document modeling ([Teh et al., 2004](#)), natural language modeling ([Liang et al., 2007](#)), object tracking ([Fox et al., 2007](#)), haplotype inference ([Xing et al., 2006](#)), natural image processing ([Kivinen et al., 2007](#)), cognitive science ([Griffiths et al., 2007](#)), and measuring similarity between musical pieces ([Hoffman et al., 2008](#)), and a dynamic hierarchical Dirichlet process (dHDP) has been proposed by [Ren et al. \(2008\)](#). (For other nonparametric approaches to modeling data in multiple related groups, see [Müller et al. \(2004\)](#), [Dunson \(2006\)](#), and [Rodriguez et al. \(2008\)](#).)

In the same way that the HDP is constructed from the Dirichlet process, one can construct a hierarchical mixture model using the corresponding distribution on



discrete measures for the MFM. We show that the resulting model, which we refer to as a hierarchical mixture of finite mixtures (HMFM), has some of the same attractive properties as the HDP mixture model. In particular, it has the same exchangeability properties, it can be represented in terms of a “franchise process” that closely parallels the Chinese restaurant franchise process of the HDP, and MCMC inference algorithms for the HDP that are based on the franchise process representation can be adapted to the HMFM.

### 6.2.1 A hierarchical variable-dimension mixture model

First, we give a succinct description of the model, employing the distribution  $\mathcal{M}(p_K, \gamma, H)$  on discrete measures  $G$  associated with the MFM; see Section 4.2.7. As before, we use  $f_G$  to denote the mixture obtained from the discrete measure  $G$ , that is,  $f_G(x) = \int f_\theta(x)G(d\theta)$ . Assume we have a family  $\{f_\theta : \theta \in \Theta\}$  and base measure  $H$  on  $\Theta$  just as before.

Suppose we have data in  $m$  groups, with  $n_r$  data points in the  $r$ th group. For instance, in the example of documents mentioned above, we have  $m$  documents with  $n_r$  words in document  $r$ . The *hierarchical mixture of finite mixtures* (HMFM) model is as follows:

$$G_0 \sim \mathcal{M}(q_0, \gamma_0, H)$$

For  $r \in \{1, \dots, m\}$  independently, given  $G_0$ :

$$G_r \sim \mathcal{M}(q_r, \gamma_r, G_0) \tag{6.2.1}$$

$$X_{rj} \sim f_{G_r}(x) \text{ independently for } j \in \{1, \dots, n_r\},$$

where  $q_0, q_1, \dots, q_m$  are p.m.f.’s on  $\{1, 2, \dots\}$  and  $\gamma_0, \gamma_1, \dots, \gamma_m > 0$ . See Figure

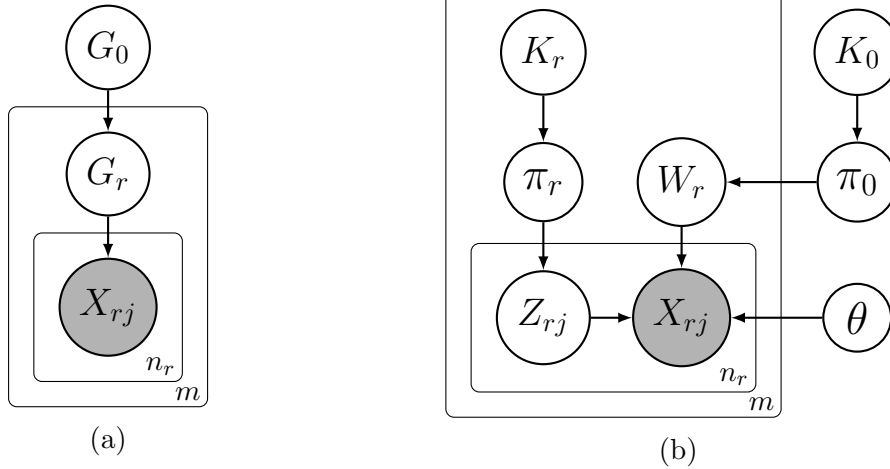


Figure 6.1: Graphical models for (a) the random discrete measure formulation, and (b) the equivalent representation including additional latent variables.

6.1(a) for a graphical model representation. The only difference between this and an HDP mixture model is that in the HDP, the  $\mathcal{M}$  distributions are replaced by Dirichlet processes: specifically,  $\mathcal{M}(q_0, \gamma_0, H)$  is replaced by  $\text{DP}(\alpha_0, H)$ , and  $\mathcal{M}(q_r, \gamma_r, G_0)$  is replaced by  $\text{DP}(\alpha_r, G_0)$  for  $r = 1, \dots, m$ .

This model has the same exchangeability properties as the HDP mixture model: the datapoints within each group are exchangeable (being conditionally i.i.d.), and if  $q_1 = \dots = q_m$  and  $\gamma_1 = \dots = \gamma_m$ , then the order of the groups does not matter, as long as the group sizes  $n_1, \dots, n_m$  are handled appropriately; for instance, the distribution is the same if the sizes are permuted, the data is sampled, and then the groups are permuted back to the original order of the sizes (or alternatively, if the sizes are taken to be infinite, then the groups are exchangeable). This model can also be extended to multi-level hierarchies in the same way as the HDP.

It is also useful to represent the model in an equivalent way, using a number of additional latent variables, as follows. Despite the flurry of symbols, this is a very

natural construction. See Figure 6.1(b) for a graphical model representation.

$$K_0 \sim q_0$$

$$\pi_0 \sim \text{Dirichlet}_{k_0}(\gamma_0, \dots, \gamma_0), \text{ given } K_0 = k_0$$

$$\theta_1, \theta_2, \dots \stackrel{\text{iid}}{\sim} H$$

For  $r = 1, \dots, m$ : (6.2.2)

$$K_r \sim q_r$$

$$\pi_r \sim \text{Dirichlet}_{k_r}(\gamma_r, \dots, \gamma_r), \text{ given } K_r = k_r$$

$$W_{r1}, W_{r2}, \dots \stackrel{\text{iid}}{\sim} \pi_0, \text{ given } \pi_0$$

$$Z_{r1}, \dots, Z_{rn_r} \stackrel{\text{iid}}{\sim} \pi_r, \text{ given } \pi_r$$

$$X_{rj} \sim f_{\theta_{Y_{rj}}}, \text{ where } Y_{rj} = W_r Z_{rj}, \text{ indep. for } j \in \{1, \dots, n_r\}, \text{ given } (\theta_i), W_r, Z_r.$$

The representation in terms of discrete measures can be recovered by setting

$$G_0 = \sum_{i=1}^{k_0} \pi_{0i} \delta_{\theta_i} \quad \text{and} \quad G_r = \sum_{i=1}^{k_r} \pi_{ri} \delta_{\theta_{W_{ri}}}$$

for  $r = 1, \dots, m$ , and the distribution on the data  $(X_{rj})$  is the same in both of these representations.

### 6.2.2 Franchise process

In the same way as an HDP mixture, this model can also be described in terms of a simple “franchise process” and an associated distribution on combinatorial structures. We use the same restaurant/franchise analogy as the HDP. We envision  $m$  different restaurants all belonging to the same franchise, and imagine the  $m$  groups of datapoints as groups of customers visiting these restaurants, respectively. Label the

$j$ th customer at restaurant  $r$  by  $(r, j)$ . For each restaurant  $r = 1, \dots, m$ , we partition the customers according to which table they are seated at, resulting in a partition  $\mathcal{C}_r$  of  $\{(r, 1), (r, 2), \dots, (r, n_r)\}$ . There are  $t_r := |\mathcal{C}_r|$  occupied tables at restaurant  $r$ , and  $L := \sum_{r=1}^m t_r$  occupied tables altogether. Now, we imagine that each occupied table is served a single dish, chosen from a franchise-wide menu. We partition the  $L$  occupied tables according to which dish they are served, and represent this as a partition  $\mathcal{C}_0$  of  $\bigcup_{r=1}^m \mathcal{C}_r$ . (In this notation,  $\mathcal{C}_0$  is a set of sets of sets of customers.)

Define

$$V_n^r(t) = \sum_{k=1}^{\infty} \frac{k^{(t)}}{(\gamma_r k)^{(n)}} q_r(k)$$

for  $r = 0, 1, \dots, m$  and  $0 \leq t \leq n$ . Consider the following process for generating  $\mathcal{C}_1, \dots, \mathcal{C}_r$  and  $\mathcal{C}_0$ :

- Initialize  $\mathcal{C}_0 \leftarrow \{\}$  and  $L \leftarrow 0$ .
- For each restaurant  $r = 1, \dots, m$ :
  - Initialize  $\mathcal{C}_r \leftarrow \{\}$ .
  - For  $j = 1, \dots, n_r$ :

(1) The  $j$ th customer at restaurant  $r$  sits ...

- \* at an existing table  $c \in \mathcal{C}_r$  with probability  $\propto |c| + \gamma_r$
- \* at a new table with probability  $\propto \frac{V_j^r(t_r + 1)}{V_j^r(t_r)} \gamma_r$

where  $t_r = |\mathcal{C}_r|$  is the number of tables occupied so far at restaurant  $r$ .

(2) If a new table is chosen in step (1), set  $L \leftarrow L + 1$  and select a dish for it from the franchise-wide menu, choosing ...

