NEW METHODS AND LARGE SAMPLE THEORY IN BAYESIAN NONPARAMETRICS AND SEMIPARAMETRICS

By

CHENG ZENG

A DISSERTATION PRESENTED TO THE GRADUATE SCHOOL
OF THE UNIVERSITY OF FLORIDA IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

UNIVERSITY OF FLORIDA

2025

Dedicated to my mom, Yan Zeling and my dad, Zeng Zhaojuan.

ACKNOWLEDGEMENTS

First and foremost, I would like to express my deepest gratitude to my advisor, Dr. Leo L. Duan. His constant stream of research ideas, boundless enthusiasm, and keen attention to detail have been a source of inspiration throughout my doctoral journey. His rigorous standards for academic work pushed me to achieve more than I thought possible, and his support in life outside of research, including shared meals and picnics, made this journey enjoyable. His dedication to both his students and his research is something I greatly admire and aspire to.

I would also like to extend my sincere thanks to my dissertation committee members: Dr. George Michailidis, Dr. Aaron J. Molstad, and Dr. Arkaprava Roy. Their insightful feedback and guidance were invaluable to my work.

A special thanks goes to my collaborator, Dr. Jeffrey W. Miller, for his critical insights and advice on our paper. His meticulous attention to detail and his writing style have deeply influenced me. I also want to thank my collaborators Dr. Jason Xu for the engaging discussions, Dr. Eleni Dilma and Yang Yaozhi, who were co-first authors on two of my dissertation works and did an outstanding job with experiments and coding, and Dr. George Michailidis, who provided many key ideas. I am also grateful to Dr. Hitoshi Iyatomi for his invaluable help with data and data processing, which was essential for our paper's success. Additionally, I appreciate the contributions of Dr. Zhang Zhengwu, who provided significant help and data for our work, even though we did not publish a paper together.

I would like to thank the professors in our department, especially Dr. Brett Presnell and Dr. Hani Doss, whose comprehensive lectures and meticulously prepared materials greatly enhanced my learning. My gratitude also extends to the professors who taught me courses: Dr. Jim Hobert, Dr. Malay Ghosh, Dr. Larry Winner, Dr. Kshitij Khare, and Dr. Michael Daniels.

I also wish to acknowledge my former and current lab mates, Dr. Xu Maoran, Yuwen Zeyu, Zheng Yu, and Yang Yaozhi, for the stimulating discussions and camaraderie. A heartfelt thanks to my fellow graduate students in the department: Dr. Liu Yanxi, Dr. Bai Yichen, Dr. Fan Xiran, Dr. Qin Zikun, Qu Xiaoda, Zhang Yiqiao, Yang Lei, Animesh Mitra, and Dr. Tian Yuhan, for the

joy and support they brought into my life, especially during the mundane or challenging times. To my dear friends Dr. Chen Ranyiliu, Yi Shun and Pu Junqian, your constant support and encouragement mean the world to me. Thank you for always standing by my side.

Finally, my deepest gratitude goes to my parents, Zeng Zhaojuan and Yan Zeling. They have supported me unconditionally from the moment I decided to pursue my doctorate, through all the ups and downs of this journey. Their unwavering love and care have been my bedrock, providing strength when I needed it most. I also want to honor my maternal grandfather, Yan Hui, and maternal grandmother, Chen Jinping whose passing during my doctoral studies was a great loss. Their emotional support and wisdom guided me through many moments of my life, and their influence will always remain with me.

TABLE OF CONTENTS

page

LIST OF TABLES

Tables page

LIST OF FIGURES

Figures page

Abstract of Dissertation Presented to the Graduate School
of the University of Florida in Partial Fulfillment of the
Requirements for the Degree of Doctor of Philosophy

NEW METHODS AND LARGE SAMPLE THEORY IN BAYESIAN NONPARAMETRICS AND SEMIPARAMETRICS

By

Cheng Zeng

August 2025

Chair: Leo L. Duan
Major: Statistics

In modern statistical analysis, the ability to extract meaningful insights from complex, high-dimensional data has become increasingly vital. This dissertation explores advanced Bayesian methodologies aimed at addressing challenges inherent in statistical estimation, particularly in high-dimensional contexts where traditional methods often fall short. Specifically, we focus on three primary contributions: Bayesian nonparametrics, generalized Bayesian frameworks, and conditional density estimation.

The first contribution presents a novel approach to mixture modeling and clustering, employing a modified Dirichlet process mixture model. By introducing a quasi-Bernoulli random variable to adjust the stick-breaking process of mixture weights, we achieve improved estimation of the number of clusters while maintaining computational efficiency. Our theoretical results demonstrate that, under certain conditions, the posterior distribution effectively converges to the true number of clusters, leading to enhanced model performance in applications such as clustering brain networks.

The second contribution develops a generalized Bayes approach that bridges optimization techniques with Bayesian modeling. This novel framework, termed bridged posterior, facilitates efficient uncertainty quantification while retaining the flexibility of Bayesian inference. Our findings reveal that the $\sqrt{n}$-adjusted posterior distribution converges to the same normal distribution as that of the canonical integrated posterior, thus dispelling prior misconceptions

about the implications of optimizing latent variables on parameter uncertainty. This approach is applied to various settings, including maximum-margin classification and latent normal models, showcasing its practical advantages.

Furthermore, we propose a posterior inference approach for model parameters that are defined as the solutions to optimization sub-problems, by using the first-order optimality. This method, termed the gradient-bridged posterior, is amenable to efficient posterior computation, and enjoys theoretical guarantees, establishing a Bernstein–von Mises theorem for asymptotic normality. This can be understood as a alternative model to the Bridged posterior framework to enable concentration around partial minimizers. The advantages of our approach are highlighted on a synthetic flow network experiment and an application to data integration using Procrustes distances.

Lastly, we address the challenge of conditional density estimation for high-dimensional responses, such as images. We propose an Augmented Posterior Conditional Density Estimation (AP-CDE) framework that leverages normalizing flow neural networks to model complex relationships between predictors and high-dimensional responses. Our method enables the effective separation of relevant variations from noise, significantly improving interpretability in applications involving image analytics.

Through these contributions, this dissertation advances the understanding and application of Bayesian methodologies, offering novel solutions for statistical estimation in nonparametric and semiparametric settings.

CHAPTER 1
INTRODUCTION

As modern data collection continues to expand in both size and complexity, the need for advanced statistical tools to understand and model data has become critical. Bayesian nonparametrics and semiparametrics have emerged as essential tools for statistical modeling, offering flexible frameworks that adapt to the complexity of real-world data. This dissertation explores large sample theory and applications in these domains, specifically focusing on density estimation, clustering, and conditional density estimation. We propose several Bayesian nonparametric models to solve some key issues, e.g., inconsistency of estimating the number of clusters, provide theoretical guarantees in large sample theory and demonstrate the advantages using simulation and data application. This dissertation also constructs a bridge between optimization and generalized Bayesian semiparametric methods, with large sample consistency theory.

1.1 Bayesian Statistics and Density Estimation

In real-world scenarios, data often comes with inherent variability due to random noise, measurement error, or other sources of randomness. To extract meaningful insights from this data, it is essential to estimate underlying parameters that characterize the population or process generating the data. These parameters, such as the mean, variance, help understand and summarize complex data. Since they are important, we not only want point estimation, but also would like to know ranges which cover the true parameters with high probability—this is the task of uncertainty quantification.

In some cases, when we characterize the generative probability distribution using infinite many parameters or we assume that the distribution is sampled from a general function space, the model is called nonparametric. In traditional parametric methods, the population density is assumed to follow a known distribution (e.g., Gaussian), but real-world data often exhibit much more complex structures that cannot be adequately captured by simple parametric models. Nonparametric approaches, on the other hand, allow the data to dictate the shape of the distribution, providing flexibility in modeling a wide range of data patterns.

In some real-world problems, we are interested in two parts of parameters, where one part is parametric (finite many parameters), while the other part is nonparametric (infinite many parameters). We refer to these models as semiparametric models.

1.1.1 Bayesian Framework

In a Bayesian framework, the parameter or function of interest is treated as a random variable, and prior knowledge or beliefs about it are encoded in a prior distribution. Once data are observed, the prior is updated using Bayes' Theorem to obtain the posterior distribution. Suppose we have independent and identically distributed (i.i.d.) observations $y = y_{1:n} = (y_1, \ldots, y_n)$ from unknown distribution $L(y; \theta)$ with parameter $\theta \in \Theta$. Given data, the posterior distribution of the parameter θ is:

$$\Pi(\theta|y) = \frac{L(y;\theta)\pi_0(\theta)}{\int_\Theta L(y;\theta)\pi_0(\theta)\, d\theta}$$

where $L(y; \theta)$ is the likelihood function and $\pi_0(\theta)$ is the prior. The posterior distribution combines the information from the prior and the observed data, allowing us to both estimate θ and quantify uncertainty.

1.1.2 Density Estimation and Mixture Model

One key problem in statistics is density estimation, where the goal is to infer the underlying probability distribution that generated a given set of data. Density estimation plays an important role in many fields, including anomaly detection, clustering, and generative modeling. For instance, in image generation, density estimation models like normalizing flows learn the underlying distribution of image datasets, allowing for the creation of realistic images by transforming simple latent variables into high-dimensional pixel data; in finance, density estimation is used to detect anomalies by modeling the distribution of normal transaction data, with outliers indicating potential fraud or irregular market activity.

Formally, given i.i.d. observations $y_{1:n} = (y_1, \ldots, y_n)$ from some unknown distribution with probability density function $f \in \mathcal{F}$, the goal of density estimation is to estimate f. Parametric methods assume that f belongs to a specific family of distributions (e.g., Gaussian), but such assumptions are often too restrictive for complex real-world data. Nonparametric methods, on the

other hand, make fewer assumptions and allow for greater flexibility in modeling the underlying data-generating process.

Among rich density estimation methods, mixture models are frequently used to analyze data with unknown group/cluster structure. Suppose

$$y_i \overset{iid}{\sim} \sum_{k=1}^{K} w_k \mathcal{F}(\theta_k)$$

for $i = 1, \ldots, n$, where $\mathcal{F}(\theta_k)$ is a distribution parameterized by θ_k, and $w_1, \ldots, w_K \geq 0$ such that $\sum_{k=1}^{K} w_k = 1$. Using the Bayesian framework that posits a prior distribution for the weights $w = w_{1:K} = (w_1, \ldots, w_K)$ and the parameters $\theta = \theta_{1:K} = (\theta_1, \ldots, \theta_K)$, one can infer the posterior distribution of the weights w and the parameters θ, as well as the assignments of data points to mixture components, which yields a clustering of the data (Fraley and Raftery, 2002).

1.1.3 Conditional Density Estimation

A conditional density characterizes the probabilistic behavior of a set of random variables, when information on a set of other variables is available. The case of a single variable y (the response) conditioned on a multivariate x (predictor) has received most attention in the literature, due to a wide range of applications. A number of methods have been proposed to address the conditional density estimation problem from observed data. Kernel density (Terrell and Scott, 1992; Botev et al., 2010; Kim and Scott, 2012) and k-nearest neighbors (Mack and Rosenblatt, 1979; Kung et al., 2012) based techniques have been extensively studied and employed in applications. Another popular approach uses a mixture model of the form $f(y_i \mid x_i) = \sum_{k=1}^{K} w_k(x_i) g[y_i \mid \theta_k(x_i)]$ over data index $i = 1, \ldots, n$, with $\sum_{k=1}^{K} w_k(x_i) = 1$ and potentially $K \to \infty$. In the mixture model and for continuous y_i, $g(\cdot)$ can be a location-scale density, such as a multivariate Gaussian one with mean μ_k and covariance matrix Σ_k. Importantly, the mixture component parameters θ_k as well as the mixture weights w_k are some deterministic transforms of the x_i's, hence allowing the mixture distribution of y_i to vary according to x_i. There

is a large literature in such a framework, see, e.g., Jiang and Tanner (1999); Geweke and Keane (2007); Villani et al. (2009); Norets (2010); Huang et al. (2022) and references therein.

The conditional density estimation framework has a wide range of statistical applications, besides serving as a nonlinear predictive model for y_i, including outlier detection and classification. In the first case, by evaluating the magnitude of $f(y_i \mid x_i)$ for each observed data point, one could identify those points in the bottom density quantile as potential outliers (Scott, 2004; Schubert et al., 2014). In the latter case, when x_i is a discrete class label, such as whether a patient is in disease status, with a class probability $p(x_i)$, $p(x_i \mid y_i) \propto p(x_i) f(y_i \mid x_i)$ could be used as a probabilistic classifier for predicting x_i (Garg and Roth, 2001).

Despite significant advances in conditional density estimation, in recent years, a number of major challenges are not satisfactorily addressed for a high-dimensional response $y_i \in \mathbb{R}^p$, such as an image. Hence, the focus of the current work is on the case of a high-dimensional response y_i—rather than low-dimensional response conditioned on a high-dimensional predictor x_i in extant literature. For the latter, a large class of solutions exist, such as BART (Chipman et al., 2010), that partition the high-dimensional space of x_i and use simple piece-wise distribution (such as spherical Gaussian, or Bernoulli) for the low-dimensional y_i in each region. Obviously, this strategy does not be applied to high-dimensional y_i.

1.2 Bayesian Nonparametric Generative Models for Density Estimation using Stick-breaking Process

In practice using the Bayesian mixture models, in addition to w_k and θ_k, we usually do not know the number of clusters either. Stick-breaking models provide an appealing solution in which the number of mixture components K is infinite, and the number of clusters in the data (that is, the number of components that the data are assigned to) can be any finite number. A general form of a stick-breaking model for the mixture weights w is

$$w_1 = v_1 \text{ and } w_k = v_k \prod_{l=1}^{k-1} (1 - v_l) \text{ for } k = 2, 3, \ldots, \tag{1-1}$$

where $v_1, v_2, \ldots$ are drawn from some prior distribution. The interpretation is that starting from a stick of length 1, for each k we break off a proportion $v_k \in [0, 1]$ from the remaining stick, and use it as the weight w_k.

Many priors have been proposed for the distribution of v_k. Perhaps the most widely used one is $v_k \sim \text{Beta}(1, \alpha)$, with which the probability distribution $\sum_{k=1}^{\infty} w_k \delta_{\theta_k}(\cdot)$ yields a realization of the Dirichlet process with concentration parameter α (Sethuraman, 1994), where $\delta_x(\cdot)$ represents the Dirac measure which satisfies $\delta_x(A) = \mathbb{1}(x \in A)$ for any measurable set A. More generally, when $v_k \sim \text{Beta}(1 - d, \alpha + kd)$ with $0 \leq d < 1$ and $\alpha > -d$, one obtains the Pitman–Yor process with discount parameter d and strength parameter α (Pitman and Yor, 1997; Ishwaran and James, 2001).

The Dirichlet process mixture model (DPMM) offers a flexible way to model data distributions without the need to specify the number of components beforehand, which makes it widely applicable in fields such as genomics, where clustering gene expression profiles is crucial. In such settings, Bayesian nonparametrics enables the discovery of latent structures in data without overfitting, as the Dirichlet process shrinks most of w_k towards zeros.

1.3 Normalizing Flow Generative Models for High-dimensional Data

When using the mixture models for density estimation, the following two difficulties arise for a high-dimensional data y: (i) specification of the component distribution g in the mixture model and (ii) the curse of dimensionality when computing a high-dimensional mixture distribution. These two points are elaborated in Chapter 5.

These challenges in high-dimensional density estimation approaches motivated the development of completely different approaches. Normalizing flow neural networks were proposed to find an invertible mapping between a random variable $y_i \in \mathbb{R}^p$ and a latent variable $z_i \sim \text{N}(0, I_p)$. The neural network is formed by stacking layers of non-linear transforms, each layer parameterized in the way such that it corresponds to a bijective transform, and the inverse transform has a closed-form or can be computed efficiently. Using a simple change-of-variable technique, one could obtain the density $f(y_i)$ as a transformed density from an independent

Gaussian one. Examples include RealNVP (Dinh et al., 2016), MADE (Germain et al., 2015), MAF (Papamakarios et al., 2017), Glow (Kingma and Dhariwal, 2018), FFJORD (Grathwohl et al., 2018) and iResNet (Behrmann et al., 2019; Chen et al., 2019). Due to the large number of parameters and expressiveness of neural networks, impressive performance has been exhibited as a generative model for y_i. For example, in image applications, after training a normalizing flow network, one could generate a new Gaussian vector $z_{i'}$ and push it forward through the trained network, with the transformed $y_{i'}$ often looking as if it were a real photo. Since there is only one neural network involved (despite a large number of parameters within it), the computation enjoys high efficiency through stochastic gradient descent. Its expressiveness as a generative model, tractability of the target density $f(y_i)$ and good computing performance make the normalizing flow a compelling alternative to mixture models for high-dimensional density estimation provided the training data set is large enough.

1.4 Generalized Bayesian Semiparametric Methods

The generalized Bayes approach is becoming increasingly popular due to its potential advantages in model simplicity and robustness. A generalized posterior can be specified based on partial information from the data, which appeals when the likelihood is inaccessible or intractable, or via a loss function that characterizes an inferential summary of the data. There is a well-established literature on partial information settings including methods based on composite likelihood (Lindsay, 1988; Varin et al., 2011), partial likelihood (Sinha et al., 2003; Dunson and Taylor, 2005), pairwise likelihood (Jensen and Künsch, 1994), and others. Recently, there has been a burgeoning interest in loss-based Bayesian models, including works involving classification loss (Polson and Scott, 2011) or distance-based losses (Duan and Dunson, 2021; Rigon et al., 2023a; Natarajan et al., 2023). Loss-based generalized Bayes models typically use a probability distribution called the Gibbs posterior (Jiang and Tanner, 2008), taking the form: $\Pi(\theta \mid y) \propto \exp\{-g(\theta, y)\}$, where $g(\theta, y)$ is some loss function with y the data and θ the parameter.

There is a vast generalized Bayes literature using the Gibbs posterior for explicit model weighting, with g chosen according to some utility function such as predictive accuracy (Lavine

et al., 2021; Tallman and West, 2024), scoring rule (Gneiting and Raftery, 2007; Dawid and Musio, 2015), fairness metrics (Chakraborty et al., 2024) or summary statistics-based divergence (Frazier and Drovandi, 2021; Frazier et al., 2023). Such an approach also lends itself to modular descriptions of data (Jacob et al., 2017), and can guard against model misspecification (Nott et al., 2023). With connections to these methods, our focus is on the case when one wants to adopt a loss g from the optimization literature for statistical modeling, while needing to quantify uncertainty beyond point estimates.

In addition to nonparametrics, this dissertation also delves into Bayesian semiparametrics, where a model combines both parametric and nonparametric components. Rather than treating θ as a high-dimensional random variable, we model $\theta = (z, \lambda)$ with only λ as a parameter in finite dimensional, while the z can be infinite dimensional. Bayesian Semiparametrics blends nonparametric flexibility with parametric structure, making it ideal for problems that require both interpretability and adaptability.

1.5 Organization of the Dissertation

The rest of this dissertation is organized as follows. In Chapter 2, we focus on Bayesian infinite mixture models, and propose a novel model using quasi-Bernoulli stick-breaking process. This addresses inconsistency on estimating number of clusters. We use large sample theoretical analysis and simulations to demonstrate the consistent estimation using our model. A data application in clustering brain networks using a mixture of low-rank probit models is presented. In Chapter 3, we develop a new generalized Bayes approach, constructing a bridge between optimization and Bayesian framework. This novel bridged posterior approach improves computational efficiency and interpretability, and addresses the challenge of high dimensionality in the data. We present a large sample theoretical finding about the $\sqrt{n}$-adjusted posterior distribution, and demonstrate the practical advantages of our approach under several settings. In Chapter 4, we extend the bridged posterior to a new gradient-bridged posterior, providing tractable posterior inference to models parameters involving implicit functions. We present the large sample theory about the posterior distribution, and show the practical advantages. In Chapter 5,

we focus on conditional density estimation problem for high-dimensional data, e.g., images. We incorporate the augmented posterior into the normalizing flow neural network. It produces a generative model for high-dimensional data that uses information from external predictors. We demonstrate the outperformed density estimation and useful dimension reduction achieved by the model. Finally, in Chapter 6, we summarize the contribution of our research work presented in this dissertation, and provide some discussion and potential future work.

CHAPTER 2
CONSISTENT MODEL-BASED CLUSTERING: USING THE QUASI-BERNOULLI STICK-BREAKING PROCESS

In this chapter, we focus on a key problem in the field of clustering and density estimation using Bayesian infinite mixture models.

2.1 Motivation

When the true data-generating distribution is a finite mixture from the assumed family, Dirichlet process mixture models and Pitman–Yor process mixture models tend to shrink most of w_k's close to zero, leading to a small number of clusters in the posterior. However, Miller and Harrison (2013, 2014) showed a striking result: neither the Dirichlet process prior nor the Pitman–Yor process prior allows the posterior distribution on the number of components K to converge to the true number of clusters as n increases, even when the family of component distributions $\mathcal{F}$ is correctly specified.

To address this issue, we develop a prior on the w_k's that yields stronger shrinkage while remaining easy to use. Specifically, we modify the canonical stick-breaking construction as follows: at each break, we multiply the remaining proportion $(1 - v_k)$ by a discrete random variable that takes value either 1 or ϵ. When the latter happens at the L-th stick, the tail probability $\sum_{k=L+1}^{\infty} w_k$ is strictly bounded by ϵ. We show that if ϵ is chosen in a sample-size-dependent way such that $\epsilon(n) = o(1/n^{2+r})$ (with r a non-negative integer that depends on α), then one can obtain posterior consistency for the number of clusters. In the special case of $\epsilon = 0$, this model reduces to a finite mixture model with a prior on the number of components (Miller and Harrison, 2018), which also yields posterior consistency for the number of clusters. Meanwhile, in the special case of $\epsilon = 1$ and $v_k \sim \text{Beta}(1, \alpha)$, this model reduces to a Dirichlet process mixture.

Therefore, using $\epsilon \in (0, 1)$ effectively interpolates between these two extremes. Compared to $\epsilon = 0$, using $\epsilon \in (0, 1)$ avoids having a singularity at $w_k = 0$. This relaxes the parameter space in a way that allows the Markov chain Monte Carlo (MCMC) sampler to more efficiently add and remove clusters, rather than employing an explicit discrete search over K. Further, it makes the

Reprinted with permission from Zeng et al. (2023).

technique compatible with more complex stick-breaking models, such as the ones involving kernels (Dunson and Park, 2008), geospatial processes (Rodríguez et al., 2010), and external predictors (Ren et al., 2011). On the other hand, compared to $\epsilon = 1$, it allows the model to behave effectively like a finite mixture with a prior on the number of components, leading to superior control on the number of clusters. We illustrate these advantages in simulations and a data application in clustering brain networks, using a mixture of low-rank probit models. A software implementation and the steps needed to replicate the results in this chapter are provided on https://github.com/zeng-cheng/quasi-bernoulli-stick-breaking.

2.2 Quasi-Bernoulli Stick-breaking Process

In the general form of the stick-breaking construction (Equation 1-1), if the proportion v_L at step L is very close to 1, then w_L will take almost all the remaining sticks, resulting in $w_k \approx 0$ for $k \geq L + 1$. Based on this intuition, we introduce the following stick-breaking process:

$$\begin{aligned} &w_1 = v_1, \quad w_k = v_k \prod_{l=1}^{k-1}(1 - v_l), \text{ for } k \geq 2, \\ &v_k = 1 - b_k\beta_k, \\ &b_k \overset{iid}{\sim} \tilde{p}\delta_1(\cdot) + (1 - \tilde{p})\delta_\epsilon(\cdot), \\ &\beta_k \overset{iid}{\sim} \text{Beta}(\alpha, 1). \end{aligned} \tag{2-1}$$

Each b_k follows a discrete distribution such that $b_k = 1$ with probability $\tilde{p} \in (0, 1)$, and $b_k = \epsilon$ with probability $1 - \tilde{p}$, for some small $\epsilon \in (0, 1)$. We refer to b_k as a quasi-Bernoulli random variable, since it resembles the standard Bernoulli supported on $\{0, 1\}$. We refer to Equation 2-1 as the quasi-Bernoulli stick-breaking process (or simply the "quasi-Bernoulli process"). With this prior on weights $w_1, w_2, \ldots$, we obtain an infinite mixture model by letting $y_i \overset{iid}{\sim} \sum_{k=1}^{\infty} w_k \mathcal{F}(\theta_k)$, where $\theta_k \overset{iid}{\sim} \mathcal{G}$ from some base measure $\mathcal{G}$; we refer to this as a quasi-Bernoulli mixture model. Moreover, this marginal representation of the mixture model can be equivalently represented in a conditional form with the introduction of the latent assignment variable c_i for each data point y_i.

Specifically, with $c_i = k$ representing the event that y_i is drawn from the mixture component k,

$$\begin{aligned} \theta_k &\overset{iid}{\sim} \mathcal{G} \text{ for } k \geq 1, \\ c_i \mid w &\overset{iid}{\sim} \text{Categorical}(w), \\ y_i \mid c_i, \theta &\overset{indep}{\sim} \mathcal{F}(\theta_{c_i}) \text{ for } i = 1, \ldots, n. \end{aligned} \tag{2-2}$$

This conditional representation is useful if one is interested in using the model to perform model-based clustering.

Note that if $\epsilon = 1$, then we would have $v_k = 1 - \beta_k \sim \text{Beta}(1, \alpha)$, yielding the stick-breaking representation of the Dirichlet process $\text{DP}(\alpha, \mathcal{G})$; whereas if $\epsilon = 0$, then we would have a random truncation on $(w_1, w_2, \ldots)$. Using $\epsilon \in (0, 1)$ yields a soft truncation that provides advantages from both of these extremes.

Before we move into more technical discussions, we first illustrate an intuition for why the Dirichlet process with a fixed α (as a data generating mechanism) tends to produce small clusters, and how our proposed prior mitigates this. Consider the scenario where we have n data points already assigned to K clusters, and there are m new data points to be assigned. Assume that there are only K clusters in the population, but we are modeling the data as drawn from a Dirichlet process mixture model. Ideally we want to assign all of the m new data into the existing K clusters. The prior probability of adding one or more new clusters can be calculated using the predictive rule (as in the "restaurant process") $\text{p}(c_i \mid c_1, \ldots, c_{i-1})$ for $i \in \{n+1, \ldots, n+m\}$ recursively for m times. For the Dirichlet process with concentration parameter α, this probability has a closed form:

$$\text{p}\left(\sum_{l=1}^{m} \mathbb{1}(c_{n+l} > K) > 0 \mid c_1, \ldots, c_n \right) = 1 - \prod_{l=1}^{m} \left(\frac{n + l - 1}{n + l - 1 + \alpha} \right),$$

which follows directly from the predictive distribution under the Dirichlet process (Blackwell and MacQueen, 1973, Equation (2)). Although the probability is small for $m = 1$, it increases rapidly and becomes non-trivial as m grows. Note that the above event includes not only assigning all m

data points into a single new cluster, but also having them scattered into several new clusters; hence this is the union of all outcomes of having small clusters.

For the quasi-Bernoulli process, although we do not have a simple closed-form expression for the above probability, we can numerically calculate it based on the partition probability function Equation 2-3 introduced in the next section. In Figure 2-1, we examine the case when given two existing clusters with $n_1 = n_2 = 50$, and assigning additional m data points. It can be seen that the prior probability of creating new cluster(s) quickly increases in the Dirichlet process, whereas the quasi-Bernoulli processes (with $\alpha = 1$ and two values for $\tilde{p}$) substantially slow down the growth of this probability. We also provide results with another rate of $\epsilon(n)$ in the appendix.

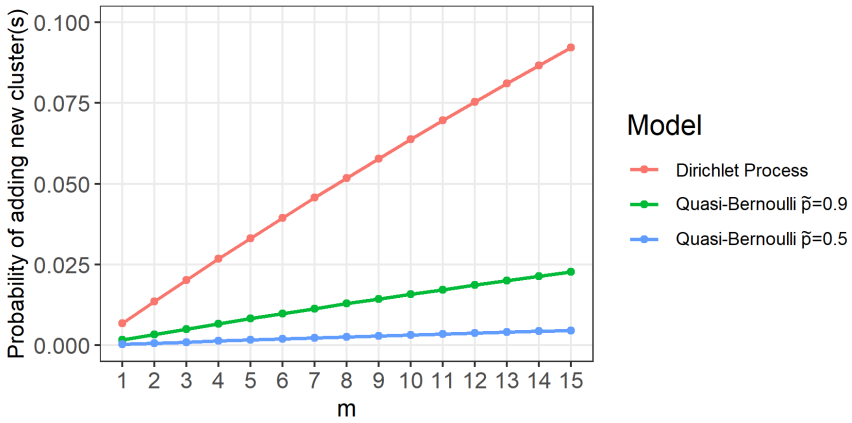


Figure 2-1. Under the prior, the Dirichlet process exhibits rapid growth in the probability of adding one or more new clusters for m future data points ($n = 100$), favoring the creation of additional clusters a priori. Meanwhile, the quasi-Bernoulli process exhibits much slower growth of this probability. For the quasi-Bernoulli process, we use $\epsilon = \epsilon(n, m) = 1/(n + m)^{2.1}$ and $\alpha = 1$ as suggested in our theory Theorem 2-4. For the Dirichlet process, we use $\alpha = 0.69$, which has the prior expected number of clusters close to the one under the quasi-Bernoulli process with $\tilde{p} = 0.9$ (see Table A-1).

We will carefully examine the posterior behaviors of the above models, including comparing with the Dirichlet process with $\alpha = \alpha(n)$ tending to zero as $n \to \infty$.

2.3 Theoretical Analysis

In this section, we provide a theoretical analysis for the quasi-Bernoulli process and the corresponding mixture model.

2.3.1 Exchangeable Random Partitioning

A large class of stick-breaking models enjoys the property of partition exchangeability—that is, the probability distribution of the induced partition only depends on the cluster sizes, and is invariant to any permutation of the cluster index (see Pitman, 1995, proposition 5). Letting $i \in \{1, \ldots, n\}$ be the data index, if there are t unique values in the cluster assignments $c = (c_1, \ldots, c_n)$, then we can form a corresponding partition $\mathcal{A} = \{A_1, \ldots, A_t\}$ in the following way: (1) initialize $A_1 = \{1\}$ and $t = 1$; (2) sequentially for $i = 2, \ldots, n$, if $c_i = c_j$ for any $j \leq i - 1$ and $j \in A_k$, then add i to the same set A_k containing j; otherwise create a new set $A_{t+1} = \{i\}$ and increment t to $t + 1$.

Theorem 2-1 (Exchangeable Partition Probability Function)**.** *The probability mass function of the random partition $\mathcal{A} = \{A_1, \ldots, A_t\}$ induced by c in the quasi-Bernoulli stick-breaking process is*

$$p_{\epsilon,n}(\mathcal{A}) = \frac{\alpha^t \Gamma(\alpha)}{\Gamma(n+\alpha)} \left(\prod_{j=1}^{t} \Gamma(n_j + 1) \right) \sum_{\sigma \in S_t} \prod_{j=1}^{t} \frac{\tilde{p} + (1-\tilde{p}) I_\epsilon(g_{j+1}(\sigma) + \alpha, n_{\sigma_j} + 1)/\epsilon^\alpha}{g_j(\sigma) + \alpha(1-\tilde{p})(1-\epsilon^{g_j(\sigma)})} \tag{2-3}$$

where $\sigma = (\sigma_1, \ldots, \sigma_t)$ is a permutation of $\{1, \ldots, t\}$, S_t is the set of all permutations of $\{1, \ldots, t\}$, $n_j = |A_j|$, $g_j(\sigma) = \sum_{l=j}^{t} n_{\sigma_l}$, $\Gamma(x) = \int_0^\infty z^{x-1} e^{-z}\, dz$ and $I_\epsilon(q_1, q_2)$ is the cumulative distribution function of Beta(q_1, q_2) evaluated at ϵ.

For conciseness, we defer all the proofs to the appendix. Using $\mathrm{p}_{\epsilon,n}(\mathcal{A})$, we can substantially simplify the model in Equations 2-1 and 2-2 into an equivalent generative process:

$$\begin{aligned} \mathcal{A} &\sim \mathrm{p}_{\epsilon,n}(\mathcal{A}), \\ \theta_k \mid \mathcal{A} &\overset{iid}{\sim} \mathcal{G} \text{ for } k = 1, \ldots, t, \\ y_i \mid \mathcal{A}, \theta &\overset{indep}{\sim} \mathcal{F}(\theta_j) \text{ for } i \in A_j,\ A_j \in \mathcal{A}. \end{aligned} \tag{2-4}$$

We now use the above representation to study the asymptotics of the clustering when $n \to \infty$.

2.3.2 Consistent Estimation of the Number of Clusters

By definition, the number of clusters is $t = |\mathcal{A}|$. We use T to denote the associated random variable in the model. Let $\mathcal{H}_t(n)$ denote the set of all partitions of $\{1, \ldots, n\}$ into t disjoint sets.

Using Equation 2-4, the marginal posterior of T is

$$\mathrm{p}_\epsilon(T = t \mid y_{1:n}) = \frac{\sum_{\mathcal{A} \in \mathcal{H}_t(n)} p(y_{1:n} \mid \mathcal{A}) \mathrm{p}_{\epsilon,n}(\mathcal{A})}{\sum_{\mathcal{A} \in \cup_{t'=1}^{n} \mathcal{H}_{t'}(n)} p(y_{1:n} \mid \mathcal{A}) \mathrm{p}_{\epsilon,n}(\mathcal{A})}, \tag{2-5}$$

where $p(y_{1:n} \mid \mathcal{A}) = \prod_{A \in \mathcal{A}} m(y_A)$, $y_A = (y_i : i \in A)$, and $m(y_A) = \int_\Theta \left(\prod_{i \in A} f_\theta(y_i)\right) d\mathcal{G}(\theta)$. Here, f_θ denotes the density of the component distribution $\mathcal{F}(\theta)$.

Suppose the data are generated from k_0 mixture components, with k_0 a fixed and finite number. We establish general conditions under which $\mathrm{p}_\epsilon(T = k_0 \mid y_{1:n}) \to 1$ as $n \to \infty$. Our proof involves two main parts: (i) we establish that this consistency property holds for the finite-dimensional model obtained by setting $\epsilon = 0$; and (ii) we bound the total variation distance between the prior distributions $\mathrm{p}_{\epsilon,n}(\mathcal{A})$ and $\mathrm{p}_{0,n}(\mathcal{A})$. We then show that this implies posterior consistency.

Consider the case when $\epsilon = 0$, that is, when b_k in Equation 2-1 is a Bernoulli random variable with success rate $\tilde{p}$. In this case, we refer to Equation 2-1 with $\epsilon = 0$ as the Bernoulli stick-breaking process. Let $K = \min\{k : b_k = 0\}$. When this occurs, we have $w_{K+1} = w_{K+2} = \cdots = 0$, and therefore, w is effectively K-dimensional. Observe that K follows a geometric distribution: $\pi_K(k) = \tilde{p}^{k-1}(1 - \tilde{p})$ for $k \in \{1, 2, \ldots\}$. Thus, the Bernoulli stick-breaking process has the following equivalent representation:

$$\begin{aligned} &K \sim \text{Geometric}(1 - \tilde{p}), \\ &v_k \overset{iid}{\sim} \text{Beta}(1, \alpha), \quad \text{for } k = 1, \ldots, K-1; \quad v_K = 1, \\ &w_1 = v_1, \quad w_k = v_k \prod_{l=1}^{k-1} (1 - v_l), \quad \text{for } k \geq 2. \end{aligned} \tag{2-6}$$

The following result establishes the exchangeable partition probability function for the Bernoulli stick-breaking process.

Lemma 2-1. *The probability mass function of the random partition $\mathcal{A} = \{A_1, \ldots, A_t\}$ in the Bernoulli stick-breaking process is*

$$p_{\epsilon=0,n}(\mathcal{A}) = \frac{\alpha^t \Gamma(\alpha)}{\Gamma(n+\alpha)} \left(\prod_{j=1}^{t} \Gamma(n_j + 1) \right) \sum_{\sigma \in S_t} \prod_{j=1}^{t} \frac{\tilde{p} + \mathbb{1}(j = t)(1-\tilde{p})/(\alpha \mathrm{B}(\alpha, n_{\sigma_t} + 1))}{g_j(\sigma) + \alpha(1-\tilde{p})}$$

where $n_j = |A_j|$, $g_j(\sigma) = \sum_{l=j}^{t} n_{\sigma_l}$, and $\mathrm{B}(q_1, q_2) = \int_0^1 z^{q_1 - 1}(1-z)^{q_2 - 1}\, dz$, the Beta function.

We now establish that the Bernoulli stick-breaking process mixture model (quasi-Bernoulli mixture model with $\epsilon = 0$) exhibits posterior consistency for K (the number of non-zero w_k's) and T (the number of clusters). To clarify, even for a large n, T could be less than the number of components K, if some component does not have any data point assigned to it. Let Ω denote the set of parameter tuples $\phi := (k, w_1, \ldots, w_k, \theta_1, \ldots, \theta_k)$ such that $k \in \{1, 2, \ldots\}$, $w_1, \ldots, w_k > 0$, $\sum_{l=1}^{k} w_l = 1$, and $\theta_1, \ldots, \theta_k \in \Theta$, the parameter space. Let Ω' denote the subset of Ω such that $\theta_i \neq \theta_j$ for all $i \neq j$. Further, let P_ϕ denote the mixture distribution $P_\phi := \sum_{l=1}^{k} w_l \mathcal{F}(\theta_l)$. When $\phi \in \Omega'$ is identifiable from P_ϕ up to permutation of the mixture components, we can define a transformation $\eta : \Omega \to \Omega'$ such that the parameter $\phi' = \eta(\phi)$ is fully identifiable from $P_{\phi'}$. See Nobile (1994, Section 3.2) for the details. Any prior on Ω defines an induced prior on Ω' through η.

Theorem 2-2. *Assume $\phi \in \Omega'$ is identifiable up to permutation of the mixture components. Let Π_0 be the prior on Ω under the model defined by Equations 2-2 and 2-6, and assume $\Pi_0(\{\phi : \exists\, i \neq j \text{ such that } \theta_i = \theta_j\}) = 0$. Let Π_0' be the corresponding prior on Ω' induced by η. Then there is a subset $\Omega_0' \subset \Omega'$ with $\Pi_0'(\Omega_0') = 1$ such that for any $\phi_0 = (k_0, w_1^0, \ldots, w_{k_0}^0, \theta_1^0, \ldots, \theta_{k_0}^0) \in \Omega_0'$, if $y_1, y_2, \ldots \mid \phi_0 \overset{iid}{\sim} P_{\phi_0}$ and the component density f_θ is continuous (with respect to θ) at each θ_k^0, then as $n \to \infty$, we have*

$$p_{\epsilon=0}(K = k_0 \mid y_{1:n}) \to 1 \quad \text{a.s.}[P_{\phi_0}],$$
$$p_{\epsilon=0}(T = k_0 \mid y_{1:n}) \to 1 \quad \text{a.s.}[P_{\phi_0}],$$

where a.s.$[P_{\phi_0}]$ *denotes almost surely convergence under the probability distribution* P_{ϕ_0}.

The first part of the result is a corollary of Nobile (1994, Proposition 3.5), the proof of which is an application of the Doob's theorem. The intuition for the second part is that since $w_1^0, \ldots, w_{k_0}^0$ are positive and do not change with n, we can expect that at least some data will be assigned to each component $k = 1, \ldots, k_0$, and thus that K and T will match in the posterior.

Now, consider the case of $\epsilon > 0$. Intuitively, when ϵ is small, we would expect the posterior to behave similarly to the case of $\epsilon = 0$. Formally, we show that this is indeed the case when $\epsilon = \epsilon(n) \to 0$ at an appropriate rate as $n \to \infty$. To show this, we employ the following bound on the distance between the partition distributions for $\epsilon > 0$ and $\epsilon = 0$ under the prior.

Theorem 2-3 (Prior equivalence as $\epsilon(n) \to 0$)**.** *Assume* $\epsilon \leq 1/n$. *Under the quasi-Bernoulli priors, the total variation distance between the partition distributions for* $\epsilon > 0$ *and* $\epsilon = 0$ *satisfies the bound*

$$\sup_{A \in \mathscr{A}} |p_{\epsilon,n}(A) - p_{0,n}(A)| \leq \sqrt{\frac{\alpha n \epsilon}{2(\alpha + 1 - \alpha\epsilon n)}}$$

where $\mathscr{A}$ *denotes the set of all subsets of* $\cup_{t=1}^{n} \mathcal{H}_t(n)$, *and* $p_{\epsilon,n}(A) = \sum_{\mathcal{A} \in A} p_{\epsilon,n}(\mathcal{A})$. *In particular, if* $\epsilon(n) = o(1/n)$, *then*

$$\sup_{A \in \mathscr{A}} |p_{\epsilon(n),n}(A) - p_{0,n}(A)| \xrightarrow[n \to \infty]{} 0.$$

The interpretation of this result is that if we control ϵ to be slightly smaller than $1/n$, then we have a stick-breaking model supported in the infinite-dimensional space that asymptotically approximates the finite-dimensional model with posterior consistency guarantees.

Now, moving to the posterior, we have the consistency of the quasi-Bernoulli model for the number of clusters.

Theorem 2-4 (Posterior consistency)**.** *Under the same assumptions and notations of Theorem 2-2, if* $\epsilon(n) = o(1/n^{2+r})$, *where* r *is the integer such that* $\max(\alpha - 1, 0) \leq r < \alpha$, *then*

$$p_{\epsilon(n)}(T = k_0 \mid y_{1:n}) \xrightarrow[n \to \infty]{} 1 \quad \text{a.s.}[P_{\phi_0}].$$

Note that here we assume $\epsilon(n) = o(1/n^{2+r})$ rather than $o(1/n)$ as in Theorem 2-3, however, we expect that the involved inequalities could be tightened further.

It should also be noted that letting $\epsilon(n)$ depend on n makes the resulting sequence of models no longer projective. That is, the model for n data points does not coincide with the distribution obtained by taking the model for $n+1$ data points and integrating over data point $n+1$. However, to achieve certain optimal asymptotic behaviors such as consistency, it is common to calibrate the prior based on the sample size (for example, see Castillo et al., 2015 on the choice of prior for variable selection). Alternatively, one could always use $\epsilon = 0$, which achieves both consistency and projectivity, but this is less flexible and less efficient in terms of computation.

2.3.3 Comparison with the Asymptotic Behavior of the Dirichlet Process

The results of Miller and Harrison (2013, 2014) on the inconsistency of the Dirichlet process mixture model show that for a fixed value of the concentration parameter α, the model asymptotically over-estimates the number of clusters on data from a finite mixture.

Our theoretic finding in the quasi-Bernoulli raises a tempting question: can we achieve consistency with a Dirichlet process mixture if we let $\alpha = \alpha(n)$ go to 0 at an appropriate rate as n increases? To our knowledge, this remains an open question. However, here we provide a partial answer (in the negative), by showing that if $\alpha(n) \to 0$ too fast, then the Dirichlet process remains inconsistent for the number of clusters.

Lemma 2-2. *Suppose the data are* $y_1, \ldots, y_n \overset{iid}{\sim} 0.5\,\mathrm{N}(0,1) + 0.5\,\mathrm{N}(\kappa, 1)$ *for any fixed* $\kappa \in \mathbb{R}$, *where* $\mathrm{N}(\mu, \sigma^2)$ *is the Gaussian distribution with density* $f_{\mu,\sigma^2}(x) \propto \exp\{-(x-\mu)^2/(2\sigma^2)\}$. *Consider a Dirichlet process mixture model with the mixture components* $\mathcal{F}(\theta) = \mathrm{N}(\theta, 1)$, *the base measure* $\mathcal{G} = \mathrm{N}(0,1)$ *(the prior on the parameter* θ*), and concentration parameter* $\alpha = \alpha(n)$ *such that* $\alpha(n) = o(\exp(-a_0\, n))$ *for some constant* $a_0 > 1/2 + \kappa^2/4$. *Then* $p(T = 2 \mid y_{1:n}) \xrightarrow{\text{a.s.}} 0$ *as* $n \to \infty$.

To clarify, the purpose of the above result is not to provide practical guidance on choosing the rate of $\alpha(n) \to 0$. Indeed, the problem of the Dirichlet process mixture is usually overestimation rather than underestimation of the number of clusters in the limit (Yang et al.,

2020a). Rather, this result shows that if consistency could be achieved for the Dirichlet process mixture model under a sample-size-dependent $\alpha(n)$, the rate of this hyper-parameter needs to satisfy both an upper and a lower bounds, which may turn out to be practically challenging for the users—there is a somewhat delicate sensitivity issue. In comparison, a strength of the quasi-Bernoulli mixture is that for obtaining consistency ϵ only needs to satisfy one upper bound $o(1/n^{2+\alpha})$, which means the user can simply use a small ϵ such as $1/n^{2+\alpha}$ (or smaller), without worrying about impacting the consistency. We provide empirical comparison via simulations with different cases of $\alpha(n)$ and $\epsilon(n)$ in the Section A.2.2.

The above result does not exclude the possibility that the Dirichlet process with $\alpha \to 0$ at a slower rate could achieve consistency. For example, recently Ohn and Lin (2023) show that setting $\alpha \approx n^{-a_1}$ for some positive number a_1 will guarantee $\text{pr}(T > Ck_0 \mid y_{1:n}) \to 0$ for some constant $C > 1$, hence preventing severe over-estimation in the number of clusters—although whether one could exactly recover k_0 is still unknown. Alternatively, another possibility is to put a hyper-prior on α. Ascolani et al. (2023) show that this method can achieve consistency when the component distribution $\mathcal{F}$ has bounded support. To our best knowledge, the consistency with general $\mathcal{F}$ remains an open question.

2.4 Posterior Sampling Algorithm

Since the quasi-Bernoulli mixture model involves a small modification to classic stick-breaking construction, we can use an efficient slice sampling algorithm [Kalli et al. (2011), as the improved version of Walker (2007)] for posterior inference. We use a sequence of decreasing positive constants $\xi_1, \xi_2, \ldots$ that converges to zero. In this article, we choose $\xi_i = 0.5^i$ for $i \geq 1$ as suggested in Kalli et al. (2011). Given c_i (the component assignment), consider a latent uniform $u_i \sim \text{Uniform}(0, \xi_{c_i})$, then we have a joint likelihood proportional to $\prod_{i=1}^{n} \mathbb{1}(u_i < \xi_{c_i}) w_{c_i} / \xi_{c_i} f_{\theta_{c_i}}(y_i)$. We define the state of the Markov chain to be (c, θ, w, u) and the target distribution is the posterior $p(c, \theta, w, u \mid y)$, where $y = y_{1:n}$, $c = c_{1:n}$, $\theta = \theta_{1:\infty}$, and $w = w_{1:\infty}$.

The slice sampler iterates the following steps:

1. *Sample c from its full conditional.* For $i = 1, \ldots, n$: sample $c_i \sim \text{Categorical}(\tilde{w})$ where

$$\tilde{w}_k = \frac{w_k / \xi_k f_{\theta_k}(y_i)}{\sum_{\{l:\xi_l > u_i\}} w_l / \xi_l f_{\theta_l}(y_i)}, \qquad \text{for } k \in \{l : \xi_l > u_i\}.$$

   Since the sequence $\xi_1, \xi_2, \ldots$ converges to zero, the index set $\{l : \xi_l > u_i\}$ is finite. Compute $n_k := \sum_i \mathbb{1}(c_i = k)$, and $m_k := \sum_i \mathbb{1}(c_i > k)$.

2. *Sample u from its full conditional.* For $i = 1, \ldots, n$: sample u_i from the uniform distribution over the interval $(0, \xi_{c_i})$.

3. *Sample w from its full conditional.* For $k \in \cup_{i=1}^{n} \{l : \xi_l > u_i\}$:

   - Sample $b_k \sim q\delta_1(\cdot) + (1-q)\delta_\epsilon(\cdot)$ where

$$q = \frac{\tilde{p}}{\tilde{p} + (1-\tilde{p})\epsilon^{-\alpha} I_\epsilon(m_k + \alpha, n_k + 1)}.$$

   - Sample β_k by drawing $X \sim \text{Beta}_{(0,b_k)}(m_k + \alpha, n_k + 1)$ and setting $\beta_k = X / b_k$, where $\text{Beta}_{(0,\epsilon)}$ denotes a Beta distribution truncated to the interval $(0, \epsilon)$.
   - Compute w_k from $b_{1:k}$ and $\beta_{1:k}$ using Equation 2-1.

4. *Sample θ from its full conditional.* For $k \in \cup_{i=1}^{n} \{l : \xi_l > u_i\}$: sample θ_k from the distribution proportional to $g(\theta_k) \prod_{i:c_i=k} f_{\theta_k}(y_i)$, where g is the density of the base measure $\mathcal{G}$.

