Probabilistic Inference When the Model is Wrong

Diana Cai

A Dissertation
Presented to the Faculty
of Princeton University
in Candidacy for the Degree
of Doctor of Philosophy

Recommended for Acceptance
by the Department of
Computer Science
Advisers: Ryan P. Adams and Barbara E. Engelhardt

September 2023

© Copyright by Diana Cai, 2023.

All Rights Reserved

Abstract

By simplifying complex real-world phenomena, probabilistic methods have proven able to accelerate applications in discovery and design. However, classical theory often evaluates models under the assumption that they are perfect representations of the observed data. There remains the danger that these simplifications might sometimes lead to failure under real-world conditions. This dissertation identifies popular data analyses that can yield unreliable conclusions—and in some cases ones that are arbitrarily unreliable—under such "misspecification." But we also show how to practically navigate misspecification. To begin, we consider clustering, a mainstay of modern unsupervised data analysis, using Bayesian finite mixture models. Some scientists are interested not only in finding meaningful groups of data but also in learning the number of such clusters. We provide novel theoretical results and empirical studies showing that, no matter how small the misspecification, some common approaches, including Bayesian robustness procedures, for learning the number of clusters give increasingly wrong answers as one receives more data. But using imperfect models need not be hopeless. For instance, we consider a popular Bayesian modeling framework for graphs based on the assumption of vertex exchangeability. A consequence of this assumption is that the resulting graph models generate dense graphs with probability 1 and are therefore misspecified for sparse graphs, a common property of many real-world graphs. To address this undesirable scaling behavior, we introduce an alternative generative modeling framework and prove that it generates a range of sparse and dense scaling behaviors; we also show empirically that it can generate graphs with sparse power law scaling behavior. Finally, we consider the case where a researcher has access to a sequence of approximate models that become arbitrarily more complex at the cost of more computation, which is common in applications with simulators of physical dynamics or models requiring numerical approximations of some fidelity. In this case, we show how to obtain estimates as though one had access to the most complex model. In particular, we propose a framework for constructing Markov chain Monte Carlo algorithms that asymptotically simulates from the most complex model while only ever evaluating models from the sequence of approximate models.

Acknowledgements

I am fortunate to have many mentors, collaborators, and friends that have enriched my time at Princeton. I feel immense gratitute towards my co-advisor, Ryan Adams, for his support and wisdom over many different stages of my research trajectory from Harvard to Princeton. Observing Ryan's big picture thinking and creativity has led me to aspire towards tackling impactful and challenging problems. My second co-advisor, Barbara Engelhardt, welcomed me into her group at Princeton, through which I found an incredible set of collaborators and friends. Barbara taught me to appreciate computational biology research, and I have looked towards her as a prime example of a mentor, teacher, and leader. Tamara Broderick has been a long-term collaborator, mentor, and inspiration. Through Tamara I gained an appreciation for theoretical and statistics research, and I have also learned countless lessons on becoming a better communicator, collaborator, and researcher. I also appreciate being included in the Broderick group activities through which I have gained another research family. Tom Griffiths and Olga Russakovsky have also provided useful advice on my work as members of my generals/dissertation committees.

I am fortunate to have many collaborators that have enriched and broadened my research experiences: Cameron Freer, Nate Ackerman, Trevor Campbell, Tamara Broderick, Michael Mitzenmacher, Ryan Adams, Rishit Sheth, Lester Mackey, Nicolo Fusi, Greg Gundersen, Chuteng Zhou, David Zoltowski, Andy Jones, Didong Li, Aishwarya Mandyam, and Barbara Engelhardt. I am also grateful for the countless friends at and beyond Princeton that I've met through being office/lab mates, classes, conferences, and internships for their endless encouragement and camaraderie.

A number of friends and family outside of Princeton have made my days immeasurably brighter. Nilesh Tripuraneni and Jeffrey Chan have been a constant source of laughter and comfort. Likewise, I am grateful to Rediet Abebe and Fiona Wood for their years of friendship, wisdom, and humor. Thank you to Spencer Liang for the steadfast support, for challenging me throughout the years, and for being a great source of joy. Finally, thank you to my family for being my greatest source of inspiration and for being there for me no matter what.

This dissertation was supported in part by a Google Ph.D. Fellowship in Machine Learning.

To my family.

Contents

List of Figures

List of Symbols and Abbreviations

i.i.d.	independent and identically distributed
ind.	independent
a.s.	almost surely
a.e.	almost everywhere
p.d.f.	probablity density function
$\stackrel{d}{=}$	equal in distribution
$\mathbb{N}$	$\{1, 2, \ldots\}$
$[N]$	$\{1, \ldots, N\}$
$X_{1:N}$	$(X_1, \ldots, X_N)$
$\lvert\mathcal{S}\rvert$	cardinality of a set $\mathcal{S}$
$X_n \stackrel{\text{a.s.}}{=} O(Y_n)$	$\limsup_{n\to\infty} \frac{X_n}{Y_n} < \infty$ a.s., for random sequences (X_n), (Y_n)
$X_n \stackrel{\text{a.s.}}{=} \Omega(Y_n)$	$Y_n \stackrel{\text{a.s.}}{=} O(X_n)$ a.s., for random sequences (X_n), (Y_n)
$X_n \stackrel{\text{a.s.}}{=} o(Y_n)$	$\lim_{n\to\infty} \frac{X_n}{Y_n} = 0$ a.s., for random sequences (X_n), (Y_n)
$X_n \stackrel{\text{a.s.}}{=} \Theta(Y_n)$	$X_n \stackrel{\text{a.s.}}{=} O(Y_n)$ and $Y_n \stackrel{\text{a.s.}}{=} O(X_n)$
$\mathcal{N}(\cdot,\cdot)$, $\mathcal{N}(\cdot \mid \cdot,\cdot)$	Normal distribution, probability density function

Chapter 1

Introduction

Machine learning is transforming the sciences with new data-driven methods to accelerate scientific discovery—from developing novel scientific understanding to suggesting new experiments or simulations. Some key desiderata of computational methods in scientific applications include the ability to incorporate scientific knowledge and theories into models, to learn interpretable, expressive representations of complex observations, and to reliably quantify the uncertainty in a method's output. *Probabilistic machine learning* provides a powerful framework for tackling these desiderata, as it is natural to encode scientific domain knowledge in models, extract interpretable, low-dimensional latent structure, and quantify and propagate uncertainty for application in downstream tasks such as prediction, inference, or decision making using statistical inference algorithms.

However, a number of challenges remain for reliably using these methods in modern scientific problems. Crucially, probabilistic methods operate on the assumption that the model is correct. Complex models are often necessary to accurately learn about sophisticated real-world phenomena, but even with careful model checking, some amount of *model misspecification* is inevitable. Additionally, probabilistic inference in scientific applications is often expensive due to costly experimental measurements, high-fidelity physical simulations, or massive data sets. This necessitates the application of approximate inference algorithms to fit these models. For instance, in variational Bayesian inference, an approximating family to the posterior is used; in models of physical systems,

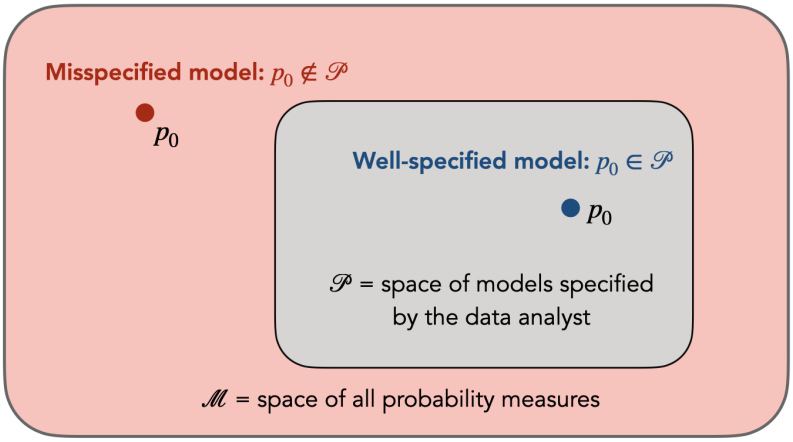


Figure 1.1.1: Schematic of well-specified and misspecified models.

simplified assumptions may be needed for computational reasons; or the complexity of the problem may require numerical approximations or data summarization algorithms.

Probabilistic inference and resulting downstream tasks are an important part of decision making—by scientists, engineers, doctors, and policy makers—that have the potential to drastically affect individuals' livelihoods. Thus, it increasingly important to develop reliable data analysis tools that account for misspecification and are computationally efficient.

1.1 A cartoon of probabilistic modeling

A probabilistic generative model is, of necessity, a simplification of the complex real-world phenomena that govern any observed data, and in many cases facilitates tractable data analysis and discovery of meaningful and actionable patterns in data. But typically any model of a real-world data set is *misspecified*. In particular, some types of misspecification can be dangerous, in that they may lead to fundamentally inaccurate or misleading inferences. Thus, it is crucial to understand what types of misspecification are leading to useful inferences, and what types of misspecification are problematic.

In Figure 1.1.1, we show a schematic that represents one modeling setting we consider in this dissertation. Here $\mathscr{M}$ (red box) depicts a space of probability measures on some set of interest. The

data analyst specifies a set of models $\mathcal{P}$ (gray box) that is some subset of this space. We assume there is some true, unknown data generating distribution p_0. When p_0 is contained in the class of models $\mathcal{P}$, we say the model is *well-specified*; otherwise, the model is *misspecified*.

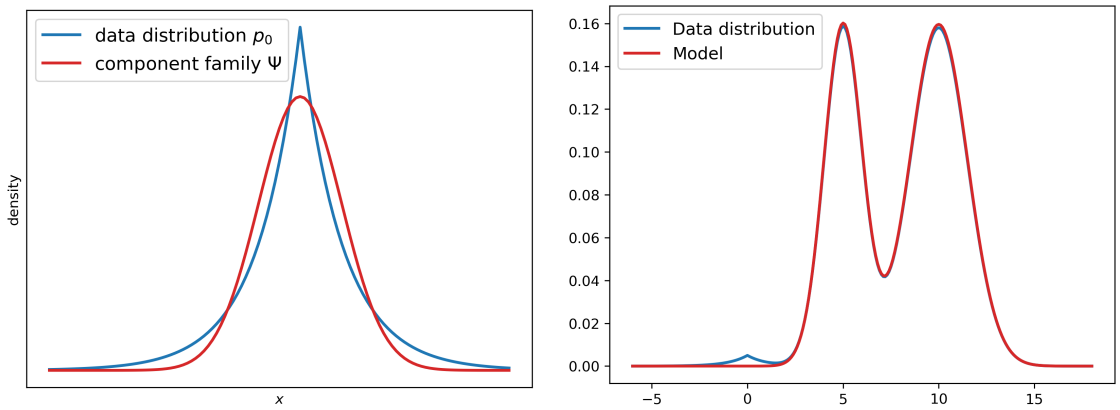
In this dissertation, we consider a few main types of misspecification: 1) Bayesian models under likelihood misspecification, 2) Bayesian models under prior misspecification, and 3) approximate Bayesian models due to limited computation. In what follows, we will discuss each of these types of misspecification in more depth, and give an overview of case studies that we expand on in later chapters of the dissertation.

1.2 Bayesian models under likelihood misspecification

One type of misspecification in Bayesian models is that of the specification of the likelihood. In this dissertation, we consider one particular example of likelihood misspecification. Latent variable models—such as (ad)mixture models and latent factor models—are applied widely in a number of scientific domains, such as genomics, neuroscience, and astronomy, to automatically discover scientifically interpretable structure in high-dimensional data by modeling the latent subpopulations (e.g., cell types or genetic subpopulations) from which the data are generated. However, the models are sensitive to the specification of the distribution of these subpopulations, which can lead to misleading inferences of important quantities of interest ranging from the properties of the latent subpopulations to the number of latent subpopulations.

A particular problem scientists and engineers are often interested in is learning the number of subpopulations, or components, present in a data set. A common suggestion is to use a finite mixture model (FMM) with a prior on the number of components, e.g., a geometric or Poisson distribution. Past work has shown the resulting distribution on the number of components is consistent; i.e., the FMM component-count posterior is consistent; that is, the posterior concentrates on the true, generating number of components. But consistency requires the assumption that the component likelihoods are perfectly specified, which is unrealistic in practice.

Indeed, empirical evidence from past work has shown that finite mixture models are sensitive to



(a) Component family misspecification in a mixture.

(b) Example of ϵ-contamination.

Figure 1.2.1: Example of model misspecification in finite mixture models for clustering.

the specification of the likelihood (Miller and Harrison, 2018) and often overestimate the number of components (Frühwirth-Schnatter, 2006). Figure 1.2.1a illustrates an example of misspecification in a finite mixture model, where there is some true data generating distribution p_0 that generated the data, for instance, a finite mixture of Gaussian distributions (or in this example a single Gaussian), but the data analyst specifies a model that is a mixture of the Laplace component family. Intuitively, the posterior will pick up on small clusters due to misspecification, and it may be unsurprising that this particular example has issues. But one might also wonder: what is the behavior of the posterior if the model is just a little bit misspecified? For instance, suppose the data analyst specifies a model given by a finite mixture of Gaussians but the true data generating distribution is a small ϵ-contamination of this model, i.e., $p_0 = (1-\epsilon)p + \epsilon q$, where p is an element of our model class, and q is some other distribution (Figure 1.2.1b).

Furthermore, there are a number of other related questions the data analyst might ask when trying to improve the specification of a finite mixture model:

- What is the behavior of the posterior under a different prior in our model?
- What happens if we use a Bayesian robustness procedure, such as a power posterior?

In Chapter 3, we address these questions in depth. First, we show that under an *arbitrarily*

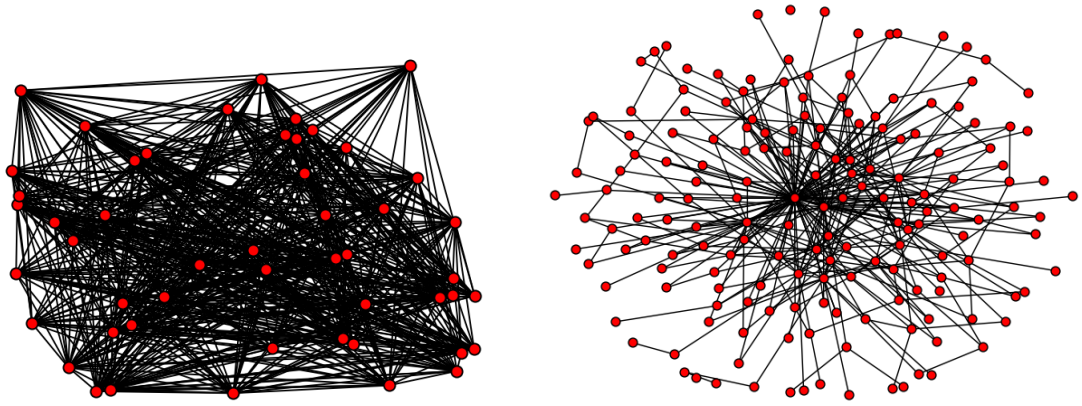
small amount of component misspecification, the posterior number of components *diverges*: i.e., the posterior probability of any particular finite number of components converges to 0 in the limit of infinite data. This result adds rigor to existing data-analysis folk wisdom, and contrary to intuition, posterior-density consistency is not sufficient to establish this result. In particular, this chapter develops novel sufficient conditions that are more realistic and easily checkable than those common in the asymptotics literature, and we introduce a new proof technique that is broadly useful beyond this application.

Finally, we explore the behavior of finite mixtures under a number of related models, such as a model with a prior that varies with the data or one where the prior has support on a finite number of components. We also begin to investigate robustness procedures for likelihood misspecification. One popular type of robustness procedure is the *power posterior*, where the likelihood is raised to a fixed power in $(0, 1)$. We show that in this case, the (power) posterior number of components still diverges, and we discuss generalizations to other robustness procedures. Throughout this chapter, we investigate posterior divergence empirically on a number of simulated data examples as well as in applications on cancer and mouse gene expression data and astronomy data of galaxy velocities.

1.3 Bayesian models under prior misspecification

Another type of misspecification in Bayesian models due to the specification of the prior. One way in which this misspecification manifests in nonparametric Bayesian models—such as models for modeling latent clusters and features (Campbell et al., 2018) in discrete data—is in terms of the resulting properties of the data generated under a model, such as scaling behavior with respect to the number of data points. In this dissertation, we consider misspecification in generative models for graphs and their scaling behavior (e.g., the number of edges in the graph as the number of vertices grows).

Graph-structured observations are ubiquitous in scientific domains such as biology, ecology, chemistry, physics, and epidemiology. One goal in developing generative models for graphs is to be able to generate graphs that can capture properties of real-world graphs, such as particular scaling



(a) Sample from a stochastic block model. (b) Sample from a vertex popularity model.

Figure 1.3.1: Example graphs generated under two different Bayesian models.

behaviors or particular network structure and symmetries. However, many popular models for graphs do not capture real-world scaling properties of interest. For instance, many Bayesian generative models for network data have been proposed that rely on the seemingly innocuous assumption of (vertex) exchangeability, in which the distribution of the graph is invariant to relabelings of the vertices. In particular, the Aldous–Hoover theorem implies that exchangeable graph models generate *dense* graphs with probability one (Orbanz and Roy, 2015); thus, such generative graph models are misspecified for *sparse* networks. Here, we say a graph is dense if the number of edges in the graph grows asymptotically as the square of the number of vertices in the graph. By contrast, we say a graph is sparse if the number of edges grows sub-quadratically as a function of the number of vertices in the graph. This disconnect between desired asymptotic behavior and model specifications motivates the development of new models that achieve sparsity rather than always generating dense graphs.

In Chapter 4, we consider an alternative modeling assumption, *edge exchangeability*, in which the distribution of a graph sequence is invariant to the order of the edges. We prove that edge-exchangeable models, unlike models that are traditionally vertex exchangeable, can exhibit sparsity. To do so, we introduce a class of *vertex popularity* generative models that are particularly convenient for inference (Campbell et al., 2018); Figure 1.3.1b visualizes a sample from one such generative

model. We verify our theoretical results through simulations and show that a particular class of models under this framework generates a range of sparse and dense scaling behaviors, in addition to power laws.

Finally, we note that the results in Chapter 4 complement a rich literature on nonparametric Bayesian models and their scaling behaviors in a number of other domains. For instance, Gnedin et al. (2007) have thoroughly characterized a wide variety of power laws that may be exhibited in clustering models and have, moreover, shown that many of these power laws are equivalent. One behavior of interest is a power law in the number of clusters as the number of data points grows. This behavior is roughly analogous to considering a power law in the number of edges as the number of vertices grows. But Gnedin et al. (2007) consider a much wider range of potential power laws. Likewise, Broderick et al. (2012) have enumerated a range of power laws for *feature allocations*, a generalization of clustering where each data point may belong to any non-negative integer number of groups—now called features instead of clusters.

Not only have previous authors studied power laws for clustering and feature allocations, but they have detailed particular, practical generative models for achieving these power laws—and these models typically lead to corresponding inference algorithms as well. For instance, the canonical power law model for clustering is the Pitman–Yor process (Goldwater et al., 2005; Pitman and Yor, 1997; Teh, 2006), and the canonical power-law model for feature allocations is the three-parameter beta process (Broderick et al., 2012; Teh and Görür, 2009). Inspired by known power laws in partitions (Gnedin et al., 2007) and feature allocations (Broderick et al., 2012), in Chapter 4, we develop new generative graph models that exhibit desirable asymptotic and power law behaviors.

1.4 Misspecification due to limited computation

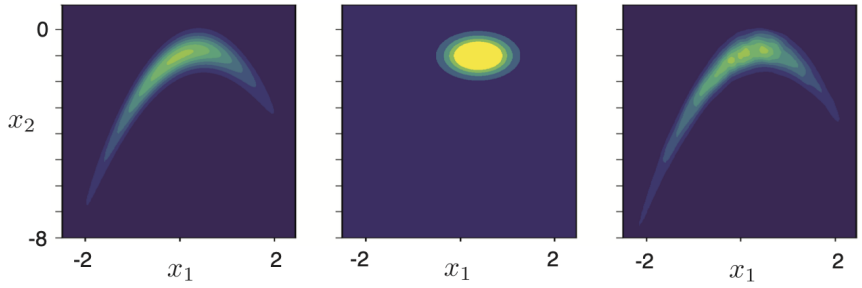
Under finite computational constraints and resources, the complexity of many modeling problems necessitates the use of approximate inference algorithms, such as MCMC or variational inference, along with other approximate algorithmic tools, such as numerical approximations or data summarization algorithms (Cai et al., 2018). Thus, due to limited computation, the data analyst is fitting

a misspecified model *by design*.

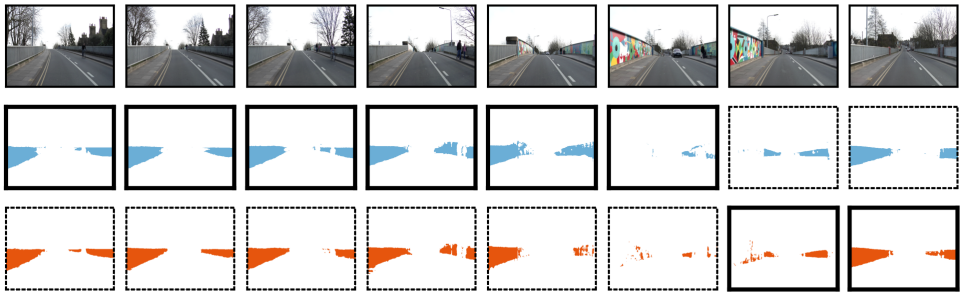
For instance, in variational inference, it is common to choose a simpler, tractable variational family as an approximation to the specified posterior. Many recent methods have been developed to facilitate variational inference in more expressive variational families; see Zoltowski et al. (2021) and references within for an overview. Another example occurs in many scientific applications, where expensive computations are needed to obtain accurate answers, e.g., a complex physical simulation or costly experiments (Gundersen et al., 2021; Jones et al., 2022, 2023). In these settings, it is common to use approximate numerical methods and low-fidelity simulations in order to make the problem tractable with existing inference algorithms. In this dissertation, we consider this problem in more depth, where the goal is to use Markov chain Monte Carlo (MCMC) for inference.

MCMC is widely used for uncertainty quantification, simulation, and optimization. A key challenge in applying MCMC to scientific domains is computation: the target density of interest is often a function of expensive computations, such as a high-fidelity physical simulation, an intractable integral, or an iterative algorithm. In these situations, MCMC quickly becomes impractical, as these expensive computations need to be evaluated at each iteration of the algorithm, and in practice, the target density uses a cheaper, low-fidelity computation, leading to bias in the resulting target density.

In Chapter 5, we propose a class of asymptotically exact *multi-fidelity MCMC algorithms* (Cai and Adams, 2022) for problems with a sequence of low-fidelity models available that approximate the expensive target density of interest arbitrarily well, e.g., a decreasing sequence of discretization sizes of a function. Using an auxiliary-variable strategy, the method uses a cheaper, randomized-fidelity unbiased estimator of the target fidelity constructed via random truncation of a telescoping series of the low-fidelity sequence of models. The key advantages of the approach are that it is easy to implement for complex MCMC algorithms, such as slice sampling and Hamiltonian Monte Carlo, and that Monte Carlo error is the only source of bias. Finally, we empirically demonstrate that the method can lead to more accurate estimates with less computation in applications such as Bayesian ODE system identification, PDE-constrained optimization, and Gaussian process modeling.



(a) An example of approximate posteriors obtained from variational inference. **Left:** Target distribution; **Middle:** Mean-field approximation; **Right:** A more expressive approximating family. Figure originally from Zoltowski, Cai, and Adams (2021).



(b) Example of high- and low-fidelity models (middle, bottom rows, respectively) of an object from video frames (top row) used to compute multi-fidelity posterior distributions for detecting changepoints in time series data. The low-fidelity model is less accurate but cheaper to compute, and the high-fidelity model is more accurate but more expensive to compute. Figure originally from Gundersen, Cai, Zhou, Engelhardt, and Adams (2021).

Figure 1.4.1: Misspecification by design due to limited computation.

1.5 Dissertation organization and relevant publications

We begin by presenting an overview of some foundations of probabilistic modeling that are relevant to later chapters in this dissertation (Chapter 2). Next we present several case studies of misspecification: in Chapter 3, we present a case study of likelihood misspecification in finite mixture models based on work; in Chapter 4, we present a case study of misspecification in graphs and their scaling behavior, and in Chapter 5, we present a case study of approximate Bayesian models under limited computation by proposing a general framework for MCMC in multi-fidelity models.

Several parts of this dissertation previously appeared in existing publications. Parts of Chapter 2

are modified versions of text that originally appeared in Cai and Adams (2022); Cai and Broderick (2015); Cai et al. (2014, 2016a,b, 2021). Chapter 3 builds on the results from Cai et al. (2020, 2021). Chapter 4 previously was published in Cai et al. (2016a), and earlier versions previously appeared in Cai and Broderick (2015) and Broderick and Cai (2015a,b). Chapter 5 is based on work from Cai and Adams (2022).

Chapter 2

Foundations of probabilistic modeling

In this chapter, we review several foundational ideas in probabilistic modeling that are referenced or applied in later chapters of this dissertation. First, we review key ideas and tools for the analysis of the posterior in the limit of infinite data (Section 2.1). Specifically, we provide an intuitive introduction to the idea of posterior consistency; then we formally define posterior consistency; and lastly, we review Schwartz's theorem, one of the key tools for establishing consistency of the posterior. In Section 2.2, we provide an overview on Bayesian generative models for graphs based on tools from Bayesian nonparametrics. We provide background on graphons, which are an equivalent characterization of the Aldous-Hoover theorem for exchangeable random graphs, and we also review the generalization to directed graphs. Finally, we discuss completely random measures as a building block for network modeling and tools for analyzing them. In Section 2.3, we review several classical sampling algorithms for performing approximate Bayesian inference.

2.1 Posterior consistency: an overview

In statistical estimation, one of the goals is to compute an estimator $\hat{\theta}_N$ of an unknown parameter θ. One of the most basic properties one might then check for is if the estimator is *consistent*: that is, as the number of data points N increases, does the estimator converge to the true parameter?

In a Bayesian analysis, one of the central goals is to compute a posterior distribution Π_N of a parameter of interest θ given N data points, and analogous ideas of consistency for the posterior have been formulated to serve as a check on a Bayesian analysis. In particular, *posterior consistency* is the idea that as the number of data observations N grows, the posterior distribution concentrates on neighborhoods of the true data parameter (or generating distribution).

In this section, we will make precise what it means for the posterior to "concentrate" and be consistent – here as the number of data points goes to infinity, the posterior mass on densities arbitrarily close to the true density will converge to 1. Then we will present Schwartz's theorem, one of the primary foundational tools for establishing posterior consistency.

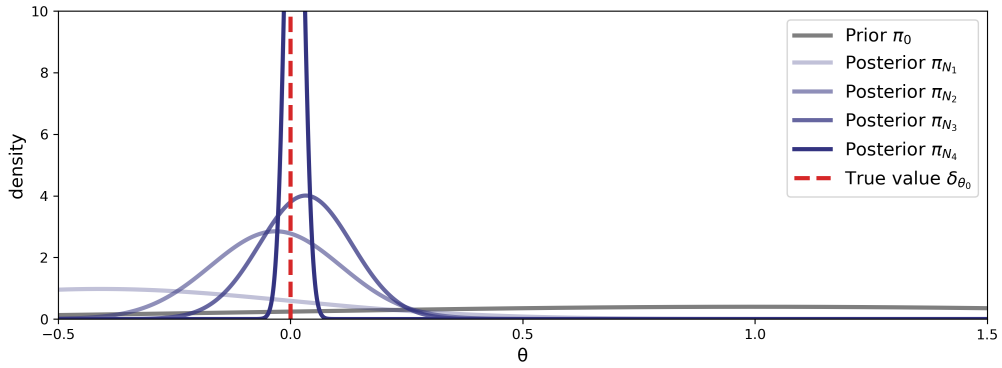
2.1.1 An informal picture of Bayesian consistency

Suppose the true data generating distribution of the data—which is unknown to the data analyst—is from a standard Gaussian with density $\mathcal{N}(\cdot \,|\, 0, 1)$. The data analyst decides to model the data as i.i.d. from the family

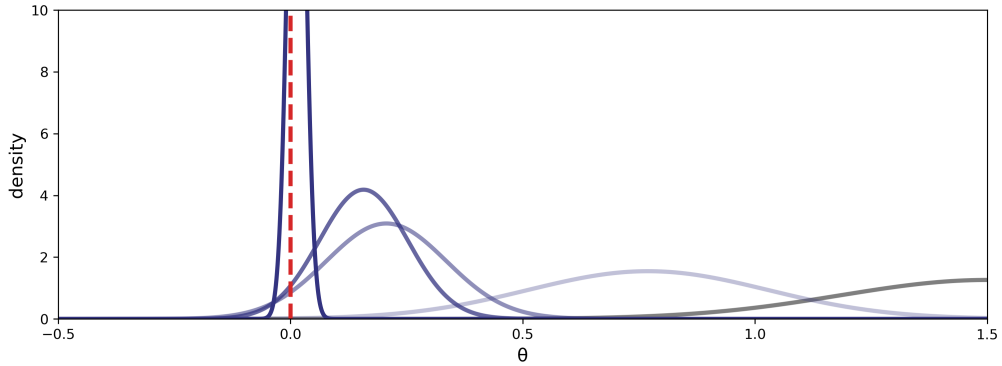
$$\mathcal{P} = \{\mathcal{N}(\cdot \,|\, \theta, 1) : \theta \in \mathbb{R}\}$$

and places a Gaussian prior on the unknown parameter θ. What happens to the posterior distribution as we get more and more data? The hope is that if one has access to an infinite data generator, the data analyst should be able to recover the "truth"; we will assume there is a true parameter θ_0 (and corresponding density p_0) that generated the data. *Posterior consistency*—which is when the posterior has the property of "concentrating" around the true generating value as the number of data points converges to infinity—formalizes this idea. Before defining posterior consistency, we first build intuition by simulating the asymptotic behavior of posteriors under this model.

Simulating consistency. We can simulate and visualize the posterior for increasing number of data set sizes for the example above (Figure 2.1.1). The model in the top panel assumes the prior $\pi_0 = \mathcal{N}(\cdot \,|\, 1.5, 1)$, and the model in the bottom panel uses $\pi_0 = \mathcal{N}(\cdot \,|\, 1.5, 0.1)$. Here each of the purple curves in Figure 2.1.1 represents a posterior density π_N computed with more data, where $N_1 = 5$, $N_2 = 50$, $N_3 = 100$, and $N_4 = 3000$ (darker curves represent larger N).



(a) Prior with more mass near true parameter



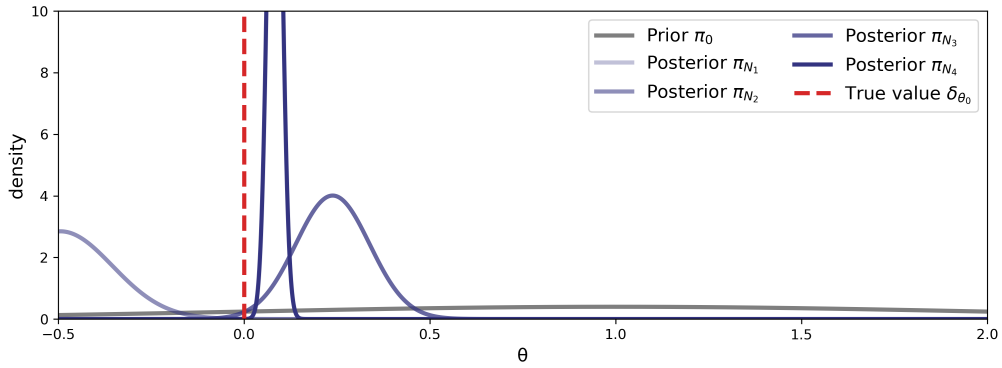
(b) Prior with less mass near true parameter

Figure 2.1.1: Simulations of a conjugate Gaussian posterior density computed using increasing numbers of data points generated from a standard Gaussian. The posterior mass becomes more concentrated around the true parameter value as more data are observed.

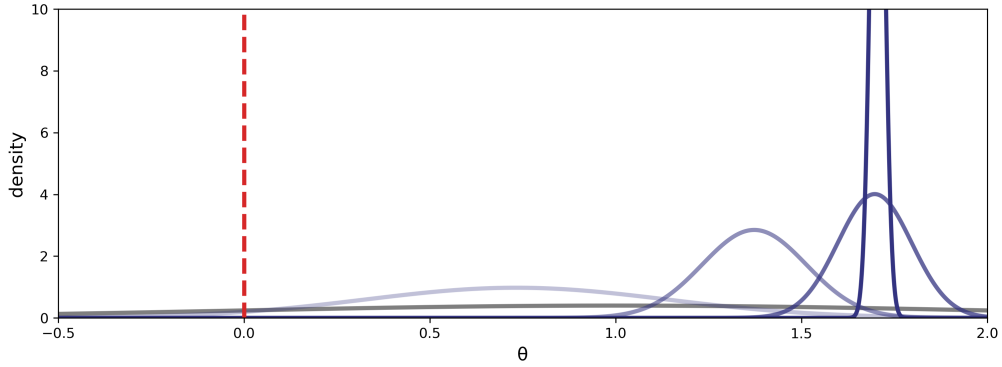
When there is no data, only the prior is present (gray curve). With increasing number of data points, the posterior mass around the true parameter value (indicated by the red dashed line) becomes more concentrated.

While the priors in these examples do not affect asymptotic property of consistency, it does affects how quickly the posteriors converge. Comparing the top and bottom panel of Figure 2.1.1, we observe that the top panel, which has more prior mass near the true parameter value, converges more quickly.

What happens if the posterior does not concentrate on at the true generating parameter? We show two examples in Figure 2.1.2, where the models are *misspecified*, i.e., $p_0 \notin \mathcal{P}$. In both examples,



(a) True density $p_0 = \mathcal{N}(\cdot \,|\, 0, 10)$.



(b) True density $p_0 = \mathcal{LN}(\cdot \,|\, 0, 1)$.

Figure 2.1.2: Examples of inconsistency of the posterior.

the posterior on θ does not concentrate on the true data generating θ_0.

Cartoon of more general consistency. In the previous example, we visualized the prior and posteriors in a finite-dimensional parameter space Θ. More generally, we can consider some true density p_0 that lives in the model space $\mathcal{P}$ and a prior on the model space. For instance in the example above, the prior on Θ induces a prior distribution Π_0 on the space $\mathcal{P} = \{N(\cdot \,|\, \theta, 1) : \theta \in \mathbb{R}\}$, and the true density $p_0 = N(\cdot \,|\, 0, 1)$ is an element of $\mathcal{P}$.

Figure 2.1.3 shows a cartoon schematic illustrating posterior concentration around p_0. Here the gray area represents the space of models (densities) given by $\mathcal{P}$, and the bulk of the posterior mass gets more tightly concentrated around the true density with more observations, as illustrated by the

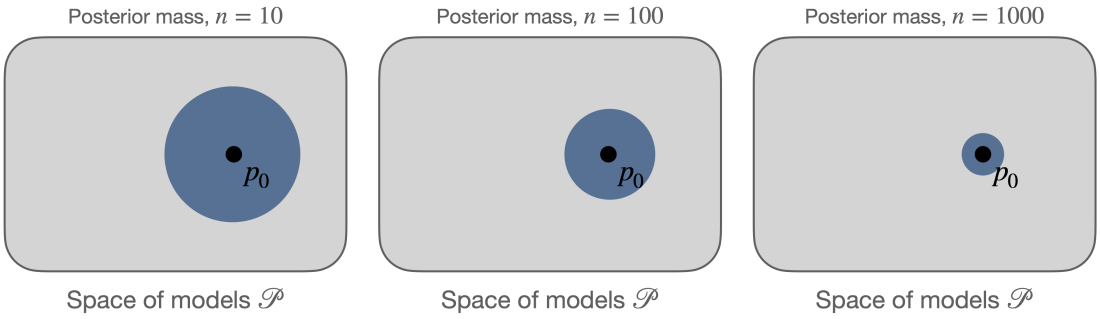


Figure 2.1.3: Cartoon schematic illustrating the posterior mass (blue circle) concentrating around the true density p_0 as the number of data points increases. The gray box incidates the space of possible models $\mathcal{P}$ on $\mathbb{X}$.

shrinking blue circles.

2.1.2 Preliminaries: the topology of weak convergence

In this dissertation, we will focus on convergence with respect to the weak topology. Throughout our discussion, we assume that $\mathbb{X}$ is a Polish space metrized by ρ. Let $\mathscr{P}(\mathbb{X})$ denote the space of probability measures on $\mathbb{X}$.

We begin by considering the Lévy-Prokhorov metric as one particular choice for ρ.

Definition 2.1.1. *Let $f, g \in \mathscr{P}(\mathbb{X})$. The Lévy-Prokhorov metric is defined as*

$$\rho(f, g) = \inf\{\epsilon > 0 : f(A) < g(A^\epsilon) + \epsilon, g(A) < f(A^\epsilon) + \epsilon\},$$

where $A^\epsilon := \{y : \rho(x, y) < \epsilon \text{ for some } x \in A\}$.

The Lévy-Prokhorov metric ρ is notable because it induces the *weak topology* on $\mathscr{P}(\mathbb{X})$.

Definition 2.1.2 (Weak convergence)**.** *A sequence of measures $f_i \in \mathscr{P}(\mathbb{X})$ converges weakly to $f \in \mathscr{P}(\mathbb{X})$ if $\rho(f_i, f) \to 0$, as $i \to \infty$. Let $f_i \Rightarrow f$ denote weak convergence.*

The Portmanteau theorem (Ghosal and van der Vaart (2017, Theorem A.2)) characterizes equivalent notions of weak convergence, and below we include the relevant portions of the Portmanteau

theorem used in later chapters.

Theorem 2.1.3 (Portmanteau (partial statement))**.** *The following statements are equivalent for any $f_i, f \in \mathscr{P}(\mathbb{X})$:*

1. $f_i \Rightarrow f$.

2. *For all bounded, uniformly continuous $h : \mathbb{X} \to \mathbb{R}$,*

$$\int h df_i \longrightarrow \int h df.$$

3. *For every closed subset C,*

$$\limsup_i f_i(C) \leq f(C).$$

Another useful result is Prokhorov's theorem (Ghosal and van der Vaart, 2017, Theorem A.4), which characterizes weakly compact subsets of $\mathscr{P}(\mathbb{X})$ in terms of a *tight* subset of measures.

Definition 2.1.4. *A subset $\Gamma \subseteq \mathscr{P}(\mathbb{X})$ is* tight *if for any $\epsilon > 0$, there exists a compact subset $K_\epsilon \subseteq \mathbb{X}$ such that for every $\psi \in \Gamma$, $\psi(K_\epsilon) \geq 1 - \epsilon$.*

Theorem 2.1.5 (Prokhorov)**.** *If $\mathbb{X}$ is a Polish space, then $\Gamma \subseteq \mathscr{P}(\mathbb{X})$ is relatively compact if and only if Γ is tight.*

2.1.3 The posterior distribution and consistency

We consider a class of models $\mathscr{M}(\mathbb{X}) \subseteq \mathscr{P}(\mathbb{X})$, and we assume $\mathscr{M}(\mathbb{X})$ is dominated by a σ-finite measure μ. In particular, define the model class of densities $\mathcal{P}$ by

$$\mathcal{P} = \left\{ \frac{df}{d\mu} : f \in \mathscr{M}(\mathbb{X}) \right\}.$$

Let Π be a prior distribution on $\mathcal{P}$, and consider the following generative model for the data

observations $X_{1:N} := X_1, \ldots, X_N$:

$$p \sim \Pi$$
$$X_1, \ldots, X_N \,|\, p \overset{i.i.d.}{\sim} p.$$

Under mild regularity conditions on the topology of $\mathcal{P}$[1], the posterior distribution exists. Using *Bayes's formula*, a version of the *posterior distribution* is given by

$$\Pi(A \,|\, X_{1:N}) = \frac{\int_A \prod_{n=1}^N p(X_n)\, d\Pi(p)}{\int_{\mathcal{P}} \prod_{n=1}^N p(X_n)\, d\Pi(p)}, \qquad A \subseteq \mathcal{P} \text{ measurable.} \tag{2.1}$$

Below, we provide a definition of posterior consistency with respect to general neighborhoods. Let $f_0^{(N)}$ denote the joint distribution with respect to N observations[2].

Definition 2.1.6. *The posterior distribution $\Pi(\cdot \,|\, X_{1:N})$ is consistent at $p_0 \in \mathcal{P}$ if for every neighborhood U of p_0, with $f_0^{(\infty)}$-probability 1,*

$$\Pi(U \,|\, X_{1:N}) \overset{N\to\infty}{\longrightarrow} 1. \tag{2.2}$$

For the rest of this discussion, we may overload the notation f_0 to denote the joint distribution with respect to N observations, and we use *a.s.* as a shorthand for *almost surely*.

Remark 2.1.7. *We note that there are several alternative definitions of posterior consistency. In some cases, the definition is stated in terms of convergence in probability (as opposed to a.s. convergence) or consistency defined explicitly with a specific topology in mind. In this chapter, we will assume the topology of weak convergence, and will refer to the above definition as* weak consistency*. However, we note that in the literature, the terminology "weak" and "strong" consistency have been used for both the type of convergence of the posterior (i.e., convergence in probability versus a.s. convergence (Ghosal and van der Vaart, 2017)) and to refer to the topology (i.e., weak topology vs strong topology).*

[1] A sufficient condition is that $\mathcal{P}$ is Polish.

[2] Define this quantity to be the i.i.d. product measure defined on $(\mathbb{X}^N, \mathcal{B}^N)$.

2.1.4 Schwartz's theorem for weak consistency

Schwartz's theorem (Schwartz, 1965) is one tool for establishing consistency at the true density with respect to the weak topology. The result is a posterior consistency theorem for the density, and thus relies on the assumption that the space of models $\mathcal{P}$ is dominated by a σ-finite measure μ.

Let $\mathrm{KL}(p; q)$ denote the Kullback-Leibler (KL) divergence between p and q.

Theorem 2.1.8 (Schwartz)**.** *Let Π be a prior on $\mathcal{P}$, and suppose that for any $\epsilon > 0$,*

$$\Pi(\{p \in \mathcal{P} : \mathrm{KL}(p_0; p) < \epsilon\}) > 0.$$

Then the posterior is weakly consistent *at p_0: i.e., for any weak neighborhood U of p_0 the sequence of posterior distributions satisfies*

$$\Pi(U \mid X_{1:N}) \overset{N\to\infty}{\longrightarrow} 1, \quad f_0\text{-a.s.} \tag{2.3}$$

Theorem 2.1.8 is remarkable in that one only needs to verify that p_0 is in the KL-support of the prior Π. Intuitively, the KL-support condition guarantees that the prior distribution places positive mass on models arbitrarily close to p_0. When more information is known about how much mass the prior puts on models near p_0, stronger results in terms of contraction rates can be derived. For instance, in Figure 2.1.1, we consider the same Gaussian family as before but under different priors. The model in Figure 2.1.1a, which has prior more mass near the true value, concentrates faster than the model in Figure 2.1.1b with respect to the number of data points N.

The above result assumes that the prior Π is fixed. Note that weak consistency also holds (with f_0-probability) for priors that vary with N, provided that the sequence Π_N satisfies the KL-support condition stated in Definition 3.5.1 (Ghosal and van der Vaart, 2017, Theorem 6.25).

2.1.5 A general version of Schwartz's theorem

Theorem 2.1.8 is a result with respect to the weak topology. More general versions of the theorem with respect to general neighborhoods have also been proven. Below, we discuss one such version

(Ghosh and Ramamoorthi, 2003).

Theorem 2.1.9 (Schwartz (general))**.** *Let* Π *be a prior on* $\mathcal{P}$*, and suppose the following two conditions hold:*

1. *p_0 is in the KL-support of the prior: for any $\epsilon > 0$, $\Pi(\{p \in \mathcal{P} : \mathrm{KL}(p_0; p) < \epsilon\}) > 0$.*
2. *Existence of a uniformly consistent sequence of tests: there exist test functions $\phi_N : \mathbb{X}^N \to [0, 1]$ such that*

$$f_0^{(N)}(\phi_N) := \int \phi_N \, df_0^{(N)} \overset{N\to\infty}{\longrightarrow} 0,$$
$$\sup_{p \in U^c} f^{(N)}(1 - \phi_N) := \sup_{p \in U^c} \int (1 - \phi_N) \, df^{(N)} \overset{N\to\infty}{\longrightarrow} 0.$$

Then the posterior is consistent at p_0: i.e., for any neighborhood U of p_0 the sequence of posterior distributions satisfies

$$\Pi(U \mid X_{1:N}) \overset{N\to\infty}{\longrightarrow} 1, \quad f_0\text{-a.s.} \tag{2.4}$$

Theorem 2.1.9 has a new condition in addition to the KL-support condition in Theorem 2.1.8: i.e., there exists a uniformly consistent sequence of tests. Intuitively, the second condition of Theorem 2.1.9 is that "the hypothesis $H_0 : p = p_0$ should be testable against complements of neighborhoods of p_0, i.e., $H_1 : p \in U^c$" (Ghosh and Ramamoorthi, 2003). The interpretation of the test ϕ_N is that the null hypothesis $H_0 : p = p_0$ is rejected with probability ϕ_N. Given the joint distribution $f^{(N)}$, then the interpretation of $f^{(N)}\phi_N$ is the probability of rejecting H_0 when the data are sampled from P.