- \* an existing “dish”  $d \in \mathcal{C}_0$  with probability  $\propto |d| + \gamma_0$  (note that  $|d|$  is the total number of tables having this dish so far, in all restaurants)

\* a new dish with probability  $\propto \frac{V_L^0(t_0 + 1)}{V_L^0(t_0)} \gamma_0$

where  $t_0 = |\mathcal{C}_0|$  is the number of dishes tried so far in all restaurants.

(3) Update  $\mathcal{C}_r$  and  $\mathcal{C}_0$  to reflect these choices.

Note that the first customer at each restaurant always sits at a new table since there are no existing tables, and similarly, the first table in the whole process always gets a new dish.

The only difference between this and the Chinese restaurant franchise process of the HDP is that for the HDP, in step (1) table  $c$  is chosen with probability  $\propto |c|$  or a new table is chosen with probability  $\propto \alpha_r$ , and in step (2) dish  $d$  is chosen with probability  $\propto |d|$  or a new dish is chosen with probability  $\propto \alpha_0$ .

For each  $r \in \{1, \dots, m\}$ , the seating of customers in restaurant  $r$  simply follows the MFM restaurant process (Section 4.2.6), independently of the other restaurants. And given the seating arrangements  $\mathcal{C}_1, \dots, \mathcal{C}_m$ , the serving of dishes also follows the MFM restaurant process, this time with occupied tables playing the role of customers. Note that it is not necessary to immediately choose a dish for each new table; in particular, it would be equivalent if we waited until all the customers at all the restaurants had been seated before choosing dishes. Therefore, the probability of obtaining  $\mathcal{C}_{0:m} := (\mathcal{C}_0, \mathcal{C}_1, \dots, \mathcal{C}_m)$  is

$$p(\mathcal{C}_{0:m}) = P_L^0(\mathcal{C}_0) \prod_{r=1}^m P_{n_r}^r(\mathcal{C}_r) \quad (6.2.3)$$

where

$$P_n^r(\mathcal{C}) = V_n^r(|\mathcal{C}|) \prod_{c \in \mathcal{C}} \gamma_r^{(|c|)} \quad (6.2.4)$$

is the MFM distribution on partitions (Equation 4.2.1), and  $L = \sum_{r=1}^m |\mathcal{C}_r|$ . Note that  $\mathcal{C}_0$  depends on  $\mathcal{C}_r$  through  $L$ . (Of course,  $p(\mathcal{C}_{0:m})$  is as above only when  $\mathcal{C}_{0:m}$  is valid, and  $p(\mathcal{C}_{0:m}) = 0$  otherwise;  $\mathcal{C}_{0:m}$  is valid when  $\mathcal{C}_r$  is a partition of  $\{(r, 1), (r, 2), \dots, (r, n_r)\}$  for  $r = 1, \dots, m$  and  $\mathcal{C}_0$  is a partition of  $\bigcup_{r=1}^m \mathcal{C}_r$ .)

For comparison, the formula for the probability of  $\mathcal{C}_{0:m}$  under the Chinese restaurant franchise process is the same, except that  $P_n^r(\mathcal{C})$  is replaced by the DP distribution on partitions,  $(\alpha_r^{|\mathcal{C}|} / \alpha_r^{(n)}) \prod_{c \in \mathcal{C}} (|c| - 1)!$ .

By exchangeability of the restaurant process, the distribution of  $\mathcal{C}_{0:m}$  is invariant to the order in which the customers enter each restaurant, and further, it is not necessary for all the customers at restaurant  $r$  to be seated and served before customers at  $r' > r$  start sitting and being served — the same distribution of  $\mathcal{C}_{0:m}$  will result if steps (1), (2), and (3) of the process are followed for any temporal ordering of the customers.

We can use this distribution on combinatorial structures  $\mathcal{C}_{0:m}$  to construct a hierarchical mixture model, as follows:

$$\begin{aligned}
 \mathcal{C}_r &\sim P_{n_r}^r \text{ independently for } r = 1, \dots, m \\
 \mathcal{C}_0 &\sim P_L^0 \text{ given } \mathcal{C}_{1:m}, \text{ where } L = \sum_{r=1}^m |\mathcal{C}_r| \\
 \phi_d &\stackrel{\text{iid}}{\sim} H \text{ for } d \in \mathcal{C}_0, \text{ given } \mathcal{C}_0 \\
 X_{rj} &\sim f_{\phi_d} \text{ independently for } (r, j) \in c, c \in d, d \in \mathcal{C}_0, \text{ given } \mathcal{C}_0.
 \end{aligned} \tag{6.2.5}$$

See Figure 6.2 for a graphical model representation. We claim that this is equivalent to the models described above; see the next section for this argument.

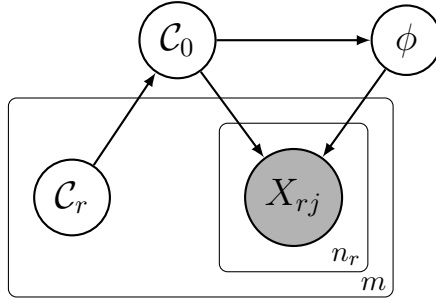


Figure 6.2: Graphical model for combinatorial representation of the model.

### 6.2.3 Equivalence of the models

Here, we show that the combinatorial model in Equation 6.2.5 is equivalent to the model in Equation 6.2.2, in the sense that it simply represents different variables in the latent structure, and gives rise to the same distribution for the data  $(X_{rj})$ .

To show this, we will start with the model as described in Equation 6.2.2, then define  $\mathcal{C}_0, \mathcal{C}_1, \dots, \mathcal{C}_m$  and  $(\phi_d : d \in \mathcal{C}_0)$  in this context, and finally, observe that we obtain the same distributions as in Equation 6.2.5. For  $r = 1, \dots, m$ , let  $\mathcal{C}_r$  be the partition of  $\{(r, 1), \dots, (r, n_r)\}$  induced by  $Z_{r1}, \dots, Z_{rn_r}$ . Note that  $\mathcal{C}_1, \dots, \mathcal{C}_m$  are independent, and by construction,  $\mathcal{C}_r$  simply has the MFM distribution on partitions,  $P_{n_r}^r$ , in the notation of Equation 6.2.4 above.

Now, let  $\mathcal{T} = \bigcup_{r=1}^m \mathcal{C}_r$ . Recalling that  $Y_{rj} = W_{rZ_{rj}}$ , note that for any  $c \in \mathcal{T}$  all of the elements  $(r, j) \in c$  have the same value of  $Y_{rj}$ ; denote this value by  $Y_c$ . Let  $\mathcal{C}_0$  be the partition of  $\mathcal{T}$  induced by  $(Y_c : c \in \mathcal{T})$ . Given  $\mathcal{C}_{1:m}$  and  $\pi_0$ , observe that  $(Y_c : c \in \mathcal{T})$  are i.i.d. from  $\pi_0$  (by applying the argument of Section 4.2.2 to each  $r$ ). Therefore, given  $\mathcal{C}_{1:m}$  (marginalizing out  $\pi_0$ ), the partition induced by  $(Y_c : c \in \mathcal{T})$  — namely  $\mathcal{C}_0$  — simply has an MFM partition distribution; specifically,  $\mathcal{C}_0 \mid \mathcal{C}_{1:m}$  has p.m.f.  $P_L^0$  in the notation of Equation 6.2.4, where  $L = |\mathcal{T}| = \sum_{r=1}^m |\mathcal{C}_r|$ .

For  $d \in \mathcal{C}_0$ , let  $\phi_d$  equal the  $\theta_i$  such that  $Y_c = i$  for all  $c \in d$ . Again, by the same argument as in Section 4.2.2,  $(\phi_d : d \in \mathcal{C}_0)$  are i.i.d. from  $H$ , given  $\mathcal{C}_0$ . Finally, it is easily verified that the distribution of  $(X_{rj}) \mid \mathcal{C}_0, (\phi_d)$  is  $X_{rj} \sim f_{\phi_d}$  independently for  $(r, j) \in c, c \in d, d \in \mathcal{C}_0$ .

### 6.2.4 Inference

Since the HMFm model has a franchise process which closely parallels that of the HDP mixture model, the Gibbs sampling algorithm for the HDP mixture (Teh et al., 2004) can be easily adapted to the HMFm. It might be interesting to see if the other inference algorithms proposed for the HDP mixture, such as the augmented sampler of Teh et al. (2006), can also be adapted to the HMFm.

## 6.3 Mixture of finite feature models

In a sense, a mixture model can be viewed as having a single discrete latent feature, controlling which component each data point comes from. Often, it is more realistic to allow for multiple latent features, enabling the model to parsimoniously represent more complex relationships. The latent features are sometimes chosen to be binary-valued (zero or one), and in this case, the feature values can be represented in a binary matrix  $Z$  such that  $Z_{ij} = 1$  when item  $i$  has feature  $j$ . Given such a matrix  $Z$ , the observed data  $X$  may be modeled in a variety of ways. A simple example is to take

$$X_i = \sum_j Z_{ij} \mu_j + \varepsilon_i$$



where the  $\mu_j$ 's and  $\varepsilon_i$ 's are multivariate normal; this can be viewed as a binary factor analysis model, where  $M := [\mu_1 \cdots \mu_k]$  is the “factor loading matrix” and  $Z_{i1}, \dots, Z_{ik}$  are the “factors” for item  $i$ .