Here, f_θ denotes the density of the component distribution $\mathcal{F}(\theta)$. Note that in the w update, we first sample b_k marginalized over β_k, then sample $\beta_k \mid b_k$, and compute the resulting value of w_k.

2.5 Simulations

In this section, we assess the empirical performance of the quasi-Bernoulli (QB) mixture model in simulation studies. We compare our method with three popular alternatives: Dirichlet process (DP) mixture, Pitman–Yor process (PY) mixture and finite mixture with a prior on the number of components (MFM).

We demonstrate the consistency of the quasi-Bernoulli mixture model for the number of clusters T when the family of component distributions is correctly specified. We set $\tilde{p} = 0.9$ for the quasi-Bernoulli probability in Equation 2-1, yielding a prior mean of no more than $1/(1-\tilde{p}) = 10$ components with mixture weights larger than ϵ. The reason is that the random variable K for the first time $b_k = \epsilon$ follows Geometric$(1-\tilde{p})$, and the weights after K are less than ϵ. Alternatively, one could place a Beta hyper-prior on $\tilde{p}$ to make the prior even more weakly informative. We set $\alpha = 1$ and $\epsilon = 1/n^{2.1}$, which satisfies the theoretical condition that $\epsilon = o(1/n^2)$ in Theorem 2-4. To have a fair comparison, we set the Dirichlet process concentration parameter α so that the expected number of clusters under the Dirichlet process prior is as close as possible to the one under the quasi-Bernoulli process prior for each n, as shown in Table A-1 in the appendix. Similarly, for the Pitman–Yor process prior PY(α, d) where each v_k follows Beta$(1-d, \alpha+kd)$, we choose α and d to match both the expectation and the variance of the number of clusters with the quasi-Bernoulli process prior. For the MFM model, we follow Miller and Harrison (2018) and set $\Pi_K(k) = p^{k-1}(1-p)$ where $p = 0.9$ for the prior on number of components, and $(w_1, \ldots, w_K) \sim \text{Dir}_K(\alpha, \ldots, \alpha)$ where $\alpha = 1$ for the prior on mixture weights.

For each experiment, we run the Markov chain for 50,000 iterations, discard the first 20,000 as burn-ins, and use thinning by keeping only every 50th iteration.

2.5.1 Simulations with Gaussian Mixtures

To compare performance in terms of consistency and MCMC mixing, we first consider simulations using Gaussian distributions $\text{N}(\mu, \Sigma)$ for the mixture components.

We first generate data with sample sizes $n \in \{50, 100, 250, 1000, 2500\}$ from a three-component univariate Gaussian mixture distribution:
$0.3\,\text{N}(-4, 1^2) + 0.3\,\text{N}(0, 1^2) + 0.4\,\text{N}(5, 1^2)$. Following Richardson and Green (1997), we use a data-dependent prior (that is, base measure $\mathcal{G}$) on the component parameters (μ, Σ):
$\mu \sim \text{N}(m_\mu, s_\mu^2)$ and $\Sigma \sim \text{Gamma}^{-1}(2, \gamma)$ where $\text{Gamma}^{-1}(a, b)$ has density
$f(x) \propto x^{-a-1}\exp(-b/x)$, with a hyper-prior $\text{Gamma}(g, h)$ on γ, where
$m_\mu = (\max\{y_{1:n}\} + \min\{y_{1:n}\})/2$, $s_\mu = \max\{y_{1:n}\} - \min\{y_{1:n}\}$, $g = 0.2$, and $h = 10/s_\mu^2$.

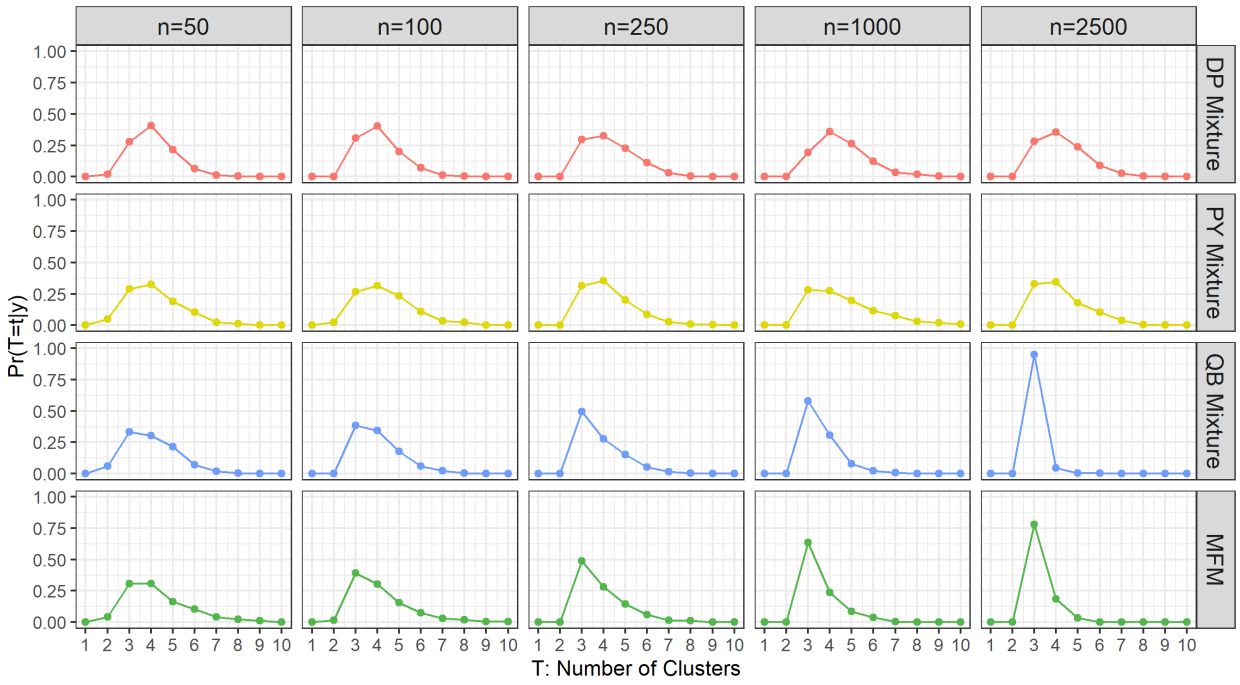
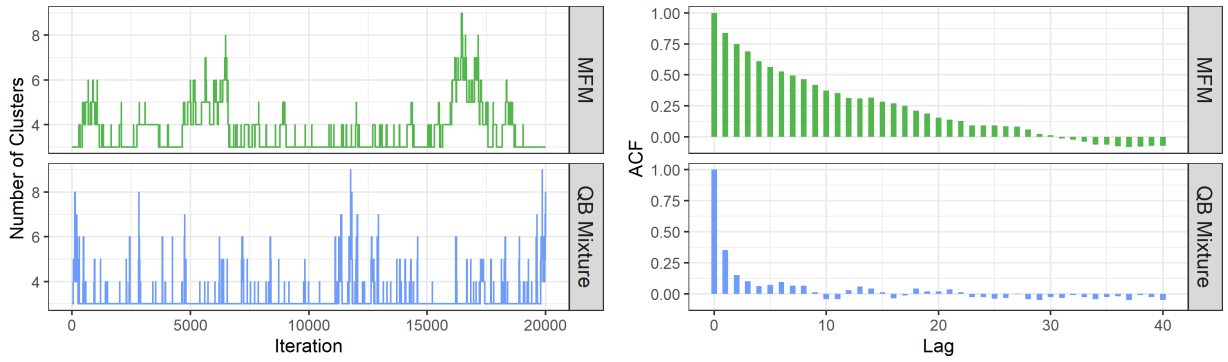


Figure 2-2. Posterior distribution of the number of clusters (T) for data from a three-component univariate Gaussian mixture. The quasi-Bernoulli mixture model correctly concentrates on three clusters, and its posterior distribution of T concentrates to a point mass at $k_0 = 3$. Coherent with our theory, at large n, the posterior distributions become almost identical to the ones using the MFM model. However, the posterior distributions of calibrated DP mixture model and PY mixture model do not concentrate to a point mass at $k_0 = 3$.

Figure 2-2 plots the posterior distribution of the number of clusters T at each n. Under the quasi-Bernoulli mixture model (shown in blue), the posterior of T concentrates to a point mass at the true number of components ($k_0 = 3$) as n grows, in accordance with our theory. Further, for both small and large n, the posterior mode of T coincides with the true number of components. Clearly, our model yields almost the same results (blue) as the MFM model (green), especially at large n. This is coherent with our theory (Equation 2-6). On the other hand, the Dirichlet process mixture model and the Pitman–Yor process mixture model fail to concentrate on the true number of components.

Despite similar performances in achieving consistency, a major strength of our model is its computational efficiency gained via the slice sampling (Kalli et al., 2011). In comparison, the existing MFM model requires a combinatorial search via the split-merge sampler (Jain and Neal,

2007), which suffers from slow mixing with high auto-correlation. As shown in Figure 2-3, with thinning, quasi-Bernoulli mixture model using slice sampler shows a drop in the auto-correlation (effective sample size 16.0%, on average of five experiments with sample size 1000), while MFM model using the split-merge sampler shows a much slower drop (effective sample size 7.8%).



(a) Trace on number of clusters

(b) Auto-correlation (based on thinning at 50)

Figure 2-3. The trace of the Markov chain on T and auto-correlation functions for univariate Gaussian mixture data with sample size 1000. Quasi-Bernoulli mixture model using the slice sampler shows much better mixing in the Markov chain, compared to the MFM model using the split-merge sampler. We discard the first 5,000 iterations as burn-ins and record the following 20,000 samples. The slice sampling is not available to the MFM model, since the change to K needs to satisfy the constraint that a non-empty cluster should not have a zero mixture weight.

Next, we consider a multivariate simulation scenario in which we generate data sets of size $n \in \{250, 1000, 2500\}$ from a three-component bivariate Gaussian mixture: $0.3\,\mathrm{N}((-4\ 1)^{\mathrm{T}}, I_2) + 0.3\,\mathrm{N}((0\ 2)^{\mathrm{T}}, I_2) + 0.4\,\mathrm{N}((5\ 3)^{\mathrm{T}}, I_2)$. We use the data-dependent prior $\mu \sim \mathrm{N}(m, C)$, $\Sigma \sim \mathrm{Wishart}_2^{-1}(C^{-1}/2, 2)$ on the component parameters, where m is the sample mean and C is the sample covariance. The results are similar to the univariate simulation scenario; see Section A.2.1.

2.5.2 Simulations with Non-Gaussian Mixtures

The quasi-Bernoulli mixture model can easily be extended to mixture models with non-Gaussian components. To illustrate that the consistency result still holds, we consider data generated from a mixture of Laplace distributions.

We generate data from a three-component Laplace mixture:
$0.35\text{Lap}(-10, 1) + 0.3\text{Lap}(0, 1.5) + 0.35\text{Lap}(10, 0.5)$, where $\text{Lap}(\mu, \lambda)$ denotes a Laplace distribution with mean μ and scale λ. We use a data-dependent prior (base measure $\mathcal{G}$) on (μ, λ): $\mu \sim \text{N}(m_\mu, \sigma_\mu^2)$ and $\lambda \sim \text{Gamma}^{-1}(2, 1)$, where $m_\mu = (\max\{y_{1:n}\} + \min\{y_{1:n}\})/2$ and $\sigma_\mu = \max\{y_{1:n}\} - \min\{y_{1:n}\}$. Figure 2-4 shows that the quasi-Bernoulli process successfully recovers the true number of components, while the Dirichlet process (red) and the Pitman–Yor process (yellow) fail to do so. Under a mixture distribution like the Laplace distribution having heavier tails than the Gaussian distribution, the Dirichlet process and the Pitman–Yor process tend to overestimate the number of clusters to a greater extent.

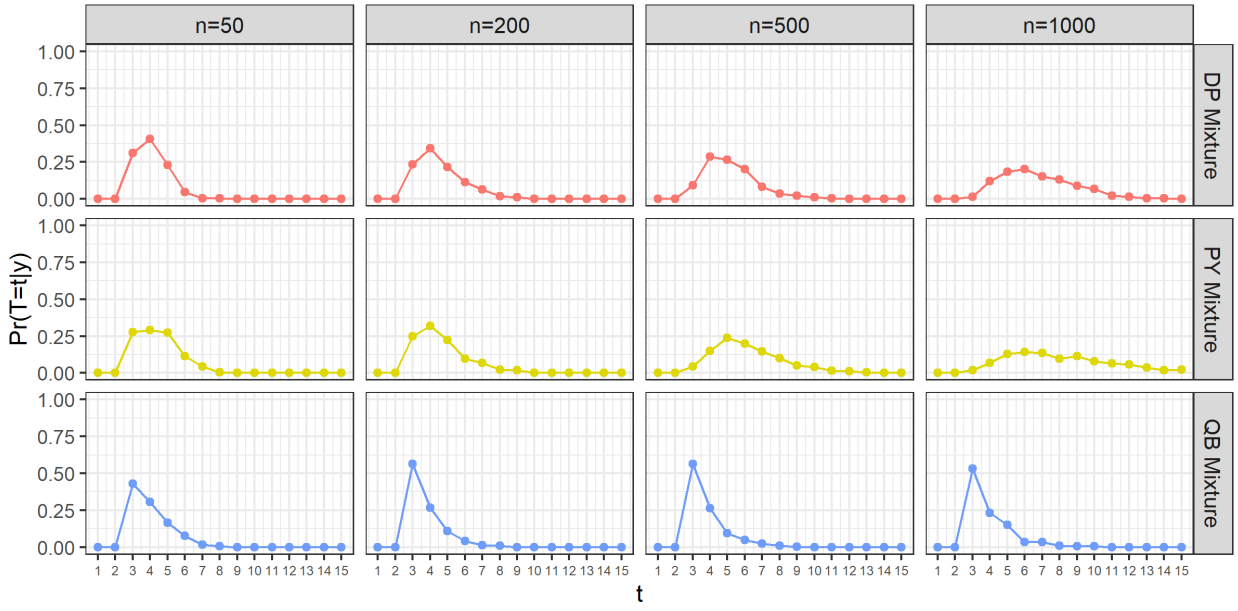


Figure 2-4. Quasi-Bernoulli mixture model correctly recovers three clusters as the ground truth when each component is from a Laplace distribution. The Dirichlet process and Pitman–Yor process overestimate the number of clusters, due to having small spurious clusters.

2.6 Data Application: Clustering Brain Networks

To demonstrate the ease of using our model in an advanced data analysis, we apply it to cluster multiple brain networks, collected from $n = 812$ subjects in the human connectome project (Marcus et al., 2011).

For each subject, resting-state functional magnetic resonance imaging (fMRI) signals were collected from $R = 50$ regions of the brain, indexed by $r = 1, \ldots, R$. The data were processed and transformed into a connectivity graph on R vertices, represented as a symmetric binary adjacency matrix $Y^{(i)} \in \{0,1\}^{R \times R}$ such that $Y^{(i)} = Y^{(i)\mathrm{T}}$.

We model these adjacency matrices using a probit-Bernoulli mixture model in which each component distribution has a low-rank latent structure. The goal of this model is to cluster the networks into disjoint groups of similar subjects, and to find a meaningful low-dimensional representation of networks within each sub-group/cluster.

Given a matrix of probabilities $\theta \in [0,1]^{R \times R}$, we write $Y \sim \text{Bernoulli}(\theta)$ to denote that $Y_{rs} \sim \text{Bernoulli}(\theta_{rs})$ independently for $r, s \in \{1, \ldots, R\}$. In this notation, we use the following infinite mixture model with a quasi-Bernoulli stick-breaking prior on the mixture weights $w = (w_1, w_2, \ldots)$,

$$\begin{aligned} Y^{(i)} \mid c_i, \mu, M &\overset{indep}{\sim} \text{Bernoulli}(\Phi(\mu + M_{c_i})), \\ c_i \mid w &\overset{iid}{\sim} \text{Categorical}(w) \text{ for } i = 1, \ldots, n, \\ M_k &= Q_k \Lambda_k Q_k^{\mathrm{T}}, \quad \text{for } k \geq 1, \\ w &\sim \text{Quasi-Bernoulli}(\tilde{p}, \epsilon, \alpha), \end{aligned} \tag{2-7}$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard Gaussian distribution, applied element-wise, and the weights $w = (w_1, w_2, \ldots)$ are drawn from Equation 2-1 which is denoted by Quasi-Bernoulli$(\tilde{p}, \epsilon, \alpha)$. The model enforces symmetry for the binary matrices $Y^{(i)}$ by only modeling the lower triangle part of them. Here, μ is a scalar that is shared by all components, and each $M_k = Q_k \Lambda_k Q_k^{\mathrm{T}}$ is a component-specific matrix such that $\Lambda_k = \text{diag}(\lambda_{k,1}, \ldots, \lambda_{k,d})$, and Q_k belongs to the Stiefel manifold $\mathcal{V}^{d,R} := \{Q \in \mathbb{R}^{R \times d} : Q^{\mathrm{T}} Q = I_d\}$ (Hoff, 2009). The role of M_k is to provide a low-rank representation for mixture component k, and the μ is a nuisance parameter that captures departures from this assumed low-rank structure.

For the other priors, we assign $Q_k \sim \text{Uniform}(\mathcal{V}^{d,R})$ for $k = 1, 2, \ldots$ with $d = 2$, use a truncated Gaussian prior $\pi(\lambda_{k,l}) \propto \text{N}(\lambda_{k,l} \mid 0, 50)$ for $l = 1, \ldots, d$, and assign a Gaussian prior $\text{N}(0, 10^2)$ on μ, following Hoff (2009). For the quasi-Bernoulli prior on w, we use $\tilde{p} = 0.9$, $\alpha = 1$

and $\epsilon = 1/n^{2.1}$. We run the MCMC sampler from Section 2.4 for 30,000 iterations and discard the first 10,000 as burn-ins.

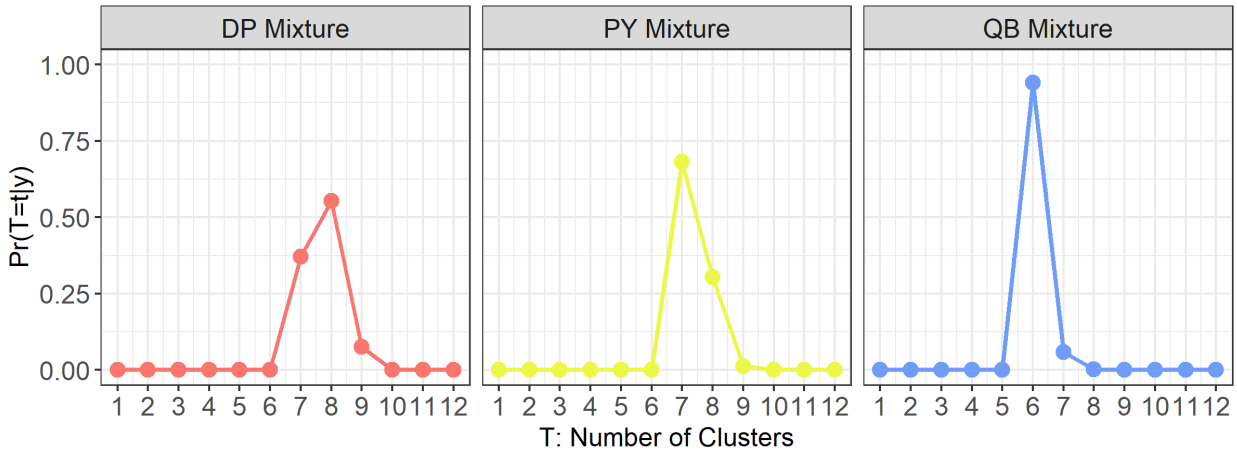
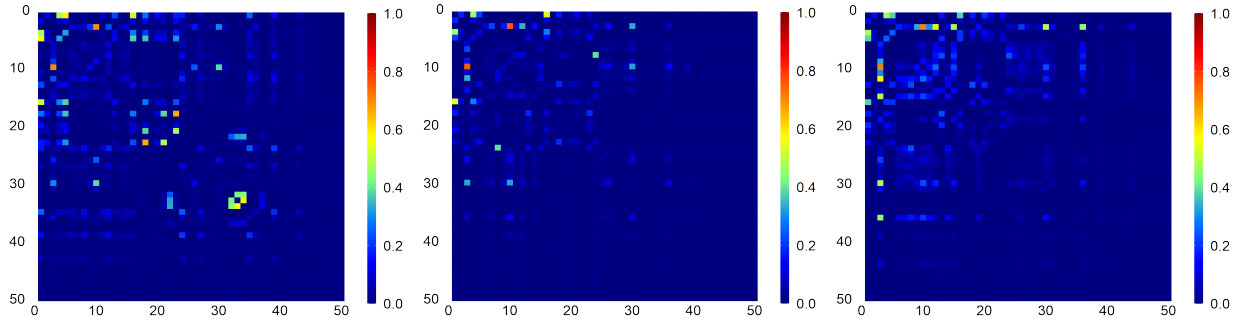
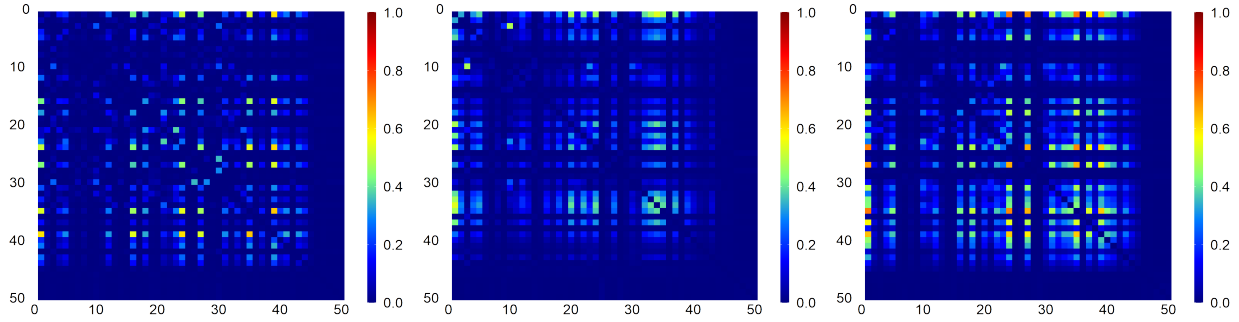


Figure 2-5. The quasi-Bernoulli model concentrates on $T = 6$ clusters on the brain connectivity data. For comparison, the posterior mode of the corresponding Dirichlet process and Pitman–Yor process mixture models are at $T = 8$ and $T = 7$.

Figure 2-5 (right) shows the posterior of the number of clusters T for the quasi-Bernoulli mixture model, which is highly concentrated on $T = 6$ clusters. For comparison, we also consider a Dirichlet process mixture and a Pitman–Yor process mixture. We use the same prior for the component parameters Q_k, Λ_k, as the one in the quasi-Bernoulli model. We set the Dirichlet process concentration parameter α so that the expected number of clusters under the Dirichlet process prior is as close as possible to the one under the quasi-Bernoulli process prior for $n = 812$. Similarly, for the Pitman–Yor process prior $\mathrm{PY}(\alpha, d)$, we choose α and d to match both the expectation and the variance of the number of clusters with the quasi-Bernoulli process prior. The α for the Dirichlet process is chosen to be 0.63, and the parameters of the Pitman–Yor process are $\alpha = 0.30$, $d = 0.11$. As the results, these two models yield posterior modes of $T = 8$ and $T = 7$ clusters (Figure 2-5, left and middle) and produce several very small clusters (the Dirichlet process mixture model produces four small groups with 5.8%, 5.5%, 5.2% and 0.1% of subjects; the Pitman–Yor process mixture model produces three small groups with 3.3%, 2.5% and 1.7% of subjects—these proportions are calculated as the average cluster sizes divided by n, with average



(a) Group 1 (46.0% of the subjects) (b) Group 2 (18.9% of the subjects) (c) Group 3 (12.9% of the subjects)



(d) Group 4 (9.0% of the subjects) (e) Group 5 (7.9% of the subjects) (f) Group 6 (5.2% of the subjects)

Figure 2-6. The posterior means of the edge connectivity probabilities $\Phi(\mu + M_k)$ over the six groups.

taken over those posterior samples with T equal to the posterior mode of T). The result from the quasi-Bernoulli model leads to a more parsimonious representation.

In the quasi-Bernoulli posterior, the subjects are clustered into six groups with high probability. Figure 2-6 shows the posterior means of the edge probabilities $\Phi(\mu + M_k)$ for each group. These results indicate that sparse connectivity is exhibited by the first three groups of subjects (accounting for 77.8% of subjects), whereas the other three groups have denser connectivity. Specifically, for each group, we examine the posterior mean proportion of node pairs with edge connectivity probabilities greater than 0.05. The proportions for the six groups are 8.6%, 4.4%, 8.6%, 15.7%, 21.1% and 28.2%, respectively.

In addition, we conduct additional experiments using two common Bayesian clustering methods: (i) Dirichlet process mixture of high-dimensional probit-Bernoulli model (without low-rank latent structure); (ii) approximation using the mixture of factor analyzers (McLachlan

et al., 2003) and treating the data as if they were continuous, and selecting the number of clusters using Bayesian information criterion (BIC). The method (i) is a popular solution, however, it is found to suffer from a curse of dimensionality (Chandra et al., 2023). Indeed, applying (i) on the data leads to only $T = 1$ cluster in the result. The method (ii) selects 3 clusters and 2 latent dimensions under BIC. The three clusters of the maximum a posteriori (MAP) estimates have 36.6%, 20.6% and 42.8% of the subjects. In the appendix, Figure A-4 shows the MAP estimate of the mean of each Gaussian component.

CHAPTER 3
BRIDGED POSTERIOR: OPTIMIZATION, PROFILE LIKELIHOOD AND A NEW APPROACH TO GENERALIZED BAYES

In this chapter, we present a new generalized Bayes approach. Rather than treating θ as a high-dimensional random variable, we model $\theta = (z, \lambda)$ with only λ as a parameter with a corresponding prior distribution. The argument z is instead treated as a latent variable that is deterministic conditional on y and λ, though importantly it remains a stochastic quantity when conditioned on y alone. As we will demonstrate in the article, this effectively reduces the dimension of θ to near that of λ, simultaneously addressing both issues surveyed above. Specifying z as the solution of an optimization subproblem allows us to retain transparent constraint conditions such as low rank, low cardinality, or combinatorial constraints.

It is natural to ask whether such an approach is consistent with Bayesian methodology, that there exists a valid generative model corresponding to a likelihood that depends only on λ. This article answers this question affirmatively. We begin with a set of profile likelihoods that partially maximize a joint likelihood $L(y; z, \lambda)$ over z, showing that each corresponds to another common likelihood where the data are modeled dependently. We then establish the theoretical result that under mild conditions, the $\sqrt{n}$-adjusted posterior distribution of the parameter under our framework converges asymptotically to the same normal limit as canonical posteriors marginalized over non-deterministic latent variables. This contribution is closely related to prior work by Polson and Scott (2016), which discovers a hierarchical duality: the scale mixture of univariate exponential or location-scale mixture of normal is proportional to another (potentially intractable) density maximized over a univariate latent variable. This perspective inspires efficient new algorithms for producing point estimates. Despite some similarities in the univariate setting, our method applies generally to multivariate problems and to settings where the latent variables may exhibit dependence. In other related work, Lee et al. (2005a) interpret the profile likelihood as resulting from an empirical prior. A key difference is that our proposed framework $L(y; z, \lambda)$

Dr. Eleni Dilma is the joint first author of this work.

can lead to a fully Bayesian method, where the latent variable z characterizes the latent dependency among the data.

3.1 Motivation

Using the Gibbs posterior introduced in Section 1.4, the point estimate $\hat{\theta} = \arg\min_\theta g(\theta, y)$ can often be efficiently computed using an iterative optimization algorithm, even under a wide range of constraints. For example, convex clustering and its variants (Tan and Witten, 2015; Chi and Lange, 2015; Chakraborty and Xu, 2023) use $g(\theta, y) = (1/2)\sum_{i=1}^n \|y_i - \theta_i\|_2^2 + \lambda \sum_{(i,j):i<j} \|\theta_i - \theta_j\|_2$ for data $y_i \in \mathbb{R}^p$, location parameter $\theta_i \in \mathbb{R}^p$, and tuning constant $\lambda > 0$. This can be understood as a relaxation of hierarchical clustering; in place of a combinatorial constraint, the penalty term encourages most of the L_2-norms $\|\hat{\theta}_i - \hat{\theta}_j\|_2$ to be zero, promoting cluster structure via a small number of unique $\hat{\theta}_i$ at the solution. The estimate $\hat{\theta}$ can be obtained using convex, continuous optimization. A popular combinatorial alternative makes use of the k-means loss (MacQueen, 1967) toward clustering, $g\{(c_1, \ldots, c_n), y\} = \sum_{k=1}^K \sum_{(i<j):c_i=c_j=k} \|y_i - y_j\|_2^2/n_k$, with $c_i \in \{1, \ldots, K\}$ where the discrete cluster assignment label $c_i = k$ if θ_k is the nearest centroid to y_i, $n_k = \sum_i 1(c_i = k)$, and $\hat{\theta}_i = \sum_{i:c_i=k} y_i/n_k$. Here too, iterative algorithms can improve performance and avoid local minima using continuous optimization techniques (Xu and Lange, 2019). Recently, Rigon et al. (2023a) form a Gibbs posterior using this loss toward quantifying the uncertainty of c_i, which shows robustness to the distributional asymmetry. Several recent works import ideas from optimization to account for constraints within a Bayesian framework (Duan et al., 2020; Presman and Xu, 2023; Zhou et al., 2024).

Under the exponential negative transformation, the Gibbs posterior distribution concentrates near the posterior mode. This induces variability around the point estimate and, in turn, enables uncertainty quantification. How to interpret this uncertainty is not immediately obvious, and one may question the authenticity of inferential procedures such as hypotheses tests or intervals based on such a posterior which may not derive from a generative likelihood model. There are several works that provide justification in the large n regime. First, Gibbs posteriors admit a coherent

update scheme for θ toward minimizing the expected loss $\int g(\theta, y)\mathcal{F}(\mathrm{d}y; \theta)$, where $\mathcal{F}$ denotes the true data generating distribution (Bissiri et al., 2016). Second, if the Gibbs posterior density is proportional to a composite likelihood, such as the conditional density under some insufficient statistics (Lewis et al., 2021) derived from a full likelihood $F(y; \theta_0)$, then the Gibbs posterior of θ concentrates toward θ_0 and enjoys asymptotic normality under mild conditions (Miller, 2021).

These methodological and theoretical breakthroughs lend a cautious optimism that loss functions from the machine learning and optimization literature have the potential to broaden the scope of Bayesian probabilistic modeling (Khare et al., 2015; Kim and Gao, 2020; Ghosh et al., 2021; Martin and Syring, 2022; Syring and Martin, 2023; Winter et al., 2023). At the same time, two pitfalls of Gibbs posteriors motivate this article. The first is computational: the Gibbs posterior is often supported on a high-dimensional space, and fails to reduce the computational burden that often plagues posterior sampling schemes such as Markov chain Monte Carlo (MCMC) in high-dimensional problems. There is a large literature characterizing the scaling limit of MCMC algorithms, which can lead to slow mixing of Markov chains as the dimension of θ increases (Roberts and Rosenthal, 2001; Belloni and Chernozhukov, 2009; Johndrow et al., 2019; Yang et al., 2020b). Meanwhile, many semi-parametric models feature a low-dimensional θ as well as a latent variable whose dimension grows with n. When closed-form marginals are not available, the necessity of sampling these latent variables can also lead to critically slow mixing. These issues have been observed in popular statistical methods such as latent normal models, and have motivated a large class of approximation methods (Rue et al., 2009) as alternatives to MCMC. This bottleneck explains in part the lack of Gibbs posterior approaches in latent variable contexts.

The second methodological gap relates to the modeling front: continuity of the Gibbs posterior distribution often yields a mismatch to constraint conditions on $\hat{\theta}$ except on a set of measure zero. To illustrate, consider the Bayesian lasso (Park and Casella, 2008), which can be viewed as the Gibbs posterior using the lasso loss. Though this promotes a sparse estimate $\hat{\theta}$, under its posterior distribution θ is non-sparse almost everywhere. A similar problem arises in a Gibbs posterior approach to support vector machines. The maximum-margin hyperplane has zero

posterior measure, which may partially explain why studies from this view have focused on point estimation (Polson and Scott, 2011), and motivates our approach in seeking a more natural quantification of the associated uncertainty. Beyond this incongruence between $\hat{\theta}$ and the samples from $\Pi(\theta \mid y)$, invariance to changes in $g(\theta, y)$ presents another consideration. To obtain an estimate $\hat{\theta}$ residing in a constrained or low dimensional space, it is common practice in optimization to employ an alternative $\tilde{g}(\theta, y)$ that has superior computational properties. For example, $\tilde{g}$ can be convex, unconstrained, or non-combinatorial, under the condition that $\arg\min_\theta \tilde{g}(\theta, y) = \arg\min_\theta g(\theta, y)$—that is, the two distinct loss functions touch at the minima. This invariance at the optimum is routinely exploited in methods such as convex relaxation, variable splitting, proximal methods, and majorization-minimization (Polson et al., 2015; Zheng and Aravkin, 2020; Landeros et al., 2023). However, the Gibbs posterior does not enjoy such an invariance, as the distribution $\Pi(\theta \mid y)$ changes whenever g changes. These issues lead us to take a marked departure from existing approaches.

3.2 Method

In this section, we formally construct the bridged posterior and show the posterior propriety. We provide several examples to illustrate the advantage of our proposed method.

3.2.1 Augmented Likelihood with Conditional Optimization

To provide background, we first review the canonical likelihood involving *latent variables*, taking the form

$$L(y; \lambda) = \int L(y, \mathrm{d}z; \lambda) = \int L(y \mid z, \lambda) \Pi_{\mathcal{L}}(\mathrm{d}z; \lambda), \tag{3-1}$$

in which we refer to $\lambda \in \mathbb{R}^d$ as the parameter, and $z \in \mathbb{R}^p$ as the latent variable. Here $\Pi_{\mathcal{L}}$ denotes the marginal latent variable distribution for z. Since z could be associated with a continuous, discrete, or degenerate distribution, we use the integration with respect to a probability measure notation $\int f(z)\mu(\mathrm{d}z)$, in which z with distribution $z \sim \mu$ is the one that we integrate over. The joint distribution $L(y, z; \lambda)$ is also known as an augmented likelihood (Tanner and Wong, 1987; Van Dyk and Meng, 2001). Examples abound in statistics: for instance, augmented likelihoods are used in characterizing dependence among discrete y via a correlated normal latent variable z

(Wolfinger, 1993; Rue et al., 2009), or model-based clustering on grouping data y via a latent discrete label z (Blei et al., 2003; Fraley and Raftery, 2002). We now consider a special case when given y and λ:

$$(z \mid y, \lambda) = \hat{z}(y, \lambda) := \arg\min_{\zeta} g(\zeta, y; \lambda) \text{ with probability 1.} \tag{3-2}$$

If the arg min is unique, then z is a *conditionally deterministic* latent variable, which we abbreviate CDLV. Otherwise, z has a conditional distribution supported on the solution set $\{\arg\min_{\zeta} g(\zeta, y; \lambda)\}$.

For simplicity of exposition, from here we focus on the case where z is the *unique* minimizer. This encompasses a large class of models and is satisfied whenever $g(\zeta, y; \lambda)$ is strictly convex in ζ for every (y, λ). Though z is conditionally deterministic form, note that when we do not condition on y, z remains randomly distributed under $\Pi_{\mathcal{L}}(z; \lambda)$. This suggests a generative view according to (3-1): we have

$$z \sim \Pi_{\mathcal{L}}(z; \lambda); \qquad y \mid z, \lambda \sim L(y \in \mathcal{Y}_{\lambda,z} \mid z, \lambda),$$
$$\text{where } \mathcal{Y}_{\lambda,z} = \left\{y : \min_{\zeta} g(\zeta, y; \lambda) = g(z, y; \lambda)\right\}.$$

That is, y is generated under the constraint given by z. To clarify, the latent z may have varying dimension p and $\Pi_{\mathcal{L}}$ according to the sample size n.

For concreteness, we present two illustrative examples based on the profile likelihood. Profile likelihoods have a frequentist origin, motivated by the convenience of testing or constructing confidence intervals for a parameter of interest λ, in the presence of other parameters ζ, often called nuisance parameters. There is a long-standing debate on whether using a profile likelihood leads to a coherent Bayes' procedure (Lee et al., 2005a; Cheng and Kosorok, 2009; Evans, 2016; Maclaren, 2018). Using the above, we can now view the profile likelihood as a special case of (3-2), taking $g(\zeta, y; \lambda) = -\log L(y, \zeta; \lambda)$.

Example 3.1. Consider linear regression with $y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n\times d}$, $\lambda \in \mathbb{R}^d$, $z > 0$, $v > 0$:

$$L(y, z; \lambda) \propto z^{-n/2} \exp\left(-\frac{\|y - X\lambda\|^2}{2z}\right) z^{-v/2-1} \exp\left(-\frac{v}{2z}\right);$$
$$z = \arg\min_\zeta\{-\log L(y, \zeta; \lambda)\}.$$

The first line has the same form as a likelihood with normal errors, with the variance z regularized by an Inverse-Gamma$(v/2, v/2)$. Instead of marginalizing out z, we maximize $\log L(y, \zeta; \lambda)$ over ζ to obtain $z = (v + \|y - X\lambda\|^2)/(v + n + 2)$. Therefore, we have

$$\Pi_{\mathcal{L}}(z; \lambda) \propto z^{-(n+v+2)/2};$$
$$L(y \mid z, \lambda) \propto \exp\left(-\frac{\|y - X\lambda\|^2}{2z}\right) \mathbb{1}\left\{\sum_{i=1}^n \left(y_i - x_i^\top \lambda\right)^2 = (v + n + 2)z - v\right\}.$$

In particular, the indicator above imposes a quadratic equality constraint on y. Upon substituting an expression in y for z, we obtain a marginal density

$$L(y; \lambda) \propto \left\{1 + \frac{\|y - X\lambda\|^2}{(v + 2)\frac{v}{v+2}}\right\}^{-(n+v+2)/2},$$

which coincides with the likelihood $L(y; \lambda)$ under an n-variate t-distribution with $v + 2$ degrees of freedom, center at $(X\lambda)$, and covariance $\{v/(v + 2)\}I$.

Example 3.2. Consider a multivariate factor model with $y = Cz + \epsilon \in \mathbb{R}^{\tilde{p}}$, $\epsilon \sim \mathrm{N}(0, I\sigma^2)$, $C \in \mathbb{R}^{\tilde{p}\times p}$. Here let $\tilde{p} \geq p$, the matrix C have rank p, $\lambda = (G, \sigma^2)$, and G positive definite:

$$L(y, z; \lambda) \propto \exp\left(-\frac{\|y - Cz\|^2}{2\sigma^2}\right) \exp\left(-\frac{1}{2} z^\top G^{-1} z\right); \quad z = \arg\min_\zeta\{-\log L(y, \zeta; \lambda)\}.$$

The first part has the same form as a likelihood with z regularized by multivariate normal distribution $\mathrm{N}(0, G)$. Here, minimization yields

$z = (C^\top C/\sigma^2 + G^{-1})^{-1} C^\top y/\sigma^2 = \{G - GC^\top (I\sigma^2 + CGC^\top)^{-1} CG\} C^\top y/\sigma^2$. Therefore,

$$\Pi_{\mathcal{L}}(z;\lambda) \propto \exp\left[-\frac{1}{2} z^\top \{GC^\top (I\sigma^2 + CGC^\top)^{-1} CG\}^{-1} z\right];$$
$$L(y \mid z, \lambda) \propto \exp\left(-\frac{\|y - Cz\|^2}{2\sigma^2}\right) \mathbb{1}\{C^\top y - (C^\top C + \sigma^2 G^{-1}) z = 0\}.$$

The indicator puts an affine equality constraint on y, and the marginal of y is

$$L(y;\lambda) \propto \exp\left\{ -\frac{1}{2} y^\top \left(I\sigma^2 + CGC^\top\right)^{-1} y \right\},$$

corresponding to a multivariate normal $\mathrm{N}\left(0, I\sigma^2 + CGC^\top\right)$.

From the above two examples, we highlight two observations: (i) partial optimization still leads to a valid probability kernel $L(y;\lambda)$ associated with a coherent generative model for y; (ii) fixing z at the conditional optimum induces dependency among the elements in y in $L(y;\lambda)$, via the constraint $\mathcal{Y}_{\lambda,z}$.

Remark 3.1. The profile likelihood-based models are an important sub-class that we will primarily focus on. Nevertheless, in general, the loss function g does not have to be the negative likelihood, and z does not have to be available in closed form. We can still specify the joint likelihood, by including an optimization problem in the equality constraint (3-2).

3.2.2 Bridged Posterior Distributions and Posterior Propriety

We now take a Bayesian approach by assigning a suitable prior distribution on λ. Denoting this prior by $\pi_0(\lambda)$, Bayes theorem provides the posterior

$$\Pi(\lambda \mid y) = \frac{\int L(y, \mathrm{d}z; \lambda)\pi_0(\lambda)}{\int\int L(y, \mathrm{d}z; \lambda)\pi_0(\mathrm{d}\lambda)}, \quad \text{subject to } z = \arg\min_\zeta g(\zeta, y; \lambda). \tag{3-3}$$

When z is the unique minimizer, we may remove the first integration from both the numerator and denominator, replacing $\mathrm{d}z$ by z. The above distribution can be viewed as obeying an *equality constraint*, which acts as a bridge between a probabilistic model and an optimization problem. Therefore, we refer to (3-3) as a bridged posterior. To clarify, the above formulation encompasses

the setting of $\lambda = (\lambda_A, \lambda_B)$, where only the first part λ_A influences the minimization of g, and λ_B corresponds to the other parameters.

For such a posterior, it is important to ensure the propriety of $\Pi(\lambda \mid y)$. To be rigorous, we note that the derived $L(y, z; \lambda)$ may be a kernel function *proportional* to a complete density of y up to some missing normalizing constant. Thus it suffices to choose an appropriate $\pi_0(\lambda)$ ensuring that $\int \Pi(\mathrm{d}\lambda \mid y) < \infty$.

A challenge arises when checking integrability (such as when verifying posterior propriety), when $L(y, z; \lambda)$ is intractable due to the lack of a closed-form solution z. Generally speaking, mathematically verifying integrability may vary from case to case; we develop a useful strategy in the case when $L(y, z; \lambda)$ is a profile likelihood. Consider the following

$$L(y, z; \lambda) = \exp\{-h(y, \lambda)\} \exp\{-\min_{\zeta} g(\zeta, y; \lambda)\}, \tag{3-4}$$

where $z = \arg\min_\zeta g(\zeta, y; \lambda)$. In the optimization literature, $\min_\zeta g(\zeta, y; \lambda)$ is often referred to as the *primal problem*, and z as the primal solution. A useful technique is to instead solve an associated dual problem $\sup_\alpha g^\dagger(\alpha, y; \lambda)$, where $\alpha \in \mathbb{R}^q$ is the dual variable. For example, the Fenchel dual for convex g is based on the conjugate function $g^\dagger(\alpha, y; \lambda) := \sup_\zeta\{\alpha^\top \zeta - g(\zeta, y; \lambda)\}$, and the Lagrangian dual for α under constraints $\tilde{c}(\alpha) \leq \vec{0}$, where $\tilde{c}(\alpha) \in \mathbb{R}^q$ and the inequality holds pointwise, is $g^\dagger(\alpha, y; \lambda) := \inf_{\zeta \in \mathbb{R}^p} g(\zeta, y; \lambda) + \alpha^\top \tilde{c}(\zeta)$, where the dual variable $\alpha \geq \vec{0}$. In many cases, the dual problem may be easier to solve, and with helpful techniques such as variable splitting (discussed further below in the context of latent normal models), one often obtains $g^\dagger(\alpha, y; \lambda)$ in a closed form, even when $\sup_\alpha g^\dagger(\alpha, y; \lambda)$ may be intractable. Duality theory provides a simple way to produce a useful bound:

Theorem 3-1. *For a likelihood* (3-4)*, consider* $\inf_\zeta g(\zeta, y; \lambda)$ *as the primal problem, and* $\sup_\alpha g^\dagger(\alpha, y; \lambda)$ *as the dual problem with* E *the feasible region of* α*. Assume that weak duality holds:* $\sup_\alpha g^\dagger(\alpha, y; \lambda) \leq \inf_\zeta g(\zeta, y; \lambda)$ *for any* λ *in the domain, given* y*. If there exists* $\tilde{\alpha} \in E$ *such that* $\int \exp[-h(y, \lambda)] \exp[-g^\dagger(\tilde{\alpha}, y; \lambda)] \pi_0(\mathrm{d}\lambda) < \infty$*, then* $\int \Pi(\mathrm{d}\lambda \mid y) < \infty$.

Remark 3.2. The proof follows rather trivially from the fact that $g^\dagger(\tilde{\alpha}, y; \lambda) \leq \sup_\alpha g^\dagger(\alpha, y; \lambda)$. On the other hand, this result leads to a very useful method for checking integrability—we do not have to solve for the optimal dual variable $\hat{\alpha}$ at which $\sup g^\dagger(\alpha, y; \lambda)$ is attained. Instead, we just need to find any $\tilde{\alpha} \in E$ that makes the product integrable. Moreover, the criteria for weak duality are straightforward to check: for Fenchel duals, g needs to be convex, and for Lagrangian duals, g can be convex or non-convex. We now illustrate the application of the theorem via a working example.

Example 3.3 (Latent normal model and latent quadratic exponential model)**.** We modify the canonical latent normal model that uses a full likelihood:

$$\tilde{L}(y, \zeta; \lambda) \propto \exp\left\{-\frac{1}{2}\zeta^\top Q^{-1}(\lambda; x)\zeta\right\} \prod_{i=1}^{n} v(y_i \mid \zeta_i), \tag{3-5}$$

where v is commonly a log-concave density of y_i conditionally independent for $i = 1, \ldots, n$, $Q(\lambda; x)$ is parameterized by a covariance kernel such as $Q(\lambda; x)_{i,j} = \tau \exp(-\|x_i - x_j\|^2/2b)$ with $x_i \in \mathbb{R}^{\tilde{d}}$ the observed predictor/location, and parameter $\lambda = (\tau, b) \in \mathbb{R}^2$. In our example, we focus on binary y_i from Bernoulli distribution under logistic link $v(y_i \mid \zeta_i) = \exp(y_i\zeta_i)/\{1 + \exp(\zeta_i)\}$. We now minimize $g(\zeta, y; \lambda) = -\log \tilde{L}(y, \zeta; \lambda)$ over $\zeta \in \mathbb{R}^n$ to induce a conditionally deterministic z, with profile likelihood:

$$L(y, z; \lambda) \propto \exp\left\{-\frac{1}{2}z^\top Q^{-1}(\lambda; x)z\right\}\left\{\prod_{i=1}^{n} v(y_i \mid z_i)\right\}, \quad z = \arg\min_\zeta\{-\log L(y, \zeta; \lambda)\}. \tag{3-6}$$

As $g(\zeta, y; \lambda)$ can be conveniently decomposed into the sum of a quadratic function and a convex function, this lends itself to variable splitting using the constraint $u = \zeta$. With $\alpha \in \mathbb{R}^n$ the Lagrange multiplier, we have the Langragian dual

$$g^\dagger(\alpha, y; \lambda) = \inf_{\zeta, u} \frac{1}{2}\zeta^\top Q^{-1}\zeta + \alpha^\top(\zeta - u) + \sum_{i=1}^{n}\left\{-y_i u_i + \log(1 + \exp(u_i))\right\},$$

where we use $Q = Q(\lambda; x)$ to ease notation. This leads to

$$\hat{\zeta} = -Q\alpha, \qquad \hat{u}_i = \log \frac{\alpha_i + y_i}{1 - (\alpha_i + y_i)} \text{ for } i = 1, \ldots, n,$$

whenever $(\alpha + y) \in (0, 1)^n$; otherwise the infimum takes $-\infty$. We have the dual function:

$$g^\dagger(\alpha, y; \lambda) = -\frac{1}{2}\alpha^\top Q\alpha - \sum_{i=1}^{n}\left\{(a_i + y_i)\log\frac{\alpha_i + y_i}{1 - (\alpha_i + y_i)} - \log\frac{1}{1 - (\alpha_i + y_i)}\right\},$$

subject to $(\alpha + y) \in (0, 1)^n$.