Note that due to the difficulty in directly verifying the uniformly consistent test condition, these conditions themselves are often not directly checked. Some equivalent conditions are given in Ghosh and Ramamoorthi (2003, Proposition 4.4.1).

As we will see in the next section, the testing condition of Schwartz's theorem is satisfied by weak neighborhoods. But for strong consistency with respect to the L1 metric, uniformly consistent

tests do not exist (Barron, 1989; LeCam, 1973). However, it is still possible to have strong posterior consistency hold via an extension of Schwartz's theorem; Barron (1988) proved a early result on this, and Ghosal and van der Vaart (2017, Theorem 6.17) presents an extended Schwartz theorem that can be used to recover the classical Schwartz theorem as well as other consistency results based on metric entropy. In addition, Schwartz's theorem has been extended to other related models, such as priors that vary with the data, misspecified models, and non-i.i.d. models.

2.1.6 The weak topology satisfies the Schwartz testing condition

Three situations in which a uniformly consistent sequence of tests exist are (1) under the weak topology, (2) when the sample space is countable, and (3) when the parameter space of the model is finite-dimensional (Ghosal and van der Vaart, 2017, Chapter 6). We now discuss the first example, which recovers Theorem 2.1.8: under the weak topology, the tests needed for the Schwartz condition exist. Thus, provided the KL-support condition holds, the posterior is consistent.

To establish weak posterior consistency, we need to show that for any weak neighborhood U of p_0,

$$\Pi(U^c|X_{1:N}) \overset{N\to\infty}{\longrightarrow} 0, \qquad f_0^{(\infty)}\text{-a.s.}, \tag{2.5}$$

and so we will verify the testing condition with respect to the complement U^c. To do so, we will verify the condition for a simpler set, which turns out to be sufficient.

A μ-weak neighborhood U of p_0 is a subset of $\mathcal{P}$ that can be represented by arbitrary unions of basis elements, given by subsets of the form

$$V = \left\{ p \in \mathcal{P} : \forall i \in \{1, \ldots, k\}, \left| \int g_i p d\mu - \int g_i p_0 d\mu \right| < \epsilon_i \right\}, \tag{2.6}$$

where for all i, g_i is a bounded, real-valued, continuous function on $\mathbb{X}$, $\epsilon_i > 0$, and $k \in \mathbb{N}$.

Note that each basis element V can be expressed as a finite intersection of subbasis elements, or

subsets of the form

$$V_i = \left\{ p \in \mathcal{P} : \left| \int g_i p d\mu - \int g_i p_0 d\mu \right| < \epsilon_i \right\}, \tag{2.7}$$

which is an intersection between two sets of the form: $A_i = \left\{ p \in \mathcal{P} : \int g_i p d\mu < \int g_i p_0 d\mu + \epsilon_i \right\}$.

Thus, the sets of the form A_i also form a subbasis for V, and the complement $V^c = \bigcup_{i=1}^{2k} A_i^c$ is then a finite union of the complements of the subbasis elements above.

Since $V \subset U$ by construction, we can decompose the posterior probability of the complement of the weak neighborhood into a sum of probabilities over the complements of the subbasis sets:

$$\Pi(U^c | X_{1:N}) \leq \Pi(V^c | X_{1:N}) \leq \sum_{i=1}^{2k} \Pi(A_i^c | X_{1:N}). \tag{2.8}$$

Thus, in order to prove that $\Pi(U^c | X_{1:N}) \to 0, (f_0^{(\infty)}$-a.s.) as $n \to \infty$, it suffices to show that the probability of the subbasis set complements vanish almost surely, i.e., for all i, $\Pi(A_i^c | X_{1:N}) \to 0$ a.s.

We will now show the Schwartz testing condition holds for sets of the form

$$A = \left\{ p \in \mathcal{P} : \int g(x)\, p(x)\, d\mu(x) < \int g(x)\, p_0(x)\, d\mu(x) + \epsilon \right\}. \tag{2.9}$$

Consider the test function

$$\phi_n(X_1, \ldots, X_n) = I \left\{ \frac{1}{n} \sum_{i=1}^{n} g(X_i) > \int g(x)\, p_0(x)\, d\mu(x) + \epsilon/2 \right\}. \tag{2.10}$$

Then the expectation of this test function can be bounded via Hoeffding's inequality: that is, since g is a bounded function (and without loss of generality, can be rescaled such that $0 \leq g \leq 1$),

$$f_0^{(n)} \phi_N = f_0^{(N)} \left(\frac{1}{N} \sum_{i=1}^{N} g(X_i) > \int g(x)\, p_0(x)\, d\mu(x) + \epsilon/2 \right) \leq e^{-N\epsilon^2/2}. \tag{2.11}$$

Similarly, another application of Hoeffding's inequality along with the property that for any

$p \in A^c$, $\int g(x)p(x)d\mu(x) - \int g(x)p_0(x)d\mu(x) > \epsilon$ implies that

$$f^{(n)}(1-\phi_N) \leq f^{(N)}\left(-\frac{1}{N}\sum_{i=1}^{n} g(X_i) > -\int g(x)\,p(x)\,d\mu(x) + \epsilon/2\right) \leq e^{-N\epsilon^2/2}. \tag{2.12}$$

2.2 Bayesian nonparametrics and random graphs models

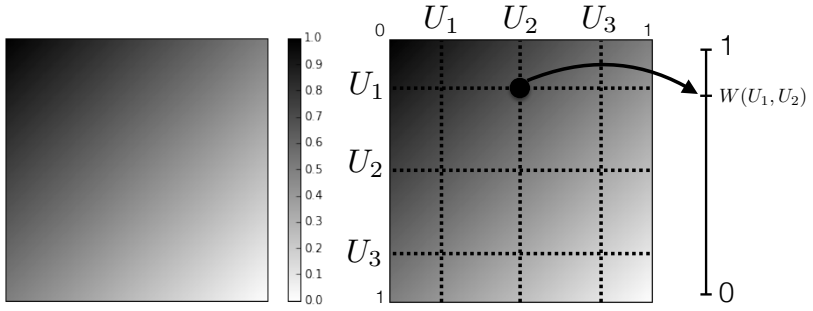
In recent years, network data has increased in both ubiquity and size. As network data appear in a growing number of applications—such as online social networks, biological networks, and networks representing communication patterns—there is growing interest in developing models and inference for such data and studying its properties. In Chapter 4, we develop new random graph models and study properties of random graphs generated under these models, and we review several foundational concepts for that chapter in this section.

First, we briefly review exchangeable random graphs and the *Aldous-Hoover theorem*; our perspective focuses on the equivalent characterization in terms of *graphons* due to Lovász and Szegedy (2006). We also discuss exchangeable random *directed graphs* and the directed analog of the Aldous-Hoover theorem, known also as *digraphons*, as a basis for generative graph models. A consequence of the Aldous-Hoover theorem is that generative models for graphs with this assumption will generate graphs that are dense or empty with probability 1.

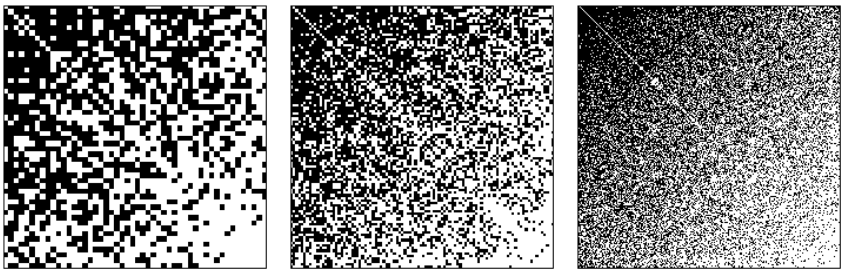
In the remainder of this section, we provide an overview on completely random measures for building generative models, which have been used extensively in models for clustering and feature allocations. Finally, we present several tools for analyzing models using completely random measures. In Chapter 4, we will use these tools to build and analyze new generative models as part of an alternative modeling framework that is able to generate both sparse and dense graphs.

2.2.1 Exchangeable random graphs: undirected and directed graphs

We define a *graph on* $[n]$ to be a graph with set of vertices $[n] := \{1, \ldots, n\}$; its adjacency matrix is the $\{0,1\}$-valued $n \times n$ matrix $(G_{ij})_{i,j\in[n]}$, where $G_{ij} = 1$ iff G has an edge between vertices i and j. Graphs on $\mathbb{N}$, and their adjacency matrices, are defined similarly. We write $x \stackrel{d}{=} y$ when two random



(a) **left**: An example of a graphon, given by the function $W(x, y) = \frac{(1-x)+(1-y)}{2}$; **right**: Schematic of sampling procedure.



(b) Samples from the finite random graphs of size 50, 100, and 500 shown as "pixel pictures" of the adjacency matrix, where black corresponds to 1 and white to 0. The samples have been resorted by increasing order of the sampled uniform random variables U_i.

Figure 2.2.1: Visualization of a graphon and the sampling procedure.

variables x and y are equal in distribution, and abbreviate *almost surely* and *almost everywhere* by a.s. and a.e., respectively.

Many popular Bayesian models for graphs assume *exchangeability*, i.e., that the joint distribution of the edges is invariant under permutations of the vertices.

Definition 2.2.1. *A random array* $(G_{ij})_{i,j\in\mathbb{N}}$ *is (jointly)* exchangeable *when*

$$(G_{ij}) \stackrel{d}{=} (G_{\sigma(i),\sigma(j)}) \tag{2.13}$$

for every permutation σ *of* $\mathbb{N}$.

By the Kolmogorov extension theorem, it is equivalent to ask for the above condition to hold only for those permutations σ that move a finite number of elements of $\mathbb{N}$.

We now define a sampling procedure that produces exchangeable graphs. In particular, foundational Aldous-Hoover theorem (Aldous, 1981; Hoover, 1979) characterizes undirected exchangeable graphs in terms of measurable functions, which can be equivalently characterized in terms of *graphons* (Lovász and Szegedy, 2006).

Definition 2.2.2. *A graphon is a symmetric, measurable function* $W\colon [0,1]^2 \to [0,1]$.

Given a graphon W, there is an associated countably infinite exchangeable graph $\mathbb{G}(\mathbb{N}, W)$ with random adjacency matrix $(G_{ij})_{i,j\in\mathbb{N}}$ defined as follows (see Figure 2.2.1):

$$\begin{aligned} U_i &\overset{\text{iid}}{\sim} \text{Uniform}[0,1] \text{ for } i \in \mathbb{N}, \\ G_{ij} \,|\, U_i, U_j &\overset{\text{ind}}{\sim} \text{Bernoulli}(W(U_i, U_j)), \text{ for } i < j, \end{aligned} \tag{2.14}$$

and set $G_{ji} = G_{ij}$ for $i < j$, and $G_{ii} = 0$.

Figure 2.2.1 shows an example of sampling from the graphon $W(x,y) = \frac{(1-x)+(1-y)}{2}$. As the number of samples increases, the limiting structure of W becomes more clear in the resorted samples (bottom row).

Every exchangeable undirected graph can be written as a mixture of such sampling procedures. For $n \in \mathbb{N}$, we write $\mathbb{G}(n, W)$ to denote the finite random undirected graph on underlying set $\{1, \ldots, n\}$ induced by this sampling procedure.

Theorem 2.2.3 (Aldous (1981); Hoover (1979))**.** *Suppose G is an exchangeable graph on $\mathbb{N}$. Then G can be written as the mixture of W-random graphs $\mathbb{G}(\mathbb{N}, W)$ for some probability measure on graphons W.*

The Aldous-Hoover representation has since been extended to higher dimensions, more general spaces of random variables, and weaker notions of symmetry; for a detailed presentation, see Kallenberg (2005).

Since every exchangeable graph is a mixture of graphon sampling procedures, many network models can be described in this way. The stochastic block model (Holland et al., 1983) is such an example; it plays a special role as one of the simplest models that can approximate arbitrary graphon

sampling procedures. Some Bayesian nonparametric models, including the eigenmodel (Hoff, 2007), Mondrian process graph model (Roy and Teh, 2008), and random function model (Lloyd et al., 2012) were built knowing the Aldous-Hoover representation. Furthermore, many other such models are naturally expressed in terms of a distribution on graphons (Lloyd et al., 2012; Orbanz and Roy, 2015), including the infinite relational model (IRM) (Kemp et al., 2006) the latent feature relational model (LFRM) (Miller et al., 2009), and the infinite latent attribute model (ILA) (Palla et al., 2012).

We now consider the analogous ideas for directed graphs. In general, for a directed graph, (G_{ij}) may be asymmetric, and we allow self-loops, which correspond to values $G_{ii} = 1$ on the diagonal.

Definition 2.2.4. *A random (infinite) directed graph G on $\mathbb{N}$ is* exchangeable *if its joint distribution is invariant under all permutations π of the vertices:*

$$(G_{ij})_{i,j\in\mathbb{N}} \stackrel{d}{=} (G_{\pi(i)\pi(j)})_{i,j\in\mathbb{N}}. \tag{2.15}$$

Such an array (G_{ij}) is sometimes called *jointly exchangeable*. The case where the distribution is preserved under permutation of each index separately, i.e., where $(G_{ij}) \stackrel{d}{=} (G_{\pi(i)\sigma(j)})$ for arbitrary permutations π and σ, is called *separately exchangeable*, and arises for adjacency matrices of bipartite graphs.

As described by Diaconis and Janson (2008), using the Aldous-Hoover theorem one may show that every exchangeable countably infinite directed graph is expressible as a mixture of $\mathbb{G}(\mathbb{N}, \mathbf{W})$ with respect to some distribution on digraphons $\mathbf{W}$.

Definition 2.2.5. *A* digraphon *is a 5-tuple $\mathbf{W} := (W_{00}, W_{01}, W_{10}, W_{11}, w)$, where $W_{ab}\colon [0,1]^2 \to [0,1]$, for $a, b \in \{0,1\}$, and $w\colon [0,1] \to \{0,1\}$ are measurable functions satisfying the following conditions for all $x, y \in [0,1]$:*

$$
\begin{aligned}
W_{00}(x,y) &= W_{00}(y,x);\\
W_{11}(x,y) &= W_{11}(y,x);\\
W_{01}(x,y) &= W_{10}(y,x);\\
W_{00}(x,y) + W_{01}(x,y) + W_{10}(x,y) + W_{11}(x,y) &= 1.
\end{aligned}
\tag{2.16}
$$

Given a digraphon $\mathbf{W}$*, write* $\mathbf{W}_4$ *for the map* $[0,1]^2 \to [0,1]^4$ *given by* $(W_{00}, W_{01}, W_{10}, W_{11})$.

The functions W_{ab} represent the joint probability of $G_{ij} = a$ and $G_{ji} = b$ for $a, b \in \{0,1\}$, i.e.,

$$
\Pr(G_{ij} = a, G_{ji} = b) = W_{ab}(U_i, U_j), \tag{2.17}
$$

conditioned on U_i and U_j. In this way, W_{00} determines the probability of having neither edge direction between vertices i and j, W_{01} of only having a single edge to j from i ("right-to-left"), W_{10} of a single edge from i to j ("left-to-right"), and W_{11} of directed edges in both directions between i to j. The function w represents the probability of G_{ii}; because it is $\{0,1\}$-valued, this merely states whether or not i has a self-loop.[3]

The adjacency matrix $(G_{ij})_{i,j\in\mathbb{N}}$ of a countably infinite random graph $\mathbb{G}(\mathbb{N}, \mathbf{W})$ is determined by the following sampling procedure:

1. Draw $U_i \overset{\text{iid}}{\sim} \text{Uniform}[0,1]$ for $i \in \mathbb{N}$.
2. For each pair of distinct vertices i, j, assign the edge values for G_{ij} and G_{ji} according to an independent Categorical($\mathbf{W}_4(U_i, U_j)$) such that Equation (2.17) holds.
3. Assign self-loops $G_{ii} = w(U_i)$ for all i.

In other words, in step 2 we assign $(G_{ij}, G_{ji}) \,|\, U_i, U_j \overset{\text{ind}}{\sim} \text{Categorical}(\mathbf{W}_4(U_i, U_j))$, where we interpret the categorical random variable as a distribution over the choices $(0,0), (0,1), (1,0), (1,1)$, in that order. Note that step 2 is well-defined by the symmetry condition in Equation (2.16). Figure 2.2.2

[3]There is an equivalent alternative set of objects that may be used to specify an exchangeable digraph, where $W_{00}, W_{01}, W_{10}, W_{11}$ are as before and $p \in [0,1]$ gives the marginal probability of a self-loop, which is independent of the other edges; see Diaconis and Janson (2008) for details.

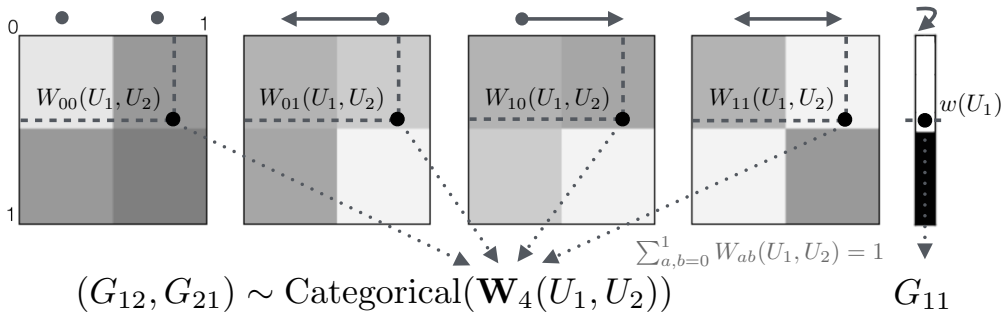


Figure 2.2.2: Schematic illustrating digraphon sampling procedure for $\mathbf{W} = (W_{00}, W_{01}, W_{10}, W_{11}, w)$.

illustrates this sampling procedure via a schematic. The x-axis is vertical and y-axis horizontal, with $(0, 0)$ in the upper left, so that the notation $W_{ab}(x, y)$ coheres with the usual (row, column) convention for matrix indexing.

An analogous sampling procedure yields *finite* random digraphs: Given $n \in \mathbb{N}$, in step 1, instead sample only U_i for $i \in [n]$. Then determine G_{ij} for $i, j \in [n]$ as before. We write $\mathbb{G}(n, \mathbf{W})$ to denote the random digraph thereby induced on $[n]$.

Diaconis and Janson (2008) derived the following corollary of the Aldous-Hoover theorem for directed graphs.

Theorem 2.2.6 (Diaconis–Janson)**.** *Every exchangeable random countably infinite directed graph is obtained as a mixture of* $\mathbb{G}(\mathbb{N}, \mathbf{W})$*; in other words, as* $\mathbb{G}(\mathbb{N}, \mathbf{W})$ *for some* random *digraphon* $\mathbf{W}$*.*

Therefore the problem of specifying the distribution of an infinite exchangeable digraph may be equivalently viewed as the problem of specifying a distribution on digraphons.

Most work involving priors on exchangeable graphs has focused on undirected graphs. For directed graphs, much of the work has extended the undirected case by using a single *asymmetric* measurable function $W_{\text{asym}} \colon [0, 1]^2 \to [0, 1]$ to model the directed graph (see Orbanz and Roy (2015, §4) for a survey of such models). While such an asymmetric function is appropriate for exchangeable bipartite graphs (Diaconis and Janson, 2008), this representation cannot express all exchangeable directed graph models (see Cai et al. (2016b) for a discussion).

The above theorem implies that exchangeable directed graphs are determined by specifying

a distribution on digraphons. Indeed, a digraphon is a more complicated representation for exchangeable directed graphs than a single asymmetric measurable function; a digraphon describes the possible directed edges between each pair of vertices *jointly*, rather than independently. In Cai et al. (2016b), random digraphons are considered as a framework for Bayesian generative models for directed graphs for a number of random structures such as directed acyclic graphs and tournaments.

2.2.2 Nonparametric Bayes and completely random measures

In Bayesian nonparametric modeling, we specify a random measure for the prior and likelihood, and then compute a posterior measure. The popular Dirichlet process is an example of a random (discrete probability) measure. The Dirichlet process is an example of normalized random measure that can be obtained from normalizing a completely random measure, which we will now define.

Let (Ω, Σ) be a measurable space with $\Omega := \mathcal{V} \times \mathbb{R}_+$ and $\Sigma := \Sigma_{\mathcal{V}} \times \mathcal{B}(\mathbb{R}_+)$. In the context of network modeling, $\mathcal{V}$ represents the space of vertices. A *random measure* W on (Ω, Σ) is a random measure-valued element such that $W(A)$ is a random variable for any measurable set $A \in \Sigma$.

Completely random measures provide an option for generating a countably infinite number of atoms in our random measure W (Kingman, 1993).

Definition 2.2.7. *A completely random measure W on (Ω, Σ) is a random measure with the additional requirement such that for any finite, disjoint measurable sets $A_1, \ldots, A_n \in \Sigma$, the random variables $W(A_1), \ldots, W(A_n)$ are (pairwise) independent.*

Completely random measures can be constructed from a Poisson point process with rate measure $\nu(d\theta, dw)$ in the following way: if we draw a sample $(\theta_i, w_i)_{i \in \mathbb{N}}$ from a Poisson point process, we construct W as follows:

$$W = \sum_{i=1}^{\infty} w_i \delta_{\theta_i}.$$

All completely random measures can be obtained in this way (along with a deterministic component and a fixed atomic component) (Kingman, 1993).

Popular CRMs include the beta process, the gamma process, the Bernoulli process, and the negative binomial process. A *beta process* $B \sim \mathrm{BP}(c, B_0)$ is a positive CRM whose rate measure depends on a concentration $c > 0$ and the base measure B_0 on $\mathcal{V}$, which is fixed. The parameter $\gamma_0 = B_0(\mathcal{V})$ is called the mass parameter. If B_0 is continuous, the rate measure is

$$\nu(dv, dw) = cw^{-1}(1-w)^{c-1} dw\, B_0(dv) \tag{2.18}$$

on $\mathcal{V} \times [0,1]$. A draw $B \sim \mathrm{BP}(\theta, B_0)$ is a discrete random measure

$$B = \sum_{i=1}^{\infty} w_i \delta_{v_i}$$

as constructed from a Poisson point process in the standard way described above. Here $w_i \in [0,1]$, and $\sum_i w_i < \infty$ whenever B_0 is finite. If B_0 is discrete, i.e., $B_0 = \sum_i u_i \delta_{v_i}$, with $u_i \in [0,1]$, then $B = \sum_i w_i \delta_{v_i}$ has atoms at the same locations as B_0, and the distribution of the weights are

$$w_i \sim \mathrm{Beta}(\theta u_i, \theta(1-u_i)).$$

To generate from a beta process, a stick-breaking process can be used:

$$\begin{aligned} W &= \sum_{i=1}^{\infty} \sum_{j=1}^{C_i} V_{i,j}^{(i)} \prod_{l=1}^{i-1} (1 - V_{i,j}^{(\ell)})\, \delta_{\psi_{i,j}}, \\ C_i &\overset{\text{iid}}{\sim} \mathrm{Pois}(\gamma_0), \\ V_{i,j}^{(\ell)} &\overset{\text{iid}}{\sim} \mathrm{beta}(1, \gamma_0), \\ \psi_{i,j} &\overset{\text{iid}}{\sim} \frac{1}{\gamma_0} B_0. \end{aligned}$$

Let B be a discrete measure on Ω. A *Bernoulli process*, denoted by $Y \sim \mathrm{BeP}(B)$, is characterized by the rate measure

$$\mu(dw, dv) = \delta_1(dw) B(dv).$$

Then a draw from a Bernoulli process with hazard measure B is the discrete random measure defined by $Y = \sum_i y_i \delta_{v_i}$, where $y_i = 1$ with Bernoulli probability w_i. See Thibaux and Jordan (2007) for additional details about the beta and Bernoulli processes.

In Chapter 4, we will consider a generalization of the beta process described above called the *three-parameter beta process*, or alternatively, the stable beta process.

2.2.3 Tools for analyzing models with completely random measures

We present two useful theorems for analyzing expectations involving random sums of functions of points from Poisson point processes. In Chapter 4, we will apply these theorems repeatedly to get expectations of graph quantities that will be used to analyze the asymptotic scaling behavior of particular generative models for graphs constructed using Poisson point processes and completely random measures.

The first theorem is Campbell's theorem, which is used to compute the moments of functionals of a Poisson process. For additional details, we refer to Kingman (1993, Sec. 3.2) for details.

Theorem 2.2.8 (Campbell's theorem)**.** *Let Π be a Poisson point process on S with rate measure ν, and let $f : S \to \mathbb{R}$ be measurable. If $\int_S \min(|f(x)|, 1)\, \nu(dx) < \infty$, then*

$$\mathbb{E}\left(\exp\left(c \sum_{x \in \Pi} f(x)\right)\right) = \exp\left(\int_S (\exp(cf(x)) - 1)\, \nu(dx)\right)$$

for any $c \in \mathbb{C}$, and furthermore,

$$\mathbb{E}\left(\sum_{x \in \Pi} f(x)\right) = \int_S f(x)\, \nu(dx).$$

The second theorem is a specific form of the Slivnyak-Mecke theorem, which is useful for computing the expected sum of a function of each point $x \in \Pi$ and $\Pi \setminus \{x\}$ over all points in a Poisson point process Π. If each point in Π is thought of as relating to a particular vertex in a graph, the Slivnyak-Mecke theorem allows us to take expectations of the sum (over all possible vertices in

the graph) of a function of each vertex and all its possible edges. For example, it is used below to compute the expected number of active vertices by taking the expected sum of vertices that have nonzero degree.

Theorem 2.2.9 (Slivnyak-Mecke theorem)**.** *Let* Π *be a Poisson point process on* S *with rate measure* ν*, and let* $f : S \times \Omega \to \mathbb{R}_+$ *be measurable. Then*

$$\mathbb{E}\left(\sum_{x\in\Pi} f(x, \Pi \setminus \{x\})\right) = \int_S \mathbb{E}\left(f(x, \Pi)\right) \nu(dx).$$

For additional details, we refer to Daley and Vere-Jones (2008, Prop. 13.1.VII) and Baddeley et al. (2007, Thm. 3.1,Thm. 3.2).

2.3 Bayesian inference and sampling algorithms

In the previous sections, we discussed constructing and analyzing particular Bayesian models. To compute the Bayesian posterior, approximate inference algorithms are often needed. In this section, we review several *Markov chain Monte Carlo* (MCMC) algorithms that are referenced in later chapters of this dissertation. We focus on reviewing the algorithms instead of the theory behind MCMC.

In MCMC, the goal is to sample from a distribution π by constructing a Markov chain with target density $\pi(\theta)$. We will assume that we may only be able to evaluate π up to a normalizing constant. While in this section, we will consider a general distribution π, we are often interested in applying MCMC in the specific setting of Bayesian inference, where π is the posterior:

$$\pi(\theta \,|\, X_{1:N}) \propto \pi_0(\theta) L_\theta(X_{1:N}), \tag{2.19}$$

where π_0 is a prior density and L_θ is a likelihood function.

Algorithm 2.3.1 Metropolis-Hastings iteration

Input: Current state θ, target density π, proposal distribution q
Propose new state $\theta' \sim q$
Compute acceptance ratio

$$R = \min\left(1, \frac{\pi(\theta')\, q(\theta|\theta')}{\pi(\theta)\, q(\theta'|\theta)}\right).$$

if rand $< R$ **then**
    Accept proposal θ'
else
    Reject proposal and set $\theta' = \theta$
end if
Output: New state θ'

2.3.1 Metropolis-Hastings and challenges

We begin by reviewing the classical *Metropolis-Hastings (M-H) algorithm* (Algorithm 2.3.1). In the Metropolis-Hastings (M-H) algorithm, a new state θ' is proposed according to a distribution q: that is, draw $\theta' \sim q(\cdot \,|\, \theta)$, where θ is the current state. The proposed state is then accepted with probability

$$\min\left(1, \frac{\pi(\theta')\, q(\theta|\theta')}{\pi(\theta)\, q(\theta'|\theta)}\right).$$

If the proposal is rejected, the new state is set to the old value. The *Gibbs sampling algorithm*—in which states are sampled in turn from their complete conditional distributions—is a special case of the M-H algorithm.

While the M-H algorithm is simple to describe and to apply to a problem, there are several reasons it can be difficult to use reliably in practice. The first challenge is choosing a suitable proposal distribution q, which often requires extensive tuning. For instance, a common choice is a proposal distribution of the form $q(\theta'|\theta) = \mathcal{N}(\theta'|\theta, \tau)$, where the parameter τ is tuned carefully. In Section 2.3.2, we describe *slice sampling*, an MCMC algorithm that does not have any tuning parameters.

A second difficulty is that M-H may have difficulties converging in high dimensional problems.

Many specialized MCMC algorithms have been proposed to to improve convergence by using other available information. For instance, if the gradient of the unnormalized target density is available, then *Hamiltonian Monte Carlo (HMC) algorithm* is a popular choice. In Section 2.3.3, we discuss one particular algorithm that is useful for sampling problems with special latent Gaussian structure.

A third challenge occurs when $\pi(\theta)$ (or its unnormalized density) itself is expensive or intractable to evaluate pointwise, this algorithm quickly becomes impractical. For example, if the likelihood is a function of an intractable integral, the likelihood needs to be approximated numerically. In Section 2.3.4, we describe a two-stage M-H algorithm that uses a low-fidelity model as a low pass filter for early rejection. In Chapter 5, we describe an alternative algorithm that uses randomized fidelity models that can be applied to more general MCMC algorithms.

2.3.2 Slice sampling

Slice sampling (Neal, 2003) is auxiliary-variable algorithm that automatically generates proposals without the need for an explicit accept/reject step. As we discuss above, slice sampling does not have any tuning parameters. Below, we describe the slice sampling algorithm in 1 dimension; one iteration is detailed in Algorithm 2.3.2.

Given the current state θ, we first sample a uniform height $u' \sim \text{unif}(0, \pi(\theta))$. A slice around the state θ is constructed via a horizontal bracket (θ_l, θ_r), and a new state θ' is proposed by sampling from the slice. If $\pi(\theta') > u'$, the proposal is accepted; otherwise the size of the bracket is decreased.

There are many ways in which the bracket can be constructed and adapted. In Chapter 5, we use the "stepping out" and "shrinking" procedures for generating and resizing the proposal bracket, as defined in MacKay (2003, Chapter 29.7) and Algorithm 2.3.2. Finally, Neal (2003) describes several ways in which slice sampling can be extended to multiple dimensions.

2.3.3 Elliptical slice sampling

Elliptical slice sampling (Murray et al., 2010) is algorithm used for inference in models with latent high-dimensional Gaussian structure and is often applied to models with Gaussian processes. Like the slice sampling algorithm described above, it requires no (or minimal) tuning.

Algorithm 2.3.2 Slice sampling iteration

1: **Input:** Current state θ, (unnormalized) target density π
2: Sample a random height

$$u' \sim \text{unif}(0, \pi(\theta))$$

3: Define a bracket (θ_l, θ_r) around θ, e.g., "stepping out": $r \sim \text{unif}(0, 1)$

$$\begin{aligned} \theta_l &= \theta - rw \\ \theta_r &= \theta + (1-r)w \\ \text{while } \pi(\theta_l) > u' &\text{: } \theta_l = \theta_l - w \\ \text{while } \pi(\theta_r) > u' &\text{: } \theta_r = \theta_r + w \end{aligned}$$

4: Draw proposal $\theta' \sim \text{unif}(\theta_l, \theta_r)$
5: **if** $\log \pi(\theta') > \log u'$ **then**
6: Accept proposal θ'
7: **else**
8: Resize bracket and generate new proposal, e.g., "shrinking":
9: **if** $\theta' > \theta$ **then**:
10: $\theta_r = \theta$
11: **else**:
12: $\theta_l = \theta$
13: **end if**
14: Goto Step 4
15: **end if**
16: **Output:** New state θ'

Let $\theta \sim N(0, \Sigma)$ denote the latent D-dimensional Gaussian variable of interest, and consider the likelihood as a function of θ, $L(\theta) = p(X_{1:N} \,|\, \theta)$. The target density of interest is

$$\pi(\theta) \propto \mathcal{N}(\theta \,|\, 0, \Sigma)\, L(\theta). \tag{2.20}$$

We note that while this model assumes a mean-zero Gaussian prior, one can perform a change of variables to obtain other priors.

The idea is to propose a new state θ' by sampling from a proposal region on an ellipse that passes through the current state θ and an auxiliary state ν:

$$\theta' = \theta \cos \vartheta + \nu \sin \vartheta,$$

Algorithm 2.3.3 Ellipitical slice sampling iteration

1: **Input:** Current state θ, log-likelihood L, covariance Σ (or its Cholesky decomposition)
2: Choose ellipse $\nu \sim \mathcal{N}(0, \Sigma)$
3: Construct log-likelihood threshold:

$$u \sim \text{unif}(0, 1), \qquad \log y = \log L(\theta) + \log u$$

4: Draw initial proposal for angle and define bracket

$$\vartheta \sim \text{unif}(0, 2\pi), \qquad [\vartheta_{\min}, \vartheta_{\max}] = [\vartheta - 2\pi, \vartheta]$$

5: Construct proposal $\theta' = \theta \cos \vartheta + \nu \sin \vartheta$
6: **if** $\log L(\theta') > \log y$ **then**
7: Accept proposal θ'
8: **else**
9: Resize bracket and generate new proposal:
10: **if** $\vartheta < 0$ **then**:
11: $\vartheta_{\min} = \vartheta$
12: **else**:
13: $\vartheta_{\max} = \vartheta$
14: **end if**
15: $\vartheta \sim \text{unif}[\vartheta_{\min}, \vartheta_{\max}]$
16: Goto Step 5
17: **end if**
18: **Output:** New state θ'

where $\vartheta \sim \text{unif}(0, 2\pi)$ and $\nu \sim \mathcal{N}(0, \Sigma)$. The proposal is accepted according to a likelihood threshold, and if rejected, the proposal region is resized. The full algorithm is summarized in Algorithm 2.3.3.

2.3.4 Two-stage Metropolis-Hastings

As we described in Section 2.3.1, the M-H algorithm is impractical if $\pi(\theta)$ is expensive to evaluate or intractable. The two-stage MH algorithm uses a low-fidelity likelihood L^{LF} as a low-pass filter to avoid the evaluation of an expensive high fidelity likelihood L^{HF}. In each iteration t, a proposal θ' is generated from the proposal distribution $q(\cdot|\theta^{(t-1)})$. The algorithm the proceeds in the following two stages.

Stage 1: The proposal is accepted for the second stage according to the acceptance probability

$$R^{\mathrm{LF}}(\theta;\theta') = \min\left(1, \frac{\pi(\theta')L^{\mathrm{LF}}(\theta')q(\theta|\theta')}{\pi(\theta)L^{\mathrm{LF}}(\theta)q(\theta'|\theta)}\right), \tag{2.21}$$

where $\theta = \theta^{(t-1)}$. If the proposal is rejected, then the value $\theta^{(t)} = \theta^{(t-1)}$.

Stage 2: In the second stage, the proposal θ' is accepted with probability

$$R^{\mathrm{HF}}(\theta;\theta') = \min\left(1, \frac{\pi(\theta')L^{\mathrm{HF}}(\theta')Q(\theta|\theta')}{\pi(\theta)L^{\mathrm{HF}}(\theta)Q(\theta'|\theta)}\right), \tag{2.22}$$

where the proposal distribution Q satisfies

$$Q(\theta'|\theta) = R(\theta',\theta)q(\theta'|\theta) + \left(1 - \int R(\theta,\theta')q(\theta'|\theta)d\theta'\right)\delta_\theta(\theta'). \tag{2.23}$$

Note that in the algorithm, the integral does not need to be explicitly computed, since if $\theta = \theta'$, the chain remains at the same value, and if $\theta \neq \theta'$, then $Q(\theta'|\theta) = R^{\mathrm{LF}}(\theta;\theta')q(\theta'|\theta)$. Furthermore, the high-fidelity acceptance probability can be computed as

$$R^{\mathrm{HF}}(\theta;\theta') = \min\left(1, \frac{L^{\mathrm{HF}}(\theta')L^{\mathrm{LF}}(\theta)}{L^{\mathrm{HF}}(\theta)L^{\mathrm{LF}}(\theta')}\right). \tag{2.24}$$

If the proposal is accepted, then the value $\theta^{(t)} = \theta'$, and otherwise, $\theta^{(t)} = \theta^{(t-1)}$.

2.3.5 Simulated annealing

Finally, we note that MCMC is often used for optimization. In simulated annealing, the goal is to sample from some distribution $P(\theta) \propto \exp(-E(\theta))$, where $E(\theta)$ is an energy function. If used for optimization, $E(\theta)$ is the function we are interested in minimizing. In the simplest simulated annealing case, we instead sample from the annealed distribution $\pi(\theta) \propto \exp(-E(\theta))^{\frac{1}{T}} = \exp(-E(\theta)/T)$.

Chapter 3

Finite mixture models do not reliably learn the number of components

Scientists and engineers are often interested in learning the number of subpopulations (or components) present in a data set. A common suggestion is to use a finite mixture model (FMM) with a prior on the number of components. Past work has shown the resulting FMM component-count posterior is consistent; that is, the posterior concentrates on the true, generating number of components. But consistency requires the assumption that the component likelihoods are perfectly specified, which is unrealistic in practice. In this chapter, we add rigor to data-analysis folk wisdom by proving that under even the slightest model misspecification, the FMM component-count posterior *diverges*: the posterior probability of any particular finite number of components converges to 0 in the limit of infinite data. Contrary to intuition, posterior-density consistency is not sufficient to establish this result. We develop novel sufficient conditions that are more realistic and easily checkable than those common in the asymptotics literature. We illustrate practical consequences of our theory on simulated and real data.

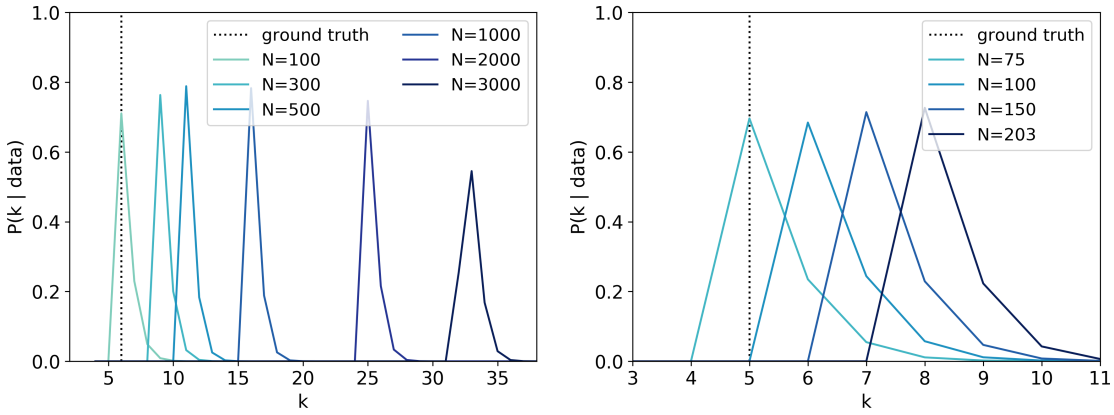
3.1 Introduction

Mixture modeling is a mainstay of statistical machine learning. In applications where the number of mixture components is unknown in advance, a principal inferential goal is to estimate and interpret this number. For example, practitioners might wish to find the number of latent genetic populations (Huelsenbeck and Andolfatto, 2007; Lorenzen et al., 2006; Pritchard et al., 2000; Tonkin-Hill et al., 2019), gene tissue profiles (Medvedovic and Sivaganesan, 2002; Yeung et al., 2001), cell types (Chan et al., 2008; Prabhakaran et al., 2016), microscopy groups (Griffié et al., 2016; Rubin-Delanchy et al., 2015), haplotypes (Xing et al., 2006), switching Markov regimes in US dollar exchange rate data (Otranto and Gallo, 2002), gamma-ray burst types (Mukherjee et al., 1998), segmentation regions in an image (e.g., tissue types in an MRI scan (Banfield and Raftery, 1993)), observed humans in radar data (Teklehaymanot et al., 2018), basketball shot selection groups (Hu et al., 2020), or communities in a social network (Geng et al., 2019; Legramanti et al., 2020).

A natural question then is: can we reliably learn the number of latent groups in a data set? To make this question concrete, we focus on a Bayesian approach. Consider the case where the true, generating number of components is known. A natural check on a Bayesian mixture analysis is to establish that the Bayesian posterior on the number of components increasingly concentrates near the truth as the number of data points becomes arbitrarily large. In the remainder, we will focus on this check—though our work has practical implications beyond Bayesian analysis.

A standard Bayesian analysis uses a component-count prior with support on all strictly-positive integers (Miller and Harrison, 2018). Nobile (1994) has shown that the component-count posterior of the resulting *finite mixture model* (FMM) does concentrate at the true number of components. But crucially, this result depends on the assumption that the component likelihoods are perfectly specified. In every application we have listed above, the true generating component likelihoods do not take a convenient parametric form that might be specified in advance. Indeed, some form of misspecification, even if slight, is typical in practice. So we must ask how the component-count posterior behaves when the component likelihoods are misspecified.

Data science folk wisdom suggests that when component likelihoods are misspecified, mixture



(a) Mouse single-cell RNA-seq data

(b) Lung tissue gene expression data

Figure 3.1.1: Posterior probability of the number of components k for Gaussian mixture models, fit to (a) mouse cortex single-cell RNA sequencing data and (b) lung tissue gene expression data. Details in Section 3.8.2.

models will tend to overestimate the number of clusters; see, e.g., Section 7.1 of Frühwirth-Schnatter (2006). This overestimation is apparent in Figure 3.1.1, which shows the component-count posterior of a Gaussian mixture model applied to two example gene expression data sets (de Souto et al., 2008; Prabhakaran et al., 2016). In fact, Figure 3.1.1 demonstrates an effect far worse than just overestimation: the posterior distribution appears to concentrate for any (large enough) fixed amount of data, but actually concentrates on increasing values as more data are observed. Therefore, inference is unreliable; the practitioner may draw quite different conclusions depending on how large the data set is.

In the present chapter, we add rigor to existing data science folk intuition by proving that this behavior occurs in a wide class of FMMs under arbitrarily small amounts of misspecification. We examine FMMs with essentially any component shape—where we make only mild, realistic, and checkable assumptions on the component likelihoods. Notably, we include univariate and multivariate Gaussian component likelihoods in our theory, but do not restrict only to these shapes. We show that under our assumptions and when the component likelihoods are not perfectly specified, the component-count posterior concentrates strictly away from the true number of components. In fact, we go further to show that the FMM posterior for the number of components *diverges*: for *any*

finite $k \in \mathbb{N}$, the posterior probability that the number of components is k converges to 0 almost surely as the amount of data grows.

We start by introducing FMMs and stating our main result in Section 3.2. We discuss our assumptions in more detail in Section 3.3 and prove our result in Section 3.4. In Section 3.5 we extend our main theorem to priors that may vary as the data set grows. We discuss related work below and in Section 3.7. The chapter concludes in Section 3.8 with empirical evidence that the FMM component-count posterior depends strongly on the amount of observed data. Our results demonstrate that, in practice, past estimates of component number may have strongly depended on the size of a particular data set.

Filling a gap in the literature. While recent work has established various asymptotic properties of mixture models, we observe that our results here are not trivial extensions of existing research. First note that, intuitively, as the number of data points grows, the posterior concentrates at the generating density (Ghosal and van der Vaart, 2017; Ghosh and Ramamoorthi, 2003; Schwartz, 1965), which can be well-approximated by an infinite mixture due in part to misspecification. However, posterior consistency for the density alone is not enough to guarantee consistency for the model parameters; parameter consistency may not hold under, for instance, a discontinuous mapping from the component parameter to the component density.

Second, note that posterior divergence for the number of components could, in principle, be obtained if parameter consistency for the mixture holds. But existing results on parameter consistency, such as Nguyen (2013), focus on obtaining rates of contraction; thus these results rely on stronger conditions that are typically verified for individual component families by imposing additional constraints, such as (second-order) strong identifiability of the mixture or a compact parameter space (Chen, 1995; Heinrich and Kahn, 2018; Nguyen, 2013). But neither of these constraints are satisfied by common families of interest such as Gaussians with unknown mean and variance. By contrast, our result uses the weakest notion of mixture identifiability (Teicher, 1961) along with a continuity condition on the component family, and we relax the requirement of a compact parameter space. To do so, we develop a novel theoretical condition that requires the component family to have *degenerate limits*. Together, these advances ensure the applicability of our theory to practical likelihood families

including, but not limited to, full Gaussians. In fact, the degenerate limits condition and its use in our analysis may be useful for extending other results on posterior asymptotics that currently rely on compact parameter spaces.