The Indian buffet process (IBP) of [Griffiths and Ghahramani \(2005\)](#) is a non-parametric model for  $Z$  with infinitely many features, finitely many of which are possessed by any given item. The IBP has seen a number of applications, including models for protein complexes ([Krause et al., 2006](#)), gene expression data ([Knowles and Ghahramani, 2007](#)), causal graph structure ([Wood et al., 2006](#)), dyadic data ([Meeds et al., 2007](#)), similarity judgments ([Navarro and Griffiths, 2007](#)), network data ([Miller et al., 2009](#)), and multiple time-series ([Fox et al., 2009](#)). The IBP has some interesting theoretical properties ([Thibaux and Jordan, 2007](#), [Teh et al., 2007](#)), and has inspired the development of a number of other models.

The IBP can be derived as a certain infinite-dimensional limit of a family of finite feature models. Here, we show that by instead placing a prior on the number of features — that is, using a variable-dimension feature model — one obtains a distribution with many of the same attractive properties as the IBP: an exchangeable distribution on equivalence classes of binary matrices, representation via a simple buffet process, and approximate inference via the same Gibbs sampling algorithms. In fact, the IBP (as well as the two-parameter generalization of [Ghahramani et al. \(2007\)](#)) can be viewed as a limiting case, corresponding to a certain corner of the parameter space. For further background on the IBP, including motivation for the use of such priors in latent feature models, and several applications, we refer to [Ghahramani et al. \(2007\)](#) and references therein (especially [Roweis and Ghahramani \(1999\)](#)).

Variable-dimension feature models have previously been developed by [Lopes and](#)

West (2004), Dunson (2006), and Ghosh and Dunson (2009), in the context of factor analysis, however, in such models the factors are continuous rather than binary-valued. For inference, Lopes and West (2004) used reversible jump Markov chain Monte Carlo, while Dunson (2006) and Ghosh and Dunson (2009) used a parameter-expanded sampler. Sparse variants of these models, in which many zeros appear in the entries of the factor loading matrix (rather than in the factors themselves), have been applied to gene expression profiling (West, 2003, Carvalho et al., 2008).

### 6.3.1 A distribution on binary matrices

Let  $p(k)$  be a p.m.f. on  $\{0, 1, 2, \dots\}$ , and let  $a, b > 0$ . Let

$$\begin{aligned} K &\sim p(k) \\ \pi_1, \dots, \pi_k &\stackrel{\text{iid}}{\sim} \text{Beta}(a, b), \text{ given } K = k \\ Z_{ij} &\sim \text{Bernoulli}(\pi_j) \text{ independently for } i \in \{1, \dots, n\}, j \in \{1, \dots, k\}, \\ &\text{given } K = k \text{ and } \pi_{1:k}. \end{aligned} \tag{6.3.1}$$

We refer to this as a *mixture of finite feature models* (MFFM). When  $K$  is fixed, say  $K = k$ , this is precisely the finite feature model described by Griffiths and Ghahramani (2005), forming the basis for the IBP, which is then obtained by setting  $b = 1$ ,  $a = \alpha/k$ , and taking  $k \rightarrow \infty$ .

### 6.3.2 A simple urn process

In studying the properties of this model, it is useful to first consider a single column, leading to a simple urn process. Let  $\pi \sim \text{Beta}(a, b)$ , and  $U_1, \dots, U_n \stackrel{\text{iid}}{\sim} \text{Bernoulli}(\pi)$

given  $\pi$ . Then

$$p(u_{1:n}) = \int p(u_{1:n}|\pi)p(\pi)d\pi = \frac{B(a+s, b+n-s)}{B(a, b)} = \frac{a^{(s)}b^{(n-s)}}{(a+b)^{(n)}} \quad (6.3.2)$$

where  $B(a, b) = \Gamma(a)\Gamma(b)/\Gamma(a+b)$  is the beta function,  $s = \sum_i u_i$ , and  $x^{(m)} = x(x+1)\cdots(x+m-1)$  with the convention that  $x^{(0)} = 1$ . Since this quantity appears so frequently, it is useful to introduce the notation

$$w_n(s) = \frac{a^{(s)}b^{(n-s)}}{(a+b)^{(n)}}. \quad (6.3.3)$$

From Equation 6.3.2, it is easy to see that  $U_1, \dots, U_n$  can also be generated by taking the first  $n$  draws in the following urn process:

$$U_1 \sim \text{Bernoulli}\left(\frac{a}{a+b}\right)$$

$$U_m | U_{m-1}, \dots, U_1 \sim \text{Bernoulli}\left(\frac{a + \sum_{i=1}^{m-1} U_i}{a+b+m-1}\right) \text{ for } m = 2, 3, \dots$$

From this perspective,  $a$  and  $b$  can be interpreted as “pseudo-counts”. Note that if  $M = \min \{i \in \{1, 2, \dots\} : U_i = 1\}$ , then for any  $m \in \{1, 2, \dots\}$ ,

$$\mathbb{P}(M = m) = \frac{ab^{(m-1)}}{(a+b)^{(m)}} = w_m(1), \text{ and} \quad (6.3.4)$$

$$\mathbb{P}(M > m) = \frac{b^{(m)}}{(a+b)^{(m)}} = w_m(0), \quad (6.3.5)$$

by Equation 6.3.2.

### 6.3.3 Equivalence classes of matrices

#### 6.3.3.1 Removing columns of all zeros

For any binary matrix  $z \in \{0, 1\}^{n \times k}$ , define  $z^*$  be the matrix obtained from  $z$  by removing any columns consisting of all zeros (and preserving the order of the remaining columns). Define  $s_j(z) = \sum_{i=1}^n z_{ij}$ . Let  $Z$  be as in Equation 6.3.1. Then, as we will see, for  $z^* \in \{0, 1\}^{n \times t}$ ,  $t \in \{0, 1, 2, \dots\}$ , such that  $s_1(z^*), \dots, s_t(z^*) > 0$ ,

$$\mathbb{P}(Z^* = z^*) = v_n(t) \prod_{j=1}^t w_n(s_j(z^*)) \quad (6.3.6)$$

where  $w_n(s)$  is as in Equation 6.3.3 above, and

$$v_n(t) = \sum_{k=0}^{\infty} \binom{k}{t} w_n(0)^{k-t} p(k). \quad (6.3.7)$$

By convention, an empty product equals 1 and  $\binom{k}{t} = 0$  when  $k < t$ . For comparison, under the IBP the corresponding distribution is

$$\mathbb{P}_{\text{IBP}}(Z^* = z^*) = \exp(-\alpha H_n) \frac{\alpha^t}{t!} \prod_{j=1}^t \frac{(s_j(z^*) - 1)! (n - s_j(z^*))!}{n!} \quad (6.3.8)$$

where  $H_n = \sum_{k=1}^n 1/k$  is the  $n$ th harmonic number.

To derive Equation 6.3.6, first note that by Equation 6.3.2, for  $z \in \{0, 1\}^{n \times k}$ ,

$$\mathbb{P}(Z = z \mid k) = \prod_{j=1}^k w_n(s_j(z)) = w_n(0)^{k-t} \prod_{j=1}^t w_n(s_j(z^*)), \quad (6.3.9)$$

where  $t$  is the number of nonzero columns in  $z$ . Then,

$$\begin{aligned}
\mathbb{P}(Z^* = z^* \mid k) &= \sum_{z_0 \in \{0,1\}^{n \times k}} \mathbb{P}(Z = z_0 \mid k) I(z_0^* = z^*) \\
&= \#\left\{z_0 \in \{0,1\}^{n \times k} : z_0^* = z^*\right\} w_n(0)^{k-t} \prod_{j=1}^t w_n(s_j(z^*)) \\
&= \binom{k}{t} w_n(0)^{k-t} \prod_{j=1}^t w_n(s_j(z^*)). \tag{6.3.10}
\end{aligned}$$

Writing  $\mathbb{P}(Z^* = z^*) = \sum_{k=0}^{\infty} \mathbb{P}(Z^* = z^* \mid k) p(k)$  yields Equation 6.3.6.

### 6.3.3.2 Staircase form

Let  $m_j(z) = \min\{i : z_{ij} = 1 \text{ or } i > n\}$ . For  $z \in \{0,1\}^{n \times k}$ , define  $z' = [B_1 \ B_2 \ \dots \ B_n]$  where  $B_i$  is the submatrix of  $z$  consisting of the columns  $j$  such that  $m_j(z) = i$ . In other words,  $z'$  is the matrix obtained by permuting the columns of  $z^*$  so that  $m_1(z') \leq \dots \leq m_t(z')$ , where  $t$  is the number of nonzero columns in  $z$ , while preserving the order of the columns within each block  $\{j : m_j(z^*) = i\}$ . We refer to this as *staircase form*.