At a given α satisfying $(\alpha + y) \in (0, 1)^n$, we have

$$\exp\{-g^\dagger(\alpha, y; \lambda)\} = \exp\left\{\frac{1}{2}\alpha^\top Q(\lambda; x)\alpha\right\}\prod_{i=1}^{n}\left[(\alpha_i + y_i)^{a_i + y_i}\{1 - (\alpha_i + y_i)\}^{1-(a_i+y_i)}\right].$$

Due to some similarity between the above form and the quadratic exponential model in McCullagh (1994), we refer to (3-6) as a latent quadratic exponential model. It is not hard to see that the above is an integrable upper bound for y, since $y_i \in \{0, 1\}$, $(\alpha_i + y_i)$ and $1 - (\alpha_i + y_i)$ are both bounded above. To find an appropriate prior for λ, the second part does not involve λ at any fixed α. It suffices to find a prior such that for a feasible $\tilde{\alpha}$:

$$\int \exp\left\{\frac{1}{2}\tilde{\alpha}^\top Q(\lambda; x)\tilde{\alpha}\right\}\pi_0(\mathrm{d}\lambda) < \infty.$$

Using $Q(\lambda; x)_{i,j} = \tau \exp(-\|x_i - x_j\|^2/2b)$, the matrix spectral norm $\|Q(\lambda; x)\|_2 \le n\tau$. As we may choose any feasible $\tilde{\alpha}$, we take $\tilde{\alpha}_i = -(1/n)\mathbb{1}(y_i = 1) + (1/n)\mathbb{1}(y_i = 0)$. Since $\tilde{\alpha}^\top Q(\lambda; x)\tilde{\alpha} \le \|\tilde{\alpha}\|_2^2\|Q(\lambda; x)\|_2 = \tau$, it suffices to assign a half-normal prior for τ proportional to $\exp(-c_1\tau^2)$ with $c_1 > 0$ and with any proper prior on $b > 0$.

Remark 3.3. For the above example, strong duality holds: $\sup_\alpha g^\dagger(\alpha, y; \lambda) = \inf_\zeta g(\zeta, y; \lambda)$. Therefore, we can use the dual ascent algorithm to find $\hat{\alpha} = \arg\max_{\alpha:(\alpha+y)\in(0,1)^n} g^\dagger(\alpha, y; \lambda)$, and then set $z = -Q\hat{\alpha}$. Note that neither the dual function (3.3) nor its gradient with respect to α

requires the inversion Q^{-1}, an $O(n^3)$ operation, so that optimization can be carried out very efficiently. At the same time, $L(y, z; \lambda)$ can be evaluated quickly since $z^\top Q^{-1} z = \hat{\alpha}^\top Q \hat{\alpha}$. In contrast, the latent normal model would involve matrix inversion and decomposition for sampling latent ζ. We defer the numerical experiments to Section 3.5.

3.2.3 Predictive Distribution

In addition to parameter estimation, one may be interested in making predictions on data $y_{(n+1):(n+k)}$ and quantifying their uncertainty, using the following distribution:

$$\begin{aligned}\Pi\{y_{(n+1):(n+k)} \mid y_{1:n}\} &\propto \int L\{y_{(n+1):(n+k)} \mid y_{1:n}, \lambda\}\Pi(\mathrm{d}\lambda \mid y_{1:n}) \\ &\propto \int \frac{L\{y_{1:(n+k)}, \hat{z}(y_{1:(n+k)}, \lambda), \lambda\}}{L\{y_{1:n}, \hat{z}(y_{1:n}, \lambda), \lambda\}}\Pi(\mathrm{d}\lambda \mid y_{1:n}),\end{aligned}$$

for which we could take each posterior sample of λ, and simulate a vector $y_{(n+1):(n+k)}$ with kernel proportional to $L\{y_{1:(n+k)}, \hat{z}(y_{1:(n+k)}, \lambda), \lambda\}$. When we lack a way to directly draw from the joint distribution of $y_{(n+1):(n+k)}$, note that

$$\frac{L\{y_{1:(n+k)}, \hat{z}(y_{1:(n+k)}, \lambda), \lambda\}}{L\{y_{1:n}, \hat{z}(y_{1:n}, \lambda), \lambda\}} = \prod_{j=1}^{k} \frac{L\{y_{1:(n+j)}, \hat{z}(y_{1:(n+j)}, \lambda), \lambda\}}{L\{y_{1:(n+j-1)}, \hat{z}(y_{1:(n+j-1)}, \lambda), \lambda\}},$$

suggesting that we can simulate y_{n+j} sequentially for $j = 1 \ldots k$. When y_{n+j} is in a low-dimensional (often one-dimensional) space, we can employ a simple algorithm such as rejection sampling. Note that *all* elements in $z = \hat{z}(y_{1:(n+k)}, \lambda) \in \mathbb{R}^p$ may vary according to $y_{1:(n+k)}$; we emphasize this by using the notation $\hat{z}(y_{1:(n+k)}, \lambda)$. This implies that when there is no closed-form solution for z, there is an additional burden to compute $\hat{z}(y_{1:(n+j)}, \lambda)$ wherever j increments to $j + 1$. Fortunately, for the problem $\hat{z}(y_{1:(n+j+1)}, \lambda) = \arg\min_\zeta g(\zeta, y_{1:(n+j+1)}; \lambda)$, we can initialize ζ at the last optimal when predicting y_{n+j}, $\hat{z}(y_{1:(n+j)}, \lambda)$, and it takes a few iterations of optimization steps to converge to $\hat{z}(y_{1:(n+j+1)}, \lambda)$. For advanced problems, there is a large literature on online optimization algorithms (Jadbabaie et al., 2015) that can be employed to efficiently obtain sequential updates.

For concreteness, we highlight a useful property of the above predictive distribution in the context of classification problems. We can find a hyperplane that not only divides the fully observed data with both predictors x_i and labels y_i for $i = 1, \ldots, n$, but also seeks to separate the observed unlabeled data $x_{i'}$ with corresponding (unobserved) label $y_{i'}$, $i' = n+1, \ldots, n+k$. This further improves the classification accuracy, and is often called the semi-supervised setting in the machine learning literature (Chapelle et al., 2010).

Example 3.4 (Bayesian Maximum Margin Classifier for Partially Labeled Data)**.** Consider the following likelihood that extends the support vector machine (Cortes and Vapnik, 1995), for n labeled data $(x_i, y_i) \in \mathbb{R}^{\tilde{p}} \times \{-1, 1\}$ and k unlabeled predictors $x_j \in \mathbb{R}^{\tilde{p}}$:

$$L\big[\{y_i\}_{i=1}^n, z = (z_w, z_b); \lambda, \{x_i\}_{i=1}^{n+k}\big] \propto \sum_{\{y_{n+j}\}_{j=1}^k \in \{-1,1\}^k} \exp\left\{ -\frac{1}{2}\lambda \|z_w\|_2^2 - \sum_{i=1}^{n+k} h(z, y_i; \lambda, x_i) \right\},$$
$$z = \arg\min_{\zeta=(\zeta_w,\zeta_b)} \frac{1}{2}\lambda\|\zeta_w\|_2^2 + \sum_{i=1}^{n+k} h(\zeta, y_i; \lambda, x_i),\ h(\zeta, y_i; \lambda, x_i) = \max\big\{1 - y_i(\zeta_w^\top x_i + \zeta_b), 0\big\}, \tag{3-7}$$

where $z_w \in \mathbb{R}^{\tilde{p}}$, $\zeta_w \in \mathbb{R}^{\tilde{p}}$ and $z_b \in \mathbb{R}$, $\zeta_b \in \mathbb{R}$. We treat x_i as fixed, so that the above likelihood is viewed as a discrete distribution for $(y_1, \ldots, y_{n+k})$. The function h is the hinge loss, which takes value zero when $y_i = 1$, $\zeta_w^\top x_i + \zeta_b \geq 1$, or when $y_i = -1$, $\zeta_w^\top x_i + \zeta_b \leq -1$. Effectively, the loss function penalizes not only the misclassified points $(x_i, y_i) : y_i(\zeta_w^\top x_i + \zeta_b) < 0$, but also the points in the *band* between two boundaries $\{x : -1 < \zeta_w^\top x_i + \zeta_b < 1\}$. The inclusion of $(1/2)\lambda\|z_w\|_2^2$ leads to a maximum distance between the two hyperplanes $\{x : z_w^\top x + z_b = 1\}$ and $\{x : z_w^\top x + z_b = -1\}$, under λ-adjusted tolerance to non-zero hinge losses.

For comparison, if we were to use a Gibbs posterior with likelihood of the form of (3-7)—without the equality constraint so that z is replaced by ζ, then it would hold that $p(y_i \mid \zeta, \lambda, x_i) \propto \exp\{-h(\zeta, y_i; \lambda, x_i)\}$ independently for $i = n+1, \ldots, n+k$. In particular, the distribution under the Gibbs posterior would yield $\tilde{L}\big[\{y_i\}_{i=1}^n, \zeta; \lambda, \{x_i\}_{i=1}^{n+k}\big] = \tilde{L}\big[\{y_i\}_{i=1}^n, \zeta; \lambda, \{x_i\}_{i=1}^n\big]$ via marginalization, which fails to incorporate any information from the observed $(x_{n+1}, \ldots, x_{n+k})$. See Liang et al. (2007) for a comprehensive

discussion. Now, under our bridged posterior approach, denote the conditional optimum

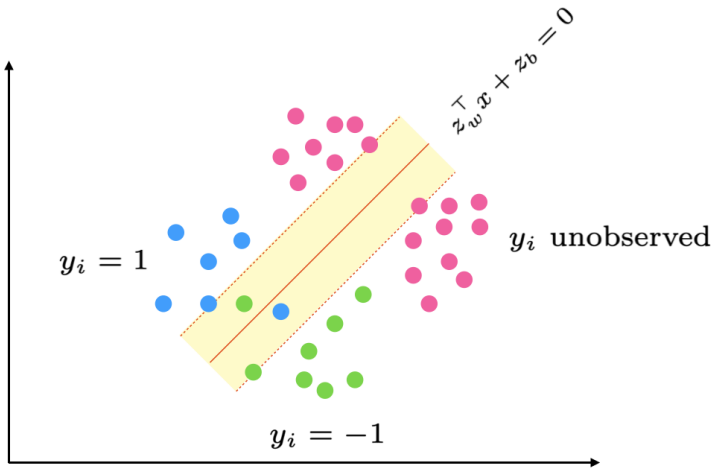


Figure 3-1. Intuition on how the Bayesian maximum margin classifier (a bridged posterior based on conditional minimization of the hinge loss) incorporates information from both the labeled (y_i observed) and unlabeled data (y_i unobserved). The posterior puts a high probability on a decision boundary with a small misclassification error among observed data (blue and green points), while trying to avoid having the decision band $\{x : -1 < \zeta_w^\top x + \zeta_b < 1\}$ cover the unlabeled predictors x_j (magenta points).

$z = \hat{z}\big(\{y_i\}_{i=1}^{n+k}, \lambda; \{x_i\}_{i=1}^{n+k}\big)$. Though a closed-form marginal for (3-7) is not available, we know from the Lagrangian dual with multiplier $\alpha \in \mathbb{R}^{n+k}$ (Chang and Lin, 2011) that the decision hyperplane $C_z = \{x : z_w^\top x + z_b = 0\}$ satisfies

$$z_w = \sum_{i=1}^{n+k} (\alpha_i y_i) x_i, \quad \text{where} \quad 0 \le \alpha_i \le \lambda^{-1},$$

and there are only a few $\alpha_i \neq 0$ for which $(\zeta_w^\top x_i + \zeta_b) y_i \le 1$—these are the so-called support vectors. Regardless of the values of $\{y_i\}_{i=n+1}^{n+k}$, the decision boundary can be influenced by the unlabeled predictor $\{x_i\}_{i=n+1}^{n+k}$. Intuitively, the bridged posterior assigns higher probability to a hyperplane C_z with a small misclassification error among observed data, while avoiding unlabeled predictors x_i in the band $\{-1 < \zeta_w^\top x + \zeta_b < 1\}$. We use Figure 3-1 to show the intuition and provide numerical results in the Section 3.5.2.

3.3 Posterior Computation

One appealing property of the bridged posterior is that the joint distribution $\Pi(\lambda, z \mid y)$ is supported on a low dimensional space relative to the ambient space, with intrinsic dimension determined by λ. This leads to efficient posterior estimation via MCMC algorithms.

3.3.1 Metropolis–Hastings with Conditional Optimization

We first focus on the case of $\lambda \in \mathbb{R}^d$ with a small d, which allows us to use simple MCMC algorithms such as Metropolis–Hastings for posterior sampling. At MCMC iteration t, we denote the posterior kernel as:

$$\Pi\{\lambda^{(t)} \mid y\} = mL\{y, z^{(t)}; \lambda^{(t)}\}\pi_0(\lambda^{(t)}), \qquad z^{(t)} = \hat{z}(y, \lambda^{(t)}) = \arg\min_{\zeta} g(\zeta, y; \lambda^{(t)}).$$

where m is the normalizing constant that does not involve $\lambda^{(t)}$ or $z^{(t)}$. We assume that π_0 has a closed form and L has a closed form as a function of $\{y^{(t)}, z^{(t)}, \lambda^{(t)}\}$, although $z^{(t)}$ may not have a closed form. This allows us to use a simple Metropolis–Hastings algorithm:

- Draw proposal $\lambda^* \sim G(\cdot; \lambda^{(t)})$
- Run optimization subroutine to find $z^* = \arg\min_{\zeta} g(\zeta, y; \lambda^*)$.
- Set $\lambda^{(t+1)} \leftarrow \lambda^*$, $z^{(t+1)} \leftarrow z^*$ with probability:

$$1 \wedge \frac{L(y, z^*; \lambda^*)\pi_0(\lambda^*)G(\lambda^{(t)}; \lambda^*)}{L\{y, z^{(t)}; \lambda^{(t)}\}\pi_0(\lambda^{(t)})G(\lambda^*; \lambda^{(t)})}.$$

  Otherwise, set $\lambda^{(t+1)} \leftarrow \lambda^{(t)}$, $z^{(t+1)} \leftarrow z^{(t)}$.

In this article, for algorithmic simplicity, we take λ as unconstrained in $\mathbb{R}^d$ under appropriate reparametrization (such as the softplus transformation for positive scalars $\tilde{\lambda}_1 = \log[1 + \exp(\lambda_1)] > 0$). We use $G(\cdot; \lambda^{(t)})$ as Uniform$(\lambda^{(t)} - s, \lambda^{(t)} + s)$, where $s \in \mathbb{R}^d_{\geq 0}$ is a tuning parameter that controls the step size in each dimension. When running MCMC for each of the examples presented, we make use of an adaptation period to tune s so that the empirical acceptance rate is close to 0.3, after which we fix s and collect Markov chain samples. This

exhibits excellent mixing performance empirically. For higher dimension d, one can use informative proposals such as the Metropolis-adjusted Langevin algorithm (MALA) or Hamiltonian Monte Carlo.

3.3.2 Diffusion-based Algorithms for Profile Likelihood-based Bridged Posterior

We first show that gradients (or sub-gradients) are readily available in cases when a profile likelihood is used, and then we discuss its use in the MALA algorithm. Under the bridged posterior, the lack of closed forms for z presents a potential challenge to these methods, leading to intractable gradients or subgradients with respect to z. However, for those based on the profile likelihood, this issue can be bypassed entirely. Consider the posterior deriving from (3-4),

$$\Pi(\lambda \mid y) \propto \pi_0(\lambda) \exp\{-h(y,\lambda)\} \exp\{-\min_{\zeta} g(\zeta, y; \lambda)\}, \quad z = \arg\min_{\zeta} g(\zeta, y; \lambda).$$

If $g(\zeta, y; \lambda)$, as an unconstrained function of three inputs, is differentiable in ζ and λ almost everywhere, then we have a very simple gradient expression provided $z = \arg\min_{\zeta} g(\zeta, y; \lambda)$ is differentiable with respect to λ:

$$\frac{\partial \min_{\zeta} g(\zeta, y; \lambda)}{\partial \lambda} = \left.\frac{\partial g(\zeta, y; \lambda)}{\partial \lambda}\right|_{\zeta = z}.$$

This is due to the envelope theorem.

When z may not be differentiable in λ but is strictly continuous in λ, the expression $\partial g(\zeta, y; \lambda)/\partial \lambda\big|_{\zeta=z}$ still holds as a subgradient of $\min_{\zeta} g(\zeta, y; \lambda)$ with respect to λ (Rockafellar and Wets, 2009, Theorem 10.49). For completeness, recall a subgradient of $f : \mathbb{R}^d \to \mathbb{R}$ at $x \in \mathbb{R}^d$ is a vector $v \in \mathbb{R}^d$ that satisfies $f(y) \geq f(x) + v^\top (y - x)$ for any y in the domain. In subgradient-based MCMC samplers (Tang and Yang, 2022), one typically refers to a local subgradient with inequality held for $y : \|y - x\| \leq \epsilon$ under a sufficiently small $\epsilon > 0$. When f is differentiable at x, there is a unique subgradient, coinciding with the usual gradient. We use $\tilde{\nabla} \log \Pi(\lambda \mid y)$ to denote a subgradient evaluated at point λ. For reversibility, in the case when there is more than one subgradient at λ, we impose a constraint that $\tilde{\nabla} \log \Pi(\lambda \mid y)$ is chosen as

one of the subgradients in a pre-determined way. This constraint is implicitly satisfied in most computing software, for example, most packages will output $\tilde{\nabla}|\lambda_1|_1 = 0$ when $\lambda_1 = 0$, even though any value $[-1, 1]$ is a subgradient. We now describe the MALA algorithm with preconditioning:

- Draw proposal $\lambda^* \sim \mathrm{N}\big[\cdot; \lambda^{(t)} + \tau M \tilde{\nabla} \log \Pi\{\lambda^{(t)} \mid y\}, 2\tau M\big]$.
- Run optimization algorithm to find $z^* = \arg\min_\zeta g(\zeta, y; \lambda^*)$.
- Set $\lambda^{(t+1)} \leftarrow \lambda^*$, $z^{(t+1)} \leftarrow z^*$ with probability:

$$1 \wedge \frac{L(y, z^*; \lambda^*)\pi_0(\lambda^*)\mathrm{N}\big\{\lambda^{(t)}; \lambda^* + \tau M \tilde{\nabla} \log \Pi(\lambda^* \mid y), 2\tau M\big\}}{L\{y, z^{(t)}; \lambda^{(t)}\}\pi_0(\lambda^{(t)})\mathrm{N}\big[\lambda^*; \lambda^{(t)} + \tau M \tilde{\nabla} \log \Pi\{\lambda^{(t)} \mid y\}, 2\tau M\big]}.$$

  Otherwise, set $\lambda^{(t+1)} \leftarrow \lambda^{(t)}$, $z^{(t+1)} \leftarrow z^{(t)}$.

In the above, $M \in \mathbb{R}^{d\times d}$ is positive definite and $\tau > 0$ is the step size.

3.4 Asymptotic Theory

Many Bayesian models satisfy a Bernstein-von Mises (BvM) theorem under suitable regularity conditions, that is the posterior distribution of $\sqrt{n}(\lambda - \lambda_n)$ where λ_n denotes the maximum likelihood estimator (MLE) converges to a normal distribution centered at 0, with covariance equal to the inverse Fisher information evaluated at λ_0, denoted by H_0^{-1}.

In a canonical Bayesian approach involving latent variable ζ (that is not conditionally determined), one could focus on the integrated posterior based on integrated likelihood (Berger et al., 1999; Severini, 2007), $\Pi(\lambda \mid y) \propto \big\{\int L(y, \mathrm{d}\zeta; \lambda)\big\}\pi_0(\lambda)$. For this integrated posterior, BvM results hold for $\sqrt{n}(\lambda - \lambda_n)$ under appropriate conditions, with asymptotic covariance H_0^{-1} (Bickel and Kleijn, 2012; Castillo and Rousseau, 2015).

Since z is now conditionally determined given (y, λ) under our bridged posterior, it may seem intuitive to expect that the posterior of λ would reflect a lower amount of uncertainty (such as having smaller marginal variances) compared to its integrated posterior counterpart. Surprisingly, we dispel this belief in the asymptotic regime—our result below proves that the bridged posterior of λ enjoys the same BvM result with covariance H_0^{-1}.

We establish sufficient conditions for BvM results under both parametric and semi-parametric cases. To be clear, the parametric setting commonly refers to when both λ and z have fixed dimensions, while the semi-parametric one does to when λ has a fixed dimension, but z has a dimension that could grow indefinitely (for instance, increasing with n). Therefore, the result developed under the semi-parametric setting can be easily extended to the parametric setting, under the same sufficient conditions while fixing the dimension of z.

In the following, we first focus on the BvM result for general bridged posterior which may or may not be based on a profile likelihood. Because we consider a broad family of distributions, we rely on relatively strong conditions here, such as differentiability of the likelihood in a parametric setting. Next, we relax the differentiability requirements and extend our scope to the semi-parametric setting. As this latter setting presents more challenging conditions, we will restrict our focus to the sub-class of bridged posteriors based on profile likelihoods in our treatment of the semi-parametric case. In both settings, we consider λ in the parameter space $\Theta \subset \mathbb{R}^d$ and that there is a fixed ground-truth λ_0, and the prior density $\pi_0(\lambda)$ to be continuous at λ_0 with $\pi_0(\lambda_0) > 0$. We use $\|\cdot\|$ as the Euclidean–Frobenius norm, and $B_r(\lambda_0) = \{\lambda \in \mathbb{R}^d : \|\lambda - \lambda_0\| < r\}$ as a ball of radius r.

3.4.1 General Bridged Posterior under Parametric Setting

For a real-valued function $\alpha(x)$ defined on $\mathbb{R}^d$, we denote first, second and third derivatives by $\alpha'(x) \in \mathbb{R}^d$, $\alpha''(x) \in \mathbb{R}^{d\times d}$ and $\alpha'''(x) \in \mathbb{R}^{d\times d\times d}$, respectively. For a vector-valued function $\alpha(x) = \{\alpha_1(x), \ldots, \alpha_m(x)\}$, we again use notations $\alpha'(x)$, $\alpha''(x)$ and $\alpha'''(x)$ to denote the derivatives, to be understood as tensors one order higher. We say a sequence of functions α_n uniformly bounded on E if the set $\{\|\alpha_n(x)\| : x \in E, n \in \mathbb{N}\}$ is bounded. We use $\xrightarrow[n\to\infty]{a.s.[y_{1:n}]}$ for almost sure convergence, and $\xrightarrow[n\to\infty]{P_{\lambda_0,\zeta_0}}$ for convergence in probability.

To ease the notation, we define

$$l_n(\lambda, \zeta) = \log L(y_{1:n}, \zeta; \lambda)/n, \qquad \hat{l}_n(\lambda) = l_n\{\lambda, \hat{z}_n(\lambda)\} = \log L\{y_{1:n}, \hat{z}_n(\lambda); \lambda\}/n,$$

where the CDLV $\hat{z}_n(\lambda) := \arg\min_\zeta g_n(\zeta, y_{1:n}; \lambda)$. Let E be an open and bounded subset of Θ such that $\lambda_0 \in E$. We first state and explain some assumptions.

(A1) The function l_n has continuous third derivatives on $E \times \hat{z}_n(E)$, $\hat{z}_n$ has continuous third derivatives on E, l_n''' is uniformly bounded on $E \times \hat{z}_n(E)$, and $\hat{z}_n'''$ is uniformly bounded on E, $a.s.[y_{1:n}]$.

(A2) The two functions $\hat{z}_n \to \hat{z}_* \ a.s.[y_{1:n}]$ on Θ for some function $\hat{z}_*$, $l_n \to l_* \ a.s.[y_{1:n}]$ for some function l_*.

(A3) The limit l_* has positive definite $l_*''\{\lambda_0, \hat{z}_*(\lambda_0)\}$ and satisfies $\frac{\partial l_*(\lambda_0,\zeta)}{\partial \zeta}|_{\zeta=\hat{z}_*(\lambda_0)} = 0$.

(A4) For some compact $K \subseteq E$ with λ_0 in the interior of K,

$$l_*(\lambda, \zeta) < l_*\{\lambda_0, \hat{z}_*(\lambda_0)\} \text{ for all } \lambda \in K \backslash \{\lambda_0\}, \zeta \in \hat{z}_*(E) \ a.s.[y_{1:n}],$$

$$\limsup_n \sup_{\lambda \in \Theta \backslash K, \zeta \in \hat{z}_n(\Theta)} l_n(\lambda, \zeta) < l_*\{\lambda_0, \hat{z}_*(\lambda_0)\} \ a.s.[y_{1:n}].$$

Conditions (A1–A2) are often imposed to enable a second-order Taylor expansion (Miller, 2021); (A3) focuses on the cases when $\lambda = \lambda_0$ and gives the local second-order optimal condition of $l_*(\lambda_0, \zeta)$ at $\zeta = \hat{z}_*(\lambda_0)$, where $\hat{z}_*(\lambda_0)$ can be produced as the minimizer of another loss function g; (A4) ensures the dominance of l_* at $\{\lambda_0, \hat{z}_*(\lambda_0)\}$ over all possible (λ, ζ) in the described neighborhood, including those points with $\lambda \neq \lambda_0$. With the above, we are ready to state the BvM result on the general bridged posterior for parametric models where $\zeta \in \mathbb{R}^p$ has a fixed and finite dimension.

Theorem 3-2. *Under (A1–A4), there is a sequence $\lambda_n \to \lambda_0$ such that $\hat{l}_n'(\lambda_n) = 0$ for all n large enough, $\hat{l}_n(\lambda_n) \to \hat{l}_*(\lambda_0)$ where $\hat{l}_*(\lambda) = l_*\{\lambda, \hat{z}_*(\lambda)\}$. Further, letting q_n be the density of $\sqrt{n}(\lambda - \lambda_n)$ when $\lambda \sim \Pi_n(\lambda \mid y)$, and $\mathcal{N}$ the normal density, we have the total variational distance $d_{TV}\{q_n, \mathcal{N}(0, H_0^{-1})\} \xrightarrow[n\to\infty]{a.s.[y_{1:n}]} 0$ with $H_0 = \hat{l}_*''(\lambda_0)$.*

The result above shows that fixing ζ to z does not impact the asymptotic variance of λ. On the other hand, since z is finite-dimensional and differentiable on E, we can use the delta method

to find out the asymptotic covariance of z. For bridged posterior using profile likelihood, we do find lower uncertainty in $\Pi(z \mid y)$ under a bridged posterior compared to $\Pi(\zeta \mid y)$ under an integrated one, as formalized below.

Corollary 3.1. *Under (A1–A4) and $g_n(\zeta, y_{1:n}; \lambda) = -L(y_{1:n}, \zeta; \lambda)$, for $j = 1, \ldots, p$, the asymptotic variance of the j-th element of $\sqrt{n}\{\zeta - \hat{z}_n(\lambda_n)\}$ is strictly greater than the one of the j-th element of $\sqrt{n}\{\hat{z}_n(\lambda) - \hat{z}_n(\lambda_n)\}$.*

3.4.2 Semi-parametric Bridged Posterior using Profile Likelihood

In the semi-parametric setting, we assume that ζ can be infinite-dimensional and live in some Hilbert space $\mathcal{H}$, and that there exists a fixed $\zeta_0 \in \mathcal{H}$. We define

$$l_n(\lambda, \zeta) = \log L(y_{1:n}, \zeta; \lambda)/n, \qquad \hat{l}_n(\lambda) = \log\{\sup_{\zeta} L(y_{1:n}, \zeta; \lambda)\}/n,$$

where the former corresponds to a full likelihood $L(y_{1:n}, \zeta; \lambda)$ with unconstrained ζ, and the latter to a profile likelihood $\sup_\zeta L(y_{1:n}, \zeta; \lambda)$. In addition to the potentially infinite dimension, another challenge is that $l_n(\lambda, \zeta)$ may not be differentiable with respect to ζ.

To facilitate analysis under these challenges, we use the approximately least-favorable submodel technique; Kosorok (2008) provides a detailed explanation. For this section to be self-contained, we overview the important definitions that are involved as the building blocks for establishing BvM results.

Submodel. For each $(\lambda, \zeta) \in \Theta \times \mathcal{H}$, consider a map $\tilde{\zeta}_t(\lambda, \zeta)$ indexed by $t \in \Theta \subset \mathbb{R}^d$:

$$l_n\{t, \tilde{\zeta}_t(\lambda, \zeta)\} \text{ is twice differentiable in } t \in \Theta, \qquad \tilde{\zeta}_{t=\lambda}(\lambda, \zeta) = \zeta. \tag{3-8}$$

Commonly, $l_n\{t, \tilde{\zeta}_t(\lambda, \zeta)\}$ is called a submodel with parameters (t, λ, ζ) (Murphy and Van der Vaart, 2000). For convenience, we use notation $\tilde{l}_n(t, \lambda, \zeta) := l_n\{t, \tilde{\zeta}_t(\lambda, \zeta)\}$.

Efficient score and Fisher information. Conventionally, the λ-score function of the full likelihood is $\dot{l}_n(\lambda, \zeta) = \frac{\partial l_n(\lambda, \zeta)}{\partial \lambda}$. Consider a direction $\delta \in \widetilde{\mathcal{H}}$ (another Hilbert space) such that a path $\{\zeta_\gamma^\delta \in \mathcal{H}\}_{\gamma \in \mathbb{R}^d}$ with $\zeta_\gamma^\delta \to \zeta_0$ as $\gamma \to \lambda_0$. We can now define the generalized ζ-score function

at $\zeta = \zeta_0$ in the direction of δ by $A^n_{\lambda_0,\zeta_0}\delta := \left.\dfrac{\partial l_n(\lambda_0, \zeta_\gamma^\delta)}{\partial \gamma}\right|_{\gamma=\lambda_0}$, where $A^n_{\lambda_0,\zeta_0} : \widetilde{\mathcal{H}} \mapsto L_2^d(P_{\lambda_0,\zeta_0})$ is a map, and $L_2^d(P_{\lambda_0,\zeta_0})$ is the space of d-dimensional vector-valued functions $\{\alpha_1(y), \ldots, \alpha_d(y)\}$ where each $\alpha_i(y)$ is L_2-integrable on $y \sim P_{\lambda_0,\zeta_0}$. In the Section B.2, we provide an illustration of the above via the Cox regression model. The efficient score function for λ at (λ_0, ζ_0) is defined by

$$\begin{aligned}
&\mathscr{Q}\dot{l}_n(\lambda_0, \zeta_0) := \dot{l}_n(\lambda_0, \zeta_0) - \mathscr{P}\dot{l}_n(\lambda_0, \zeta_0),\\
&\mathscr{P}\dot{l}_n(\lambda_0, \zeta_0) := \arg\min_k \mathbb{E}_{\lambda_0,\zeta_0}\left\|\dot{l}_n(\lambda_0, \zeta_0) - \kappa\right\|^2, \quad \kappa \in \text{closed linear span of } A^n_{\lambda_0,\zeta_0}\delta.
\end{aligned}$$

The efficient Fisher information at (λ_0, ζ_0) is defined as $\tilde{I}_0 := \mathbb{E}_{\lambda_0,\zeta_0}\left\{\mathscr{Q}\dot{l}_n(\lambda_0, \zeta_0)\mathscr{Q}\dot{l}_n(\lambda_0, \zeta_0)^\top\right\}$. Equivalently, $\mathscr{P}\dot{l}_n(\lambda_0, \zeta_0)$ is the projection of the score function for λ_0 onto the closed linear space spanned by the set $\{A^n_{\lambda_0,\zeta_0}\delta\}_{\delta\in\widetilde{\mathcal{H}}}$.

Least favorable model. To connect the two topics above, notice that if (3-8) further satisfies $\left.\dfrac{\partial \tilde{l}_n(t,\lambda_0,\zeta_0)}{\partial t}\right|_{t=\lambda_0} = \mathscr{Q}\dot{l}_n(\lambda_0, \zeta_0)$, then we have the submodel $\tilde{l}_n(t, \lambda, \zeta)$ least favorable at $t = \lambda_0$. This is because among all submodels $\tilde{l}_n(t, \lambda_0, \zeta_0)$, this submodel has the smallest Fisher information on each dimension of $t \in \Theta \subset \mathbb{R}^d$ by the definition of $\mathscr{Q}\dot{l}_n(\lambda_0, \zeta_0)$. Since our focus is on the asymptotic regime, we only need the least favorable model condition to hold in a limiting sense. This leads to the approximately least favorable model. With these ingredients, we are ready to derive our results. We first show that the profile $\hat{l}_n(\lambda)$ is locally asymptotically normal (LAN). We require the following sufficient conditions.

(B1) There exists a neighborhood $V \subset \Theta \times \Theta \times \mathcal{H}$ containing $(\lambda_0, \lambda_0, \zeta_0)$ such that

- $\sup_{(t,\lambda,\zeta)\in V}\left\|\frac{\partial^2 \tilde{l}_n(t,\lambda,\zeta)}{\partial t^2} + H_0\right\| \xrightarrow[n\to\infty]{P_{\lambda_0,\zeta_0}} 0$ for some symmetric $H_0 \in \mathbb{R}^{d\times d}$; and
- $\sup_{(t,\lambda,\zeta)\in V}\sqrt{n}\left\|\frac{\partial \tilde{l}_n(t,\lambda,\zeta)}{\partial t} - \mathbb{E}_{\lambda_0,\zeta_0}\frac{\partial \tilde{l}_n(t,\lambda,\zeta)}{\partial t} - h_n\right\| \xrightarrow[n\to\infty]{P_{\lambda_0,\zeta_0}} 0$ for a sequence of random variables $h_n \in \mathbb{R}^d$,

where P_{λ_0,ζ_0} and $\mathbb{E}_{\lambda_0,\zeta_0}$ are defined based on the ground-truth distribution of $y_{1:n}$.

(B2) The function $\hat{z}_n(\lambda)$ converges to ζ_0 when $\lambda \to \lambda_0$ and $n \to \infty$.

(B3) There exists a neighborhood $U \subset \Theta$ containing λ_0 such that

$$\mathbb{E}_{\lambda_0,\zeta_0} \frac{\partial \tilde{l}_n\{t, \lambda, \hat{z}_n(\lambda)\}}{\partial t}\bigg|_{t=\lambda_0} = o_{P_{\lambda_0,\zeta_0}}(1)\left(\|\lambda - \lambda_0\| + n^{-1/2}\right) \tag{3-9}$$

holds for all $\lambda \in U$, $o_{P_{\lambda_0,\zeta_0}}(1)$ refers to a term that converges to 0 in P_{λ_0,ζ_0} as $n \to \infty$.

Conditions (B1–B3) are the approximately least favorable submodel conditions. A similar result for iid data given (ζ_0, λ_0) has been previously shown in Murphy and Van der Vaart (2000, Theorem 1). However, our result is more general and holds regardless of whether $(y_{1:n} \mid \zeta_0, \lambda_0)$ are iid or not, and may be of independent interest outside the context of establishing BvM results.

Lemma 3-1. *Under (B1–B3), there exists a neighborhood $B_\epsilon(\lambda_0)$ for some $\epsilon > 0$ such that*

$$\hat{l}_n(\lambda) - \hat{l}_n(\lambda_0) = (\lambda - \lambda_0)^\top h_n - \frac{1}{2}(\lambda - \lambda_0)^\top H_0(\lambda - \lambda_0) + o_{P_{\lambda_0,\zeta_0}}(1)\left\{\left(\|\lambda - \lambda_0\| + n^{-1/2}\right)^2\right\} \tag{3-10}$$

holds for all $\lambda \in B_\epsilon(\lambda_0)$.

With the LAN condition for $\hat{l}_n(\lambda)$, we make the the BvM result statement.

Theorem 3-3. *Assume (3-10) holds with positive definite H_0. Suppose that the maximum likelihood estimator $\hat{\lambda}_n$ exists and converges to λ_0 when $n \to \infty$; for any $\epsilon > 0$, there exists $\delta > 0$ such that*

$$P_{\lambda_0,\zeta_0}\left[\inf_{\|\lambda - \hat{\lambda}_n\| \geq \epsilon} \left\{\hat{l}_n(\hat{\lambda}_n) - \hat{l}_n(\lambda)\right\} \geq \delta\right] \xrightarrow[n\to\infty]{} 1. \tag{3-11}$$

Then letting π_n be the density of λ when $\lambda \sim \Pi_n(\lambda \mid y)$, we have

$$\int_{B_\epsilon(\lambda_0)} \pi_n(\lambda)\, \mathrm{d}\lambda \xrightarrow[n\to\infty]{P_{\lambda_0,\zeta_0}} 1 \text{ for all } \epsilon > 0, \tag{3-12}$$

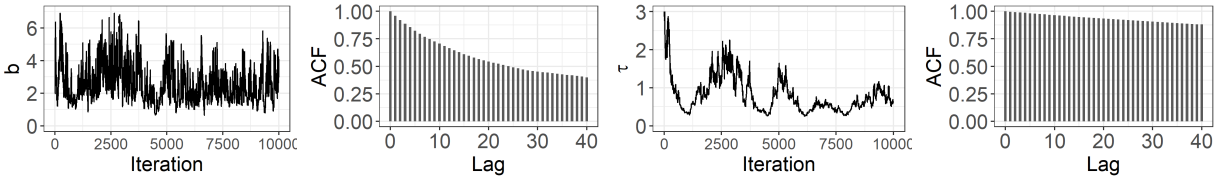
and letting q_n be the density of $\sqrt{n}(\lambda - \hat{\lambda}_n)$, we have $d_{TV}\left\{q_n, \mathcal{N}\left(0, H_0^{-1}\right)\right\} \xrightarrow[n\to\infty]{P_{\lambda_0,\zeta_0}} 0$.

We provide a detailed comparison with existing BvM results on semi-parametric models in Section B.3.

3.5 Simulations

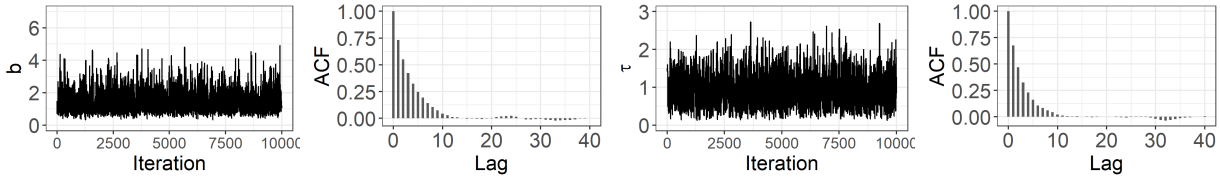
3.5.1 Latent Quadratic Exponential Model

We compare a latent normal model and a latent quadratic exponential model, based on Example 3.3. To simulate data for benchmarking, we generate random locations $x_1, \ldots, x_{1000} \sim \text{Uniform}(-6, 6)$, and ground-truth mean from a latent curve $\tilde{z}_i = \cos(x_i)$. At each x_i, we generate a binary observation $y_i \sim \text{Bernoulli}(1/\{1 + \exp(-\tilde{z}_i)\})$.



Traceplot of b | Autocorrelation of b | Traceplot of τ | Autocorrelation of τ

(a) The MCMC samples produced by the data augmentation Gibbs sampler applied on the latent normal model.



Traceplot of b | Autocorrelation of b | Traceplot of τ | Autocorrelation of τ

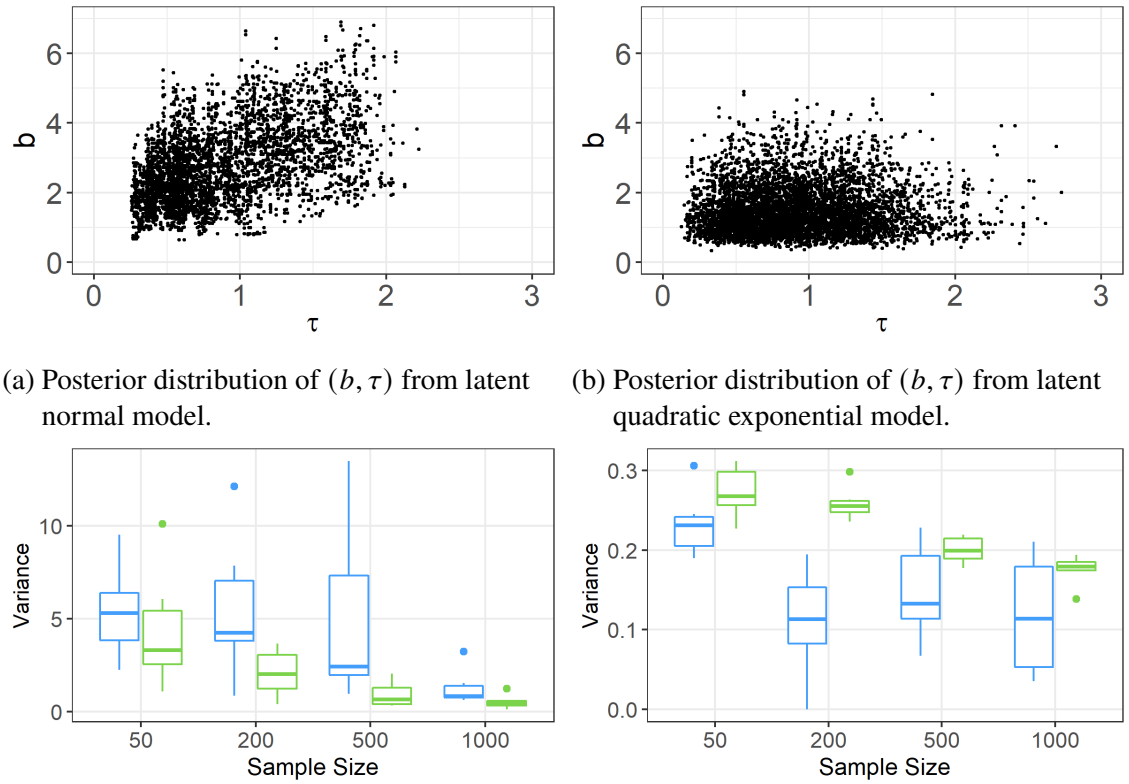
(b) The MCMC samples produced by the random walk Metropolis applied on the latent quadratic exponential model.

Figure 3-2. Compared to the latent normal model using data augmentation Gibbs sampler, the latent quadratic exponential model (a bridged posterior model) can be estimated using a much simpler random walk Metropolis, while enjoying faster mixing of the Markov chains.

We fit the latent quadratic exponential model (3-6) and the latent normal model (3-5) to the simulated data. For both models, we assign half-normal $\text{N}_+(0, 1)$ prior on τ and Inverse-Gamma$(2, 5)$ prior on b. We use random walk Metropolis for the latent quadratic exponential model. For the latent normal model with binary observations, we follow Polson et al. (2013) and use the following data augmentation:

$$L(\lambda, \zeta, \eta; y) \propto \exp\left\{-\frac{1}{2}\zeta^\top Q^{-1}(\lambda; x)\zeta\right\} \prod_{i=1}^{n} \exp\{(y_i - 1/2)\zeta_i\} \exp\{-\eta_i(\zeta_i)^2/2\}\text{PG}(\eta_i; 1, 0)\pi_0(\lambda),$$

where $\mathrm{PG}(\cdot; 1, 0)$ is the density of Pólya-Gamma$(1, 0)$ distribution. This leads to closed-form update of ζ from a normal full conditional distribution. Then a Gibbs sampler is used.



(a) Posterior distribution of (b, τ) from latent normal model.

(b) Posterior distribution of (b, τ) from latent quadratic exponential model.

(c) Boxplots of posterior variances of b at different sample sizes.

(d) Boxplots of posterior variances of τ at different sample sizes.

Figure 3-3. The posterior distributions of the covariance kernel parameters from the latent normal model (Panel a) and the latent quadratic exponential model (Panel b), collected from two experiments under sample size 1000. The experiments are repeated under different sample sizes, and the posterior variances of b and τ from the latent quadratic exponential model (green) and the latent normal model (blue) are shown in Panels c and d.

We run each MCMC algorithm for $10,000$ iterations and discard the first $2,000$ as burn-ins. The latent quadratic exponential model takes about 8.37 minutes, and the latent normal model takes about 11.87 minutes on a 12-core laptop. Figure 3-2 compares the mixing of MCMC algorithms for those two models. Clearly, the latent quadratic exponential model mixes better, while taking less runtime. In terms of effective sample size for (b, τ) per time unit (10 seconds

wall time) (ESS/time), the latent quadratic exponential model achieves 0.100 and 0.129, while the latent normal model yields only 0.0079 and 0.0009.

Next, we compare the posterior distributions of parameters (τ, b). As can be seen in Figure 3-3, these two distributions show a similar range of τ and b in the high posterior probability region. Since these two distributions correspond to two distinct models, we do not expect the distributions of τ or b to match exactly. On the other hand, we can see that the posterior variances are on the same scale, with $\text{Var}(b \mid y) = 1.08^2$ and $\text{Var}(\tau \mid y) = 0.41^2$ for the latent normal model, and $\text{Var}(b \mid y) = 0.66^2$ and $\text{Var}(\tau \mid y) = 0.41^2$ for the latent quadratic exponential model.

We repeat the experiments and compare the variances under different sample sizes. We use a simulation experiment of latent quadratic exponential model to show that the profile likelihood-based bridged model has an asymptotic posterior variance for the parameter λ that is equal to that of the Bayesian model based on the full likelihood.

We generate random locations $x_1, \ldots, x_S \sim \text{Uniform}(-6, 6)$ where $S \in \{50, 200, 500, 1000\}$ is the sample size, and ground-truth means from 6 latent curves $\tilde{z}_{ji} = f_j(x_i)$ where $j = 1, \ldots, 6$. At each x_i for each curve, we generate a binary $y_{ji} \sim \text{Bernoulli}(1/\{1 + \exp(-\tilde{z}_{ji})\})$.

For each group of data $y_{j1}, \ldots, y_{jS}$ from the j-th curve, we fit both the latent quadratic exponential model and the latent normal model. For both model, we assign a half-normal $\text{N}_+(0, 1)$ prior on τ and Inverse-Gamma$(2, 5)$ prior on b. We use the same algorithm that is used in Section 5. Since the latent quadratic exponential model enjoys much better mixing performance compared to the latent normal model, for the former model, we run the MCMC algorithm for 5000 iterations, discard the first 2000 as burn-ins and the samples are thinned at 10, while for the latter one, we run the MCMC algorithm for 13000 iterations, discard the first 4000 as burn-ins, and the samples are thinned at 30.

3.5.2 Bayesian Maximum Margin Classifier

To illustrate the strengths of our approach in terms of uncertainty quantification and borrowing information from unlabeled data, we apply the Bayesian maximum margin classifier (Example 4) to prediction on heart failure-related deaths. The dataset we consider comprises 299

total patients who had a previous occurrence of heart failure. For each patient, there are 12 measured clinical features, with binary outcomes y_i on whether the patient died during a follow-up care period between April and December 2015, at the Faisalabad Institute of Cardiology and at the Allied Hospital in Faisalabad, Pakistan. There are 194 men and 105 women between age 40 and 95.

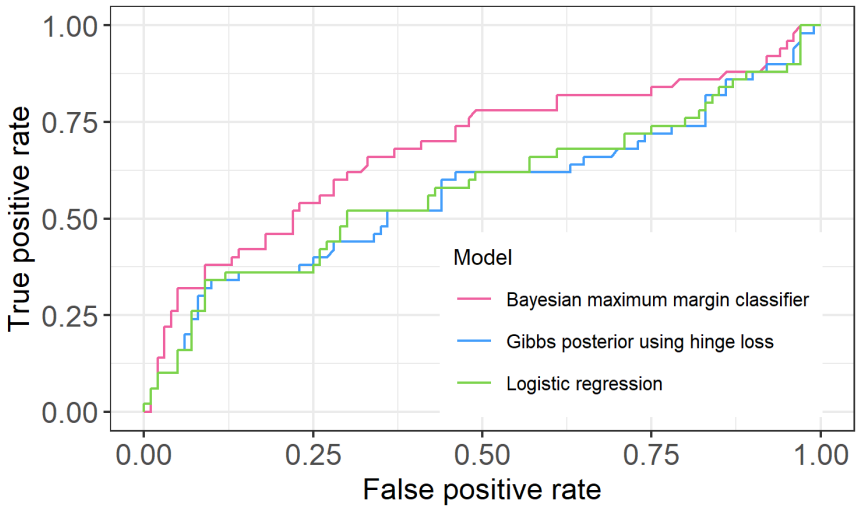


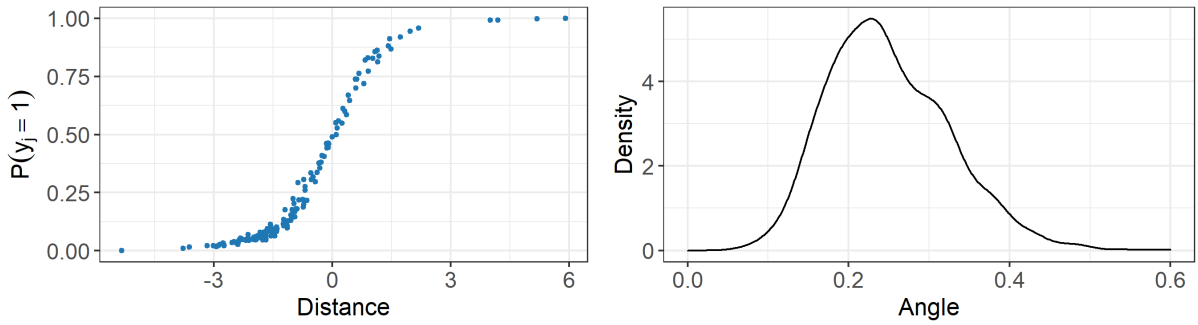
Figure 3-4. The prediction receiver operating characteristic curves from the three models.