Finally, Miller and Harrison (2013, 2014) have shown that typical uses of Dirichlet process mixture models (DPMMs) inconsistently estimate the true, generating number of components. But Miller and Harrison (2013, 2014) focus on the DPMM prior instead of the FMM and on perfectly specified likelihoods. The DPMM is misspecified in a different sense than the one we focus on in the present chapter: namely, the DPMM uses infinitely many components though we assume finitely many generated the data. For this reason, practitioners typically invoke the DPMM posterior on the number of *clusters* (Huelsenbeck and Andolfatto, 2007; Pella and Masuda, 2006), i.e., components represented in the observed data, rather than the component-count posterior directly. Indeed, Miller and Harrison (2018) recommend using the FMM we study here to resolve the difficulties of the DPMM. Finally, observe that the work of Miller and Harrison (2018) demonstrates that nonparametrically estimating component shape with a DPMM would not provide a simple resolution of the FMM divergence issue.

3.2 Main result

We begin with a brief description of the finite mixture model used in this work. In this section, we provide just enough detail to state Definition 3.2.1 and leave the precise probabilistic details for Section 3.3. Let g be a mixing measure $g := \sum_{j=1}^{k} p_j \delta_{\theta_j}$ on a parameter space Θ with $p_j \in [0, 1]$ and $\sum_{j=1}^{k} p_j = 1$, and let $\Psi = \{\psi_\theta : \theta \in \Theta\}$ be a family of component distributions dominated by a σ-finite measure μ. We can express a finite mixture f of the components as

$$f = \int_\Theta \psi_\theta dg(\theta) = \sum_{j=1}^{k} p_j \psi_{\theta_j}.$$

Consider a Bayesian model with prior distribution Π on the set of all mixing measures $\mathbb{G}$ on Θ with finitely many atoms, i.e., $g \sim \Pi$, and likelihood corresponding to conditionally i.i.d. data from

$f = \int \psi_\theta dg(\theta)$. The model assumes the likelihood is f, but the model is *misspecified*; i.e., the observations $X_{1:N} := (X_1, \ldots, X_N)$ are generated conditionally i.i.d. from a finite mixture f_0 of distributions *not* in Ψ.

Our main result is that under this misspecification of the likelihood, the posterior on the number of components $\Pi(k \,|\, X_{1:N})$ diverges; i.e., for any finite $k \in \mathbb{N}$, $\Pi(k \,|\, X_{1:N}) \to 0$ as $N \to \infty$. We make only two requirements of the mixture model to guarantee this result: (1) the true data-generating distribution f_0 must be arbitrarily well-approximated by finite mixtures of Ψ, and (2) the family Ψ must satisfy mild regularity conditions that hold for popular mixture models (e.g., the family Ψ of Gaussians parametrized by mean and variance). We provide precise definitions of the assumptions needed for Definition 3.2.1 to hold in Section 3.3, and a proof in Section 3.4.

Theorem 3.2.1 (Main result)**.** *Suppose observations $X_{1:N}$ are generated i.i.d. from a distribution f_0 that is not a finite mixture of Ψ. Assume that:*

Definition 3.3.1: f_0 is in the KL-support of the prior Π,

Definition 3.3.6: Ψ is continuous, is mixture-identifiable, and has degenerate limits.

Then the posterior on the number of components diverges; i.e., for all $k \in \mathbb{N}$,

$$\Pi(k \,|\, X_{1:N}) \overset{N\to\infty}{\longrightarrow} 0 \quad f_0\text{-a.s.} \tag{3.1}$$

Note that the conditions of the theorem—although technical—are satisfied by a wide class of models used in practice. Definition 3.3.1 requires that the prior Π places enough mass on mixtures near the true generating distribution f_0. Definition 3.3.6 enforces regularity of the component family and is satisfied by many popular models used in practice, such as the multivariate Gaussian family.

Proposition 3.2.2. *Let $\Psi = \left\{\mathcal{N}(\nu, \Sigma) : \nu \in \mathbb{R}^d, \Sigma \in \mathbb{S}^d_{++}\right\}$ be the multivariate Gaussian family, where $\mathbb{S}^d_{++} := \{\Sigma \in \mathbb{R}^{d\times d} : \Sigma = \Sigma^\top,\ \Sigma \succ 0\}$ is the set of $d \times d$ symmetric, positive definite matrices. Then Ψ satisfies Definition 3.3.6.*

Thus, provided that f_0 is in the KL-support of the prior, under a misspecified Gaussian mixture model, our main result implies that the posterior number of components diverges. While Definition 3.2.2 is stated for Gaussian component distributions, we generalize it to mixture-identifiable location-scale families Ψ in Definition B.2.2.

Additionally, we note that the divergence of the posterior given in Equation (3.1) is stronger than the behavior described in Miller and Harrison (2013) for DPMMs: namely, Miller and Harrison (2013) show that the posterior probability converges to 0 at the *true* number of components. In contrast, here we show that the posterior probability converges to 0 for any finite number of components. We conjecture that posterior divergence also holds for DPMMs, but the proof is outside of the scope of this chapter.

Extension: Priors that vary with N. While the result of Definition 3.2.1 assumes that the model uses a fixed prior Π, in practical modeling scenarios one may specify a prior Π_N that depends on the observed data $X_{1:N}$. For instance, these priors can arise in empirical Bayes; see Sections 3.5 and 3.8 for examples. In Section 3.5 we show that if f_0 satisfies a modified KL-support condition with respect to the sequence of priors Π_N, the number of components also diverges in this setting.

Extension: Priors with an upper bound on the number of components. Definition 3.2.1 is designed for priors that place full support on any positive integer number of components. One might instead use a prior that has support on at most $\tilde{k}$ components, with $\tilde{k}$ finite. In this case, the posterior number of components will not diverge to infinity but instead typically concentrate on the upper bound, $\tilde{k}$. A precise statement of this behavior appears in Definition B.1.1 (in Appendix B.1) as an analog of our Definition 3.2.1. Definition B.1.1 shows that posterior inference does not improve over the baseline estimate of the number of components provided by $\tilde{k}$. If $\tilde{k}$ is already a good estimate of the number components, posterior concentration at $\tilde{k}$ does not improve the estimate. In practice $\tilde{k}$ is often chosen as some large upper bound of convenience; then $\tilde{k}$ is not a good estimate of the number of components, and concentration at $\tilde{k}$ is undesirable.

3.3 Precise setup and assumptions in Theorem 3.2.1

This section makes the details of the modeling setup and each of the conditions in Definition 3.2.1 precise.

3.3.1 Notation and setup

Let $\mathbb{X}$ and Θ be Polish spaces for the observations and parameters, respectively, and endow both with their Borel σ-algebra. For a topological space $(\cdot)$, let $\mathscr{C}(\cdot)$ be the bounded continuous functions from $(\cdot)$ into $\mathbb{R}$, and $\mathscr{P}(\cdot)$ be the set of probability measures on $(\cdot)$ endowed with the weak topology metrized by the Lévy-Prokhorov distance d (Definition 2.1.1). We use $f_i \Rightarrow f$ and $f_i \iff f_i'$ to denote $\lim_{i\to\infty} d(f_i, f) = 0$ and $\lim_{i\to\infty} d(f_i, f_i') = 0$, respectively, for $f_i, f_i', f \in \mathscr{P}(\cdot)$. We assume that the family of distributions $\Psi = \{\psi_\theta : \theta \in \Theta\}$ is absolutely continuous with respect to a σ-finite base measure μ, i.e., $\psi_\theta \ll \mu$ for all $\theta \in \Theta$, and that for measurable $A \subseteq \mathbb{X}$, $\psi_\theta(A)$ is a measurable function on Θ. Define the measurable mapping $F : \mathscr{P}(\Theta) \to \mathscr{P}(\mathbb{X})$ from mixing measures to mixtures of Ψ, $F(g) = \int \psi_\theta dg(\theta)$. Let $\mathbb{G}$ be the set of atomic probability measures on Θ with finitely many atoms, and let $\mathbb{F}$ be the set of finite mixtures of Ψ.

In the Bayesian finite mixture model from Section 3.2, a mixing measure $g \sim \Pi$ is generated from a prior measure Π on $\mathbb{G}$, and $f = F(g)$ is a likelihood distribution.

The posterior distribution on the mixing measure is, for all measurable $A \subseteq \mathbb{G}$,

$$\Pi(A \,|\, X_{1:N}) = \frac{\int_A \prod_{n=1}^N \frac{df}{d\mu}(X_n)\, d\Pi(g)}{\int_{\mathbb{G}} \prod_{n=1}^N \frac{df}{d\mu}(X_n)\, d\Pi(g)}, \tag{3.2}$$

where $\frac{df}{d\mu}$ is the density of $f = F(g)$ with respect to μ. This posterior on the mixing measure $g \in \mathbb{G}$ induces a posterior on the number of components $k \in \mathbb{N}$ by counting the number of atoms in g, and it also induces a posterior on mixtures $f \in \mathbb{F}$ via the pushforward through the mapping F. We overload the notation $\Pi(\cdot \,|\, X_{1:N})$ to refer to all of these posterior distributions and $\Pi(\cdot)$ to refer to prior distributions; the meaning should be clear from context.

3.3.2 Model assumptions

The first assumption of Definition 3.2.1 is that while the true data-generating distribution f_0 is not contained in the model class $f_0 \notin \mathbb{F}$, it lies on the boundary of the model class. In particular, we assume f_0 is in the *KL-support* of the prior Π. Denote the Kullback-Leibler (KL) divergence between probability measures f_0 and f as

$$\mathrm{KL}(f_0, f) := \begin{cases} \int \log\left(\frac{df_0}{df}\right) df_0 & f_0 \ll f \\ \infty & \text{otherwise} \end{cases}.$$

Assumption 3.3.1. *For all $\epsilon > 0$, the prior distribution Π satisfies*

$$\Pi(f \in \mathbb{F} : \mathrm{KL}(f_0, f) < \epsilon) > 0.$$

We use Definition 3.3.1 in the proof of Definition 3.2.1 primarily to ensure that the Bayesian posterior is consistent for f_0. Note that Definition 3.3.1 is fairly weak in practice. Intuitively, it just requires that the family Ψ is rich enough so that mixtures of Ψ can approximate f_0 arbitrarily well, and that the prior Π places sufficient mass on those mixtures close to f_0. For Bayesian mixture modeling, Ghosal et al. (1999, Theorem 3), Tokdar (2006, Theorem 3.2), Wu and Ghosal (2008, Theorem 2.3), and Petralia et al. (2012, Theorem 1) provide conditions needed to satisfy Definition 3.3.1.

The second assumption of Definition 3.2.1 is that the family of component distributions Ψ is well-behaved. This assumption has three stipulations. First, the mapping $\theta \mapsto \psi_\theta$ must be continuous; this condition essentially asserts that similar parameter values θ must result in similar component distributions ψ_θ.

Definition 3.3.2. *The family Ψ is* continuous *if the map $\theta \mapsto \psi_\theta$ is continuous.*

Second, the family Ψ must be *mixture-identifiable*, which guarantees that each mixture $f \in \mathbb{F}$ is associated with a unique mixing measure $G \in \mathbb{G}$.

Definition 3.3.3 (Teicher (1961, 1963))**.** *The family* Ψ *is* mixture-identifiable *if the mapping* $F(g) = \int \psi_\theta dg(\theta)$ *restricted to finite mixtures* $F : \mathbb{G} \to \mathbb{F}$ *is a bijection.*

In practice, one should always use an identifiable mixture model for clustering; without identifiability, the task of learning the number of components is ill posed. And many models satisfy mixture-identifiability, such as finite mixtures of the multivariate Gaussian family (Yakowitz and Spragins, 1968), the Cauchy family (Yakowitz and Spragins, 1968), the gamma family (Teicher, 1963), the generalized logistic family, the generalized Gumbel family, the Weibull family, and von Mises family (Ho and Nguyen, 2016, Theorem 3.3). A number of authors (e.g. Chen, 1995; Guha et al., 2021; Heinrich and Kahn, 2018; Ho and Nguyen, 2016; Ishwaran et al., 2001; Nguyen, 2013) appeal to stronger notions of identifiability for mixtures than Definition 3.3.3. But, to show posterior divergence in the present work, we do not require conditions stronger than Definition 3.3.3.

The third stipulation—that the family Ψ has *degenerate limits*—guarantees that a "poorly behaved" sequence of parameters $(\theta_i)_{i\in\mathbb{N}}$ creates a likewise "poorly behaved" sequence of distributions $(\psi_{\theta_i})_{i\in\mathbb{N}}$. This condition allows us to rule out such sequences in the proof of Definition 3.2.1, and is the essential regularity condition to guarantee that a sequence of finite mixtures of at most k components cannot approximate f_0 arbitrarily closely.

Definition 3.3.4. *A sequence of distributions* $(\psi_i)_{i=1}^\infty$ *is* μ-wide *if for any closed set* C *such that* $\mu(C) = 0$ *and any sequence of distributions* $(\phi_i)_{i=1}^\infty$ *such that* $\psi_i \iff \phi_i$*,*

$$\limsup_{i\to\infty} \phi_i(C) = 0.$$

Definition 3.3.5. *The family* Ψ *has* degenerate limits *if for any tight,* μ*-wide sequence* $(\psi_{\theta_i})_{i\in\mathbb{N}}$*, we have that* $(\theta_i)_{i\in\mathbb{N}}$ *is relatively compact.*

The contrapositive of Definition 3.3.5 provides an intuitive explanation of the condition: as $i \to \infty$, for any sequence of parameters θ_i that eventually leaves every compact set $K \subseteq \Theta$, either the ψ_{θ_i} become "arbitrarily flat" (not tight) or "arbitrarily peaky" (not μ-wide). For example, consider the family Ψ of Gaussians on $\mathbb{R}$ with Lebesgue measure μ. If the variance of ψ_{θ_i} shrinks as i grows,

the sequence of distributions converges weakly to a sequence of point masses (not dominated by the Lebesgue measure). If either the variance or the mean diverges, the distributions flatten out and the sequence is not tight. We use the fact that these are the only two possibilities when a sequence of parameters is poorly behaved (not relatively compact) in the proof of Definition 3.2.1.

These three stipulations together yield Definition 3.3.6.

Assumption 3.3.6. *The mixture component family Ψ is continuous, is mixture-identifiable, and has degenerate limits.*

3.4 Proof of Theorem 3.2.1

The proof has two essential steps. The first is to show that the Bayesian posterior is weakly consistent for the mixture f_0; i.e., for any weak neighborhood U of f_0 the sequence of posterior distributions satisfies

$$\Pi(U \mid X_{1:N}) \overset{N\to\infty}{\longrightarrow} 1, \quad f_0\text{-a.s.} \tag{3.3}$$

By Schwartz's theorem (Definition 2.1.8), weak consistency for f_0 is guaranteed directly by Definition 3.3.1 and the fact that Ψ is dominated by a σ-finite measure μ. The second step is to show that for any finite $k \in \mathbb{N}$, there exists a weak neighborhood U of f_0 containing no mixtures of the family Ψ with at most k components. Together, these steps show that the posterior probability of the set of all k-component mixtures converges to 0 f_0-a.s. as the amount of observed data grows.

We provide a proof of the second step. To begin, note that Definition 3.3.1 has two additional implications about f_0 beyond Equation (3.3). First, f_0 must be absolutely continuous with respect to the dominating measure μ; if it were not, then there exists a measurable set A such that $f_0(A) > 0$ and $\mu(A) = 0$. Since μ dominates Ψ, any $f \in \mathbb{F}$ satisfies $f(A) = 0$. Therefore $\mathrm{KL}(f_0, f) = \infty$, and the prior support condition cannot hold. Second, it implies that f_0 can be arbitrarily well-approximated by finite mixtures under the weak metric, i.e., there exists a sequence of finite mixtures $f_i \in \mathbb{F}$, $i \in \mathbb{N}$ such that $f_i \Rightarrow f_0$ as $i \to \infty$. This holds because $\sqrt{\frac{1}{2}\mathrm{KL}(f_0, f)} \geq \mathrm{TV}(f_0, f) \geq d(f_0, f)$.

Now suppose the contrary of the claim for the second step, i.e., that there exists a sequence $(f_i)_{i=1}^{\infty}$ of mixtures of at most k components from Ψ such that $f_i \Rightarrow f_0$. By mixture-identifiability, we have a sequence of mixing measures g_i with at most k atoms such that $F(g_i) = f_i$. Suppose first that the atoms of the sequence $(g_i)_{i\in\mathbb{N}}$ either stay in a compact set or have weights converging to 0. More precisely, suppose there exists a compact set $K \subseteq \Theta$ such that

$$g_i\left(\Theta \setminus K\right) \to 0. \tag{3.4}$$

Decompose each $g_i = g_{i,K} + g_{i,\Theta\setminus K}$ such that $g_{i,K}$ is supported on K and $g_{i,\Theta\setminus K}$ is supported on $\Theta \setminus K$. Define the sequence of probability measures $\hat{g}_{i,K} = \frac{g_{i,K}}{g_{i,K}(\Theta)}$ for sufficiently large i such that the denominator is nonzero. Then Equation (3.4) implies

$$F\left(\hat{g}_{i,K}\right) \Rightarrow f_0.$$

Since Ψ is continuous and mixture-identifiable, the restriction of F to the domain $\mathbb{G}$ is continuous and invertible; and since K is compact, the elements of $(\hat{g}_{i,K})_{i\in\mathbb{N}}$ are contained in a compact set $\mathbb{G}_K \subseteq \mathbb{G}$ by Prokhorov's theorem (Definition 2.1.5). Therefore $F(\mathbb{G}_K) = \mathbb{F}_K$ is also compact, and the map F restricted to the domain $\mathbb{G}_K$ is uniformly continuous with a uniformly continuous inverse by Rudin (1976, Theorems 4.14, 4.17, 4.19). Next since $F(\hat{g}_{i,K}) \Rightarrow f_0$, the sequence $F(\hat{g}_{i,K})$ is Cauchy in $\mathbb{F}_K$; and since F^{-1} is uniformly continuous on $\mathbb{F}_K$, the sequence $\hat{g}_{i,K}$ must also be Cauchy in $\mathbb{G}_K$. Since $\mathbb{G}_K$ is compact, $\hat{g}_{i,K}$ converges in $\mathbb{G}_K$. Definition 3.4.1 below guarantees that the convergent limit g_K is also a mixing measure with at most k atoms; continuity of F implies that $F(g_K) = f_0$, which is a contradiction, since by assumption f_0 is not representable as a finite mixture of Ψ.

Lemma 3.4.1. *Suppose $\phi, (\phi_i)_{i\in\mathbb{N}}$ are Borel probability measures on a Polish space such that $\phi_i \Rightarrow \phi$ and $\sup_i |\operatorname{supp}\phi_i| \leq k \in \mathbb{N}$. Then $|\operatorname{supp}\phi| \leq k$.*

Proof. Suppose $|\operatorname{supp}\phi| > k$. Then we can find $k+1$ distinct points $x_1, \ldots, x_{k+1} \in \operatorname{supp}\phi$. Pick any metric ρ on the Polish space, and denote the minimum pairwise distance between the

points 2ϵ. Then for each point $j = 1, \ldots, k+1$ define the bounded, continuous function $h_j(x) = 0 \vee \big(1 - \epsilon^{-1}\rho(x, x_j)\big)$. Since $x_j \in \operatorname{supp}\phi$, we have that $\int h_j d\phi > 0$. Weak convergence $\phi_i \Rightarrow \phi$ therefore implies $\min_{j=1,\ldots,k+1} \liminf_{i\to\infty} \int h_j d\phi_i > 0$. But the h_j are nonzero on disjoint sets, and each ϕ_i only has k atoms; the pigeonhole principle yields a contradiction. □

Now we consider the remaining case: for all compact sets $K \subseteq \Theta$, $g_i(\Theta \setminus K) \not\to 0$. Therefore there exists a sequence of parameters $(\theta_i)_{i=1}^{\infty}$ that is not relatively compact such that $\limsup_{i\to\infty} g_i(\{\theta_i\}) > 0$. By Definition 3.3.6, the sequence $(\psi_{\theta_i})_{i\in\mathbb{N}}$ is either not tight or not μ-wide. If $(\psi_{\theta_i})_{i\in\mathbb{N}}$ is not tight then $f_i = F(g_i)$ is not tight, and by Prokhorov's theorem f_i cannot converge to a probability measure, which contradicts $f_i \Rightarrow f_0$. If $(\psi_{\theta_i})_{i\in\mathbb{N}}$ is not μ-wide then $f_i = F(g_i)$ is not μ-wide. Denote $(\phi_i)_{i\in\mathbb{N}}$ to be the singular sequence associated with $(f_i)_{i\in\mathbb{N}}$ and C to be the closed set such that $\limsup_{i\to\infty} \phi_i(C) > 0$, $\mu(C) = 0$, and $\phi_i \iff f_i$ per Definition 3.3.4. Since $f_0 \ll \mu$, $f_0(C) = 0$. But $f_i \Rightarrow f_0$ implies that $\phi_i \Rightarrow f_0$, so $\limsup_{i\to\infty} \phi_i(C) = f_0(C) = 0$ by the Portmanteau theorem (Definition 2.1.3). This is a contradiction.

3.5 Extension to priors that vary with N

Our main result (i.e., Definition 3.2.1) applies to the setting of a fixed prior Π. However, it is often natural to specify a prior distribution that changes with N (e.g., Roeder and Wasserman, 1997; Richardson and Green, 1997; and Miller and Harrison, 2018, Section 7.2.1). Definition 3.5.2 below demonstrates that a result nearly identical to Definition 3.2.1 holds for priors that are allowed to vary with N, provided that f_0 is in the KL-support of the *sequence* of priors Π_N. The only difference is that our result in this case is slightly weaker: we show that the posterior number of components diverges in probability rather than almost surely.

Assumption 3.5.1. *For all $\epsilon > 0$, the sequence of prior distributions Π_N satisfies*

$$\liminf_{N\to\infty} \Pi_N(f : \mathrm{KL}(f_0, f) < \epsilon) > 0.$$

Corollary 3.5.2. *Suppose in the setting of Definition 3.2.1 we replace Definition 3.3.1 with Definition 3.5.1. Then the posterior on the number of components diverges in f_0-probability: i.e., for all $k \in \mathbb{N}$,*

$$\Pi(k \mid X_{1:N}) \overset{N\to\infty}{\longrightarrow} 0 \quad \textit{in } f_0\textit{-probability.}$$

Proof. Since for any $\epsilon > 0$, $\liminf_{N\to\infty} \Pi_N(f : \mathrm{KL}(f_0, f) < \epsilon) > 0$, Ghosal and van der Vaart (2017, Theorem 6.17, Lemma 6.26, and Example 6.20) imply that the posterior is weakly consistent at f_0 in probability: i.e., for any weak neighborhood U of f_0,

$$\Pi(U \mid X_{1:N}) \overset{N\to\infty}{\longrightarrow} 1 \quad \text{in } f_0\text{-probability.}$$

Definition 3.5.1 also implies that for sufficiently large N, f_0 is a weak limit of finite mixtures in $\mathbb{F}$. The remainder of the proof is identical to that of Definition 3.2.1. □

3.6 Extension to power posteriors

One proposal to address likelihood misspecification is to use an α-posterior, or *power posterior*, where we replace the likelihood with the same likelihood but raised to a fixed power $\alpha > 0$, often between 0 and 1 (Ghosh and Sudderth, 2012; Grünwald and van Ommen, 2017; Grünwald, 2006; Holmes and Walker, 2017; Royall and Tsou, 2003). Power posteriors have much more general application than just to mixture models; we focus only on mixture models here and note that behavior of power posteriors for other models may be very different than for mixture models. In a separate line of work, Miller and Dunson (2019) propose a *coarsened posterior*, which they show can be closely approximated by a variant of the power posterior with exponent $\alpha_N \to 0$ as the number of data points $N \to \infty$; in fact, they use this approximation in all of their experiments. Note that we use the terminology "power posterior" or α-posterior throughout to refer to the *fixed* power case. Here we find that the component-count power posterior, with power that is constant in N, diverges in the same way.

The α-posterior distribution on the mixing measure is

$$\forall \text{ measurable } A \subseteq \mathbb{G}, \quad \Pi^{(\alpha)}(A \mid X_{1:N}) = \frac{\int_A \prod_{n=1}^N (\frac{df}{d\mu})^\alpha(X_n)\, d\Pi(g)}{\int_{\mathbb{G}} \prod_{n=1}^N (\frac{df}{d\mu})^\alpha(X_n)\, d\Pi(g)}, \tag{3.5}$$

where $\frac{df}{d\mu}$ is the density of $f = F(g)$ with respect to μ and $\alpha > 0$. The α-posterior on the mixing measure $g \in \mathbb{G}$ induces an α-posterior on the number of components $k \in \mathbb{N}$ by counting the number of atoms in g, and it also induces a posterior on mixtures $f \in \mathbb{F}$ via the pushforward through the mapping F. We overload the notation $\Pi^{(\alpha)}(\cdot \mid X_{1:N})$ to refer to all of these α-posterior distributions and $\Pi(\cdot)$ to refer to prior distributions; the meaning should be clear from context.

Let $\Pi^{(\alpha)}(k \mid X_{1:N})$ denote the posterior marginal number of components induced by raising the likelihood to a fixed power $\alpha > 0$. Our main result is that under this misspecification of the likelihood, for any $\alpha \in (0, 1]$, the α-posterior on the number of components $\Pi^{(\alpha)}(k \mid X_{1:N})$ *diverges*; i.e., for any finite $k \in \mathbb{N}$, $\Pi^{(\alpha)}(k \mid X_{1:N}) \to 0$ as $N \to \infty$.

We make only two requirements of the mixture model to guarantee this result: (1) the true data-generating distribution f_0 must be arbitrarily well-approximated by finite mixtures of Ψ, and (2) the family Ψ must satisfy mild regularity conditions that hold for popular mixture models (e.g., the family Ψ of Gaussians parametrized by mean and variance). We provide precise definitions of the assumptions needed for Definition 3.2.1 to hold in Section 3.3, and a proof in Section 3.4.

Theorem 3.6.1 (Main result)**.** *Suppose observations $X_{1:N}$ are generated i.i.d. from a distribution f_0 that is not a finite mixture of Ψ. Assume that:*

Definition 3.3.1: f_0 is in the KL-support of the prior Π, and

Definition 3.3.6: Ψ is continuous, is mixture-identifiable, and has degenerate limits.

Then for any $\alpha \in (0, 1]$, the α-posterior on the number of components diverges; i.e., for all $k \in \mathbb{N}$,

$$\Pi^{(\alpha)}(k \mid X_{1:N}) \overset{N\to\infty}{\longrightarrow} 0 \quad f_0\text{-a.s.} \tag{3.6}$$

We use a similar analysis as before to lend insight into the power posterior for finite mixtures;

again the proof has two essential steps. The first is to show that for any $\alpha \in (0, 1]$, the α-posterior is weakly consistent for the mixture f_0; i.e., for any weak neighborhood U of f_0 the sequence of posterior distributions satisfies

$$\Pi^{(\alpha)}(U \mid X_{1:N}) \overset{N\to\infty}{\longrightarrow} 1, \quad f_0\text{-a.s.} \tag{3.7}$$

Weak consistency for f_0 is guaranteed directly by Definition 3.3.1 and the fact that Ψ is dominated by a σ-finite measure μ. For $\alpha = 1$, Definition 3.3.1 implies weak consistency for f_0, i.e., Equation (3.7) (Ghosh and Ramamoorthi, 2003, Theorem 4.4.2). For $\alpha \in (0, 1)$, if Definition 3.3.1 holds, by Ghosal and van der Vaart (2017, Theorem 6.54), with f_0-probability 1, for any Hellinger neighborhood V of f_0 and any $\alpha \in (0, 1)$, $\Pi^{(\alpha)}(V|X_{1:N}) \overset{N\to\infty}{\longrightarrow} 1$. Thus, since $V \subseteq U$,

$$\Pi^{(\alpha)}(U \mid X_{1:N}) \geq \Pi^{(\alpha)}(V \mid X_{1:N}) \overset{N\to\infty}{\longrightarrow} 1, \qquad f_0\text{-a.s.}$$

The second step follows the same proof as before.

3.7 Related work

In this work, we consider FMMs with a prior on the number of components. In the broader Bayesian mixture modeling literature, posterior consistency for the mixture density (Ghosal et al., 1999; Kruijer et al., 2010; Lijoi et al., 2004) and the mixing measure (Guha et al., 2021; Ho and Nguyen, 2016; Nguyen, 2013) is well established. But posterior consistency for the number of components is not as thoroughly characterized. There are several results establishing consistency for the number of components in well-specified FMMs. Nobile (1994, Proposition 3.5) and Guha et al. (2021, Theorem 3.1a) demonstrate that FMMs exhibit posterior consistency for the number of components when the model is well specified and Ψ is mixture-identifiable. The present work characterizes the behavior of the FMM posterior on the number of components under component misspecification. Under misspecification of the component family or the support of the true mixing measure, Guha et al. (2021, Theorem 4.1, Theorem 4.3) establish posterior rates of contraction for the mixing

measure for Gaussian and Laplace location mixtures with compact parameter spaces. Our results, which rely on posterior density consistency results, assume weaker conditions on the prior and hold for more general classes of component families, such as multivariate Gaussians parameterized by a mean and covariance.

A related approach for handling a finite but unknown number of components is to specify a prior with a finite upper bound on the number of components (e.g. Chambaz and Rousseau, 2008; Frühwirth-Schnatter and Malsiner-Walli, 2019; Ishwaran et al., 2001; Malsiner-Walli et al., 2016; Rousseau and Mengersen, 2011; Zhang et al., 2018). In the setting of overfitted FMMs with well-specified component densities, Rousseau and Mengersen (2011, Theorem 1) show that under a stronger identifiability condition than mixture-identifiability and additional regularity assumptions on the model, the posterior will concentrate properly by emptying the extra components. Ishwaran et al. (2001, Theorem 1) consider the setting of estimating the number of components with the assumption of a known upper bound on the true number of components and well-specified components, and show that the posterior does not asymptotically underestimate the number of components when assuming a stronger identifiability condition than mixture-identifiability and a KL-support condition on the prior. Under a weaker notion of (second-order) strong identifiability (Chen, 1995) and a well-specified model, Chambaz and Rousseau (2008, Theorem 4) provide upper bounds on the underestimation and overestimation error of the number of components; furthermore, they show that their conditions are satisfied by univariate Gaussians with bounded mean and variance (Chambaz and Rousseau, 2008, Corollary 1). Notably, all of these methods with finite-support priors assume well-specified component families. By contrast, we show in Theorem B.1 that even for these finite-support priors, misspecified component families yield unreliable estimates of the number of components.

Frühwirth-Schnatter (2006) provides a wide-ranging review of methodology for finite mixture modeling. In (e.g.) Section 7.1, Frühwirth-Schnatter (2006) observes that, in practice, the learned number of mixture components will generally be higher than the true generating number of components when the likelihood is misspecified—but does not prove a result about the number of components under misspecification. Similarly, Miller and Harrison (2018, Section 7.1.5) discuss the issue of estimating the number of components in FMMs under model misspecification and state that

the posterior number of components is expected to diverge to infinity as the number of samples increases, but no proof of this asymptotic behavior is provided.

Finally, a growing body of work is focused on developing more robust FMMs and related mixture models. In order to address the issue of component misspecification, a number of authors propose using finite mixture models with nonparametric component densities, e.g. Gaussian-mixture components (Bartolucci, 2005; Di Zio et al., 2007; Malsiner-Walli et al., 2017) or overfitted-mixture components (Aragam et al., 2020). However, for these finite mixture models that have mixtures as components, the posterior number of components and its asymptotic behavior have yet to be characterized.

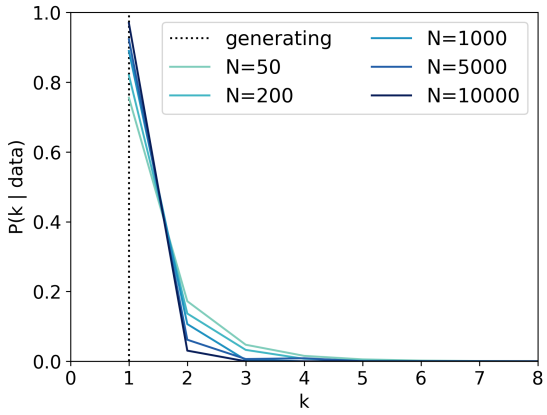
3.8 Experiments

In this section, we demonstrate one of the primary practical implications of our theory: the inferred number of components can change drastically depending on the amount of observed data in misspecified finite mixture models. For all experiments below, we use a finite mixture model with a multivariate Gaussian component family having diagonal covariance matrices and a conjugate prior on each dimension. In particular, consider number of components k, mixture weights $p \in \mathbb{R}^k$, Gaussian component precisions $\tau \in \mathbb{R}_+^{k \times D}$ and means $\theta \in \mathbb{R}^{k \times D}$, labels $Z \in \{1, \ldots, k\}^N$, and data $X \in \mathbb{R}^{N \times D}$.

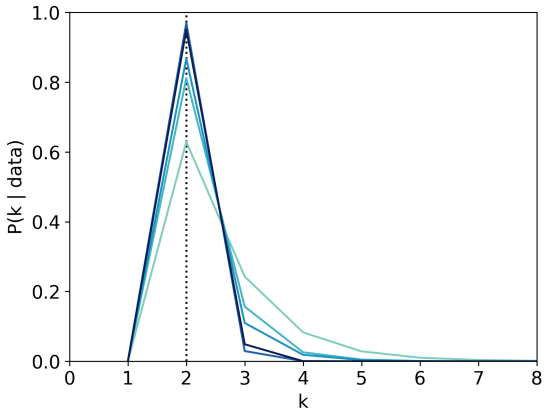
Then the probabilistic generative model is

$$
\begin{aligned}
k &\sim \text{Geom}(r) & p &\sim \text{Dirichlet}_k(\gamma, \ldots, \gamma) \\
\tau_{jd} &\overset{\text{i.i.d.}}{\sim} \text{Gam}(\alpha, \beta) & \theta_{jd} &\overset{\text{i.i.d.}}{\sim} \mathcal{N}(m, \kappa_{jd}^{-1}) \\
Z_n &\overset{\text{i.i.d.}}{\sim} \text{Categorical}(p) & X_{nd} &\overset{\text{ind}}{\sim} \mathcal{N}(\theta_{z_n d}, \tau_{z_n d}^{-1}),
\end{aligned}
$$

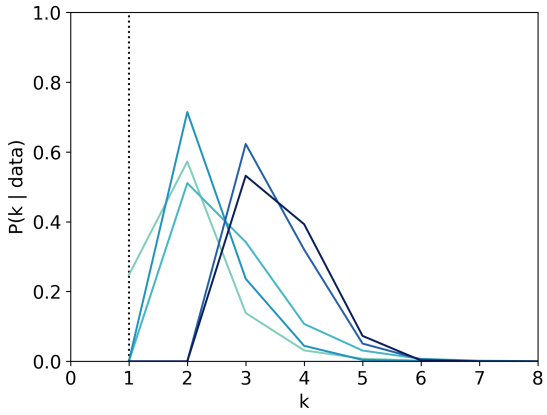
where j ranges from $1, \ldots, k$, d ranges from $1, \ldots, D$, and n ranges from $1, \ldots, N$. For posterior inference, we use a Julia implementation of split-merge collapsed Gibbs sampling (Jain and Neal,



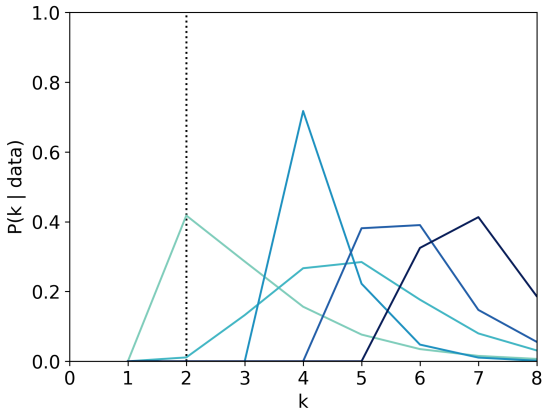
(a) Gaussian data, 1 component



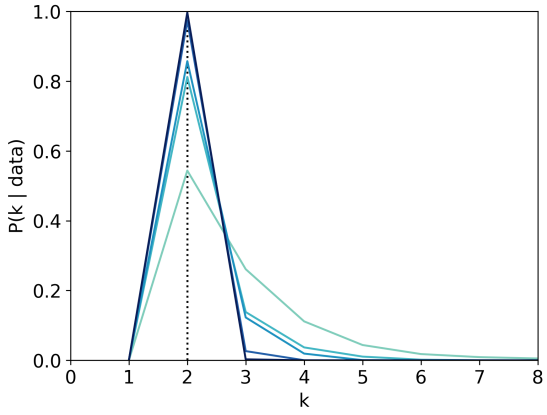
(b) Gaussian data, 2 components



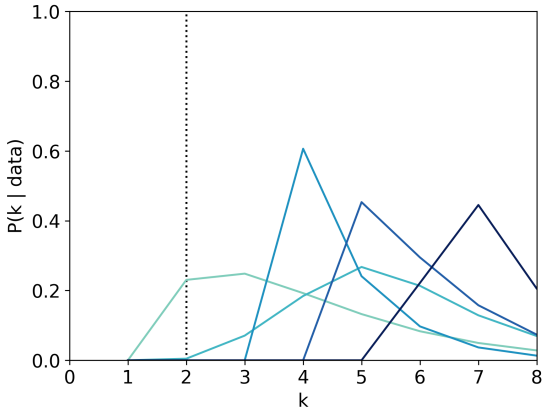
(c) Laplace data, 1 component



(d) Laplace data, 2 components



(e) Gaussian data, varying prior



(f) Laplace data, varying prior

Figure 3.8.1: *Upper and middle rows*: Posterior probability of the number of components k for Gaussian mixture models with a fixed prior fit to (a,b) univariate data generated from a Gaussian mixture model and (c,d) a Laplace mixture model, *Lower row*: Posterior probability of the number of components of Gaussian mixtures with a varying prior fit to (e) 2-component univariate data from a Gaussian mixture model and (f) 2-component univariate data from a Laplace mixture model.

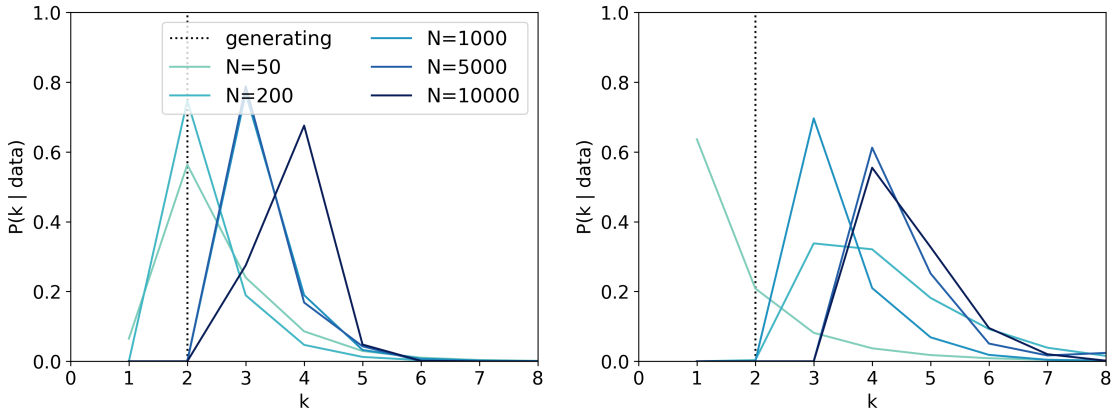
2004; Neal, 2000) from Miller and Harrison (2018).[1] The model and inference algorithm are described in more detail in Miller and Harrison (2018, Sec. 7.2.2, Algorithm 1). Note that we use this model primarily to illustrate the problem of posterior divergence under model misspecification; it should not be interpreted as a carefully-specified model for the data examples that we study. Also note that while the empirical examples below involve Gaussian FMMs, our theory applies to a more general class of component distributions.

3.8.1 Synthetic data

Gaussian and Laplace mixtures Our first experiments on synthetic data are inspired by Figure 3 of Miller and Dunson (2019), which investigates the posterior of a mixture of perturbed Gaussians. Here we study the effects of varying data set sizes under both well-specified and misspecified models. We generated data sets of increasing size $N \in \{50, 200, 1000, 5000, 10000\}$ from 1- and 2-component univariate Gaussian and Laplace mixture models, where the 1-component distributions have mean 0 and scale 1, and the 2-component distributions have means $(-5, 5)$, scales $(1.5, 1)$, and mixing weights $(0.4, 0.6)$. We generated the sequence of data sets such that each was a subset of the next, larger data set in the sequence. Following Miller and Harrison (2018, Section 7.2.1), we set the hyperparameters of the Bayesian finite mixture model as follows: $m = \frac{1}{2}(\max_{n\in[\tilde{N}]} X_n + \min_{n\in[\tilde{N}]} X_n)$ where $\tilde{N} = 10{,}000$, $\kappa = (\max_{n\in[\tilde{N}]} X_n - \min_{n\in[\tilde{N}]} X_n)^{-2}$, $\alpha = 2$, $r = 0.1$, $\gamma = 1$, and $\beta \sim \text{Gam}(0.2, 10/\kappa)$. We refer to Miller and Harrison (2018, Section 7.2.1) for additional details on the choice of model hyperparameters and the sampling of β. We ran a total of 100,000 Markov chain Monte Carlo iterations per data set; we discarded the first 10,000 iterations as burn-in.

The results of the simulations are shown in Figure 3.8.1. For the data generated from the 1-component models, the posterior on the number of components concentrates around 1 in the case of Gaussian-generated data as the sample size increases (Figure 3.8.1a), whereas the posterior on the number of components diverges for the Laplace data (Figure 3.8.1c). We observe similar behavior in the 2-component case, where the posterior concentrates around the correct value in the Gaussian

[1] Code available at https://github.com/jwmi/BayesianMixtures.jl.



(a) Data generated with $\epsilon = 0.01$

(b) Data generated with $\epsilon = 0.1$

Figure 3.8.2: Posterior probability of the number of components k for Gaussian mixture models with a fixed prior fit to data generated from an ϵ-contaminated 2-component Gaussian mixture model, where ϵ is the proportion of data generated from a Laplace distribution.

case (Figure 3.8.1b) but not the Laplace case (Figure 3.8.1d).

Priors that vary with N Next, we considered the same finite Gaussian mixture model described above but with a prior that varies with the data. Specifically, for the prior on the means, we set the hyperparameters to $m_N = \frac{1}{2}(\max_{n\in[N]} X_n + \min_{n\in[N]} X_n)$ and $\kappa_N = (\max_{n\in[N]} X_n - \min_{n\in[N]} X_n)^{-2}$, which is the setting considered by Miller and Harrison (2018, Section 7.2.1); the other hyperparameters were otherwise set to the same values above. We used the 2-component Gaussian and Laplace data sets constructed above for the fixed prior case. The bottom row of Figure 3.8.1 shows the results of the posterior number of components under this prior for the well-specified and misspecified cases; again we observe that the posterior diverges under model misspecification.

ϵ-contamination Finally, in order to study the posterior number of components under a very slight amount of misspecification, we applied the fixed-prior Gaussian mixture model above to data generated with ϵ-contamination. That is, we generated the data according to the ϵ-contaminated distribution $f_0 = (1-\epsilon)f + \epsilon q$, where f is a 2-component Gaussian mixture distribution with

means $(5, 10)$, variances $(1, 1.5)$, and mixing weights $(0.4, 0.6)$, and q is a Laplace distribution with location 0 and scale 1. We generated two data sets: one with $\epsilon = 0.01$ and one with $\epsilon = 0.1$. In Figure 3.8.2, we observe that even under very small amounts of misspecification, the posterior number of components diverges.

3.8.2 Gene expression data

Computational biologists are interested in classifying cell types by applying clustering techniques to gene expression data (de Souto et al., 2008; McLachlan et al., 2002; McNicholas and Murphy, 2010; Medvedovic and Sivaganesan, 2002; Medvedovic et al., 2004; Rasmussen et al., 2008; Yeung et al., 2001). In our next set of experiments, we apply the Gaussian finite mixture model to two gene expression data sets: (1) single-cell RNA sequencing data from mouse cortex and hippocampus cells (Zeisel et al., 2015) with the same feature selection as Prabhakaran et al. (2016) ($N = 3008$, $D = 558$, 11,000 Gibbs sampling steps with 1,000 of those as burn-in) and (2) mRNA expression data from human lung tissue (Bhattacharjee et al., 2001) ($N = 203, D = 1543$, and 10,000 Gibbs sampling steps with 1,000 of those burn-in). Our experiments here represent a simplified version of previous mixture model analyses for these and other related data sets (Armstrong et al., 2001; de Souto et al., 2008; Miller and Harrison, 2018; Prabhakaran et al., 2016).

As these gene expression data sets contain counts, we first transformed the data to real numerical values. In particular, we used a base-2 log transform followed by standardization—such that each dimension of the data had zero mean and unit variance—per standard practices (e.g., Miller and Harrison (2018)). Then to examine the effect of increasing data set size on inferential results, we randomly sampled subsets of increasing size without replacement; each smaller subset was contained in the next larger data set. For both data sets, we used hyperparameters $\alpha = 1$, $\beta = 1$, $m = 0$, $\kappa_{jd} = \tau_{jd}$, $r = 0.1$, and $\gamma = 1$.