Then, the distribution of  $Z'$  is as follows: for  $z' \in \{0,1\}^{n \times t}$ ,  $t \in \{0,1,2,\dots\}$ , such that  $m_1(z') \leq \dots \leq m_t(z') \leq n$ ,

$$\mathbb{P}(Z' = z') = \frac{t!}{t_1! \dots t_n!} v_n(t) \prod_{j=1}^t w_n(s_j(z')) \tag{6.3.11}$$

where  $t_i = \#\{j : m_j(z') = i\}$ . This is easily derived from Equation 6.3.6, since there are  $t!/(t_1! \dots t_n!)$  matrices  $z^*$  giving rise to  $z'$ , each of which has probability  $v_n(t) \prod_{j=1}^t w_n(s_j(z'))$ .

Likewise,  $\mathbb{P}_{\text{IBP}}(Z' = z')$  is given by Equation 6.3.8 times  $t!/(t_1! \cdots t_n!)$ ; for reference, this is Equation (5) of Griffiths and Ghahramani (2005).

A case of particular interest arises when  $p(k) = \text{Poisson}(k|\lambda)$ . In this case, we have the closed-form expression

$$v_n(t) = \exp(\lambda w_n(0)) \text{Poisson}(t|\lambda), \quad (6.3.12)$$

by a straightforward calculation. For later reference, we note that in this case,

$$\mathbb{P}(Z' = z') = \frac{\lambda^t}{t_1! \cdots t_n!} \exp\left(-\lambda \sum_{m=1}^n w_m(1)\right) \prod_{j=1}^t w_n(s_j(z')) \quad (6.3.13)$$

since  $1 - w_n(0) = \sum_{m=1}^n w_m(1)$  by Equations 6.3.4 and 6.3.5.

### 6.3.3.3 Left-ordered form

Following Griffiths and Ghahramani (2005), we also consider the reduction to “left-ordered form”. Each column can be viewed as representing a number in binary, taking the first entry to be the most significant bit. For  $z \in \{0, 1\}^{n \times k}$ , define  $z^l$  to be the matrix obtained by permuting the columns of  $z^*$  so that the associated binary numbers are nonincreasing. Such a matrix is said to be in *left-ordered form*.

Similar to the reduction to staircase form, the distribution of  $Z^l$  is

$$\mathbb{P}(Z^l = z^l) = \frac{t!}{\prod_{h=1}^{2^n-1} \tau_h!} v_n(t) \prod_{j=1}^t w_n(s_j(z^l)) \quad (6.3.14)$$

for  $z^l$  in left-ordered form (with no empty columns), where  $\tau_h$  is the number of columns of  $z^l$  representing  $h$  in binary, and  $t = \sum_{h=1}^{2^n-1} \tau_h$  is the total number of

columns in  $z^l$ .

Consider any two matrices  $z_1, z_2 \in \bigcup_{k=0}^{\infty} \{0, 1\}^{n \times k}$  to be equivalent if they have the same left-ordered form, i.e.,  $z_1^l = z_2^l$ . Note that for any permutation  $\sigma$  of the rows, we have  $z_1^l = z_2^l$  if and only if  $\sigma(z_1)^l = \sigma(z_2)^l$ ; consequently, any such  $\sigma$  induces a permutation on the set of equivalence classes. In the same way as the IBP, it is straightforward to show that the distribution on equivalence classes induced by the distribution on left-ordered matrices above (Equation 6.3.14) is exchangeable in the rows, in the sense that it is invariant under such permutations.

### 6.3.4 Buffet process

When  $p(k) = \text{Poisson}(k|\lambda)$ , the distribution of  $Z'$  can be described by a simple buffet process bearing a close resemblance to the Indian buffet process (IBP). Consider the following buffet process:

For  $m = 1, \dots, n$ : Customer  $m$  ...

- (1) tries each previously-trying dish  $j$  with probability  $\frac{a + s_j}{a + b + m - 1}$ , where  $s_j$  is the number of previous customers trying dish  $j$ , and
- (2) tries a  $\text{Poisson}(\lambda w_m(1))$  number of new dishes.

(Customer 1 simply tries a  $\text{Poisson}(\lambda w_1(1))$  number of dishes, since there are no previously-trying dishes.) In the same way as the Indian buffet process, this gives rise to a binary matrix  $z'$  in the staircase form described above; specifically, labeling dishes in the order they are tried,  $z'_{mj} = 1$  if and only if customer  $m$  tries dish  $j$ . Note in particular that, in the notation of Section 6.3.3.2,  $t_m$  is the number of new

dishes tried by customer  $m$ , and  $t = \sum_{m=1}^n t_m$  is the total number of dishes tried.

We claim that a random matrix generated by the buffet process above has the same distribution as  $Z'$  (Equation 6.3.11); see below for this argument.

The only difference between the buffet process above and the IBP is that in the IBP,  $\frac{a + s_j}{a + b + m - 1}$  is replaced by  $s_j/m$ , and  $\lambda w_m(1)$  is replaced by  $\alpha/m$ . In a sense, the IBP can be viewed as a special case obtained by setting  $b = 1$ ,  $\lambda = \alpha/a$ , and taking the limit as  $a \rightarrow 0$ . Further, the two-parameter generalization of the IBP described by Ghahramani et al. (2007) can be obtained similarly by setting  $b = \beta$ ,  $\lambda = \alpha\beta/a$ , and taking  $a \rightarrow 0$ . This two-parameter generalization was motivated by the desire to control the prior on the total number of dishes tried, separately from the number of dishes per customer. In addition to providing such control, the buffet process described above has a further degree of freedom through the parameter  $a$ . This has two effects: (1) the urn process for each column is controlled by both parameters  $a$  and  $b$ , instead of just  $b$ , and (2) the rate of decay of the mean number of new dishes per customer can be controlled by  $a$ , since

$$\lambda w_m(1) = \lambda \frac{ab^{(m-1)}}{(a+b)^{(m)}} \sim \frac{\lambda a \Gamma(a+b)}{m^{a+1} \Gamma(b)}$$

as  $m \rightarrow \infty$  (by Stirling's approximation for the gamma function); in contrast, in the two-parameter IBP the mean number of new dishes for customer  $m$  is  $\alpha\beta/(\beta+m-1)$ .

With this increased flexibility, a diverse variety of matrices can be generated by the MFFM. To illustrate, Figure 6.3 shows samples of  $Z'$  for various parameter settings. To control the total number of dishes tried, we reparametrize using

$$c := \sum_{m=1}^n \lambda w_m(1) = \lambda(1 - w_n(0)) = \lambda \left( 1 - \frac{b^{(n)}}{(a+b)^{(n)}} \right)$$



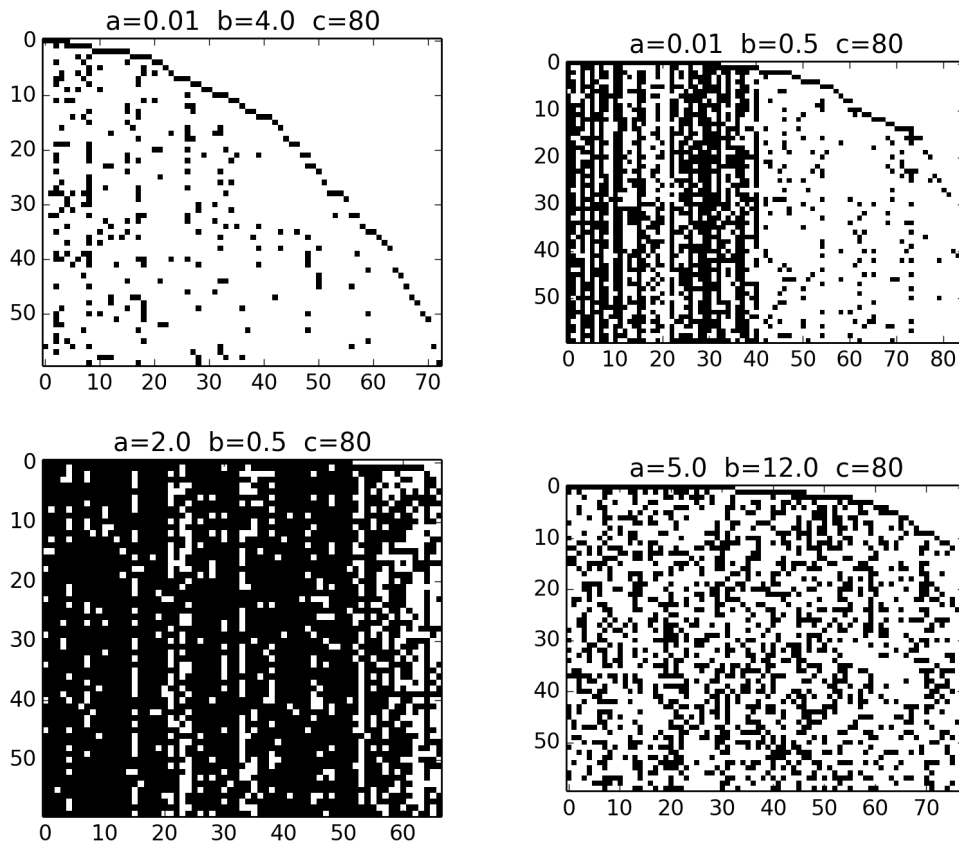


Figure 6.3: Sample matrices from the MFFM buffet process

instead of  $\lambda$ , so that the total number of dishes tried is a  $\text{Poisson}(c)$  random variable.