We mask the outcomes of randomly chosen 97 men and 52 women (corresponding to roughly 50% missing labels), and fit the data under the (i) Bayesian maximum margin classifier model, (ii) a Gibbs posterior model using hinge loss, and (iii) logistic regression. We specify the priors $\lambda \sim \text{Gamma}(3, 2)$ for models (i) and (ii), and $\zeta_w \sim \text{N}(0, 3^2 I)$ and $\zeta_b \sim \text{N}(0, 3^2)$ for models (ii) and (iii). For each model, we run MCMC for $1,500$ iterations and discard the first 500 as burn-in. At each iteration, we make a binary prediction on each unlabeled x_j, using the average as a posterior estimate for predicting $P(y_j = 1 \mid y_{1:n})$ for $j = n+1, \ldots, n+k$. Comparing each of these prediction probabilities with the true y_j produces the prediction receiver operating characteristic (ROC) curves, displayed in Figure 3-4. For binary estimates, we use 0.5 as the threshold probability and report classification accuracy. Figure 3-4 reveals a barely noticeable difference between logistic regression and the Gibbs posterior using hinge loss. In contrast, the

Bayesian maximum margin classifier clearly produces higher area under the curve (AUC). This advantage is also apparent in terms of classification accuracy, displayed in Table 3-1.

Table 3-1. Prediction accuracy for heart failure dataset using four methods.

Method	Area Under ROC Curve	Classification Accuracy
Bayesian maximum margin classifier	0.681	0.707
Gibbs posterior using hinge loss	0.568	0.653
Support vector machine	-	0.653
Logistic regression	0.577	0.673



(a) Posterior prediction $P(y_j = 1)$ versus distance to the posterior mean of decision boundary hyperplane.

(b) Posterior distribution of the absolute angle (in radians) between z_w and the posterior mean $\bar{z}_w$.

Figure 3-5. Uncertainty estimates for the Bayesian maximum margin classifier applied on the heart failure dataset.

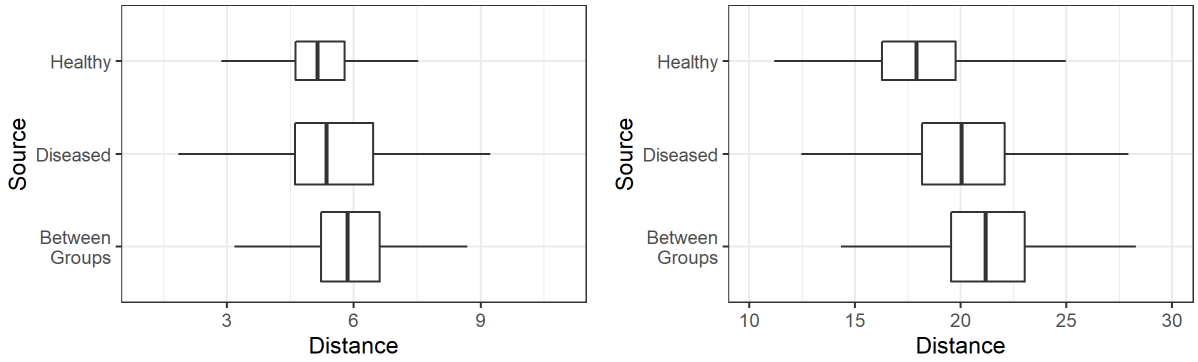
To see that these gains are largely due to borrowing information from the unlabeled data, we also fit a support vector machine only using the labeled part, and hold out the unlabeled portion for prediction. Here, the classification accuracy falls to similar levels as the other two Bayesian models, with the threshold probability at 0.5.

Finally, in addition to ROC curves, Figure 3-5 shows the other uncertainty estimates that describe how the posterior prediction $P(y_j = 1)$ changes with the distance between x_j and the posterior mean of the decision boundary hyperplane $\{x : z_w^\top x + z_b = 0\}$. We can also consider how the posterior distribution describing this decision boundary varies around the posterior mean, in terms of angle between z_w and $\bar{z}_w$.

3.6 Application on Functional Connectivity Graphs

We now use the proposed method to model a collection of raw functional connectivity graphs. The graphs were extracted from resting-state functional magnetic resonance imaging (rs-fMRI) scans, collected from $S = 166$ subjects, of whom 64 are healthy subjects and 102 are at various stages of Alzheimer's disease. For each subject, a functional connectivity matrix was produced via a standard neuroscience pre-processing pipeline (Ding et al., 2006), summarized in the form of a symmetric, weighted adjacency matrix, denoted by $A^{(s)} \in \mathbb{R}^{R \times R}_{\geq 0}$ between $R = 116$ regions of interests (ROIs) for subjects $s = 1, \ldots, S$; there are no self-loops—that is, $A^{(s)}_{i,i} = 0$ for all $i = 1, \ldots, R$.

The graph Laplacian $\mathcal{L}^{(s)} = D^{(s)} - A^{(s)}$ is a routinely used one-to-one transform of $A^{(s)}$, where $D^{(s)}$ is a diagonal matrix $D^{(s)}_{ii} = \sum_{j=1}^{R} A^{(s)}_{i,j}$. Compared to the adjacency matrix, the Laplacian enjoys a few appealing properties, namely: (i) $\mathcal{L}^{(s)}$ is always positive semidefinite, (ii) the number of zero eigenvalues equals the number of disjoint component sub-graphs (each known as a community); (iii) the first smallest non-zero eigenvalues quantify the connectivity (normalized graph cut) in each component sub-graph. Because of these properties, we can quantify the difference between two graphs via the geodesic distance in the interior of the positive definite cone (Lim et al., 2019). For two positive semidefinite matrices X and Y of equal size, $\mathrm{dist}(X, Y) = \lim_{\eta \to 0_+} \left(\sum_{j=1}^{R} \log^2 \left[\xi_j \{ (X + I\eta)^{-1} (Y + I\eta) \} \right] \right)^{1/2}$, where $\xi_j(\cdot)$ is the j-th eigenvalue.



(a) Boxplot of pairwise distances of the observed Laplacian matrices.

(b) Boxplot of pairwise distances of the low-rank smoothed Laplacian matrices using different λ_s for each subject.

Figure 3-6. Boxplots of the pairwise distances among the observed Laplacian matrices, and that among the smoothed Laplacian matrices.

Figure 3-6(a) plots the pairwise distances between the observed $\mathcal{L}^{(s)}$'s, using three boxplots corresponding to subjects within the diseased group, subjects within the healthy group, and those between the two groups. Though we see that the within-group distances have slightly smaller means than the between-group distances, there is significant overlap among the three boxplots. This is understandable since each observed $\mathcal{L}^{(s)}$ is of rank $(R-1)$ corresponding to having no disjoint components (equivalent to one community only). It is reasonable to take a reduced-rank smoothing by solving the following convex problem:

$$Z^{(s)} = \arg\min_{\zeta} \frac{1}{2}\|\mathcal{L}^{(s)} - \zeta\|^2 + \tilde{\lambda}_s \|\zeta\|_*$$
$$\text{subject to } \zeta \in \mathbb{R}^{n\times n}, \zeta_{i,i} = -\sum_{j:j\neq i} \zeta_{i,j}, \zeta_{i,j} = \zeta_{j,i} \leq 0 \text{ for } i \neq j,$$

where $\|\zeta\|_*$ is the nuclear norm of matrix ζ (the sum of singular values of ζ), and $\tilde{\lambda}_s > 0$ is the tuning parameter. As the result, each $Z^{(s)}$ with rank $(R - K^{(s)})$ is the graph Laplacian of a $K^{(s)}$-community graph, where $K^{(s)}$ is monotonically non-decreasing in $\tilde{\lambda}_s > 0$.

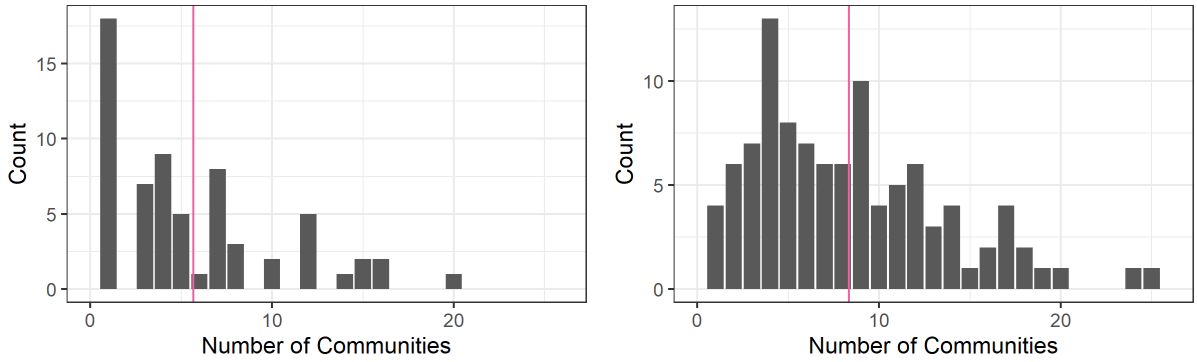
Note that treating $Z^{(s)}$ here as a latent variable is especially appealing—if we had viewed it as a parameter in $\mathbb{R}^{n\times n}$, we would incur a high modeling and computational cost. The question boils down to a meaningful choice of $\tilde{\lambda}_s$. Due to the heterogeneity of $\mathcal{L}^{(s)}$, the same value of

$\tilde{\lambda}_s = \tilde{\lambda}_{s'}$ may yield quite different $Z^{(s)}$ and $Z^{(s')}$. As a result, for the purpose of data harmonization, instead of assigning an independent prior or equal value on $\tilde{\lambda}_s$, we assign a dependent likelihood based on the pairwise distances among $Z^{(s)}$:

$$
\begin{aligned}
&L\left[\{\mathcal{L}^{(s)}, Z^{(s)}\}_{s=1}^{S}; \{\tilde{\lambda}_s\}_{s=1}^{S}, \sigma^2, \tau\right] \\
&\propto \left[\prod_{s=1}^{S} (\sigma^2)^{-1/2} \exp\left\{-\frac{\|\mathcal{L}^{(s)} - Z^{(s)}\|_F^2}{2\sigma^2}\right\} \frac{\tilde{\lambda}_s}{\sigma^2} \exp\left\{-\frac{\tilde{\lambda}_s \|Z^{(s)}\|_*}{\sigma^2}\right\}\right] \\
&\quad \times \left[\prod_{s=1}^{S} \tau^{-1/2} \exp\left\{-\frac{\sum_{k:k\neq s} \mathrm{dist}^2\left(Z^{(k)}, Z^{(s)}\right)/(S-1)}{2\tau}\right\}\right].
\end{aligned}
$$

The second line is a pairwise kernel via the average total squared geodesic distance between each $Z^{(s)}$ and other $Z^{(s')}$, so that it borrows information across subjects to reduce the heterogeneity. We clarify that group information is not used above; hence it can serve as a data harmonization tool, even in the absence of group labels.

To calculate $Z^{(s)}$ at different $\tilde{\lambda}_s$, we use the alternating direction method of multipliers (ADMM) algorithm (details are provided in the appendix). To facilitate computation, for each $\tilde{\lambda}_s$, we assign a discrete uniform equally spread over 10 values in $(0, 5]$, so that the possible values of $Z^{(s)}$ as well as their associated pairwise geodesic distances can be precomputed before running MCMC. We specify an Inverse-Gamma$(2, 1)$ prior for σ^2 and Inverse-Gamma$(2, 1)$ prior for τ. Running MCMC for $10,000$ iterations takes 46.5 minutes on a 12-core laptop; the first $2,000$ samples are discarded as burn-in.



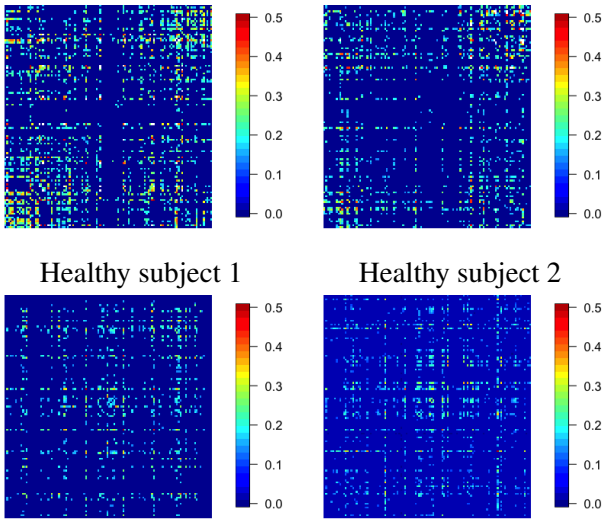
(a) Healthy group (64 subjects).

(b) Diseased group (102 subjects).

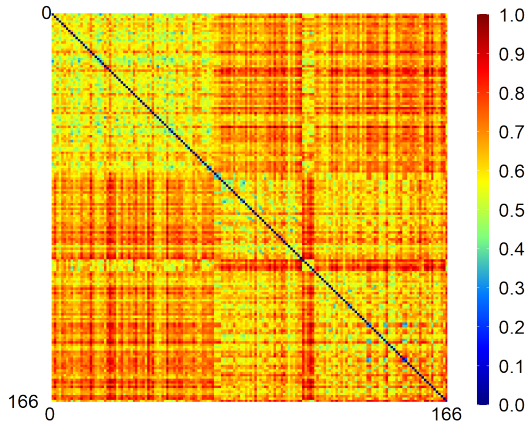
Figure 3-7. The barplots on the number of communities in $Z^{(s)}$ at each subject's posterior mean λ_s. The vertical line is the mean of the number of communities over all subjects.

Using the smoothed Laplacian $Z^{(s)}$, we calculate posterior mean of the the distance matrix $\{\text{dist}(Z^{(s)}, Z^{(s')})\}_{\text{all } s,s'}$ and show re-calculated boxplots of geodesic distances in Figure 3-6(b). Clearly, between the low-rank smoothed $Z^{(s)}$, the healthy group now has much lower pairwise distances than the diseased group, while the diseased group has slightly lower pairwise distances compared to the between-group. We compute the Kolmogorov–Smirnov (KS) statistical metric between the empirical distribution of geodesic distances. When we switch from using raw $\mathcal{L}^{(s)}$ to smoothed $Z^{(s)}$, the KS metric between the diseased and healthy increases from 0.143 to 0.299.

By calculating the number of zero eigenvalues of the $Z^{(s)}$, we find $K^{(s)}$ as the number of communities for each subject. Figure 3-7 shows histograms of $K^{(s)}$ evaluated at each subject's posterior mean $\tilde{\lambda}^s$. The average number of communities for the healthy subjects is 5.77 while it is 8.41 for the diseased subjects. This is consistent with the known fact that a diseased brain tends to be more fragmented than a healthy one, due to the disruptions caused by Alzheimer's disease.



(a) Smoothed adjacency matrices for four subjects.



(b) Posterior mean of the pairwise geodesic distance among the smoothed Laplacians. Top-left 64×64: healthy subjects, bottom-right 102×102: diseased subjects.

Figure 3-8. Illustration of the smoothed graph estimates. Panel a plots the smoothed adjacency matrices for four subjects, based on $\tilde{A}^{(s)} = -Z^{(s)}$ (the diagonal elements are masked) with $Z^{(s)}$ obtained at the posterior mean of $\lambda^{(s)}$ for each s, Panel b shows the posterior mean matrix of the geodesic distances between $Z^{(1)}, \ldots, Z^{(S)}$.

Figure 3-8 shows the smoothed adjacency matrices for two subjects chosen from the healthy group, and two from the diseased group, and the posterior mean of the pairwise geodesic distances. To validate the result, we further apply spectral clustering on the pairwise distance matrix, and cluster the subjects into two groups. Based on the posterior mean distance matrix among $Z^{(s)}$, 96.9% of the subjects in the healthy group are correctly grouped together, and 89.2% for the diseased group. Using the distance matrix among raw $\mathcal{L}^{(s)}$, these numbers are 87.5% and 89.2%.

CHAPTER 4
GRADIENT-BRIDGED POSTERIOR: BAYESIAN INFERENCE FOR MODELS WITH IMPLICIT FUNCTIONS

In this chapter, motivated to develop a conceptually simple and computationally efficient alternative to the bridged posterior, we propose a continuous posterior distribution that concentrates around the partial minimizer of h over z for any β, for the class of models where loss in the sub-problem is differentiable. We propose to exploit the equivalence of $\hat{z}_\beta = \arg\min_z h(\beta, z; y)$ and $\nabla_z h(\beta, z; y) \mid_{z=\hat{z}_\beta} = 0$ over z in an open set. This is known as the first-order optimality condition in optimization, and motivates a shrinkage kernel we impose on the gradient norm to induce a continuous distribution surrounding the partial minimizer $z \approx \hat{z}_\beta$. In doing so, the likelihood becomes tractable up to an unimportant normalizing constant, enabling straightforward inference. Elaborating on a simple core idea, we carefully establish statistical guarantees via a Bernstein-von Mises theorem, and detail several nuanced cases including its applicability in non-open domains and dual formulations. The method is amenable to efficient posterior computation via gradient-based samplers, and its merits are showcased empirically via synthetic and real data case studies.

4.1 Motivation

In modern statistical applications, it is often desirable to posit that some parameters arise as a solution to an optimization sub-problem. Here, a sub-problem refers to a separate loss function, which may or may not be the same as the negative-log-likelihood that characterizes the stochastic component of a model. For example, in Procrustes data analysis widely used for examining shape or microarray data, the Procrustes distance is defined as the metric between two sets of points after applying optimal rotation and reflection; thus, the rotation and reflection parameters are solutions to a Procrustes alignment sub-problem (Dryden and Mardia, 2016; Goodall, 1991). Similarly, in flow network analysis, as encountered in transportation or communication systems, it is common to model the observed flow as a noisy version of an optimal flow. The latter sub-problem is routinely defined as a linear program (Bazaraa et al., 2011, chapter 12).

Yang Yaozhi is the joint first author of this work.

The need to conduct statistical inference under these settings, especially under finite, moderate sample sizes, makes it appealing to consider a Bayesian approach. To be concrete, we may use the following likelihood,

$$L(y;\beta,z_\beta) \propto g(y;\beta,\hat{z}_\beta), \qquad \text{subject to } \hat{z}_\beta = \arg\min_z h(\beta,z;y), \tag{4-1}$$

with an appropriate prior $\beta \sim \pi_0$. Under the Bayesian paradigm, estimation uncertainty is characterized through the posterior $\Pi(\beta \mid y)$. When the sub-problem loss $h = -\log g$, (4-1) becomes a profile likelihood (Murphy and Van der Vaart, 2000; Maclaren, 2018); there have been many studies that justify this in the Bayesian setting (Lee et al., 2005a; Cheng and Kosorok, 2008, 2009). More generally, Zeng et al. (2024a) coined the term *bridged posterior* for (4-1), as it bridges a divide between Bayesian and optimization methodologies. The bridged posterior can be broadly viewed as a generalization of the projected posterior, for which h is a projection loss $h(\beta,z;y) = \text{dist}(z,\beta)$ for some appropriate distance and z in some constrained space (Astfalck et al., 2024; Chakraborty and Ghosal, 2022; Lee et al., 2023).

Conceptually, the posterior associated with (4-1) is an equality-constrained posterior (Gelfand et al., 1992). In principle, one can fit the bridged posterior computationally by simply running an optimization algorithm nested within a Markov chain Monte Carlo sampler. However, the lack of explicit forms for $\hat{z}_\beta$ and hence $L(y;\beta,\hat{z}_\beta)$ limit interpretability. These also translate into obstacles to inference, complicating procedures that are typically straightforward from a Bayesian perspective such as model-based imputation for missing values. In some cases, these difficulties can be reconciled: for instance, for some pairs g and h involving convex conjugate functions, Polson and Scott (2016) showed that an equivalent mixture representation may exist where $g(y;\beta,\hat{z}_\beta) \propto \int \tilde{g}(y;\beta,\gamma)\pi(\gamma)\mathrm{d}\gamma$, with both g and π tractable. Our contributions seek a more comprehensive treatment to fill this gap when such specific representations are unavailable.

When it comes to constructing a Bayesian model motivated by a loss function, the Gibbs posterior (Jiang and Tanner, 2008) is one of the most popular choices. By exponentiating a

negative loss function $\exp\{-\lambda\tilde{h}(\theta; y)\}$ with some $\lambda > 0$, mimicking the usual role of the log-likelihood in standard Bayesian inference, one defines a distribution that concentrates around the minimizer $\hat{\theta} = \arg\min_\theta \tilde{h}(\theta; y)$. Application of Gibbs posteriors, also falling under the generalized Bayes label, covers a wide range of models (Bissiri et al., 2016; Martin and Syring, 2022; Bhattacharya and Martin, 2022; Rigon et al., 2023b; West, 2024).

Unfortunately, the Gibbs posterior cannot accommodate the implicit function setting of (4-1). This is due to the two-way dependence of $h(\beta, z; y)$ on z and β, given y. A Gibbs posterior approach would force both β and z to concentrate near a point $(\hat{\beta}, \hat{z}) = \arg\min_{(\beta,z)} h(\beta, z; y)$. Instead, we desire to properly account for the variability of β, as (4-1) is based on a *partial* minimization of $h(\beta, z; y)$ over z only. To illustrate, consider a simple loss $h(\beta, z; y) = \tau_1\|z - \beta\|_2^2 + \tau_2\|y - z\|_2^2/2$, with $\beta, z, y \in \mathbb{R}^d$, and $\tau_1, \tau_2 > 0$. We can immediately see that partial minimization of $h(\beta, z; y)$ over z yields $z = (\tau_1\beta + \tau_2 y)/(\tau_1 + \tau_2)$, a convex combination of y and β. However, the Gibbs posterior $\exp\{-\lambda h(\beta, z; y)\}$ instead concentrates at $(\hat{\beta}, \hat{z}) : \hat{\beta} = \hat{z} = y$ for large values of λ.

4.2 Method

4.2.1 Gradient-bridged Posteriors Via First-order Optimality

Throughout this article, we focus on differentiable $h(\beta, z; y)$. In this section, we first consider the case where the feasible region of h is an open and convex set $\mathcal{Z} \subseteq \mathbb{R}^d$, and the minimizer $\arg\min_z h(\beta, z; y)$ exists and satisfies the first-order optimality condition:

$$\hat{z}_\beta = \arg\min h(\beta, z; y) \text{ if and only if } \nabla_z h(\beta, \hat{z}_\beta; y) = 0. \tag{4-2}$$

Since h also depends on β and y, we assume the above condition holds almost surely with respect to the distribution of β for all observed data y.

We propose the following posterior distribution:

$$\begin{aligned} &\Pi(\beta, z \mid y) \propto L(y, z; \beta)\pi_0(\beta), \\ &L(y, z; \beta) \propto g(y; \beta, z) \exp\left\{-\lambda\|\nabla_z h(\beta, z; y)\|_2^2\right\}, \end{aligned} \tag{4-3}$$

where g is a kernel, and the hyper-parameter $\lambda > 0$ controls the degree that z concentrates around $\hat{z}_\beta$. We see that (4-3) promotes the gradient stationarity condition in (4-2) via regularization, and can be viewed as a relaxation of the constraint that $\nabla_z h = 0$ (Duan et al., 2020). We follow Presman and Xu (2023) and choose a Gaussian-type kernel incorporating the squared norm $\|\nabla_z h(\beta, z; y)\|_2^2$, a choice which will facilitate gradient-based posterior computation downstream. In particular, its derivative with respect to (β, z) is $2[\nabla^2_{z\beta} h(\beta, z; y), \nabla^2_{zz} h(\beta, z; y)]^{\mathrm{T}}[\nabla_z h(\beta, z; y)]$, which remains numerically stable when $\|\nabla_z h(\beta, z; y)\| \approx 0$.

Conceptually, $L(y; \beta) = \int L(y, z; \beta) \mathrm{d}z$ can be interpreted as the marginal likelihood. From this lens, z is a latent variable defined under the likelihood and does not require a prior. We refer to (4-3) as the *gradient-bridged posterior*, and $\exp\{-\lambda \|\nabla_z h(\beta, z; y)\|_2^2\}$ as the shrinkage kernel.

At a large but finite λ, we effectively create a concentration of z near its conditional optimal, that is, a ball $\mathbb{B}_\beta(\epsilon) = \{z : \|\nabla_z h(\beta, z; y)\|_2 \leq \epsilon\}$ has a high posterior probability for some $\epsilon > 0$ given β. Indeed, as $\lambda \to \infty$, we have $z \to \hat{z}_\beta$ due to $\|\nabla_z h(\beta, z; y)\|_2 \to 0$, recovering the bridged posterior (4-1). The case of convex h lends more intuition: by the first-order characterization of convexity, we know for any $z \in \mathbb{B}_\beta(\epsilon)$, the optimality gap is bounded above by

$$h(\beta, z; y) - h(\beta, z_\beta; y) \leq (z - \hat{z}_\beta)^{\mathrm{T}} \nabla_z h(\beta, z; y) \leq \epsilon \|z - \hat{z}_\beta\|_2.$$

Further, if h is μ-strongly convex with Lipschitz gradient, then $h(\beta, z; y) - h(\beta, \hat{z}_\beta; y) \leq 2\epsilon^2/\mu$. We will develop an interesting characterization on the set $\mathbb{B}_\beta(\epsilon)$ in Section 2.3. To be illustrative for now, we first examine a simple example related to the normal means problem.

Example 4.1. The normal means problem considers $y_i \sim \mathrm{No}(z_i, \tau)$, $z_i \sim \mathrm{No}(0, \beta)$ independently for $i = 1, \ldots, n$, with parameter $\beta > 0$ and with $\tau > 0$ a known constant. Due to the large number of z_i, one may think about reducing its variability via minimizing a loss $h(\beta, z; y) = \|z - y\|_2^2/(2\tau) + \|z\|^2/(2\beta)$, with partial derivative $\nabla_z h(\beta, z; y) = (z - y)/\tau + z/\beta$. In this toy example, the minimization problem is tractable with minimizer $\hat{z}_\beta = \{1 - \tau/(\tau + \beta)\}y$, which coincides with the James–Stein shrinkage estimator if $(\tau + \beta)$ is assigned the empirical

estimate $\|y\|_2^2/(n-2)$. Ignoring that we know the closed form solution for now, it is interesting to examine the gradient-bridged posterior under some finite λ:

$$\begin{aligned}\Pi(z \mid y, \beta) &\propto \exp\left(- \frac{\|z-y\|_2^2}{2\tau} - \frac{\|z\|^2}{2\beta} - \lambda \left\| \frac{z-y}{\tau} + \frac{z}{\beta} \right\|_2^2 \right), \\ &\propto \exp\left\{ - \frac{\|z-\hat{z}_\beta\|_2^2}{2(1/\tau+1/\beta)^{-1}} - \lambda \frac{\|z-\hat{z}_\beta\|^2}{(1/\tau+1/\beta)^{-2}} \right\}.\end{aligned}$$

We see that as the degree of relaxation decreases when λ increases from 0 to ∞, the distribution of z gradually changes from $\mathrm{No}[\hat{z}_\beta, (1/\tau+1/\beta)^{-1}]$ to a point mass at the solution $\hat{z}_\beta$.

Remark 4.1. We clarify that though first-order optimality is often used to characterize the solution of convex problems, the scope of our method applies beyond convex h. First, the condition (4-2) applies broadly to some non-convex functions. For example, the mode of the inverse gamma distribution is associated with the loss $f(z) = \alpha \log z + \beta/z$, which is pseudo-convex but non-convex. Second, for non-convex problems where (4-2) does not hold, we can often find its Lagrange dual objective function, which is always concave and whose maximizer determines $\hat{z}_\beta$ (under suitable conditions). For those problems, we can form (4-3) using the dual form of h. Further details are discussed in Section 4.2.5.

An immediate advantage of the gradient-bridged posterior (4-3) over its exact counterpart under (4-1) lies in its associated computation. By relaxing the equality constraint, we avoid having to solve a constrained optimization problem within a Bayesian procedure. Instead, one can directly apply canonical Markov chain Monte Carlo algorithms (such as random-walk Metropolis or Hamiltonian Monte Carlo), as well as alternatives such as variational inference, to estimate the posterior. Details on posterior sampling are discussed in Section 3.

4.2.2 Adjustment of Loss Function for Boundary Optimum

We now extend the applicable range of (4-3) to allow for cases where the feasible region $\mathcal{Z}$ may not be open. Many problems have a convex but non-open $\mathcal{Z}$, in which the optimal solution may fall on the boundary. For example, in the linear programming problem where $\mathcal{Z} = \{z : Mz \le c\}$ for some matrix M and vector c, one often has optimal $M_j^{\mathrm{T}} z = c_j$ for one or

more indices j. In such cases, with $\tilde{h}$ the loss function, the first order optimality condition for $\hat{z}$ becomes $(z-\hat{z})^{\mathrm{T}}\nabla_z \tilde{h}(\beta, \hat{z}; y) \geq 0$ for any $z \in \mathcal{Z}$, rendering the shrinkage kernel in (4-3) unhelpful.

We address these cases with an interior point approach. Suppose we have the original loss function $\tilde{h}$ with boundary optimum. We adjust $\tilde{h}$ with log-barrier function

$$h(\beta, z; y) = \tilde{h}(\beta, z; y) - \frac{1}{t}\sum_{j=1}^{m} \log[r_j(z)],$$

where the $r_j(z)$ correspond to all inequality constraints $r_j(z) \geq 0$ defining $\mathcal{Z}$, with hyperparameter $t > 0$. Since $-\log\{r_j(z)\} \to \infty$ as $r_j(z) \to 0$, the minimizer $\arg\min_z h(\beta, z; y)$ must shift from the boundary of $\mathcal{Z}$ into its interior. Therefore, $\hat{z}_\beta = \arg\min h(\beta, z; y)$ is equivalent to

$$\nabla_z h(\beta, z; y) = \nabla_z \tilde{h}(\beta, z; y) - \frac{1}{t}\sum_{j=1}^{m} \frac{\nabla_z r_j(z)}{r_j(z)} = 0, \tag{4-4}$$

enabling us to use $\exp\{-\lambda\|\nabla_z h(\beta, z; y)\|_2^2\}$ for use in the gradient-bridged posterior. Typically, one chooses a relatively large t, so that $\arg\min \tilde{h}(\beta, z; y) \approx \arg\min h(\beta, z; y)$; in this article, we use $t = 1000$.

We now present our first example of a Bayesian model defined with an implicit function that lacks a closed-form solution. This problem is motivated by applications in flow network data.

Example 4.2. A flow network consists of weighted values z_{ij} on a directed network $G = (V, E)$ with V the set of nodes, and E the set of uni-directed edges $(i \to j)$. The problem of finding the maximum flow possible from a designated source node s to a sink node t, subject to edge capacities is a well-studied linear program:

$$\begin{aligned}
\text{maximize} \quad & f(z) = \sum_{j:\,(s\to j)\in E} z_{sj} \\
\text{subject to} \quad & \sum_{j:\,(i\to j)\in E} z_{ij} - \sum_{k:\,(k\to i)\in E} z_{ji} = 0, \quad \forall i \in V \setminus \{s, t\}, \\
& 0 \leq z_{ij} \leq \beta_{ij}, \quad \forall (i \to j) \in E.
\end{aligned}$$

In the above, the objective function is the total amount of flow entering the network. The first equality constraint is the flow conservation: no flow is lost or created at any node except at the source and sink. The second inequality incorporates the *edge capacities* β_{ij} together with non-negativity of flows. It is clear that the solution $\arg\max_z f(z)$ depends on the value of β, and is denoted by $\hat{z}_\beta$. For networks with more than one source, an auxiliary *super source* can be introduced to cast the problem equivalently; a similar treatment can be done for networks with more than one sink.

In fields such as transportation science, it is often reasonable to assume that the flow network has nearly reached its maximum during a particular time period (such as rush hour). On the other hand, the observed flows y_{ij} are realistically a noisy measurement of z_{ij}, and the true capacities β_{ij} may differ from a *designed capacity* c_{ij}. For instance, if c_{ij} are lanes, congestion or closure may cause the number of usable lanes $\beta_{ij} < c_{ij}$, so that β_{ij} may be better treated as unknown. Accounting for these considerations leads to a statistical problem; we can consider the likelihood

$$L(y, z; \beta) \propto \exp\left\{-\frac{\sum_{(i\to j)\in E}(y_{ij} - z_{ij})^2}{2\sigma_y^2} - \frac{\sum_{(i\to j)\in E}(c_{ij} - \beta_{ij})^2}{2\sigma_c^2}\right\} \exp\left\{-\lambda \|\nabla_z h(\beta, z; y)\|_2^2\right\}, \quad (4\text{-}5)$$

where it is natural to define a subproblem loss $h(\beta, z; y) = -f(z) - (1/t)\sum_{l=1}^m \log[r_l(z)]$ as the negative objective in addition to log-barrier terms. Here $r_l(z) = z_{ij}$ to ensure $z_{ij} > 0$, and $r_l(z) = \beta_{ij} - z_{ij}$ to ensure $z_{ij} < \beta_{ij}$, for all edges $(i \to j) \in E$. To meet the equality constraint, we use a simple reparameterization: for each $i \in V \setminus \{s, t\}$, we choose one inflow edge $k^* \to i$ and parameterize $z_{k^*i} = (\sum_{j:\,(i\to j)\in E} z_{ij}) - (\sum_{k\neq k^*:\,(k\to i)\in E} z_{ki})$ as a transformation of other flows associated with i. We will revisit this example in our numerical study in Section 5.

4.2.3 Implicit Function Manifold and Its Relaxation

We now further characterize the posterior defined under the implicit function and its relaxation. In this section, we restrict our scope to the class of models where h is twice continuously differentiable in z, and $\nabla_z h(\beta, z; y)$ is continuously differentiable in β. Within this class, the function $\nabla_z h(\beta, z; y)$ can be nicely characterized by the implicit function theorem, so that $\{(\beta, \hat{z}_\beta) : \beta \in \Theta\}$ is a smooth manifold that we will call the *implicit function manifold*.

We first briefly review the implicit function theorem using our notation. Let (β_0, z_0) be a point satisfying $\nabla_z h(\beta_0, z_0; y) = 0$, and assume that the Hessian matrix of h with respect to z, $\nabla^2_{zz} h(\beta_0, z_0; y)$ is invertible. Then, by the implicit function theorem, there exists a neighborhood of β_0 in which we can define a continuously differentiable function

$$\hat{z}_\beta := \zeta(\beta) : \nabla_z h(\beta, \zeta(\beta); y) = 0.$$

Two useful results follow. First, the function $\zeta(\beta)$ and hence $\hat{z}_\beta$ is unique. Second, the change rate of $\hat{z}_\beta$ with respect to β is given by

$$\frac{\partial \hat{z}_\beta}{\partial \beta} = -\left\{\nabla^2_{zz} h(\beta, \zeta(\beta); y)\right\}^{-1} \nabla^2_{z\beta} h(\beta, \zeta(\beta); y),$$

where $\nabla^2_{z\beta} h(\beta, z; y)$ denotes the matrix of mixed partial derivatives of h with respect to z and β.

Now, recall the gradient-bridged posterior allows $\|\nabla_z h(\beta, z; y)\|$ to be small but non-zero, and hence the high-density posterior region of (β, z) is a continuous relaxation from the implicit function manifold. Naturally, one may wonder how much relaxation is induced at some $\|\nabla_z h(\beta, z; y)\| \leq \epsilon$. The following theorem provides a bound on the amount of relaxation.

Theorem 4-1. *At any β with $\nabla^2_{zz} h(\beta, z_0; y)$ invertible and $z_0 = \hat{z}_\beta$, there exists a Hessian-based neighborhood of z_0:*

$$\mathbb{B}(z_0, k; \beta) = \{z : \|I - \nabla^2_{zz} h(\beta, z_0; y)^{-1} \nabla^2_{zz} h(\beta, z; y)\|_{op} \leq k < 1\}.$$

Further, for any $z \in \mathbb{B}(z_0, k; \beta)$, if $\|\nabla_z h(\beta, z; y)\| \leq \epsilon$, we have

$$\|z - z_0\| \leq \frac{\epsilon}{(1-k)\lambda_{\min}\{\nabla^2_{zz} h(\beta, z_0; y)\}},$$

where $\lambda_{\min}$ denotes the smallest eigenvalue and $\|\cdot\|_{op}$ denotes the operator norm.

Remark 4.2. In the above we started with a Hessian-based neighborhood of z_0, which should not be confused with the Euclidean neighborhood of z_0. This Hessian-based neighborhood can be quite large when the Hessian is slowly changing as we move away from z_0.

Roughly speaking, the above theorem shows that $\|z - \hat{z}_\beta\| = O\{\|\nabla_z h(\beta, z; y)\|\}$, justifying that controlling the gradient norm of h governing the implicit function manifold is an effective alternative to solving for $\hat{z}_\beta$.

4.2.4 Suitability for Canonical Bayesian Inference

The likelihood in the gradient-bridged posterior is designed to be tractable (up to a constant). As a result, it is convenient to apply canonical Bayesian inference procedures in addition to parameter estimation under a fixed sample size n, such as handling missing values or model selection via Bayes factor.

To illustrate, consider the missing data setting where $y_D = \{y_i : i \in D\}$ denotes the set of observed data, and $y_M = \{y_i : i \in M\}$ those with missing values, so that the index sets $D \cup M = (1, \ldots, n) =: [n]$. Under the gradient-bridged setup, the joint distribution

$$\Pi(y_M, \beta \mid y_D, z) \propto \pi_0(\beta) g(\{y_D, y_M\}; \beta, z) \exp\left\{-\lambda \|\nabla_z h(\beta, z; \{y_D, y_M\})\|_2^2\right\},$$

is amenable to simultaneous imputation of y_M and estimation of β using the same approaches that one would take in a canonical Bayesian model, for instance via data-augmented Markov chain Monte Carlo (Tanner and Wong, 1987; Van Dyk and Meng, 2001). As one can imagine, the imputation would become considerably more complex if we replace z with $\hat{z}_\beta$, due to a constraint resulting from the deterministic relationship between $\hat{z}_\beta$ and y_D, and in turn between $\hat{z}_\beta$ and y_M.

This is just one example of how our framework inherits straightforward inference under fixed sample size n. Extra caution should be exercised for inference tasks under changing n; prediction is one such prominent example. In particular, Bayesian prediction tasks typically assign the same form of likelihood (4-3) for both sample sizes n and $(n + m)$, with m the number of points to be predicted. A coherent prediction rule should ensure the marginalization property

$L(y_{[n]};\beta) = \int L(y_{[n+m]};\beta)\,\mathrm{d}(y_{n+1},\ldots,y_{n+m})$ holds, as exhibited by a sequential generative process.

Unfortunately, we do not expect this condition to hold in complete generality for gradient-bridged posterior models. This is a common limitation to loss-based generalized Bayes approaches (Natarajan et al., 2024; Rigon et al., 2023b). Nonetheless, we can identify salient sufficient conditions that ensure the marginalization condition, such as separability of the loss over i. To be concrete, if (i) each y_i has a corresponding latent z_i so that the loss $h(\beta, z_{[n]}; y_{[n]}) = \sum_{i=1}^{n} s(z_i, y_i; \beta)$, and (ii) the marginalization condition

$$\int \frac{g(y_{[n+1]};\beta, z_{[n+1]})}{g(y_{[n]};\beta, z_{[n]})} \exp\{-\lambda\|\nabla_{z_{n+1}} s(z_{n+1}, y_{n+1};\beta)\|_2^2\}\,\mathrm{d}(y_{n+1}, z_{n+1}) \tag{4-6}$$

does not depend on β, then it follows that $L(y_{[n]}, z_{[n]};\beta) = \int L(y_{[n+1]}, z_{[n+1]};\beta)\,\mathrm{d}(y_{n+1}, z_{n+1})$. Integrating both sides over $z_{[n]}$ reveals the marginalization property.

Remark 4.3. In the case that (4-6) depends on β, which we abbreviate as some function $m(\beta)$, one can calibrate the function g using a term involving $1/m(\beta)$. To illustrate, in Example 4.1, the normalizing term of $g(y_{[n]};\beta, z_{[n]})$ is $\beta^{-n/2}$. If we instead model by replacing this with $\{m(\beta)\}^{-n}$, $m(\beta) = \tau\beta\{\tau\beta + 2\lambda(\tau+\beta)\}^{-1/2}$, then (4-6) will be equal to a constant in terms of β.

4.2.5 Non-convex Problems and Duality

We now consider (4-3) in the non-convex setting, which includes the case where the function h is non-convex as well as the possibility that $\mathcal{Z}$ is a non-convex set. Though first-order stationarity no longer suffices for optimality, progress can be made when the loss enjoys a manageable dual form. Recall we seek to minimize what we now refer to as the *primal* objective or loss

$$f(z) := h(\beta, z; y),$$

over a feasible region $\mathcal{Z} = \{z : r(z) \leq 0,\ s(z) = 0\}$.

The functions $r : \mathbb{R}^d \to \mathbb{R}^{m_1}$ and $s : \mathbb{R}^d \to \mathbb{R}^{m_2}$ may depend on y and β as well. Using Lagrange multipliers $\gamma \in \{(\gamma_1, \gamma_2) : \gamma_1 \in \mathbb{R}^{m_1}, \gamma_1 \geq 0, \gamma_2 \in \mathbb{R}^{m_2}\} := \Gamma$, we have the dual function

(that we will refer to as the *dual* objective):

$$f^\dagger(\gamma) = \min_{z \in \mathbb{R}^d} \{ f(z) + \gamma_1^{\mathrm{T}} r(z) + \gamma_2^{\mathrm{T}} s(z) \}.$$

Here, the minimization of z is now unconstrained over $\mathbb{R}^d$. To provide some background, for any feasible z, we have $\gamma_1^{\mathrm{T}} r(z) \leq 0$ and $\gamma_2^{\mathrm{T}} s(z) = 0$ for any $\gamma \in \Gamma$. It always holds that $f^\dagger(\gamma) \leq \min_{z \in \mathcal{Z}} f(z)$: this is known as *weak duality*, which applies for any primal problem regardless of whether it is convex. The inequality implies two important consequences. First, the optimal dual objective $\max_\gamma f^\dagger(\gamma)$ serves as a lower-bound for the optimal primal loss. As a result, the quantity $\{f(z) - \max_\gamma f^\dagger(\gamma)\}$ provides a practical upper-bound estimate on $\{f(z) - \min_{z \in \mathcal{Z}} f(z)\}$ for any value z. Second, if we have $f(z^*) = f^\dagger(\gamma)$, for some pair $z^* \in \mathcal{Z}$ and $\gamma \in \Gamma$, then z^* must be the optimal primal solution. This latter scenario is also known as *strong duality*.

Importantly, $f^\dagger(\gamma)$ is always concave, as it comprises an affine function of γ composed with a pointwise infimum over z. The concavity of $f^\dagger(\gamma)$ means that we can find the optimal γ via the first order optimality (4-2), or approximately in the interior of Γ via (4-4), provided $f^\dagger$ is tractable and differentiable. Using $h^\dagger(\beta, \gamma; y) = f^\dagger(\gamma)$ or its log-barrier modification in Section 4.2.2, the shrinkage kernel $\exp\{-\lambda \|\nabla_\gamma h^\dagger(\beta, \gamma; y)\|_2^2\}$ is useful toward a relaxed solution of $\arg\max_\gamma f^\dagger(\gamma)$.

Remark 4.4. For simplicity, we will focus on primal f that enjoys strong duality $\max_{\gamma \in \Gamma} f^\dagger(\gamma) = \min_{z \in \mathcal{Z}} f(z)$, and hence we do not need to worry about the gap between $f(z)$ and $f^\dagger(\gamma)$.

To illustrate this approach in action, we consider the orthogonal Procrustes problem (Gower, 1975), which will also play a role in our case study. The primal problem here is non-convex, but we show how the form of the dual objective enables us to construct a shrinkage kernel using (4-4).

Example 4.3. The orthogonal Procrustes problem aims to solve

$$\min_R \ f(R) = \|Ry - \beta\|_F^2, \quad \text{subject to } R^{\mathrm{T}} R = I_p,$$

for two matrices $y \in \mathbb{R}^{p\times m}$ and $\beta \in \mathbb{R}^{p\times m}$, where R can include rotations and/or reflections of each column of y. The primal problem is not convex due to the non-convexity of the orthonormal space, although owing to von Neumann's trace inequality, the primal solution given β is tractable using the singular value decomposition $\beta y^{\mathrm{T}} = U\Sigma V^{\mathrm{T}}$ and $\hat{R} = UV^{\mathrm{T}}$. On the other hand, when β is not known (or contains missing values) as in a Bayesian model, simultaneous updating of R and β becomes challenging.

To derive the dual form, we introduce a symmetric multiplier $\gamma \in \mathbb{R}^{p\times p}$ in the Lagrangian:

$$\mathcal{L}(R,\gamma) = \|Ry - \beta\|_F^2 + \mathrm{tr}\{\gamma(R^{\mathrm{T}}R - I)\},$$

which is bounded from below whenever $(\gamma + yy^{\mathrm{T}}) \succ 0$ and is minimized at $\hat{R} = \beta y^{\mathrm{T}}(\gamma + yy^{\mathrm{T}})^{-1}$. The finite dual function and its derivative are given by

$$\begin{aligned} f^{\dagger}(\gamma) &= \mathrm{tr}(\beta^{\mathrm{T}}\beta) - \mathrm{tr}(\gamma) - \mathrm{tr}\{y\beta^{\mathrm{T}}\beta y^{\mathrm{T}}(\gamma + yy^{\mathrm{T}})^{-1}\}, \\ \nabla_{\gamma} f^{\dagger}(\gamma) &= -I + (yy^{\mathrm{T}} + \gamma)^{-1}\,(y\beta^{\mathrm{T}}\beta y^{\mathrm{T}})\,(yy^{\mathrm{T}} + \gamma)^{-1} \;=\; -I + \hat{R}^{\mathrm{T}}\hat{R}. \end{aligned} \tag{4-7}$$

Now, from the first-order optimality condition on the dual $\nabla_{\gamma} f^{\dagger}(\gamma) = 0$, rearranging (4-7) shows that the solution satisfies the constraint $\hat{R}^{\mathrm{T}}\hat{R} = I_p$. That is $\hat{R}$ achieves primal feasibility, and strong duality holds: $\max f^{\dagger}(\gamma) = \min_{R:R^{\mathrm{T}}R=I} f(R)$. In practice, we further may avoid explicit matrix inversion by parameterizing a lower-triangular matrix W with positive diagonal entries, taking $(yy^{\mathrm{T}} + \gamma)^{-1} = WW^{\mathrm{T}}$ which is always positive definite. Doing so, we arrive at the shrinkage kernel for the orthogonal Procrustes problem

$$\exp\{-\lambda\|WW^{\mathrm{T}}(y\beta^{\mathrm{T}}\beta y^{\mathrm{T}})WW^{\mathrm{T}} - I\|_F^2\} \;:=\; \exp\{-\lambda\|R^{\mathrm{T}}R - I\|_F^2\} \tag{4-8}$$

with primal variable $R = \beta y^{\mathrm{T}}(WW^{\mathrm{T}})$.

4.3 Hamiltonian Monte Carlo Near Implicit Function Manifold

The gradient-bridged posterior is designed for efficient computing, and benefits from the large body of work on gradient-based samplers and widely adopted auto-differentiation algorithms. We first review the Hamiltonian Monte Carlo algorithm, focusing on how the choice of the mass matrix can improve its efficiency. For simplicity, we abbreviate the parameter by $\theta = (z, \beta)$.

Hamiltonian Monte Carlo augments the parameters θ with an auxiliary momentum variable p, and makes use of Hamiltonian dynamics in the augmented space to generate informed proposals to new states. Denote the joint density

$$\Pi(\theta, p \mid y) \propto \Pi(\theta \mid y) \exp\left(-\frac{1}{2} p^{\mathrm{T}} M^{-1} p\right),$$

where M is a positive-definite mass matrix. A new proposal is generated by first sampling the momentum $p \sim \mathrm{No}(0, M)$, and then simulating Hamiltonian dynamics, typically using a numerical integrator. The leapfrog algorithm is a prevalent such approach: given the current state (θ_t, p_t), we perform L steps of the following, with step-size ϵ, to propose (θ^*, p^*):

(B1) Half-step for momentum: $p(t + \epsilon/2) = p_t + (\epsilon/2)\nabla_\theta \log \Pi(\theta_t \mid y)$.