For the single-cell RNAseq data set, the posterior on the number of components is shown in Figure 3.1.1a. Here the ground truth number of clusters is captured when the data set size is $N = 100$. But as predicted by our theory, as we increase the number of data points, the posterior number of components diverges.

The posterior on the number of components for the lung gene expression data is shown in Figure 3.1.1b. Again we find that on the smallest data subsets, the posterior appears to capture the ground truth number of clusters, but that as we examine more and more data, the posterior diverges.

The diagonal covariance Gaussian components are a particularly simple form of cluster shape. But no matter how complex the component model, one could wonder whether an even-more complex model might solve the issue that the number of components diverge. In the typical real-world situation that the component model cannot be specified in absolute perfection, our theory confirms that the divergence problem will remain. Thus, these examples suggest the need for more robust analyses.

3.8.3 Power posterior results

For all experiments below, we use a finite mixture model with a Gaussian component family and a conjugate prior. In particular, consider number of components k, mixture weights $p \in \mathbb{R}^k$, Gaussian component precisions $\tau \in \mathbb{R}_+^k$ and means $\theta \in \mathbb{R}^k$, labels $Z \in \{1, \ldots, k\}^N$, and data $X \in \mathbb{R}^N$. Then the probabilistic generative model is

$$
\begin{aligned}
k &\sim \text{Geom}(r) & w &\sim \text{Dirichlet}_k(\gamma, \ldots, \gamma) \\
\tau_j &\overset{\text{i.i.d.}}{\sim} \text{Gam}(\alpha, \beta) & \theta_j &\overset{\text{i.i.d.}}{\sim} \mathcal{N}(m, \kappa^{-1}) \\
Z_n &\overset{\text{i.i.d.}}{\sim} \text{Categorical}(w) & X_n &\overset{\text{ind}}{\sim} \mathcal{N}(\theta_{z_n}, \tau_{z_n}^{-1}),
\end{aligned}
$$

where j ranges from $1, \ldots, k$, and n ranges from $1, \ldots, N$.

For posterior inference, we used a Gibbs sampler (Miller and Harrison, 2018, Sec. 7.2.2, Algorithm 1), coupled with the approximation described in Miller and Dunson (2019, Section 5). Note that we use this model primarily to illustrate the problem of α-posterior divergence under model misspecification; it should not be interpreted as a carefully-specified model for the data examples that we study. Also note that while the empirical examples below involve Gaussian FMMs, our theory applies to a more general class of component distributions.

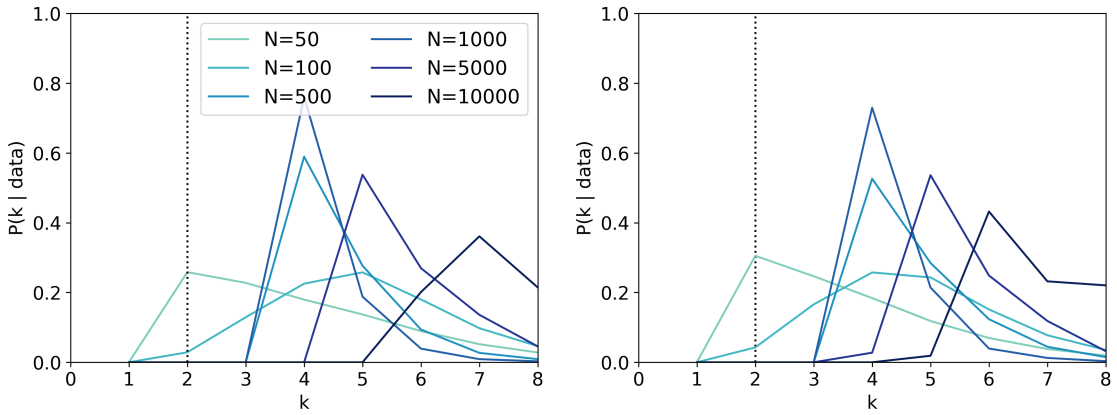
In this section, we consider the behavior of the α-posterior only for fixed $\alpha \in (0, 1)$.

Synthetic mixture data Here we study the effects of varying data set sizes under a misspecified model. We generated data sets of increasing size $N \in \{50, 100, 500, 1000, 5000, 10000\}$ from a 2-component Laplace mixture models, where the 2-component distributions have means $(-5, 5)$, scales $(1.5, 1)$, and mixing weights $(0.4, 0.6)$. We generated the sequence of data sets such that each was a subset of the next, larger data set in the sequence. Following Miller and Harrison (2018, Section 7.2.1), we set the hyperparameters of the Bayesian finite mixture model as follows: $m = \frac{1}{2}(\max_{n\in[\tilde{N}]} X_n + \min_{n\in[\tilde{N}]} X_n)$ where $\tilde{N} = 10{,}000$, $\kappa = (\max_{n\in[\tilde{N}]} X_n - \min_{n\in[\tilde{N}]} X_n)^{-2}$, $\alpha = 2$, $r = 0.1$, $\gamma = 1$, and $\beta = 1$. We ran a total of 150,000 Markov chain Monte Carlo iterations per data set; we discarded the first 50,000 iterations as burn-in.

In Figure 3.8.3, we show the α-posterior number of components resulting from fixed $\alpha = 1, 0.8, 0.5, 0.2$. Note that $\alpha = 1$ is the usual posterior distribution on the number of components. The figures show that as α decreases, the posterior mass tends to shift to smaller numbers of components and also becomes less concentrated. However, as in the usual posterior ($\alpha = 1$), the α-posterior still diverges, though more slowly with lower values of α.

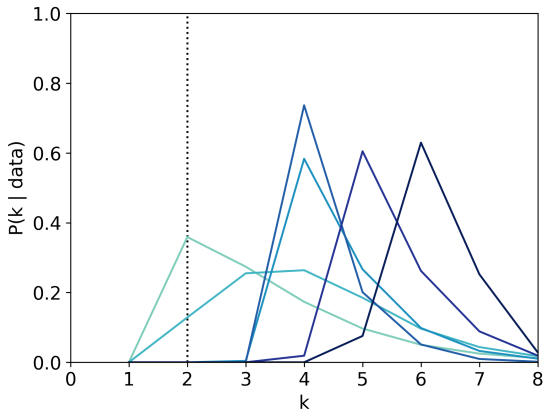
Galaxy mixture data Mixture models are used in astronomy to characterize stellar populations (Nemec and Nemec, 1991), including analysis of star and galaxy clusters. We study the Shapley galaxy data set (Drinkwater et al., 2004), which contains measurements of redshifts (i.e., velocities in km/s) for 4215 galaxies in the Shapley Concentration regions. To examine the effect of increasing data set size on inferential results, we randomly sampled subsets of increasing size without replacement with $N \in \{100, 200, 500, 1000, 2000, 4215\}$; each smaller subset was contained in the next larger data set. We set the hyperparameters of the Bayesian finite mixture model as follows: $m = \frac{1}{2}(\max_{n\in[\tilde{N}]} X_n + \min_{n\in[\tilde{N}]} X_n)$ where $\tilde{N} = 4215$, $\kappa = (\max_{n\in[\tilde{N}]} X_n - \min_{n\in[\tilde{N}]} X_n)^{-2}$, $\alpha = 2$, $r = 0.1$, $\gamma = 1$, and $\beta = 1$. We ran a total of 100,000 Markov chain Monte Carlo iterations per data set; we discarded the first 50,000 iterations as burn-in.

The α-posterior number of components for this model is displayed in Figure 3.8.4. For $\alpha = 1$, we find that as we examine more and more data, the posterior diverges. Similar behavior occurs with

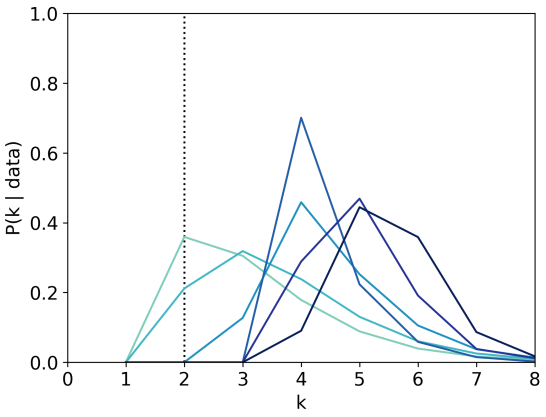


(a) $\alpha = 1$

(b) $\alpha = 0.8$

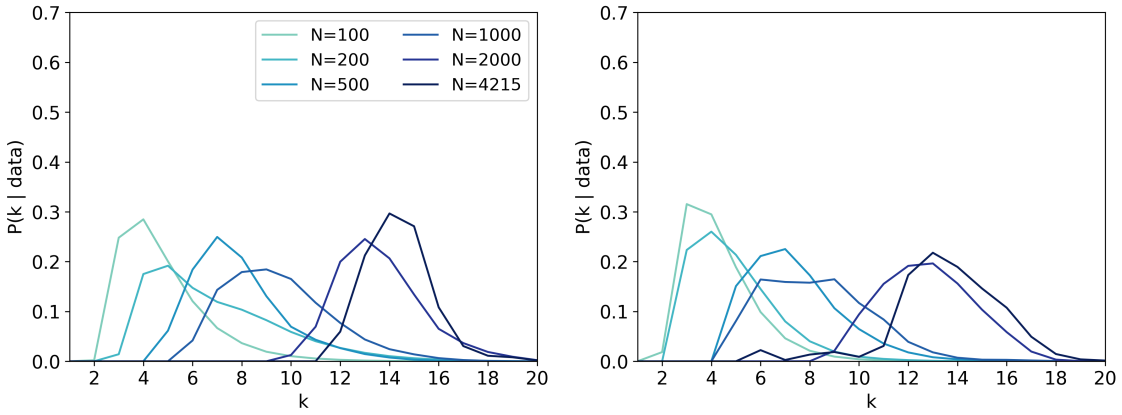


(c) $\alpha = 0.5$



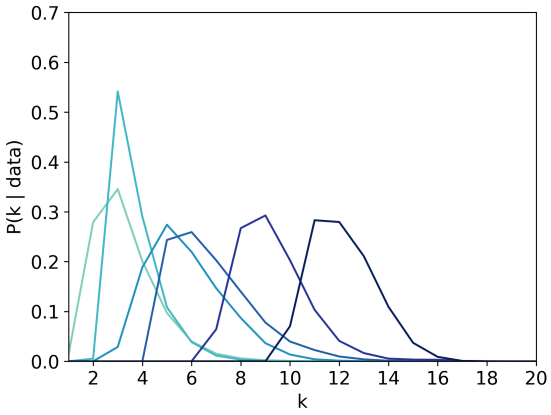
(d) $\alpha = 0.2$

Figure 3.8.3: Synthetic data generated from a 2-component Laplace mixture model. Curves are α-posteriors on number of components (with fixed α) as N varies. The vertical black dotted line denotes the generating number of components.

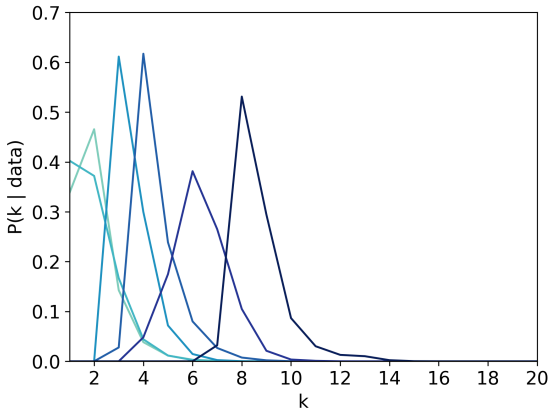


(a) $\alpha = 1$

(b) $\alpha = 0.8$



(c) $\alpha = 0.5$



(d) $\alpha = 0.2$

Figure 3.8.4: Shapley galaxy data. Curves are α-posteriors on the number of components (with fixed α) as N varies.

fractional α-posteriors. With smaller values of α, the α-posterior diverges more slowly.

3.9 Discussion

We have shown that the posterior distribution for the number of components in finite mixtures diverges when the mixture component family is misspecified. Since misspecification is almost unavoidable in real applications, it follows that finite mixture models are typically unreliable for estimating the number of components. In practice, our conclusion implies that inferences on the

number of components can change drastically depending on the size of the data set, calling into question the usefulness of these counts in application.

Since our analysis is inherently asymptotic, it is possible that the Bayesian component-count posterior may still provide useful inferences for a finite sample—for instance if care is taken to account for the dependence of inferential conclusions on data set size. Indeed, a number of authors have recently proposed robust Bayesian inference methods to mitigate likelihood misspecification (Bissiri et al., 2016; Grünwald and van Ommen, 2017; Holmes and Walker, 2017; Huggins and Miller, 2019; Jewson et al., 2018; Knoblauch et al., 2019; Miller and Dunson, 2019; Rigon et al., 2023; Rodriguez and Dunson, 2011; Wang et al., 2017; Woo and Sriram, 2006, 2007); it remains to better understand connections between our results and these methods.

Chapter 4

Edge-exchangeable graphs and sparsity

Many popular network models rely on the assumption of *(vertex) exchangeability*, in which the distribution of the graph is invariant to relabelings of the vertices. However, the Aldous-Hoover theorem guarantees that these graphs are dense or empty with probability one, whereas many real-world graphs are sparse. We present an alternative notion of exchangeability for random graphs, which we call *edge exchangeability*, in which the distribution of a graph sequence is invariant to the order of the edges. We demonstrate that edge-exchangeable models, unlike models that are traditionally vertex exchangeable, can exhibit sparsity. To do so, we outline a general framework for graph generative models; by contrast to the pioneering work of Caron and Fox (2017), models within our framework are stationary across steps of the graph sequence. In particular, our model grows the graph by instantiating more latent atoms of a single random measure as the dataset size increases, rather than adding new atoms to the measure.

4.1 Introduction

In recent years, network data have appeared in a growing number of applications, such as online social networks, biological networks, and networks representing communication patterns. As a result, there is growing interest in developing models for such data and studying their properties. Crucially, individual network data sets also continue to increase in size; we typically assume that the number of vertices is unbounded as time progresses. We say a graph sequence is *dense* if the number of edges grows quadratically in the number of vertices, and a graph sequence is *sparse* if the number of edges grows sub-quadratically as a function of the number of vertices. Sparse graph sequences are more representative of real-world graph behavior. However, many popular network models (see, e.g., Lloyd et al. (2012) for an extensive list) share the undesirable scaling property that they yield dense sequences of graphs with probability one. The poor scaling properties of these models can be traced back to a seemingly innocent assumption: that the vertices in the model are *exchangeable*, that is, any finite permutation of the rows and columns of the graph adjacency matrix does not change the distribution of the graph. Under this assumption, the Aldous-Hoover theorem (Aldous, 1981; Hoover, 1979) implies that such models generate dense or empty graphs with probability one (Orbanz and Roy, 2015).

This fundamental model misspecification motivates the development of new models that can achieve sparsity. One recent focus has been on models in which an additional parameter is employed to uniformly decrease the probabilities of edges as the network grows (e.g., Bollobás et al. (2007); Borgs et al. (2019, 2021); Wolfe and Olhede (2013)). While these models allow sparse graph sequences, the sequences are no longer *projective*. In projective sequences, vertices and edges are added to a graph as a graph sequence progresses—whereas in the models above, there is not generally any strict subgraph relationship between earlier graphs and later graphs in the sequence. Projectivity is natural in streaming modeling. For instance, we may wish to capture new users joining a social network and new connections being made among existing users—or new employees joining a company and new communications between existing employees.

Caron and Fox (2017) have pioneered initial work on sparse, projective graph sequences. Instead

of the *vertex exchangeability* that yields the Aldous-Hoover theorem, they consider a notion of graph exchangeability based on the idea of independent increments of subordinators (Kallenberg, 2005), explored in depth by Veitch and Roy (2015). However, since this Kallenberg-style exchangeability introduces a new countable infinity of latent vertices at every step in the graph sequence, its generative mechanism seems particularly suited to the non-stationary domain. By contrast, we are here interested in exploring *stationary* models that grow in complexity with the size of the data set. Consider classic Bayesian nonparametric models as the Chinese restaurant process (CRP) and Indian buffet process (IBP); these engender growth by using a single infinite latent collection of parameters to generate a finite but growing set of instantiated parameters. Similarly, we propose a framework that uses a single infinite latent collection of vertices to generate a finite but growing set of vertices that participate in edges and thereby in the network. We believe our framework will be a useful component in more complex, non-stationary graphical models—just as the CRP and IBP are often combined with hidden Markov models or other explicit non-stationary mechanisms. Additionally, Kallenberg exchangeability is intimately tied to continuous-valued labels of the vertices, and here we are interested in providing a characterization of the graph sequence based solely on its topology.

In this work, we introduce a new form of exchangeability, distinct from both vertex exchangeability and Kallenberg exchangeability. In particular, we say that a graph sequence is *edge exchangeable* if the distribution of any graph in the sequence is invariant to the *order* in which edges arrive—rather than the order of the vertices. We will demonstrate that edge exchangeability admits a large family of sparse, projective graph sequences.

In the remainder of the chapter, we start by defining dense and sparse graph sequences rigorously. We review vertex exchangeability before introducing our new notion of edge exchangeability in Section 4.2, which we also contrast with Kallenberg exchangeability in more detail in Section 4.4. We define a family of models, which we call *graph frequency models*, based on random measures in Section 4.3. We use these models to show that edge-exchangeable models can yield sparse, projective graph sequences via theoretical analysis in Section 4.5 and via simulations in Section 4.6. Along the way, we highlight other benefits of the edge exchangeability and graph frequency model frameworks.

4.2 Exchangeability in graphs: old and new

Let $(G_n)_n := G_1, G_2, \ldots$ be a sequence of graphs, where each graph $G_n = (V_n, E_n)$ consists of a (finite) set of vertices V_n and a (finite) multiset of edges E_n. Each edge $e \in E_n$ is a set of two vertices in V_n. We assume the sequence is *projective*—or growing—so that $V_n \subseteq V_{n+1}$ and $E_n \subseteq E_{n+1}$. Consider, e.g., a social network with more users joining the network and making new connections with existing users. We say that a graph sequence is *dense* if $|E_n| = \Omega(|V_n|^2)$, i.e., the number of edges is asymptotically lower bounded by $c \cdot |V_n|^2$ for some constant c. Conversely, a sequence is *sparse* if $|E_n| = o(|V_n|^2)$, i.e., the number of edges is asymptotically upper bounded by $c \cdot |V_n|^2$ for all constants c. In what follows, we consider random graph sequences, and we focus on the case where $|V_n| \to \infty$ almost surely.

4.2.1 Vertex-exchangeable graph sequences

If the number of vertices in the graph sequence grows to infinity, the graphs in the sequence can be thought of as subgraphs of an "infinite" graph with infinitely many vertices and a correspondingly infinite adjacency matrix. Traditionally, exchangeability in random graphs is defined as the invariance of the distribution of any finite submatrix of this adjacency matrix—corresponding to any finite collection of vertices—under finite permutation. Equivalently, we can express this form of exchangeability, which we henceforth call *vertex exchangeability*, by considering a random sequence of graphs $(G_n)_n$ with $V_n = [n]$, where $[n] := \{1, \ldots, n\}$. In this case, only the edge sequence is random. Let π be any permutation of the integers $[n]$. If $e = \{v, w\}$, let $\pi(e) := \{\pi(v), \pi(w)\}$. If $E_n = \{e_1, \ldots, e_m\}$, let $\pi(E_n) := \{\pi(e_1), \ldots, \pi(e_m)\}$.

Definition 4.2.1. *Consider the random graph sequence $(G_n)_n$, where G_n has vertices $V_n = [n]$ and edges E_n. $(G_n)_n$ is (infinitely)* vertex exchangeable *if for every $n \in \mathbb{N}$ and for every permutation π of the vertices $[n]$, $G_n \overset{d}{=} \tilde{G}_n$, where $\tilde{G}_n$ has vertices $[n]$ and edges $\pi(E_n)$.*

A great many popular models for graphs are vertex exchangeable; see Appendix C.2 and Lloyd et al. (2012) for a list. However, it follows from the Aldous-Hoover theorem (Aldous, 1981; Hoover, 1979) that any vertex-exchangeable graph is a mixture of sampling procedures from *graphons*.

Further, any graph sampled from a graphon is almost surely dense or empty (Orbanz and Roy, 2015). Thus, vertex-exchangeable random graph models are misspecified models for sparse network datasets, as they generate dense graphs.

4.2.2 Edge-exchangeable graph sequences

Vertex-exchangeable sequences have distributions invariant to the order of vertex arrival. We introduce *edge-exchangeable* graph sequences, which will instead be invariant to the order of edge arrival. As before, we let $G_n = (V_n, E_n)$ be the nth graph in the sequence. Here, though, we consider only *active vertices*—that is, vertices that are connected via some edge. That lets us define V_n as a function of E_n; namely, V_n is the union of the vertices in E_n. Note that a graph that has sub-quadratic growth in the number of edges as a function of the number of active vertices will necessarily have sub-quadratic growth in the number of edges as a function of the number of all vertices, so we obtain strictly stronger results by considering active vertices. In this case, the graph G_n is completely defined by its edge set E_n.

As above, we suppose that $E_n \subseteq E_{n+1}$. We can emphasize this projectivity property by augmenting each edge with the step on which it is added to the sequence. Let E'_n be a collection of tuples, in which the first element is the edge and the second element is the step (i.e., index) on which the edge is added: $E'_n = \{(e_1, s_1), \ldots, (e_m, s_m)\}$. We can then define a *step-augmented graph sequence* $(E'_n)_n = (E'_1, E'_2, \ldots)$ as a sequence of step-augmented edge sets. Note that there is a bijection between the step-augmented graph sequence and the original graph sequence.

Example 4.2.2. *In the setup for vertex exchangeability, we assumed $V_n = [n]$ and every edge is introduced as soon as both of its vertices are introduced. In this case, the step of any edge in the step-augmented graph is the maximum vertex value. For example, in Figure 4.2.1, we have*

$$E'_1 = \emptyset, E'_2 = E'_3 = \{(\{1,2\},2)\}, E'_4 = \{(\{1,2\},2), (\{1,4\},4), (\{2,4\},4), (\{3,4\},4)\}.$$

In general step-augmented graphs, though, the step need not equal the max vertex, as we see next.

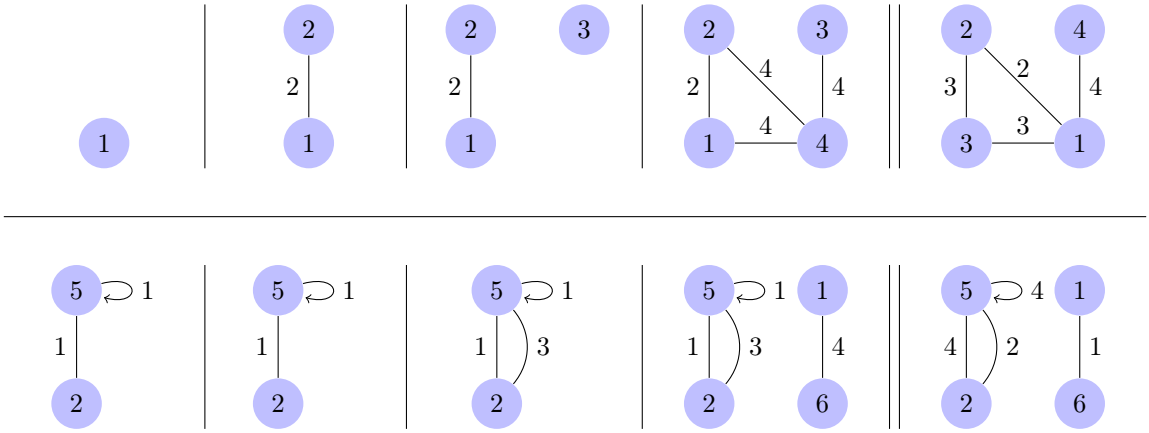


Figure 4.2.1: *Upper, left four*: Step-augmented graph sequence from Ex. 4.2.2. At each step n, the step value is always at least the maximum vertex index. *Upper, right two*: Two graphs with the same probability under vertex exchangeability. *Lower, left four*: Step-augmented graph sequence from Ex. 4.2.3. *Lower, right two*: Two graphs with the same probability under edge exchangeability.

Example 4.2.3. *Suppose we have a graph given by the edge sequence (see Figure 4.2.1):*

$$E_1 = E_2 = \{\{2,5\},\{5,5\}\}, E_3 = E_2 \cup \{\{2,5\}\}, E_4 = E_3 \cup \{\{1,6\}\}.$$

The step-augmented graph E_4' *is* $\{(\{2,5\},1),(\{5,5\},1),(\{2,5\},3),(\{1,6\},4)\}$.

Roughly, a random graph sequence is edge exchangeable if its distribution is invariant to finite permutations of the steps. Let π be a permutation of the integers $[n]$. For a step-augmented edge set $E_n' = \{(e_1,s_1),\ldots,(e_m,s_m)\}$, let $\pi(E_n') = \{(e_1,\pi(s_1)),\ldots,(e_m,\pi(s_m))\}$.

Definition 4.2.4. *Consider the random graph sequence* $(G_n)_n$*, where* G_n *has step-augmented edges* E_n' *and* V_n *are the active vertices of* E_n*.* $(G_n)_n$ *is (infinitely)* edge exchangeable *if for every* $n \in \mathbb{N}$ *and for every permutation* π *of the steps* $[n]$*,* $G_n \overset{d}{=} \tilde{G}_n$*, where* $\tilde{G}_n$ *has step-augmented edges* $\pi(E_n')$ *and associated active vertices.*

See Figure 4.2.1 for visualizations of both vertex exchangeability and edge exchangeability. It remains to show that there are non-trivial models that are edge exchangeable (Section 4.3) and that edge-exchangeable models admit sparse graphs (Section 4.5).

4.3 Edge-exchangeable graph frequency models

We next demonstrate that a wide class of models, which we call *graph frequency models*, exhibit edge exchangeability. Consider a latent infinity of vertices indexed by the positive integers $\mathbb{N} = \{1, 2, \ldots\}$, along with an infinity of edge labels $(\theta_{\{i,j\}})$, each in a set Θ, and positive edge rates (or frequencies) $(w_{\{i,j\}})$ in $\mathbb{R}_+$. We allow both the $(\theta_{\{i,j\}})$ and $(w_{\{i,j\}})$ to be random, though this is not mandatory. For instance, we might choose $\theta_{\{i,j\}} = (i, j)$ for $i \leq j$, and $\Theta = \mathbb{R}^2$. Alternatively, the $\theta_{\{i,j\}}$ could be drawn iid from a continuous distribution such as Unif$[0, 1]$. For any choice of $(\theta_{\{i,j\}})$ and $(w_{\{i,j\}})$,

$$W := \sum_{\{i,j\}:i,j\in\mathbb{N}} w_{\{i,j\}}\delta_{\theta_{\{i,j\}}} \tag{4.1}$$

is a *measure* on Θ. Moreover, it is a discrete measure since it is always atomic. If either $(\theta_{\{i,j\}})$ or $(w_{\{i,j\}})$ (or both) are random, W is a *discrete random measure* on Θ since it is a random, discrete-measure-valued element. Given the edge rates (or frequencies) $(w_{\{i,j\}})$ in W, we next show some natural ways to construct edge-exchangeable graphs.

Single edge per step If the rates $(w_{\{i,j\}})$ are normalized such that $\sum_{\{i,j\}:i,j\in\mathbb{N}} w_{\{i,j\}} = 1$, then $(w_{\{i,j\}})$ is a distribution over all possible vertex pairs. In other words, W is a probability measure. We can form an edge-exchangeable graph sequence by first drawing values for $(w_{\{i,j\}})$ and $(\theta_{\{i,j\}})$—and setting $E_0 = \emptyset$. We recursively set $E_{n+1} = E_n \cup \{e\}$, where e is an edge $\{i, j\}$ chosen from the distribution $(w_{\{i,j\}})$. This construction introduces a single edge in the graph each step, although it may be a duplicate of an edge that already exists. Therefore, this technique generates multigraphs one edge at a time. Since the edge every step is drawn conditionally iid given W, we have an edge-exchangeable graph.

Multiple edges per step Alternatively, the rates $(w_{\{i,j\}})$ may not be normalized. Then W may not be a probability measure. Let $f(m|w)$ be a distribution over non-negative integers m given some rate $w \in \mathbb{R}_+$. We again initialize our sequence by drawing $(w_{\{i,j\}})$ and $(\theta_{\{i,j\}})$ and setting $E_0 = \emptyset$. In this case, recursively, on the nth step, start by setting $F = \emptyset$. For every possible edge $e = \{i, j\}$,

we draw the multiplicity of the edge e in this step as $m_e \overset{\text{ind}}{\sim} f(\cdot|w_e)$ and add m_e copies of edge e to F. Finally, $E_{n+1} = E_n \cup F$. This technique potentially introduces multiple edges in each step, in which edges themselves may have multiplicity greater than one and may be duplicates of edges that already exist in the graph. Therefore, this technique generates multigraphs, multiple edges at a time. If we restrict f and W such that finitely many edges are added on every step almost surely, we have an edge-exchangeable graph, as the edges in each step are drawn conditionally iid given W.

Given a sequence of edge sets $E_0, E_1, \ldots$ constructed via either of the above methods, we can form a binary graph sequence $\bar{E}_0, \bar{E}_1, \ldots$ by setting $\bar{E}_i$ to have the same edges as E_i except with multiplicity 1. Although this binary graph is not itself edge exchangeable, it inherits many of the properties (such as sparsity, as shown in Section 4.5) of the underlying edge-exchangeable multigraph.

The choice of the distribution on the measure W has a strong influence on the properties of the resulting edge-exchangeable graph sampled via one of the above methods. For example, one choice is to set $w_{\{i,j\}} = w_i w_j$, where the $(w_i)_i$ are a countable infinity of random values generated according to a *Poisson point process* (PPP). We say that $(w_i)_i$ is distributed according to a Poisson point process parameterized by rate measure ν, $(w_i)_i \sim \text{PPP}(\nu)$, if (a) $\#\{i : w_i \in A\} \sim \text{Poisson}(\nu(A))$ for any set A with finite measure $\nu(A)$ and (b) $\#\{i : w_i \in A_j\}$ are independent random variables across any finite collection of disjoint sets $(A_j)_{j=1}^J$. In Section 4.5 we examine a particular example of this graph frequency model, and demonstrate that sparsity is possible in edge-exchangeable graphs.

4.4 Related work and connection to nonparametric Bayes

Given a unique label θ_i for each vertex $i \in \mathbb{N}$, and denoting $g_{ij} = g_{ji}$ to be the number of undirected edges between vertices i and j, the graph itself can be represented as the discrete random measure $G = \sum_{i,j} g_{ij} \delta_{(\theta_i, \theta_j)}$ on $\mathbb{R}_+^2$. A different notion of exchangeability for graphs than the ones in Section 4.2 can be phrased for such atomic random measures: a point process G on $\mathbb{R}_+^2$ is (jointly) exchangeable if, for all finite permutations π of $\mathbb{N}$ and all $h > 0$,

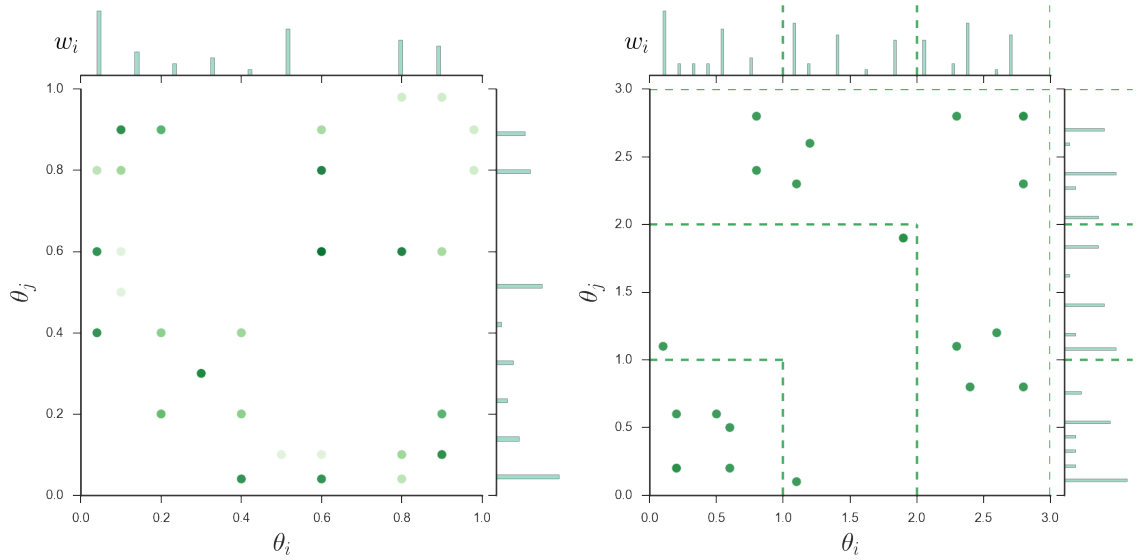
$$G(A_i \times A_j) \overset{\text{d}}{=} G(A_{\pi(i)} \times A_{\pi(j)}), \text{ for } (i,j) \in \mathbb{N}^2, \qquad \text{where } A_i := [h \cdot (i-1), h \cdot i].$$

This form of exchangeability, which we refer to as *Kallenberg exchangeability*, can intuitively be viewed as invariance of the graph distribution to relabeling of the vertices, which are now embedded in $\mathbb{R}_+^2$. As such it is analogous to vertex exchangeability, but for discrete random measures (Caron and Fox, 2017, Sec. 4.1). Exchangeability for random measures was introduced by Aldous (Aldous, 1985), and a representation theorem was given by Kallenberg (Kallenberg, 1990, 2005, Ch. 9). The use of Kallenberg exchangeability for modeling graphs was first proposed by Caron and Fox (2017), and then characterized in greater generality by Veitch and Roy (2015) and Borgs et al. (2018). Edge exchangeability is distinct from Kallenberg exchangeability, as shown by the following example.

Example 4.4.1 (Edge exchangeable but not Kallenberg exchangeable)**.** *Consider the graph frequency model developed in Section 4.3, with* $w_{\{i,j\}} = (ij)^{-2}$ *and* $\theta_{\{i,j\}} = \{i, j\}$*. Since the edges at each step are drawn iid given* W*, the graph sequence is edge exchangeable. However, the corresponding graph measure* $G = \sum_{i,j} n_{ij}\delta_{(i,j)}$ *(where* $n_{ij} = n_{ji} \sim \text{Binom}(N, (ij)^{-2})$*) is not Kallenberg exchangeable, since the probability of generating edge* $\{i, j\}$ *is directly related to the positions* (i, j) *and* (j, i) *in* $\mathbb{R}_+^2$ *of the corresponding atoms in* G *(in particular, the probability is decreasing in* ij*).*

Our graph frequency model is reminiscent of the Caron and Fox (2017) generative model, but has a number of key differences. At a high level, this earlier model generates a weight measure $W = \sum_{i,j} w_{ij}\delta_{(\theta_i,\theta_j)}$ (Caron and Fox (2017) used, in particular, the outer product of a completely random measure), and the graph measure G is constructed by sampling g_{ij} *once* given w_{ij} for each pair i, j. To create a finite graph, the graph measure G is restricted to the subset $[0, y] \times [0, y] \subset \mathbb{R}_+^2$ for $0 < y < \infty$; to create a projective growing graph sequence, the value of y is increased. By contrast, in the analogous graph frequency model of the present work, y is fixed, and we grow the network by *repeatedly* sampling the number of edges g_{ij} between vertices i and j and summing the result. Thus, in the Caron and Fox (2017) model, a latent infinity of vertices (only finitely many of which are active) are added to the network each time y increases. In our graph frequency model, there is a single collection of latent vertices, which are all gradually activated by increasing the number of samples that generate edges between the vertices. See Figure 4.4.1 for an illustration.

Increasing n in the graph frequency model has the interpretation of both (a) time passing and (b)



(a) Graph frequency model (fixed y, n steps) (b) Caron–Fox, PPP on $[0, y] \times [0, y]$ (1 step, y grows)

Figure 4.4.1: A comparison of a graph frequency model (Section 4.3 and Equation (4.2)) and the generative model of Caron and Fox (2017). Any interval $[0, y]$ contains a countably infinite number of atoms with a nonzero weight in the random measure; a draw from the random measure is plotted at the top (and repeated on the right side). Each atom corresponds to a latent vertex. Each point (θ_i, θ_j) corresponds to a latent edge. Darker point colors on the left occur for greater edge multiplicities. On the *left*, more latent edges are instantiated as more steps n are taken. On the *right*, the edges within $[0, y]^2$ are fixed, but more edges are instantiated as y grows.

new individuals joining a network because they have formed a connection that was not previously there. In particular, only latent individuals that will eventually join the network are considered. This behavior is analogous to the well-known behavior of other nonparametric Bayesian models such as, e.g., a Chinese restaurant process (CRP). In this analogy, the Dirichlet process (DP) corresponds to our graph frequency model, and the clusters instantiated by the CRP correspond to the vertices that are active after n steps. In the DP, only latent clusters that will eventually appear in the data are modeled. Since the graph frequency setting is stationary like the DP/CRP, it may be more straightforward to develop approximate Bayesian inference algorithms, e.g., via truncation (Campbell et al., 2016).

Edge exchangeability first appeared in work by Crane and Dempsey (2015a,b); Williamson (2016), and Broderick and Cai (2015a,b); Cai and Broderick (2015). Broderick and Cai (2015a,b) established

the notion of edge exchangeability used here and provided characterizations via exchangeable partitions and feature allocations, as in Appendix C.3. Broderick and Cai (2015a); Cai and Broderick (2015) developed a frequency model based on weights $(w_i)_i$ generated from a Poisson process and studied several types of power laws in the model. Crane and Dempsey (2015a) established a similar notion of edge exchangeability in the context of a larger statistical modeling framework. Crane and Dempsey (2015a,b) provided sparsity and power law results for the case where the weights $(w_i)_i$ are generated from a Pitman-Yor process and power law degree distribution simulations. Williamson (2016) described a similar notion of edge exchangeability and developed an edge-exchangeable model where the weights $(w_i)_i$ are generated from a Dirichlet process, a mixture model extension, and an efficient Bayesian inference procedure. In work concurrent to the present chapter, Crane and Dempsey (2016) re-examined edge exchangeability, provided a representation theorem, and studied sparsity and power laws for the same model based on Pitman-Yor weights. By contrast, we here obtain sparsity results across all Poisson point process-based graph frequency models of the form in Equation (4.2) below, and use a specific three-parameter beta process rate measure only for simulations in Section 4.6.

4.5 Sparsity in Poisson process graph frequency models

We now demonstrate that, unlike vertex exchangeability, edge exchangeability allows for sparsity in random graph sequences. We develop a class of sparse, edge-exchangeable multigraph sequences via the Poisson point process construction introduced in Section 4.3, along with their binary restrictions.

Model Let $\mathcal{W}$ be a Poisson process on $[0, 1]$ with a nonatomic, σ-finite rate measure ν satisfying $\nu([0, 1]) = \infty$ and $\int_0^1 w\nu(\mathrm{d}w) < \infty$. These two conditions on ν guarantee that $\mathcal{W}$ is a countably infinite collection of rates in $[0, 1]$ and that $\sum_{w\in\mathcal{W}} w < \infty$ almost surely. We can use $\mathcal{W}$ to construct the set of rates: $w_{\{i,j\}} = w_i w_j$ if $i \neq j$, and $w_{\{i,i\}} = 0$. The edge labels $\theta_{\{i,j\}}$ are unimportant in characterizing sparsity, and so can be ignored.

To use the multiple-edges-per-step graph frequency model from Section 4.3, we let $f(\cdot|w)$ be

Bernoulli with probability w. Since edge $\{i,j\}$ is added in each step with probability $w_i w_j$, its multiplicity $M_{\{i,j\}}$ after n steps has a binomial distribution with parameters $n, w_i w_j$. Note that self-loops are avoided by setting $w_{\{i,i\}} = 0$. Therefore, the graph after n steps is described by:

$$\mathcal{W} \sim \mathrm{PPP}(\nu) \qquad\qquad M_{\{i,j\}} \overset{\text{ind}}{\sim} \mathrm{Binom}(n, w_i w_j) \;\text{ for } i < j \in \mathbb{N}. \tag{4.2}$$

As mentioned earlier, this generative model yields an edge-exchangeable graph, with edge multiset E_n containing $\{i,j\}$ with multiplicity $M_{\{i,j\}}$, and active vertices $V_n = \{i : \sum_j M_{\{i,j\}} > 0\}$. Although this model generates multigraphs, it can be modified to sample a binary graph $(\bar{V}_n, \bar{E}_n)$ by setting $\bar{V}_n = V_n$ and $\bar{E}_n$ to the set of edges $\{i,j\}$ such that $\{i,j\}$ has multiplicity ≥ 1 in E_n. We can express the number of vertices and edges, in the multi- and binary graphs respectively, as

$$|\bar{V}_n| = |V_n| = \sum_i \mathbb{1}\left(\sum_{j \neq i} M_{\{i,j\}} > 0\right), \quad |E_n| = \frac{1}{2}\sum_{i \neq j} M_{\{i,j\}}, \quad |\bar{E}_n| = \frac{1}{2}\sum_{i \neq j} \mathbb{1}\left(M_{\{i,j\}} > 0\right).$$

Moments Recall that a sequence of graphs is considered *sparse* if $|E_n| = o(|V_n|^2)$. Thus, sparsity in the present setting is an *asymptotic* property of a random graph sequence. Rather than consider the asymptotics of the (dependent) random sequences $|E_n|$ and $|V_n|$ in concert, Definition 4.5.1 allows us to consider the asymptotics of their first moments, which are deterministic sequences and can be analyzed separately. We use $\sim$ to denote asymptotic equivalence, i.e., $a_n \sim b_n \iff \lim_{n\to\infty} \frac{a_n}{b_n} = 1$. For details on our asymptotic notation and proofs for this section, see Appendix C.4.

Lemma 4.5.1. *The number of vertices and edges for both the multi- and binary graphs satisfy*

$$|\bar{V}_n| = |V_n| \overset{a.s.}{\sim} \mathbb{E}\left(|V_n|\right), \qquad |E_n| \overset{a.s.}{\sim} \mathbb{E}\left(|E_n|\right), \qquad |\bar{E}_n| \overset{a.s.}{\sim} \mathbb{E}\left(|\bar{E}_n|\right), \qquad n \to \infty.$$

Thus, we can examine the asymptotic behavior of the random numbers of edges and vertices by examining the asymptotic behavior of their expectations, which are provided by Definition 4.5.2.

Lemma 4.5.2. *The expected numbers of vertices and edges for the multi- and binary graphs are*

$$\mathbb{E}\left(|\bar{V}_n|\right) = \mathbb{E}\left(|V_n|\right) = \int \left[1 - \exp\left(-\int (1-(1-wv)^n)\nu(\mathrm{d}v)\right)\right]\nu(\mathrm{d}w),$$
$$\mathbb{E}\left(|E_n|\right) = \frac{n}{2}\iint wv\,\nu(\mathrm{d}w)\nu(\mathrm{d}v), \qquad \mathbb{E}\left(|\bar{E}_n|\right) = \frac{1}{2}\iint (1-(1-wv)^n)\,\nu(\mathrm{d}w)\nu(\mathrm{d}v).$$

Sparsity We are now equipped to characterize the sparsity of this random graph sequence:

Theorem 4.5.3. *Suppose ν has a regularly varying tail, i.e., there exist $\alpha \in (0,1)$ and $\ell : \mathbb{R}_+ \to \mathbb{R}_+$ s.t.*

$$\int_x^1 \nu(\mathrm{d}w) \sim x^{-\alpha}\ell(x^{-1}), \quad x \to 0 \qquad \textit{and} \qquad \forall c > 0,\ \lim_{x\to\infty}\frac{\ell(cx)}{\ell(x)} = 1.$$

Then as $n \to \infty$,

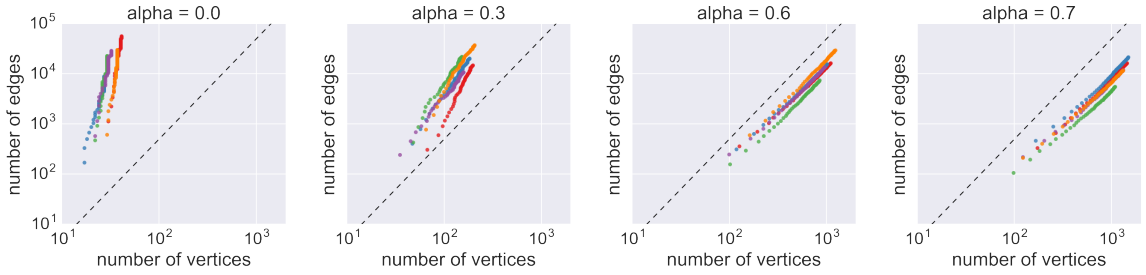
$$|V_n| \overset{a.s.}{=} \Theta(n^\alpha \ell(n)), \qquad |E_n| \overset{a.s.}{=} \Theta(n), \qquad |\bar{E}_n| \overset{a.s.}{=} O\left(\ell(n^{1/2})\min\left(n^{\frac{1+\alpha}{2}}, \ell(n)n^{\frac{3\alpha}{2}}\right)\right).$$

Definition 4.5.3 implies that the multigraph is sparse when $\alpha \in (1/2, 1)$, and that the restriction to the binary graph is sparse for any $\alpha \in (0,1)$. See Remark C.4.8 for a discussion. Thus, edge-exchangeable random graph sequences allow for a wide range of sparse and dense behavior.