#### 6.3.4.1 Justification of the buffet process

We show that the buffet process above generates  $Z'$  when  $p(k) = \text{Poisson}(k|\lambda)$ . To see this, fix a  $z'$  obtained via the buffet process. Note that there is a single sequence of choices leading to  $z'$ , so the probability of getting  $z'$  is simply the product of the probabilities of these choices. Suppose  $m_j$  is the first customer to try dish  $j$ , i.e.,  $m_j = m_j(z')$ . For each customer after  $m_j$ , the choice of whether or not to try dish  $j$  follows the urn process of Section 6.3.2, and therefore, defining  $U_1, U_2, \dots$  as in

Section 6.3.2, these choices contribute the factor

$$\begin{aligned} & \mathbb{P}(U_n = z'_{nj}, \dots, U_{m_j+1} = z'_{m_j+1,j} \mid U_{m_j} = 1, U_i = 0 \forall i < m_j) \\ &= \frac{\mathbb{P}(U_{1:n} = z'_{1:n,j})}{\mathbb{P}(U_{m_j} = 1, U_i = 0 \forall i < m_j)} = \frac{w_n(s_j(z'))}{w_{m_j}(1)} \end{aligned} \quad (6.3.15)$$

to the probability of  $z'$ , where the last step follows by applying Equation 6.3.2 to both numerator and denominator. Multiplying these for  $j = 1, \dots, t$ , where  $t$  is the number of columns in  $z'$ , accounts for the choices in step (1) of the buffet process.

For the choices in step (2), let  $t_m$  denote the number of new dishes tried by customer  $m$ , and note that  $t = \sum_{m=1}^n t_m$ . The probability of choosing  $t_1, \dots, t_n$  is

$$\begin{aligned} \prod_{m=1}^n \text{Poisson}(t_m \mid \lambda w_m(1)) &= \prod_{m=1}^n \exp(-\lambda w_m(1)) \frac{(\lambda w_m(1))^{t_m}}{t_m!} \\ &= \exp\left(-\lambda \sum_{m=1}^n w_m(1)\right) \frac{\lambda^t}{t_1! \dots t_n!} \prod_{m=1}^n w_m(1)^{t_m}. \end{aligned}$$

Multiplying this by Equation 6.3.15 for each  $j = 1, \dots, t$ , the  $w_m(1)$  factors cancel and we obtain Equation 6.3.13, as desired.

#### 6.3.4.2 Exchangeable buffet process

As with the IBP, although customers are not exchangeable under the buffet process above, this can be achieved by a certain modification (in the same way as the IBP). This involves generating a matrix in left-ordered form with the distribution of  $Z^l$ , and considering the associated equivalence class. Suppose that the matrix generated by the customers before customer  $m$  has  $\tau_h$  columns representing  $h$  in binary, for  $h = 1, \dots, 2^{m-1} - 1$ . For each  $h$ , customer  $m$  draws  $L \sim \text{Binomial}\left(\tau_h, \frac{a + s_h}{a + b + m - 1}\right)$  and tries the first  $L$  of the  $\tau_h$  associated dishes, where  $s_h$  is the number of ones in

the binary representation of  $h$ . Then, as before, he tries a  $\text{Poisson}(\lambda w_m(1))$  number of new dishes. It can be shown that the probability of obtaining  $z^l$  by this process is  $\mathbb{P}(Z^l = z^l)$  (as in Equations 6.3.14 and 6.3.12) when  $p(k) = \text{Poisson}(k|\lambda)$ .

In the distribution on equivalence classes as in Section 6.3.3.3, the rows are exchangeable, and therefore the customers in this process are exchangeable in the sense that following the process for any ordering of the customers will yield the same distribution on equivalence classes.

### 6.3.5 Inference

The Gibbs sampling algorithm for the IBP described by Griffiths and Ghahramani (2005) is easily adapted to the MFFM. However, it is well-known that mixing can be very slow when using Gibbs sampling with the IBP, and the same issue is present with the MFFM. Other inference algorithms have also been proposed for the IBP (e.g., Teh et al. (2007), Doshi-Velez et al. (2008)), and it might be interesting to see if these approaches can also be adapted to MFFM.

# CHAPTER SEVEN

---

## Conclusion

In this thesis, we have seen that variable-dimension models can be an appealing alternative to the commonly-used infinite-dimensional nonparametric models. In closing, we will speculatively discuss some open questions and possibilities for future work.

The inconsistency results of Chapters 2 and 3 — and more fundamentally, the properties of the partition distribution that lead to this inconsistency — clearly indicate that when the data more closely resembles a finite mixture than an infinite mixture from the assumed family, naively estimating heterogeneity using the number of clusters in an infinite mixture can be problematic. In these situations, estimating heterogeneity using a variable-dimension model may be preferable to using an infinite-dimensional model, however, the effect of misspecification of the component distributions needs to be carefully considered; finding a principled way to handle misspecification would be interesting.

Although we have shown that the DPM posterior on the number of clusters does not concentrate, we have not been able to determine the precise limiting behavior. As a theoretical question, this is challenging and might be interesting to know, although the practical relevance may be limited.

The empirical studies of Chapter 5 indicate that for density estimation, the MFM and the DPM seem essentially indistinguishable. On the other hand, they can be quite different for clustering, and perhaps the most significant difference is in the posteriors on the number of components and clusters, although interestingly, using a DPM with random  $\alpha$  seems to be more similar to the MFM than a DPM with fixed  $\alpha$ . It would be interesting to see practical examples of applications where these differences matter.

Regarding inference algorithms, there are several questions that would be interesting to explore. The slow mixing time of the incremental samplers when  $n$  is large is troublesome, but it seems likely that using one of the existing split-merge algorithms would resolve this issue. Using a smaller value of  $\gamma$  may also help in this regard. With the non-conjugate sampler, are there practical examples in which either DPM or MFM mixing can be significantly improved by approximating the single-cluster posterior, rather than sampling from the prior, for the auxiliary variable distribution? Would it make sense to use stick-breaking representations for inference in variable-dimensional models?

It is striking that so many of the elegant properties of the three nonparametric models considered here — the Dirichlet process, the hierarchical Dirichlet process, and the Indian buffet process — are also exhibited by their variable-dimension counterparts (Chapters 4 and 6). It would be interesting to see how far this can be taken. Do all nonparametric models have a variable-dimension counterpart that exhibits similar properties? Is there a general principle at play here? Is there a broader perspective from which all of these models can be viewed?

# Bibliography

- M. Aitkin. Likelihood and Bayesian analysis of mixtures. *Statistical Modelling*, 1(4): 287–304, 2001.
- C. E. Antoniak. Mixtures of Dirichlet processes with applications to Bayesian non-parametric problems. *The Annals of Statistics*, 2(6):1152–1174, November 1974.
- A. Azzalini and A. Capitanio. Statistical applications of the multivariate skew-normal distribution. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3):579–602, 1999.
- A. Azzalini and A. Dalla Valle. The multivariate skew-normal distribution. *Biometrika*, 83(4):715–726, 1996.
- A. Barron, M. J. Schervish, and L. Wasserman. The consistency of posterior distributions in nonparametric problems. *The Annals of Statistics*, 27(2):536–561, 1999.
- D. Barry and J. A. Hartigan. Product partition models for change point problems. *The Annals of Statistics*, pages 260–279, 1992.
- D. A. Berry and R. Christensen. Empirical Bayes estimation of a binomial parameter via mixtures of Dirichlet processes. *The Annals of Statistics*, pages 558–568, 1979.
- J. Besag and P. J. Green. Spatial statistics and Bayesian computation. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 25–37, 1993.
- A. Bhattacharya and D. B. Dunson. Nonparametric Bayesian density estimation on manifolds with applications to planar shapes. *Biometrika*, 97(4):851–865, 2010.
- D. Blackwell and J. B. MacQueen. Ferguson distributions via Pólya urn schemes. *The Annals of Statistics*, pages 353–355, 1973.
- D. M. Blei and M. I. Jordan. Variational inference for Dirichlet process mixtures. *Bayesian Analysis*, 1(1):121–143, 2006.
- L. Breiman. *Probability (Classics in Applied Mathematics)*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1992.
- T. Broderick, M. I. Jordan, and J. Pitman. Beta processes, stick-breaking and power laws. *Bayesian Analysis*, 7(2):439–476, 2012.