(B2) Full-step for position: $\theta(t + \epsilon) = \theta_t + \epsilon\, M^{-1} p(t + \epsilon/2)$.

(B3) Another half-step for momentum: update the momentum using the new position: $p(t + \epsilon) = p(t + \epsilon/2) + (\epsilon/2)\nabla_\theta \log \Pi(\theta(t + \epsilon) \mid y)$.

Since the leapfrog integrator is reversible and volume preserving (Neal, 2011), the acceptance ratio calculation is straightforward: θ^* is accepted with probability

$$\alpha\big\{(\theta_t, p_t), (\theta^*, p^*)\big\} = \min\Big[1, \exp\big\{-H(\theta^*, p^*) + H(\theta_t, p_t)\big\}\Big],$$

where the Hamiltonian is given by $H(\theta, p) = -\log \Pi(\theta \mid y) + \frac{1}{2} p^{\mathrm{T}} M^{-1} p$. By simulating the Hamiltonian dynamics, the algorithm is able to propose distant moves with high acceptance rates,

especially when the mass matrix M is chosen judiciously to match the local geometry of the target distribution.

An extension of Hamiltonian Monte Carlo, known as the No-U-Turn Sampler, adaptively determines the number of leapfrog steps to avoid inefficient trajectories that double back on themselves, thereby eliminating the need to pre-specify L (Hoffman et al., 2014). We adopt the No-U-Turn Sampler in this article. We now focus on the choice of M specifically for the gradient-bridged posterior. Combining the discrete moves in a leapfrog step, we have

$$\theta(t+\epsilon) = \theta_t + M^{-1}\left\{\epsilon\, p_t + \, (\epsilon^2/2)\nabla_\theta \log \Pi(\theta_t \mid y)\right\}.$$

For large λ, we have small $\|\nabla_z h(\beta, z; y)\|_2^2 \approx 0$ with high probability. This suggests that $\theta(t)$ lies near the manifold, and thus after a small time ϵ, $\theta(t+\epsilon)$ should tend to stay near the manifold as well due to conservation of energy in the exact Hamiltonian dynamics. Since the leapfrog algorithm provides only an approximation to the exact Hamiltonian dynamics, it is sensible to choose M so to avoid an abrupt change of $\|\nabla_z h(\beta, z; y)\|_2^2$.

To this end, we consider linearizing the gradient. Let $(\Delta\beta, \Delta z) = \theta(t+\epsilon) - \theta(t)$: then the gradient at $\theta(t+\epsilon)$ can be approximated by

$$\nabla_z h(\beta + \Delta\beta, z + \Delta z; y) \approx \nabla_z h(\beta, z; y) + G_{\beta,z}^{\mathrm{T}}(\Delta\beta, \Delta z),$$

where $G_{\beta,z} = [\nabla^2_{z\beta} h(\beta, z; y), \nabla^2_{zz} h(\beta, z; y)]$ is the submatrix of the Hessian containing the partials with respect to both (β, z). Since having $G_{\beta,z}^{\mathrm{T}}(\Delta\beta, \Delta z) = 0$ would be useful for reducing changes in $\nabla_z h(\beta + \Delta\beta, z + \Delta z; y)$, this leads to a natural choice of M^{-1} as a null-space projection $I - G_{\beta,z}(G_{\beta,z}^{\mathrm{T}} G_{\beta,z})^{-1} G_{\beta,z}^{\mathrm{T}}$.

On the other hand, the leapfrog integrator would lose reversibility if M^{-1} varies with the parameters, and the momentum distribution would become degenerate if M^{-1} is rank-deficient.

With these practical considerations in mind, we use

$$M^{-1} = (1+\tau)I - \hat{G}_{\beta,z}(\hat{G}_{\beta,z}^{\mathrm{T}}\hat{G}_{\beta,z})^{-1}\hat{G}_{\beta,z}^{\mathrm{T}}, \tag{4-9}$$

where $\tau > 0$ is a small constant chosen to make M^{-1} strictly positive definite (we use $\tau = 10^{-3}$ in this article), and $\hat{G}_{\beta,z}$ is an estimate of $E(G_{\beta,z})$, the sub-matrix of the negative Fisher information with expectation taken over $\Pi(\beta, z \mid y)$. In practice, to use its sample counterpart $\hat{G}_{\beta,z}$, one can take the Hessian matrix evaluated at the posterior mode of β and the minimizer z, or the observed Fisher information matrix by averaging samples $G_{\beta,z}$ over the burn-in period of the Markov chain Monte Carlo. In this article, we use the Hessian at the posterior mode.

Remark 4.5. The constant-value specification of (4-9) is suitable for problems where the Hessian changes relatively slowly in the high posterior density region. For problems with $G_{\beta,z}$ rapidly changing, the linearization may provide a poor approximation, and thus $G_{\beta,z}^{\mathrm{T}}M^{-1}$ may be further from zero, rendering (4-9) ineffective. Broadly speaking, rapidly changing Hessians pose a challenge for existing Hamiltonian Monte Carlo algorithms. One remedy involves location-dependent specifications of $M(\theta)$ becomes necessary—it tends to require intensive computation as an alternative to the leapfrog integrator. One can find discussion of those integrators in Girolami and Calderhead (2011); Lan et al. (2015); Nishimura and Dunson (2017). As an alternative, variational inference can be used to bypass the challenges, at the cost of additional posterior approximation.

4.4 Asymptotic Theory

We now study the asymptotic properties of the proposed method, establishing normality in the large sample limit in a Bernstein–von Mises sense. In Bayesian analysis, the Bernstein–von Mises theorem asserts that, under mild regularity conditions, the posterior distribution of the centered and scaled maximum likelihood estimator $\sqrt{n}(\beta - \hat{\beta}_n)$ converges as sample size n increases to a normal distribution centered at 0, with covariance matrix equal to the inverse Fisher information. In models involving a nuisance parameter z, the primary focus often lies in the marginal posterior distribution of the parameter of interest, β (Berger et al., 1999; Severini, 2007).

This marginal posterior is related to the complete posterior by integrating out the nuisance variable z: that is, $\Pi(\beta \mid y) \propto \int_{\mathcal{H}} L(y, z; \beta)\pi_0(\beta)\,\mathrm{d}z$. The domain of integration $z \in \mathcal{H}$ need not be Euclidean, in which case the Bayesian model is a semi-parametric one, so that the Bernstein–von Mises theorem provides a feasible way to characterize the asymptotic posterior when the parameter z lies in any Hilbert space.

Bickel and Kleijn (2012) established a Bernstein–von Mises theorem for the marginal posterior under the existence of the least favorable model, provided that the posterior concentrates around this model. Building upon this foundation, we establish a Bernstein–von Mises result for the gradient-bridged posterior. Our approach hinges on the condition that the posterior distribution concentrates around the optimal value of the parameter z, as determined by the the first-order optimality condition. This is satisfied due to the shrinkage kernel.

We assume that $\beta \in \Theta \subset \mathbb{R}^d$, and $z \in \mathcal{H}$ lies in a Hilbert space. Let β_0, z_0 be the fixed, ground-truth values. The prior $\pi_0(\beta)$ is assumed to be continuous at β_0 with $\pi_0(\beta_0) > 0$. We define the log-likelihood as $l_n(\beta, z) = \log L_n(\beta, z; y)$, where the subscript indicates the sample size. We use $\|\cdot\|$ to denote the Euclidean–Frobenius norm. We denote the generative distribution of the data by $P_{\beta,z}$ with parameters β and z. We use $H(P, Q)$ to denote the Hellinger distance between two probability measure P and Q, and define a metric d_H on the space of z by $d_H(z_1, z_2; \beta) = H(P_{\beta,z_1}, P_{\beta,z_2})$. We denote the unique minimizer of the gradient-bridge function $h(\beta, z; y)$ by $\hat{z}(\beta) := \hat{z}_\beta$. We now state the sufficient conditions for establishing the Bernstein–von Mises theorem.

Assumption 4.1. *There exists a decreasing sequence $\{\rho_n\}$ converging to 0 such that for every bounded sequence h_n, with $\beta_n = \hat{\beta}_n + h_n/\sqrt{n}$,*

$$\Pi[d_H\{z, \hat{z}(\beta); \beta\} > \rho_n \mid \beta = \beta_n; y] \xrightarrow[n\to\infty]{P_{\beta_0,z_0}} 0\,,$$

where the above denotes convergence in probablity with respect to P_{β_0,z_0}.

Assumption 4.2. *There exists a symmetric $H_n(z)$ such that for $\beta_n = \hat{\beta}_n + h_n/\sqrt{n}$,*

$$l_n(\beta_n, z) = l_n\{\hat{\beta}_n, \hat{z}(\hat{\beta}_n)\} - \frac{1}{2} h_n^{\mathrm{T}} H_n(z) h_n + r_n(h_n, z)$$

holds for all z satisfying $d_H\{z, \hat{z}(\beta_n); \beta_n\} < \delta$, with remainder term $r_n(h_n, z) = o_{P_{\beta_0, z_0}}(1)$, and $H_n\{\hat{z}(\hat{\beta}_n)\} \to H_0$.

Assumption 4.3. *There exists a constant C such that for $\beta_n = \hat{\beta}_n + h_n/\sqrt{n}$,*

$$\sup_{z: d_H\{z, \hat{z}(\beta_n); \beta_n\} < \rho_n} \int L_n(\beta_n, z; y) \, \mathrm{d}P_{\beta_0, z_0}(y) \leq C.$$

We refer to the set of models when $z = \hat{z}(\beta)$ as the constrained favorable submodels. Assumption 4.1 posits that the conditional posterior concentrates around the constrained favorable submodels. One way for this assumption to be satisfied is by letting $\lambda = \lambda_n \to \infty$. Assumption 4.2 is the stochastic local asymptotic normality of the likelihood around the constrained favorable submodels, which is standard for proving Bernstein–von Mises results. In parametric settings when z belongs to a Euclidean space, the condition can often be proven using Taylor expansion, provided conditions on the derivatives of the likelihood function. Assumption 4.3 is the domination condition of the likelihood around the constrained favorable submodels.

Let the integrated likelihood $S_n(\beta) = \int_{\mathcal{H}} L_n(\beta, z; y) \, \mathrm{d}\Pi_0(z)$, and $s_n(\beta) = \log S_n(\beta)$. We have the following lemma for the locally asymptotically normal property of the integrated likelihood.

Lemma 4-1. *Under Assumption 4.1–4.3, for every bounded sequence h_n and $\beta_n = \hat{\beta}_n + h_n/\sqrt{n}$,*

$$s_n(\beta_n) = s_n(\hat{\beta}_n) - \frac{1}{2} h_n^{\mathrm{T}} H_0 h_n + o_{P_{\beta_0, z_0}}(1). \tag{4-10}$$

With the locally asymptotically normal condition for $s_n(\beta)$, we make the Bernstein–von Mises result statement.

Theorem 4-2. *Assume* (4-10) *holds with positive definite* H_0. *Suppose that the maximum likelihood estimator* $\hat{\beta}_n$ *exists and converges to* β_0 *when* $n \to \infty$; *there exists* $\delta > 0$ *such that for any sequence of positive numbers* $M_n \to \infty$,

$$p_{\beta_0, z_0}\left[\inf_{z \in \mathcal{H}} \inf_{\sqrt{n}\|\beta - \hat{\beta}_n\| \geq M_n} \left\{l_n(\hat{\beta}_n) - l_n(\beta)\right\} \geq \frac{\delta M_n^2}{n}\right] \xrightarrow[n \to \infty]{} 1.$$

Then letting π_n *be the density of* β *when* $\beta \sim \Pi_n(\beta \mid y)$, *we have*

$$\int_{B_\epsilon(\beta_0)} \pi_n(\beta)\, \mathrm{d}\beta \xrightarrow[n \to \infty]{P_{\beta_0, z_0}} 1 \text{ for all } \epsilon > 0.$$

Moreover, letting q_n *denote the density of* $\sqrt{n}(\beta - \hat{\beta}_n)$, *we have* $d_{TV}\left\{q_n, \mathcal{N}\left(0, H_0^{-1}\right)\right\} \to 0$ *in probability with respect to* P_{β_0, z_0} *as* $n \to \infty$.

4.5 Simulation Study

We empirically illustrate the gradient-bridged posterior with application to the flow network described in Example 4.2. The directed network used in the synthetic experiments is shown in Fig. 4-1. We use package `networkx` to generate a flow network, which consists of 7 nodes and 10 directed edges. Node 0 is the source and node 6 is the sink, and assign ground-truth edge capacities β_{ij}^0 depicted in Figure 4-1. Given this network, we begin by computing the optimal flow values z^0 under β^0 by solving the maximum flow optimization problem. Next, we simulate observed flow data y_{ij}^s for $s = 1, \ldots, 1000$, where each observation s represents a noisy measurement of the optimal flow. We simulate the designed capacity c_{ij} from $\mathrm{No}(\beta_{ij}^0, 0.5^2)$.

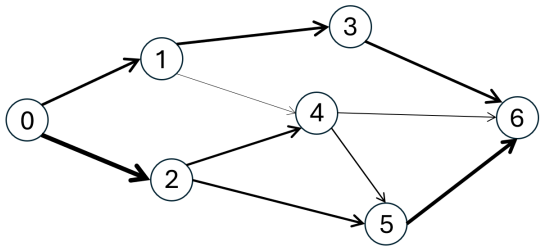
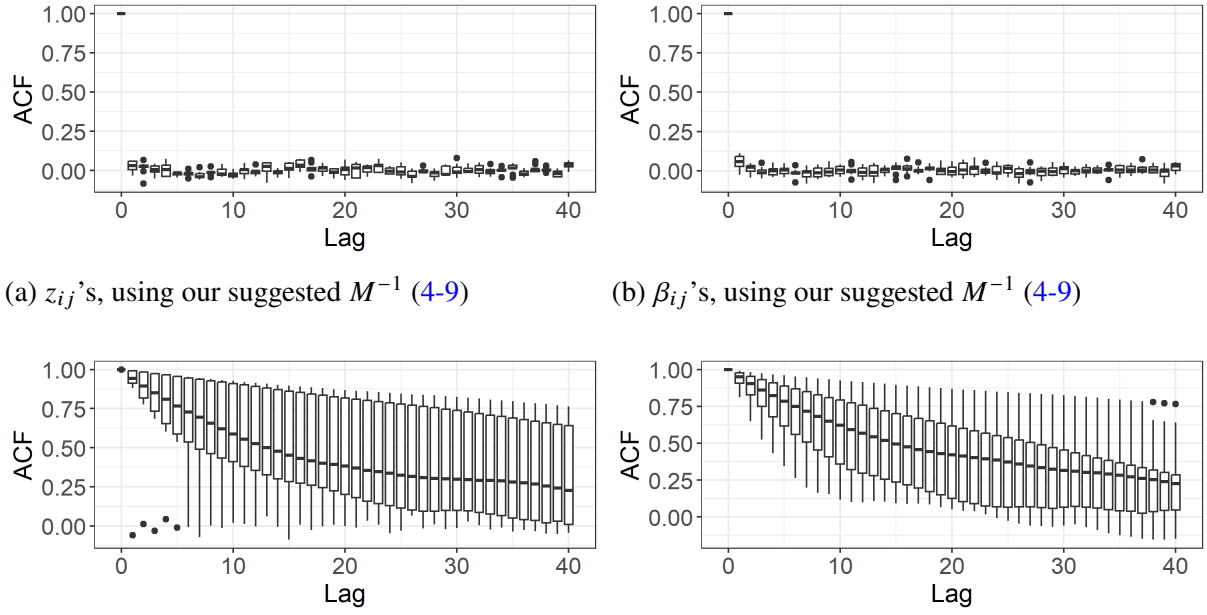


Figure 4-1. The directed network we use for the experiment on the maximum flow problem. The width of the edges is proportional to the magnitude of the optimal flow.

We assign independent half-normal $\mathrm{No}_+(0, 10^2)$ priors on the β_{ij}'s and z_{ij}'s, assign $\mathrm{Ga}^{-1}(2, 5)$ prior on the σ_y^2 and assign $\mathrm{Ga}^{-1}(5, 2)$ prior on σ_c^2. Using our suggested choice of M^{-1} matrix (4-9), we use the No-U-Turn sampler (Hoffman et al., 2014) to sample from the posterior distribution under (4-5) for 12,000 iterations. The first 2000 iterations are treated as burn-in, and we apply a thining factor of 10. Figure 4-2(a)–(b) shows the autocorrelation for all coordinates of z and β; its rapid decay indicates good mixing. In contrast, we also consider the No-U-Turn Sampler under the default choice of adaptive M^{-1}. It is visually clear from panels 4-2(c)–(d) that our choice of M^{-1} significantly improves performance in terms of mixing.



(a) z_{ij}'s, using our suggested M^{-1} (4-9)

(b) β_{ij}'s, using our suggested M^{-1} (4-9)

(c) z_{ij}'s, using default choice of M^{-1}

(d) β_{ij}'s, using default choice of M^{-1}

Figure 4-2. Autocorrelation of the Markov chain from No-U-Turn Samplers using different choices of the inverse mass matrix. Panels (a) and (b) use our suggested M^{-1} following (4-9). Panels (c) and (d) use the default choice of M^{-1}.

Figures 4-3 and 4-4 present the posterior distributions of the flow variables z and the capacity parameters β. An immediate takeaway is that empirically, the gradient-bridged posterior is able to recover the ground truth, assigning high posterior mass concentrated around the ground truth z^0 and β^0 values. To highlight the benefits of our proposed method, we compare performance to inference using existing Gibbs posterior approaches. A canonical Gibbs posterior,

corresponding to the loss (4-5) taking $\lambda = 0$, also results in a posterior that assigns mass covering the ground truth values. While estimates of z are quite similar to those under our approach, estimation of β_{ij} exhibits strikingly higher variability (Figure 4-4). To understand this phenomenon, recall that there are 1000 noisy flow measurements y^s, but only one observed measurement of the capacity c. The standard Gibbs posterior approach makes good use of y to estimate z, but fails to borrow information in the relationship between z and β. Instead, our shrinkage kernel exploits the relationship between β and z via the implicit function, making the posterior of β more accurate.

To be complete, we additionally consider a Gibbs posterior approach that attempts to account for the relationship between β and z via the loss-based kernel $\exp\{-\lambda h(\beta, z; y)\}$; that is, performing joint shrinkage in lieu of the proposed methodology allowing for *conditional* shrinkage via partial minimization. Empirically, we see that such joint regularization is a poor substitute for accommodating the desired sub-problem. Performance here is arguably even worse, with posteriors now exhibiting bias in both z as well as β. We see that penalizing $h(\beta, z; y)$ can lead to poor estimation and inflated posterior variance in both Figures 4-3 and 4-4, while the gradient-bridged posterior is free from this issue due to partial minimization over z.

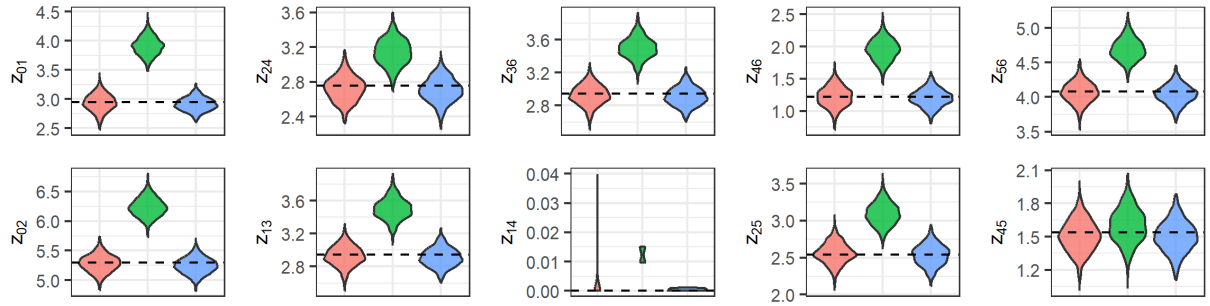


Figure 4-3. Flow value posteriors z_{ij} for each network edge from the Gibbs posterior with $\lambda = 0$ (red), the Gibbs posterior with shrinkage kernel $\exp\{-\lambda h(\beta, z; y)\}$ (green), and the gradient-bridged posterior (blue); horizontal dashed lines depict ground-truth values z^0_{ij}.

Additional simulation results related to the flow network modeling problem are detailed in the Supplement. There we also include another synthetic experiment on the latent quadratic

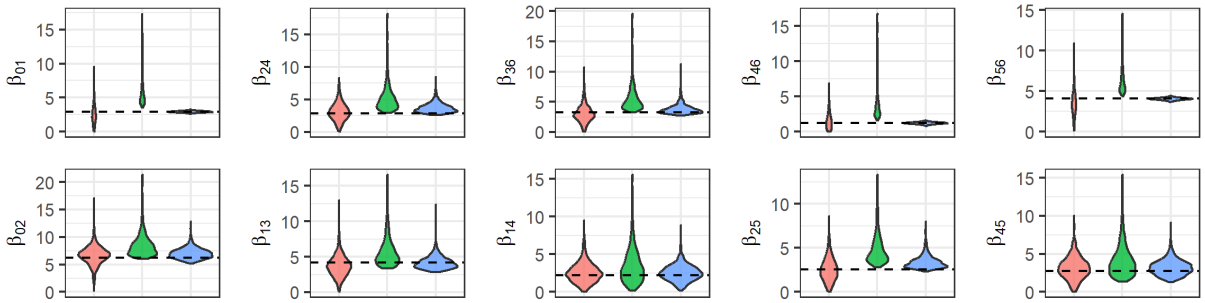


Figure 4-4. Capacity parameter posteriors β_{ij} for each network edge from the Gibbs posterior with $\lambda = 0$ (red), the Gibbs posterior with shrinkage kernel $\exp\{-\lambda h(\beta, z; y)\}$ (green), and the gradient-bridged posterior (blue); horizontal dashed lines depict true values β^0_{ij}.

model (Zeng et al., 2024a), a computationally efficient alternative to the latent Gaussian model, again finding that the gradient-bridged posterior allows for excelleng mixing performance and accurate inference in a relatively high-dimensional setting with $z \in \mathbb{R}^{1000}$.

4.6 Data Application: Data Integration for Single-cell Data

We now consider a case study on data integration task. For data that are reproduced across studies or otherwise collected in batches, study-specific sources of variation may arise artifactually, and should be accounted for to better estimate the desired shared signal. Many Bayesian studies of integrating data across studies take a natural factor analytic approach (De Vito et al., 2021; Roy et al., 2021; Avalos-Pacheco et al., 2022; Chandra et al., 2024). While some of these studies give special attention to various possible identifiability issues, due to the rotationally invariant nature of the problem formulation, they all require alignment of the estimated loadings via post-processing. These studies have proposed post-hoc alignment via orthogonal procrustes problems as well as Varimax rotation (Aßmann et al., 2016; Rohe and Zeng, 2023).

The gradient-bridged posterior now enables a straightforward way to address the alignment problem directly. We can include an orthogonal procrustes sub-problem in defining our model likelihood (Gower, 1975; Ma et al., 2024). Here, we illustrate how just using a distance-based objective, the methodology allows us to generate aligned samples toward a unified representation that mitigates batch effects while preserving information from cell types. We remark that this

machinery can be combined within model-based approaches such as factor models to ameliorate rotational ambiguity.

We consider human pancreatic data (Stoeckius et al., 2017) consisting of transcriptomic profiles from multiple types of human pancreatic cells. These profiles are obtained through $B = 5$ batches of sequencing technologies, each with distinct technical biases and variations in sequencing depth. For each batch, the data were processed and transformed into a matrix $X_b \in \mathbb{R}^{d\times n}$ $(b = 1, \ldots, B)$, where d is the number of features, n is the sample size; hence $X_{b,(.,i)}$ is the i-th observation in the b-th batch. The data pre-processing procedures are provided in the Supplementary Materials. To align these batches, we introduce a shared centering parameter $u \in \mathbb{R}^{d\times n}$, a scaling parameter $s_b > 0$ for adjusting the magnitude of the data and an approximate rotation-reflection matrix $R_b \in \mathbb{R}^{d\times d}$ $(i = 1, \ldots, B)$ for each batch. Following our result described in Example 4.3, our point of departure is the primal problem

$$\min_{R_b} \|R_b X_b - s_b u\|_F^2, \quad \text{subject to } R_b^{\mathrm{T}} R_b = I_d, \tag{4-11}$$

over batches $b = 1, \ldots, B$. Then, following Equation (4-8) we construct the shrinkage kernel

$$\exp\{-\lambda \sum_{b=1}^{B} \|W_b W_b^{\mathrm{T}} (s_b^2 X_b u^{\mathrm{T}} u X_b^{\mathrm{T}}) W_b W_b^{\mathrm{T}} - I\|_F^2\},$$

with $R_b = s_b u X_b^{\mathrm{T}} (W_b W_b^{\mathrm{T}})$. This positions us to quantify measure of fit under the remaining contribution to the likelihood

$$g(y; \beta, z) = \prod_{b=1}^{n} (\sigma^2)^{-B/2} \exp\left\{-\frac{\|R_b X_b - s_b u\|_F^2}{2\sigma^2}\right\}.$$

At a large λ, the gradient-bridged posterior drives the gradient of the dual objective nearly to zero, thereby approximately satisfying the orthogonality constraint for R_b while aligning it closely with its conditional optimum. In contrast, if $\lambda = 0$ and $R_b^{\mathrm{T}} R_b = I_d$ exactly, we would have a Gibbs posterior based on the generalized Procrustes loss. The Gibbs posterior strictly enforces the

orthogonality constraint, yet it lacks a strong concentration of R_b near its posterior mode, unless σ^2 is forced to be near zero. The comparison in Fig. 4-5 shows an empirical performance difference.

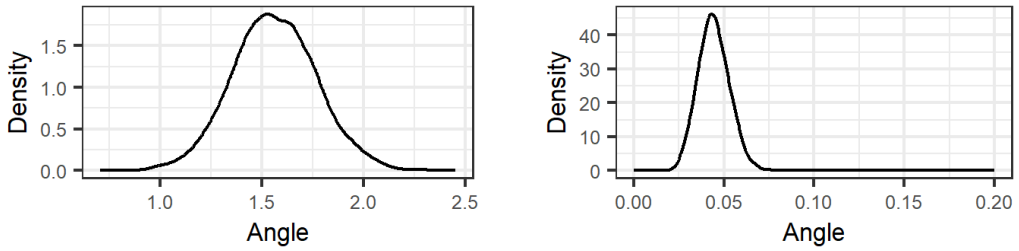


Figure 4-5. Density of the angles between the posterior samples R_b and the optimum of (4-11), from Gibbs posterior (left) and gradient-bridged posterior (right).

We choose standard half-normal priors for s_b, $\mathrm{Ga}^{-1}(2, 1)$ prior for σ^2 and a standard normal prior for u. We run the No-U-Turn Sampler for $100,000$ iterations, and discard the first $10,000$ iterations as burn-ins, and we thin the chain at 20. For each sample of $\{s_b, R_b\}_{b=1}^B$, we can produce a sample of the unified representation (aligned data) $(R_b X_b / s_b)_{b=1}^B$.

To assess batch effect mitigation, we use the batch-associated Davies–Bouldin index (Davies and Bouldin, 1979) which quantifies batch separation relative to within-batch compactness. Lower values indicate better batch mixing, which in our case indicates reduced batch effects and thus more successful data integration. Because one index is associated with each aligned sample, the distribution of Davies-Bouldin indices further allows us to evaluate the variability of integration outcomes. We compare results using the gradient-bridged posterior to the same index evaluated on (i) the raw data, (ii) the aligned data via solving the generalized Procrustes problem under loss function (4-11) under $u = X_1$, and (iii) the samples from a Gibbs posterior as described above taking $\lambda = 0$ (using a Gibbs sampler with the help of the `rstiefel` package, Hoff and Franks, 2021). To facilitate comparison to data under (i) and (ii), we derive a point estimate from samples under the gradient-bridged posterior and the Gibbs posterior under a decision-theoretic framework. Specifically, we select the empirical maximizer of the normalized mutual information

computed with respect to the cell-type labels. Normalized mutual information is chosen because its values effectively reflect how well cell-type signals are preserved.

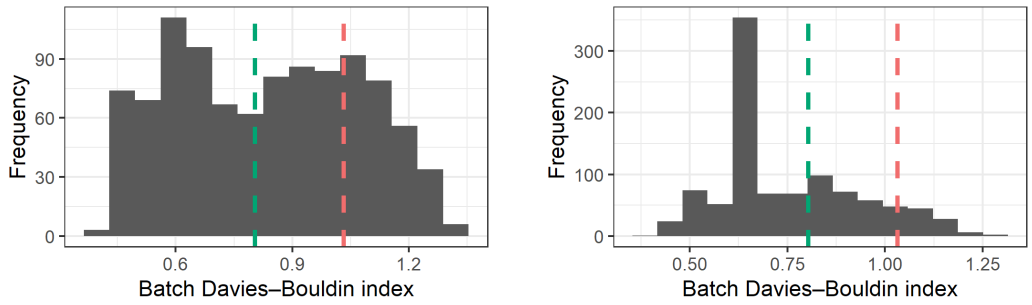


Figure 4-6. Histograms of posterior samples of batch-associated Davies–Bouldin index, from Gibbs posterior (left) and gradient-bridged posterior (right). The dashes indicate the corresponding values from the raw data (red) and the generalized Procrustes analysis (green).

Figure 4-6 displays the posterior distributions of the batch-associated Davies–Bouldin index. The gradient-bridged posterior shows a substantial improvement in the batch-associated Davies–Bouldin index compared to raw data and other two methods. To visualize the aligned data, we embed the point estimates of the unified representations into the space of the first two principal components. In Figure 4-7, we can see clear batch effects from the raw data and those processed using generalized Procrustes analysis, as the batches, labeled by color, can still be spotted as fairly distinct clusters. The Gibbs posterior improves the batch mixing, but arguably still exhibits noticeable batch separation. The gradient-bridged posterior achieves the best batch mixing. In more detail, these findings are supported quantitatively as presented in Table 4-1.

Table 4-1. Comparison of point estimates of unified representation in terms of batch-associated Davies–Bouldin index

Data representation	Davies–Bouldin Index
Raw data	1.032
Generalized Procrustes analysis	0.803
Gibbs posterior	0.834
Gradient-bridged posterior	0.748

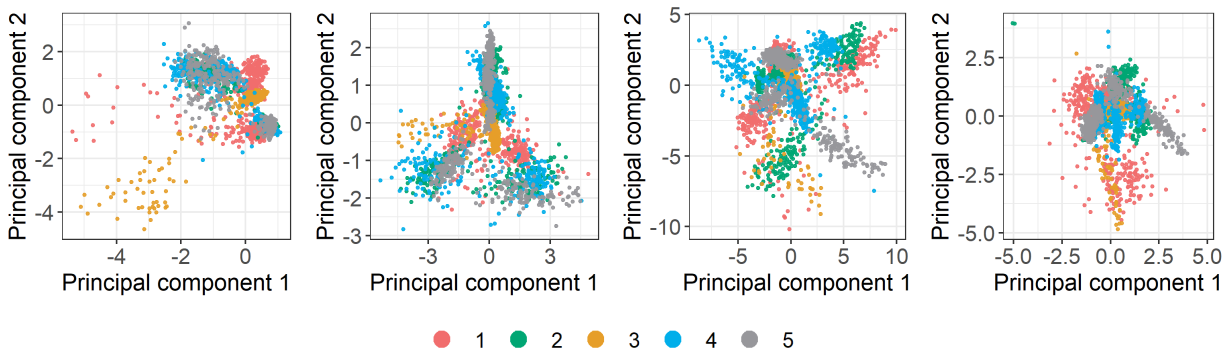


Figure 4-7. Integrated data represented in two-dimension using principle component analysis colored by batch. Left to right: raw data, generalized Procrustes analysis, Gibbs posterior, gradient-bridged posterior.

CHAPTER 5
NORMALIZING FLOW TO AUGMENTED POSTERIOR: CONDITIONAL DENSITY ESTIMATION WITH INTERPRETABLE DIMENSION REDUCTION FOR HIGH DIMENSIONAL DATA

In this chapter, we focus on the conditional density estimation (CDE) problem involving a high-dimensional y_i and low-dimensional x_i; for example, x_i could correspond to labels, or a continuous vector providing context information for the observed y_i. Specifically, we use the normalizing flow to form an invertible transform of data y_i and an "augmented posterior (AP)" based z_i that comprises of two components $[z_{P,i}, z_{N,i}]$; $z_{P,i}$ is a low-dimensional subvector that forms a joint distribution with x_i, such as via simple logistic/linear regression likelihood $f(x_i \mid z_{P,i})$. Effectively, the distribution of $[z_{P,i}, z_{N,i}]$ is the posterior distribution of $(z_{P,i} \mid x_i)$ augmented by independent Gaussian $z_{N,i}$. Note that there has been some recent work on normalizing flow based CDE, e.g., Papamakarios et al. (2017). However, the proposed approach enjoys several unique advantages: first, it produces a supervised dimension reduction in those $z_{P,i}$ within the CDE framework; second, it requires only simple modification of the common normalizing flow networks, so it is very easy to implement; last, it produces a single latent variable z_i for each data point y_i, and hence the unconditional density calculation does not involve summation or integration over the space of the predictor variable. We will illustrate these advantages as well as the outperformed density estimation results in this chapter.

5.1 Motivation

When using the mixture models for CDE, the following two difficulties arise for a high-dimensional data y. First, a parametric specification of the component distribution g can be unsatisfactory, since location-scale distribution is often an over-simplification for high-dimensional data. For example, for a collection of face photos from one subject, the mean of a Gaussian density g is often a poor summary characterization for this group of data points, since there are other factors (such as unknown lighting conditions) that contribute to the within-group variability, besides just pixel-wise random noise. To address this issue, it is often useful to assume that the high-dimensional data point lie close to several manifolds, each having an intrinsic low

Reprinted with permission from Zeng et al. (2024b).

dimension. One of the well-known solutions is the mixture of factor analyzers (MFA) (McLachlan and Peel, 2000; Tang et al., 2012), or mixture of probabilistic principal component analysis (PCA) (Tipping and Bishop, 1999), which retains the multivariate Gaussian density $g[y_i \mid \mu_k(x_i), \Sigma_k]$; further, the covariance matrix takes the form $\Sigma_k = U_k U_k' + \Lambda_k$, with U_k some $p \times d$ matrix with d small, and Λ_k a diagonal positive matrix. Effectively this parameterization assumes that for some h, $y_i - \mu_k(x_i)$ lies near a linear subspace spanned by the columns of U_k. However, it is rather difficult to extend this technique to non-linear manifolds through non-linear mappings. Alternatively, a popular idea is to find a low-dimensional representation, or "embeddings" that preserve some (often not all) relational characteristics of the high-dimensional data, such as pairwise distances or local neighborhoods. Such embedding is denoted by $z_i \in \mathbb{R}^d$ for each data point. Examples include Sammon's mapping (Sammon, 1969), kernel PCA (Schölkopf et al., 1997), Laplacian eigenmaps (Belkin and Niyogi, 2003), locally-linear embeddings (Roweis and Saul, 2000), Gaussian process latent variables (GPLV) (Lawrence, 2003), t-distributed stochastic neighbor embeddings (t-SNE) (Van der Maaten and Hinton, 2008), uniform manifold approximation and projection (UMAP) (McInnes et al., 2018), just to name a few. After obtaining such z_i's, one could calculate the conditional density $f(z_i \mid x_i)$ for z_i, as a surrogate the corresponding conditional density for y_i. Despite some success in visualization tasks and cluster analysis, a key issue among the aforementioned methods is that the procedure of dimension reduction often lacks a generative distribution (except for GPLV); consequently, a density for $f(y_i \mid x_i)$ can not be obtained.

The second difficulty is the curse of dimensionality when computing a high-dimensional mixture distribution, which is often overlooked in the literature. Common algorithms used for mixture model estimation involve a discrete latent variable $z_i = h$ with probability w_k, given that $(y_i \mid z_i = h, x_i)$ comes from a component distribution $g[\cdot \mid \theta_k(x_i)]$. Since $p(z_i = h \mid y_i, \theta_k) \propto w_k f[y_i \mid \theta_k(x_i)]$, it allows iterating through the following two steps: (a) sampling of z_i via a multinomial distribution (or taking expectation $\tilde{z}_i = p(z_i = h \mid y_i, \theta_k)$ in the EM algorithm (Fraley and Raftery, 2002)); (ii) updating $\theta_k(x_i)$ conditioned on z_i or $\tilde{z}_i$.

Nevertheless, since for $h = 1, \ldots, K$, the high-dimensional y_i creates large magnitude of densities, thus resulting in $f[y_i \mid \theta_k(x_i)]/f[y_i \mid \theta_l(x_i)] \approx 0$ or ∞, unless $\theta_k(x_i) \approx \theta_l(x_i)$. Consequently, each $p[z_i \mid y_i, \theta_k(x_i)]$ is very likely to be stuck at 1 for a given h^*, and close to 0 for all the other $h \neq h^*$, with the end result being that the estimation algorithm "gets stuck" at the initial assignment z_i's.

The curse-of-dimensionality for density estimation is a well known issue and the reason that algorithms such as rejection sampling, importance sampling, and Metropolis-Hastings fail in high dimensions (Gelfand, 2000; Zuev et al., 2011). A similar problem was recently discovered in high-dimensional clustering (Chandra et al., 2023), which is closely related to the unconditional density estimation problem.

As introduced in Section 1.3, the normalizing flow models were proposed for high-dimensional data, e.g., images, where we assume that for each data y_i, there is a latent z_i following a base distribution, and the two variables are connected by an invertible mapping. Since one can inversely obtain z_i as a deterministic transform of y_i, a number of interesting directions of exploration for z_i arise. One of them is whether a more interpretable modeling structure for z_i can be used, other than just being drawn from an independent Gaussian distribution. Early examples includes using a mixture of Gaussian distribution (Izmailov et al., 2020), or a mixture of subspace structure (Peng et al., 2020) to name a select few. Although some interpretable results, including improved clustering accuracy were reported, it was later discovered that most of the improved results were largely due to the specific pre-processing of the reported data sets (Haeffele et al., 2020), instead of the selected distribution for z_i. This cautionary tale serves as a good warning, that it is quite difficult — if not impossible — to rely on unsupervised normalizing flow (that is, using y_i alone) to find structure in the latent z_i. Naturally, this motivates us to consider external information from x_i, and create a normalizing flow-based conditional density estimator. This motivates us to propose our new method.

5.2 Preliminary: Generative Models based on Normalizing Flows

This section provides some preliminary on the normalizing flow neural networks. Suppose there is a one-to-one and differentiable almost everywhere mapping $T_\theta : \mathbb{R}^p \to \mathbb{R}^p$ (with θ the parameters within it), that can transform random variable y into another latent continuous $z \in \mathbb{R}^p$ following a simple distribution f_z. With a slight abuse of notation, $f_\cdot$ is used to represent both a density and a probability function. After a change of variable, for $i = 1, \ldots, n$:

$$f_y(y_i) = f_z[T_\theta(y_i)]|\nabla_y T_\theta(y_i)|,$$

with $|\nabla_y T_\theta(\cdot)|$ being the determinant of the Jacobian matrix with the gradient taken with respect to y_i. To estimate T_θ, one typically minimizes the Kullback–Leibler (KL) divergence between the target distribution $f_y^*(y)$ and the normalizing flow-based $f_y(y)$, which is

$$\mathrm{KL}[f_y^*(y), f_z[T_\theta(y)]|\nabla_y T_\theta(y)|] \approx \frac{1}{n}\sum_{i=1}^{n}\{-\log f_z[T_\theta(y_i)]|\nabla_y T_\theta(y_i)|\} + \text{constant},$$

where the right hand side is the empirical KL, equaling to the loss function to be minimized over θ.

To flexibly parameterize the transform while maintaining invertibility, one uses a special multilayer neural network ("invertible neural network") with $T = T_1 \circ T_2 \circ \cdots \circ T_m$, and each $T_k : \mathbb{R}^p \to \mathbb{R}^p$ is a layer of an invertible transform of relatively simple operations; for example, in RealNVP (Dinh et al., 2016), the input of T_k is equally partitioned to $r = [r_A, r_B]$, and $T_k([r_A, r_B]) = [r_A, r_B \odot s(r_A) + l(r_A)]$, with s and l functions that produce the location-scale change to r_B. Then in the next layer, one alternates by using $T_{k+1}([r_A, r_B]) = [r_A \odot s(r_B) + l(r_B), r_B]$. Other types of neural networks include autoregressive flows and residual flows. We refer readers to Papamakarios et al. (2021) as a review for all types of flows employed. Besides invertibility, these neural networks are also carefully designed so that the term of the determinant of the Jacobian and the inverse mapping of the neural networks can be computed at low cost.

Obtaining an approximate solution of T_θ via an invertible neural network leads to a transport map $T_{\hat{\theta}}$, that gives a density estimator for the data $\hat{f}_y(y_i) = f_z[T_{\hat{\theta}}(y_i)]|\nabla_y T_{\hat{\theta}}(y_i)|$. This method is commonly referred to as "normalizing flow", since one often assigns a standard independent Gaussian distribution to $z \sim \mathrm{N}(0, I_p)$ for simplicity, which is often called the base distribution. As a generative model, one can sample $z_{i'} \sim \mathrm{N}(0, I_p)$, and then $y_{i'} = T^{-1}_{\hat{\theta}}(z_{i'})$ produces new generated data from the estimated distribution f_y.

5.3 Method

The section begins by introducing notation used in the sequel. Let $y_i \in \mathbb{R}^p$ be a continuous response with distribution f_y, and $x_i \in \mathcal{X} \subseteq \mathbb{R}^m$ be the corresponding predictor variable (could be discrete, continuous, or a mix of both) drawn from another distribution f_x. This chapter focuses on the case of large p and small m.

5.3.1 Augmented Posterior for CDE

To enable CDE for $f_{y|x}(y_i \mid x_i)$, as well as making the latent variable z_i more interpretable, we consider a joint distribution between z and x,

$$f_{z,x}(z_i, x_i) = f_{z_N}(z_{N,i}) f_{z_P}(z_{P,i}) f_{x|z}(x_i \mid z_{P,i}; \beta), \tag{5-1}$$

independently for $i = 1, \ldots, n$, where $z_i = [z_{P,i}, z_{N,i}]$. The first component $z_{P,i} \in \mathbb{R}^d$ is a low-dimensional subvector and is used in a predictive model for x_i, whereas the second component $z_{N,i} \in \mathbb{R}^{p-d}$ is a high-dimensional subvector unrelated to x_i, and β is viewed as a non-random parameter that will be estimated. If one considers $f_{x|z}$ to be the likelihood for x, and f_z to be the prior distribution for z, then it is not hard to see that,

$$f_{z|x}(z_i \mid x_i) = f_{z_N}(z_{N,i}) \frac{f_{z_P}(z_{P,i}) f_{x|z}(x_i \mid z_{P,i}; \beta)}{\int_{\mathbb{R}^d} f_{z_P}(t) f_{x|z}(x_i \mid t; \beta) \mathrm{d}t},$$

where the second part is the posterior of z_P, $f_{z_P|x}(z_{P,i} \mid x_i; \beta)$, and $f_{z_N}(z_{N,i})$ is an independent random variable that augments this posterior, to make z_i match the dimension of y_i. Therefore, each $z_i = (z_{P,i}, z_{N,i})$ is referred to as a sample point from an "augmented posterior".

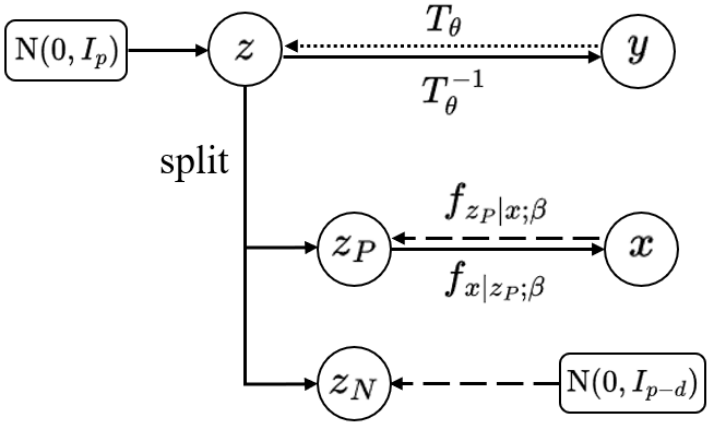


Figure 5-1. The diagram of the architecture of AP-CDE. The solid lines show the generative process for the data (y, x). The dashed lines show how to generate a new latent variable z from the augmented posterior.

Figure 5-1 shows the architecture of the proposed model. Note that for simplicity, we retain the standard independent Gaussian density for f_{z_N} and f_{z_P} as in a typical normalizing flow, while employing a generalized linear model for $f_{x|z}$. For example, if $x_i \in \mathbb{R}^m$ is continuous, then one can use $x_i = \beta_0 + \beta_1 z_i + \epsilon_i$, with $\beta_0 \in \mathbb{R}^m$, β_1 an $m \times d$ matrix, and $\epsilon_i \overset{iid}{\sim} \mathrm{N}[\vec{0}, \mathrm{diag}(\beta_2)]$ with β_2 positive vector. For each univariate discrete $x_{i,j} \in \{1, \ldots, K\}$, one can use a multinomial logistic regression in $f_{x|z}$, $p(x_{i,j} = k \mid z_{P,i}) \propto \exp(\beta_{0,k} + \beta_k' z_{P,i})$, with $\beta_{0,k} \in \mathbb{R}$, $\beta_k \in \mathbb{R}^d$ for $k = 1, \ldots, K-1$, and β_K and $\beta_{0,K}$ fixed to 0-value as common in logistic regression. In general, the conditional probability function $f_{x|z}$ could be proportional to the original function to the power λ for the purpose of regularization. This is commonly used in robust Bayesian models (Grünwald and Van Ommen, 2017) and hybrid deep generative models (Nalisnick et al., 2019). In the proposed model, the normalizing constant is merged into the denominator of the posterior $f_{z_P|x}$ naturally, and the model is still a generative one.

In the case of x_i having both continuous and (potentially more than one) discrete elements, $x_i = [x_{A,i}, x_{B,i}]$, one can partition $z_{P,i} = [z_{P_A,i}, z_{P_B,i}]$, and use a separable likelihood function $f_{x|z_P}(x_i \mid z_{P,i}) = f_{x_A|z_{P_A}}(x_{A,i} \mid z_{P_A,i}) f_{x_B|z_{P_B}}(x_{B,i} \mid z_{P_B,i})$ — where the first term on the right models the continuous part, and the second term does the discrete part. This separation in z_P is motivated by applications under consideration, in which commonly the discrete part corresponds

to the class label of an image, whereas the continuous part to other conditions (such as lighting) that are unrelated to the labeling information.

Next, if there is an invertible mapping T_θ that connects z_i and y_i, by applying change-of-variable, one has

$$f_{y|x}(y_i \mid x_i) = f_{z|x}[T_\theta(y_i) \mid x_i; \beta]|\nabla_y T_\theta(y_i)|.$$

To estimate this mapping T_θ as well as the parameter β in the predictive model of x_i, one minimizes the empirical KL divergence between the target distribution $f^*_{y|x}(y \mid x)$ and the $f_{y|x}(y \mid x)$ in the above, leading to:

$$\begin{aligned}\min_{\theta,\beta} \frac{1}{n}\sum_{i=1}^{n}\Big\{ &- \log f_z[T_\theta(y_i)]|\nabla_y T_\theta(y_i)| \\ &- \log f_{x|z}[x_i \mid T^P_\theta(y_i); \beta] + \log \int_{\mathbb{R}^d} f_{z_P}(t) f_{x|z}(x_i \mid t; \beta)\mathrm{d}t\Big\},\end{aligned} \tag{5-2}$$

where $T^P_\theta(z_i)$ [or, $T^N_\theta(z_i)$] means taking the subvector (corresponding to z_P, or z_N) from the output of $T_\theta(z_i)$. Although the last integral above is often intractable, as the stochastic gradient descent technique (the commonly used optimization algorithm in normalizing flow) only requires an approximate gradient via taking a random subset of size n_b instead of n, one can replace the last term via a Monte Carlo estimate, $\log\big[\sum_{l=1}^{M} f_{x|z}(x_i \mid t_l; \beta)/M\big]$, with each $t_l \overset{iid}{\sim} f_{z_P}$ [in this chapter, $\mathrm{N}(\vec{0}, I_d)$]. The n_b integrals to be estimated can share those M random samples t_l's across i in the subset of size n_b, and hence the estimation is inexpensive.