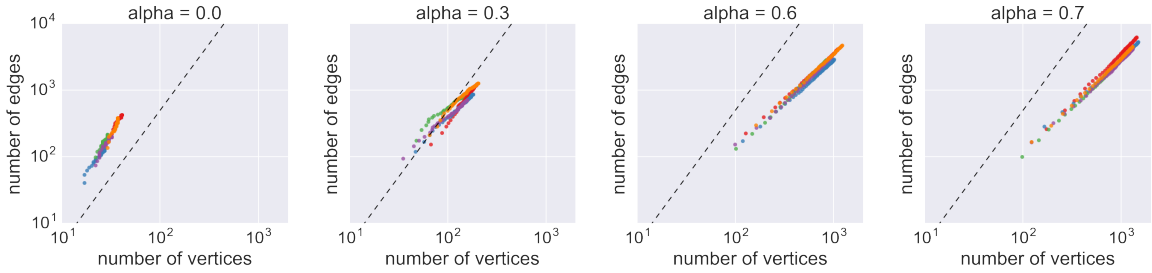
4.6 Simulations

In this section, we explore the behavior of graphs generated by the model from Section 4.5 via simulation, with the primary goal of empirically demonstrating that the model produces sparse graphs. We consider the case when the Poisson process generating the weights in Equation (4.2) has the rate measure of a *three-parameter beta process* (3-BP) on $(0,1)$ (Broderick et al., 2012; Teh and Görür, 2009):

$$\nu(dw) = \gamma\frac{\Gamma(1+\beta)}{\Gamma(1-\alpha)\Gamma(\alpha+\beta)}w^{-1-\alpha}(1-w)^{\alpha+\beta-1}\,dw, \tag{4.3}$$



(a) Multigraph edges vs. active vertices



(b) Binary graph edges vs. active vertices

Figure 4.6.1: Data simulated from a graph frequency model with weights generated according to a 3-BP. Colors represent different random draws. The dashed line has a slope of 2.

with mass $\gamma > 0$, concentration $\beta > 0$, and discount $\alpha \in (0, 1)$. In order for the 3-BP to have finite total mass $\sum_j w_j < \infty$, we require that $\beta > -\alpha$. We draw realizations of the weights from a 3-BP(γ, β, α) according to the stick-breaking representation given by Broderick, Jordan, and Pitman (2012). That is, the w_i are the atom weights of the measure W for

$$W = \sum_{i=1}^{\infty} \sum_{j=1}^{C_i} V_{i,j}^{(i)} \prod_{l=1}^{i-1} (1 - V_{i,j}^{(\ell)}) \delta_{\psi_{i,j}}, \qquad C_i \overset{\text{iid}}{\sim} \text{Pois}(\gamma),$$

$$V_{i,j}^{(\ell)} \overset{\text{ind}}{\sim} \text{Beta}(1 - \alpha, \beta + \ell\alpha), \qquad \psi_{i,j} \overset{\text{iid}}{\sim} B_0$$

and any continuous (i.e., non-atomic) choice of distribution B_0.

Since simulating an infinite number of atoms is not possible, we truncate the outer summation in i to 2000 rounds, resulting in $\sum_{i=1}^{2000} C_i$ weights. The parameters of the beta process were fixed to $\gamma = 3$ and $\theta = 1$, as they do not influence the sparsity of the resulting graph frequency model, and we varied the discount parameter α. Given a single draw W (at some specific discount α), we then

simulated the edges of the graph, where the number of Bernoulli draws N varied between 50 and 2000.

Figure 4.6.1a shows how the number of edges varies versus the total number of active vertices for the multigraph, with different colors representing different random seeds. To check whether the generated graph was sparse, we determined the exponent by examining the slope of the data points (on a log-scale). In all plots, the black dashed line is a line with slope 2. In the multigraph, we found that for the discount parameter settings $\alpha = 0.6, 0.7$, the slopes were below 2; for $\alpha = 0, 0.3$, the slopes were greater than 2. This corresponds to our theoretical results; for $\alpha < 0.5$ the multigraph is dense with slope greater than 2, and for $\alpha > 0.5$ the multigraph is sparse with slope less than 2. Furthermore, the sparse graphs exhibit *power law* relationships between the number of edges and vertices, i.e., $|E_N| \overset{\text{a.s.}}{\sim} c\,|V_N|^b$, $N \to \infty$, where $b \in (1, 2)$, as suggested by the linear relationship in the plots between the quantities on a log-scale. Note that there are necessarily fewer edges in the binary graph than in the multigraph, and thus this plot implies that the binary graph frequency model can also capture sparsity. Figure 4.6.1b confirms this observation; it shows how the number of edges varies with the number of active vertices for the binary graph. In this case, across $\alpha \in (0, 1)$, we observe slopes that are less than 2. This agrees with our theory from Section 4.5, which states that the binary graph is sparse for any $\alpha \in (0, 1)$.

4.7 Discussion

We have proposed an alternative form of exchangeability for random graphs, which we call *edge exchangeability*, in which the distribution of a graph sequence is invariant to the order of the edges. We have demonstrated that edge-exchangeable graph sequences, unlike traditional vertex-exchangeable sequences, can be sparse by developing a class of edge-exchangeable graph frequency models that provably exhibit sparsity. Simulations using edge frequencies drawn according to a three-parameter beta process confirm our theoretical results regarding sparsity. Our results suggest that a variety of future directions would be fruitful—including theoretically characterizing different types of power laws within graph frequency models, characterizing the use of truncation within graph frequency

models as a means for approximate Bayesian inference in graphs, and understanding the full range of distributions over sparse, edge-exchangeable graph sequences.

Chapter 5

Multi-fidelity Monte Carlo: a pseudo-marginal approach

Markov chain Monte Carlo (MCMC) is an established approach for uncertainty quantification and propagation in scientific applications. A key challenge in applying MCMC to scientific domains is computation: the target density of interest is often a function of expensive computations, such as a high-fidelity physical simulation, an intractable integral, or a slowly-converging iterative algorithm. Thus, using an MCMC algorithms with an expensive target density becomes impractical, as these expensive computations need to be evaluated at each iteration of the algorithm. In practice, these computations often approximated via a cheaper, low-fidelity computation, leading to bias in the resulting target density. Multi-fidelity MCMC algorithms combine models of varying fidelities in order to obtain an approximate target density with lower computational cost. In this chapter, we describe a class of asymptotically exact multi-fidelity MCMC algorithms for the setting where a sequence of models of increasing fidelity can be computed that approximates the expensive target density of interest. We take a pseudo-marginal MCMC approach for multi-fidelity inference that utilizes a cheaper, randomized-fidelity unbiased estimator of the target fidelity constructed via random truncation of a telescoping series of the low-fidelity sequence of models. Finally, we discuss

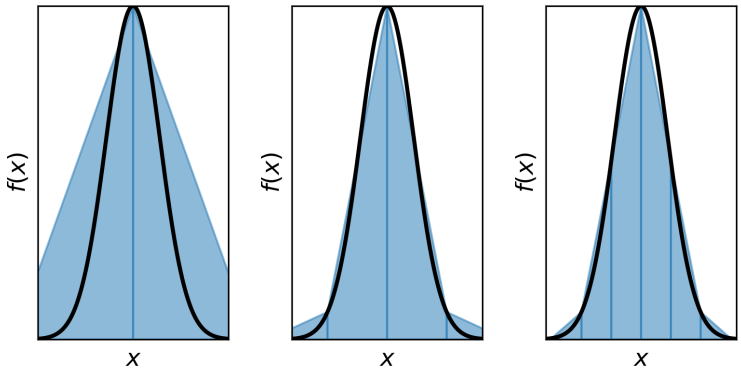
and evaluate the proposed multi-fidelity MCMC approach on several applications, including log-Gaussian Cox process modeling, Bayesian ODE system identification, PDE-constrained optimization, and Gaussian process parameter inference.

5.1 Introduction

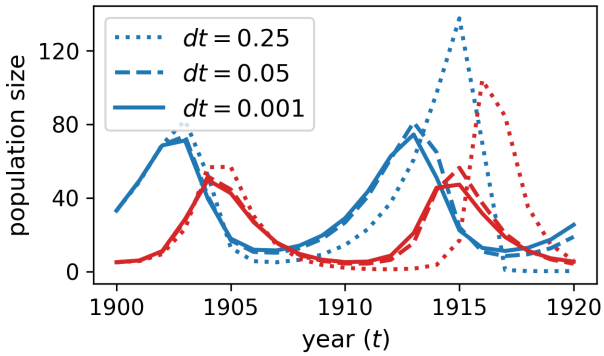
Simulation and computational modeling play a key role in science, engineering, economics, and many other areas. When these models are high-quality and accurate, they are important for scientific discovery, design, and data-driven decision making. However, the ability to accurately model complex physical phenomena often comes with a significant cost—many models involve expensive computations that then need to be evaluated repeatedly in, for instance, a sampling or optimization algorithm. Examples of model classes with expensive computations include intractable integrals or sums, expensive quantum simulations (Troyer and Wiese, 2005), expensive numerical simulations arising from partial differential equations (PDEs) (Raissi et al., 2017) and large systems of ordinary equations (ODEs).

In many situations, one has the ability to trade off computational cost against *fidelity* or accuracy in the result. Such a tradeoff might arise from the choice of discretization or the number of basis functions when solving a PDE, or the number of quadrature points when estimating an integral. It is often possible to leverage lower-fidelity models to help accelerate high-quality solutions, e.g., by using multigrid methods (Hackbusch, 2013) for spatial discretizations. More generally, *multi-fidelity* methods combine multiple models of varying cost and fidelity to accelerate computational algorithms and have been applied to solving inverse problems (Cui et al., 2015; Higdon et al., 2002; Raissi et al., 2017), trust region optimization (Alexandrov et al., 1998; Arian et al., 2000; Fahl and Sachs, 2003; March and Willcox, 2012; Robinson et al., 2008), Bayesian optimization (Brevault et al., 2020; Gramacy and Lee, 2009; Jones et al., 1998; Li et al., 2020; Song et al., 2019; Wu et al., 2020), Bayesian quadrature (Gessner et al., 2020; Xi et al., 2018), and sequential learning (Gundersen et al., 2021; Palizhati et al., 2022).

One critically important tool for scientific and engineering computation is Markov chain Monte



(a) Trapezoid rule with $2k$ trapezoids ($k = 1, 2, 3$). Sequence trapezoid quadrature estimates I_k, where I_k is the trapzoid rule with $2k$ trapezoids.



(b) Lotka-Volterra ODE solutions, $dt = 1/k$. Lotka-Volterra ODE solutions for prey $u(t)$ (blue) and predator $v(t)$ (red) using Euler's method with step size dt.

Figure 5.1.1: Examples of low-fidelity sequences of models.

Carlo (MCMC), which is widely used for uncertainty quantification, optimization, and integration. MCMC methods are recipes for constructing a Markov chain with some desired target distribution as the limiting distribution. Pseudo-random numbers are used to simulate transitions of the Markov chain in order to produce samples from the target distribution. However, MCMC often becomes impractical for high-fidelity models, where a single step of the Markov chain may, for instance, involve a numerical simulation that takes hours or days to complete. Multi-fidelity methods for MCMC focus on constructing Markov chain transition operators that are sometimes able to use inexpensive

low-fidelity evaluations instead of expensive high-fidelity evaluations. The goal is to increase the effective number of samples generated by the algorithm, given a constrained computational budget. A large focus of the multi-fidelity MCMC literature is on two-stage Metropolis-Hastings (M-H) methods (Christen and Fox, 2005; Efendiev et al., 2006), which use a single low-fidelity model for early rejection of a proposed sample, thereby often short-circuiting the evaluation of the expensive, high-fidelity model.

However, there are several limitations of two-stage multi-fidelity Monte Carlo. First, in many applications, a *hierarchy* of cheaper, low-fidelity models is available; for instance, in the case of integration, k-point quadrature estimates form a hierarchy of low-fidelity models, and in the case of a PDE, varying the discretization. Thus, the two-stage approach does not fully utilize the availability of a hierarchy of fidelities and may be more suitable for settings where the high- and low-fidelity models are not hierarchically related, e.g., semi-empirical methods vs. Hartree-Fock in computational chemistry. In addition, in such applications, there is often a limiting model of interest, such as a continuous function that the low-fidelity discretizations approximate. Two-stage MCMC does not asymptotically sample from this limiting target density and will at best sample from an approximation of the biased, high-fidelity posterior. Finally, the two-stage method is unnatural to generalize to more sophisticated MCMC algorithms such as slice sampling and Hamiltonian Monte Carlo (HMC).

We propose a class of multi-fidelity MCMC methods designed for applications with a hierarchy of low-fidelity models available. More specifically, we assume access to a sequence of low-fidelity models that converge to a "perfect-fidelity" model in the limit. Within an MCMC algorithm, we can approximate the perfect-fidelity target density with an unbiased estimator constructed from a randomized truncation of the infinite telescoping series of low-fidelity target densities. This class of multi-fidelity MCMC is an example of a pseudo-marginal MCMC (PM-MCMC) algorithm—the unbiased estimator essentially guarantees that the algorithm is asymptotically exact in that the limiting distribution recovers the perfect-fidelity target distribution as its marginal distribution. Our approach introduces the fidelity of a model as an auxiliary random variable that is evolved separately from the target variable according to its own conditional target distribution; this technique

can be used in conjunction with any suitable MCMC update that leaves the conditional update for the target variable of interest invariant, such as M-H, slice sampling, elliptical slice sampling, or Hamiltonian Monte Carlo. We apply the pseudo-marginal multi-fidelity MCMC approach to several problems, including log-Gaussian Cox process modeling, Bayesian ODE system identification, PDE-constrained optimization, and Gaussian process parameter inference.

5.1.1 Related work

Multi-fidelity MCMC methods are commonly applied in a two-stage procedure, where the goal is to reduce the computational cost of using a single expensive high-fidelity model by using a cheap low-fidelity model as a low-pass filter for a delayed acceptance/rejection algorithm (Christen and Fox, 2005; Cui et al., 2015; Efendiev et al., 2006); see Peherstorfer et al. (2018) for a survey. Higdon et al. (2002) propose coupling a high-fidelity Markov chain with a low-fidelity Markov chain via a product chain. In constrast, our approach aims to sample from a "perfect-fidelity" target density while reducing computational cost; two-stage MCMC algorithms result in biased estimates with respect to this target density. A related class of methods is multilevel Monte Carlo (Dodwell et al., 2015; Giles, 2008, 2013; Warne et al., 2021), which uses a hierarchy of multi-fidelity models for Monte Carlo estimation by expressing the expectation of a high-fidelity model as a telescoping sum of low-fidelity models. Dodwell et al. (2015) use the M-H algorithm to form the multilevel Monte Carlo estimates, simulating from a separate Markov chain for each level of the telescoping sum. In practice multilevel Monte carlo requires choosing a finite number of fidelities, inducing bias in the estimator with respect to the (limiting) perfect-fidelity model. In contrast, our method uses a randomized fidelity within a single Markov chain with the perfect-fidelity model as the target.

Our approach applies pseudo-marginal MCMC to multi-fidelity problems. There is a rich literature developing pseudo-marginal MCMC methods (Andrieu and Roberts, 2009; Beaumont, 2003) for so-called "doubly-intractable" likelihoods, which are likelihoods that are intractable to evaluate. Several approaches in the pseudo-marginal MCMC literature are particular relevant to our work. The first are the PM-MCMC methods introduced by Lyne et al. (2015), which describes a class of pseudo-marginal M-H methods that use Russian roulette estimators to obtain unbiased

estimators of the likelihood. However, this method samples the variable of interest jointly with the auxillary randomness, which often leads to sticking.

Alternatively, several methods have considered sampling the randomness separately. The idea of clamping random numbers is explored in depth by Andrieu et al. (2010) and Murray and Graham (2016); the latter applies to this pseudo-marginal slice sampling. In particular, our approach applies these ideas to the specific setting of multi-fidelity models, where the random fidelity is treated as an auxillary variable. Finally, while our approach applies to doubly-intractable problems, we are also motivated by a larger class of multi-fidelity problems studied in the computational sciences that may not even be inference problems, such as quantum simulations and PDE-constrained optimization.

5.2 Multi-fidelity MCMC

Monte Carlo methods approximate integrals and sums that can be expressed an expectation:

$$\mathbb{E}_\pi(h(\theta)) = \int h(\theta)\,\pi(\theta)\,d\theta \approx \frac{1}{T}\sum_{t=1}^{T} h(\theta^{(t)}), \qquad \text{where} \quad \theta^{(t)} \sim \pi, \tag{5.1}$$

and where $\pi : \Theta \to \mathbb{R}_+$ is the *target density* that may only be known up to a constant, $h(\theta)$ is a function of interest, and $\{\theta^{(t)}\}_{t=1}^{T}$ are samples from π. Markov chain Monte Carlo methods are then used to generate samples $\theta^{(t)}$ from π by simulating from a Markov chain with target π.

In many settings, pointwise evaluations of the target function $\pi(\theta)$ are expensive or even intractable; from here on we will assume that the goal is to compute statistics of a quantity of interest $h(\theta)$ with respect to a *perfect-fidelity* target density $\pi_\infty(\theta)$. In practice, the estimate in Equation (5.1) is instead estimated using a cheaper, low-fidelity density $\pi_k(\theta)$, where $k \in \mathbb{N} := \{1, 2, \ldots\}$. In particular, we consider settings where there is a *sequence* of low-fidelity densities available that converge to the target, i.e., $\pi_k(\theta) \overset{k\to\infty}{\longrightarrow} \pi_\infty(\theta)$. We assume that as k increases, the model becomes higher in fidelity (with respect to π_∞) but more costly to evaluate, increasing in expense super-linearly with k.

For instance, π_∞ could represent a target density that depends on an intractable integral, the

solution of a PDE, the solution of a large system of ODEs, the solution of a large system of linear equations, or the minimizer of a function. Thus, a typical evaluation of π_∞ requires an approximation at a fidelity k with a tolerable level of bias for a given computational budget. Here increasing k could correspond to finer discretizations of differential equations, increasing numbers of quadrature points, or performing a larger number of iterations in a linear solver or optimization routine.

In the multi-fidelity setting, the goal is to combine several models of varying fidelity within an MCMC algorithm to reduce the computational cost of estimating Equation (5.1). In this chapter, we describe a class of MCMC algorithms that leverages the sequence of low-fidelity models π_k. Our strategy for multi-fidelity MCMC (MF-MCMC) will be to construct an unbiased estimator of $\pi_\infty(\theta)$ using random choices of the fidelity K and then to include K in the Markov chain as an auxiliary variable. By carefully constructing such a Markov chain, it will be possible to asymptotically estimate the functional in Equation (5.1) as though the samples were taken from the perfect-fidelity model; each step of the Markov chain will nevertheless only require a finite amount of computation. Finally, our approach allows us to essentially plug in any valid MCMC algorithm, and we apply this strategy to develop multi-fidelity variants of a number of MCMC algorithms, such as M-H and slice sampling.

5.2.1 Pseudo-marginal MCMC for the multi-fidelity setting

Pseudo-marginal MCMC (Andrieu and Roberts, 2009; Beaumont, 2003) is a class of auxillary-variable MCMC algorithms that replaces the target density $\pi(\theta)$ with an estimator $\hat{\pi}(\theta)$ that is a function of a random variable. If the estimator is nonnegative and unbiased, i.e., for all $\theta \in \Theta$, $\hat{\pi}(\theta) \geq 0$ and $\mathbb{E}[\hat{\pi}(\theta)] = \pi(\theta)$, then MCMC transitions that use the estimator still have $\pi(\theta)$ as their invariant distribution. This property is sometimes referred to as "exact-approximate" MCMC as the transitions are approximate but the limiting distribution is exact. Estimators can be constructed from a variety of methods, including particle filtering (Andrieu and Roberts, 2009); our approach will use randomized series truncations, which has been consider in pseudo-marginal MCMC methods such as Lyne et al. (2015), Georgoulas et al. (2017), and Biron-Lattes et al. (2022).

We now apply the pseudo-marginal approach to the multi-fidelity setting. Here the target density

estimator arises from a random choice of the fidelity $K \in \mathbb{N}$ that is governed by a distribution μ on $\mathbb{N}$. We denote the estimator using $\hat{\pi}_K(\theta)$ to make the dependence on the random fidelity K explicit. The estimator is constructed such that it is unbiased with respect to μ, i.e.,

$$\sum_{k=1}^{\infty} \mu(k)\hat{\pi}_k(\theta) = \pi_\infty(\theta)\,. \tag{5.2}$$

The distribution μ is also constructed by the user: ideally, the estimator $\hat{\pi}_K(\theta)$ will prefer smaller values of K while having sufficiently low variance as to allow the Markov chain to mix effectively. Thus the simulations can be run at inexpensive low-fidelities, while the estimates will be as though the perfect-fidelity model were being used.

The standard pseudo-marginal MCMC approach is to construct a Markov chain that has the following joint density as its stationary distribution:

$$\pi(\theta, K) = \mu(K)\hat{\pi}_K(\theta). \tag{5.3}$$

Observe that while Equation (5.3) does not depend on the perfect-fidelity target density π_∞, it returns the desired marginal π_∞ via Equation (5.2). As a concrete example, a pseudo-marginal M-H algorithm generates a new state θ' and fidelity K' jointly using $q(\theta';\theta)$ as the proposal for θ', $q(K';K) = \mu(K')$ as the proposal distribution for the fidelity, and accepts/rejects the state according to

$$a = \frac{\pi(\theta', K')q(\theta;\theta')q(K;K')}{\pi(\theta, K)q(\theta';\theta)q(K';K)} = \frac{\hat{\pi}_{K'}(\theta')q(\theta;\theta')}{\hat{\pi}_K(\theta)q(\theta';\theta)}, \tag{5.4}$$

where the equality holds since the distribution terms for K and K' cancel. Note that the right-hand side of Equation (5.4) is the standard M-H ratio but that the target density π is replaced with the estimator $\hat{\pi}_K$.

However, standard pseudo-marginal MCMC using joint proposals of the state and fidelity can "get stuck" when the estimator is noisy and fail to accept new states. Thus, we apply the approach in Murray and Graham (2016) that augments the Markov chain to include the randomness of the

estimator via a separate update; here the randomness of the estimator arises from the fidelity K. Concretely, we construct a Markov chain that simulates from Equation (5.3) by alternating sampling between the conditional target densities $\pi(K|\theta)$ and $\pi(\theta|K)$ (steps 5 and 6 of Algorithm 5.3.1, respectively). We refer to this strategy as *multi-fidelity MCMC* (MF-MCMC), since by conditioning on $K = k$, the update for the state θ becomes a standard deterministic update applied to a low-fidelity model $\hat{\pi}_k(\theta)$, and any appropriate MCMC update can be used here, making it straightforward to use complex MCMC methods, such as slice sampling and HMC. Similarly, any suitable MCMC update for the fidelity K can be used using the conditional target $\pi(K|\theta)$.

Many techniques can be used to construct an unbiased estimator of π_∞ with randomness K; we describe a general approach in the next section. However, it is generally difficult to guarantee the estimator is nonnegative, as required by pseudo-marginal MCMC. One technique considered by Lin et al. (2000) and Lyne et al. (2015) is to instead sample from the target distribution induced by the absolute value of the estimator and applying a sign-correction to the final Monte Carlo estimate in Equation (5.1), an approach borrowed from the quantum Monte Carlo literature where it is necessary for modeling fermionic particles. This approach has been applied to the M-H algorithm, but we note that this general approach can be applied much more broadly, as we do in this work.

In problems where the estimator may be negative, we sample from the conditional target distributions using the absolute value of the estimator $|\hat{\pi}_K(\theta)|$, and we denote these conditionals with $\tilde{\pi}(K \,|\, \theta) \propto \mu(K)|\hat{\pi}_K(\theta)|$ and $\tilde{\pi}(\theta|K = k) \propto |\hat{\pi}_k(\theta)|$. The estimate in Equation (5.1) is then corrected using the signs $\sigma(\theta, k)$ of evaluations of $\hat{\pi}_k(\theta)$,

$$\int h(\theta)\, \pi(\theta)\, d\theta \approx \frac{\sum_{t=1}^{T} h(\theta^{(t)})\sigma(\theta^{(t)}, K^{(t)})}{\sum_{t=1}^{T}\sigma(\theta^{(t)}, K^{(t)})} =: \hat{I}_T, \tag{5.5}$$

where $\{(\theta^{(t)}, K^{(t)})\}_{t=1}^{T}$ are the sampled values from the joint distribution $\tilde{\pi}(\theta, K) \propto |\hat{\pi}_K(\theta)|\mu(K)$.

Importantly, the sign-corrected estimate still asymptotically leads to the desired estimate of the functional of interest. Let $\sigma(\theta, k)$ denote the sign of the estimator such that $\hat{\pi}_k(\theta) = \sigma(\theta, k)|\hat{\pi}_k(\theta)|$. The estimator $\hat{I}_T$ in Equation (5.5) is formed using a Monte Carlo estimate of the functional after

expanding it into its joint distribution, i.e.,

$$\int h(\theta)\pi_\infty(\theta)d\theta = \int \sum_{k=1}^{\infty} h(\theta)\hat{\pi}_k(\theta)\mu(k)d\theta = \frac{\int \sum_{k=1}^{\infty} h(\theta)\sigma(\theta,k)\tilde{\pi}(\theta,k)d\theta}{\int \sum_{k=1}^{\infty} \sigma(\theta,k)\tilde{\pi}(\theta,k)d\theta}. \tag{5.6}$$

The full multi-fidelity MCMC algorithm with sign correction summarized in Algorithm 5.3.1. We note that while the Markov chain no longer converges to a target with the marginal π_∞, the final estimate after sign-correction—which is the downstream goal of interest—converges to the quantity of interest due to Equation (5.6). While this may seem limiting if one is interested in the posterior itself, useful unbiased posterior summaries may be still be obtained via the functional, such as the posterior mean, variance, quantiles, and histograms that may be used to visualize marginal distributions.

5.3 Unbiased low-fidelity estimators via randomized truncations

In this section, we discuss how to construct an unbiased estimator of $\pi_\infty(\theta)$, given a sequence of low-fidelity likelihoods with the property $\pi_k(\theta) \to \pi_\infty(\theta)$ as $k \to \infty$. This estimator has the property that it requires a finite amount of computation with probability one, and it also has a tunable amount of expected computation per estimate, i.e., it uses low-fidelity density evaluations to estimate the perfect-fidelity target density. The central idea of this estimator has been used for decades, going back to John von Neumann and Stanislaw Ulam. More recently it has found use in applications of inference and optimization in related work such as Glynn and Rhee (2014), Lyne et al. (2015), Beatson and Adams (2019), and Jacob et al. (2020).

First note that we can express the perfect-fidelity model as a telescoping sum of low-fidelity models: let $\pi_0(\theta) = 0$ and write

$$\pi_\infty(\theta) = \sum_{k=1}^{\infty} \pi_k(\theta) - \pi_{k-1}(\theta). \tag{5.7}$$

The estimator $\hat{\pi}_K$ is then constructed by taking a random truncation $K \sim \mu$ of the infinite telescoping series. The sampled terms in the sum are then reweighted to ensure the estimator remains unbiased:

$$\hat{\pi}_K(\theta) = \sum_{k=1}^{K} w_{k,K}(\pi_k(\theta) - \pi_{k-1}(\theta)) \,. \tag{5.8}$$

Two approaches are commonly used to ensure that the resulting estimator is unbiased: weighted single-term estimators and Russian roulette estimators. The single-term estimator (Lyne et al., 2015) is constructed by importance sampling a term from the series in Equation (5.7): the truncation level is drawn as $K \sim \mu$, and the Kth term is used to form the estimate

$$\hat{\pi}_K(\theta) = \mu(K)^{-1}(\pi_K(\theta) - \pi_{K-1}(\theta)) \,. \tag{5.9}$$

Thus, the weight in Equation (5.8) is $W_{k,K} = \mu(K)^{-1}\mathbb{1}(K = k)$. In the Russian roulette estimator, the remaining terms in the estimator are reweighted by their survival probabilities, i.e., $W_{k,K} = (1 - \sum_{k'=1}^{k-1} \mu(k'))^{-1}\mathbb{1}(K \geq k)$. The distribution μ controls the number of terms in the estimator, and a good proposal distribution should match the tails of the sequence of low-fidelity densities (Beatson and Adams, 2019; Lyne et al., 2015; Potapczynski et al., 2021).

The ability to use cheaper models is a key feature of multi-fidelity inference, and the low-fidelity estimator provides a means to reduce the computational cost of multi-fidelity Monte Carlo. However, these estimators are an example of a class of methods that explores a compute-variance tradeoff: computationally cheaper estimates leads to high variability. The resulting increase in variance slows down the convergence of the MCMC procedure and could lead to an overall less efficient method due to a reduced effective sample size.

5.4 Summary of the multi-fidelity MCMC recipe

Here we summarize the recipe for constructing a multi-fidelity Markov chain Monte Carlo algorithm.

First, identify a sequence of increasing-fidelity target densities with the property that their limit is the desired "perfect-fidelity" density. Low-fidelity densities should be cheap with the cost

Algorithm 5.3.1 Multi-fidelity Monte Carlo with sign-correction

Input: Initial state θ and fidelity K, truncation distribution μ
for $t = 1, \ldots, T$ **do**
    Given current K and θ, form estimator $\hat{\pi}_K(\theta) = \sum_{k=1}^{K} w_{k,K}(\pi_k(\theta) - \pi_{k-1}(\theta))$
    Save sign $\sigma(\theta, K) = \text{sign}(\hat{\pi}_K(\theta))$
    Update fidelity K leaving invariant the target conditional

$$\tilde{\pi}(K|\theta) \propto \mu(K)|\hat{\pi}_K(\theta)|$$

    Update state θ leaving invariant the target conditional

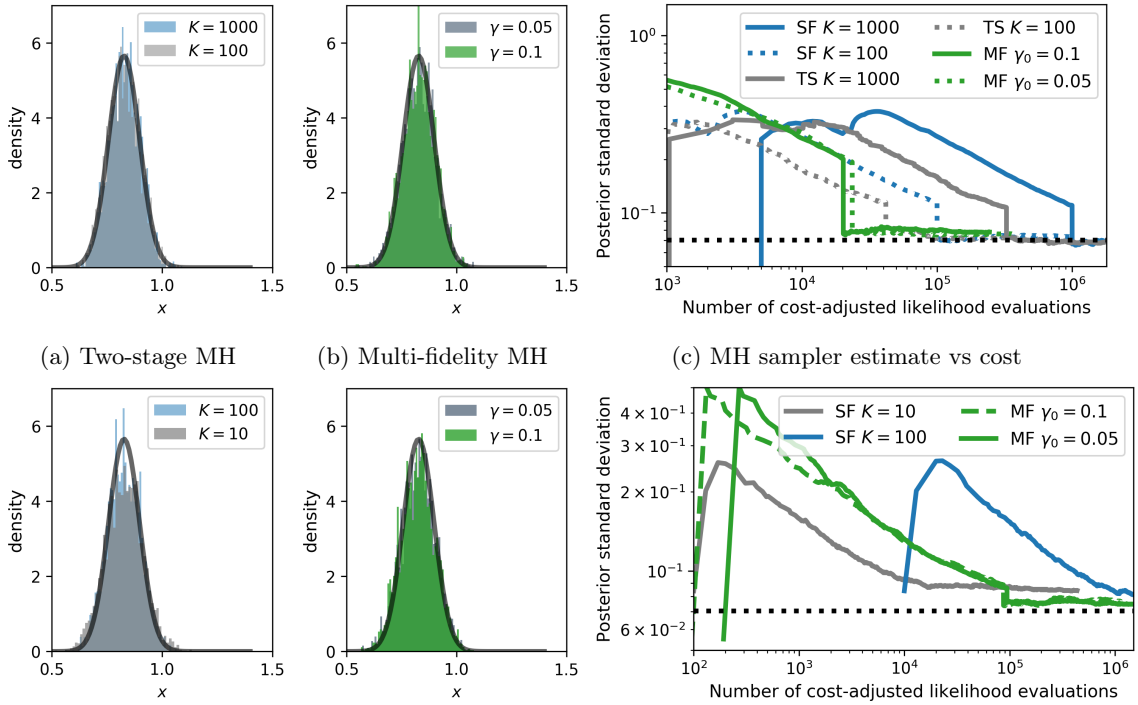
$$\tilde{\pi}(\theta|K = k) \propto |\hat{\pi}_k(\theta)|$$

end for
Output: Samples $\{(\theta^{(t)}, K^{(t)})\}$ and estimate $\hat{I}_T = \left(\sum_{t=1}^{T} \sigma^{(t)} h(\theta^{(t)})\right) / \left(\sum_{t=1}^{T} \sigma^{(t)}\right)$

rapidly increasing within the sequence. In the context of Bayesian inference, it may be appropriate to focus the multi-fidelity aspects on the likelihood term and construct the target densities via, e.g., $\pi_k(\theta; \mathcal{D}) \propto \pi_0(\theta) L_k(\theta\,;\, \mathcal{D})$, where π_0 is the prior, L_k is a low-fidelity likelihood, and $\mathcal{D}$ is the set of observations. This likelihood-based version is what we use in several of the experiments.

Next, introduce a truncation distribution μ on $\mathbb{N}$. This truncation distribution should be chosen to balance between expected cost and variance of the resulting estimator; our overall goal is to mostly use cheap low-fidelity densities, but high-variance estimators will presumably damage the mixing time and/or the asymptotic variance.

Initialize the Markov chain with a reasonable choice for θ and a draw of K from the distribution μ. Each step of the Markov chain simulation consists of an update to θ given K and an update of K given θ. The update of θ given K can be performed using any standard MCMC algorithm, e.g., M-H, slice sampling, or HMC, applied to the low-fidelity estimator. It is important to use the absolute value of the estimator and keep track of its sign. The update of K given θ is also flexible, but it is reasonable to construct the update so that only a few K are considered in each step, as each of those fidelities will need to be evaluated. By default, we consider a simple random walk on the positive integers for our experiments. After running a sufficient number of steps of the Markov chain, use the sign corrected-estimator in Equation (5.5) to compute the expectation of the function $h(\theta)$.



(a) Two-stage MH (b) Multi-fidelity MH (c) MH sampler estimate vs cost

(d) Single-fidelity slice (e) Multi-fidelity slice (f) Slice sampler estimate vs cost

Figure 5.5.1: Demonstration of multi-fidelity MCMC on a conjugate Gaussian model. *Left:* Histograms for M-H (a,b) and slice sampling (d,e). *Right:* Comparison of posterior standard deviation estimate vs computation for M-H (c) and slice sampling (d) methods.

5.5 Experiments

In all experiments, we use a random-walk M-H update to sample from the conditional $K|\theta$, and truncation distribution $\mu(K) = \text{geometric}(K; \gamma_0)$. Additional experimental details are in Appendix D.3.

5.5.1 Toy conjugate Gaussian models

In order to understand the behavior of MF-MCMC on a simple example of Bayesian inference, we first examine an example where the computational cost of evaluating the sequence of low-fidelity likelihoods does not increase with k. Consider a perfect-fidelity likelihood $L_\infty(\theta) = \mathcal{N}(x; \theta, \sigma_\infty)$ and a low-fidelity likelihood $L_k(\theta) = \mathcal{N}(x; \theta, \sigma_k)$, where $\sigma_k^2 \to \sigma_\infty^2$. The prior is $\pi_0(\theta) = N(\theta|0, 1)$, and so

a closed-form posterior density can be computed. Here we consider the sequence $\sigma_k^2 = 1 + 2/k^2$ and $\sigma_\infty^2 = 1$. In Figure 5.5.1 we compare the results of single-fidelity and multi-fidelity M-H and slice sampling as well as the two-stage M-H algorithm summarized in Section 2.3.4. We consider 2 two-stage M-H settings with high and low fidelities of $\{k^{\mathrm{HF}}, k^{\mathrm{LF}}\} = \{1000, 10\}$ and $\{k^{\mathrm{HF}}, k^{\mathrm{LF}}\} = \{100, 5\}$. The histograms show the bias of each method after simulating 10,000 samples, and the solid gray curve denotes the exact posterior density. We also compute a measure of total cost and a running average of the estimate of the posterior standard deviation functional, where the dotted black line denotes the true value. The number of cost-adjusted likelihoods was computed by upweighting each likelihood evaluation by the fidelity. Here the multi-fidelity methods typically converge to a similar value as the single high-fidelity methods but in fewer cost-adjusted likelihood evaluations.

5.5.2 Log-Gaussian Cox processes

We examine an application of MF-MCMC to the log Gaussian Cox process (LGCP) model (Møller et al., 1998), where the perfect-fidelity model is a function of an integral and the lower-fidelity sequence of models arises from k-point quadrature estimates. Let $\log f \sim \mathrm{GP}(0, \kappa_\ell)$, where $\kappa_\ell(x, x') = \exp\left(-\frac{1}{2\ell^2}\|x - x'\|_2^2\right)$ and where ℓ is a lengthscale hyperparameter. Consider an inhomogenous Poisson process on $\mathbb{X} \subseteq \mathbb{R}^D$ with intensity $\lambda(x) = e^{f(x)}$. Given a random set of points $\{X_n\}_{n=1}^N$, the perfect-fidelity likelihood is

$$L_\infty(f) = \exp\left(\int_{\mathbb{X}} (1 - e^{f(x)}) dx\right) \prod_{n=1}^{N} e^{f(X_n)}. \tag{5.10}$$

Typically, inference in the LGCP uses a grid-based approximation of Equation (5.10), where the points are binned into counts and modeled with a Poisson likelihood (Diggle et al., 2013; Murray et al., 2010; Teng et al., 2017), resulting in a biased posterior. Because the likelihood depends on a high-dimensional latent Gaussian vector, we perform inference for f using the elliptical slice sampling (ESS) algorithm (see Section 2.3.3). We approximate the integral in Equation (5.10) with a trapezoidal quadrature rule I_k, where the number of quadrature points is a linear function of k.

We apply multi-fidelity and single-fidelity ESS algorithms to a coal mining disasters data set

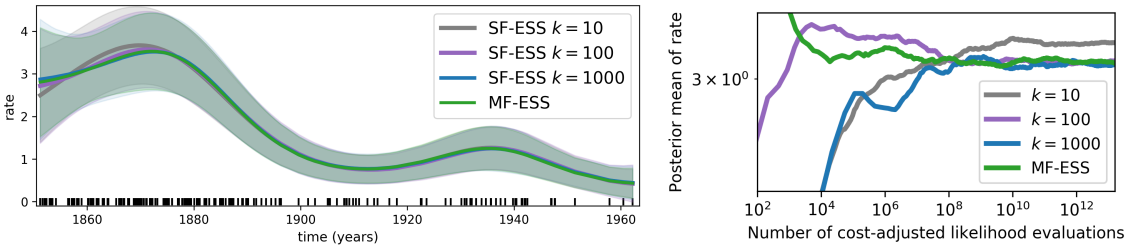


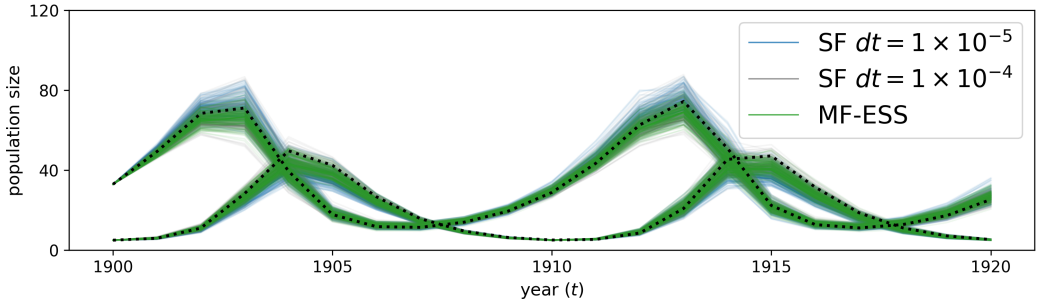
Figure 5.5.2: Coal mining disasters 1850–1963. *Left:* Posterior mean of the rate function at the observed data points. *Right:* Posterior mean of the rate function at $T = 1862$ vs computation.

(Carlin et al. (1992)). The data contain the dates of 191 coal mine explosions that killed ten or more men in Britain between March 15, 1851 and March 22, 1962. Figure 5.5.2 (left) shows the estimated mean intensity and standard deviation on coal mining disasters data between one run of multi-fidelity ESS and single-fidelity ESS with $k = 10, 100, 1000$ quadrature points. In this plot, the high- ($k = 1000$) and multi-fidelity posterior mean and standard deviation estimates match well, and the bias in the lowest fidelity ($k = 10$) estimate is apparent. We also computed the cost-adjusted number of likelihood evaluations performed in each iteration of MF-ESS and SF-ESS. Figure 5.5.2 (right) shows the average estimated mean intensity at the time step $t = 1862$ on the three models against the average cost-adjusted number of likelihood evaluations per iteration. We observe that the multi-fidelity and high-fidelity estimates are close after many iterations of sampling, but that the multi-fidelity estimate converges with less computation.

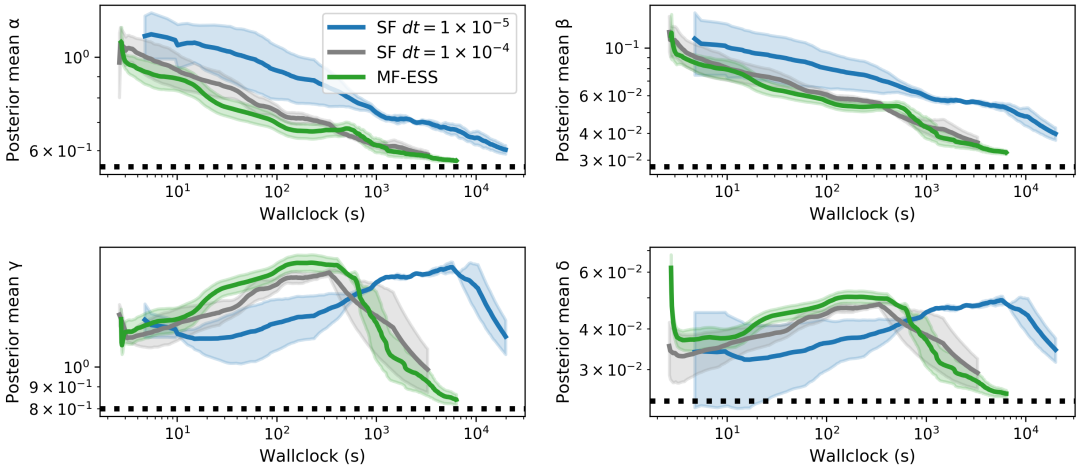
5.5.3 Bayesian ODE system identification

We now apply the MF-MCMC approach to Bayesian system identification for the Lotka-Volterra ODE. Let $u(t) \geq 0$ represent the population size of the prey species at time t, and $v(t) \geq 0$ represent the population size of the predator species. The dynamics of the two species given parameters $\alpha, \beta, \gamma, \delta \geq 0$ are given by a pair of first-order ODEs:

$$\frac{d}{dt}u = (\alpha - \beta v)u = \alpha u - \beta uv, \qquad \frac{d}{dt}v = (-\gamma - \delta u)v = -\gamma v - \delta uv. \tag{5.11}$$



(a) Sampled posterior trajectories under a fixed computational budget



(b) Posterior mean estimate vs computational cost

Figure 5.5.3: Lotka-Volterra system parameter identification. The fidelity represents (a function of) the step size dt of the ODE solver.

System identification solves the inverse problem by estimating the parameters of the ODE system $\theta = (\alpha, \beta, \gamma, \delta)$. Taking a Bayesian approach, we specify a noise model for the observed data and priors on the parameters, and we use MCMC to infer a distribution over the solution. For simplicity, we assume that the initial conditions are known and fix $\sigma = 0.25$.

Define $z_n := (u(t_n), v(t_n))$ and let $z_1(\theta), \ldots, z_N(\theta)$ be the solutions to the Lotka-Volterra differential equations at times $t_1, \ldots, t_N$ given the initial conditions and the system parameters $\theta = (\alpha, \beta, \gamma, \delta)$. Suppose we have observations arising from $\log(y_n) = \log(z_n) + \epsilon_n$, where $\epsilon \sim N(0, \sigma^2 I)$. The low-fidelity likelihood is a function of a numerical solution of the ODE using a

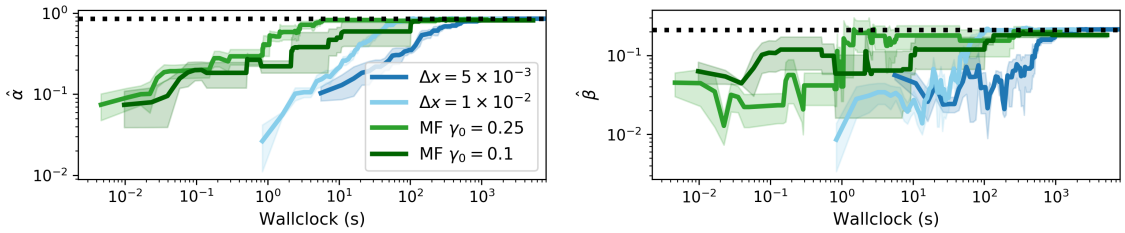


Figure 5.5.4: PDE-constrained optimization with a linear heat equation. Estimate for α (left) and β (right) vs computation. The black dotted lines denote the true values of α, β.

time step of size dt (Equation (D.3)). We compared the performance of a multi-fidelity elliptical slice sampler to single-fidelity ellipitical slice samplers with step size $dt = 1 \times 10^{-5}, 1 \times 10^{-4}$. For the ODE solver, we considered both an Euler solver and an 4th-order Runge-Kutta solver (Figure D.3.1). Figure 5.5.3a shows 200 trajectories corresponding to parameters sampled from the posterior distributions under each method using the Euler solver. Figure 5.5.3b shows posterior mean estimates for each system parameter. We observe that the estimates from the multi-fidelity slice sampler approach the estimates reported by Howard (2009) (black dotted lines) in less time than the single-fidelity samplers.