- C. A. Bush and S. N. MacEachern. A semiparametric Bayesian model for randomised block designs. *Biometrika*, 83(2):275–285, 1996.
- C. M. Carvalho, J. Chang, J. E. Lucas, J. R. Nevins, Q. Wang, and M. West. High-dimensional sparse factor modeling: applications in gene expression genomics. *Journal of the American Statistical Association*, 103(484), 2008.
- A. Cerquetti. Generalized Chinese restaurant construction of exchangeable Gibbs partitions and related results. *arXiv:0805.3853*, 2008.
- A. Cerquetti. Conditional  $\alpha$ -diversity for exchangeable Gibbs partitions driven by the stable subordinator. *arXiv:1105.0892*, 2011.
- H. Chen, P. L. Morrell, V. E. T. M. Ashworth, M. de la Cruz, and M. T. Clegg. Tracing the geographic origins of major avocado cultivars. *Journal of Heredity*, 100(1):56–65, 2009.
- J. Chen, P. Li, and Y. Fu. Inference on the order of a normal mixture. *Journal of the American Statistical Association*, 107(499):1096–1105, 2012.
- W.-C. Chen. On the weak form of Zipf’s law. *Journal of Applied Probability*, pages 611–622, 1980.
- Y. Chung and D. B. Dunson. Nonparametric Bayes conditional distribution modeling with variable selection. *Journal of the American Statistical Association*, 104(488), 2009.
- D. B. Dahl. An improved merge-split sampler for conjugate Dirichlet process mixture models. *Technical Report, Department of Statistics, University of Wisconsin Madison*, 2003.
- D. B. Dahl. Sequentially-allocated merge-split sampler for conjugate and nonconjugate Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 11, 2005.
- D. B. Dahl. Model-based clustering for expression data via a Dirichlet process mixture model. *Bayesian inference for gene expression and proteomics*, pages 201–218, 2006.
- D. B. Dahl. Modal clustering in a class of product partition models. *Bayesian Analysis*, 4(2):243–264, 2009.
- N. G. De Bruijn. *Asymptotic Methods in Analysis*. North-Holland Publishing Co., Amsterdam, 1970.
- P. Diaconis and D. Freedman. On the consistency of Bayes estimates. *The Annals of Statistics*, 14(1):1–26, 1986.
- P. Diaconis and D. Ylvisaker. Conjugate priors for exponential families. *The Annals of Statistics*, 7(2):269–281, 1979.
- J. L. Doob. Application of the theory of martingales. In *Actes du Colloque International Le Calcul des Probabilités et ses applications (Lyon, 28 Juin – 3 Juillet, 1948)*, pages 23–27. Paris CNRS, 1949.



- F. Doshi-Velez, K. T. Miller, J. Van Gael, and Y. W. Teh. Variational inference for the Indian buffet process. In *AISTATS 2009*, pages 137–144, 2008.
- M. J. Drinkwater, Q. A. Parker, D. Proust, E. Slezak, and H. Quintana. The large scale distribution of galaxies in the Shapley supercluster. *Publications of the Astronomical Society of Australia*, 21(1):89–96, 2004.
- D. B. Dunson. Efficient Bayesian model averaging in factor analysis. *Technical Report, ISDS, Duke University*, 2006.
- D. B. Dunson and J.-H. Park. Kernel stick-breaking processes. *Biometrika*, 95(2):307–323, 2008.
- R. Durrett. *Probability: Theory and Examples*, volume 2. Cambridge University Press, 1996.
- R. G. Edwards and A. D. Sokal. Generalization of the Fortuin-Kasteleyn-Swendsen-Wang representation and Monte Carlo algorithm. *Physical Review D*, 38:2009–2012, 1988.
- M. D. Escobar. *Estimating the means of several normal populations by nonparametric estimation of the distribution of the means*. PhD thesis, Yale University, 1988.
- M. D. Escobar and M. West. Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, 90(430):577–588, 1995.
- M. D. Escobar and M. West. Computing nonparametric hierarchical models. In D. Dey, P. Müller, and D. Sinha, editors, *Practical Nonparametric and Semiparametric Bayesian Statistics*, pages 1–22. Springer-Verlag, New York, 1998.
- S. Favaro, A. Lijoi, and I. Pruenster. On the stick-breaking representation of normalized inverse Gaussian priors. *Biometrika*, 99(3):663–674, 2012.
- P. Fearnhead. Particle filters for mixture models with an unknown number of components. *Statistics and Computing*, 14(1):11–21, 2004.
- T. S. Ferguson. A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, pages 209–230, 1973.
- T. S. Ferguson. Bayesian density estimation by mixtures of normal distributions. In M. H. Rizvi, J. Rustagi, and D. Siegmund, editors, *Recent Advances in Statistics*, pages 287–302. Academic Press, 1983.
- E. B. Fox, E. B. Sudderth, and A. S. Willsky. Hierarchical Dirichlet processes for tracking maneuvering targets. In *10th International Conference on Information Fusion, 2007*, pages 1–8. IEEE, 2007.
- E. B. Fox, E. B. Sudderth, M. I. Jordan, and A. S. Willsky. Sharing features among dynamical systems with beta processes. In *Advances in Neural Information Processing Systems*, pages 549–557, 2009.
- D. A. Freedman. On the asymptotic behavior of Bayes’ estimates in the discrete case. *The Annals of Mathematical Statistics*, pages 1386–1403, 1963.

- Z. Ghahramani, T. L. Griffiths, and P. Sollich. Bayesian nonparametric latent feature models. 2007.
- S. Ghosal. The Dirichlet process, related priors and posterior asymptotics. In N. L. Hjort, C. Holmes, P. Müller, and S. G. Walker, editors, *Bayesian Nonparametrics*, pages 36–83. Cambridge University Press, 2010.
- S. Ghosal and A. Van der Vaart. Posterior convergence rates of Dirichlet mixtures at smooth densities. *The Annals of Statistics*, 35(2):697–723, 2007.
- S. Ghosal and A. W. Van der Vaart. Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities. *The Annals of Statistics*, pages 1233–1263, 2001.
- S. Ghosal, J. K. Ghosh, and R. V. Ramamoorthi. Posterior consistency of Dirichlet mixtures in density estimation. *The Annals of Statistics*, 27(1):143–158, 1999.
- J. Ghosh and D. B. Dunson. Default prior distributions and efficient posterior computation in Bayesian factor analysis. *Journal of Computational and Graphical Statistics*, 18(2):306–320, 2009.
- J. K. Ghosh and R. V. Ramamoorthi. *Bayesian Nonparametrics*. Springer–Verlag, New York, 2003.
- S. K. Ghosh and S. Ghosal. Semiparametric accelerated failure time models for censored data. *Bayesian Statistics and its Applications*, pages 213–229, 2006.
- A. Gnedin. A species sampling model with finitely many types. *Elect. Comm. Probab.*, 15:79–88, 2010.
- A. Gnedin and J. Pitman. Exchangeable Gibbs partitions and Stirling triangles. *Journal of Mathematical Sciences*, 138(3):5674–5685, 2006.
- E. G. Gonzalez and R. Zardoya. Relative role of life-history traits and historical factors in shaping genetic population structure of sardines (*Sardina pilchardus*). *BMC evolutionary biology*, 7(1):197, 2007.
- R. L. Graham, D. E. Knuth, and O. Patashnik. *Concrete Mathematics*. Addison-Wesley, 1989.
- P. J. Green and S. Richardson. Modeling heterogeneity with and without the Dirichlet process. *Scandinavian Journal of Statistics*, 28(2):355–375, June 2001.
- J. E. Griffin and M. J. Steel. Order-based dependent Dirichlet processes. *Journal of the American Statistical Association*, 101(473):179–194, 2006.
- T. L. Griffiths and Z. Ghahramani. Infinite latent feature models and the Indian buffet process. In *Advances in Neural Information Processing Systems*, volume 18, pages 475–482, 2005.
- T. L. Griffiths, K. R. Canini, A. N. Sanborn, and D. J. Navarro. Unifying rational models of categorization via the hierarchical Dirichlet process. In *Proceedings of the 29th Annual Cognitive Science Society*, pages 323–328, 2007.

- B. Hansen and J. Pitman. Prediction rules for exchangeable sequences related to species sampling. *Statistics & probability letters*, 46(3):251–256, 2000.
- J. Henna. On estimating of the number of constituents of a finite mixture of continuous distributions. *Annals of the Institute of Statistical Mathematics*, 37(1):235–240, 1985.
- J. Henna. Estimation of the number of components of finite mixtures of multivariate distributions. *Annals of the Institute of Statistical Mathematics*, 57(4):655–664, 2005.
- N. L. Hjort. Bayesian analysis for a generalised Dirichlet process prior. *Technical Report, University of Oslo*, 2000.
- M.-W. Ho, L. F. James, and J. W. Lau. Gibbs partitions (EPPF's) derived from a stable subordinator are Fox H and Meijer G transforms. *arXiv:0708.0619*, 2007.
- W. Hoeffding. The strong law of large numbers for U-statistics. *Institute of Statistics, Univ. of N. Carolina, Mimeograph Series*, 302, 1961.
- M. D. Hoffman, D. M. Blei, and P. R. Cook. Content-based musical similarity computation using the hierarchical Dirichlet process. In *ISMIR*, pages 349–354, 2008.
- J. Hoffmann-Jørgensen. *Probability with a View toward Statistics*, volume 2. Chapman & Hall, 1994.
- J. P. Huelsenbeck and P. Andolfatto. Inference of population structure under a Dirichlet process model. *Genetics*, 175(4):1787–1802, 2007.
- H. Ishwaran and L. F. James. Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, 96(453), 2001.
- H. Ishwaran and L. F. James. Generalized weighted Chinese restaurant processes for species sampling mixture models. *Statistica Sinica*, 13(4):1211–1236, 2003.
- H. Ishwaran and M. Zarepour. Markov chain Monte Carlo in approximate Dirichlet and beta two-parameter process hierarchical models. *Biometrika*, 87(2):371–390, 2000.
- H. Ishwaran, L. F. James, and J. Sun. Bayesian model selection in finite mixtures by marginal density decompositions. *Journal of the American Statistical Association*, 96(456), 2001.
- S. Jain and R. M. Neal. A split-merge Markov chain Monte Carlo procedure for the Dirichlet process mixture model. *Journal of Computational and Graphical Statistics*, 13(1), 2004.
- S. Jain and R. M. Neal. Splitting and merging components of a nonconjugate Dirichlet process mixture model. *Bayesian Analysis*, 2(3):445–472, 2007.
- L. F. James. Large sample asymptotics for the two-parameter Poisson–Dirichlet process. In *Pushing the Limits of Contemporary Statistics: Contributions in Honor of Jayanta K. Ghosh*, pages 187–199. Institute of Mathematical Statistics, 2008.