After the KL divergence is minimized, a conditional density estimator is obtained by

$$\hat{f}_{y|x}(y_i \mid x_i) = f_{z|x}\big[T_{\hat{\theta}}(y_i) \mid x_i; \hat{\beta}\big]|\nabla_y T_{\hat{\theta}}(y_i)|.$$

Further, to calculate the marginal density of y for a new data point for which only the response y_i is available, one can simply marginalize over x_i in Equation 5-1, and obtain

$$f_y(y_i) = f_{z_N}\left[T^N_{\hat{\theta}}(y_i)\right] f_{z_P}\left[T^P_{\hat{\theta}}(y_i)\right] \left|\nabla_y T_{\hat{\theta}}(y_i)\right|.$$

Therefore, after the optimization, both conditional and marginal density estimation can be accomplished in a computational efficient manner.

5.3.2 Supervised Dimension Reduction and Validation

Besides conditional density estimates, another advantage of using the augmented posterior is that it produces a low-dimensional representation $z_{P,i}$ that is related to the variations in x_i. Indeed, after the optimization step concludes, one obtains a computational form that produces low-dimensional $z_{P,i} = T^P_{\hat{\theta}}(y_i)$, and $T^P_{\hat{\theta}}$ is estimated under supervising information from x.

On the other hand, one may further wonder if $z_{P,i}$ has captured all the useful information that connects y_i and x_i. To formalize, if the true data generating mechanism for (y_i, x_i) is indeed based on $z_i \sim f_z$, $y_i = T^{-1}_{\theta}(z_i)$, $x_i \sim f_{x|z_P}$, then one would have the following sufficient dimension reduction outcome:

$$x \perp\!\!\!\perp y \mid T^P_{\theta}(y),$$

as the ideal result.

There are ways to test conditional independence in classical linear models (Cook and Ni, 2005) and non-linear low dimensional models (Su and White, 2007, 2008). However, the combination of nonlinearity and high dimensionality poses significant challenges. Fortunately, for high-dimensional data with y_i and x_i, this can be validated via synthesizing new data and predicting x_i via another neural network G, independently trained with (y_i, x_i)'s.

Specifically, for each $z_i = T_{\hat{\theta}}(y_i)$ produced, one fixes $z_{P,i}$ while replacing $z_{N,i}$ with an independently sampled $\tilde{z}_{N,i,j} \sim \mathrm{N}(0, I_{p-d})$ for $j = 1, \ldots, J$. Then synthesized $\tilde{y}_{i,j} = T^{-1}_{\hat{\theta}}[z_{P,i}, \tilde{z}_{N,i,j}]$ is obtained. One predicts the x_i using $\tilde{y}_{i,j}$ via the separately trained network G, and observes if each predicted $\hat{x}_{i,j}$ differ from the observed x_i — in the ideal case, $\hat{x}_{i,j}$ should

not differ much from x_i (since $z_{P,i}$ should contain most of the predictive information about x_i), and the conditional independence can be quantified by the error rate.

5.3.3 Parameterization Details

In this chapter, T_θ is parameterized using the state-of-art normalizing flow Glow (Kingma and Dhariwal, 2018). We choose Glow in the experiments for its high accuracy on the density estimation and the ease of implementation. It employs a multi-scale architecture (Dinh et al., 2016), which contains L levels. After each level, half of the dimensions of latent z_i are immediately modeled as Gaussians, while the remaining half are further transformed by the flows. This significantly improves the computation efficiency. Each level consists of K (depth) steps of flow that share an identical structure which contains an activation normalization, an invertible 1×1 convolution and an affine coupling layer. In this chapter, we use the additive coupling layer as a special case of the affine coupling layer, in which the number of channels of the hidden layers (convolutional neural networks) is set to be 512.

Under this kind of multi-scale architecture, the following important problem is how to choose the way of splitting z_i into $[z_{P,i}, z_{N,i}]$. First, to make the Monte Carlo estimation of the integral in Equation 5-2 accurate, efficient and stable, one does not expect d, the dimension of $z_{P,i}$, too large. Second, since the outputs from the later layers experience more transforms compared to the ones from the earlier layers, choosing dimensions of $z_{P,i}$ from the relatively later layers will improve the model performance. This is illustrated in the numerical experiments.

5.4 Numerical Experiments

The following set of numerical experiments demonstrates the AP-CDE's advantages for density estimation and dimensional reduction. The following competing models are used to compare its performance: (i) the Glow normalizing flow based on independent Gaussian $z \sim \mathrm{N}(\vec{0}, I_p)$, without using any information from x; (ii) a modified Glow based on a Gaussian mixture (named Glow-Mix), $z_i \sim \sum_{k=1}^{K} w_k \mathrm{N}(\mu_k, I_p)$, with K set to the ground-truth number of classes in the data; and (iii) a CDE extended from Glow (named Glow-CDE) by adding x_i as an input in each additive coupling layer (Papamakarios et al., 2017, Section 3.4). When the predictor

variable x_i is discrete, e.g., class label, the comparison is also made with (iv) a naive CDE model based on Glow where for each value x' of x_i a Glow is trained on those data y_i with $x_i = x'$ (named Glow-NCDE). For a fair comparison, we set the four competitors to have the same numbers of levels L and the same depth K, as in the AP-CDE model. The first two models perform unconditional density estimation with producing latent variables, whereas the last two models are conditional. Unlike the proposed model, the two CDE models compared do not have a unique corresponding latent variable for a new data. Note that the fourth model is inefficient because one has to train models of number of classes of x.

We use the Adam optimizer (Kingma and Ba, 2014) provided in the PyTorch framework to train all models, with a mini batch size of $n_b = 64$ and learning rate at 0.0005 for all models. All models are trained for 200 epochs, where the optimizer goes through the whole training dataset exactly once in each epoch. For the first 10 epochs as warm-up, the learning rate linearly increases to 0.0005 after each batch training. Then the learning rate goes down to 10^{-4} using cosine annealing schedule (Loshchilov and Hutter, 2016). We monitor the loss function of all models and ensure that convergence is achieved by all models. Finally, $M = 1000$ is set in the Monte Carlo estimator for the integral in Equation 5-2.

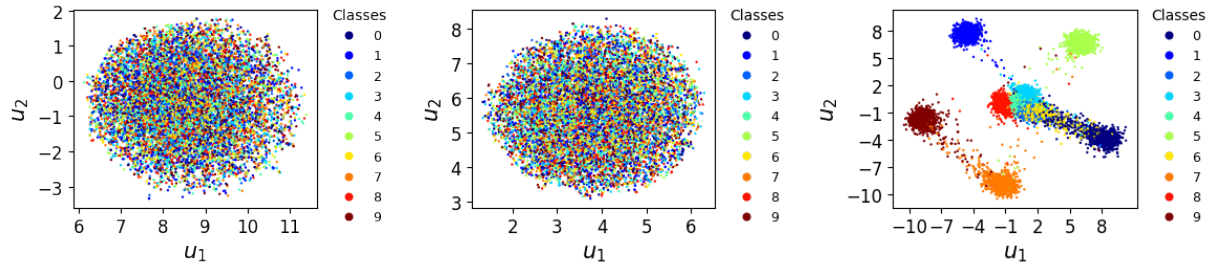
The experiments are based on the following two datasets: the FashionMNIST one of the fashion products (Xiao et al., 2017) and the Extended Yale Face B one of face images (Lee et al., 2005b). These images have a single color channel, containing pixel values $\{0, \ldots, 255\}$. We follow Papamakarios et al. (2017); Dinh et al. (2016) to dequantize the pixel values by adding standard uniform noise onto every pixel and scaling the values to $(0, 1)$ by dividing 256. Additional numerical experiments are shown in the Chapter D.

5.4.1 FashionMNIST Images of Fashion Products

The FashionMNIST data is used to illustrate the CDE when x_i is discrete. This dataset contains $70,000$ processed images of fashion products, each having 28×28 pixels. Among them, $60,000$ is used for training purposes and the remaining is used for testing. Each image y_i is associated with a discrete label x_i with values from 0 to 9 recording the ground-truth fashion

products, including T-shirt, sandal, bag, etc. Each image is padded to dimension 32×32 by adding 2 more dimensions of 0 in each of four direction, and then extend each image to 3 channels by repeating the image in each of channel.

For the Glow model, levels $L = 3$ and depth $K = 32$ are set. For the sub-model $f_{x|z}$ in AP-CDE, we use the likelihood function of the multinomial logistic regression, as stated in Section 5.3.1. Since the labels are relatively balanced across classes, to improve interpretation on the latent $z_{P,i}$'s, all intercept terms $\beta_{0,k}$'s are set to be zeros. We compare several choices of the dimensions of $z_{P,i}$. To be clear, in the multi-scale architecture, when the data y_i has dimension $3 \times 32 \times 32$, the output $z_i^{(1)}$ from the first level has dimension $6 \times 16 \times 16$, the output $z_i^{(2)}$ from the second level has dimension $12 \times 8 \times 8$, while the final output $z_i^{(3)}$ has dimension $48 \times 4 \times 4$. The following choices of $z_{P,i}$ are compared: (1) $z^{(3)}_{1:2,1,1}$; (2) $z^{(2)}_{1:2,1,1}$; (3) $z^{(1)}_{1:2,1,1}$; (4) $z^{(3)}_{1:16,1,1}$; (5) $z^{(2)}_{1:4,1:2,1:2}$; (6) $z^{(1)}_{1:4,1:2,1:2}$; (7) $z^{(3)}_{1:48,1:2,1:2}$; (8) $z^{(2)}_{1:12,1:4,1:4}$; and (9) $z^{(1)}_{1:3,1:8,1:8}$.



(a) Glow (b) Glow-Mix (c) AP-CDE

Figure 5-2. Latent representations estimated by the three models applied on the FashionMNIST training set. For the Glow and the Glow-Mix models, UMAP is used to reduce the dimensions to 2.

Figure 5-2 plots the latent representations produced by Glow, Glow-Mix and AP-CDE with the choice (1) for the $z_{P,i}$. Recall that for Glow and Glow-Mix, the latent z_i has the same dimensionality as the images, UMAP (McInnes et al., 2018) is used to reduce the dimension and plot its output in 2D. For AP-CDE, the plot of the latent $z_{P,i}$ is provided. As expected, the latent variable produced by Glow follows a simple spherical Gaussian, and thus is not interpretable. Somewhat surprising, Glow-Mix does not produce a meaningful result either, despite using a mixture of 10 Gaussians (that corresponds to the true number of classes) — instead, the

Glow-Mix model converges to only one component containing a mixture of z_i's from all classes. This negative finding is in accordance to an early critique on clustering with deep neural networks (Haeffele et al., 2020), where it was reported that in an unsupervised setting, imposing a modeling structure on the latent variable (such as a mixture of Gaussians) does not lead to a clear separation of data from different classes. Using AP-CDE and the supervising label information form x_i, we obtain a good separation of the ten products based on the low-dimensional representation $z_{P,i} \in \mathbb{R}^2$ for the model $z^{(3)}_{1:2,1,1}$ (shown in Figure 5-2(c)).

Table 5-1. The averages of bits per dimension on the training and the testing sets of FashionMNIST for all models. Lower BPM means higher density.

Models	Glow	Glow-Mix	Glow-CDE	Glow-NCDE	AP-CDE
Training set	1.02	1.04	1.03	1.31	1.02
Testing set	1.03	1.05	1.04	1.32	1.03

Table 5-1 depicts the average bits per dimension (BPM) on the training and the testing sets for all models. Here the BPM is the negative log-densities divided by the number of dimensions, which is broadly used in the literature because it has similar scale for different resolution of images. For AP-CDE model, we choose the best model considering the error rate and the density estimation performance among the all choices of $z_{P,i}$. The model of using $z^{(3)}_{1:48,1:2,1:2}$ is chosen. Clearly, for this dataset, AP-CDE, likely due to a better group-wise concentration, produces overall higher (or equal) marginal densities compared to its competitors.

Table 5-2 depicts the average bits per dimension on the training and the testing sets for all AP-CDE models as well as the classification error rates on the testing set, where one obtains $z_{P,i}$ and predict the label via the trained logistic regression model using $\arg\max_{k\in\{0,\ldots,9\}} f_{x|z_P}(k \mid T^P_{\hat{\theta}}(y_i))$. It shows that when one chooses more dimensions of the same level for the $z_{P,i}$, the classification error rate will be lower, while the density estimation performances do not vary significantly. Moreover, even when the $z_{P,i}$ has the same number of dimensions, the higher level can provide more accurate classification. This bunch of experiments is an important guidance on how to choose the dimensions for $z_{P,i}$ — the whole output of the last level is always not a bad choice.

Table 5-2. The averages of bits per dimension on the training and the testing sets and the classification error rates on the testing sets for all AP-CDE models on the FashionMNIST data.

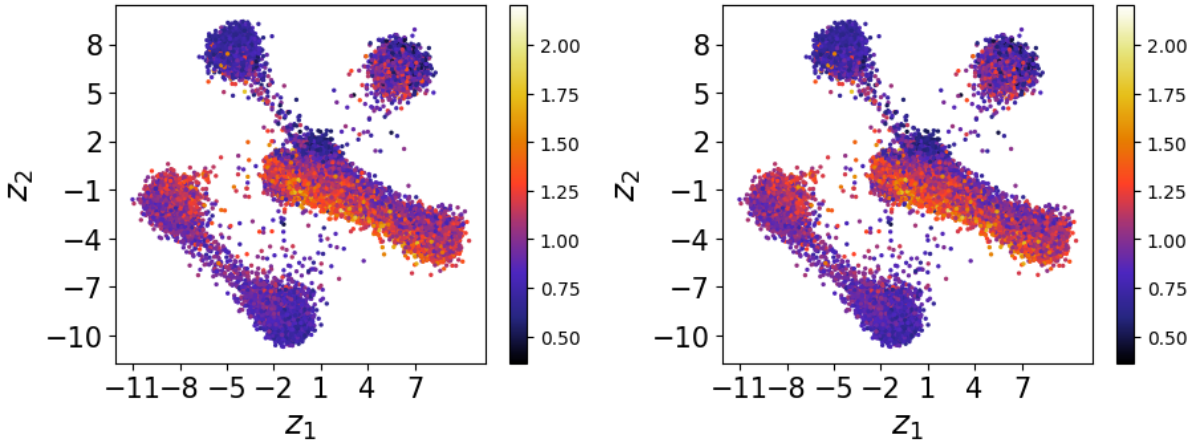
Models	BPM (Training set)	BPM (Testing set)	Error rate (%)
$z^{(3)}_{1:2,1,1}$	1.02	1.03	50.96
$z^{(2)}_{1:2,1,1}$	1.03	1.03	60.53
$z^{(1)}_{1:2,1,1}$	1.01	1.02	89.69
$z^{(3)}_{1:16,1,1}$	1.05	1.06	6.89
$z^{(2)}_{1:4,1:2,1:2}$	1.06	1.06	7.48
$z^{(1)}_{1:4,1:2,1:2}$	1.02	1.02	80.33
$z^{(3)}_{1:48,1:2,1:2}$	1.02	1.03	6.46
$z^{(2)}_{1:12,1:4,1:4}$	1.04	1.05	7.26
$z^{(1)}_{1:3,1:8,1:8}$	1.01	1.02	89.9

Empirically it shows that the low-dimensional representation $z_{P,i}$ contains almost all the information to separate the different classes when the dimensions of $z_{P,i}$ are well chosen. For the model $z^{(3)}_{1:48,1:2,1:2}$, as described in Section 5.3.2, if one fixes $z_{P,i}$, but replaces $z_{N,i}$ with independently sampled realizations from a Gaussian distribution, and then through $T^{-1}_{\hat{\theta}}$ one can obtain 10 new images for each original observation i. We then employ the ResNet101 (He et al., 2016) (a convolutional neural network separately trained on the training set) to classify these artificially generated images, and find that 95.26% of them are still classified to the same class label as the y_i's. Hence, it is concluded that the $z_{N,i}$ largely corresponds to within-class variation, whereas $z_{P,i}$ captures between-class variation.

Further, we color the latent $z_{P,i}$ using the magnitude of the density (Figure 5-3). In the result, those points with relatively low density values tend to correspond to images of low quality or higher ambiguity regarding the product class. To show this, we plot a few sampled fashion products in Figure 5-4 and sort each row by the density value in increasing order. It can be seen that the images on the left tend to be harder to assign to a class, compared to the ones on the right.

5.4.2 Human Face Photos

To illustrate the AP-CDE with continuous and mixed-type x_i, experiments are also run on the Yale face dataset. There are $2,414$ face photos, each containing a single color channel with a



(a) Training set.

(b) Testing set.

Figure 5-3. The first two dimensions of the latent variables from AP-CDE model $z^{(3)}_{1:2,1,1}$ on FashionMNIST, colored by the estimated densities in the scale of BPM.

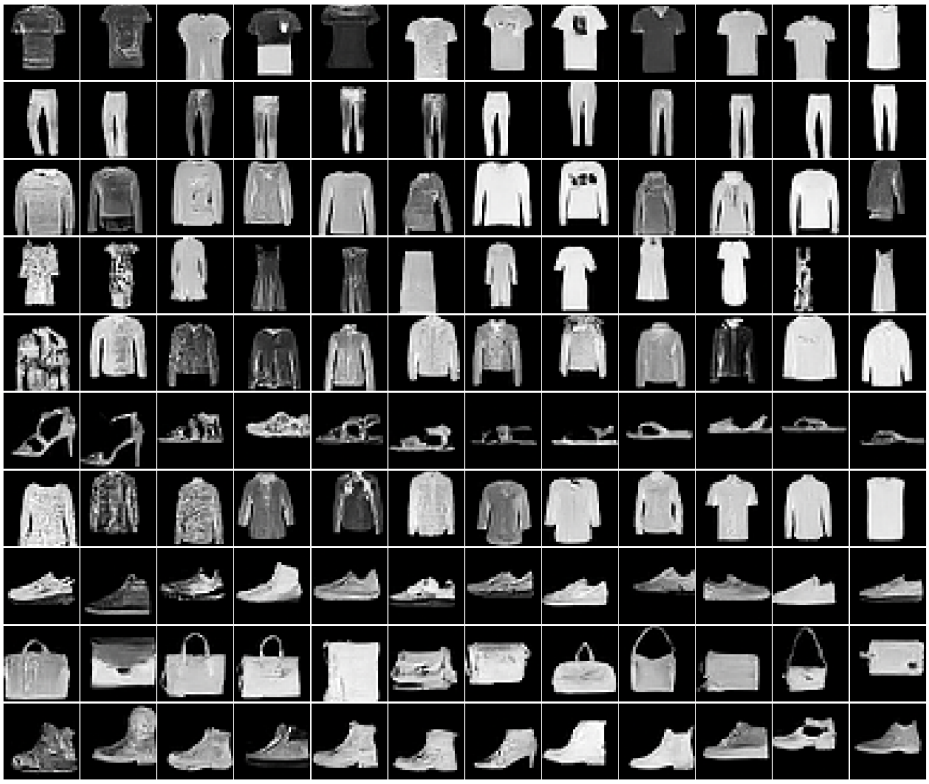
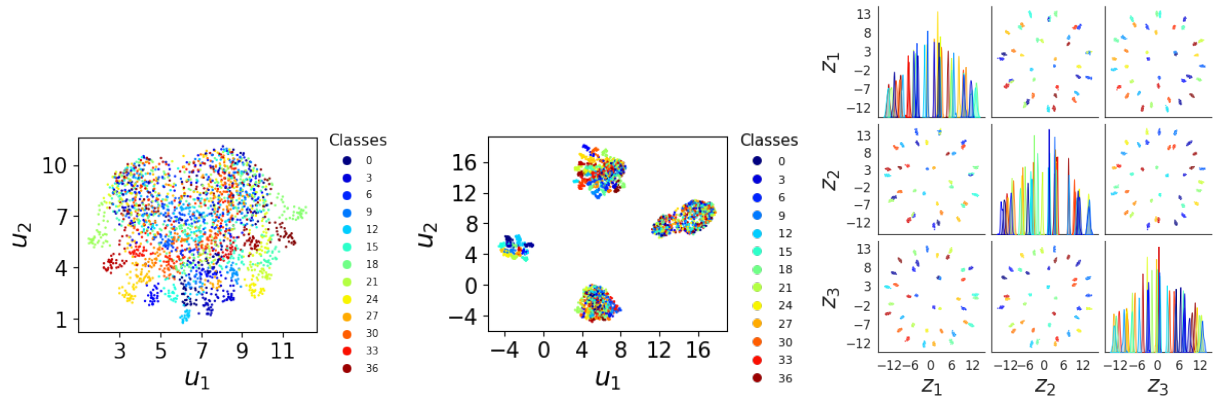


Figure 5-4. Sample images from the AP-CDE model $z^{(3)}_{1:48,1:2,1:2}$ trained by FashionMNIST data, with each row sorted in the increasing order of estimated densities.

168×192 pixel resolution. The images are resized to 28×32 to reduce computation cost, while maintaining clarity of the photos. The images come from 38 people, and this information is used

as a discrete class variable $x_{A,i}$. Further, the photos were taken under different light conditions, recorded as azimuth angle and elevation. This information is used as two continuous variables $x_{B,i}$ and $x_{C,i}$.

For the Glow model, the parameters are set to $L = 3$ and $K = 32$. A logistic regression likelihood $f_{x_A|z_{P_A}}$ is used for the discrete x_A, that depends on 3 dimensions of $z^{(3)}_{1:3,1,1}$. Here, $z^{(3)}$ is the output from the third level. Linear regression $x_{B,i} = \beta^B_0 + \beta^B_1 z_{P_B,i} + \epsilon^B_i$, $x_{C,i} = \beta^C_0 + \beta^C_1 z_{P_C,i} + \epsilon^C_i$ are employed for the other two covariates, where both $z_{P_B,i} = z^{(3)}_{4,1,1}$ and $z_{P_C,i} = z^{(3)}_{5,1,1}$ are one dimensional. Assume $\epsilon^B_i \overset{iid}{\sim} \mathrm{N}(0, 0.01)$ and $\epsilon^C_i \overset{iid}{\sim} \mathrm{N}(0, 0.01)$; note the low value of the variance selected, which forces higher correlation between (x_A, z_{P_A}) and (x_B, z_{P_B}). In this case, the last integral in Equation 5-2 has closed form because the integrand is the product of two Gaussian densities. The integral is $1/\sqrt{2\pi(\beta_2^2 + \beta_1^2)}\exp\{-(x_i - \beta_0)^2/[2(\beta_2^2 + \beta_1^2)]\}$.



(a) Glow (b) Glow-Mix (c) AP-CDE

Figure 5-5. The latent representations estimated from the Yale face data. The UMAP is used to reduce the dimensions to 2 for Glow and Glow-Mix. For the AP-CDE model, all the dimensions of z_P are shown in pairs plot.

Competing methods include Glow and Glow-Mix. As can be seen in Figure 5-5, AP-CDE leads to a clear separation of latent representations due to the use of label information, whereas unsupervised Glow and Glow-Mix fail to do so. Further, the Glow model does not produce a group-mixed sphere as did in the FashionMNIST experiment. The Glow-Mix model produces four clusters in the latent space but none of them represents some class of people.

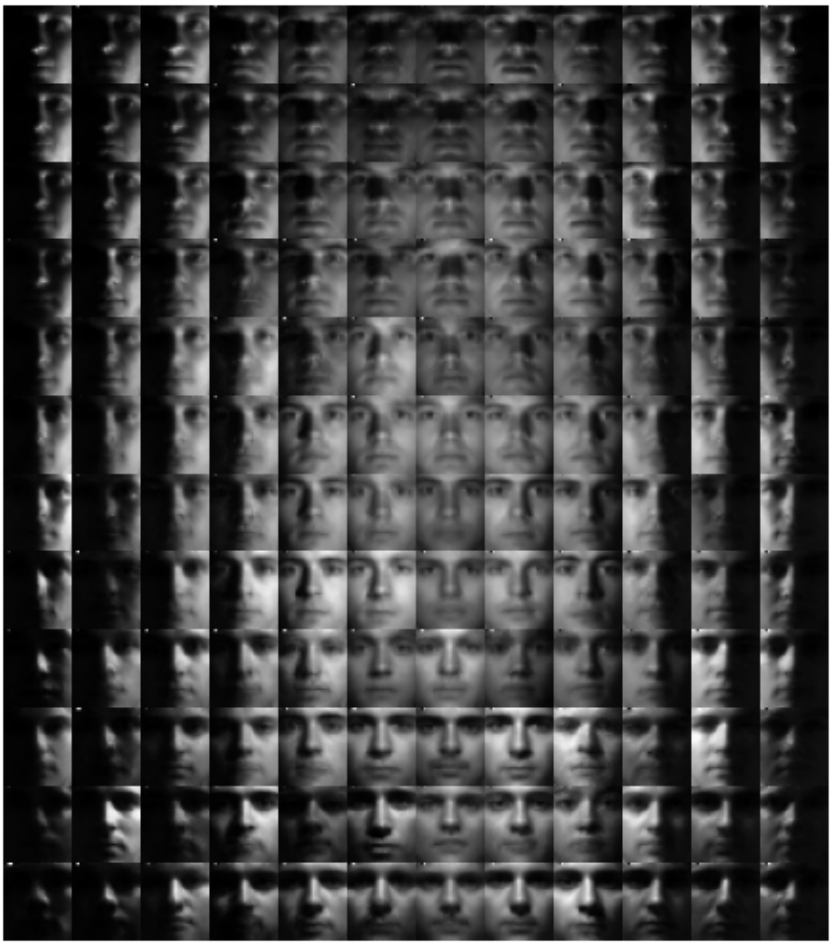


Figure 5-6. Synthesized face photos with gradually changing light azimuth angles (left to right) and elevations (top to bottom).

To show the expressiveness of AP-CDE as a generative model, the following procedure is used to synthesize new artificial images. One selects a range for Azimuth angles (-100 to 100) and a range for elevations (-60 to 60) (with interpolation), and form a 12×12 grid. For each grid cell, one draws a $Z_{P_A,i}$ that corresponds to a person's identity, from the empirical posterior distribution from the AP-CDE estimates. Further, one randomly draws the $Z_{N,i}$ component from a $\mathrm{N}(\vec{0}, I)$ distribution. With this choice of $\tilde{z}_i$, we synthesize new images for the i-th subject $\tilde{y}_i = T_{\hat{\theta}}^{-1}(\tilde{z}_i)$. As shown in Figure 5-6, there is a clear trend of change in the lighting conditions, caused by the changing values of $(x_{B,i}, x_{C,i})$.

5.5 Data Application

In the application study, we use 18001 images of leaves from strawberry plants, which either are healthy or have one of the three types of diseases: powdery mildew, anthracnose and fusarium wilt. This forms the labels of the four classes, which can be treated as the predictor variable x. The AP-CDE model is expected to conditionally estimate the densities of the images as well as do a supervised dimensional reduction.

Considering that the backgrounds of the images cause overfitting problem because the similarities between backgrounds rather than between the characteristics of the diseases take account for the major contribution on the distinction of the several groups (Saikawa et al., 2019), we use the method mentioned in Saikawa et al. (2019) to remove the backgrounds of all images. This segmentation also enhanced the quality of density estimation by reducing the redundant information. The images are resized to 128×128 resolution. To improve visual quality in generating samples, we follow Kingma and Dhariwal (2018) to use 5-bit images. Then they dequantize the pixel values as stated in the Section 5.4. The data is split into training set which includes 16000 images and the testing set which includes 2001 images.

For the parameters of Glow, the parameters are set to $L = 6$ and $K = 32$. For the conditional likelihood of $x \mid z_P$, again, we use $f_{x|z} \propto g^{\lambda}_{x|z}$, where g is the likelihood function of the multinomial logistic regression through the origin and $\lambda = 1000$. The subvector $z_{P,i}$ is chosen to be the output of the last level, which has dimension $384 \times 2 \times 2$.

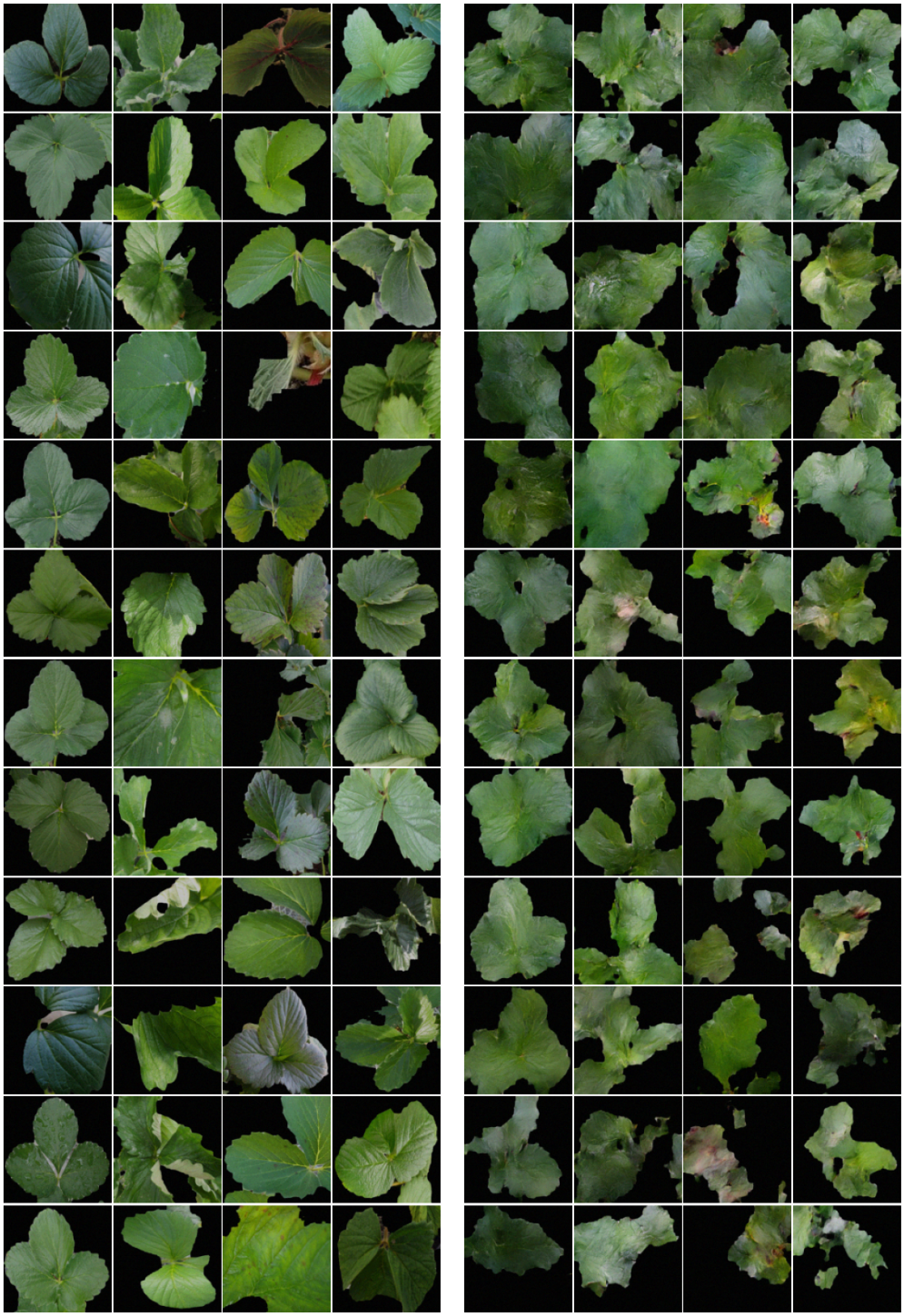
Table 5-3. The averages bits per dimension on the training and the testing sets of the leaves of strawberry plants data for all models.

Models	Glow	Glow-Mix	Glow-CDE	Glow-NCDE	AP-CDE
Training set	4.16	4.25	4.19	4.18	4.16
Testing set	4.17	4.26	4.22	4.18	4.17

Table 5-3 shows the average bits per dimension results for all models on the training and testing sets. The proposed AP-CDE model outperforms the other competitors on density estimation, while gets results as good as Glow. Figure 5-7(b) shows generated images for each of the four classes. As a comparison, Figure 5-7(a) shows the real images for each of the four classes.

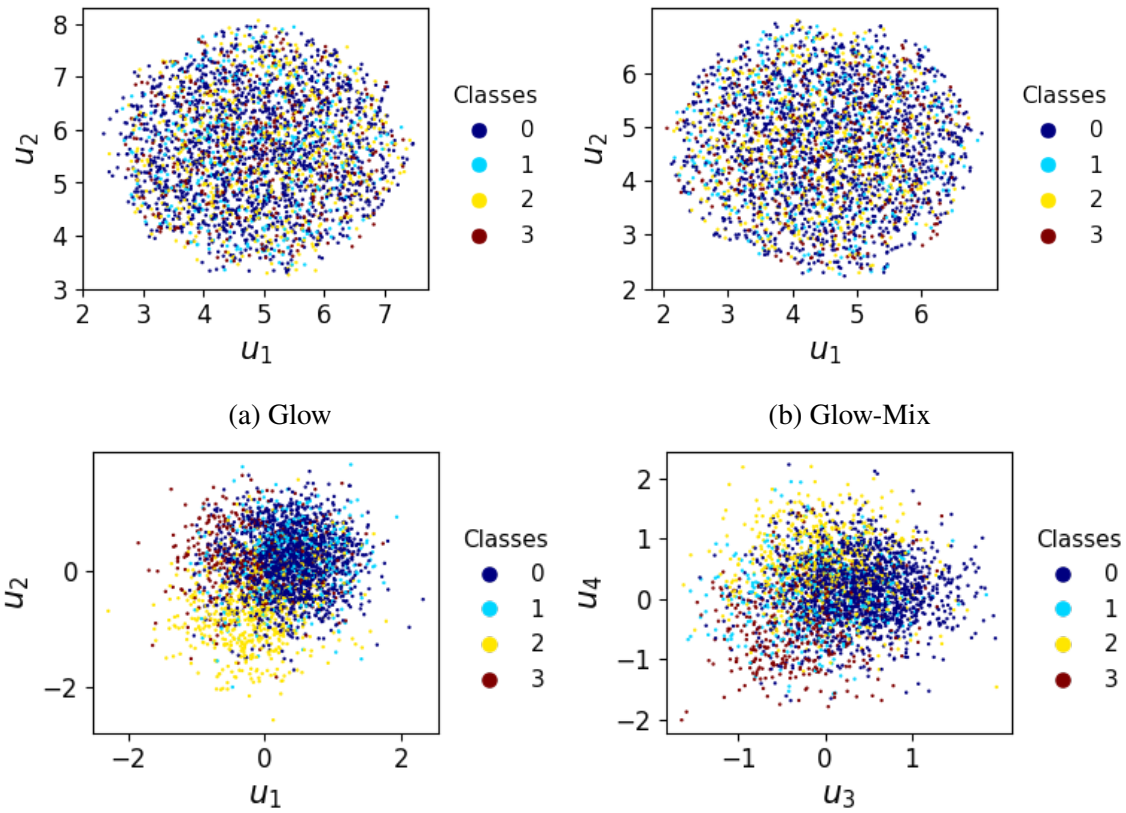
The second column shows the generated powdery mildew images, which clearly have some white spots on the leaves; the third column shows the generated anthracnose images, which have purple spots on the leaves; and the fourth column shows the generated fusarium wilt images.

For this complex dataset, the proposed model still captures some features of the diseases of the strawberry plants, and gives 20.8% classification error on the testing set, where the label is predicted via the logistic regression model on the $z_{P,i}$. We also validate the model using the method stated in Section 5.3.2 and the independent classification model for which the ResNet101 is used gets 82.7% accuracy on the new generated images. This means the AP-CDE model is helpful for conditional generating more images of leaves of strawberry plants. Figure 5-8 plots the latent representations produced by these models. For Glow and Glow-Mix, the UMAP (McInnes et al., 2018) is used to reduce the dimension of latent z_i and plot its output in 2D. For the latent $z_{P,i}$ produced by AP-CDE, the plot between the first and the second dimension is provided, as well as the plot between the third and the fourth dimension. The Glow and Glow-Mix models do not provide any useful dimension reduction, while the latent variables provided by AP-CDE are separated over the four classes.



(a) Real images (b) Generated images

Figure 5-7. Real images and generated images using the AP-CDE model for the leaves of strawberry plants.



(a) Glow

(b) Glow-Mix

(c) AP-CDE, the first two dimensions

(d) AP-CDE, the third and fourth dimensions

Figure 5-8. The latent variables mapping from the leaves of strawberry plants images produced from the models. The UMAP is used to reduce the dimensions to 2 for Glow and Glow-Mix models.

CHAPTER 6
DISCUSSION AND FUTURE WORK

In Chapter 2, we propose a modification to the canonical stick-breaking construction, leading to an infinite mixture model that provides consistency for the number of clusters (like the MFM model) as well as easy implementation in posterior MCMC computation (like the Dirichlet process mixture model). Heiner et al. (2019) similarly proposes a tweak to the breaking proportion v_k under the framework of the Bayesian finite mixture prior, while we focus on the infinite mixture model and construct the theoretical properties on the number of clusters.

There are several extensions worth further pursuing. First, recovery of the true number of clusters under a *misspecified* model is still an open problem. In a recent work (Cai et al., 2021), it is theoretically shown that even a small amount of misspecification of the components will tend to result in over-estimation of the number of clusters in overfitted finite mixture models. Intuitively, this suggests that in addition to controlling the mixture weights to avoid small spurious clusters, it is also important to ensure that the family of component distributions is flexible enough to avoid severe model misspecification issues. Second, it would be interesting to investigate whether the combination of the quasi-Bernoulli infinite mixture framework and distance clustering approaches, such as the Laplacian-based approach (Rohe et al., 2011), can lead to a consistency result for the number of clusters.

The popular mixture models (such as Dirichlet process and Pitman–Yor process mixtures) are completely fine for the task of density estimation; nevertheless in some sense, a non-parametric mixing measure entails some misspecification under an indefinitely increasing n, which inherently assumes the number of clusters growing with n, hence conflicting with the popular modeling view where there is a fixed ground-truth k_0. Our insight is that for any fixed n, we have an exchangeable partition probability function (which is calculated by integrating out the mixture weights of those non-occupied components) that gives a discrete distribution for $K \in \{1, \dots, n\}$. Therefore, we can calibrate the asymptotic behavior of the probability function to produce a consistent estimator, via either controlling α in the Dirichlet process mixture or ϵ as we do in our quasi-Bernoulli model.

In Chapter 3, we present an approach for using optimization as a modeling tool to form a class of augmented likelihoods. These likelihoods enjoy a generative interpretation, with the use of latent variables z and constraints via the conditional distribution of y given z. Hence, they are amenable to the inference in the Bayesian framework, and in turn allow uncertainty quantification. We demonstrate several computational and modeling advantages over related Gibbs posterior alternatives in the literature.

In our present examples, we have focused on well-behaved loss functions with unique optima, which can be obtained efficiently with high numerical accuracy. Moving beyond this relatively clean setting merits future study as extensions and generalizations of the bridged posterior. Many problems, such as those with non-convex objectives functions, entail losses featuring multiple local optima (Zhang et al., 2020). The solution z returned by the algorithm at convergence may depend on the choice of initialization $\tilde{z}$ in these cases, where our approach has made use of a well-defined solution as input in a hierarchy. One generalization that may be fruitful is to assign a probability distribution over $\tilde{z}$, enabling us to view the optimization procedure as an algorithm mapping to another distribution for z. Second, many popular optimization algorithms, including stochastic variants of schemes such as gradient descent and early stopping, may produce approximately optimal solutions. In such cases, it may be more appropriate to model $\Pi(z \mid y, \lambda)$ to carry some uncertainty reflecting numerical errors or stopping criteria, in place of the point mass used in our current formulation. It is interesting to explore further connections to areas including Bayesian probabilistic numerical methodology (Cockayne et al., 2019) and optimization-based frequentist confidence intervals for constrained problems (Batlle et al., 2023).

In Chapter 4, we introduce an intuitive approach to account for sub-problems within a Bayesian statistical model. The idea of using a continuous shrinkage kernel on the partial gradient of the sub-problem loss is designed to yield straightforward inference and computational routines. Rather than requiring bespoke samplers to handle constraints or optimization sub-problems, the methodology is designed so that standard gradient-based posterior sampling can be readily

applied. The methodology is supported via asymptotic theory, and its performance in finite samples is supported empirically through simulation and case studies.

While this method is equipped to handle a broad variety of problems, it is currently limited to differentiable sub-problem losses h. A natural avenue for future work, then, would extend these ideas to non-differentiable losses such as hinge loss. Proximal mappings provide one route for handling such cases (Polson et al., 2015), and recent ideas in the Bayesian literature for handling constraints and priors using these optimization-centric ideas (Heng et al., 2023; Zhou et al., 2024) are likely to be fruitful in similarly extending the gradient-bridged framework. Additionally, for non-convex problems such as phase retrieval (Bahmani and Romberg, 2017), it may be advantageous to solve a convex relaxation of the problem while incorporating additional priors to keep z near the approximate solution.

In Chapter 5, we extends the normalizing flow neural network to the task of CDE. It produces a generative model for high-dimensional data that can incorporate information from external predictors. Importantly, by using only a subset of the one-to-one transform from the high-dimensional data, a useful dimension reduction is achieved, in which the low-dimensional representation is empirically sufficient to characterize the changes of the response variable due to the predictor.

A number of neural network-based models for CDE have appeared in the literature (see, e.g., recent reviews Ambrogioni et al. (2017); Rothfuss et al. (2019)). Nevertheless, the authors want to emphasize the versatility and simplicity of the proposed approach. AP-CDE can work with any existing normalizing flow network architecture, with only a modification of the base distribution density from a normal one to the product of a prior distribution and a likelihood function.

There are several interesting directions for future work. First, there is a connection of the strategy for new photo synthesis —"keeping z_P, sampling $\tilde{z}_N$ and pulling back via T^{-1}"— to the popular practice of "data augmentation in deep learning" (Shorten and Khoshgoftaar, 2019). Conventionally, to counter the small training sample problem, especially in image modeling, one relies on techniques such as geometric transformations, color space augmentation and random

erasing. Nevertheless, there is a recent trend in using another neural network for data augmentation, such as adversarial training and neural style transfer. The proposed AP-CDE can be considered another solution. Second, the normalizing flow networks can be too flexible, in the sense that they could transform a data distribution approximately to *any* latent distribution. This is likely why the unsupervised mixture of Gaussian latent distribution fails to explain the variations in the observed space. The proposal of using the predictor-conditional distribution shows that there is room to make the latent variable more interpretable; nevertheless, caution should be taken and additional validation methods could be developed.

APPENDIX A
APPENDIX FOR CHAPTER 2

A.1 Proofs

Proof of Theorem 2-1.

The conditional probability mass function of the assignment variables $c = (c_1, \ldots, c_n)$ is

$$\mathrm{p}(c \mid b_1, b_2, \ldots, \beta_1, \beta_2, \ldots) = \prod_{k=1}^{\infty} (1-\beta_k b_k)^{n_k} (\beta_k b_k)^{m_k},$$

where $n_k = \sum_{i=1}^{n} \mathbb{1}(c_i = k)$ and $m_k = \sum_{i=1}^{n} \mathbb{1}(c_i > k)$. Define $M(c) := \max\{c_1, \ldots, c_n\}$ and $Q_k := \tilde{p} + (1-\tilde{p}) I_\epsilon(m_k + \alpha, n_k + 1)/\epsilon^\alpha$. Then

$$\begin{aligned}
\mathrm{p}(c) &= \prod_{k=1}^{\infty} \int_0^1 \left(\tilde{p}(1-\beta_k)^{n_k}(\beta_k)^{m_k} + (1-\tilde{p})(1-\epsilon\beta_k)^{n_k}(\epsilon\beta_k)^{m_k} \right) \alpha \beta_k^{\alpha-1} \, d\beta_k \\
&= \prod_{k=1}^{\infty} \left(\tilde{p}\alpha \mathrm{B}(n_k+1, m_k+\alpha) + (1-\tilde{p}) \frac{\alpha}{\epsilon^\alpha} \mathrm{B}(n_k+1, m_k+\alpha) \int_0^\epsilon \frac{(1-x)^{n_k}(x)^{m_k+\alpha-1}}{\mathrm{B}(n_k+1, m_k+\alpha)} \, dx \right) \\
&\overset{\text{(a)}}{=} \prod_{k=1}^{M(c)} \left(\tilde{p}\alpha \mathrm{B}(n_k+1, m_k+\alpha) + (1-\tilde{p}) \frac{\alpha}{\epsilon^\alpha} \mathrm{B}(n_k+1, m_k+\alpha) I_\epsilon(m_k+\alpha, n_k+1) \right) \\
&= \prod_{k=1}^{M(c)} \frac{\alpha \Gamma(n_k+1)\Gamma(m_k+\alpha)}{\Gamma(n_k+m_k+\alpha+1)} Q_k \\
&\overset{\text{(b)}}{=} \left(\prod_{k=1}^{M(c)} \Gamma(n_k+1) \right) \prod_{k=1}^{M(c)} \frac{\alpha Q_k \Gamma(m_k+\alpha)}{(m_{k-1}+\alpha)\Gamma(m_{k-1}+\alpha)} \\
&= \frac{\Gamma(\alpha)}{\Gamma(n+\alpha)} \left(\prod_{k=1}^{M(c)} \Gamma(n_k+1) \right) \prod_{k=1}^{M(c)} \frac{\alpha Q_k}{m_{k-1}+\alpha}.
\end{aligned}$$

In step (a), we use the fact that for all $k > M(c)$, we have $n_k = 0$ and $m_k = 0$, and thus, everything cancels for such k since $B(1, \alpha) = 1/\alpha$ and $I_\epsilon(\alpha, 1) = \epsilon^\alpha$. In step (b), we use the fact that $n_k + m_k = m_{k-1}$, and thus, $\Gamma(n_k + m_k + \alpha + 1) = (m_{k-1} + \alpha)\Gamma(m_{k-1} + \alpha)$.

Define $g_k := \sum_{i=1}^{n} \mathbb{1}(c_i \geq k) = \sum_{l=k}^{\infty} n_l$, and note that $g_k = m_{k-1}$. Let $\mathcal{A}_c$ be the partition of $\{1, \ldots, n\}$ induced by c. Fix a partition $\mathcal{A} = \{A_1, \ldots, A_t\}$. When $\mathcal{A}_c = \mathcal{A}$, there are exactly t unique values among $c_1, \ldots, c_n$. Let $k_1 < k_2 < \cdots < k_t$ denote these unique values, and set $k_0 = 0$. For $k_{j-1} < k < k_j$, we have $n_k = 0$ and $g_k = g_{k+1} = g_{k_j}$, and thus $Q_k = \tilde{p} + (1-\tilde{p})\epsilon^{g_{k_j}}$.