5.5.4 PDE-constrained optimization

We now consider global optimization of a PDE-constrained objective, where an expensive physical simulation is run repeatedly in an outer loop problem. A common approach for global optimization is simulated annealing, which has been applied to constrained global optimization (Romeijn and Smith, 1994). Consider a model for heat flow in a thin rod of length L with spatial coordinates $x \in [0, L]$. Let $u(x, t)$ represent the temperature in the rod at position x and time t, and let $\bar{u}$ represent a desired target temperature. The goal is to minimize a loss function subject to u satisfying a linear heat equation. This objective along with an initial condition and homogenous boundary

conditions can be summarized as:

$$\begin{aligned} &\text{minimize}_u \quad \|u - \bar{u}\|_2^2 \\ \text{subject to} \quad &\frac{\partial u}{\partial t} = \alpha \cdot \frac{\partial^2 u}{\partial x^2} + 2\beta \cdot u, \\ &u(x,0) = \sin\left(\pi x/2\right), \\ &u(0,t) = u(L,t) = 0, \quad x \in [0,L], t \in [0,T], \end{aligned} \tag{5.12}$$

where $\alpha, \beta > 0$ are the system parameters. The goal is to find $\theta = (\alpha, \beta)$ that minimizes the objective and satisfies the constraints. To solve the PDE, we discretize the domain into a grid of size Δx and represent the second derivative using the central difference formula. This induces a system of ODEs that we solve numerically using a Tsitouras 5/4 Runge-Kutta method, setting $\Delta t = 0.4\Delta x^2$ so as to satisfy a CFL stability condition. Here the fidelity of the problem is given by the size of the spatial discretization Δx, which in turn controls Δt. We compared against two single-fidelity discretizations of the spatial coordinate, where $\Delta x = 5 \times 10^{-3}, 1 \times 10^{-2}$. The results are in Figure 5.5.4, where we plot two of the MF results with $\gamma_0 = 0.1, 0.25$. In these examples, the multi-fidelity estimates converge faster than the single-fidelity estimates in wallclock time.

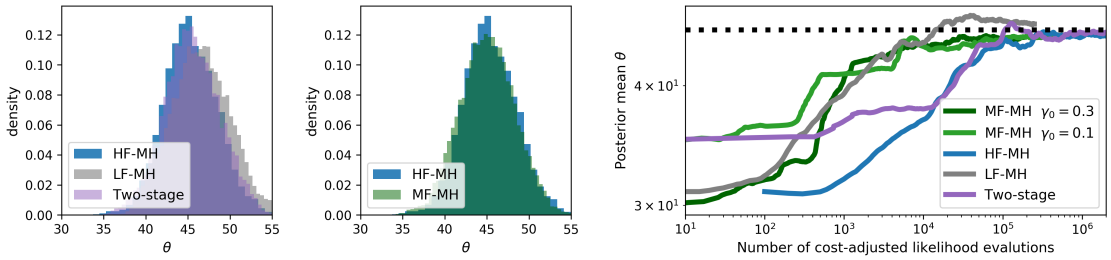
5.5.5 Gaussian process regression parameter inference

Let $X \in \mathbb{R}^{N \times D}$ and consider a Gaussian process regression model with a squared exponential kernel:

$$f \sim \text{GP}(0, k_\theta), \quad y = f(X) + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma_0^2), \quad k_\theta(x, x') = \exp\left(-\frac{1}{2\theta^2}\|x - x'\|_2^2\right), \tag{5.13}$$

where we assume σ_0^2 is known. Let $\Sigma_\theta := [k_\theta(x_i, x_j)]_{i \in [N], j \in [N]}$.

In many applications of Gaussian process modeling, one is interested in integrating out the parameters θ using MCMC. Computing the posterior $\pi(\theta|X, y)$ is expensive because each evaluation of the likelihood $p(y|X, \theta) = \mathcal{N}(y \,|\, 0, \Sigma_\theta + \sigma_0^2 I)$ involves computing a determinant and solving a linear system with respect to the matrix $\Sigma_\theta + \sigma_0^2 I$, which has an $O(N^3)$ computational cost associated with standard methods (e.g., Cholesky decomposition). For simplicity, we will only consider the



(a) Histograms of high, low, two-stage, and multi-fidelity (b) Posterior mean estimate vs computation

Figure 5.5.5: Parameter inference in a Gaussian process regression model. *Left:* The posterior distribution of the parameter θ. *Right:* The posterior mean estimate vs computional cost.

fidelity of solution to the linear system, but we note that the determinant can considered using the approach described in Potapczynski et al. (2021). Additional derivations and details are in Appendix D.3.5.

We generate synthetic data from the GP model with $N = 100$, $\sigma_0^2 = 1$, and lengthscale $\theta_0 = 45$. For the GP model, we use a log Normal prior on θ given above in Equation (D.5) with parameters $\nu_0 = 3.8, \nu_1 = 0.03$. In Figure 5.5.5, we compare these approaches using single-fidelity, multi-fidelity, and two-stage M-H samplers. The low-fidelity likelihood sequence was constructed by computing the solution to the linear system using a preconditioned conjugate gradient solver with k steps. The single-fidelity likelihoods were a high-fidelity likelihood ($k = N$) and a low-fidelity likelihood ($k = k_N \ll N$), and the multi-fidelity M-H samplers used $\gamma_0 = 0.1$. The two-stage M-H approach used high and low fidelities of $k \in \{100, 5\}$. For all methods, we use a M-H sampler with $T = 50000$ iterations. In the histograms, we observe that the high-fidelity, multi-fidelity and two-stage approaches tend to lead to similar posteriors, while the low-fidelity sampler has more noticable bias with respect to the high-fidelity histogram. The estimate produced by the multi-fidelity samplers converged in fewer cost-adjusted likelihood evaluations than the high-fidelity and two-stage approaches.

5.6 Discussion

In this work, we introduced a class of multi-fidelity MCMC that uses a low-fidelity unbiased estimator to reduce the computational cost of sampling while still maintaining the desired limiting target distribution of the Markov chain. In particular, we have demonstrated the use of our framework on more advanced MCMC algorithms beyond M-H, such as slice sampling, and to additional settings such as optimization. Our results show a reduction in computation while producing accurate solutions in comparison with high-fidelity models when it is possible to construct a target estimator that is not too noisy. Many future directions remain. First, applying MF-MCMC to large-scale expensive applications has many computational challenges. Making the method more robust to specialized problems is important, especially if the estimator is heavy-tailed. Thus, constructing good proposal distributions matching properties of the low-fidelity target sequence is crucial, especially for application to high-dimensional problems. In addition, we have thus far focused on target densities where there is a single computation whose fidelity is varied. However, in many settings, there may be target densities with multiple computations that converge at different rates, for example, if the target density includes both an intractable integral and a solution of a linear system. Our framework can be extended to that setting by adjusting the proposal distribution, and it is useful to understand how these rates impact the convergence properties of the sampler.

Chapter 6

Conclusion and future directions

In this chapter, we briefly review the contributions of the dissertation, and we conclude by discussing future avenues of research.

6.1 Summary of contributions

In this dissertation, we consider several case studies of popular probabilistic machine learning methods under misspecification. In the first case study, we study a popular Bayesian clustering model consisting of a finite mixture model with a prior on the number of components. First, we add rigor to existing data-analysis folk wisdom by proving that no matter the amount of misspecification, the posterior number of components diverges, i.e., the posterior probability of any particular finite number of components converges to 0 in the limit of infinite data. In particular, this result uses novel sufficient conditions that are more easily verifiable than those typical in the asymptotics literature.

Next, we study misspecification of scaling behavior in Bayesian network models. Many real-world graphs are *sparse* in the scaling of the number of edges relative to the number of vertices, but a class of popular probabilistic models for graphs based on exchangeability is misspecified for sparse graphs. We develop a framework that considers an alternative notation of exchangeability, called edge exchangeability, and show that a large class of generative models under this framework can

generate sparse graphs. Empirically, we also observe sparse power law scaling behaviors.

Finally, we consider Bayesian models with expensive or intractable likelihoods or target densities, a scenario that arises commonly in scientific computing applications. This is particularly problematic when using MCMC to infer the posterior, as each iteration typically requires an evaluation of the (unnormalized) target density. In these scenarios, it is common to approximate the target density, adding a source of bias to the posterior. We show that in problems in which we have access to a sequence of low-fidelity target densities that converge to the "infinite"-fidelity target, we can construct a Markov chain with the infinite-fidelity target while only ever having to evaluate elements in the lower-fidelity sequence of target densities. Our approach allows us to construct general-purpose MCMC algorithms, and we apply this framework to a number of complex algorithms, including slice sampling and elliptical slice sampling.

6.2 Future directions

Robustness to misspecification in latent variable models. In this dissertation, we only began to explore the behavior of finite mixture models under component misspecification, and many future directions remain. We expect that under many related models, estimates for the number of components will have similar conclusions to those studied in this dissertation, including for non-Bayesian mixture modeling methods. One future direction is to analyze the behavior of the estimate of the number of components selected by information criterion methods, such as AIC and BIC. In addition, because our analysis is inherently asymptotic, it is possible that the posterior on the number of components may still provide useful inferences for a finite sample. Finally, we analyzed the behavior of the power posterior under a fixed finite power. We point to similar power-posterior behavior even when the power changes in N but converges to a constant in $(0, 1)$. However, Miller and Dunson (2019) consider powers that converge to 0 in the limit of $N \to \infty$. It remains to investigate this case, especially for more general sequences of powers converging to zero, beyond the particular sequence suggested by Miller and Dunson (2019).

Accelerating scientific applications with multi-fidelity MCMC. In Chapter 5, we outlined a recipe for constructing multi-fidelity MCMC algorithms and demonstrated the approach on a few proof-of-concept examples. More specialized study of this approach in particular problem subclasses and for specific scientific applications remains. In particular, we expect many applications in the physical sciences, including applications in the quantum sciences, material science, structural design, and astrophysics. One extension of the current method is to apply the method to models with multiple types of computations that need to be approximated. Incorporating randomized fidelities for each computation into the Markov chain is then a straightforward extension, but choosing an appropriate truncation distribution will require more care.

Concluding remarks. While there have been many exciting advances in probabilistic machine learning in recent years, there remains much work to be done in order to understand how these methods behave when various assumptions are violated, which is often the case in real-world applications. Insufficient understanding of these behaviors could lead to misleading or fundamentally inaccurate inferences that may be then propagated into downstream tasks, such as high-stakes decisions. Thus, it increasingly important to develop reliable data analysis tools that account for misspecification.

Appendix A

Supplementary material: foundations of probabilistic modeling

A.1 The general version of Schwartz's theorem

We now provide a proof of Theorem 2.1.9. First, rewrite the posterior distribution of the complement of the neighborhood U as:

$$\Pi(U^c \,|\, X_1, \ldots, X_n) = \frac{\int_{U^c} \prod_{i=1}^n p(X_i)\, d\Pi(p)}{\int_{\mathcal{P}} \prod_{i=1}^n p(X_i)\, d\Pi(p)} = \frac{\int_{U^c} \prod_{i=1}^n \frac{p(X_i)}{p_0(X_i)}\, d\Pi(p)}{\int_{\mathcal{P}} \prod_{i=1}^n \frac{p(X_i)}{p_0(X_i)}\, d\Pi(p)}. \tag{A.1}$$

Since the test functions $\phi_n \in [0, 1]$, we can upper bound the posterior from above as

$$\Pi(U^c \,|\, X_1, \ldots, X_n) \leq \Pi(U^c \,|\, X_1, \ldots, X_n) + \phi_n(1 - \Pi(U^c \,|\, X_1, \ldots, X_n)) \tag{A.2}$$

$$= \phi_n + \frac{(1 - \phi_n) \int_{U^c} \prod_{i=1}^n \frac{p(X_i)}{p_0(X_i)}\, d\Pi(p)}{\int_{\mathcal{P}} \prod_{i=1}^n \frac{p(X_i)}{p_0(X_i)}\, d\Pi(p)}. \tag{A.3}$$

By Markov's inequality, the assumption $f_0^{(n)}(\phi_n) \leq e^{-Cn}$ implies that $\sum_{n \geq 1} f_0^{(n)}(\phi_n > e^{-Cn}) <$

∞. And so the Borel–Cantelli lemma then implies that

$$f_0^{(\infty)}(\phi_n > e^{-Cn} \text{ occurs for infinitely many } n) = 0.$$

Hence, the first term in the sum above $\phi_n \to 0$ almost surely under $f_0^{(\infty)}$ as $n \to \infty$.

It remains to show that the second term has the appropriate behavior. The goal is to control the behavior of the numerator and the denominator of the second term of the sum such that (with $f_0^{(\infty)}$-probability 1) this ratio converges to 0.

Step 1: Define $K_\epsilon(p_0) := \{p \in \mathcal{P} : \mathrm{KL}(p_0; p) < \epsilon\}$. The KL-support condition on the prior Π – i.e., for all $\epsilon > 0$, $\Pi(K_\epsilon(p_0)) > 0$ – is used to show that for any $\beta > 0$,

$$\liminf_{n\to\infty} \; e^{n\beta} \int_p \prod_{i=1}^n \frac{p(X_i)}{p_0(X_i)} \, d\Pi(p) = \infty \quad f_0^{(\infty)}\text{-a.s.} \tag{A.4}$$

First we rewrite the expression as an exponential of the log of the product, which allows us to transform this into a sum of logarithms:

$$e^{n\beta} \int_p \exp\left(\log\left(\prod_{i=1}^n \frac{p(X_i)}{p_0(X_i)}\right)\right) d\Pi(p) = e^{n\beta} \int_p \exp\left(-\sum_{i=1}^n \log\left(\frac{p_0(X_i)}{p(X_i)}\right)\right) d\Pi(p) \tag{A.5}$$

$$\geq e^{n\beta} \int_{K_\epsilon(p_0)} \exp\left(-\sum_{i=1}^n \log\left(\frac{p_0(X_i)}{p(X_i)}\right)\right) d\Pi(p). \tag{A.6}$$

The second line holds because we are restricting the integral to a smaller set $K_\epsilon(p_0)$.

Fatou's lemma gives us an inequality between the liminf outside and the liminf inside the integral:

$$\liminf_{n\to\infty} e^{n\beta} \int_p \exp\left(\log\left(\prod_{i=1}^n \frac{p(X_i)}{p_0(X_i)}\right)\right) d\Pi(p) \tag{A.7}$$

$$\geq \liminf_{n\to\infty} e^{n\beta} \int_{K_\epsilon(p_0)} \exp\left(-\sum_{i=1}^n \log\left(\frac{p_0(X_i)}{p(X_i)}\right)\right) d\Pi(p), \tag{A.8}$$

$$\geq \int_{K_\epsilon(p_0)} \liminf_{n\to\infty} \exp\left(n\beta - \sum_{i=1}^n \log\left(\frac{p_0(X_i)}{p(X_i)}\right)\right) d\Pi(p), \tag{A.9}$$

Now consider the integrand in the line above and its limiting behavior (thus evaluating its liminf). The strong law of large numbers implies that with $f_0^{(\infty)}$-probability 1, for all $p \in K_\epsilon(p_0)$,

$$-\frac{1}{n}\sum_{i=1}^n \log\left(\frac{p_0(X_i)}{p(X_i)}\right) \overset{n\to\infty}{\longrightarrow} -f_0 \log\left(\frac{p_0}{p}\right) = -\int \log\left(\frac{p_0(x)}{p(x)}\right) p_0(x) d\mu(x) = -K(p_0, p) \geq -\epsilon.$$

where the inequality holds since $p \in K_\epsilon(p_0)$,

This then implies that with $f_0^{(\infty)}$-probability 1,

$$\exp\left(n\left[\beta - \frac{1}{n}\sum_{i=1}^n \log\left(\frac{p_0(X_i)}{p(X_i)}\right)\right]\right) \overset{n\to\infty}{\longrightarrow} \infty. \tag{A.10}$$

Plugging this into the liminf above and using the KL support condition, i.e., $\Pi(K_\epsilon(p_0)) > 0$ and applying Fubini's theorem, we have

$$\liminf_{n\to\infty} e^{n\beta} \int_p \prod_{i=1}^n \frac{p(X_i)}{p_0(X_i)} d\Pi(p) \tag{A.11}$$

$$\geq \int_{K_\epsilon(p_0)} \liminf_{n\to\infty} \exp\left(n\left[\beta - \frac{1}{n}\sum_{i=1}^n \log\left(\frac{p_0(X_i)}{p(X_i)}\right)\right]\right) d\Pi(p) = \infty \quad f_0^{(\infty)}\text{-a.s.}, \tag{A.12}$$

and so the conclusion holds.

Step 2: The existence of a uniformly sequence of tests is used to show that

$$\lim_{n\to\infty} e^{n\beta_0}(1-\phi_n)\int_{U^c}\prod_{i=1}^{n}\frac{p(X_i)}{p_0(X_i)}\,d\Pi(p)=0 \quad f_0^{(\infty)}\text{-a.s.} \tag{A.13}$$

By Fubini's theorem, we can exchange the expectation and integral

$$\begin{aligned}
f_0^{(n)}\left((1-\phi_n)\int_{U^c}\prod_{i=1}^{n}\frac{p(X_i)}{p_0(X_i)}\,d\Pi(p)\right) &= \int(1-\phi_n)\int_{U^c}\prod_{i=1}^{n}\frac{p(X_i)}{p_0(X_i)}\,d\Pi(p)\,df_0^{(n)} && \text{(A.14)}\\
&= \int_{U^c}\int(1-\phi_n)\prod_{i=1}^{n}\frac{p(X_i)}{p_0(X_i)}\,df_0^{(n)}\,d\Pi(p) && \text{(A.15)}\\
&= \int_{U^c}\int(1-\phi_n)\prod_{i=1}^{n}p(X_i)d\mu\,d\Pi(p) && \text{(A.16)}\\
&= \int_{U^c}\int(1-\phi_n)\,df^{(n)}\,d\Pi(p) && \text{(A.17)}\\
&= \int_{U^c}f^{(n)}(1-\phi_n)\,d\Pi(p) && \text{(A.18)}\\
&\lesssim e^{-Cn}, && \text{(A.19)}
\end{aligned}$$

where the last line follows from our assumption that for some $C>0$,

$$\sup_{p\in U^c} f^{(n)}(1-\phi_n) < e^{-Cn}.$$

And so for some $\beta_0>0, C>0$,

$$f_0^{(n)}\left(e^{n\beta_0}(1-\phi_n)\int_{U^c}\prod_{i=1}^{n}\frac{p(X_i)}{p_0(X_i)}\,d\Pi(p)\right) < e^{-n(C-\beta_0)}. \tag{A.20}$$

Since Markov's inequality implies that $\sum_{n\geq 1} e^{-n(C-\beta_0)}<\infty$ for $C>\beta_0$, by the Borel–Cantelli lemma, the numerator goes to 0 almost surely. Finally, choosing $\beta=\beta_0$, the ratio goes to 0 a.s.

Appendix B

Supplementary material: finite mixture models

B.1 Finite mixture models with an upper bound on the number of components

In this section we consider a modification of the setting from the main chapter in which the prior Π has support on only those finite mixtures with at most $\tilde{k}$ components. We start by stating and proving our main result in this finite-support case. Then we discuss why our conditions have changed slightly from Definition 3.2.1. Finally we demonstrate our finite-support theory in practice with an experiment.

B.1.1 Result and proof

Let $\mathbb{F}(k)$ be the set of finite mixtures with exactly k components for $k \leq \tilde{k}$. We can apply the same proof technique in Section 3.4 to the present case, provided that the mixture-density posterior concentrates on weak neighborhoods of some compact subset of $\tilde{k}$-mixtures.

Theorem B.1.1. *Suppose that the prior* Π *has support on only those mixtures with at most* $\tilde{k}$ *components. Assume that:*

1. *The posterior concentrates on weak neighborhoods of a weak-compact subset of* $\mathbb{F}(\tilde{k})$*, and*
2. Ψ *is continuous, is mixture-identifiable, and has degenerate limits.*

Then the posterior on the number of components concentrates on $\tilde{k}$*:*

$$\Pi(\tilde{k} \,|\, X_{1:N}) \overset{N\to\infty}{\longrightarrow} 1 \quad f_0\text{-}a.s. \tag{B.1}$$

Proof Sketch. By assumption, the posterior concentrates on weak neighborhoods of some weak-compact subset $\mathcal{A} \subseteq \mathbb{F}(\tilde{k})$. It remains to show that there exists a weak neighborhood U of $\mathcal{A}$ that, for all $k < \tilde{k}$, contains no k-mixtures of the family of Ψ. Suppose the contrary, i.e., that every such neighborhood contains a mixture of strictly less than $\tilde{k}$ components; then we can construct a sequence $(f_i)_{i=1}^{\infty}$ of mixtures of strictly less than $\tilde{k}$ components such that $f_i \Rightarrow \mathcal{A}$ (in the sense that the infimum of the weak metric between f_i and elements of $\mathcal{A}$ converges to 0). Let g_i be the corresponding sequence of mixing measures such that $f_i = F(g_i)$. Now we follow step 2 of the proof of the main theorem, with some slight modifications to account for the fact that f_i converges weakly to a set rather than a single density. Suppose that $g_i(\Theta \setminus K) \to 0$ for some compact subset $K \subseteq \Theta$. Then following the proof of the main theorem, we have that $\mathbb{F}_K$ and $\mathbb{G}_K$ are compact, and so there is a weak-convergent subsequence of $F(\hat{g}_{i,K})$ that converges to some f_0; since $\mathcal{A}$ is weak-closed, $f_0 \in \mathcal{A}$. The remainder of this branch of the proof then follows the proof main theorem directly. Now for the other branch, suppose $g_i(\Theta \setminus K) \not\to 0$ for any compact $K \subseteq \Theta$. Then as in the main proof there is a sequence of parameters that is not relatively compact; so the corresponding sequence of components ψ_i is either not tight or not μ-wide. Since $\mathcal{A}$ is weak-compact, by Prokhorov's theorem $\mathcal{A}$ is tight, so f_i must be tight, so ψ_i must be tight. On the other hand, ψ_i also must be μ-wide, since otherwise replacing it with the singular sequence ϕ_i shows that f_i would not converge weakly to $\mathcal{A}$. This concludes the second branch of the proof, and the result follows. □

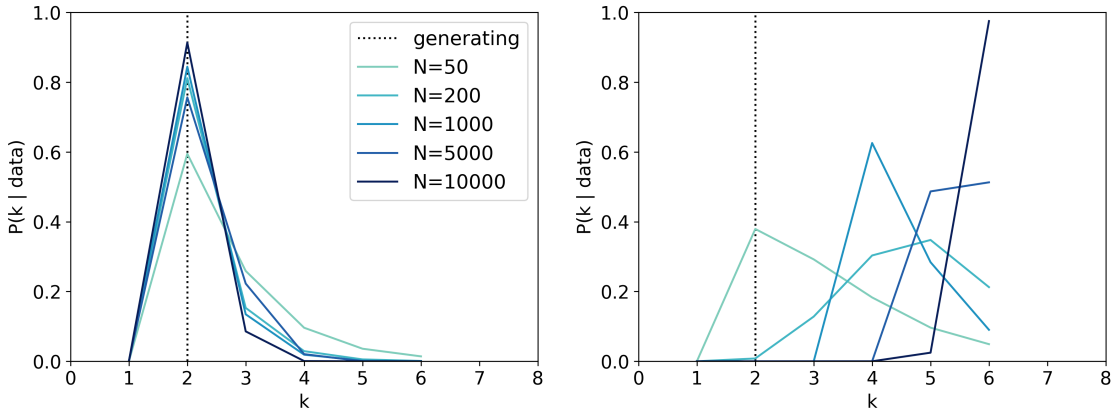
B.1.2 Discussion of the weak concentration condition

Our main result in Definition 3.2.1 uses a KL support condition to guarantee weak concentration of the posterior. In contrast, in Definition B.1.1, we do not impose any KL support condition and instead just directly assume weak posterior concentration for the mixture density. First we discuss why this assumption remains reasonable and then discuss why we chose to change the condition.

Reasonableness of the condition Note that the new weak-concentration assumption is actually weaker than the KL condition in the main chapter—albeit potentially substantially more difficult to verify. As a simple example of why this assumption is reasonable, suppose we obtain data generated from a Laplace distribution, and we use a mixture model with Gaussian components and a prior that asserts that the mixture has at most 10 components. Then we expect the posterior to concentrate on mixture densities that have exactly 10 components, and in particular, the set of KL-closest mixture densities to the Laplace. Although many examples will have a single closest such density, we state Definition B.1.1 in such a way that it allows for the case where the posterior concentrates on a *compact set* of densities (usually due to symmetry in the model).

Why change the condition In the main text, we assume—via the KL support condition, Definition 3.3.1—that the infimum of the KL divergence from the data generating distribution f_0 to mixture distributions from the model is 0. In other words, we must be able to approximate f_0 arbitrarily well using mixture distributions from the model. However, in the setting with a bounded number of components, this assumption typically does not hold. In particular, the infimum KL from the data-generating distribution f_0 to mixture distributions in the model is nonzero. For example, in the previous Laplace versus Gaussian mixture example, we require an unbounded number of components to achieve a vanishing KL divergence. If we are limited to 10 components, the infimum KL is nonzero.

Demonstrating weak consistency with a reasonable amount of generality when the KL support condition does not hold is challenging; see for instance, Kleijn (2003, Lemma 2.8) and Ramamoorthi et al. (2015, Remark 4). Thus, we opt to require that weak concentration be verified directly for each



(a) Gaussian mixture data

(b) Laplace mixture data

Figure B.1.1: Well-specified and misspecified component families that use a prior with an upper bound on the number of components given by $k \sim \text{Unif}\{1, \ldots, 6\}$. Posterior values for component counts k with $k > 6$ are all zero, so we do not plot them.

particular applied setting of interest, rather than attempting to develop a general set of sufficient conditions. The fact that we directly require weak convergence also means that we do not need to make any stipulations about how data are generated. Therefore, in contrast to the main theorem, we do not impose any such assumptions.

B.1.3 Experiments

Now we demonstrate that the asymptotic behavior described by our theory occurs in practice. In order to the study both the well-specified and misspecified cases, we consider the same 2-component Gaussian and Laplace data described in Section 3.8.1. Here, we set the prior on the number of components to be a uniform distribution on $\{1, \ldots, 6\}$. The resulting posterior number of components appears in Figure B.1.1. Here the well-specified model (Gaussian data) is consistent and concentrates on the true generating number of components as N grows (Rousseau and Mengersen, 2011). On the other hand, in the misspecified model (Laplace data), the posterior concentrates on the largest possible number of components under the prior, in this case given by $\tilde{k} = 6$.

B.2 Proof of Proposition 2.2

Consider the multivariate Gaussian family $\Psi = \left\{\mathcal{N}(\nu, \Sigma) : \nu \in \mathbb{R}^d, \Sigma \in \mathbb{S}^d_{++}\right\}$ with parameter space $\Theta = \mathbb{R}^d \times \mathbb{S}^d_{++}$, equipped with the topology induced by the Euclidean metric. Let $(\lambda_j(\Sigma))_{j=1}^d$ denote the eigenvalues of the covariance matrix $\Sigma \in \mathbb{S}^d_{++}$ that satisfy $\infty > \lambda_1(\Sigma) \geq \ldots \geq \lambda_d(\Sigma) > 0$. Since the family of Gaussians is continuous and mixture-identifiable (Yakowitz and Spragins, 1968, Proposition 2), the main condition we need to verify is that the family has degenerate limits (Definition 3.3.5). A useful fact is that if a sequence of Gaussian distributions is tight, then the sequence of means and the eigenvalues of the covariance matrix is bounded.

Lemma B.2.1. *Let $(\psi_i)_{i\in\mathbb{N}}$ be a sequence of Gaussian distributions with mean $\nu_i \in \mathbb{R}^d$ and covariance $\Sigma_i \in \mathbb{S}^d_{++}$. If $(\psi_i)_{i\in\mathbb{N}}$ is a tight sequence of measures, then the sequences $(\nu_i)_{i\in\mathbb{N}}$ and $(\lambda_1(\Sigma_i))_{i\in\mathbb{N}}$ are bounded.*

Proof. Let Y_i denote a random variable with distribution ψ_i. For each covariance matrix Σ_i, consider its eigenvalue decomposition $\Sigma_i = U_i \Lambda_i U_i^\top$, where $U_i \in \mathbb{R}^{d\times d}$ is an orthonormal matrix and $\Lambda_i \in \mathbb{R}^{d\times d}$ is a diagonal matrix. Then the random variable $Z_i = U_i^\top Y_i$ has distribution $\mathcal{N}(U_i^\top \nu_i, \Lambda_i)$. If either $\|\nu_i\|_2 = \|U_i^\top \nu_i\|_2$ is unbounded or $\|\Lambda_i\|_F$ is unbounded, then Z_i is not tight (Billingsley, 1986, Example 25.10). Since Z_i and Y_i lie in any ball centered at the origin with the same probability, Y_i is not tight. □

We now show that the multivariate Gaussian family has degenerate limits.

Proof of Proposition 3.2.2. If the parameters $(\theta_i)_{i\in\mathbb{N}}$ are not a relatively compact subset of Θ, then either some coordinate of the sequence of means ν_i diverges, $\lambda_1(\Sigma_i) \to \infty$, or $\lambda_d(\Sigma_i) \to 0$. If some coordinate of the mean ν_i diverges or the maximum eigenvalue diverges, i.e., $\lambda_1(\Sigma_i) \to \infty$, then the sequence (ψ_{θ_i}) is not tight by Definition B.2.1. On the other hand, if $\lambda_d(\Sigma_i) \to 0$ as $i \to \infty$, then ψ_{θ_i} converges weakly to a sequence of degenerate Gaussian measures that concentrate on $C_i = \left\{x \in \mathbb{R}^d : (x - \nu_i)^\top u_{d,i} = 0\right\}$, where $u_{d,i}$ is the d^{th} eigenvector of Σ_i. Note that $\mu(C_i) = 0$ for Lebesgue measure μ; so if we define $C = \cup_i C_i$ in the setting of Definition 3.3.4, the sequence is not μ-wide. □

We can generalize Definition 3.2.2 beyond multivariate Gaussians to mixture-identifiable location-scale families, as shown in Definition B.2.2. Examples of such families include the multivariate Gaussian family, the Cauchy family, the logistic family, the von Mises family, and generalized extreme value families. The proof is similar to that of Definition 3.2.2.

Proposition B.2.2. *Suppose Ψ is a location-scale family that is mixture-identifiable and absolutely continuous with respect to Lebesgue measure μ, i.e.,*

$$\frac{d\Psi}{d\mu} = \left\{ |\Sigma|^{-1/2} \varphi\left(\Sigma^{-1/2}(x-\nu)\right) : \nu \in \mathbb{R}^d, \Sigma \in \mathbb{S}^d_{++} \right\},$$

where $\varphi : \mathbb{R}^d \to \mathbb{R}$ is a probability density function. Then Ψ satisfies Definition 3.3.6.

B.3 Additional related work

Priors for microclustering behavior have been a recent focus in the Bayesian nonparametrics literature (Klami and Jitta, 2016; Zanella et al., 2016). Since having a fixed number of components across dataset sizes N would be incompatible with sublinear growth (in N) of cluster size across all clusters, we expect divergence issues similar in flavor to those in Miller and Harrison (2013, 2014).

Appendix C

Supplementary material: edge-exchangeable graphs and sparsity

C.1 Overview

In Appendix C.2, we provide more examples of graph models that are either vertex exchangeable or Kallenberg exchangeable. In Appendix C.3, we establish characterizations of edge exchangeability in graphs via existing notions of exchangeability for combinatorial structures such as random partitions and feature allocations. In Appendix C.4, we provide full proof details for the theoretical results in the main text.

C.2 More exchangeable graph models

Many popular graph models are vertex exchangeable. These models include the classic Erdős–Rényi model (Erdős and Rényi, 1959), as well as Bayesian generative models for network data, such as

the stochastic block model (Holland et al., 1983), the mixed membership stochastic block model (Airoldi et al., 2008), the infinite relational model (Kemp et al., 2006; Xu et al., 2007), the latent space model (Hoff et al., 2002), the latent feature relational model (Miller et al., 2009), the infinite latent attribute model (Palla et al., 2012), and the random function model (Lloyd et al., 2012). See Orbanz and Roy (2015) and Lloyd et al. (2012) for more examples and discussion.

Recently, a number of extensions to the Kallenberg-exchangeable model of Caron and Fox (2017), which builds on early work on bipartite graphs by Caron (2012), have also been developed. These models include extensions to stochastic block models (Herlau et al., 2016), mixed membership stochastic block models (Todeschini et al., 2016), and dynamic network models (Palla et al., 2016).

C.3 Characterizations of edge-exchangeable graph sequences

We introduced edge exchangeability, a new notion of exchangeability for graphs. Just as the Aldous-Hoover theorem provides a characterization of the distribution of vertex-exchangeable graphs, it is desirable to provide a characterization of edge exchangeability in graphs. Below we show how characterization theorems that already exist for other combinatorial structures can be readily applied to provide characterizations for edge exchangeability in graphs.

We first develop mappings from edge-exchangeable graph sequences to familiar combinatorial structures—such as partitions (Pitman, 1995), feature allocations (Broderick et al., 2013b), and trait allocations (Broderick et al., 2015; Campbell et al., 2018)—showing that edge exchangeability in the graph corresponds to exchangeability in those structures. In this manner, we provide characterizations of the case where one edge is added to the graph per step in Appendix C.3.1, where multiple unique edges may be added per step in Appendix C.3.1, and where multiple (non)unique edges may be added in Appendix C.3.1.

A limitation of these connections is that it is not immediately clear how to recover the connectivity in the graph from the mapped combinatorial object; for instance, given a particular feature allocation, the graph to which it corresponds is not identifiable. This issue has been addressed in a purely combinatorial context via *vertex allocations* and the *graph paintbox* (Campbell et al., 2018) using

the general theory of trait allocations. In Appendix C.3.2, we provide an alternative connection to *ordered* combinatorial structures (Broderick et al., 2013b; Campbell et al., 2018) under the assumption that vertex labels are provided. This assumption is often reasonable in the setting of network data where the vertices and edges are observed directly. By contrast, it is unusual to assume that labels are provided for blocks in the case of partitions, feature allocations, and trait allocations since, in these cases, the combinatorial structure is typically entirely latent in real data analysis problems. For instance, in clustering applications, finding parameters that describe each cluster is usually part of the inference problem. In the graph case, though, the use of an ordered structure identifies the particular pair of vertices corresponding to each edge in the graph, allowing recovery of the graph itself.

C.3.1 The step collection sequence and connections to other forms of combinatorial exchangeability

In order to analyze edge-exchangeable graphs using the existing combinatorial machinery of random partitions, feature allocations, and trait allocations, we introduce a new combinatorial structure, the step collection sequence, which can take the form of a sequence of partitions, feature allocations, or trait allocations. As we will now see, the step collection sequence can be constructed from the step-augmented graph sequence in the following way.

Suppose we assign a unique label ϕ to each pair of vertices. Then if a pair of vertices is labeled ϕ, we may imagine that any particular edge between this pair of vertices is assigned label ϕ when it appears. Let ϕ_j be the jth such unique edge label.

Recall that we consider a sequence of graphs defined by its step-augmented edge sequence E'_n. Let S_j be the set of steps up to the current step n in which any edge labeled ϕ_j was added. If m edges labeled ϕ_j were added in a single step s, s appears in S_j with multiplicity m. So each element $s \in S_j$ is an element of $[n]$. Let K_n be the number of unique vertex pairs seen among edges introduced up until the current step n. Then we may define C_n to be the collection of step sets

across edges that have appeared by step n:

$$C_n = \{S_1, \ldots, S_{K_n}\}.$$

Finally, we can define the *step collection sequence* $C = (C_1, C_2, \ldots)$ as the sequence of C_n for $n = 1, 2, \ldots$. Note that it is not clear how to recover the original edge connectivity of the graph from the step collection sequence, or whether it is possible to modify the sequence (or the labels ϕ_j) such that it is easy to recover connectivity while maintaining the (non-trivial) connections to combinatorial exchangeability provided in Appendix C.3.1 below.

Example C.3.1. *Suppose we have the edge sequence*

$$\begin{aligned}
E_1 &= \{\{2,3\},\{3,6\}\},\\
E_2 &= \{\{2,3\},\{3,6\}\},\\
E_3 &= \{\{2,3\},\{3,6\},\{6,6\},\{3,6\}\},\\
E_4 &= \{\{2,3\},\{1,4\},\{3,6\},\{6,6\},\{3,6\}\},
\end{aligned}$$

with step-augmentation

$$E_4' = \{(\{2,3\},1),(\{1,4\},4),(\{3,6\},1),(\{6,6\},3),(\{3,6\},3)\}$$

for E_4. Now we label the unique edges in E_n'. Using an order of appearance scheme Broderick et al. (2013b) to index the labels, E_4' becomes

$$\{(\phi_1,1),(\phi_2,1),(\phi_3,3),(\phi_1,3),(\phi_4,4)\},$$

where the labels ϕ_j correspond to the four unique vertex pairs: $\phi_1 = \{3,6\}, \phi_2 = \{2,3\}, \phi_3 =$

$\{6,6\}, \phi_3 = \{1,4\}$. *The step collection sequence for* $C_1, \ldots, C_4$ *is*

$$C_1 = \{\underbrace{\{1\}}_{\phi_1}\}, \qquad C_2 = \{\underbrace{\{1\}}_{\phi_1}\}, \qquad C_3 = \{\underbrace{\{1,3\}}_{\phi_1}, \underbrace{\{3\}}_{\phi_3}\}, \qquad C_4 = \{\underbrace{\{1,3\}}_{\phi_1}, \underbrace{\{3\}}_{\phi_3}, \underbrace{\{4\}}_{\phi_4}\}.$$

Here each element of C_n *is a set corresponding to one of the four unique labels* ϕ_j *and contains all step indices up to step* n *in which an edge with that label was added to the graph sequence.*

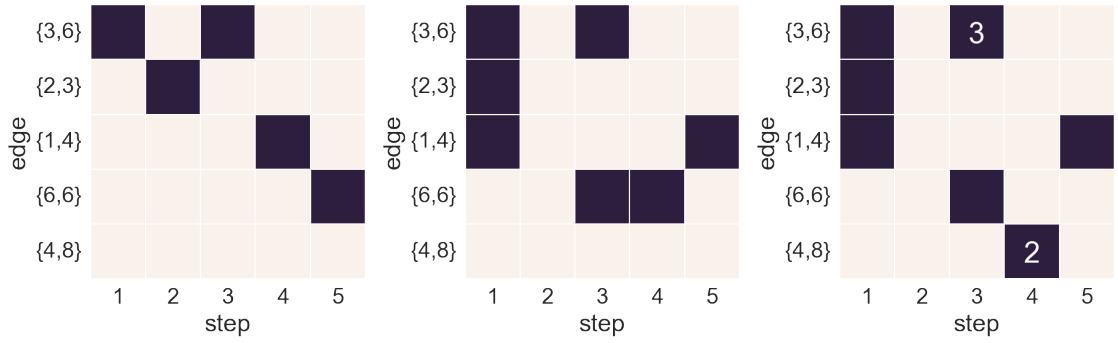
To see that the step collection sequence can be interpreted as a familiar combinatorial object, we recall the following definitions. A *partition* C_n of $[n]$ is a set $\{S_1, \ldots, S_{K_n}\}$ whose blocks, or *clusters*, are mutually exclusive, i.e., $S_i \cap S_j = \emptyset, i \neq j$, and exhaustive, i.e., $\bigcup_j S_j = [n]$. Feature allocations relax the definition of partitions by no longer requiring the blocks to be mutually exclusive and exhaustive. A *feature allocation* C_n of $[n]$ is a multiset $\{S_1, \ldots, S_{K_n}\}$ of subsets of $[n]$, such that any datapoint in $[n]$ occurs in finitely many *features* S_j (Broderick et al., 2013b). A *trait allocation* generalizes the feature allocation where now each S_j, called a *trait*, may itself be a multiset (Broderick et al., 2015; Campbell et al., 2018).

We see that the step collection C_n can be interpreted as follows. If a single edge is added to the graph at each round, C_n is a partition of $[n]$, and the step collection sequence is a projective partition sequence. If at most one edge is added between any pair of vertices at each step, C_n is a feature allocation of $[n]$, and the step collection sequence is a projective sequence of feature allocations. In the most general case, when multiple edges may be added between any pair of vertices at each step, C_n is a trait allocation of $[n]$, and the step collection sequence is a projective sequence of trait allocations.

In the following examples, corresponding to Figure C.3.1, we show different step collection sequences that correspond to a partition, a feature allocation, and a trait allocation.

Example C.3.2 (Partition)**.** *Consider the step collection* $C_5 = \{\{1,3\}, \{2\}, \{4\}, \{5\}\}$. *The edges form a partition of the steps. Here exactly one edge arrives in each step.*

Example C.3.3 (Feature allocation)**.** *Consider the step collection* $C_5 = \{\{1,3\}, \{1\}, \{1,5\}, \{3,4\}\}$. *This step collection forms a feature allocation of the steps. Thus in this case, there may be multiple*



(a) Partition (b) Feature allocation (c) Trait allocation

Figure C.3.1: Connection of edge-exchangeable graphs with partitions, feature allocations, and trait allocations. Light blocks represent 0, dark blocks either represent 1 or the specified count. In a partition, exactly one edge arrives in each step. In a feature allocation, multiple edges may arrive at each step, but at most one edge arrives between any two vertices at each step. In a trait allocation, there may be multiple edges of any type.

unique *edges arriving in each step.*

Example C.3.4 (Trait allocation)**.** *In a trait allocation, there may be multiple edges (not necessarily unique) at each step. Consider the step collection* $C_5 = \{\{1,3,3,3\},\{1\},\{1,5\},\{3\},\{4,4\}\}$*. This collection forms a trait allocation of the steps, where elements of* C_5 *are now multisets.*

In this section, we have connected certain types of edge-exchangeable graphs to partitions and feature allocations. In the next two sections, we make use of known characterizations of these combinatorial objects to characterize edge exchangeability in graphs.

Partition connection

First consider the connection to partitions. In this case, suppose that each index in $[n]$ appears exactly once across all of the subsets of C_n. This assumption on C_n is equivalent to assuming that in the original graph sequence $E_1, E_2, \ldots$, we have that E_{n+1} always has exactly one more edge than E_n. In this case, C_n is exactly a *partition* of $[n]$; that is, C_n is a set of mutually exclusive and exhaustive subsets of $[n]$. If the edge sequence (E_n) is random, then (C_n) is random as well.

We say that a partition sequence $C_1, C_2, \ldots$, where C_n is a (random) partition of $[n]$ and $C_m \subseteq C_n$ for all $m \leq n$, is infinitely exchangeable if, for all n, permuting the indices in $[n]$ does not change the

distribution of the (random) partitions (Pitman, 1995). Permuting the indices $[n]$ in the partition sequence (C_m) corresponds to permuting the order in which edges are added in our graph sequence (E_m). As an example of a model that generates a step collection sequence corresponding to a partition sequence, consider the frequency model we introduced in Section 4.3 where the weights are normalized. At each step, we choose a single edge according the resulting probability distribution over pairs of vertices.

Given this connection to exchangeable partitions, the *Kingman paintbox theorem* (Kingman, 1978) provides a characterization of edge exchangeability in graph sequences that introduce one edge per step: in particular, it guarantees that a graph sequence that adds exactly one edge per step is edge exchangeable if and only if the associated step collection sequence (C_n) has a Kingman paintbox representation. An alternate characterization of edge exchangeability in graph sequences that introduce one edge per step is provided by *exchangeable partition probability functions (EPPFs)* (Pitman, 1995). In particular, a graph sequence that introduces one edge per step is edge-exchangeable if and only if the marginal distribution of C_n (the step collection at step n) is given by an EPPF for all n.

Feature allocation connection

Next we notice that it need not be the case that exactly one edge is added at each step of the graph sequence, e.g. between E_n and E_{n+1}. If we allow multiple unique edges at any step, then the step collection C_n is just a set of subsets of $[n]$, where each subset has at most one of each index in $[n]$. Suppose that any m belongs to only finitely many subsets in C_n for any n. That is, we suppose that only finitely many edges are added to the graph at any step. Then C_n is an example of a *feature allocation* (Broderick et al., 2013b). Again, if (E_n) is random, then (C_n) is random as well.