- L. F. James, C. E. Priebe, and D. J. Marchette. Consistent estimation of mixture complexity. *The Annals of Statistics*, pages 1281–1296, 2001.
- G. H. Jang, J. Lee, and S. Lee. Posterior consistency of species sampling priors. *Statistica Sinica*, 20(2):581, 2010.
- A. Jasra, C. Holmes, and D. Stephens. Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling. *Statistical Science*, pages 50–67, 2005.
- M. Kalli, J. E. Griffin, and S. G. Walker. Slice sampling mixture models. *Statistics and computing*, 21(1):93–105, 2011.
- C. Keribin. Consistent estimation of the order of mixture models. *Sankhya Ser. A*, 62(1):49–66, 2000.
- S. Khazaei, J. Rousseau, and F. Balabdaoui. Nonparametric Bayesian estimation of densities under monotonicity constraint. (*preprint*), 2012.
- S. Kim, M. G. Tadesse, and M. Vannucci. Variable selection in clustering via Dirichlet process mixture models. *Biometrika*, 93(4):877–893, 2006.
- J. J. Kivinen, E. B. Sudderth, and M. I. Jordan. Learning multiscale representations of natural scenes using Dirichlet processes. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–8. IEEE, 2007.
- A. W. Knappp. *Basic Real Analysis*. Birkhäuser, 2005.
- D. Knowles and Z. Ghahramani. Infinite sparse factor analysis and infinite independent components analysis. In *Independent Component Analysis and Signal Separation*, pages 381–388. Springer, 2007.
- S. G. Krantz. *Function Theory of Several Complex Variables*. AMS Chelsea Publishing, Providence, 1992.
- R. Krause, D. L. Wild, W. Chu, and Z. Ghahramani. Identifying protein complexes in high-throughput protein interaction screens using an infinite latent feature model. In *Pacific Symposium on Biocomputing*, volume 11, pages 231–242, 2006.
- W. Kruijjer. *Convergence rates in nonparametric Bayesian density estimation*. PhD thesis, Department of Mathematics, Vrije Universiteit Amsterdam, 2008.
- W. Kruijjer, J. Rousseau, and A. Van der Vaart. Adaptive Bayesian density estimation with location-scale mixtures. *Electronic Journal of Statistics*, 4:1225–1257, 2010.
- N. Lartillot and H. Philippe. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Molecular Biology and Evolution*, 21(6):1095–1109, 2004.
- A. D. Leaché and M. K. Fujita. Bayesian species delimitation in West African forest geckos (*Hemidactylus fasciatus*). *Proceedings of the Royal Society B: Biological Sciences*, 277(1697):3071–3077, 2010.

- J. K. Lee, V. Dančák, and M. S. Waterman. Estimation for restriction sites observed by optical mapping using reversible-jump Markov chain Monte Carlo. *Journal of Computational Biology*, 5(3):505–515, 1998.
- B. G. Leroux. Consistent estimation of a mixing distribution. *The Annals of Statistics*, 20(3):1350–1360, 1992.
- P. Liang, S. Petrov, M. I. Jordan, and D. Klein. The infinite PCFG using hierarchical Dirichlet processes. In *EMNLP-CoNLL*, pages 688–697, 2007.
- A. Lijoi and I. Prünster. Models beyond the Dirichlet process. *Bayesian Nonparametrics*, 28:80, 2010.
- A. Lijoi, R. H. Mena, and I. Prünster. Hierarchical mixture modeling with normalized inverse-Gaussian priors. *Journal of the American Statistical Association*, 100(472):1278–1291, 2005a.
- A. Lijoi, I. Prünster, and S. G. Walker. On consistency of nonparametric normal mixtures for Bayesian density estimation. *Journal of the American Statistical Association*, 100(472):1292–1296, 2005b.
- A. Lijoi, R. H. Mena, and I. Prünster. Bayesian nonparametric estimation of the probability of discovering new species. *Biometrika*, 94(4):769–786, 2007.
- A. Lijoi, I. Prünster, and S. G. Walker. Bayesian nonparametric estimators derived from conditional Gibbs structures. *The Annals of Applied Probability*, 18(4):1519–1547, 2008.
- J. S. Liu. The collapsed Gibbs sampler in Bayesian computations with applications to a gene regulation problem. *Journal of the American Statistical Association*, 89(427):958–966, 1994.
- A. Y. Lo. On a class of Bayesian nonparametric estimates: I. Density estimates. *The Annals of Statistics*, 12(1):351–357, 1984.
- H. F. Lopes and M. West. Bayesian model assessment in factor analysis. *Statistica Sinica*, 14(1):41–68, 2004.
- E. D. Lorenzen, P. Arctander, and H. R. Siegismund. Regional genetic structuring and evolutionary history of the impala *Aepyceros melampus*. *Journal of Heredity*, 97(2):119–132, 2006.
- S. N. MacEachern. Estimating normal means with a conjugate style Dirichlet process prior. *Communications in Statistics-Simulation and Computation*, 23(3):727–741, 1994.
- S. N. MacEachern. Computational methods for mixture of Dirichlet process models. In *Practical nonparametric and semiparametric Bayesian statistics*, pages 23–43. Springer, 1998.
- S. N. MacEachern. Dependent dirichlet processes. *Unpublished manuscript, Department of Statistics, The Ohio State University*, 2000.

- S. N. MacEachern and P. Müller. Estimating mixture of Dirichlet process models. *Journal of Computational and Graphical Statistics*, 7(2):223–238, 1998.
- S. N. MacEachern, M. Clyde, and J. S. Liu. Sequential importance sampling for nonparametric Bayes models: The next generation. *Canadian Journal of Statistics*, 27(2):251–267, 1999.
- S. N. MacEachern, A. Kottas, and A. Gelfand. Spatial nonparametric Bayesian models. In *Proceedings of the 2001 Joint Statistical Meetings*, volume 3, 2001.
- J. Marin, M. Rodriguez-Bernal, and M. Wiper. Using Weibull mixture distributions to model heterogeneous survival data. *Communications in Statistics Simulation and Computation*, 34(3):673–684, 2005.
- P. McCullagh and J. Yang. How many clusters? *Bayesian Analysis*, 3(1):101–120, 2008.
- M. Medvedovic and S. Sivaganesan. Bayesian infinite mixture model based clustering of gene expression profiles. *Bioinformatics*, 18(9):1194–1206, 2002.
- E. Meeds, Z. Ghahramani, R. M. Neal, and S. T. Roweis. Modeling dyadic data with binary latent factors. *Advances in Neural Information Processing Systems*, 19:977, 2007.
- S. P. Meyn and R. L. Tweedie. *Markov Chains and Stochastic Stability*. Cambridge University Press, 2009.
- J. W. Miller and M. T. Harrison. A simple example of Dirichlet process mixture inconsistency for the number of components. In *Advances in Neural Information Processing Systems, Vol. 26*, 2013a.
- J. W. Miller and M. T. Harrison. Inconsistency of Pitman–Yor process mixtures for the number of components. *arXiv:1309.0024*, 2013b.
- K. T. Miller, T. L. Griffiths, and M. I. Jordan. Nonparametric latent feature models for link prediction. In *Advances in Neural Information Processing Systems*, volume 9, pages 1276–1284, 2009.
- P. Müller and F. Quintana. Random partition models with regression on covariates. *Journal of Statistical Planning and Inference*, 140(10):2801–2808, 2010.
- P. Müller, F. Quintana, and G. Rosner. A method for combining inference across related nonparametric Bayesian models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(3):735–749, 2004.
- P. Müller, F. Quintana, and G. L. Rosner. A product partition model with regression on covariates. *Journal of Computational and Graphical Statistics*, 20(1), 2011.
- L. Murino, C. Angelini, I. Bifulco, I. De Feis, G. Raiconi, and R. Tagliaferri. Multiple clustering solutions analysis through least-squares consensus algorithms. In *Computational Intelligence Methods for Bioinformatics and Biostatistics*, pages 215–227. Springer, 2010.