Meanwhile, for $k = k_j$, we have $n_k = n_{k_j}$ and $g_{k+1} = g_{k_{j+1}}$, and thus it follows that $Q_k = \tilde{p} + (1-\tilde{p})I_\epsilon(g_{k_{j+1}} + \alpha, n_{k_j} + 1)/\epsilon^\alpha$. Hence, for all c such that $\mathcal{A}_c = \mathcal{A}$, we have

$$\mathrm{p}(c) = \frac{\Gamma(\alpha)}{\Gamma(n+\alpha)}\left(\prod_{j=1}^{t}\Gamma(n_{k_j}+1)\right)\prod_{j=1}^{t}U_j(c)$$

where

$$U_j(c) = \left(\frac{\alpha\tilde{p} + \alpha(1-\tilde{p})I_\epsilon(g_{k_{j+1}}+\alpha, n_{k_j}+1)/\epsilon^\alpha}{g_{k_j}+\alpha}\right)\left(\frac{\alpha\tilde{p}+\alpha(1-\tilde{p})\epsilon^{g_{k_j}}}{g_{k_j}+\alpha}\right)^{d_j}$$

where $d_j = k_j - k_{j-1} - 1$. Since there is a unique permutation $\sigma = (\sigma_1, \ldots, \sigma_t)$ of $\{1, \ldots, t\}$ such that $A_{\sigma_j} = \{i : c_i = k_j\}$ for all $j \in \{1, \ldots, t\}$, the mapping between $\{c : \mathcal{A}_c = \mathcal{A}\}$ and $\{(\sigma, d_1, \ldots, d_t) : \sigma \in S_t,\ d_1, \ldots, d_t \in \mathbb{N}\}$ is a bijection. (Here, S_t is the set of all permutations of $\{1, \ldots, t\}$, and $\mathbb{N} := \{0, 1, 2, \ldots\}$.) Let $n^*_j := |A_j|$ and $g^*_j(\sigma) := \sum_{l=j}^{t} n^*_{\sigma_l}$. For the value of $(\sigma, d_{1:t})$ that corresponds to c, we have $g_{k_j} = g^*_j(\sigma)$ and $n_{k_j} = |A_{\sigma_j}| = n^*_{\sigma_j}$, and thus, $U_j(c) = U^*_j(\sigma, d_{1:t})$ where

$$U^*_j(\sigma, d_{1:t}) := \left(\frac{\alpha\tilde{p} + \alpha(1-\tilde{p})I_\epsilon(g^*_{j+1}(\sigma)+\alpha, n^*_{\sigma_j}+1)/\epsilon^\alpha}{g^*_j(\sigma)+\alpha}\right)\left(\frac{\alpha\tilde{p}+\alpha(1-\tilde{p})\epsilon^{g^*_j(\sigma)}}{g^*_j(\sigma)+\alpha}\right)^{d_j}.$$

Summing over d_j and using $\sum_{d=0}^{\infty} x^d = 1/(1-x)$ for $|x| < 1$, we have

$$\sum_{d_j=0}^{\infty} U^*_j(\sigma, d_{1:t}) = \frac{\alpha\tilde{p} + \alpha(1-\tilde{p})I_\epsilon(g^*_{j+1}(\sigma)+\alpha, n^*_{\sigma_j}+1)/\epsilon^\alpha}{g^*_j(\sigma) + \alpha(1-\tilde{p})(1-\epsilon^{g^*_j(\sigma)})}.$$

Note that $U^*_j(\sigma, d_{1:t})$ depends on $d_{1:t}$ only through d_j. Therefore,

$$\begin{aligned}
\mathrm{p}_{\epsilon,n}(\mathcal{A}) &= \sum_{c\,:\,\mathcal{A}_c=\mathcal{A}} \mathrm{p}(c) = \frac{\Gamma(\alpha)}{\Gamma(n+\alpha)}\left(\prod_{j=1}^{t}\Gamma(n^*_j+1)\right)\sum_{\sigma\in S_t}\sum_{d_1=0}^{\infty}\cdots\sum_{d_t=0}^{\infty}\prod_{j=1}^{t}U^*_j(\sigma, d_{1:t}) \\
&= \frac{\Gamma(\alpha)}{\Gamma(n+\alpha)}\left(\prod_{j=1}^{t}\Gamma(n^*_j+1)\right)\sum_{\text{all }\sigma}\prod_{j=1}^{t}\sum_{d_j=0}^{\infty}U^*_j(\sigma, d_{1:t}) \\
&= \frac{\alpha^t\Gamma(\alpha)}{\Gamma(n+\alpha)}\left(\prod_{j=1}^{t}\Gamma(n^*_j+1)\right)\sum_{\text{all }\sigma}\prod_{j=1}^{t}\frac{\tilde{p}+(1-\tilde{p})I_\epsilon(g^*_{j+1}(\sigma)+\alpha, n^*_{\sigma_j}+1)/\epsilon^\alpha}{g^*_j(\sigma)+\alpha(1-\tilde{p})(1-\epsilon^{g^*_j(\sigma)})}.
\end{aligned}$$

This proves the result. □

Proof of Lemma 2-1.

The general approach of the proof follows the technique of Miller (2019). The conditional probability mass function of the assignment variables is

$$\mathrm{p}(c \mid b_1, b_2, \ldots, \beta_1, \beta_2, \ldots) = \prod_{k=1}^{\infty} (1-\beta_k b_k)^{n_k} (\beta_k b_k)^{g_{k+1}},$$

where $n_k = \sum_{i=1}^{n} \mathbb{1}(c_i = k)$ and $g_k = \sum_{i=1}^{n} \mathbb{1}(c_i \geq k)$. Let $M(c) = \max\{c_1, \ldots, c_n\}$. Then

$$\begin{aligned}
\mathrm{p}(c) &= \prod_{k=1}^{\infty} \int_0^1 \Big(\tilde{p}(1-\beta_k)^{n_k}(\beta_k)^{g_{k+1}} + (1-\tilde{p})\mathbb{1}(g_{k+1}=0)\Big)\alpha\beta_k^{\alpha-1}\, d\beta_k \\
&= \prod_{k=1}^{M(c)} \Big(\tilde{p}\alpha \mathrm{B}(n_k+1, g_{k+1}+\alpha) + (1-\tilde{p})\mathbb{1}(g_{k+1}=0)\Big) \\
&= \left(\prod_{k=1}^{M(c)-1} \frac{\tilde{p}\alpha\Gamma(n_k+1)\Gamma(g_{k+1}+\alpha)}{\Gamma(n_k+g_{k+1}+\alpha+1)}\right)\big(\tilde{p}\alpha\mathrm{B}(n_{M(c)}+1,\alpha)+1-\tilde{p}\big) \\
&= \left(\prod_{k=1}^{M(c)-1} \Gamma(n_k+1)\right)\left(\prod_{k=1}^{M(c)-1} \frac{\tilde{p}\alpha\Gamma(g_{k+1}+\alpha)}{\Gamma(g_k+\alpha+1)}\right)\big(\tilde{p}\alpha\mathrm{B}(n_{M(c)}+1,\alpha)+1-\tilde{p}\big) \\
&= \frac{\Gamma(\alpha)}{\Gamma(n+\alpha)}\left(\prod_{k=1}^{M(c)} \Gamma(n_k+1)\right)\left(\prod_{k=1}^{M(c)-1} \frac{\tilde{p}\alpha}{g_k+\alpha}\right)\left(\frac{\tilde{p}\alpha + (1-\tilde{p})/\mathrm{B}(n_{M(c)}+1,\alpha)}{n_{M(c)}+\alpha}\right).
\end{aligned}$$

Let $\mathcal{A}(c)$ denote the partition of $\{1, \ldots, n\}$ corresponding to c. For fixed $\mathcal{A} = \{A_1, \ldots, A_t\}$, when $\mathcal{A}(c) = \mathcal{A}$, there are exactly t unique values among $c_1, \ldots, c_n$. Let $k_1 < k_2 < \cdots < k_t$ denote these unique values, and set $k_0 = 0$. For $k_{j-1} < k \leq k_j$, we have $g_k = g_{k_j}$. Hence, for c satisfying $\mathcal{A}(c) = \mathcal{A}$, we have $\mathrm{p}(c) = (\Gamma(\alpha)/\Gamma(n+\alpha))\big(\prod_{j=1}^{t}\Gamma(n_{k_j}+1)\big)\prod_{j=1}^{t} U_j(c)$, where $U_j(c) = (\tilde{p}\alpha/(g_{k_j}+\alpha))^{d_j}$ for $1 \leq j < t$ and

$$U_t(c) = \frac{\tilde{p}\alpha + (1-\tilde{p})/\mathrm{B}(n_{k_t}+1,\alpha)}{n_{k_t}+\alpha}\left(\frac{\tilde{p}\alpha}{g_{k_t}+\alpha}\right)^{d_t-1} = \left(1 + \frac{(1-\tilde{p})/\tilde{p}\alpha}{\mathrm{B}(n_{k_t}+1,\alpha)}\right)\left(\frac{\tilde{p}\alpha}{g_{k_t}+1}\right)^{d_t}$$

where $d_j = k_j - k_{j-1}$ for $j = 1, \ldots, t$.

Since there is a unique permutation $\sigma = (\sigma_1, \ldots, \sigma_t) \in S_t$ such that $A_{\sigma_j} = \{i : c_i = k_j\}$, the mapping between $\{c : \mathcal{A}(c) = \mathcal{A}\}$ and $\left\{(\sigma, d_1, \ldots, d_t) : \sigma \in S_t,\ d_1, \ldots, d_t \in \{1, 2, \ldots\}\right\}$ is a bijection. Letting $n_j = |A_j|$, we have

$$\mathrm{p}_{0,n}(\mathcal{A}) = \sum_{c\,:\,\mathcal{A}(c)=\mathcal{A}} \mathrm{p}(c) = \frac{\Gamma(\alpha)}{\Gamma(n+\alpha)} \left(\prod_{j=1}^{t} \Gamma(n_j + 1) \right) \sum_{\sigma \in S_t} \sum_{d_1=1}^{\infty} \cdots \sum_{d_t=1}^{\infty} \prod_{j=1}^{t} U_j(c), \tag{A-1}$$

treating c as a function of $\sigma, d_1, \ldots, d_t$. Changing the order of summations and multiplication, defining $g_j(\sigma) = \sum_{l=j}^{t} n_{\sigma_l}$, and using the geometric series $\sum_{d=1}^{\infty} x^d = x/(1-x)$ for $x \in (0,1)$,

$$\begin{aligned}
\sum_{d_1=1}^{\infty} \cdots \sum_{d_t=1}^{\infty} \prod_{j=1}^{t} U_j(c) &= \left(1 + \frac{(1-\tilde{p})/\tilde{p}\alpha}{\mathrm{B}(n_{\sigma_t}+1, \alpha)}\right) \prod_{j=1}^{t} \sum_{d_j=1}^{\infty} \left(\frac{\tilde{p}\alpha}{g_j(\sigma)+\alpha}\right)^{d_j} \\
&= \left(1 + \frac{(1-\tilde{p})/\tilde{p}\alpha}{\mathrm{B}(n_{\sigma_t}+1, \alpha)}\right) \prod_{j=1}^{t} \frac{\tilde{p}\alpha}{g_j(\sigma)+\alpha(1-\tilde{p})} \\
&= \alpha^t \prod_{j=1}^{t} \frac{\tilde{p} + \mathbb{1}(j=t)(1-\tilde{p})/(\alpha \mathrm{B}(n_{\sigma_t}+1, \alpha))}{g_j(\sigma)+\alpha(1-\tilde{p})}.
\end{aligned}$$

Combining with Equation A-1, this proves the result. □

We provide the complete statements of the two results from Nobile (1994).

Theorem A-1. *Nobile (1994, Corollary 3.1) Assume $\phi \in \Omega'$ is identifiable up to permutation of the mixture components. Let Π_0 be the prior on Ω under the model defined by Equations 2-2 and 2-6, and assume $\Pi_0(\{\phi : \exists\, i \neq j \text{ such that } \theta_i = \theta_j\}) = 0$. Let Π'_0 be the corresponding prior on Ω' induced by η. Then there is a subset $\Omega'_0 \subset \Omega'$ with $\Pi'_0(\Omega'_0) = 1$ such that for any $\phi_0 = (k_0, w_1^0, \ldots, w_{k_0}^0, \theta_1^0, \ldots, \theta_{k_0}^0) \in \Omega'_0$, if $y_1, y_2, \ldots \mid \phi_0 \overset{iid}{\sim} P_{\phi_0}$, then as $n \to \infty$, we have*

$$p_{\epsilon=0}(\phi \in D \mid y_{1:n}) \to \mathbb{1}(\phi_0 \in D') \quad \text{a.s.}[P_{\phi_0}],$$

for any measurable subset $D' \subset \Omega'$ and $D = \{\phi \in \Omega : \eta(\phi) \in D'\}$.

Theorem A-2. *Nobile (1994, Proposition 3.5) Under the same assumptions of Theorem A-1,*

$$p_{\epsilon=0}(K = k \mid y_{1:n}) \to \mathbb{1}(k_0 = k) \quad \text{a.s.}[P_{\phi_0}].$$

Proof of Theorem 2-2.

The first result is proved by Theorem A-2. The second result can be proved as follows. Using the Theorem A-1, there is a subspace Ω_0' as described in the theorem such that if $\phi_0 \in \Omega_0'$ then the posterior distribution of (w, θ) given $K = k_0$ and $y_{1:n}$ will converge to a uniform distribution in which with equal probability the (w, θ) is one of the permutations of $(w_1^0, \theta_1^0), \ldots, (w_{k_0}^0, \theta_{k_0}^0)$. This is because the transformation η maps all of the permutations of $(w_1^0, \theta_1^0), \ldots, (w_{k_0}^0, \theta_{k_0}^0)$ into the same one with a specific order. Define the random variables $N_k = \sum_{i=1}^n \mathbb{1}(c_i = k)$ for $k = 1, \ldots, k_0$. Then

$$\begin{aligned}
&p_{\epsilon=0}(N_k = 0 \mid K = k_0, y_{1:n}) \\
&= \mathrm{p}_{\epsilon=0}(\cap_{i=1}^n \{c_i \neq k\} \mid K = k_0, y_{1:n}) \\
&= \int \mathrm{p}_{\epsilon=0}(\cap_{i=1}^n \{c_i \neq k\} \mid K = k_0, w, \theta, y_{1:n}) \mathbb{P}_{\epsilon=0}(dw, d\theta \mid K = k_0, y_{1:n}) \\
&= \int \prod_{i=1}^n \left(1 - \frac{w_k f_{\theta_k}(y_i)}{\sum_{l=1}^{k_0} w_l f_{\theta_l}(y_i)}\right) \mathbb{P}_{\epsilon=0}(dw, d\theta \mid K = k_0, y_{1:n}) \\
&\leq \int \prod_{i=1}^{n_0} \left(1 - \frac{w_k f_{\theta_k}(y_i)}{\sum_{l=1}^{k_0} w_l f_{\theta_l}(y_i)}\right) \mathbb{P}_{\epsilon=0}(dw, d\theta \mid K = k_0, y_{1:n}) \qquad \text{(A-2)}
\end{aligned}$$

for any given positive integer n_0 and $n \geq n_0$. Using the weak convergence of the posterior distribution of the (w, θ) and the fact that the integrand is bounded and f_θ is continuous at all θ_k^0's, the Equation A-2 converges to

$$\sum_{k=1}^{k_0} \frac{1}{k_0} \prod_{i=1}^{n_0} \left(1 - \frac{w_k^0 f_{\theta_k^0}(y_i)}{\sum_{l=1}^{k_0} w_l^0 f_{\theta_l^0}(y_i)}\right)$$

a.s.$[P_{\phi_0}]$ when $n \to \infty$. Since Equation A-2 holds for any positive integer n_0, we have

$$\limsup_{n\to\infty} p_{\epsilon=0}(N_k = 0 \mid K = k_0, y_{1:n}) \leq \sum_{k=1}^{k_0} \frac{1}{k_0} \prod_{i=1}^{\infty} \left(1 - \frac{w_k^0 f_{\theta_k^0}(y_i)}{\sum_{l=1}^{k_0} w_l^0 f_{\theta_l^0}(y_i)}\right).$$

For every $k = 1, \ldots, k_0$, there exists a measurable set D_k with non-zero measure such that $f_{\theta_k^0}(y) \geq \delta_k > 0$ when $y \in D_k$, and all $f_{\theta_l^0}(y)$ $(l = 1, \ldots, k_0)$ are finite on D_k. For every $k = 1, \ldots, k_0$, there exists a sequence $n_{k1}, n_{k2}, \ldots$ such that $y_{n_{ki}} \in D_k$ $(i = 1, 2, \ldots)$, a.s.$[P_{\phi_0}]$. Hence, for all $i \geq 1$, $f_{\theta_k^0}(y_{n_{ki}}) \geq \delta_k$ and $\sum_{l=1}^{k_0} w_l^0 f_{\theta_l^0}(y_{n_{ki}}) \leq M$ for some M. Therefore,

$$\prod_{i=1}^{\infty} \left(1 - \frac{w_k^0 f_{\theta_k^0}(y_i)}{\sum_{l=1}^{k_0} w_l^0 f_{\theta_l^0}(y_i)}\right) \leq \prod_{i=1}^{\infty} \left(1 - \frac{w_k^0 f_{\theta_k^0}(y_{n_{ki}})}{\sum_{l=1}^{k_0} w_l^0 f_{\theta_l^0}(y_{n_{ki}})}\right) \leq \prod_{i=1}^{\infty} \left(1 - \frac{w_k^0 \delta}{M}\right) = 0,$$

which leads to $p_{\epsilon=0}(N_k = 0 \mid K = k_0, y_{1:n}) \xrightarrow{\text{a.s.}} 0$ as $n \to \infty$. Hence, given $K = k_0$, the posterior probability of having k_0 clusters is

$$\begin{aligned}
&\mathrm{p}_{\epsilon=0}(T = k_0 \mid K = k_0, y_{1:n}) = \mathrm{p}_{\epsilon=0}(N_1 > 0, \ldots, N_{k_0} > 0 \mid K = k_0, y_{1:n}) \\
&= 1 - \mathrm{p}_{\epsilon=0}(\cup_{k=1}^{k_0} \{N_k = 0\} \mid K = k_0, y_{1:n}) \\
&\geq 1 - \sum_{k=1}^{k_0} \mathrm{p}_{\epsilon=0}(N_k = 0 \mid K = k_0, y_{1:n}) \to 1
\end{aligned}$$

a.s.$[P_{\phi_0}]$ as $n \to \infty$. Therefore,

$$\begin{aligned}
\mathrm{p}_{\epsilon=0}(T = k_0 \mid y_{1:n}) &= \sum_{k=k_0}^{\infty} \mathrm{p}_{\epsilon=0}(T = k_0 \mid K = k, y_{1:n}) \mathrm{p}_{\epsilon=0}(K = k \mid y_{1:n}) \\
&\geq \mathrm{p}_{\epsilon=0}(T = k_0 \mid K = k_0, y_{1:n}) \mathrm{p}_{\epsilon=0}(K = k_0 \mid y_{1:n}) \to 1
\end{aligned}$$

a.s.$[P_{\phi_0}]$ as $n \to \infty$. □

Proof of Theorem 2-3.

For a given partition $\mathcal{A}$, let $t = |\mathcal{A}|$. Define the following notation to represent the factors in $\mathrm{p}_{0,n}(\mathcal{A})$ and $\mathrm{p}_{\epsilon,n}(\mathcal{A})$, respectively:

$$U_j(\sigma) := \frac{\tilde{p} + \mathbb{1}(j = t)(1 - \tilde{p})/(\alpha \mathrm{B}(\alpha, n_{\sigma_t} + 1))}{g_j(\sigma) + \alpha(1 - \tilde{p})}$$
$$V_j(\sigma) := \frac{\tilde{p} + (1 - \tilde{p}) I_\epsilon(g_{j+1}(\sigma) + \alpha, n_{\sigma_j} + 1)/\epsilon^\alpha}{g_j(\sigma) + \alpha(1 - \tilde{p})(1 - \epsilon^{g_j(\sigma)})}$$

for $j = 1, \ldots, t$. When $j < t$, we have

$$U_j(\sigma) \leq V_j(\sigma) \tag{A-3}$$

since $(1 - \tilde{p}) I_\epsilon(g_{j+1}(\sigma) + \alpha, n_{\sigma_j} + 1)/\epsilon^\alpha > 0$ and
$g_j(\sigma) + \alpha(1 - \tilde{p}) > g_j(\sigma) + \alpha(1 - \tilde{p})(1 - \epsilon^{g_j(\sigma)}) > 0$.

Meanwhile, for the case of $j = t$, we have

$$\begin{aligned}
\frac{U_t(\sigma)}{V_t(\sigma)} &\overset{\text{(a)}}{\leq} \frac{\tilde{p} + (1 - \tilde{p})/(\alpha \mathrm{B}(\alpha, n_{\sigma_t} + 1))}{\tilde{p} + (1 - \tilde{p}) I_\epsilon(\alpha, n_{\sigma_t} + 1)/\epsilon^\alpha} \\
&\overset{\text{(b)}}{\leq} \frac{\tilde{p}\alpha \mathrm{B}(\alpha, n_{\sigma_t} + 1) + 1 - \tilde{p}}{\tilde{p}\alpha \mathrm{B}(\alpha, n_{\sigma_t} + 1) + (1 - \tilde{p})(1 - \alpha\epsilon n_{\sigma_t}/(\alpha + 1))} \\
&= 1 + \frac{(1 - \tilde{p})\alpha\epsilon n_{\sigma_t}/(\alpha + 1)}{\tilde{p}\alpha \mathrm{B}(\alpha, n_{\sigma_t} + 1) + (1 - \tilde{p})(1 - \alpha\epsilon n_{\sigma_t}/(\alpha + 1))} \\
&\overset{\text{(c)}}{\leq} 1 + \frac{\alpha n \epsilon}{\alpha + 1 - \alpha\epsilon n},
\end{aligned} \tag{A-4}$$

where (a) uses $g_{\sigma_k} + \alpha(1 - \tilde{p}) > g_{\sigma_k} + \alpha(1 - \tilde{p})(1 - \epsilon^{g_{\sigma_k}}) > 0$, (b) uses

$$\frac{\alpha \mathrm{B}(\alpha, n_{\sigma_t} + 1) I_\epsilon(\alpha, n_{\sigma_t} + 1)}{\epsilon^\alpha} = \frac{\alpha}{\epsilon^\alpha} \int_0^\epsilon x^{\alpha - 1}(1 - x)^{n_{\sigma_t}}\, dx \geq \frac{\alpha}{\epsilon^\alpha} \int_0^\epsilon x^{\alpha - 1}(1 - n_{\sigma_t} x)\, dx$$
$$= 1 - \frac{\alpha\epsilon n_{\sigma_t}}{\alpha + 1},$$

and (c) uses $\tilde{p}\alpha \mathrm{B}(\alpha, n_{\sigma_t} + 1) > 0$, $n_{\sigma_t} \leq n$, and the assumption that $\epsilon \leq 1/n$.

Using the exchangeable partition probability functions in Theorem 2-1 and Lemma 2-1, along with Equations A-3, A-4 and A-7, we have

$$\frac{\mathrm{p}_{0,n}(\mathcal{A})}{\mathrm{p}_{\epsilon,n}(\mathcal{A})} = \frac{\sum_{\sigma\in S_t}\prod_{j=1}^{t} U_j(\sigma)}{\sum_{\sigma\in S_t}\prod_{j=1}^{t} V_j(\sigma)} \leq 1 + \frac{\alpha n\epsilon}{\alpha + 1 - \alpha\epsilon n} \tag{A-5}$$

for all partitions $\mathcal{A}$. Therefore, the Kullback–Leibler divergence satisfies

$$D_{\mathrm{KL}}(\mathrm{p}_{0,n}\|\mathrm{p}_{\epsilon,n}) = \sum_{\mathcal{A}\in\cup_{t=1}^{n}\mathcal{H}_t(n)} \mathrm{p}_{0,n}(\mathcal{A}) \log\frac{\mathrm{p}_{0,n}(\mathcal{A})}{\mathrm{p}_{\epsilon,n}(\mathcal{A})} \leq \sum_{\mathcal{A}} \mathrm{p}_{0,n}(\mathcal{A})\frac{\alpha n\epsilon}{\alpha + 1 - \alpha\epsilon n} = \frac{\alpha n\epsilon}{\alpha + 1 - \alpha\epsilon n},$$

where the sum is over all partitions of $\{1, \ldots, n\}$ and the inequality uses $\log(x) \leq x - 1$. The result follows by Pinsker's inequality. □

In the proof of Theorem 2-4, we employ the following two inequalities. Let $x_i, y_i \geq 0$ for all i in some countable set I, such that $\sum_{i\in I} x_i > 0$ and $\sum_{i\in I} y_i > 0$. First, if $A \subseteq I$ then

$$\begin{aligned}\left|\frac{\sum_{i\in A} x_i}{\sum_{i\in I} x_i} - \frac{\sum_{i\in A} y_i}{\sum_{i\in I} y_i}\right| &= \frac{|\sum_{i'\in A}\sum_{i\in I} x_{i'}y_i - \sum_{i'\in A}\sum_{i\in I} x_i y_{i'}|}{\sum_{i',i\in I} x_{i'}y_i} \\ &= \frac{|\sum_{i'\in A}\sum_{i\in A^c} x_{i'}y_i - \sum_{i'\in A}\sum_{i\in A^c} x_i y_{i'}|}{\sum_{i',i\in I} x_{i'}y_i} \\ &\leq \frac{\sum_{i'\in A}\sum_{i\in A^c}|x_{i'}y_i - x_i y_{i'}|}{\sum_{i',i\in I} x_{i'}y_i}\end{aligned} \tag{A-6}$$

where $A^c = I \setminus A$. Second, if $a_i \geq 0$, $y_i > 0$ for all $i \in I$, $A \subseteq I$, and $\sum_{i\in I} a_i y_i > 0$, then

$$\frac{\sum_{i\in A} a_i x_i}{\sum_{i\in I} a_i y_i} = \frac{\sum_{i\in A} a_i (x_i/y_i) y_i}{\sum_{i\in I} a_i y_i} \leq \frac{\sum_{i\in A} a_i (\max_{j\in A} x_j/y_j) y_i}{\sum_{i\in I} a_i y_i} \leq \max_{i\in A} x_i/y_i. \tag{A-7}$$

Proof of Theorem 2-4.

In Theorem 2-2, we proved posterior consistency of the number of clusters in the case of $\epsilon = 0$, so we only need to show that $|\mathrm{p}_{\epsilon(n)}(T = t \mid y_{1:n}) - \mathrm{p}_{\epsilon=0}(T = t \mid y_{1:n})| \to 0$ as $n \to \infty$. We abbreviate $y = y_{1:n}$ to reduce notational clutter. First, using Equation 2-5, for any integer

$1 \le t \le n$,

$$
\begin{aligned}
&|\mathrm{p}_\epsilon(T = t \mid y) - \mathrm{p}_{\epsilon=0}(T = t \mid y)| \\
&= \left| \frac{\sum_{\mathcal{A} \in \mathcal{H}_t(n)} \mathrm{p}(y \mid \mathcal{A}) \mathrm{p}_{\epsilon,n}(\mathcal{A})}{\sum_{\mathcal{A} \in \cup_{t=1}^n \mathcal{H}_t(n)} \mathrm{p}(y \mid \mathcal{A}) \mathrm{p}_{\epsilon,n}(\mathcal{A})} - \frac{\sum_{\mathcal{A} \in \mathcal{H}_t(n)} \mathrm{p}(y \mid \mathcal{A}) \mathrm{p}_{0,n}(\mathcal{A})}{\sum_{\mathcal{A} \in \cup_{t=1}^n \mathcal{H}_t(n)} \mathrm{p}(y \mid \mathcal{A}) \mathrm{p}_{0,n}(\mathcal{A})} \right| \\
&\overset{(a)}{\le} \frac{\sum_{\mathcal{A}' \in \mathcal{H}_t(n)} \sum_{\mathcal{A} \in \cup_{l \ne t} \mathcal{H}_l(n)} \mathrm{p}(y \mid \mathcal{A}') \mathrm{p}(y \mid \mathcal{A}) |\mathrm{p}_{\epsilon,n}(\mathcal{A}') \mathrm{p}_{0,n}(\mathcal{A}) - \mathrm{p}_{0,n}(\mathcal{A}') \mathrm{p}_{\epsilon,n}(\mathcal{A})|}{\sum_{\mathcal{A}', \mathcal{A} \in \cup_{t=1}^n \mathcal{H}_t(n)} \mathrm{p}(y \mid \mathcal{A}') \mathrm{p}(y \mid \mathcal{A}) \mathrm{p}_{\epsilon,n}(\mathcal{A}') \mathrm{p}_{0,n}(\mathcal{A})} \\
&\overset{(b)}{\le} \max_{\mathcal{A}' \in \mathcal{H}_t(n)} \max_{\mathcal{A} \in \cup_{l \ne t} \mathcal{H}_l(n)} \left| 1 - \frac{\mathrm{p}_{0,n}(\mathcal{A}') \mathrm{p}_{\epsilon,n}(\mathcal{A})}{\mathrm{p}_{\epsilon,n}(\mathcal{A}') \mathrm{p}_{0,n}(\mathcal{A})} \right|,
\end{aligned}
\tag{A-8}
$$

where (a) and (b) are by Equations A-6 and A-7, respectively. To ensure that the denominators in the preceding display are nonzero, there needs to exist at least one partition $\mathcal{A} \in \cup_{t=1}^n \mathcal{H}_t(n)$ such that $\mathrm{p}(y \mid \mathcal{A}) > 0$, and this is indeed the case since (1) with probability 1, $p(y \mid \phi_0) > 0$, and (2) we assume the support of the prior $\mathcal{G}(\theta)$ contains a neighborhood of each θ_k^0 and the component density f_θ is continuous (with respect to θ) at each θ_k^0.

We use the same notation as in the proof of Theorem 2-3, where we have already proved that

$$
\frac{\mathrm{p}_{0,n}(\mathcal{A})}{\mathrm{p}_{\epsilon,n}(\mathcal{A})} \le 1 + \frac{\alpha n \epsilon}{\alpha + 1 - \alpha \epsilon n} \tag{A-9}
$$

for any partition $\mathcal{A}$. Next, we construct an upper bound on the reciprocal, $\mathrm{p}_{\epsilon,n}(\mathcal{A})/\mathrm{p}_{0,n}(\mathcal{A})$, by upper bounding $V_j(\sigma)/U_j(\sigma)$ for each j. We split the analysis into two cases: $j \le t-1$ and $j = t$.

Case 1: $j \le t - 1$. This implies $g_{j+1}(\sigma) \ge t - j \ge 1$ since every cluster has at least one element. Letting r denote the integer such that $0 < \alpha - r \le 1$ (or equivalently,

$\max(\alpha-1,0) \le r < \alpha$),

$$
\begin{aligned}
&\frac{1}{\epsilon^\alpha} I_\epsilon(g_{j+1}(\sigma)+\alpha, n_{\sigma_j}+1) \\
&\le \frac{1}{\epsilon^\alpha \mathrm{B}(g_{j+1}(\sigma)+\alpha, n_{\sigma_j}+1)} \int_0^\epsilon x^{g_{j+1}(\sigma)+\alpha-1}\,dx \\
&= \frac{\Gamma(g_j(\sigma)+\alpha+1)\epsilon^{g_{j+1}(\sigma)}}{\Gamma(g_{j+1}(\sigma)+\alpha)\Gamma(n_{\sigma_j}+1)(g_{j+1}(\sigma)+\alpha)} \\
&= \epsilon^{g_{j+1}(\sigma)} \frac{(g_j(\sigma)+\alpha)(g_j(\sigma)+\alpha-1)\cdots(\alpha-r)}{(g_{j+1}(\sigma)+\alpha)(g_{j+1}(\sigma)+\alpha-1)\cdots(\alpha-r)} \frac{1}{n_{\sigma_j}!} \\
&= \epsilon^{g_{j+1}(\sigma)} \left(\prod_{m=0}^{g_{j+1}(\sigma)+r} \frac{n_{\sigma_j}+\alpha-r+m}{\alpha-r+m} \right) \left(\prod_{m=1}^{n_{\sigma_j}} \frac{m+\alpha-r-1}{m} \right) \\
&\overset{\text{(a)}}{\le} \epsilon^{g_{j+1}(\sigma)} \left(1+\frac{n_{\sigma_j}}{\alpha-r}\right)^{g_{j+1}(\sigma)+r+1} \\
&\overset{\text{(b)}}{\le} \epsilon^{t-j} \left(1+\frac{n}{\alpha-r}\right)^{t-j+r+1}
\end{aligned}
$$

for all n sufficiently large, where (a) results from

$(n_{\sigma_j}+\alpha-r+m)/(\alpha-r+m) \le (n_{\sigma_j}+\alpha-r)/(\alpha-r)$ for $m = 0, 1, \ldots, g_{j+1}(\sigma)+r$ and

$(m+\alpha-r-1)/m \le 1$ for $m = 1, \ldots, n_{\sigma_j}$, and (b) uses $n_{\sigma_j} \le n$, $\epsilon(1+n/(\alpha-r)) < 1$ for all n sufficiently large since $\epsilon = \epsilon(n) = o(1/n)$, and $g_{j+1}(\sigma) \ge t-j$.

Combining this with

$$
\frac{g_j(\sigma)+\alpha(1-\tilde{p})}{g_j(\sigma)+\alpha(1-\tilde{p})(1-\epsilon^{g_j(\sigma)})} = 1 + \frac{\alpha(1-\tilde{p})\epsilon^{g_j(\sigma)}}{g_j(\sigma)+\alpha(1-\tilde{p})(1-\epsilon^{g_j(\sigma)})} \le 1+\alpha(1-\tilde{p})\epsilon, \qquad \text{(A-10)}
$$

we have

$$
\frac{V_j(\sigma)}{U_j(\sigma)} \le \left(1+\frac{1-\tilde{p}}{\tilde{p}}\epsilon^{t-j}\left(1+\frac{n}{\alpha-r}\right)^{t-j+r+1}\right)\left(1+\alpha(1-\tilde{p})\epsilon\right).
$$

Case 2: $j = t$. We have

$$
\frac{V_t(\sigma)}{U_t(\sigma)} \le 1+\alpha(1-\tilde{p})\epsilon
$$

since Equation A-10 holds when $j = t$ and

$$\frac{I_\epsilon(\alpha, n_{\sigma_t} + 1)}{\epsilon^\alpha} \leq \frac{1}{\epsilon^\alpha \mathrm{B}(\alpha, n_{\sigma_t} + 1)} \int_0^\epsilon x^{\alpha-1}\, dx = \frac{1}{\alpha \mathrm{B}(\alpha, n_{\sigma_t} + 1)}.$$

Thus, using the same expression as in Equation A-5,

$$\begin{aligned}
\frac{\mathrm{p}_{\epsilon,n}(\mathcal{A})}{\mathrm{p}_{0,n}(\mathcal{A})} &\leq \max_{\sigma \in S_t} \prod_{j=1}^{t} \frac{V_j(\sigma)}{U_j(\sigma)} \\
&\leq \big(1 + \alpha(1-\tilde{p})\epsilon\big)^t \prod_{j=1}^{t-1} \left(1 + \frac{1-\tilde{p}}{\tilde{p}} \epsilon^{t-j} \left(1 + \frac{n}{\alpha - r}\right)^{t-j+r+1}\right) \\
&\leq \big(1 + \alpha(1-\tilde{p})\epsilon\big)^n \left(1 + \frac{1-\tilde{p}}{\tilde{p}} \epsilon^2 \left(1 + \frac{n}{\alpha - r}\right)^{3+r}\right)^n \left(1 + \frac{1-\tilde{p}}{\tilde{p}} \epsilon \left(1 + \frac{n}{\alpha - r}\right)^{2+r}\right) \qquad \text{(A-11)}
\end{aligned}$$

since $t = |\mathcal{A}| \leq n$ and $\epsilon(1 + n/(\alpha - r)) < 1$ for n sufficiently large, because $\epsilon = \epsilon(n) = o(1/n)$.

Since $\epsilon(n) = o(1/n^{2+r})$, the upper bound on $\mathrm{p}_{\epsilon(n),n}(\mathcal{A})/\mathrm{p}_{0,n}(\mathcal{A})$ in Equation A-11 converges to 1 as $n \to \infty$. Meanwhile, Equation A-9 provides a lower bound that also converges to 1. Since these upper and lower bounds depend only on α, r, $\tilde{p}$, ϵ, and n, they hold uniformly over all $\mathcal{A} \in \cup_{t=1}^n \mathcal{H}_t(n)$. Hence, $\mathrm{p}_{\epsilon(n),n}(\mathcal{A})/\mathrm{p}_{0,n}(\mathcal{A}) \to 1$ uniformly over $\mathcal{A}$, as $n \to \infty$. Therefore, along with Equation A-8, this implies that

$$|\mathrm{p}_{\epsilon(n)}(T = t \mid y) - \mathrm{p}_{\epsilon=0}(T = t \mid y)| \leq \max_{\mathcal{A}' \in \mathcal{H}_t(n)} \max_{\mathcal{A} \in \cup_{l \neq t} \mathcal{H}_l(n)} \left| 1 - \frac{\mathrm{p}_{0,n}(\mathcal{A}')\mathrm{p}_{\epsilon(n),n}(\mathcal{A})}{\mathrm{p}_{\epsilon(n),n}(\mathcal{A}')\mathrm{p}_{0,n}(\mathcal{A})} \right| \longrightarrow 0$$

as $n \to \infty$. □

Proof of Lemma 2-2.

To study the posterior distribution of the number of clusters $p(T = t \mid y_{1:n})$, we will focus on

$$p(y_{1:n}, T = t) = \sum_{\mathcal{A} \in \mathcal{H}_t(n)} p(y_{1:n} \mid \mathcal{A}) p(\mathcal{A}) \qquad \text{(A-12)}$$

where $\mathcal{A} = \{A_1, \ldots, A_t\}$, $p(y_{1:n} \mid \mathcal{A}) = \prod_{A \in \mathcal{A}} m(y_A)$, $y_A = (y_i : i \in A)$, and $m(y_A) = \int_\Theta \big(\prod_{i \in A} f_\theta(y_i)\big)\, d\mathcal{G}(\theta)$, as described at Equation 2-5. In this lemma,

$f_\theta(y_i) = \mathrm{N}(y_i \mid \theta, 1)$ and

$$m(y_A) = \frac{1}{\sqrt{|A|+1}} f_0(y_A) \exp\left(\frac{(\sum_{i\in A} y_i)^2}{2(|A|+1)}\right), \tag{A-13}$$

where $f_0(y_A) = \prod_{j\in A} \mathrm{N}(y_j \mid 0, 1)$. Under the Dirichlet process prior, the probability mass function on partitions is $p(\mathcal{A}) = \frac{\alpha^t}{\alpha^{(n)}} \prod_{i=1}^t (|A_i|-1)!$ where $\alpha^{(n)} = \alpha(\alpha+1)\cdots(\alpha+n-1)$. Hence,

$$\begin{aligned}
\frac{p(y_{1:n}, T=2)}{p(y_{1:n}, T=1)} &\overset{(a)}{=} \frac{\sum_{\mathcal{A}=\{A_1,A_2\}\in\mathcal{H}_2(n)} p(\mathcal{A}) m(y_{A_1}) m(y_{A_2})}{\mathbb{P}(\mathcal{A} = \{\{1,\ldots,n\}\}) m(y_{1:n})} \\
&\overset{(b)}{=} \sum_{\mathcal{A}=\{A_1,A_2\}\in\mathcal{H}_2(n)} \left[\frac{\alpha^2 (|A_1|-1)!(|A_2|-1)!}{\alpha(n-1)!} \times \frac{\sqrt{n+1}}{\sqrt{|A_1|+1}\sqrt{|A_2|+1}} \right. \\
&\qquad \left. \times \exp\left(\frac{(\sum_{i\in A_1} y_i)^2}{2(|A_1|+1)}\right) \exp\left(\frac{(\sum_{i\in A_2} y_i)^2}{2(|A_2|+1)}\right) \exp\left(-\frac{(\sum_{i=1}^n y_i)^2}{2(n+1)}\right)\right] \\
&\overset{(c)}{\leq} \frac{\alpha\sqrt{n+1}}{(n-1)!} \sum_{\mathcal{A}\in\mathcal{H}_2(n)} \frac{(|A_1|-1)!(|A_2|-1)!}{\sqrt{|A_1|+1}\sqrt{|A_2|+1}} \exp\left(\frac{\sum_{i\in A_1} y_i^2}{2} + \frac{\sum_{i\in A_2} y_i^2}{2}\right) \\
&\overset{(d)}{\leq} \frac{\alpha\sqrt{n+1}}{2(n-1)!} \exp\left(\frac{\sum_{i=1}^n y_i^2}{2}\right) \sum_{\mathcal{A}\in\mathcal{H}_2(n)} (|A_1|-1)!(|A_2|-1)! \\
&= \frac{\alpha\sqrt{n+1}}{2(n-1)!} \exp\left(\frac{\sum_{i=1}^n y_i^2}{2}\right) \sum_{i=1}^{n-1} \frac{1}{2}\binom{n}{i} (i-1)!(n-i-1)! \\
&= \frac{\alpha\sqrt{n+1}}{2} \exp\left(\frac{\sum_{i=1}^n y_i^2}{2}\right) \sum_{i=1}^{n-1} \frac{n}{2i(n-i)} \\
&= \frac{\alpha\sqrt{n+1}}{2} \exp\left(\frac{\sum_{i=1}^n y_i^2}{2}\right) \left(1 + \frac{1}{2} + \cdots + \frac{1}{n-1}\right) \\
&\leq \frac{\alpha\sqrt{n+1}}{2} \exp\left(\frac{\sum_{i=1}^n y_i^2}{2}\right) \left(\log(n-1) + 1\right)
\end{aligned}$$

where (a) is using Equation A-12 for both numerator and denominator, (b) is using Equation A-13 and the probability mass function of $\mathcal{A}$, (c) follows from $(\sum_{j=1}^n y_j)^2 \geq 0$ and Jensen's inequality, (d) is using $A_1 \cup A_2 = \{1,\ldots,n\}$ and both $|A_1|$ and $|A_2|$ are greater than or equal to 1, and the last inequality is induced from $1/k \leq \log(k) - \log(k-1)$ for $k = 2,\ldots,n-1$.

By the strong law of large numbers,

$$\frac{1}{n}\sum_{i=1}^{n} y_i^2 \xrightarrow[n\to\infty]{\text{a.s.}} E(y_1^2) = 1 + \kappa^2/2.$$

Let $\delta = C - (1/2 + \kappa^2/4)$, where by the assumption of the theorem, $C > 1/2 + \kappa^2/4$ such that $\alpha = \alpha(n) = o(\exp(-C\,n))$. Then, almost surely, for all n sufficiently large, $\frac{1}{2n}\sum_{i=1}^{n} y_i^2 \le 1/2 + \kappa^2/4 + \delta/2 = C - \delta/2$. Hence, almost surely,

$$\begin{aligned}\limsup_{n\to\infty} \frac{p(y_{1:n}, T=2)}{p(y_{1:n}, T=1)} &\le \frac{\alpha(n)\sqrt{n+1}}{2} \exp\big(n(C-\delta/2)\big)\big(\log(n-1)+1\big) \\ &= \big(\alpha(n)\exp(C\,n)\big)\exp(-n\delta/2)\frac{\sqrt{n+1}}{2}\big(\log(n-1)+1\big) \longrightarrow 0\end{aligned}$$

as $n \to \infty$. Therefore, we have the conclusion,

$$p(T=2 \mid y_{1:n}) = \frac{p(y_{1:n}, T=2)}{\sum_{t=1}^{\infty} p(y_{1:n}, T=t)} \le \frac{p(y_{1:n}, T=2)}{p(y_{1:n}, T=1)} \xrightarrow{\text{a.s.}} 0.$$

□

A.2 Additional Simulation Results

In this section, we first provide the simulation results when the component distribution is the bivariate Gaussian distribution. Then we show a comparison between the quasi-Bernoulli mixture and the Dirichlet process mixture with $\alpha(n) \to 0$. Finally, we provide information on convergence diagnostics and the running time of the algorithm.

A.2.1 Simulation Results with Bivariate Gaussian Mixtures

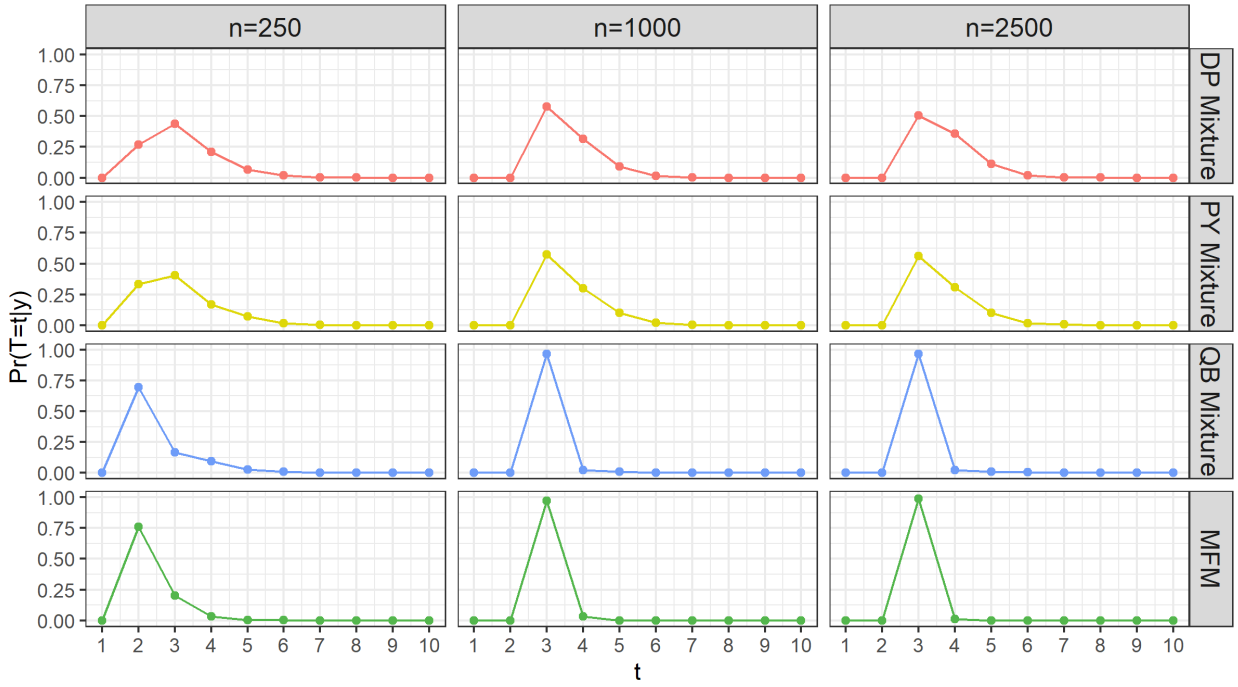
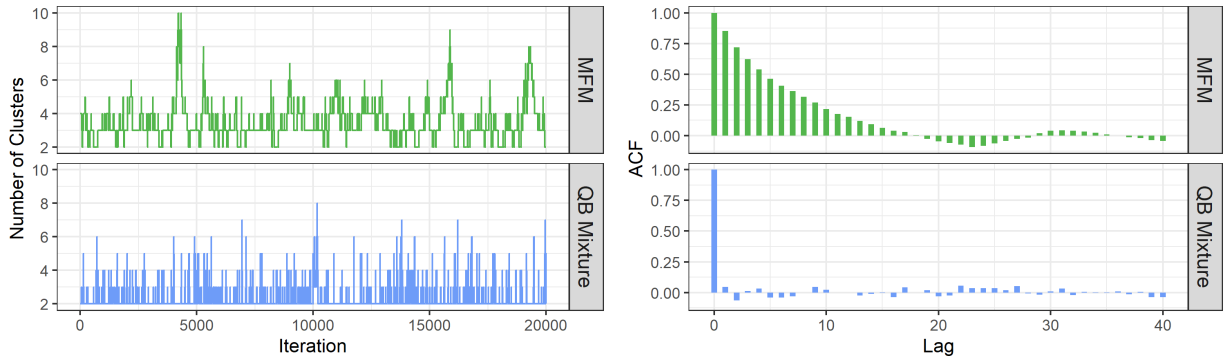


Figure A-1. Posterior distribution on the number of clusters for data generated from a three-component Gaussian mixture in $\mathbb{R}^2$. Using the quasi-Bernoulli mixture model, the posterior probabilities of T concentrate to a point mass at $k_0 = 3$ for large n. However, the posterior distributions of the calibrated Dirichlet process mixture model and Pitman–Yor process mixture model do not concentrate to a point mass at $k_0 = 3$.

Figure A-1 plots the posterior distribution of the number of clusters T at each n. Under the quasi-Bernoulli mixture model (shown in blue), the posterior of T concentrates to a point mass at the true number of components ($k_0 = 3$) as n grows, in accordance with our theory. On the other hand, the posterior distributions of T with the Dirichlet process mixture model and the Pitman–Yor process mixture model fail to concentrate to a point mass at the true number of components.

As shown in Figure A-2, the MFM model suffers from slow mixing with high auto-correlation even after thinning (effective sample size 15.3% on average of five experiments with sample size 250); whereas the quasi-Bernoulli mixture model quickly shows a much faster drop in the auto-correlation within a few lags (effective sample size 57.2%).



(a) Trace on number of clusters

(b) Auto-correlation (based on thinning at 25)

Figure A-2. The trace of the Markov chain on T and auto-correlation functions for bivariate Gaussian mixture data with sample size 250. Quasi-Bernoulli mixture model shows much better mixing in the Markov chain, compared to the MFM model. We discard the first 5,000 iterations as burn-ins and record the following 20,000 samples.

A.2.2 Simulations on the Dirichlet Process Mixture with $\alpha(n) \to 0$

To compare the quasi-Bernoulli mixture model with the Dirichlet process mixture model under different rates of $\alpha(n) \to 0$, we conduct additional experiments with univariate Gaussian mixtures. The experimental settings are similar to the ones in Section 2.5.1, except that we generate data from mixtures with smaller distances between the component centers: $0.3\,\mathrm{N}(-2, 1^2) + 0.4\,\mathrm{N}(0, 1^2) + 0.3\,\mathrm{N}(2, 1^2)$ under sample sizes $n \in \{100, 250, 1000, 2500\}$. The purpose is to examine if each model shows the trend of converging to the ground truth $T = k_0$ as n increases, when the component distribution $\mathcal{F}$ has an unbounded support and clusters have large overlaps.

For Dirichlet process mixture models, we use three rates $\alpha_1(n) = \exp(-n/10)$, $\alpha_2(n) = 4/\log(n)$ and $\alpha_3(n) = 20/n$. For quasi-Bernoulli mixture models, we use two rates $\epsilon_1(n) = n^{-2.1}$ and $\epsilon_2(n) = n^{-3.1}$. Figure A-3 shows the posterior distributions of the number of clusters.