We say that a (random) feature allocation sequence (C_m) is infinitely exchangeable if, for any n, permuting the indices of $[n]$ does not change the distribution of the (random) feature allocations Broderick et al. (2013a,b). Permuting the indices $[n]$ in the sequence (C_m) corresponds to permuting the steps when edges are added in the edge sequence (E_m). Consider the following example of a graph frequency model that produces a step collection sequence corresponding to an exchangeable

feature allocation. For $n = 1, 2, \ldots$, we draw whether the graph has an edge $\{i, j\}$ at time step n as Bernoulli with probability $w_{\{i,j\}} = w_i w_j$. Thus, in each step, we draw at most one edge per unique vertex pair. But we may draw multiple edges in the same step.

Similarly to the partition case in Appendix C.3.1, we can apply known results from feature allocations to characterize edge exchangeability in graph models of this form. For instance, we know that the *feature paintbox* Broderick et al. (2013b); Campbell et al. (2018) characterizes distributions over exchangeable feature allocations (and therefore the step collection sequence for graphs of this form) just as the Kingman paintbox characterizes distributions over exchangeable partitions (and therefore the step collection sequence for edge-exchangeable graphs with exactly one new edge per step).

We may also consider feature paintbox distributions with extra structure. For instance, the step collection sequence is said to have an *exchangeable feature probability function* (EFPF) (Broderick et al., 2013b) if the probability of each step collection C_n in the sequence can be expressed as a function only of the total number of steps n and the subset sizes within C_n (i.e. the edge multiplicities in the graph), and is symmetric in the subset sizes. As another example, the step collection sequence is said to have a *feature frequency model* if there exists a (random) sequence of probabilities $(w_j)_{j=1}^{\infty}$ associated with edges $j = 1, 2, \ldots$ and a number $\lambda > 0$, conditioned on which the step collection sequence arises from the graph built by adding edge j at each step independently[1] with probability w_j for all values of $j \in \mathbb{N}$, along with an additional Poiss(λ) number of edges that never share a vertex with any other edge in the sequence. In other words, the graph is constructed with a graph frequency model as in the main text of the present work (modulo the aforementioned additional Poisson number of edges). Theorem 17 ("Equivalence of EFPFs and feature frequency models") from Broderick et al. (2013b) shows that these two examples are actually equivalent: if the step collection sequence has an EFPF, it has a feature frequency model, and vice versa.

[1]This is conditional independence since the (w_j) may be random.

Further extensions

Finally, we may consider the case where at every step, any non-negative (finite) number of edges may be added *and* those edges may have non-trivial (finite) multiplicity; that is, the multiplicity of any edge at any step can be any non-negative integer. By contrast, in Appendix C.3.1, each unique edge occurred at most once at each step. In this case, the step collection C_n is a set of subsets of $[n]$. The subsets need not be unique or exclusive since we assume any number of edges may be added at any step. And the subsets themselves are multisets since an edge may be added with some multiplicity at step n. We say that C_n is a *trait allocation*, which we define as a generalization of a feature allocation where the subsets of C_n are multisets. As above, if (E_n) is random, (C_n) is as well.

We say that a (random) trait allocation sequence (C_m) is infinitely exchangeable if, for any n, permuting the indices of $[n]$ does not change the distribution of the (random) trait allocation. Here, permuting the indices of $[n]$ corresponds to permuting the steps when edges are added in the edge sequence (E_m). A graph frequency model that generates a step collection sequence as a trait allocation sequence is the multiple-edge-per-step frequency model sampling procedure described in Section 4.3. Here, at each step, multiple edges can appear each with multiplicity potentially greater than 1, requiring the full generality of a trait allocation sequence.

Campbell et al. (2018) characterize exchangeable trait allocations via, e.g., probability functions and paintboxes and thereby provide a characterization over the corresponding step collection sequences of such edge-exchangeable graphs.

C.3.2 Connections to exchangeability in ordered combinatorial structures

As noted earlier, it is not immediately clear how to recover the connectivity in an edge-exchangeable graph from the step collection sequence, nor how to do so in a way that preserves non-trivial connections to other exchangeable combinatorial structures. Campbell et al. (2018) considers an alternative to the step collection sequence in which the (multi)subsets in the combinatorial

structure correspond to *vertices* rather than edges, known as a *vertex allocation*. This allows for the characterization of edge-exchangeable graphs via the *graph paintbox* using the general theory of trait allocations, while maintaining an explicit representation of the structure of the graph, i.e., the connection between edges that share a vertex.

If we are willing to eschew the unordered nature of the step collection sequence, and assume that we have an a priori labeling on the vertices, there is yet another alternative using the *ordered step collection sequence*. The availability of labeled vertices is often a reasonable assumption in the setting of network data, where the vertices and edges are typically observed directly. Suppose the vertices are labeled using the natural numbers $1, 2, \ldots$. Then we can use the ordering of the vertex labels to order the vertex pairs in a diagonal manner, i.e. $\{1,1\}$, $\{1,2\}$, $\{2,2\}$, $\{1,3\}$, $\{2,3\}, \ldots$. Note that, for the purpose of building this diagonal ordering, we consider the lowest-valued index in each vertex pair first. We build the step collection sequence (C_n) in the same manner as before, except that each step collection C_n is no longer an unordered collection of subsets; the subsets derive their order from the vertex pairs they represent. For example, if we observe edges at vertex pairs $\{1,1\}$ and $\{1,2\}$ at step 1, and edges at vertex pairs $\{1,1\}$ and $\{2,3\}$ at step 2, then

$$C_1 = \big(\{1\}, \{1\}, \emptyset, \emptyset, \ldots\big)$$

and

$$C_2 = \big(\{1,2\}, \{1\}, \emptyset, \emptyset, \{2\}, \emptyset, \ldots\big)\,.$$

Since we know the order of the subsets in each C_n as they relate to the vertex pairs in the graph and their connectivity, we can recover the graph sequence from the ordered step collection sequence (C_n). Exchangeability in an ordered step collection sequence means that the distribution is invariant to permutations of the indices within the subsets (although the ordering of the subsets themselves cannot be changed). Given this notion of exchangeability, the earlier connections to exchangeable partitions, feature allocations, and trait allocations remain true, modulo the fact that they must themselves be ordered. Broderick et al. (2013b) provides a paintbox characterization of ordered exchangeable feature allocations, thereby providing characterizations (via the earlier connections to

partitions and feature allocations) of edge-exchangeable graphs that add either one or multiple unique edges per step. Note that, in these cases, this is a full characterization of edge-exchangeable graphs, by contrast to Appendix C.3.1, where we provided a characterization only of edge exchangeability in graphs. We suspect that a similar characterization of edge-exchangeable graphs with multiple (non)unique edges per step is available by examining characterizations of exchangeable ordered trait allocations.

C.4 Proofs

The proof of the main theorem (Theorem 4.5.3) follows from a collection of lemmas below. Lemma 4.5.2 characterizes the expected number of vertices and edges; Lemma C.4.2 establishes a useful transformation of those expectations; and Lemma C.4.3 shows that the two sets of expectations are asymptotically equivalent, so it is enough to consider the transformed expectation. Lemma C.4.5 provides the asymptotics of the transformed expectations. Finally, Lemma 4.5.1 shows that the random sequences converge almost surely to their expectations, yielding the final result.

C.4.1 Preliminaries

Notation We first define the asymptotic notation used in the main chapter and appendix. We use the notation "a.s." to mean almost surely, or with probability 1. Let $(X_n)_{n\in\mathbb{N}}, (Y_n)_{n\in\mathbb{N}}$ be two random sequences. We say that $X_n \overset{\text{a.s.}}{=} O(Y_n)$ if $\limsup_{n\to\infty} \frac{X_n}{Y_n} < \infty$ a.s., and that $X_n \overset{\text{a.s.}}{=} \Omega(Y_n)$ if $Y_n \overset{\text{a.s.}}{=} O(X_n)$ a.s. We say that $X_n \overset{\text{a.s.}}{=} o(Y_n)$ if $\lim_{n\to\infty} \frac{X_n}{Y_n} = 0$ a.s. Lastly, we say that $X_n \overset{\text{a.s.}}{=} \Theta(Y_n)$ if $X_n \overset{\text{a.s.}}{=} O(Y_n)$ and $Y_n \overset{\text{a.s.}}{=} O(X_n)$.

Let V_n, E_n be the respective sets of active vertices and edges at step n in the multigraph, and $|V_n|, |E_n|$ be their respective cardinalities, as defined in the main text. We use the notation $\bar{V}_n$ and $\bar{E}_n$ to represent these analogous vertex and edge sets for the binary graph. Note that $\bar{V}_n$ is the same as V_n.

C.4.2 Graph moments

In this section, we give the expected number of vertices and expected number of edges for the multi- and binary graph cases. We begin by defining the degree D_i of vertex i in the multigraph and the degree $\bar{D}_i$ of vertex i in the binary graph, respectively, as

$$D_i = \sum_j M_{\{i,j\}} \qquad\qquad \bar{D}_i = \sum_j \mathbb{1}\left(M_{\{i,j\}} > 0\right). \tag{C.1}$$

Now we present the expected number of edges and vertices. We note that both the multi- and binary graphs have the same number of (active) vertices, and so their expectations are the same.

Lemma C.4.1 (4.5.2, main text)**.** *The expected number of vertices and edges for the multi- and binary graphs are*

$$\begin{aligned}
\mathbb{E}\left(|\bar{V}_n|\right) = \mathbb{E}\left(|V_n|\right) &= \int \left[1 - \exp\left(-\int 1 - (1-wv)^n \, \nu(\mathrm{d}v)\right)\right] \nu(\mathrm{d}w), \\
\mathbb{E}\left(|E_n|\right) &= \frac{n}{2} \iint wv \; \nu(\mathrm{d}w)\, \nu(\mathrm{d}v), \\
\mathbb{E}\left(|\bar{E}_n|\right) &= \frac{1}{2} \iint (1 - (1-wv)^n)\, \nu(\mathrm{d}w)\, \nu(\mathrm{d}v).
\end{aligned}$$

Proof. Using the tower property of conditional expectation and Fubini's theorem, we have that the expected number of vertices is

$$\mathbb{E}\left(|V_n|\right) = \mathbb{E}\left(\mathbb{E}\left(\sum_i \mathbb{1}(D_i > 0) \,\middle|\, \mathcal{W}\right)\right) = \mathbb{E}\left(\sum_i \mathbb{P}\left(D_i > 0 \,\middle|\, \mathcal{W}\right)\right),$$

followed by the definition of degree in Equation (C.1) and the binomial density,

$$\mathbb{E}\left(|V_n|\right) = \mathbb{E}\left(\sum_i \left[1 - \prod_j \mathbb{P}\left(M_{\{i,j\}} = 0 \,|\, \mathcal{W}\right)\right]\right) = \mathbb{E}\left(\sum_{w \in \mathcal{W}} \left[1 - \prod_{v \in \mathcal{W}\setminus\{w\}} (1-wv)^n\right]\right).$$

Using the Slivnyak-Mecke theorem (Definition 2.2.9),

$$\begin{aligned}\mathbb{E}\left(|V_n|\right) &= \int \mathbb{E}\left(1 - \prod_{v\in\mathcal{W}} (1-wv)^n\right)\nu(\mathrm{d}w) \\ &= \int \left[1 - \mathbb{E}\left(\exp\left(n\sum_{v\in\mathcal{W}}\log(1-wv)\right)\right)\right]\nu(\mathrm{d}w),\end{aligned}$$

and finally by Campbell's theorem (Definition 2.2.8) on the inner expectation,

$$\mathbb{E}\left(|V_n|\right) = \int \left[1 - \exp\left(-\int (1-(1-wv)^n)\,\nu(\mathrm{d}v)\right)\right]\nu(\mathrm{d}w).$$

For the expected number of edges, we can again apply the tower property and Fubini's theorem followed by repeated applications of Slivnyak-Mecke to the expectations to get:

$$\mathbb{E}(|E_n|) = \mathbb{E}\left(\mathbb{E}\left(\frac{1}{2}\sum_{i\neq j} M_{\{i,j\}}\big|\mathcal{W}\right)\right) = \frac{1}{2}\int \mathbb{E}\left(\sum_{v\in\mathcal{W}} nwv\right)\nu(dw) = \frac{n}{2}\int wv\nu(dw)\nu(dv).$$

The expected number of edges for the binary case is obtained similarly via Fubini and Slivnyak-Mecke:

$$\begin{aligned}\mathbb{E}(|\bar{E}_n|) &= \mathbb{E}\left(\frac{1}{2}\sum_{i\neq j} P(M_{\{i,j\}} > 0|\mathcal{W})\right) = \frac{1}{2}\,\mathbb{E}\left(\sum_{w\in\mathcal{W}, v\in\mathcal{W}\setminus\{w\}} (1-(1-wv)^n)\right) \\ &= \frac{1}{2}\int\int (1-(1-wv)^n)\,\nu(dw)\,\nu(dv).\end{aligned}$$

□

The asymptotic behavior of these quantities is difficult to derive directly due to the discreteness of the indices n. Therefore, we rely on a technique called *Poissonization*, which allows us to bypass this difficulty by instead considering a continuous analog of the quantities in order to get asymptotic behaviors. Below, we introduce primed notation $V_t', E_t', \bar{E}_t', D_{t,i}'$ to represent the Poissonized quantities for the vertices, multigraph edges, binary edges, and the degree of a vertex,

where the index t now represents a continuous quantity. These will be defined such that V'_N has the same asymptotic behavior as V_N, E'_N has the same asymptotic behavior as E_N, and so on.

Given $\mathcal{W}$, let Π_{ij} be the Poisson process generated with rate $w_i w_j$ if $i < j$ and rate 0 if $i = j$, and let $\Pi_{ji} = \Pi_{ij}$. Let $\Pi_i := \bigcup_{j=1}^{\infty} \Pi_{ij}$, which is a Poisson process with rate $u_i := \sum_{j:j\neq i} w_i w_j$ via Poisson process superposition (Kingman, 1993, Sec. 2.2). If we think of t as continuous time passing, the process Π_{ij} represents the times at which new edges are added between vertices i and j, and Π_i represents the times at which any new edges involving vertex i are added.

Thus, we define the Poissonized degree of vertex i in the multi- and binary graph cases, respectively, to be a function of the continuous parameter $t > 0$,

$$D'_{t,i} = |\Pi_i \cap [0,t]|, \qquad \bar{D}'_{t,i} = \sum_j \mathbb{1}\left(|\Pi_{ij} \cap [0,t]| > 0\right).$$

We can define the Poissonized graph quantities of interest using these two quantities:

$$|\bar{V}'_t| = |V'_t| = \sum_i \mathbb{1}(D'_{t,i} > 0), \qquad |E'_t| = \frac{1}{2}\sum_{i=1}^{\infty} D'_{t,i}, \qquad |\bar{E}'_t| = \frac{1}{2}\sum_i \bar{D}'_{t,i}.$$

Lemma C.4.2. *The expected number of Poissonized vertices and edges for the multi- and binary graphs is*

$$\begin{aligned}
\mathbb{E}\left(|V'_t|\right) &= \int \left[1 - \exp\left(-\int (1 - e^{-twv})\,\nu(dv)\right)\right] \nu(dw) \\
\mathbb{E}\left(|E'_t|\right) &= \frac{t}{2}\iint wv\,\nu(\mathrm{d}w)\,\nu(\mathrm{d}v) \\
\mathbb{E}\left(|\bar{E}'_t|\right) &= \frac{1}{2}\iint (1 - \exp(-twv))\,\nu(\mathrm{d}w)\,\nu(\mathrm{d}v).
\end{aligned}$$

Proof. For the expected number of Poissonized vertices, we apply the tower property and Fubini's theorem to get

$$\mathbb{E}\left(|V'_t|\right) = \mathbb{E}\left(\mathbb{E}\left(\sum_i \mathbb{1}(D'_{t,i} > 0) \,\middle|\, \mathcal{W}\right)\right) = \mathbb{E}\left(\sum_i 1 - \mathbb{P}\left(D_{t,i} = 0 \,|\, \mathcal{W}\right)\right).$$

Using the fact that $D'_{t,i}|\mathcal{W}$ is Poisson-distributed,

$$\mathbb{E}\left(|V'_t|\right) = \mathbb{E}\left(\sum_i 1 - \exp\left(-tu_i\right)\right) = \mathbb{E}\left(\sum_{w\in\mathcal{W}} 1 - \exp\left(-tw \sum_{v\in\mathcal{W}\setminus\{w\}} v\right)\right).$$

Finally, by the Slivnyak-Mecke theorem and Campbell's theorem,

$$\begin{aligned}\mathbb{E}\left(|V'_t|\right) &= \int \mathbb{E}\left(1 - \exp\left(-tw \sum_{v\in\mathcal{W}} v\right)\right) \nu(dw) \\ &= \int \left[1 - \exp\left(\int (e^{-twv} - 1)\,\nu(dv)\right)\right] \nu(dw).\end{aligned}$$

For the expected number of Poissonized edges, after applying Fubini's theorem and Slivnyak-Mecke we have

$$\begin{aligned}\mathbb{E}\left(|E'_t|\right) &= \mathbb{E}\left(\frac{1}{2}\sum_i D'_{t,i}\right) = \mathbb{E}\left(\frac{1}{2}\sum_i \mathbb{E}\left(D'_{t,i}|\mathcal{W}\right)\right) \\ &= \mathbb{E}\left(\frac{1}{2}\sum_i u_i\right) = \mathbb{E}\left(\frac{1}{2}\sum_{w\in\mathcal{W}, v\in\mathcal{W}\setminus\{w\}} wv\right) \\ &= \frac{1}{2}\int\int wv\,\nu(dw)\,\nu(dv).\end{aligned}$$

For the expected number of Poissonized edges in the binary case, noting that $|\Pi_{ij} \cap [0,t]|$ is Poisson-distributed with rate tw_iw_j, and applying Fubini's theorem and Slivnyak-Mecke, we have:

$$\begin{aligned}\mathbb{E}(|\bar{E}'_t|) &= \mathbb{E}\left(\mathbb{E}\left(\sum_i \bar{D}'_{t,i}|\mathcal{W}\right)\right) = \mathbb{E}\left(\sum_{w\in\mathcal{W}, v\in\mathcal{W}\setminus\{w\}} (1 - \exp(-twv))\right) \\ &= \int\int (1 - \exp(-twv))\,\nu(dw)\,\nu(dv).\end{aligned}$$

□

C.4.3 Asymptotics

We have defined the expected number of vertices and edges for the multigraph and binary graph cases (Lemma 4.5.2) and presented the Poissonized version of these expectations (Lemma C.4.2). We now show in Lemma C.4.3 that the expected graph quantities and their Poissonized expectations are asymptotically equivalent.

Lemma C.4.3. *The Poissonized expectations for the number of vertices and the number of edges in the multi- and binary graphs are asymptotically equivalent to the original expectations; i.e., as* $n \to \infty$,

$$\begin{aligned}
\mathbb{E}\left(|V_n'|\right) &\sim \mathbb{E}\left(|V_n|\right), \\
\mathbb{E}\left(|E_n'|\right) &\sim \mathbb{E}\left(|E_n|\right), \\
\mathbb{E}\left(|\bar{E}_n'|\right) &\sim \mathbb{E}\left(|\bar{E}_n|\right).
\end{aligned}$$

Proof. For the vertices, we have

$$\mathbb{E}\left(|V_n| - |V_n'|\right) = \int \left[\exp\left(-\int (1 - e^{-nwv})\,\nu(\mathrm{d}v)\right) - \exp\left(-\int (1 - (1 - wv)^n)\,\nu(\mathrm{d}v)\right)\right] \nu(\mathrm{d}w).$$

Using the elementary inequalities

$$\begin{aligned}
&0 \le e^{-nx} - (1-x)^n \le nx^2 e^{-nx}, \qquad x \in [0,1],\ n > 0 \\
&0 \le e^{-a} - e^{-b} \le b - a, \qquad 0 \le a \le b,
\end{aligned}$$

we have

$$0 \le \mathbb{E}\left(|V_n| - |V_n'|\right) \le \iint n(wv)^2 e^{-nwv}\,\nu(\mathrm{d}v)\,\nu(\mathrm{d}w). \tag{C.2}$$

Finally, note that

$$\forall n > 0, \forall w, v \in [0,1], \quad nwve^{-nwv} \leq e^{-1}$$

and

$$\iint e^{-1} wv \, \nu(\mathrm{d}w) \, \nu(\mathrm{d}v) = e^{-1} \left(\int w \, \nu(\mathrm{d}w) \right)^2 < \infty.$$

Therefore by Lebesgue dominated convergence,

$$0 \leq \lim_{n\to\infty} \mathbb{E}\left(|V_n| - |V_n'|\right) \leq \iint \lim_{n\to\infty} n(wv)^2 e^{-nwv} \, \nu(\mathrm{d}v) \, \nu(\mathrm{d}w) = 0,$$

so we have that $\lim_{n\to\infty} \mathbb{E}\left(|V_n| - |V_n'|\right) = 0$. Since $\mathbb{E}(|V_n|)$, $\mathbb{E}(|V_n'|)$ are monotonically increasing by inspection, $\mathbb{E}(|V_n|) \sim \mathbb{E}(|V_n'|)$, $n \to \infty$, as required.

For the binary graph edges,

$$\mathbb{E}\left(|\bar{E}_n| - |\bar{E}_n'|\right) = \frac{1}{2} \iint (\exp(-nwv) - (1-wv)^n) \, \nu(\mathrm{d}v) \, \nu(\mathrm{d}w).$$

Using the earlier elementary inequalities,

$$0 \leq \mathbb{E}\left(|\bar{E}_n| - |\bar{E}_n'|\right) = \frac{1}{2} \iint n(wv)^2 e^{-nwv} \, \nu(\mathrm{d}v) \, \nu(\mathrm{d}w).$$

This is (modulo the constant factor of 1/2) the exact expression in Equation (C.2). Therefore, the same analysis can be performed, and the result holds.

For multigraph edges,

$$\mathbb{E}\left(|E_n| - |E_n'|\right) = \frac{n}{2} \iint (wv - wv) \, \nu(\mathrm{d}v) \, \nu(\mathrm{d}w) = 0,$$

so $\mathbb{E}\left(|E_n|\right) \sim \mathbb{E}\left(|E_n'|\right)$, $n \to \infty$. □

Lemma C.4.3 allows us to analyze the asymptotics of the Poissonized expectations and apply the result directly to the asymptotics of the original graph quantities. To achieve the desired asymptotics for the Poissonized expectations, we will make a further assumption on the rate measure ν generating the vertex weights in Equation (4.2). Namely, we assume that the tails of ν decay at a rate that will yield the appropriate weight decay in the weights (w_j)—and thereby the appropriate decay in vertex creation to finally yield sparsity in the graph itself. In particular, the tail of a measure ν is said to be *regularly varying* if there exists a function $\ell : \mathbb{R}_+ \to \mathbb{R}_+$ and $\alpha \in (0, 1)$ such that

$$\int_x^1 \nu(\mathrm{d}w) \sim x^{-\alpha}\ell(x^{-1}), \quad x \to 0, \qquad\qquad \forall\, c > 0,\ \lim_{x\to\infty} \frac{\ell(cx)}{\ell(x)} = 1. \tag{C.3}$$

The condition on the function ℓ is equivalent to saying that ℓ is *slowly varying*. For additional details on regular and slow variation, see Feller (1971, VIII.8). An important equivalent formulation of Equation (C.3) that we will use in our following proof of the asymptotics of Poissonized expectations is provided by Lemma C.4.4 (see Gnedin et al. (2007, Prop. 13) and Broderick et al. (2012, Prop. 6.1) for the proof).

Lemma C.4.4 (Broderick et al. (2012); Gnedin et al. (2007))**.** *The tail of ν is regularly varying iff there exists a function $\ell : \mathbb{R}_+ \to \mathbb{R}_+$ and $\alpha \in (0, 1)$ such that*

$$\int_0^x u\nu(du) \sim x^{1-\alpha}\ell(x^{-1}), \quad x \to 0, \qquad\qquad \forall\, c > 0,\ \lim_{x\to\infty} \frac{\ell(cx)}{\ell(x)} = 1. \tag{C.4}$$

Lemma C.4.4 is often easier to use than Equation (C.3) when checking whether a particular measure ν has a regularly varying tail. For example, for the three-parameter beta process, we have

$$\begin{aligned}
\int_0^x u\nu(du) &= \gamma \frac{\Gamma(1+\beta)}{\Gamma(1-\alpha)\Gamma(\beta+\alpha)} \int_0^x u^{-\alpha}(1-u)^{\beta+\alpha-1} du \\
&\sim \gamma \frac{\Gamma(1+\beta)}{\Gamma(1-\alpha)\Gamma(\beta+\alpha)} \int_0^x u^{-\alpha} du, \quad x \downarrow 0 \\
&= \gamma \frac{\Gamma(1+\beta)}{\Gamma(1-\alpha)\Gamma(\beta+\alpha)} \frac{1}{1-\alpha} x^{1-\alpha},
\end{aligned}$$

so the tail of ν is regularly varying when the discount parameter α satisfies $\alpha \in (0, 1)$ with $\ell(x^{-1})$

equal to the constant function

$$\ell(x^{-1}) = \frac{\gamma}{1-\alpha}\frac{\Gamma(1+\beta)}{\Gamma(1-\alpha)\Gamma(\beta+\alpha)}. \tag{C.5}$$

Note that the two-parameter beta process does not exhibit this behavior (since in this case, $\alpha = 0$).

Given the two formulations of a measure ν with a regularly varying tail above, we are ready to characterize the asymptotics of the earlier Poissonized expectations.

Lemma C.4.5. *If the tail of ν is regularly varying as per Equation* (C.3)*, then as $n \to \infty$,*

$$\mathbb{E}\left(|V_n'|\right) = \Theta(n^\alpha \ell(n)), \qquad \mathbb{E}\left(|E_n'|\right) = \Theta(n), \qquad \mathbb{E}\left(|\bar{E}_n'|\right) = O\left(\ell(\sqrt{n}) \min\left(n^{\frac{1+\alpha}{2}}, \ell(n) n^{\frac{3\alpha}{2}}\right)\right).$$

Proof. Throughout this proof we use c to denote a constant; the precise value of c changes but is immaterial. We also define the tail of ν as $\bar{\nu}(x) := \int_x^1 \nu(\mathrm{d}w)$, for notational brevity. Furthermore, recall that we assume the rate measure ν satisfies $\int w\nu(dw) < \infty$.

We first examine the expected number of Poissonized vertices,

$$\mathbb{E}\left(|V_n'|\right) = \int \left[1 - \exp\left(-\int (1 - e^{-nwv})\nu(dv)\right)\right] \nu(dw),$$

by splitting the integral into two parts. First, by the assumption that the tail of ν is regularly varying,

$$\int_{n^{-1}}^1 \left[1 - \exp\left(-\int (1 - e^{-nwv})\nu(dv)\right)\right] \nu(dw) \leq \int_{n^{-1}}^1 \nu(dw) \sim cn^\alpha \ell(n). \tag{C.6}$$

Next, we upper bound the integral term

$$\begin{aligned}\int_0^{n^{-1}} \left[1 - \exp\left(-\int (1 - e^{-nwv})\nu(dv)\right)\right] \nu(dw) &\leq \int_0^{n^{-1}} \int (1 - e^{-nwv})\nu(\mathrm{d}v)\nu(\mathrm{d}w) \\ &\leq \int_0^{n^{-1}} \int nwv\nu(\mathrm{d}v)\nu(\mathrm{d}w) \\ &\leq \left(\int v\nu(\mathrm{d}v)\right) n \int_0^{n^{-1}} w\nu(\mathrm{d}w) \\ &\sim cn^{\alpha}\ell(n), \end{aligned} \tag{C.7}$$

where the asymptotic behavior in the last line follows from Lemma C.4.4. Thus, combining the upper bounds on Equation (C.6) and Equation (C.7) gives the bound for the entire integral: $\mathbb{E}\left(|V_n'|\right) = O(n^{\alpha}\ell(n))$.

Now we bound the integral below:

$$\begin{aligned}&\int_{n^{-1}}^{1} \left[1 - \exp\left(-\int (1 - e^{-nwv})\nu(dv)\right)\right] \nu(dw) \\ &\geq \int_{n^{-1}}^{1} \left[1 - \exp\left(-\int (1 - e^{-v})\nu(dv)\right)\right] \nu(dw) \\ &= \left(\int_{n^{-1}}^{1} \nu(\mathrm{d}w)\right) \left(1 - \exp\left(-\int (1 - e^{-v})\nu(dv)\right)\right) \\ &\sim cn^{\alpha}\ell(n),\end{aligned}$$

where the last line follows from the assumption that the tail of ν is regularly varying. The second piece of the integral on $[0, n^{-1}]$ is bounded below by 0, and in combination, we have that $n^{\alpha}\ell(n) = O\left(\mathbb{E}\left(|V_n'|\right)\right)$. Now combining this with the previous upper bound result, we have $\mathbb{E}\left(|V_n'|\right) = \Theta(n^{\alpha}\ell(n))$.

The expected number of Poissonized multigraph edges satisfies $\mathbb{E}\left(E_n'\right) = \Theta(n)$, since

$$\mathbb{E}(|E_n'|) = \frac{n}{2} \iint wv\nu(dw)\nu(dv) = \frac{n}{2} \int w\nu(dw) \int v\nu(dv) = \frac{c^2}{2}n.$$

For the Poissonized binary graph edges, we split the integral into two pieces. We first upper bound the integral on the interval $[0, n^{-1/2}]$ and apply Definition C.4.4 to get the following asymptotic behavior:

$$\begin{aligned}
\frac{1}{2}\int_0^{n^{-1/2}}\int (1-\exp(-nwv))\,\nu(\mathrm{d}w)\,\nu(\mathrm{d}v) &\leq \frac{1}{2}\int_0^{n^{-1/2}}\int nwv\,\nu(\mathrm{d}w)\,\nu(\mathrm{d}v) \\
&= \frac{n}{2}\left(\int w\nu(\mathrm{d}w)\right)\int_0^{n^{-1/2}} v\nu(\mathrm{d}v) \\
&\sim cn(n^{-1/2})^{1-\alpha}\ell(n^{1/2}) \\
&= cn^{\frac{1+\alpha}{2}}\ell(n^{1/2}).
\end{aligned}$$

We then bound the second portion on the interval $[n^{-1/2}, 1]$ by linearizing at $v = n^{-1/2}$. Using integration by parts and an Abelian theorem (Feller, 1971, Sec. XIII.5, Thm. 4) for the Laplace transform, for some constants $b, d > 0$, we have

$$\begin{aligned}
&\frac{1}{2}\int_{n^{-1/2}}^1\int (1-\exp(-nwv))\,\nu(\mathrm{d}w)\,\nu(\mathrm{d}v) \\
&\leq \frac{1}{2}\int_{n^{-1/2}}^1\int \left(1-\exp\left(-n^{1/2}w\right)+nw\exp\left(-n^{1/2}w\right)(v-n^{-1/2})\right)\nu(\mathrm{d}w)\,\nu(\mathrm{d}v) \\
&= \frac{1}{2}\left(\int_{n^{-1/2}}^1 \nu(\mathrm{d}v)\right)\int n^{1/2}\exp(-n^{1/2}w)\,\bar{\nu}(w)\,\mathrm{d}w \\
&\quad + \frac{1}{2}\int_{n^{-1/2}}^1 (nv-n^{1/2})\,\nu(\mathrm{d}v)\int w\exp(-n^{1/2}w)\,\nu(\mathrm{d}w) \\
&\sim bn^{\alpha}\ell^2(n^{1/2}) + \frac{1}{2}\int_0^1 v\,\nu(\mathrm{d}v)\,n^{1/2}\int n^{1/2}\left(\exp(-n^{1/2}w)-n^{1/2}w\exp\left(-n^{1/2}w\right)\right)\bar{\nu}(w)\,\mathrm{d}w \\
&\leq bn^{\alpha}\ell^2(n^{1/2}) + \frac{1}{2}\int_0^1 v\,\nu(\mathrm{d}v)\;n^{1/2}\int n^{1/2}\exp(-n^{1/2}w)\,\bar{\nu}(w)\,\mathrm{d}w \\
&\sim bn^{\alpha}\ell^2(n^{1/2}) + dn^{1/2}n^{\alpha/2}\ell(n^{1/2}) \\
&= O(n^{\frac{1+\alpha}{2}}\ell(n^{1/2})).
\end{aligned}$$

Therefore we have that $\mathbb{E}\left(|\bar{E}'_n|\right) = O(n^{\frac{1+\alpha}{2}}\ell(n^{1/2}))$.

To get the other bound, we split the integral into three pieces. First,

$$\begin{aligned}
&\frac{1}{2}\int_0^{n^{-1}}\int (1-\exp(-nwv))\,\nu(\mathrm{d}w)\,\nu(\mathrm{d}v)\\
&\leq \frac{1}{2}\int_0^{n^{-1}}\int nwv\,\nu(\mathrm{d}w)\,\nu(\mathrm{d}v)\\
&= \frac{n}{2}\left(\int w\,\nu(\mathrm{d}w)\right)\int_0^{n^{-1}} v\,\nu(\mathrm{d}v)\\
&\sim cn(n^{-1})^{1-\alpha}\ell(n) = cn^{\alpha}\ell(n).
\end{aligned}$$

Next, integration by parts yields

$$\begin{aligned}
&\frac{1}{2}\int_{n^{-1/2}}^{1}\int (1-\exp(-nwv))\,\nu(\mathrm{d}w)\,\nu(\mathrm{d}v)\\
&\leq \frac{1}{2}\int_{n^{-1/2}}^{1}\int (1-\exp(-nw))\,\nu(\mathrm{d}w)\,\nu(\mathrm{d}v)\\
&= \frac{1}{2}\left(\int_{n^{-1/2}}^{1}\nu(\mathrm{d}v)\right)\int n\exp(-nw)\,\bar{\nu}(w)\,\mathrm{d}w\\
&\sim c\left(n^{-1/2}\right)^{-\alpha}\ell(n^{1/2})n^{\alpha}\ell(n)\\
&= cn^{\frac{3\alpha}{2}}\ell(n)\ell(n^{1/2}).
\end{aligned}$$

Finally, integration by parts yields the final upper bound

$$\begin{aligned}
&\frac{1}{2}\int_{n^{-1}}^{n^{-1/2}}\int (1-\exp(-nwv))\,\nu(\mathrm{d}w)\,\nu(\mathrm{d}v)\\
&\leq \frac{1}{2}\int_{n^{-1}}^{n^{-1/2}}\int (1-\exp(-n^{1/2}w))\,\nu(\mathrm{d}w)\,\nu(\mathrm{d}v)\\
&= \frac{1}{2}\left(\int_{n^{-1}}^{n^{-1/2}}\nu(\mathrm{d}v)\right)\int n^{1/2}\exp\left(-n^{1/2}w\right)\,\bar{\nu}(w)\,\mathrm{d}w\\
&\sim \left(c_1n^{\alpha}\ell(n)-c_2n^{\frac{\alpha}{2}}\ell(n^{1/2})\right)\left(c_3n^{\alpha/2}\ell(n^{1/2})\right)\\
&\sim cn^{\frac{3\alpha}{2}}\ell(n)\ell(n^{1/2}).
\end{aligned}$$

Therefore $\mathbb{E}\left(|\bar{E}'_n|\right) = O(\ell(n)\ell(n^{1/2})\,n^{\frac{3\alpha}{2}})$. □

Finally, we show that $|E_n|$, $|\bar{E}_n|$, and $|V_n|$ are asymptotically equivalent to their expectations almost surely; thus, the asymptotic results for the expectation sequences applies to the random sequences.

Lemma C.4.6 (4.5.1, main text)**.** *The number of edges and vertices for both the multi- and binary graphs satisfy*

$$|E_n| \overset{a.s.}{\sim} \mathbb{E}\left(|E_n|\right), \qquad |\bar{E}_n| \overset{a.s.}{\sim} \mathbb{E}\left(|\bar{E}_n|\right) \qquad |\bar{V}_n| = |V_n| \overset{a.s.}{\sim} \mathbb{E}\left(|V_n|\right), \qquad n \to \infty.$$

Proof. We use X_n to refer to $|E_n|$, $|\bar{E}_n|$, or $|V_n|$, since the proof technique is the same for all. Since we need to show $X_n/\mathbb{E}\left(X_n\right) \overset{\text{a.s.}}{\to} 1$, by the Borel-Cantelli lemma it is sufficient to show that for any $\epsilon > 0$,

$$\sum_n P(\left|X_n - \mathbb{E}\left(X_n\right)\right| > \epsilon \mathbb{E}\left(X_n\right)) < \infty.$$

By the union bound, and the fact that X_n can be expressed as a countable sum of indicators combined with the note after Theorem 4 in Freedman (1973),

$$\begin{aligned}
&P(\left|X_n - \mathbb{E}\left(X_n\right)\right| > \epsilon \mathbb{E}\left(X_n\right)) \\
&\leq P(X_n > (1+\epsilon)\mathbb{E}\left(X_n\right)) + P(X_n < (1-\epsilon)\mathbb{E}\left(X_n\right)) \\
&\leq 2\exp\left(-\frac{\epsilon^2 \mathbb{E}\left(X_n\right)}{2}\right).
\end{aligned}$$

Since $\mathbb{E}(X_n) \geq n^\beta$ for some $\beta > 0$, the expression is summable and the result holds. □

Combining the results of Lemmas 4.5.1, C.4.3, and C.4.5 gives us the main theorem, which we state here for completeness.

Theorem C.4.7 (4.5.3, main text)**.** *If the tail of ν is regularly varying as per Equation* (C.3)*, then*

as $n \to \infty$,

$$|V_n| \overset{a.s.}{=} \Theta(n^\alpha \ell(n)), \qquad |E_n| \overset{a.s.}{=} \Theta(n), \qquad |\bar{E}_n| \overset{a.s.}{=} O\left(\ell(n^{1/2}) \min\left(n^{\frac{1+\alpha}{2}}, \ell(n) n^{\frac{3\alpha}{2}}\right)\right).$$

Remark C.4.8. *Finally, to conclude that there exists a class of sparse, edge-exchangeable graphs, we examine the asymptotics from this result in more detail. In the multigraph case, we see that the number of vertices increases at the same rate as* $n^\alpha \ell(n)$*, and the number of edges increases linearly in* n*. So* $|E_n|$ *grows at the same rate as* $|V_n|^{1/\alpha} \ell(n)^{-1/\alpha}$*. When* $\alpha \in (1/2, 1)$*, the exponent* $1/\alpha$ *lies in the range* $(1, 2)$*, and thus this parameter range for* α *results in sparse graph sequences. For binary graphs, the number of edges* $|\bar{E}_n|$ *grows at a rate that is bounded by* $\ell(\sqrt{n}) \min\left\{|V_n|^{\frac{1+\alpha}{2\alpha}} \ell(n)^{-\frac{1+\alpha}{2\alpha}}, |V_n|^{\frac{3}{2}} \ell(n)^{-\frac{1}{2}}\right\}$*. Since* $\min\left\{\frac{1+\alpha}{2\alpha}, \frac{3}{2}\right\} \leq 3/2 < 2$*, binary graphs are sparse for any* $\alpha \in (0, 1)$*. Note that* $\ell(n)$ *does not affect the growth rate throughout since it is a slowly-varying function; i.e., for all* $c > 0$*,* $\ell(cn) \sim \ell(n)$*. For the three-parameter beta process, which we examined in our simulations, the function* ℓ *is a constant function, as in Equation* (C.5)*.*

We have shown that edge exchangeability admits sparse graphs by proving the existence of sparse graph sequences in a wide subclass of graph frequency models: those frequency models with weights generated from Poisson point processes whose rate measures have power law tails. Notably, we have shown the existence of a range of sparse and dense behavior in this wide class of graph frequency models, as desired.

Appendix D

Supplementary material: multi-fidelity MCMC

D.1 Additional related work

Approximate Bayesian computation (ABC) is a related class of methods to pseudo-marginal MCMC for implicit likelihoods. We note that the asymptotic target for ABC is an approximation to the target density of interest. Prescott and Baker (2020) propose a multi-fidelity approach to ABC. A number of multilevel ABC approaches (Guha and Tan, 2017; Lester, 2018; Warne et al., 2021) have also been proposed in recent work.

In our work, the Russian roulette estimator is used to construct an unbiased, low-fidelity likelihood. The Russian roulette estimator has been used recently in a number of applications for optimization and inference (Beatson and Adams, 2019; Luo et al., 2020; Potapczynski et al., 2021). In particular, Potapczynski et al. (2021) apply similar techniques to get an unbiased estimate of the gradient of the marginal likelihood for Gaussian process regression; this can viewed as an optimization analog to our approach.

D.2 Multi-fidelity simulated annealing

To adapt simulated annealing Section 2.3.5 to a multi-fidelity method, we consider energy functions of fidelity k, denoted by $E_k(\theta)$, and the corresponding low-fidelity targets $\pi_k(\theta) \propto \exp(-E_k(\theta)/T)$. To sample from $\pi(\theta|K=k)$, we accept or reject a proposal θ' based on:

$$R = \exp\left(-\frac{E_k(\theta') - E_k(\theta)}{T}\right).$$

To sample from $\pi(K|\theta)$, we accept or reject a proposal K' based on:

$$R = \exp\left(-\frac{E_{K'}(\theta) - E_K(\theta)}{T}\right)\left(\frac{\mu(K')}{\mu(K)}\right)^{\frac{1}{T}}.$$

D.3 Experiments: additional experiments and method details

In this section, we provide additional details for the methods used in our experiments along with additional details of the setup of each experiment.

Methods compared We will use the abbreviations *SF* to denote a single-fidelity algorithm, e.g., SF M-H, *MF* to refer to the pseudo-marginal MF-MCMC method proposed in this work, and *TS* to refer to the two-stage M-H algorithm described in Section 2.3.4. The primary sampling algorithms used to update the state $\theta|K$ are Metropolis-Hastings (M-H), (line) slice sampling (SS), and elliptical slice sampling (ESS).

Target estimator $\hat{\pi}$ In our experiments, by default we consider the Russian roulette estimator with $\mu = \text{geometric}(\gamma_0)$, unless stated otherwise.

Sampling the fidelity $K|\theta$ To sample the fidelity from the conditional target $K|\theta$, we consider the following random walk M-H move. Here the target is

$$\pi(K|\theta) \propto \mu(K)\hat{\pi}_K(\theta). \tag{D.1}$$

To propose a new fidelity, we consider a random walk on the positive integers: flip a fair coin to determine a new candidate location $k^* = k \pm 1$, where k is the current value. Then we can compute the following ratio and decide to accept/reject this candidate value:

$$R = \min\left(1, \frac{\mu(k^*)\hat{\pi}_{k^*}(\mathcal{D})}{\mu(k)\hat{\pi}_k(\mathcal{D})}\right).$$

In problems where the estimator may return negative values, we compute the absolute value of the estimator $|\hat{\pi}|$, as summarized in Algorithm 5.3.1.

D.3.1 Toy conjugate sequence

In this example, we consider a toy conjugate Bayesian model, where the data are assumed to arise i.i.d. from a perfect-fidelity model $L_\infty(\theta) = \mathcal{N}(x; \theta, \sigma_\infty)$, and a conjugate prior on θ, $\mathcal{N}(\theta|0, 1)$; conjugacy leads to a closed form Gaussian posterior density that we can compute and compare to the posterior samples obtained from the methods that we compare. Thus, the perfect-fidelity target is $\pi_\infty(\theta) \propto \mathcal{N}(\theta|0, 1)\prod_{n=1}^{N}\mathcal{N}(X_n; \theta, \sigma_\infty)$.

Now suppose that we only have access to the sequence of low-fidelity models $L_k(\theta) = \mathcal{N}(x; \theta, \sigma_k)$, where $\sigma_k^2 \to \sigma_\infty^2$. Here we consider the sequence $\sigma_k^2 = 1 + 2/k^2$ and $\sigma_\infty^2 = 1$. In this example, we consider the performance of (1) SF M-H, MF M-H, and two-stage M-H, and (2) SF slice sampling and slice sampling (there is not an analogous two-stage MCMC algorithm for slice sampling). We generate $N = 200$ observations $\mathcal{D}|\theta_0$ from the perfect-fidelity likelihood with true mean $\theta_0 \sim \mathcal{N}(0, 1)$.

To compute the "cost" of a likelihood evaluation, we pretend that the likelihood evaluation L_k has cost k. This is to demonstrate the cost of the method for problems where the cost of an evaluation increases linearly with k.

In what follows, we first compare the low-fidelity estimators, and then we compare the sampling methods on one choice of estimator.

Comparing SF-MCMC, MF-MCMC, and two-stage M-H We also compare to the two-stage M-H algorithm summarized in Section 2.3.4; here we consider 2 two-stage setups of $k =$

$\{1000, 10\}$ and $k = \{100, 5\}$. For all methods, we ran 4 chains initialized from the prior with $T = 10000$ iterations. We discarded 2000 burn-in samples and the subsequently collected every other sample.

D.3.2 Log Gaussian Cox Process

In this section, we provide details for the LGCP experiment on the coal mining disasters data set.