- D. J. Navarro and T. L. Griffiths. A nonparametric Bayesian method for inferring features from similarity judgments. *Advances in Neural Information Processing Systems*, 19:1033, 2007.
- R. M. Neal. Bayesian mixture modeling. In *Maximum Entropy and Bayesian Methods*, pages 197–211. Springer, 1992.
- R. M. Neal. Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9(2):249–265, 2000.
- D. P. Nguyen, L. M. Frank, and E. N. Brown. An application of reversible-jump Markov chain Monte Carlo to spike classification of multi-unit extracellular recordings. *Network (Bristol, England)*, 14(1):61–82, 2003.
- X. L. Nguyen. Convergence of latent mixing measures in finite and infinite mixture models. *The Annals of Statistics*, 41(1):370–400, 2013.
- A. Nobile. *Bayesian Analysis of Finite Mixture Distributions*. PhD thesis, Department of Statistics, Carnegie Mellon University, Pittsburgh, PA, 1994.
- A. Nobile. Bayesian finite mixtures: a note on prior specification and posterior computation. *Technical Report, Department of Statistics, University of Glasgow*, 2005.
- A. Nobile and A. T. Fearnside. Bayesian finite mixtures with an unknown number of components: The allocation sampler. *Statistics and Computing*, 17(2):147–162, 2007.
- A. Onogi, M. Nurimoto, and M. Morita. Characterization of a Bayesian genetic clustering algorithm based on a Dirichlet process prior and comparison among Bayesian clustering methods. *BMC Bioinformatics*, 12(1):263, 2011.
- E. Otranto and G. M. Gallo. A nonparametric Bayesian approach to detect the number of regimes in Markov switching models. *Econometric Reviews*, 21(4):477–496, 2002.
- J. W. Paisley, A. K. Zaas, C. W. Woods, G. S. Ginsburg, and L. Carin. A stick-breaking construction of the beta process. In *Proceedings of the 27th International Conference on Machine Learning*, pages 847–854, 2010.
- O. Papaspiliopoulos and G. O. Roberts. Retrospective Markov chain Monte Carlo methods for Dirichlet process hierarchical models. *Biometrika*, 95(1):169–186, 2008.
- J.-H. Park and D. B. Dunson. Bayesian generalized product partition model. *Statistica Sinica*, 20:1203–1226, 2010.
- D. Pati, D. B. Dunson, and S. T. Tokdar. Posterior consistency in conditional distribution estimation. *Journal of Multivariate Analysis*, 116:456–472, 2013.
- J. Pella and M. Masuda. The Gibbs and split-merge sampler for population mixture analysis from genetic data with incomplete baselines. *Canadian Journal of Fisheries and Aquatic Sciences*, 63(3):576–596, 2006.

- M. Perman, J. Pitman, and M. Yor. Size-biased sampling of Poisson point processes and excursions. *Probability Theory and Related Fields*, 92(1):21–39, 1992.
- D. B. Phillips and A. F. M. Smith. Bayesian model comparison via jump diffusions. In *Markov chain Monte Carlo in Practice*, pages 215–239. Springer, 1996.
- J. Pitman. Exchangeable and partially exchangeable random partitions. *Probability Theory and Related Fields*, 102(2):145–158, 1995.
- J. Pitman. Some developments of the Blackwell-MacQueen urn scheme. *Lecture Notes-Monograph Series*, pages 245–267, 1996.
- J. Pitman. *Combinatorial Stochastic Processes*. Springer-Verlag, Berlin, 2006.
- J. Pitman and M. Yor. The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *The Annals of Probability*, 25(2):855–900, 1997.
- J. K. Pritchard, M. Stephens, and P. Donnelly. Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–959, 2000.
- F. A. Quintana. A predictive view of Bayesian clustering. *Journal of Statistical Planning and Inference*, 136(8):2407–2429, 2006.
- F. A. Quintana and P. L. Iglesias. Bayesian clustering and product partition models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(2):557–574, 2003.
- L. Ren, D. B. Dunson, and L. Carin. The dynamic hierarchical Dirichlet process. In *Proceedings of the 25th International Conference on Machine Learning*, pages 824–831. ACM, 2008.
- C. M. Richards, G. M. Volk, A. A. Reilley, A. D. Henk, D. R. Lockwood, P. A. Reeves, and P. L. Forsline. Genetic diversity and population structure in *Malus sieversii*, a wild progenitor species of domesticated apple. *Tree Genetics & Genomes*, 5(2):339–347, 2009.
- S. Richardson and P. J. Green. On Bayesian analysis of mixtures with an unknown number of components. *Journal of the Royal Statistical Society, Series B*, 59(4):731–792, 1997.
- C. P. Robert and G. Casella. *Monte Carlo Statistical Methods*, volume 319. Springer, 2004.
- R. T. Rockafellar. *Convex Analysis*. Princeton University Press, 1970.
- A. Rodriguez and D. B. Dunson. Nonparametric Bayesian models through probit stick-breaking processes. *Bayesian Analysis*, 6(1), 2011.
- A. Rodriguez, D. B. Dunson, and A. E. Gelfand. The nested Dirichlet process. *Journal of the American Statistical Association*, 103(483), 2008.
- K. Roeder. Density estimation with confidence sets exemplified by superclusters and voids in the galaxies. *Journal of the American Statistical Association*, 85(411):617–624, 1990.



- K. Roeder. A graphical technique for determining the number of components in a mixture of normals. *Journal of the American Statistical Association*, 89(426):487–495, 1994.
- J. Rousseau and K. Mengersen. Asymptotic behaviour of the posterior distribution in overfitted mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(5):689–710, 2011.
- S. Roweis and Z. Ghahramani. A unifying review of linear Gaussian models. *Neural Computation*, 11(2):305–345, 1999.
- T. Sapatinas. Identifiability of mixtures of power-series distributions and related characterizations. *Annals of the Institute of Statistical Mathematics*, 47(3):447–459, 1995.
- C. Scricciolo. Adaptive Bayesian density estimation using Pitman–Yor or normalized inverse-Gaussian process kernel mixtures. *arXiv:1210.8094*, 2012.
- J. Sethuraman. A constructive definition of Dirichlet priors. *Statistica Sinica*, 4:639–650, 1994.
- J. Sethuraman and R. C. Tiwari. Convergence of Dirichlet measures and the interpretation of their parameter. *Technical Report, Department of Statistics, Florida State University*, 1981.
- M. Stephens. Bayesian analysis of mixture models with an unknown number of components—An alternative to reversible jump methods. *The Annals of Statistics*, 28(1):40–74, 2000.
- Y. Tang and S. Ghosal. Posterior consistency of Dirichlet mixtures for estimating a transition density. *Journal of Statistical Planning and Inference*, 137(6):1711–1726, June 2007.
- Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Sharing clusters among related groups: hierarchical Dirichlet processes. In *Advances in Neural Information Processing Systems*, pages 1385–1392, 2004.
- Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.
- Y. W. Teh, D. Görür, and Z. Ghahramani. Stick-breaking construction for the Indian buffet process. In *International Conference on Artificial Intelligence and Statistics*, pages 556–563, 2007.
- H. Teicher. Identifiability of finite mixtures. *The Annals of Mathematical Statistics*, 34(4):1265–1269, 1963.
- R. Thibaux and M. I. Jordan. Hierarchical beta processes and the Indian buffet process. In *International Conference on Artificial Intelligence and Statistics*, pages 564–571, 2007.
- L. Tierney. Markov chains for exploring posterior distributions. *The Annals of Statistics*, pages 1701–1728, 1994.

- S. T. Tokdar. Posterior consistency of Dirichlet location-scale mixture of normals in density estimation and regression. *Sankhyā: The Indian Journal of Statistics*, pages 90–110, 2006.
- G. Tomlinson and M. Escobar. Analysis of densities. *Technical Report, University of Toronto*, 1999.
- C. Vogl, F. Sanchez-Cabo, G. Stocker, S. Hubbard, O. Wolkenhauer, and Z. Trajanoski. A fully Bayesian model to cluster gene-expression profiles. *Bioinformatics*, 21(suppl 2):ii130–ii136, 2005.
- S. G. Walker, A. Lijoi, and I. Prünster. On rates of convergence for posterior distributions in infinite-dimensional models. *The Annals of Statistics*, 35(2):738–746, 2007.
- M. West. Hyperparameter estimation in Dirichlet process mixture models. *ISDS Discussion Paper #92-A03, Duke University*, 1992.
- M. West. Bayesian factor regression models in the “large p, small n” paradigm. In J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith, and M. West, editors, *Bayesian Statistics 7*. Oxford University Press, 2003.
- M. West, P. Müller, and M. D. Escobar. Hierarchical priors and mixture models, with application in regression and density estimation. In P. Freeman and A. F. Smith, editors, *Aspects of Uncertainty: A Tribute to D. V. Lindley*, pages 363–386. Wiley, 1994.
- M.-J. Woo and T. Sriram. Robust estimation of mixture complexity for count data. *Computational Statistics and Data Analysis*, 51(9):4379–4392, 2007.
- M.-J. Woo and T. N. Sriram. Robust estimation of mixture complexity. *Journal of the American Statistical Association*, 101(476), 2006.
- F. Wood, T. L. Griffiths, and Z. Ghahramani. A non-parametric Bayesian method for inferring hidden causes. In *Proceedings of the 22nd Conference on Uncertainty in Artificial Intelligence (UAI)*, 2006.
- Y. Wu and S. Ghosal. The  $L_1$ -consistency of Dirichlet mixtures in multivariate Bayesian density estimation. *Journal of Multivariate Analysis*, 101(10):2411–2419, November 2010.
- E. P. Xing, K. A. Sohn, M. I. Jordan, and Y. W. Teh. Bayesian multi-population haplotype inference via a hierarchical Dirichlet process mixture. In *Proceedings of the 23rd International Conference on Machine Learning*, pages 1049–1056, 2006.
- S. J. Yakowitz and J. D. Spragins. On the identifiability of finite mixtures. *The Annals of Mathematical Statistics*, 39(1):209–214, 1968.