We approximate the integral in Equation (5.10) with a trapezoidal quadrature rule I_k: i.e., given k points $\tilde{x}_1, \ldots, \tilde{x}_k \in \mathbb{X}$ and observed points $\{X_1, \ldots, X_N\}$, the low-fidelity likelihood is:

$$L_k(f) = \exp\left(I_k(f(\tilde{x}_1), \ldots, f(\tilde{x}_K))\right) \prod_{n=1}^{N} e^{f(X_n)}, \tag{D.2}$$

where I_k is a trapezoid quadrature rule with $2k + c$ quadrature points and c is a constant offset parameter. When computing L_k for a grid of values different than the vector of latent function values currently available, we draw new function values conditioned on the existing values of f.

For all samplers, we used a squared-exponential kernel with lengthscale $\ell = 20$ and variance of 1. For the low-fidelity estimator $\hat{L}_k$, we used a Russian roulette estimator and set the offset $c = 10$. The truncation parameter of the MF model was fixed at $\gamma_0 = 0.08$. The results in the rightmost figure in Figure 5.5.2 are computed with respect to an average over 3 chains initialized from the prior with $T = 10000$ samples. The estimates with MF-ESS in Figure 5.5.2 were adjusted for negative signs; empirically, we observed roughly 2.5% of negative signs in our experiments.

D.3.3 Bayesian ODE system identification

Given a set of parameters θ and initial conditions, we can solve the ODE at a fidelity k to obtain the solution $z_n^{(k)}$. Thus, the likelihood of fidelity k is given by:

$$L_k(\theta) = \prod_{n=1}^{N} \prod_{j=1}^{2} \text{LogNormal}(\log(z_{n,j}^{(k)}(\theta)), \sigma), \tag{D.3}$$

where k represents the fidelity of the ODE solver for obtaining the solution $z_n(\theta)$. We use the following priors on the parameters

$$(\log \alpha, \log \beta, \log \gamma, \log \delta) \sim \mathcal{N}(\theta_0, \sigma_0 I), \qquad \theta_0 = [0, -2, 0, -3]^\top, \quad \sigma_0 = 0.1. \tag{D.4}$$

In order to apply elliptical slice sampling, which requires the prior to have mean 0, we apply a change of variables: define $L_k(\bar{\theta}) = L_k(\theta + \theta_0)$, and then transform the sampled values $\theta^{(t)} = \bar{\theta}^{(t)} + \theta_0$. In our experiments, we first verified the sampler was recovering values on synthetic data generated with initial conditions $z_0 = [1.0, 1.0]$, system parameters $\alpha = 1.5, \beta = 1.0, \gamma = 3.0, \delta = 1.0$, and noise parameter $\sigma = 0.8$ at a grid of N solution values.

We then applied the method to the Hudson's Bay Lynx-Hare data set, which documents the canadian lynx and showshoe hare populations between 1900 and 1920, based on the data collefted by the Hudson's Bay company. We compared two single-fidelity models with ODE step size $dt = 1 \times 10^{-5}, 1 \times 10^{-4}$. For the multi-fidelity ESS sampler, we visualize the results of $\gamma_0 = 0.12$, and the step size for the low-fidelity target sequence was computed as $dt(k) = 1/(sk + c)$, where we set $s = 10$ and $c = 50$.

The results using Euler's method to solve the ODE are in Figure 5.5.3, and the results of the 4th-order Runge Kutta solver are in Figure D.3.1. The maximum number of iterations of each ODE solver was set to 1×10^8 iterations.

In the top row of each figure, the black vertical dotted line denotes maximum likelihood estimates reported by Howard (2009).[1] In the bottom row of each figure, we report the posterior mean estimates of the system parameters averaged over 4 chains initialized from the prior. The wallclock time in seconds of each iteration was measured and the average per iteration was reported. Here the first 5000 samples of each chain were discarded and then every third sample was collected. Overall, we observe that the single-fidelity models can both be quite expensive; while they are able to recover the posterior mean well, they require quite a bit more computation than the multi-fidelity approach.

[1] Our model is a modification of the one proposed in a Stan case study, which compares their Bayesian estimates to the reported maximum likelihood results. See `https://mc-stan.org/users/documentation/case-studies/lotka-volterra-predator-prey.html` for further discussion.

Empirically, we observed roughly 1% of negative signs in our experiments.

D.3.4 PDE-constrained optimization

In the problem setting, the spatial domain is $[0, L]$ and the time domain is $[0, T]$. For our experiments, we chose $L = 10$ and $T = 1$.

To solve the PDE, we discretize the spatial domain into a grid of size Δx: thus, we can consider points $x_1, \ldots, x_I$ and $u_1(t), \ldots, u_I(t)$, where $u_i(t) = u(x_i, t)$. Then, we represent the second derivative using the central difference formula for the second degree derivative:

$$\frac{\partial^2 u(x,t)}{\partial^2 x} \approx \left[\frac{u_{i+1}(t) - 2u_i(t) + u_{i-1}(t)}{\Delta x^2}\right]_{i=1}^{I}.$$

Thus, we now consider the system of equations (with the appropriate boundary conditions imposed):

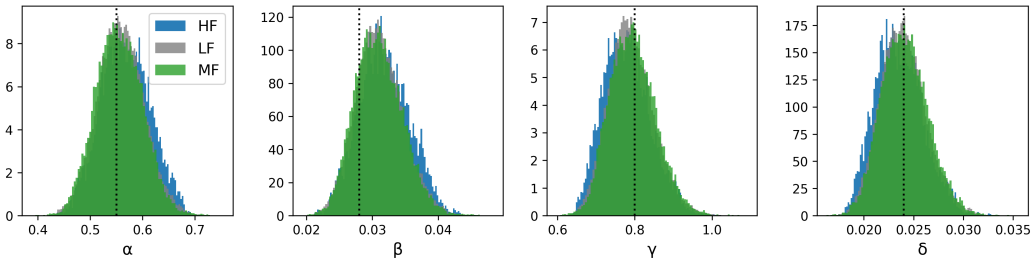
$$\frac{u_{i+1}(t) - 2u_i(t) + u_{i-1}(t)}{\Delta x^2} = \frac{du_i(t)}{dt}.$$

We solve the system with the Tsitouras 5/4 Runge-Kutta method, setting $\Delta t = 0.4\Delta x^2$ so as to satisfy a CFL stability condition. Here the fidelity of the problem is given by the size of the spatial discretization Δx, which in turn controls the discretization of Δt.

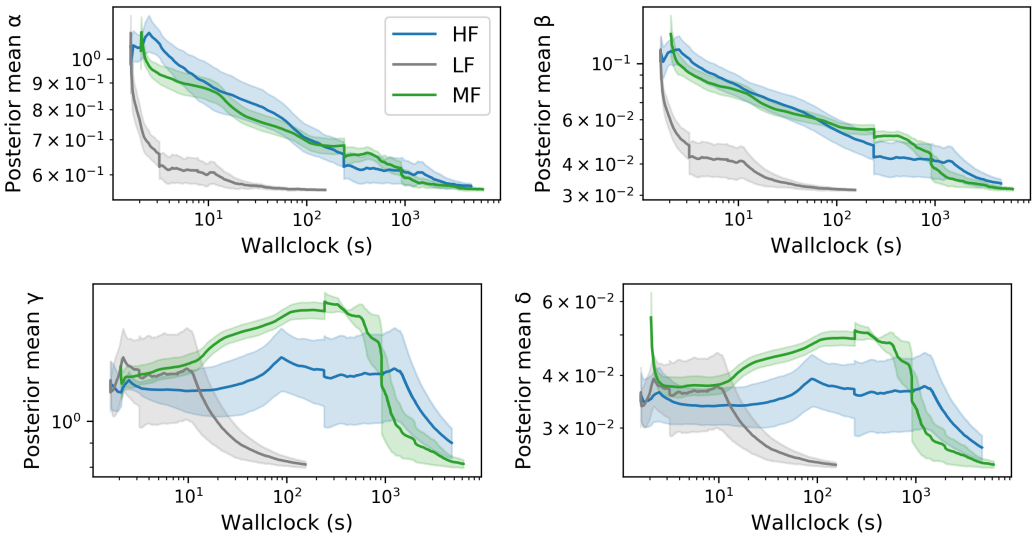
The target temperature $\bar{u}$ was constructed by solving the PDE with parameters $\alpha_0 = 0.85$ and $\beta_0 = 0.21$. For the simulated annealing algorithm, we use a Metropolis-Hastings algorithm as the base sampler; all methods used a truncated Normal proposal with scale set to 0.3 and a logarithmic temperature schedule.

In the top row of Figure 5.5.4, we visualization the target $\bar{u}$ solutions recovered by a number of methods. The low-fidelity solution in target (c) is given by a crude step size of $\Delta x = 2$; note that we do not evaluate the cost of this given how poorly the solution is recovered at this state.

In the bottom row of Figure 5.5.4, we compare the MF-ESS approach with two single-fidelity step sizes, $\Delta x = 5 \times 10^{-3}, 1 \times 10^{-2}$. In the multi-fidelity method, the low-fidelity target sequence was chosen using the discretization sequence $\Delta x(k) = 1/(k + c)$, where $c = 8$. The results are



(a) Marginal densities of system parameters



(b) Posterior mean estimate vs computational cost

Figure D.3.1: Lotka-Volterra system parameter identification with a 4th-order Runge Kutta ODE solver. The fidelity represents (a function of) the step size of the ODE solver. *Top:* Marginal distributions of system parameters. *Bottom:* Posterior mean estimates of the parameters vs wallclock.

averaged over random seeds using the initialization $[0, 0]$. The horizontal dotted lines in each plot denote the values of α_0, β_0, and we plot the current minimum at each iteration.

D.3.5 Gaussian process regression parameter inference

In many applications of GPs, the goal is to integrate out the parameters θ via a Monte Carlo approximation that uses MCMC to sample $\{\theta^{(t)}\}$ from the target density

$$\pi_\infty(\theta \,|\, \mathcal{D} = (X, y)) \propto \pi(\theta) L_\infty(\theta) = \text{logNormal}(\theta \,|\, \nu_0, \nu_1) \times \mathcal{N}(y \,|\, 0, \Sigma_\theta + \sigma_0^2 I). \tag{D.5}$$

Note that the Gaussian pdf has the form

$$L_\infty(\theta) = |2\pi(\Sigma_\theta + \sigma_0^2 I)|^{-\frac{1}{2}} \exp\left(-\frac{1}{2} y^\top (\Sigma_\theta + \sigma_0^2 I)^{-1} y\right), \tag{D.6}$$

and so when N is large, the linear system and determinant above become expensive.

Let the low-fidelity likelihood $L_k(\theta)$ denote the computation of the likelihood with k iterations of (preconditioned) conjugate gradient. That is, suppose, $z^{(k)}$ is the k^{th} iteration of the CG with respect to the linear system $(\Sigma_\theta + \sigma_0^2 I)z = y$. Thus, the low-fidelity likelihood is

$$L_k(\theta) = |2\pi(\Sigma_\theta + \sigma_0^2 I)|^{-\frac{1}{2}} \exp\left(-\frac{1}{2} y^\top z^{(k)}\right).$$

In practice, the determinant also needs to be approximated with another low-fidelity computation. Our goal here is to show a proof of concept, and so we only consider the linear system above; however, we note that the determinant can be iteratively computed as a byproduct of conjugate gradient as in Potapczynski et al. (2021). Note that we can compute the likelihood recursively in that each $z^{(k)}$ reuses computation from the previous step $z^{(k-1)}$, and thus a Russian roulette estimator also can reuse computation for each term in the sum.

We generate synthetic data from the GP model with $N = 100$, $\sigma_0^2 = 1$, and lengthscale $\theta_0 = 45$. For the GP model, we use the Log Normal prior on θ given above in Equation (D.5) with parameters $\nu_0 = 3.8, \nu_1 = 0.03$. We compare several likelihoods: a high-fidelity likelihood ($K = 100$), low-fidelity

likelihood ($K = 5$), and the multi-fidelity approach we describe with $\gamma_0 = 0.1$. The low-fidelity likelihood sequence was constructed by computing the solution to the linear system using a conjugate gradient solver with k steps. Finally, we also compare to a two-stage M-H approach with $k \in \{100, 5\}$. For all methods, we use a M-H sampler with $T = 50000$ iterations. The results are in Figure 5.5.5.

Bibliography

E. M. Airoldi, D. M. Blei, S. E. Fienberg, and E. P. Xing. Mixed membership stochastic blockmodels. *Journal of Machine Learning Research*, 9:1981–2014, 2008.

D. J. Aldous. Representations for partially exchangeable arrays of random variables. *Journal of Multivariate Analysis*, 11(4):581–598, 1981.

D. J. Aldous. Exchangeability and related topics. In *École d'été de probabilités de Saint-Flour, XIII—1983*, volume 1117 of *Lecture Notes in Math.*, pages 1–198. Springer, Berlin, 1985.

N. M. Alexandrov, J. E. Dennis, R. M. Lewis, and V. Torczon. A trust-region framework for managing the use of approximation models in optimization. *Structural Optimization*, 15(1):16–23, 1998.

C. Andrieu and G. O. Roberts. The pseudo-marginal approach for efficient Monte Carlo computations. *The Annals of Statistics*, 37(2):697–725, 2009.

C. Andrieu, A. Doucet, and R. Holenstein. Discussion of "Particle Markov chain Monte Carlo methods". *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(3): 269–342, 2010.

B. Aragam, C. Dan, E. P. Xing, and P. Ravikumar. Identifiability of nonparametric mixture models and Bayes optimal clustering. *Annals of Statistics*, 48(4):2277–2302, 2020.

E. Arian, M. Fahl, and E. W. Sachs. Trust-region proper orthogonal decomposition for flow control. In *IEEE Conference on Decision and Control*, 2000.

S. A. Armstrong, J. E. Staunton, L. B. Silverman, R. Pieters, M. L. d. Boer, M. D. Minden, S. E. Sallan, E. S. Lander, T. R. Golub, and S. J. Korsmeyer. MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. *Nature Genetics*, 30(1):41, 2001.

A. Baddeley, I. Bárány, and R. Schneider. Spatial point processes and their applications. *Stochastic Geometry: Lectures given at the CIME Summer School held in Martina Franca, Italy, September 13–18, 2004*, pages 1–75, 2007.

J. D. Banfield and A. E. Raftery. Model-based Gaussian and non-Gaussian clustering. *Biometrics*, 49(3):803–821, 1993.

A. R. Barron. *The exponential convergence of posterior probabilities with implications for Bayes estimators of density functions.* Department of Statistics, University of Illinois Champaign, IL, 1988.

A. R. Barron. Uniformly powerful goodness of fit tests. *The Annals of Statistics*, pages 107–124, 1989.

F. Bartolucci. Clustering univariate observations via mixtures of unimodal normal mixtures. *Journal of Classification*, 22(2):203–219, 2005.

A. Beatson and R. P. Adams. Efficient optimization of loops and limits with randomized telescoping sums. In *International Conference on Machine Learning*, pages 534–543, 2019.

M. A. Beaumont. Estimation of population growth or decline in genetically monitored populations. *Genetics*, 164(3):1139–1160, 2003.

A. Bhattacharjee, W. G. Richards, J. Staunton, C. Li, S. Monti, P. Vasa, C. Ladd, J. Beheshti, R. Bueno, M. Gillette, M. Loda, G. Weber, E. Mark, E. Lander, W. Wong, B. Johnson, T. Golub, D. Sugarbaker, and M. Meyerson. Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proceedings of the National Academy of Sciences*, 98(24):13790–13795, 2001.

P. Billingsley. *Probability and Measure.* John Wiley and Sons, third edition, 1986.

M. Biron-Lattes, A. Bouchard-Côté, and T. Campbell. Pseudo-marginal inference for CTMCs on infinite spaces via monotonic likelihood approximations. *Journal of Computational and Graphical Statistics*, 2022.

P. G. Bissiri, C. C. Holmes, and S. G. Walker. A general framework for updating belief distributions. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, 78(5):1103, 2016.

B. Bollobás, S. Janson, and O. Riordan. The phase transition in inhomogeneous random graphs. *Random Structures Algorithms*, 31(1):3–122, 2007.

C. Borgs, J. T. Chayes, H. Cohn, and N. Holden. Sparse exchangeable graphs and their limits via graphon processes. *Journal of Machine Learning Research*, 18:1–71, 2018.

C. Borgs, J. Chayes, H. Cohn, and Y. Zhao. An L^p theory of sparse graph convergence I: Limits, sparse random graph models, and power law distributions. *Transactions of the American Mathematical Society*, 372(5):3019–3062, 2019.

C. Borgs, J. T. Chayes, H. Cohn, and S. Ganguly. Consistent nonparametric estimation for heavy-tailed sparse graphs. *The Annals of Statistics*, 49(4):1904–1930, 2021.

L. Brevault, M. Balesdent, and A. Hebbal. Overview of Gaussian process based multi-fidelity techniques with variable relationship between fidelities. *arXiv e-print 2006.16728*, 2020.

T. Broderick, M. I. Jordan, and J. Pitman. Cluster and feature modeling from combinatorial stochastic processes. *Statistical Science*, 2013a.

T. Broderick, J. Pitman, and M. I. Jordan. Feature allocations, probability functions, and paintboxes. *Bayesian Analysis*, 8(4):801–836, 2013b.

T. Broderick and D. Cai. Edge-exchangeable graphs, sparsity, and power laws. In *NIPS 2015 Workshop on Bayesian Nonparametrics: The Next Generation*, 2015a.

T. Broderick and D. Cai. Edge-exchangeable graphs and sparsity. In *NIPS 2015 Workshop on Networks in the Social and Informational Sciences*, 2015b.

T. Broderick, M. I. Jordan, and J. Pitman. Beta processes, stick-breaking and power laws. *Bayesian Analysis*, 7(2):439–475, 2012.

T. Broderick, A. C. Wilson, and M. I. Jordan. Posteriors, conjugacy, and exponential families for completely random measures. *Bernoulli*, 2015.

D. Cai and R. P. Adams. Multi-fidelity Monte Carlo: a pseudo-marginal approach. In *Advances in Neural Information Processing Systems*, 2022.

D. Cai and T. Broderick. Completely random measures for modeling power laws in sparse graphs. In *NIPS 2015 Workshop on Networks in the Social and Informational Sciences*, 2015.

D. Cai, C. Freer, and N. Ackerman. An iterative step-function estimator for graphons. *arXiv e-print 1412.2129*, 2014.

D. Cai, T. Campbell, and T. Broderick. Edge-exchangeable graphs and sparsity. In *Advances in Neural Information Processing Systems*, 2016a.

D. Cai, C. Freer, and N. Ackerman. Priors on exchangeable directed graphs. *Electronic Journal of Statistics*, 10(2):3490–3515, 2016b.

D. Cai, M. Mitzenmacher, and R. P. Adams. A Bayesian nonparametric view on count-min sketch. In *Advances in Neural Information Processing Systems*, 2018.

D. Cai, T. Campbell, and T. Broderick. Power posteriors do not reliably learn the number of components in a finite mixture. In *NeurIPS Workshop: I Can't Believe It's Not Better*, 2020.

D. Cai, T. Campbell, and T. Broderick. Finite mixture models do not reliably learn the number of components. In *International Conference of Machine Learning*, 2021.

T. Campbell, J. Huggins, J. How, and T. Broderick. Truncated random measures. *arXiv e-print 1603.00861*, 2016.

T. Campbell, D. Cai, and T. Broderick. Exchangeable trait allocations. *Electronic Journal of Statistics*, 12(2):2290–2322, 2018.

B. P. Carlin, A. E. Gelfand, and A. F. Smith. Hierarchical Bayesian analysis of changepoint problems. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 41(2):389–405, 1992.

F. Caron. Bayesian nonparametric models for bipartite graphs. In *Advances in Neural Information Processing Systems*, 2012.

F. Caron and E. B. Fox. Sparse graphs using exchangeable random measures. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 79(5):1295–1366, 2017.

A. Chambaz and J. Rousseau. Bounds for Bayesian order identification with application to mixtures. *The Annals of Statistics*, 36(2):938–962, 2008.

C. Chan, F. Feng, J. Ottinger, D. Foster, M. West, and T. B. Kepler. Statistical mixture modeling for cell subtype identification in flow cytometry. *Cytometry Part A*, 73(8):693–701, 2008.

J. Chen. Optimal rate of convergence for finite mixture models. *The Annals of Statistics*, 23(1): 221–233, 1995.

J. A. Christen and C. Fox. Markov chain Monte Carlo using an approximation. *Journal of Computational and Graphical Statistics*, 14(4):795–810, 2005.

H. Crane and W. Dempsey. A framework for statistical network modeling. *arXiv e-print 1509.08185*, 2015a.

H. Crane and W. Dempsey. Atypical scaling behavior persists in real world interaction networks. *arXiv e-print 1509.08184*, 2015b.

H. Crane and W. Dempsey. Edge exchangeable models for network data. *arXiv e-print 1603.04571*, 2016.

T. Cui, Y. M. Marzouk, and K. E. Willcox. Data-driven model reduction for the Bayesian solution of inverse problems. *International Journal for Numerical Methods in Engineering*, 102(5):966–990, 2015.

D. J. Daley and D. Vere-Jones. *An introduction to the theory of point processes. Vol. II.* Probability and its Applications (New York). Springer, New York, second edition, 2008.

M. C. de Souto, I. G. Costa, D. S. de Araujo, T. B. Ludermir, and A. Schliep. Clustering cancer gene expression data: a comparative study. *BMC Bioinformatics*, 9(1):497, 2008.

M. Di Zio, U. Guarnera, and R. Rocci. A mixture of mixture models for a classification problem: The unity measure error. *Computational Statistics & Data Analysis*, 51(5):2573–2585, 2007.

P. Diaconis and S. Janson. Graph limits and exchangeable random graphs. *Rendiconti di Matematica e delle sue Applicazioni. Serie VII*, 28(1):33–61, 2008.

P. J. Diggle, P. Moraga, B. Rowlingson, and B. M. Taylor. Spatial and spatio-temporal log-Gaussian Cox processes: extending the geostatistical paradigm. *Statistical Science*, 28(4):542–563, 2013.

T. J. Dodwell, C. Ketelsen, R. Scheichl, and A. L. Teckentrup. A hierarchical multilevel Markov chain Monte Carlo algorithm with applications to uncertainty quantification in subsurface flow. *SIAM/ASA Journal on Uncertainty Quantification*, 3(1):1075–1108, 2015.

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.

Y. Efendiev, T. Hou, and W. Luo. Preconditioning Markov chain Monte Carlo simulations using coarse-scale models. *SIAM Journal on Scientific Computing*, 28(2):776–803, 2006.

P. Erdős and A. Rényi. On random graphs. I. *Publicationes Mathematicae Debrecen*, 6:290–297, 1959.

M. Fahl and E. W. Sachs. Reduced order modelling approaches to PDE-constrained optimization based on proper orthogonal decomposition. In *Large-scale PDE-constrained Optimization*, pages 268–280. Springer, 2003.

W. Feller. *An introduction to probability theory and its applications. Vol. II.* John Wiley & Sons, Inc., New York-London-Sydney, second edition, 1971.

D. Freedman. Another note on the Borel-Cantelli lemma and the strong law, with the Poisson approximation as a by-product. *The Annals of Probability*, 1(6):910–925, 1973.

S. Frühwirth-Schnatter. *Finite mixture and Markov switching models*. Springer Series in Statistics, 2006.

S. Frühwirth-Schnatter and G. Malsiner-Walli. From here to infinity: sparse finite versus Dirichlet process mixtures in model-based clustering. *Advances in Data Analysis and Classification*, 13(1): 33–64, 2019.

J. Geng, A. Bhattacharya, and D. Pati. Probabilistic community detection with unknown number of communities. *Journal of the American Statistical Association*, 114(526):893–905, 2019.

A. Georgoulas, J. Hillston, and G. Sanguinetti. Unbiased Bayesian inference for population Markov jump processes via random truncations. *Statistics and Computing*, 27(4):991–1002, 2017.

A. Gessner, J. Gonzalez, and M. Mahsereci. Active multi-information source Bayesian quadrature. In *Uncertainty in Artificial Intelligence*, pages 712–721, 2020.

S. Ghosal and A. van der Vaart. *Fundamentals of Nonparametric Bayesian Inference*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2017.

S. Ghosal, J. Ghosh, and R. Ramamoorthi. Posterior consistency of Dirichlet mixtures in density estimation. *The Annals of Statistics*, 27(1):143–158, 1999.

J. Ghosh and R. Ramamoorthi. *Bayesian Nonparametrics*. Springer Series in Statistics, 2003.

S. Ghosh and E. B. Sudderth. Nonparametric learning for layered segmentation of natural images. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2272–2279. IEEE, 2012.

M. B. Giles. Multilevel Monte Carlo path simulation. *Operations Research*, 56(3):607–617, 2008.

M. B. Giles. Multilevel Monte Carlo methods. *Monte Carlo and Quasi-Monte Carlo Methods*, pages 83–103, 2013.

P. W. Glynn and C.-h. Rhee. Exact estimation for Markov chain equilibrium expectations. *Journal of Applied Probability*, 51(A):377–389, 2014.

A. Gnedin, B. Hansen, and J. Pitman. Notes on the occupancy problem with infinitely many boxes: general asymptotics and power laws. *Probability Surveys*, 4:146–171, 2007.

S. Goldwater, M. Johnson, and T. L. Griffiths. Interpolating between types and tokens by estimating power-law generators. In *Advances in Neural Information Processing Systems*, pages 459–466, 2005.

R. B. Gramacy and H. K. Lee. Adaptive design and analysis of supercomputer experiments. *Technometrics*, 51(2):130–145, 2009.

J. Griffié, M. Shannon, C. L. Bromley, L. Boelen, G. L. Burn, D. J. Williamson, N. A. Heard, A. P. Cope, D. M. Owen, and P. Rubin-Delanchy. A Bayesian cluster analysis method for single-molecule localization microscopy data. *Nature Protocols*, 11(12):2499, 2016.

P. Grünwald and T. v. Ommen. Inconsistency of Bayesian inference for misspecified linear models, and a proposal for repairing it. *Bayesian Analysis*, 12(4):1069–1103, 2017.

P. D. Grünwald. Bayesian inconsistency under misspecification. In *World Meeting of the International Society for Bayesian Analysis*, 2006.

A. Guha, N. Ho, and X. Nguyen. On posterior contraction of parameters and interpretability in bayesian mixture modeling. *Bernoulli*, 27(4):2159–2188, 2021.

N. Guha and X. Tan. Multilevel approximate Bayesian approaches for flows in highly heterogeneous porous media and their applications. *Journal of Computational and Applied Mathematics*, 317: 700–717, 2017.

G. W. Gundersen, D. Cai, C. Zhou, B. E. Engelhardt, and R. P. Adams. Active multi-fidelity Bayesian online changepoint detection. In *Uncertainty in Artificial Intelligence*, pages 1916–1926, 2021.

W. Hackbusch. *Multi-grid methods and applications*, volume 4. Springer Science & Business Media, 2013.

P. Heinrich and J. Kahn. Strong identifiability and optimal minimax rates for finite mixture estimation. *The Annals of Statistics*, 46(6A):2844–2870, 2018.

T. Herlau, M. Schmidt, and M. Mørup. Completely random measures for modelling block-structured networks. In *Advances in Neural Information Processing Systems*, 2016.

D. Higdon, H. Lee, and Z. Bi. A Bayesian approach to characterizing uncertainty in inverse problems using coarse and fine-scale information. *IEEE Transactions on Signal Processing*, 50(2):389–399, 2002.

N. Ho and X. Nguyen. On strong identifiability and convergence rates of parameter estimation in finite mixtures. *Electronic Journal of Statistics*, 10(1):271–307, 2016.

P. Hoff. Modeling homophily and stochastic equivalence in symmetric relational data. In *Advances in Neural Information Processing Systems*, pages 657–664, 2007.

P. D. Hoff, A. E. Raftery, and M. S. Handcock. Latent space approaches to social network analysis. *Journal of the American Statistical Association*, 97(460):1090–1098, 2002.

P. W. Holland, K. B. Laskey, and S. Leinhardt. Stochastic blockmodels: first steps. *Social Networks*, 5(2):109–137, 1983.

C. Holmes and S. Walker. Assigning a value to a power likelihood in a general Bayesian model. *Biometrika*, 104(2):497–503, 2017.

D. N. Hoover. Relations on probability spaces and arrays of random variables. Preprint, Institute for Advanced Study, Princeton, NJ, 1979.

P. Howard. Modeling basics. *Lecture Notes for Math*, 442, 2009.

G. Hu, H.-C. Yang, and Y. Xue. Bayesian group learning for shot selection of professional basketball players. *Stat*, page e324, 2020.

J. P. Huelsenbeck and P. Andolfatto. Inference of population structure under a Dirichlet process model. *Genetics*, 175(4):1787–1802, 2007.

J. H. Huggins and J. W. Miller. Using bagged posteriors for robust inference and model criticism. *arXiv e-print 1912.07104*, 2019.

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):1316–1332, 2001.

P. E. Jacob, J. O'Leary, and Y. F. Atchadé. Unbiased Markov chain Monte Carlo methods with couplings. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(3): 543–600, 2020.

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):158–182, 2004.

J. Jewson, J. Q. Smith, and C. Holmes. Principles of Bayesian inference using general divergence criteria. *Entropy*, 20(6):442, 2018.

A. Jones, D. Cai, and B. Engelhardt. Multi-fidelity Bayesian experimental design using power posteriors. *In NeurIPS Workshop on Gaussian Processes, Spatiotemporal Modeling, and Decision-making Systems*, 2022.

A. Jones, D. Cai, D. Li, and B. Engelhardt. Optimizing the design of spatial genomics experiments. *bioRxiv doi: 10.1101/2023.01.29.526115*, 2023.

D. R. Jones, M. Schonlau, and W. J. Welch. Efficient global optimization of expensive black-box functions. *Journal of Global Optimization*, 13(4):455–492, 1998.

O. Kallenberg. Exchangeable random measures in the plane. *Journal of Theoretical Probability*, 3 (1):81–136, 1990.

O. Kallenberg. *Probabilistic symmetries and invariance principles.* Probability and its Applications. Springer, New York, 2005.

C. Kemp, J. B. Tenenbaum, T. L. Griffiths, T. Yamada, and N. Ueda. Learning systems of concepts with an infinite relational model. In *AAAI 21*, 2006.

J. F. C. Kingman. *Poisson processes*, volume 3 of *Oxford Studies in Probability*. The Clarendon Press, Oxford University Press, New York, 1993. Oxford Science Publications.

J. F. C. Kingman. The representation of partition structures. *Journal of the London Mathematical Society*, 2(2):374–380, 1978.

A. Klami and A. Jitta. Probabilistic size-constrained microclustering. In *Proceedings of the Conference of Uncertainty in Artificial Intelligence*, 2016.

B. Kleijn. *Bayesian asymptotics under misspecification*. PhD thesis, Vrije Universiteit Amsterdam, 2003.

J. Knoblauch, J. Jewson, and T. Damoulas. Generalized variational inference: Three arguments for deriving new posteriors. *arXiv e-print 1904.02063*, 2019.

W. Kruijer, J. Rousseau, and A. van der Vaart. Adaptive Bayesian density estimation with location-scale mixtures. *Electronic Journal of Statistics*, 4:1225–1257, 2010.

L. LeCam. Convergence of estimates under dimensionality restrictions. *The Annals of Statistics*, pages 38–53, 1973.

S. Legramanti, T. Rigon, D. Durante, and D. B. Dunson. Extended stochastic block models with application to criminal networks. *arXiv e-print 2007.08569*, 2020.

C. Lester. Multi-level approximate Bayesian computation. *arXiv e-print 1811.08866*, 2018.

S. Li, W. Xing, R. Kirby, and S. Zhe. Multi-fidelity Bayesian optimization via deep neural networks. *Advances in Neural Information Processing Systems*, 33:8521–8531, 2020.

A. Lijoi, I. Prünster, and S. Walker. Extending Doob's consistency theorem to nonparametric densities. *Bernoulli*, 10(4):651–663, 2004.

L. Lin, K. Liu, and J. Sloan. A noisy Monte Carlo algorithm. *Physical Review D*, 61(7):074505, 2000.

J. R. Lloyd, P. Orbanz, Z. Ghahramani, and D. M. Roy. Random function priors for exchangeable arrays with applications to graphs and relational data. In *Advances in Neural Information Processing Systems*, 2012.

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.

L. Lovász and B. Szegedy. Limits of dense graph sequences. *Journal of Combinatorial Theory. Series B*, 96(6):933–957, 2006.

Y. Luo, A. Beatson, M. Norouzi, J. Zhu, D. Duvenaud, R. P. Adams, and R. T. Chen. SUMO: Unbiased estimation of log marginal probability for latent variable models. *International Conference on Learning Representations*, 2020.

A.-M. Lyne, M. Girolami, Y. Atchadé, H. Strathmann, and D. Simpson. On Russian roulette estimates for Bayesian inference with doubly-intractable likelihoods. *Statistical Science*, 30(4): 443–467, 2015.

D. J. MacKay. *Information theory, inference and learning algorithms*. Cambridge University Press, 2003.

G. Malsiner-Walli, S. Frühwirth-Schnatter, and B. Grün. Model-based clustering based on sparse finite Gaussian mixtures. *Statistics and Computing*, 26(1-2):303–324, 2016.

G. Malsiner-Walli, S. Frühwirth-Schnatter, and B. Grün. Identifying mixtures of mixtures using Bayesian estimation. *Journal of Computational and Graphical Statistics*, 26(2):285–295, 2017.

A. March and K. Willcox. Constrained multifidelity optimization using model calibration. *Structural and Multidisciplinary Optimization*, 46(1):93–109, 2012.

G. J. McLachlan, R. Bean, and D. Peel. A mixture model-based approach to the clustering of microarray expression data. *Bioinformatics*, 18(3):413–422, 2002.

P. D. McNicholas and T. B. Murphy. Model-based clustering of microarray expression data via latent gaussian mixture models. *Bioinformatics*, 26(21):2705–2712, 2010.

M. Medvedovic and S. Sivaganesan. Bayesian infinite mixture model based clustering of gene expression profiles. *Bioinformatics*, 18(9):1194–1206, 2002.

M. Medvedovic, K. Y. Yeung, and R. E. Bumgarner. Bayesian mixture model based clustering of replicated microarray data. *Bioinformatics*, 20(8):1222–1232, 2004.

J. W. Miller and D. B. Dunson. Robust Bayesian inference via coarsening. *Journal of the American Statistical Association*, 114(527):1113–1125, 2019.

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*, pages 199–206, 2013.

J. W. Miller and M. T. Harrison. Inconsistency of Pitman-Yor process mixtures for the number of components. *Journal of Machine Learning Research*, 15(1):3333–3370, 2014.

J. W. Miller and M. T. Harrison. Mixture models with a prior on the number of components. *Journal of the American Statistical Association*, 113(521):340–356, 2018.

K. T. Miller, T. L. Griffiths, and M. I. Jordan. Nonparametric latent feature models for link prediction. In *Advances in Neural Information Processing Systems*, pages 1276–1284, 2009.

J. Møller, A. R. Syversveen, and R. P. Waagepetersen. Log Gaussian Cox processes. *Scandinavian Journal of Statistics*, 25(3):451–482, 1998.

S. Mukherjee, E. D. Feigelson, G. J. Babu, F. Murtagh, C. Fraley, and A. Raftery. Three types of gamma-ray bursts. *The Astrophysical Journal*, 508(1):314, 1998.

I. Murray and M. Graham. Pseudo-marginal slice sampling. In *Artificial Intelligence and Statistics*, pages 911–919, 2016.

I. Murray, R. Adams, and D. MacKay. Elliptical slice sampling. In *Artificial Intelligence and Statistics*, pages 541–548, 2010.

R. M. Neal. Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9(2):249–265, 2000.

R. M. Neal. Slice sampling. *The Annals of Statistics*, 31(3):705–767, 2003.

J. Nemec and A. F. L. Nemec. Mixture models for studying stellar populations. I. Univariate mixture models, parameter estimation, and the number of discrete population components. *Publications of the Astronomical Society of the Pacific*, 103(659):95, 1991.

X. 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, Carnegie Mellon University, 1994.

P. Orbanz and D. M. Roy. Bayesian models of graphs, arrays and other exchangeable random structures. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(2):437–461, 2015.

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.

A. Palizhati, S. B. Torrisi, M. Aykol, S. K. Suram, J. S. Hummelshøj, and J. H. Montoya. Agents for sequential learning using multiple-fidelity data. *Scientific Reports*, 12(1):1–13, 2022.

K. Palla, D. A. Knowles, and Z. Ghahramani. An infinite latent attribute model for network data. In *International Conference on Machine Learning*, 2012.

K. Palla, F. Caron, and Y. W. Teh. A Bayesian nonparametric model for sparse dynamic networks. *arXiv e-print 1607.01624*, 2016.

B. Peherstorfer, K. Willcox, and M. Gunzburger. Survey of multifidelity methods in uncertainty propagation, inference, and optimization. *SIAM Review*, 60(3):550–591, 2018.

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.

F. Petralia, V. Rao, and D. B. Dunson. Repulsive mixtures. In *Advances in Neural Information Processing Systems*, pages 1889–1897, 2012.

J. Pitman. Exchangeable and partially exchangeable random partitions. *Probability theory and related fields*, 102(2):145–158, 1995.

J. Pitman and M. Yor. The two-parameter poisson-dirichlet distribution derived from a stable subordinator. *The Annals of Probability*, pages 855–900, 1997.

A. Potapczynski, L. Wu, D. Biderman, G. Pleiss, and J. P. Cunningham. Bias-free scalable Gaussian processes via randomized truncations. In *International Conference on Machine Learning*, pages 8609–8619, 2021.

S. Prabhakaran, E. Azizi, A. Carr, and D. Pe'er. Dirichlet process mixture model for correcting technical variation in single-cell gene expression data. In *International Conference on Machine Learning*, pages 1070–1079, 2016.

T. P. Prescott and R. E. Baker. Multifidelity approximate Bayesian computation. *SIAM/ASA Journal on Uncertainty Quantification*, 8(1):114–138, 2020.

J. K. Pritchard, M. Stephens, and P. Donnelly. Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–959, 2000.

M. Raissi, P. Perdikaris, and G. E. Karniadakis. Inferring solutions of differential equations using noisy multi-fidelity data. *Journal of Computational Physics*, 335:736–746, 2017.

R. Ramamoorthi, K. Sriram, and R. Martin. On posterior concentration in misspecified models. *Bayesian Analysis*, 10(4):759–789, 2015.

C. Rasmussen, B. de la Cruz, Z. Ghahramani, and D. Wild. Modeling and visualizing uncertainty in gene expression clusters using Dirichlet process mixtures. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 6(4):615–628, 2008.

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

T. Rigon, A. H. Herring, and D. B. Dunson. A generalized Bayes framework for probabilistic clustering. *Biometrika*, 2023.

T. Robinson, M. S. Eldred, K. E. Willcox, and R. Haimes. Surrogate-based optimization using multifidelity models with variable parameterization and corrected space mapping. *AIAA journal*, 46(11):2814–2822, 2008.

A. Rodriguez and D. B. Dunson. Nonparametric Bayesian models through probit stick-breaking processes. *Bayesian Analysis*, 6(1):145–178, 2011.

K. Roeder and L. Wasserman. Practical Bayesian density estimation using mixtures of normals. *Journal of the American Statistical Association*, 92(439):894–902, 1997.

H. E. Romeijn and R. L. Smith. Simulated annealing for constrained global optimization. *Journal of Global Optimization*, 5(2):101–126, 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.

D. M. Roy and Y. W. Teh. The Mondrian process. In *Advances in Neural Information Processing Systems*, pages 1377–1384, 2008.

R. Royall and T.-S. Tsou. Interpreting statistical evidence by using imperfect models: robust adjusted likelihood functions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(2):391–404, 2003.

P. Rubin-Delanchy, G. L. Burn, J. Griffié, D. J. Williamson, N. A. Heard, A. P. Cope, and D. M. Owen. Bayesian cluster identification in single-molecule localization microscopy data. *Nature Methods*, 12(11):1072–1076, 2015.

W. Rudin. *Principles of Mathematical Analysis*. McGraw-Hill, 1976.

L. Schwartz. On Bayes procedures. *Zeitschrift für Wahrscheinlichkeitstheorie*, 4:10–26, 1965.

J. Song, Y. Chen, and Y. Yue. A general framework for multi-fidelity Bayesian optimization with gaussian processes. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 3158–3167, 2019.

Y. W. Teh and D. Görür. Indian buffet processes with power-law behavior. In *Advances in Neural Information Processing Systems*, 2009.

Y. W. Teh. A hierarchical bayesian language model based on pitman-yor processes. In *Proceedings of the 21st International Conference on Computational Linguistics*, pages 985–992. Association for Computational Linguistics, 2006.

H. Teicher. Identifiability of mixtures. *The Annals of Mathematical Statistics*, 32(1):244–248, 1961.

H. Teicher. Identifiability of finite mixtures. *The Annals of Mathematical Statistics*, pages 1265–1269, 1963.

F. K. Teklehaymanot, A.-K. Seifert, M. Muma, M. G. Amin, and A. M. Zoubir. Bayesian target enumeration and labeling using radar data of human gait. In *26th European Signal Processing Conference*, pages 1341–1346. IEEE, 2018.

M. Teng, F. Nathoo, and T. D. Johnson. Bayesian computation for Log-Gaussian Cox processes: A comparative analysis of methods. *Journal of Statistical Computation and Simulation*, 87(11): 2227–2252, 2017.

R. Thibaux and M. I. Jordan. Hierarchical beta processes and the Indian buffet process. In *Artificial Intelligence and Statistics*, pages 564–571, 2007.

A. Todeschini, X. Miscouridou, and F. Caron. Exchangeable random measures for sparse and modular graphs with overlapping communities. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(2):487–520, 2016.

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. Tonkin-Hill, J. A. Lees, S. D. Bentley, S. D. Frost, and J. Corander. Fast hierarchical Bayesian analysis of population structure. *Nucleic Acids Research*, 47(11):5539–5549, 2019.

M. Troyer and U.-J. Wiese. Computational complexity and fundamental limitations to fermionic quantum Monte Carlo simulations. *Physical Review Letters*, 94(17):170201, 2005.

V. Veitch and D. M. Roy. The class of random graphs arising from exchangeable random measures. *arXiv e-print 1512.03099*, 2015.

Y. Wang, A. Kucukelbir, and D. M. Blei. Reweighted data for robust probabilistic models. In *International Conference on Machine Learning*, page 3646–3655, 2017.

D. J. Warne, T. P. Prescott, R. E. Baker, and M. J. Simpson. Multifidelity multilevel Monte Carlo to accelerate approximate Bayesian parameter inference for partially observed stochastic processes. *arXiv e-print 2110.14082*, 2021.

S. Williamson. Nonparametric network models for link prediction. *Journal of Machine Learning Research*, 17:1–21, 2016.

P. J. Wolfe and S. C. Olhede. Nonparametric graphon estimation. *arXiv e-print 1309.5936*, 2013.

M.-J. Woo and T. Sriram. Robust estimation of mixture complexity. *Journal of the American Statistical Association*, 101(476):1475–1486, 2006.

M.-J. Woo and T. Sriram. Robust estimation of mixture complexity for count data. *Computational Statistics & Data Analysis*, 51(9):4379–4392, 2007.

J. Wu, S. Toscano-Palmerin, P. I. Frazier, and A. G. Wilson. Practical multi-fidelity Bayesian optimization for hyperparameter tuning. In *Uncertainty in Artificial Intelligence*, pages 788–798, 2020.

Y. Wu and S. Ghosal. Kullback Leibler property of kernel mixture priors in Bayesian density estimation. *Electronic Journal of Statistics*, 2:298–331, 2008.

X. Xi, F.-X. Briol, and M. Girolami. Bayesian quadrature for multiple related integrals. In *International Conference on Machine Learning*, pages 5373–5382, 2018.

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 *International Conference on Machine Learning*, pages 1049–1056, 2006.

Z. Xu, V. Tresp, S. Yu, K. Yu, and H. Kriegel. Fast inference in infinite hidden relational models. In *Proceedings of Mining and Learning with Graphs*, 2007.

S. J. Yakowitz and J. D. Spragins. On the identifiability of finite mixtures. *The Annals of Mathematical Statistics*, 39(1):209–214, 1968.

K. Y. Yeung, D. R. Haynor, and W. L. Ruzzo. Validating clustering for gene expression data. *Bioinformatics*, 17(4):309–318, 2001.

G. Zanella, B. Betancourt, H. Wallach, J. Miller, A. Zaidi, and R. C. Steorts. Flexible models for microclustering with application to entity resolution. In *Advances in Neural Information Processing Systems*, 2016.

A. Zeisel, A. B. Muñoz-Manchado, S. Codeluppi, P. Lönnerberg, G. La Manno, A. Juréus, S. Marques, H. Munguba, L. He, C. Betsholtz, C. Rolny, G. Castelo-Branco, J. Hjerling-Leffler, and S. Linnarsson. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science*, 347(6226):1138–1142, 2015.

Z. Zhang, G. Zhang, H. Goyal, L. Mo, and Y. Hong. Identification of subclasses of sepsis that

showed different clinical outcomes and responses to amount of fluid resuscitation: a latent profile analysis. *Critical Care*, 22(1):1–11, 2018.

D. Zoltowski, D. Cai, and R. P. Adams. Slice sampling reparameterization gradients. In *Advances in Neural Information Processing Systems*, 2021.