This empirical result suggests that one may be able to obtain posterior consistency on estimating T for the Dirichlet process mixture model with general $\mathcal{F}$, by choosing an $\alpha \to 0$ faster than $4/\log(n)$ but slower than $20/n$, although the theory remains an open question. On the other hand, the quasi-Bernoulli mixture models show almost no difference in the trend of convergence.

This is as expected, since both $n^{-2.1}$ and $n^{-3.1}$ satisfy the rate condition that guarantees consistency on estimating T.

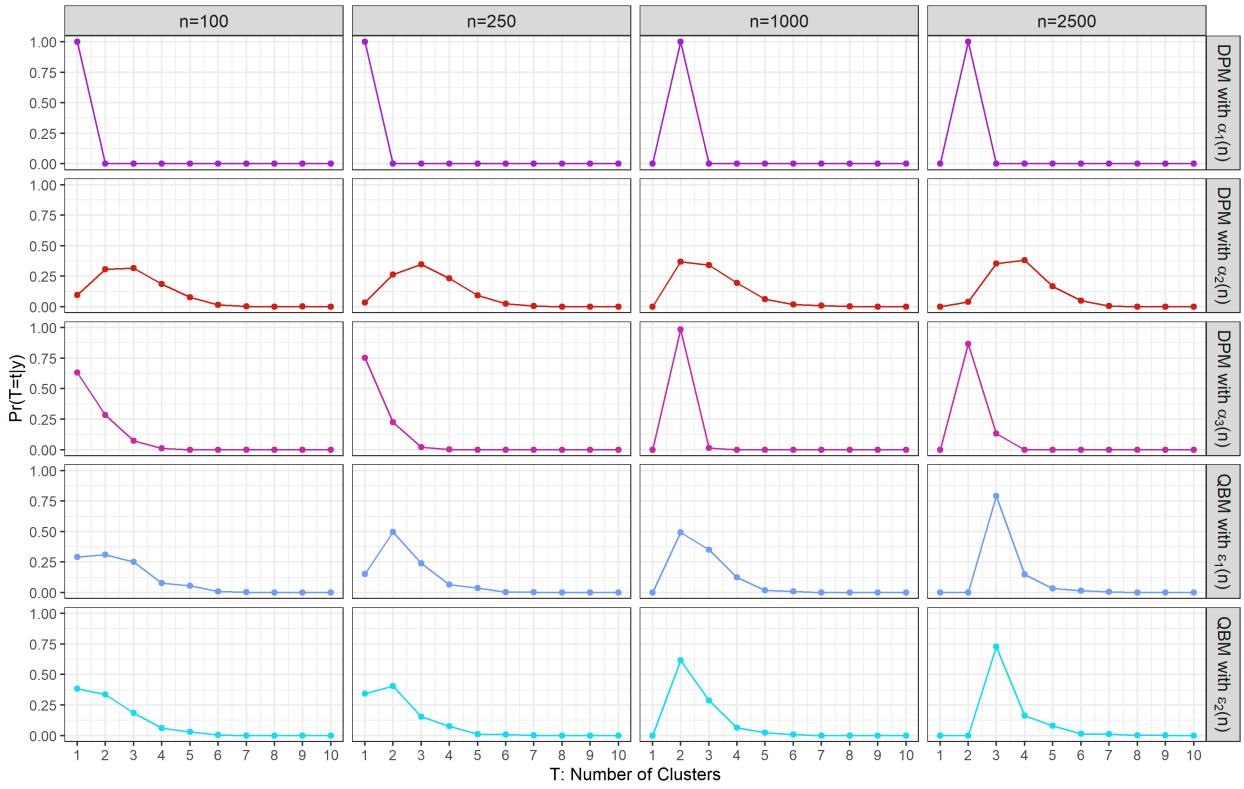


Figure A-3. Posterior distribution on the number of clusters for data generated from a three-component univariate Gaussian mixture.

A.2.3 Convergence Diagnostics and Timing Information

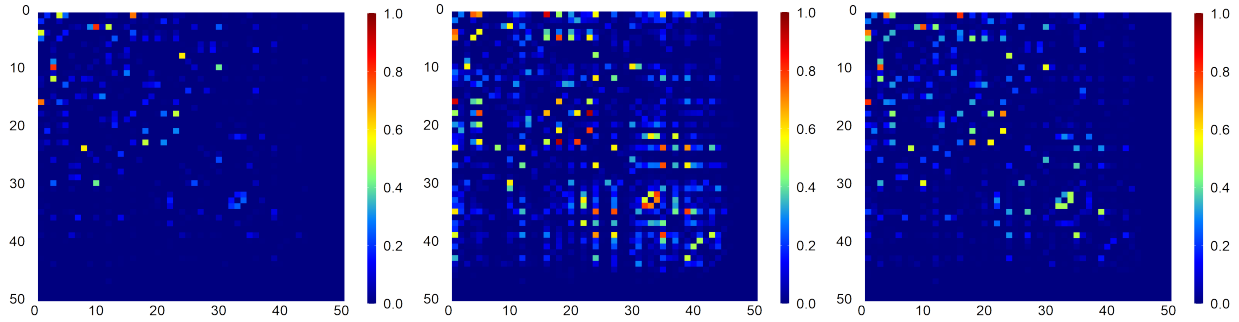
We use the Markov chain sample of T (the number of clusters) for convergence diagnosis. We choose T because it is often the slowest-changing variable. As shown in Figure 3-2, the auto-correlations for T show a quick drop to an insignificant level, within as few lags after thinning at 50. For each experiment, we run multiple chains from 5 randomly initialized points, and compute the $\hat{R}$ statistic (Gelman and Rubin, 1992). All of the experiments get $\hat{R}$ close to 1, which means that the Markov chains have converged.

We provide the timing information of our posterior sampling algorithms. The algorithms are implemented in R, and run on a 4.0 GHz processor. For the model with univariate Gaussian components, each iteration costs 0.0005, 0.0006, 0.0010, 0.0035, 0.0078 seconds for the sample

size 50, 100, 250, 1000, 2500, respectively. For the bivariate Gaussian case, the algorithm runs 0.0029, 0.0236, 0.0945 seconds for each iteration for sample size 250, 1000, 2500, respectively. When the component distribution is the Laplace distribution, the algorithm takes 0.0055, 0.1018, 0.2269, 0.3083 seconds for each iteration for sample size 50, 200, 500, 1000, respectively. For the network model used in the data application, each iteration takes around 0.3 seconds.

A.3 Other Useful Results

Some useful results for understanding the Chapter 2 are provided.



(a) Group 1 (36.6% of the subjects) (b) Group 2 (20.6% of the subjects) (c) Group 3 (42.8% of the subjects)

Figure A-4. The MAP estimation of the mean of each Gaussian component under the mixture of factor analyzers model.

Table A-1. Settings of the hyper-parameters for the Dirichlet processes and the Pitman–Yor processes when sample sizes $n \in \{50, 200, 500, 1000, 2500\}$. We match the expectations of the number of clusters T under the three priors, and also make the variances of T close under the Pitman-Yor process prior and the quasi-Bernoulli process prior (with $\tilde{p} = 0.9$). Each expectation and variance are approximated based on 2×10^5 samples from the prior.

	Dirichlet process		Pitman–Yor process				quasi-Bernoulli process	
	α	$\mathbb{E}(T)$	α	d	$\mathbb{E}(T)$	$\mathrm{Var}(T)$	$\mathbb{E}(T)$	$\mathrm{Var}(T)$
50	0.71	3.62	0.48	0.09	3.62	3.07	3.64	3.09
100	0.69	4.04	0.46	0.09	4.10	3.95	4.06	3.97
250	0.67	4.58	0.38	0.10	4.56	5.39	4.59	5.40
1000	0.63	5.25	0.29	0.11	5.25	8.25	5.28	7.92
2500	0.61	5.69	0.29	0.10	5.74	9.51	5.69	9.79

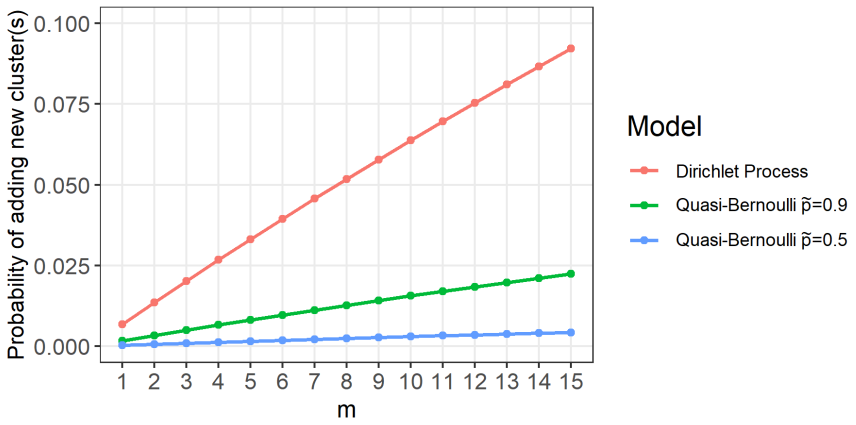


Figure A-5. The probability of adding one or more new clusters for m future data points ($n = 100$). All of the parameters are chosen to be the same as Figure 2-1 except for $\epsilon = \epsilon(n, m) = 1/(n+m)^5$. Under the prior, the Dirichlet process exhibits rapid growth in this probability, favoring the creation of additional clusters *a priori*. Meanwhile, the quasi-Bernoulli process exhibits much slower growth of this probability.

APPENDIX B
APPENDIX FOR CHAPTER 3

B.1 Proofs

The proof of Theorem 3-2 uses a theorem from Miller (2021). We provide the complete statement of that theorem as following.

Theorem B-1. *Miller (2021, Theorem 5) Let $\Theta \subseteq \mathbb{R}^D$. Let $E \subseteq \Theta$ be open (in $\mathbb{R}^D$) and bounded. Fix $\theta_0 \in E$ and let $\pi : \Theta \to \mathbb{R}$ be a probability density with respect to Lebesgue measure such that π is continuous at θ_0 and $\pi(\theta_0) > 0$. Let $f_n : \Theta \to \mathbb{R}$ have continuous third derivatives on E. Suppose $f_n \to f$ pointwise for some $f : \Theta \to \mathbb{R}$, $f''(\theta_0)$ is positive definite, and (f_n''') is uniformly bounded on E. If either of the following two assumptions is satisfied:*

1. *$f(\theta) > f(\theta_0)$ for all $\theta \in K \backslash \{\theta_0\}$ and $\liminf_n \inf_{\theta \in \Theta \backslash K} f_n(\theta) > f(\theta_0)$ for some compact $K \subseteq E$ with θ_0 in the interior of K, or*

2. *each f_n is convex and $f'(\theta_0) = 0$,*

then there is a sequence $\theta_n \to \theta_0$ such that $f_n'(\theta_n) = 0$ for all n sufficiently large, $f_n(\theta_n) \to f(\theta_0)$, defining $m_n = \int_{\mathbb{R}^D} \exp(-n f_n(\theta))\pi(\theta)\, \mathrm{d}\theta$ and $\pi_n(\theta) = \exp(-n f_n(\theta))\pi(\theta)/m_n$, we have $\int_{B_\varepsilon(\theta_0)} \pi_n(\theta)\, \mathrm{d}\theta \xrightarrow[n\to\infty]{} 1$ for all $\varepsilon > 0$, that is, π_n concentrates at θ_0, and letting q_n be the density of $\sqrt{n}(\theta - \theta_n)$ when $\theta \sim \pi_n$, we have $\int_{\mathbb{R}^D} \left| q_n(x) - \mathcal{N}\left(x \mid 0, H_0^{-1}\right) \right| \mathrm{d}x \xrightarrow[n\to\infty]{} 0$, that is, q_n converges to $\mathcal{N}(0, H_0^{-1})$ in total variation, where $H_0 = f''(\theta_0)$. Further, 2 $\Rightarrow$ 1 under the assumptions of the theorem.

Proof of Theorem 3-2. We show the following properties in the sense of $a.s.[y_{1:n}]$.

1. $\hat{l}_n$ has continuous third derivatives on E. Since l_n has continuous third derivatives on $E \times \hat{z}_n(E)$ and $\hat{z}_n$ has continuous third derivative, we have

$$\begin{aligned}
\hat{l}_n'(\lambda) &= \left.\frac{\partial l_n(\lambda,\zeta)}{\partial \lambda}\right|_{\zeta=\hat{z}_n(\lambda)} + \left.\frac{\partial l_n(\lambda,\zeta)}{\partial \zeta}\right|_{\zeta=\hat{z}_n(\lambda)} \hat{z}_n'(\lambda), \\
\hat{l}_n''(\lambda) &= \left.\frac{\partial^2 l_n(\lambda,\zeta)}{\partial \lambda^2}\right|_{\zeta=\hat{z}_n(\lambda)} + 2\left.\frac{\partial^2 l_n(\lambda,\zeta)}{\partial \lambda \partial \zeta}\right|_{\zeta=\hat{z}_n(\lambda)} \hat{z}_n'(\lambda) \\
&\quad + [\hat{z}_n'(\lambda)]^\top \left.\frac{\partial^2 l_n(\lambda,\zeta)}{\partial \zeta^2}\right|_{\zeta=\hat{z}_n(\lambda)} \hat{z}_n'(\lambda) + \left.\frac{\partial l_n(\lambda,\zeta)}{\partial \zeta}\right|_{\zeta=\hat{z}_n(\lambda)} \hat{z}_n''(\lambda).
\end{aligned}$$

   Then it is not hard to see that the property is satisfied.

2. $\hat{l}_n \to \hat{l}_*$ pointwise on Θ and $\hat{l}_*''(\lambda_0)$ is positive definite. The first part is obvious as $l_n \to l_*$ and $\hat{z}_n \to \hat{z}_*$. To show $\hat{l}_*''(\lambda_0)$ is positive definite, we have

$$\begin{aligned}
\hat{l}_*''(\lambda_0) &= \left.\frac{\partial^2 l_*(\lambda_0,\zeta)}{\partial \lambda^2}\right|_{\zeta=\hat{z}_*(\lambda_0)} + 2\left.\frac{\partial^2 l_*(\lambda_0,\zeta)}{\partial \lambda \partial \zeta}\right|_{\zeta=\hat{z}_*(\lambda_0)} \hat{z}_*'(\lambda_0) \\
&\quad + [\hat{z}_*'(\lambda_0)]^\top \left.\frac{\partial^2 l_*(\lambda_0,\zeta)}{\partial \zeta^2}\right|_{\zeta=\hat{z}_*(\lambda_0)} \hat{z}_*'(\lambda_0) + \left.\frac{\partial l_*(\lambda_0,\zeta)}{\partial \zeta}\right|_{\zeta=\hat{z}_*(\lambda_0)} \hat{z}_*''(\lambda_0) \\
&= \begin{bmatrix} I_d & [\hat{z}_*'(\lambda_0)]^\top \end{bmatrix} l_*''(\lambda_0, \hat{z}_*(\lambda_0)) \begin{bmatrix} I_d \\ \hat{z}_*'(\lambda_0) \end{bmatrix}
\end{aligned}$$

   where the second equation is using $\left.\frac{\partial l_*(\lambda_0,\zeta)}{\partial \zeta}\right|_{\zeta=\hat{z}_*(\lambda_0)} = 0$ to cancel out the last term.

3. $\hat{l}_n'''$ is uniformly bounded on E. Since l_n''' and $\hat{z}_n'''$ are uniformly bounded, by the theorem 7 of Miller (2021), l_n', l_n'', $\hat{z}_n'$ and $\hat{z}_n''$ are all uniformly bounded. Hence, $\hat{l}_n'''$ is uniformly bounded

by the expansion of $\hat{l}_n'''$:

$$\begin{aligned}\hat{l}_n'''(\lambda) = &\left.\frac{\partial^3 l_n(\lambda,\zeta)}{\partial\lambda^3}\right|_{\zeta=\hat{z}_n(\lambda)} + 3\left.\frac{\partial^3 l_n(\lambda,\zeta)}{\partial\lambda^2\partial\zeta}\right|_{\zeta=\hat{z}_n(\lambda)} \hat{z}_n'(\lambda) \\ &+ 3[\hat{z}_n'(\lambda)]^\top \left.\frac{\partial^3 l_n(\lambda,\zeta)}{\partial\lambda\partial\zeta^2}\right|_{\zeta=\hat{z}_n(\lambda)} \hat{z}_n'(\lambda) + [\hat{z}_n'(\lambda)]^\top \left.\frac{\partial^3 l_n(\lambda,\zeta)}{\partial\zeta^3}\right|_{\zeta=\hat{z}_n(\lambda)} [\hat{z}_n'(\lambda)]^2 \\ &+ 3[\hat{z}_n'(\lambda)]^\top \left.\frac{\partial^2 l_n(\lambda,\zeta)}{\partial\zeta^2}\right|_{\zeta=\hat{z}_n(\lambda)} \hat{z}_n''(\lambda) + 3\left.\frac{\partial^2 l_n(\lambda,\zeta)}{\partial\lambda\partial\zeta}\right|_{\zeta=\hat{z}_n(\lambda)} \hat{z}_n''(\lambda) \\ &+ \left.\frac{\partial l_n(\lambda,\zeta)}{\partial\zeta}\right|_{\zeta=\hat{z}_n(\lambda)} \hat{z}_n'''(\lambda).\end{aligned}$$

4. for some compact $K \subseteq E$ with λ_0 in the interior of K, $\hat{l}_*(\lambda) > \hat{l}_*(\lambda_0)$ for all $\lambda \in K\backslash\{\lambda_0\}$ and $\liminf_n \inf_{\lambda\in\Theta\backslash K} \hat{l}_n(\lambda) > \hat{l}_*(\lambda_0)$.

By Theorem B-1 with $f_n = -\hat{l}_n$ and $f = -\hat{l}_*$, 1–4 complete the proof. □

Proof of Corollary 1. We use the delta method to find the asymptotic distribution of $\sqrt{n}$-adjusted $\hat{z}_n(\lambda)$. Since convergence in total variation implies convergence in distribution (weak convergence), the random vector $\sqrt{n}(\lambda - \lambda_n) \rightsquigarrow \mathrm{N}(0, H_0^{-1})$. Using delta method, we can prove $\sqrt{n}\{\hat{z}_n(\lambda) - \hat{z}_n(\lambda_n)\} \rightsquigarrow \mathrm{N}(0, H_z^{-1})$ where $H_z^{-1} = \hat{z}_*'(\lambda_0) H_0^{-1} \hat{z}_*'(\lambda_0)^\top$.

When $g_n(\zeta, y_{1:n}; \lambda) = -L(y_{1:n}, \zeta; \lambda)$ with that the bridged posterior coincides with the profile likelihood, the asymptotic variance above can be represented by the second derivatives of l_*. We first show that $H_0 = \hat{l}_*''(\lambda_0) = l_{*0,\lambda\lambda} - l_{*0,\lambda\zeta} l_{*0,\zeta\zeta}^{-1} l_{*0,\zeta\lambda}$, where $l_{*0,\lambda\lambda}, l_{*0,\zeta\zeta}, l_{*0,\zeta\lambda}, l_{*0,\lambda\zeta}$ are respectively the second partial derivatives $\frac{\partial^2 l_*(\lambda,\zeta)}{\partial\lambda^2}, \frac{\partial^2 l_*(\lambda,\zeta)}{\partial\zeta^2}, \frac{\partial^2 l_*(\lambda,\zeta)}{\partial\zeta\partial\lambda}, \frac{\partial^2 l_*(\lambda,\zeta)}{\partial\lambda\partial\zeta}$ evaluating at $\lambda = \lambda_0, \zeta = \hat{z}_*(\lambda_0)$.

Since in this case $\frac{\partial l_n(\lambda,\zeta)}{\partial\zeta}|_{\zeta=\hat{z}_n(\lambda)} = 0$, we have

$$0 = \frac{\partial}{\partial\lambda}\left\{\left.\frac{\partial l_n(\lambda,\zeta)}{\partial\zeta}\right|_{\zeta=\hat{z}_n(\lambda)}\right\} = \left.\frac{\partial^2 l_n(\lambda,\zeta)}{\partial\zeta\partial\lambda}\right|_{\zeta=\hat{z}_n(\lambda)} + \left.\frac{\partial^2 l_n(\lambda,\zeta)}{\partial\zeta^2}\right|_{\zeta=\hat{z}_n(\lambda)} \hat{z}_n'(\lambda).$$

Hence

$$\hat{z}_n'(\lambda) = -\left\{\left.\frac{\partial^2 l_n(\lambda,\zeta)}{\partial\zeta^2}\right|_{\zeta=\hat{z}_n(\lambda)}\right\}^{-1} \left.\frac{\partial^2 l_n(\lambda,\zeta)}{\partial\zeta\partial\lambda}\right|_{\zeta=\hat{z}_n(\lambda)}.$$

By Miller (2021, Theorem 7), we have $l_n'' \to l_*''$. Letting $n \to \infty$ and $\lambda = \lambda_0$, we have $\hat{z}_*'(\lambda_0) = -l_{*0,\zeta\zeta}^{-1} l_{*0,\zeta\lambda}$. Now

$$\hat{l}_n''(\lambda) = \frac{\partial}{\partial \lambda} \left\{ \frac{\partial l_n(\lambda, \zeta)}{\partial \lambda}\bigg|_{\zeta=\hat{z}_n(\lambda)} \right\} = \frac{\partial^2 l_n(\lambda, \zeta)}{\partial \lambda \partial \zeta}\bigg|_{\zeta=\hat{z}_n(\lambda)} \hat{z}_n'(\lambda) + \frac{\partial^2 l_n(\lambda, \zeta)}{\partial \lambda^2}\bigg|_{\zeta=\hat{z}_n(\lambda)}.$$

Letting $n \to \infty$ and $\lambda = \lambda_0$, we have the result $\hat{l}_*''(\lambda_0) = l_{*0,\lambda\lambda} - l_{*0,\lambda\zeta} l_{*0,\zeta\zeta}^{-1} l_{*0,\zeta\lambda}$.

We assume that both $l_{*0,\lambda\lambda}$ and $l_{*0,\zeta\zeta}$ are positive definite. The asymptotic variance of $\sqrt{n}\{\hat{z}_n(\lambda) - \hat{z}_n(\lambda_n)\}$ is thus $\hat{z}_*'(\lambda_0) H_0^{-1} \hat{z}_*'(\lambda_0)^\top = l_{*0,\zeta\zeta}^{-1} l_{*0,\zeta\lambda} (l_{*0,\lambda\lambda} - l_{*0,\lambda\zeta} l_{*0,\zeta\zeta}^{-1} l_{*0,\zeta\lambda})^{-1} l_{*0,\lambda\zeta} l_{*0,\zeta\zeta}^{-1}$.

If we treat the latent variable ζ as non-deterministic in the likelihood $L(y, \zeta; \lambda)$ with some prior, then Miller (2021) proves $\sqrt{n}\big([\lambda\ \zeta]^\top - [\lambda_n\ \zeta_n]^\top\big) \rightsquigarrow N(0, \tilde{H}_0^{-1})$ for some sequences λ_n and ζ_n when $(\lambda, \zeta) \sim \Pi(\lambda, \zeta \mid y)$, where $\tilde{H}_0 = l_*''(\lambda_0, \zeta_0)$ and ζ_0 is a fixed ground-truth of ζ. Marginally, the asymptotic variance of $\sqrt{n}(\zeta - \zeta_n)$ is the ζ-block of $\tilde{H}_0^{-1}$, which is equal to

$$(l_{*0,\zeta\zeta} - l_{*0,\zeta\lambda} l_{*0,\lambda\lambda}^{-1} l_{*0,\lambda\zeta})^{-1} = l_{*0,\zeta\zeta}^{-1} + l_{*0,\zeta\zeta}^{-1} l_{*0,\zeta\lambda} (l_{*0,\lambda\lambda} - l_{*0,\lambda\zeta} l_{*0,\zeta\zeta}^{-1} l_{*0,\zeta\lambda})^{-1} l_{*0,\lambda\zeta} l_{*0,\zeta\zeta}^{-1}.$$

Hence, the asymptotic variance of the j-th element of $\sqrt{n}(\zeta - \zeta_n)$ is strictly greater than the one of the j-th element of $\sqrt{n}\{\hat{z}_n(\lambda) - \hat{z}_n(\lambda_n)\}$. □

Proof of Lemma 3-1. By the definition of the profile likelihood, we have $\hat{l}_n(\lambda) = l_n(\lambda, \hat{z}_\lambda) \geq l_n\{\lambda, \tilde{\zeta}_\lambda(\lambda_0, \hat{z}_{\lambda_0})\}$, so

$$\begin{aligned}
\hat{l}_n(\lambda) - \hat{l}_n(\lambda_0) &= l_n(\lambda, \hat{z}_\lambda) - l_n(\lambda_0, \hat{z}_{\lambda_0}) \\
&\geq l_n\{\lambda, \tilde{\zeta}_\lambda(\lambda_0, \hat{z}_{\lambda_0})\} - l_n\{\lambda_0, \tilde{\zeta}_{\lambda_0}(\lambda_0, \hat{z}_{\lambda_0})\} \\
&= \tilde{l}_n(\lambda, \lambda_0, \hat{z}_{\lambda_0}) - \tilde{l}_n(\lambda_0, \lambda_0, \hat{z}_{\lambda_0}) \qquad \text{(B-1)}
\end{aligned}$$

by using $\tilde{\zeta}_\lambda(\lambda,\zeta)=\zeta$ for the second term. Similarly, $\hat{l}_n(\lambda_0)=l_n(\lambda_0,\hat{z}_{\lambda_0})\geq l_n\{\lambda_0,\tilde{\zeta}_{\lambda_0}(\lambda,\hat{z}_\lambda)\}$, and hence

$$\begin{aligned}\hat{l}_n(\lambda)-\hat{l}_n(\lambda_0)&=l_n(\lambda,\hat{z}_\lambda)-l_n(\lambda_0,\hat{z}_{\lambda_0})\\&\leq l_n\{(\lambda,\tilde{\zeta}_\lambda(\lambda,\hat{z}_\lambda)\}-l_n\{\lambda_0,\tilde{\zeta}_{\lambda_0}(\lambda,\hat{z}_\lambda)\}\\&=\tilde{l}_n(\lambda,\lambda,\hat{z}_\lambda)-\tilde{l}_n(\lambda_0,\lambda,\hat{z}_\lambda)\end{aligned}\tag{B-2}$$

by using $\tilde{\zeta}_\lambda(\lambda,\zeta)=\zeta$ for the first term. Both (B-1) and (B-2) are differences between function l_n evaluating at λ and the one at λ_0 while keeping the other arguments unchanged. They have the Taylor expansion of the function $\tilde{l}_n$ with respect to the first argument,

$$(\lambda-\lambda_0)^\top\frac{\partial\tilde{l}_n(t,\psi,\hat{z}_\psi)}{\partial t}\bigg|_{t=\lambda_0}+\frac{1}{2}(\lambda-\lambda_0)^\top\frac{\partial^2\tilde{l}_n(t,\psi,\hat{z}_\psi)}{\partial t^2}\bigg|_{t=\tilde{t}}(\lambda-\lambda_0)\tag{B-3}$$

where $\tilde{t}$ is somewhere between λ and λ_0, and ψ can be λ or λ_0. By the assumption 1 in (B1), the second term is equal to $-(\lambda-\lambda_0)^\top H_0(\lambda-\lambda_0)/2+o_{P_{\lambda_0,\zeta_0}}(1)(\|\lambda-\lambda_0\|^2)$. By the assumption 2 in (B1), the first term is equal to

$$(\lambda-\lambda_0)^\top h_n+(\lambda-\lambda_0)^\top\mathbb{E}_{\lambda_0,\zeta_0}\frac{\partial\tilde{l}_n(t,\lambda,\hat{z}_\lambda)}{\partial t}\bigg|_{t=\lambda_0}+o_{P_{\lambda_0,\zeta_0}}(1)\big(\|\lambda-\lambda_0\|n^{-1/2}\big).$$

Combining with (3-9) and $\|\lambda-\lambda_0\|n^{-1/2}\leq\big(\|\lambda-\lambda_0\|+n^{-1/2}\big)^2$, the first term of (B-3) becomes $(\lambda-\lambda_0)^\top h_n+o_{P_{\lambda_0,\zeta_0}}(1)\big\{\big(\|\lambda-\lambda_0\|+n^{-1/2}\big)^2\big\}$, and hence (3-10) is proved. □

Proof of Theorem 3-3. We have $\pi_n(\lambda)=\exp\{n\hat{l}_n(\lambda)\}\pi_0(\lambda)/m_n$

where $m_n=\int_{\mathbb{R}^d}\exp\{n\hat{l}_n(\lambda)\}\pi_0(\lambda)\,\mathrm{d}\lambda$, and $q_n(x)=\pi_n(\lambda_n+x/\sqrt{n})n^{-d/2}$. Let

$$g_n(x)=q_n(x)\exp\{-n\hat{l}_n(\lambda_n)\}n^{d/2}m_n=\exp[n\{\hat{l}_n(\lambda_n+x/\sqrt{n})-\hat{l}_n(\lambda_n)\}]\pi_0(\lambda_n+x/\sqrt{n})$$

and define $g_0(x)=\exp\{-x^\top H_0x/2\}\pi_0(\lambda_0)$. We first show that $\int_{\mathbb{R}^d}|g_n(x)-g_0(x)|\,\mathrm{d}x\xrightarrow[n\to\infty]{P_{\lambda_0,\zeta_0}}0$. Denote the ϵ chosen from Lemma 3-1 as ϵ_0. Since π_0 is continuous at λ_0, we choose sufficiently

small $\epsilon \in (0, \epsilon_0/2)$ such that $\pi_0(\lambda) \le 2\pi_0(\lambda_0)$ for all $\lambda \in B_{2\epsilon}(\lambda_0)$. Let δ be the number from (3-11).

Murphy and Van der Vaart (2000, Corollary 1) shows that $\sqrt{n}\|\lambda_n - \lambda_0\|$ is bounded in probability and for all $\lambda \in B_{\epsilon_0}(\lambda_0)$ and large enough n such that $\lambda_n \in B_{\epsilon_0}(\lambda_0)$,

$$\hat{l}_n(\lambda) - \hat{l}_n(\lambda_n) = -\frac{1}{2}(\lambda - \lambda_n)^\top H_0(\lambda - \lambda_n) + o_{P_{\lambda_0,\zeta_0}}(1)\left\{\left(\|\lambda - \lambda_n\| + n^{-1/2}\right)^2\right\}.$$

Letting $\lambda = \lambda_n + x/\sqrt{n}$ with $x \in B_{\epsilon\sqrt{n}}(0)$ and large enough n such that $\lambda_n \in B_{\epsilon_0/2}(0)$, we have

$$n\left\{\hat{l}_n(\lambda_n + x/\sqrt{n}) - \hat{l}_n(\lambda_n)\right\} = -\frac{1}{2}x^\top H_0 x + o_{P_{\lambda_0,\zeta_0}}(1)(\|x\| + 1)^2.$$

Combining with π_0 is continuous at λ_0 and $\lambda_n + x/\sqrt{n} \to \lambda_0$, we have $g_n(x) \to g_0(x)$ pointwise with probability converge to 1. Consider $n > 1/\epsilon^2$ sufficiently large such that the term $o_{P_{\lambda_0,\zeta_0}}(1) < \alpha/4$ where α is less than the smallest eigenvalue of H_0. We denote $A_0 = H_0 - \alpha I$, and define

$$h_n(x) = \begin{cases} \exp(-x^\top A_0 x/2 + \alpha/2)2\pi_0(\lambda_0) & \text{if } \|x\| < \epsilon\sqrt{n}, \\ \exp(-n\delta/2)\pi_0(\lambda_n + x/\sqrt{n}) & \text{if } \|x\| \ge \epsilon\sqrt{n}. \end{cases}$$

When $\|x\| < \epsilon\sqrt{n}$, for n large enough we have $\|(\lambda_n + x/\sqrt{n}) - \lambda_0\| < \|\lambda_n - \lambda_0\| + \epsilon < 2\epsilon$. By the choice of ϵ, we have $\pi_0(\lambda_n + x/\sqrt{n}) \le 2\pi_0(\lambda_0)$. Since $(\|x\| + 1)^2 \le 2\|x\|^2 + 2$, we have $o_{P_{\lambda_0,\zeta_0}}(1)(\|x\| + 1)^2 \le \alpha(\|x\|^2 + 1)/2 = \alpha x^\top x/2 + \alpha/2$. Hence $g_n(x) \le h_n(x)$ with probability converge to 1 for n sufficiently large, combining with (3-11) when $\|x\| \ge \epsilon\sqrt{n}$.

Also, $h_n(x) \to h_0(x) = \exp\{-x^\top A_0 x/2 + \alpha/2\}2\pi_0(\lambda_0)$ pointwise. Now,

$$\begin{aligned} &\int_{\mathbb{R}^d} h_n(x)\,\mathrm{d}x \\ &= \int_{\|x\|<\epsilon\sqrt{n}} \exp(-x^\top A_0 x/2)\exp(\alpha/2)2\pi_0(\lambda_0)\,\mathrm{d}x + \int_{\|x\|\ge\epsilon\sqrt{n}} \exp(-n\delta/2)\pi_0(\lambda_n + x/\sqrt{n})\,\mathrm{d}x. \end{aligned}$$

The second term is less than

$\int_{\mathbb{R}^d} \exp(-n\delta/2)\pi_0(\lambda_n + x/\sqrt{n})\,\mathrm{d}x = \exp(-n\delta/2)\int_{\mathbb{R}^d} \pi_0(\lambda)n^{d/2}\,\mathrm{d}\lambda = \exp(-n\delta/2)n^{d/2} \to 0$,

while the first term monotonically converges to $\int_{\mathbb{R}^d} h_0(x)\,\mathrm{d}x$. Since g_n, g_0, h_n, h_0 are integrable, by the generalized dominated convergence theorem (the version for convergence in probability), we have $\int_{\mathbb{R}^d} |g_n(x) - g_0(x)|\,\mathrm{d}x \xrightarrow[n\to\infty]{P_{\lambda_0,\zeta_0}} 0$ and $\int_{\mathbb{R}^d} g_n(x)\,\mathrm{d}x \xrightarrow[n\to\infty]{P_{\lambda_0,\zeta_0}} \int_{\mathbb{R}^d} g_0(x)\,\mathrm{d}x$.

Let $a_n = 1/\int_{\mathbb{R}^d} g_n(x)\,\mathrm{d}x$ and $a_0 = 1/\int_{\mathbb{R}^d} g_0(x)\,\mathrm{d}x$. Then $a_n \to a_0$ in P_{λ_0,ζ_0}-probability, and thus

$$\begin{aligned}
&\int_{\mathbb{R}^d} \left| q_n(x) - \mathcal{N}\left(x \mid 0, H_0^{-1}\right) \right| \mathrm{d}x = \int_{\mathbb{R}^d} |a_n g_n(x) - a_0 g_0(x)|\,\mathrm{d}x \\
&\le \int_{\mathbb{R}^d} |a_n g_n(x) - a_n g_0(x)|\,\mathrm{d}x + \int_{\mathbb{R}^d} |a_n g_0(x) - a_0 g_0(x)|\,\mathrm{d}x \\
&\le |a_n| \int_{\mathbb{R}^d} |g_n(x) - g_0(x)|\,\mathrm{d}x + |a_n - a_0| \int_{\mathbb{R}^d} |g_0(x)|\,\mathrm{d}x \xrightarrow[n\to\infty]{P_{\lambda_0,\zeta_0}} 0.
\end{aligned}$$

This proves $\mathrm{d_{TV}}\left\{q_n, \mathcal{N}\left(0, H_0^{-1}\right)\right\} \xrightarrow[n\to\infty]{P_{\lambda_0,\zeta_0}} 0$, and (3-12) follows from (Miller, 2021, Lemma 28). □

B.2 An Illustration of Least Favorable Submodel

The Hilbert space $\widetilde{\mathcal{H}}$ highly depends on the parameter space $\mathcal{H}$. To illustrate how this works, we show an example from the Cox regression model without censoring.

Example B.1 (Cox regression model without censoring)**.** Consider the density function of the survival time $(y_1, \ldots, y_n)$ where $y_i \in \mathbb{R}_+ : \mathbb{R} \cap [0, \infty]$ with covariates $(x_1, \ldots, x_n)$ where $x_i \in \mathbb{R}$:

$$L(y_{1:n}, \zeta; \lambda) = \prod_{i=1}^{n} \exp(\lambda x_i)\zeta(y_i) \exp\{-\exp(\lambda x_i)Z(y_i)\},$$

where the parameter $\lambda \in \mathbb{R}$ and $Z(y) = \int_0^y \zeta(y)\,\mathrm{d}y$. The latent variable ζ is a "hazard function", which is a non-negative integrable function on $\mathbb{R}_+$, and belongs to the Hilbert space $\mathcal{H} = L^1(\mathbb{R}_+)$. The "cumulative hazard function" Z is a non-negative and non-decreasing function on $\mathbb{R}_+$. In regard to the model, we have

$$l_n(\lambda, \zeta) = \sum_{i=1}^{n} \{\lambda x_i + \log \zeta(y_i) - \exp(\lambda x_i)Z(y_i)\}/n,$$

and the λ-score function is $\dot{l}_n(\lambda, \zeta) = \sum_{i=1}^{n} \{x_i - x_i \exp(\lambda x_i)Z(y_i)\}/n$.

Let $\widetilde{H} = L^2(\mathbb{R}_+)$. Given a fixed function ζ_0 and a bounded function $\delta \in \widetilde{H}$, we can define a path $\{\zeta_\gamma^\delta \in \mathcal{H}\}_{\gamma\in\mathbb{R}}$ by $\zeta_\gamma^\delta(y) = \{1 + (\gamma - \lambda_0)\delta(y)\}\zeta_0(y)$ for all $y \in \mathbb{R}_+$. It satisfies that $\zeta_\gamma^\delta \to \zeta_0$ in $\mathcal{H}$ when $\gamma \to \lambda_0$. Also, we define $Z_\gamma^\delta(y) = \int_0^y \{1 + (\gamma - \lambda_0)\delta(y)\}\zeta_0(y)\,\mathrm{d}y$ correspondingly. Now, plugging ζ_γ^δ into $l_n(\lambda_0, \zeta)$ as ζ and differentiating it at $\gamma = \lambda_0$, we get the ζ-score function at $\zeta = \zeta_0$ in the direction of δ by

$$A^n_{\lambda_0,\zeta_0}\delta = \sum_{i=1}^n \{\delta(y_i) - \exp(\lambda_0 x_i)\int_0^{y_i} \delta(y)\zeta_0(y)\,\mathrm{d}y\}/n.$$

B.3 Comparison with Existing BvM Results on Semi-parametric Models

There is a rich theory literature on BvM results on semi-parametric models. Naturally, it is of interest to compare our results with them, contextualizing and clarifying our contribution. The existing results can roughly be divided into two categories. The first is similar to our focused setting where the posterior is obtained under a profile likelihood. Lee et al. (2005a) showed that $\mathbb{E}_{\lambda\sim\Pi(\lambda|y)} g\{\sqrt{n}(\lambda - \lambda_n)\}$ converges to $\mathbb{E}_{u\sim\mathcal{N}(0,\tilde{I}_0^{-1})} g(u)$ in probability, assuming a Taylor expansion form and for iid data. Their condition is similar to (3-10), and on the other hand, they do not give the result of the posterior density converging to normal density in total variation. Cheng and Kosorok (2008) showed a BvM result for the posterior induced from profile likelihood for iid probability model, under the assumption that the third derivatives exist. Compared to their result, ours is general in the sense that it is applicable to non-iid data and under potential non-differentiability.

The second category of BvM results relate to canonical Bayesian methodology involving integrated posterior $\Pi(\lambda \mid y) = \int \Pi(\lambda, \mathrm{d}\zeta \mid y)$ over a non-deterministic ζ. Bickel and Kleijn (2012) proved a BvM result for marginal posterior distribution of λ using the LAN property for the marginal likelihood of λ, which has similar form with (3-10). On the other hand, they additionally assume that the marginal posterior probability of λ inside the neighborhood $B_{M_n/\sqrt{n}}(\lambda_0)$ converges to 1 for every $M_n \to \infty$. This condition is similar to but arguably stronger than (3-11). Castillo and Rousseau (2015) proved a BvM result on a functional of the parameters,

under an essentially necessary no-bias condition which is related to both the likelihood and the prior specification of (λ, ζ).

In Bickel and Kleijn (2012), to get the LAN property for the marginal likelihood of λ, they first assume that a neighborhood of z has enough prior mass, and the parameter space of z has bounded Hellinger metric entropy (covering number). They assume that for z outside a neighborhood of the fixed z_0, the Hellinger distances between the likelihood at $\lambda_0 + h_n/\sqrt{n}$ and the one at λ_0 are uniformly (with respect to z) infinitesimal for every bounded h_n. Finally, assuming that the least favorable submodel exists, that is a submodel $l_n(\lambda, z_\lambda)$ with parameters (λ, z_λ) satisfying $\dot{\mathscr{Q}}l_n(\lambda, z_\lambda) = \dfrac{\partial l_n(\lambda, z_\lambda)}{\partial \lambda}$ for all λ in a neighborhood of λ_0, the conditional posterior distribution of z can be shown to concentrate around the parameter of the least favorable submodel z_λ, that is the probability that the Hellinger distance between z and z_λ is greater than a positive number converges to zero. This leads to the marginal LAN property assuming that the full likelihood has LAN property in the direction of λ when the z is perturbed around the least favorable submodel, that is $z = z_\lambda + \zeta$ for all ζ in a neighborhood of 0.

In Castillo and Rousseau (2015), let h be the least favorable direction satisfying $\dot{\mathscr{P}}l_n(\lambda_0, z_0) = A_{\lambda_0, z_0} h$. Let $\lambda_t = \lambda - t\tilde{I}_0^{-1}/\sqrt{n}$, and $z_t = z + th\tilde{I}_0^{-1}/\sqrt{n}$. Assume over the direction of h, the log full likelihood has the LAN property with a remainder term, and the difference between the remainder terms at (λ, z) and at (λ_t, z_t) converges to zero. Further, suppose that the ratio between the integral of the likelihood under the prior at (λ_t, z_t) and the one at (λ, z) converge to 1 for all (λ, z). They proved the BvM theorem where the mean is λ_0 plus the first order term of the LAN expansion.

B.4 ADMM for Optimization Problem in Data Application

To solve the optimization problem

$$\min_{\zeta} \frac{1}{2}\|\mathcal{L} - \zeta\|^2 + \tilde{\lambda}\|\zeta\|_* \quad \text{subject to } \zeta \in \mathbb{R}^{n\times n}, \zeta_{i,i} = \sum_{j:j\neq i} \zeta_{i,j}, \zeta_{i,j} = \zeta_{j,i} \leq 0 \text{ for } i \neq j,$$

we use ADMM under constraints and log-barrier:

$$\min_{\zeta,Z} \frac{1}{2}\|\mathcal{L}-\zeta\|^2 + \rho \sum_{(i,j):i\neq j} \{-\log(-\zeta_{i,j})\} + \tilde{\lambda}\|Z\|_* + \frac{\eta}{2}\|\zeta - Z + W\|^2$$
$$\text{subject to } \zeta_{i,i} = -\sum_{j:j\neq i} \zeta_{i,j}, \zeta_{i,j} = \zeta_{j,i} \leq 0 \text{ for } i \neq j,$$

where $W = W^T$ is the Lagrangian multiplier. The ADMM algorithm iterates the following steps:

1. Constrained gradient descent for ζ: set ζ to be

$$\arg\min_{\zeta} \frac{1}{2}\|\mathcal{L}-\zeta\|^2 + \rho \sum_{(i,j):i\neq j} \{-\log(-\zeta_{i,j})\} + \frac{\eta}{2}\|\zeta - Z + W\|^2$$
$$\text{subject to } \zeta_{i,i} = -\sum_{j:j\neq i} \zeta_{i,j}, \zeta_{i,j} = \zeta_{j,i} \text{ for } i \neq j.$$

   The constraints are easy to satisfy, by restricting the free parameters to $\{\zeta_{i,j}\}_{i>j}$, and setting $\zeta_{i,i} = -\sum_{j<i} \zeta_{i,j} - \sum_{j>i} \zeta_{j,i}$.

2. Minimizing over Z: set $Z = S_{\tilde{\lambda}/\eta}(\zeta + W)$ where $S_{\tilde{\lambda}/\eta}(X) = \sum_{i=1}^{n} (\sigma_i - \tilde{\lambda}/\eta)_+ u_i v_i^T$, and $X = U\text{diag}(\sigma_i)V$ is the singular value decomposition. The solution of this step satisfies the conditions of symmetry and the rows add to zero.

3. Updating W: set W to be $\zeta - Z + W$.

APPENDIX C
APPENDIX FOR CHAPTER 4

C.1 Proofs

Proof of Theorem 1. Let $A_0 := \nabla^2_{zz} h(\beta_0, z_0; y)$, and consider a mapping:

$$G(\beta, z) = z - A_0^{-1} \nabla_z h(\beta, z; y),$$

for which a fixed point z of $G(\beta, \cdot)$ satisfies $G(\beta, z) = z \Rightarrow \nabla_z h(\beta, z; y) = 0$. We can now show $G(\beta, \cdot)$ is a contraction for z inside a neighborhood of z_0 defined via the Hessian $\nabla^2_{zz} h(\beta, z; y)$.

Since $\nabla_z G(\beta_0, z_0) = I - A_0^{-1} A_0 = 0$, by the continuity of $\nabla^2_{zz} h(\beta, z; y)$, we can find z in a neighborhood of z_0, $\mathbb{B}(z_0, k; \beta_0) := \{z : \|\nabla_z G(\beta_0, z)\|_{op} = \|I - A_0^{-1} \nabla^2_{zz} h(\beta_0, z; y)\|_{op} \leq k < 1\}$, by the mean value theorem, $G(\beta_0, \cdot)$ is a contraction in $\mathbb{B}(z_0, k; \beta_0)$, i.e. $\|G(\beta_0, z_0) - G(\beta_0, z)\| \leq k\|z_0 - z\|$. It follows that

$$\begin{aligned}
\|z - z_0\| &= \|G(\beta_0, z) - G(\beta_0, z_0) + z - G(\beta_0, z)\| \\
&\leq k\|z - z_0\| + \|z - G(\beta_0, z)\| \\
&= k\|z - z_0\| + \|A_0^{-1} \nabla_z h(\beta, z; y)\| \\
&\leq k\|z - z_0\| + \|A_0^{-1}\|_{op} \epsilon \\
&= k\|z - z_0\| + \lambda^{-1}_{\min}(A_0)\epsilon.
\end{aligned}$$

Rearranging terms yields the result. □

Proof of Lemma 1. Let η, γ be any given positive numbers. By the Assumption 1, there exists a large enough N such that

$$P_{\beta_0, z_0}\Big(\Pi[d_H\{z, \hat{z}(\beta); \beta\} < \rho_n \mid \beta = \beta_n; y] > e^{-\eta}\Big) > 1 - \gamma$$

for all $n > N$. Let $D(\beta_n, \rho_n) = \{z : d_H\{z, \hat{z}(\beta_n); \beta_n\} < \rho_n\}$. Note that $\Pi(z \mid \beta = \beta_n) = L_n(\beta_n, z; y) / \int_{\mathcal{H}} L_n(\beta_n, z; y)\, \mathrm{d}\Pi_0(z) = L_n(\beta_n, z; y)/S_n(\beta_n)$, we have

$\Pi[d_H\{z, \hat{z}(\beta); \beta\} < \rho_n \mid \beta = \beta_n; y] = \int_{D(\beta_n,\rho_n)} L_n(\beta_n, z; y)\, \mathrm{d}\Pi_0(z)/S_n(\beta_n)$. Hence,

$$P_{\beta_0,z_0}\left(\log \int_{D(\beta_n,\rho_n)} L_n(\beta_n, z; y)\, \mathrm{d}\Pi_0(z) - \log S_n(\beta_n) > -\eta\right) > 1 - \gamma\,.$$

We only need to prove $\log \int_{D(\beta_n,\rho_n)} L_n(\beta_n, z; y)\, \mathrm{d}\Pi_0(z)$ satisfies the conclusion of Lemma 1.

We define the events $F_n(z, \epsilon) = \{\sup_{h_n} |h_n^{\mathrm{T}} H_n(z) h_n - h_n^{\mathrm{T}} H_0 h_n| \le \epsilon\}$. Then

$$\begin{aligned}
&\int_{D(\beta_n,\rho_n)} L_n(\beta_n, z; y)\, \mathrm{d}\Pi_0(z) \\
&= \int_{D(\beta_n,\rho_n)} L_n(\beta_n, z; y) 1_{F_n(z,\epsilon)}\, \mathrm{d}\Pi_0(z) + \int_{D(\beta_n,\rho_n)} L_n(\beta_n, z; y) 1_{F_n^c(z,\epsilon)}\, \mathrm{d}\Pi_0(z)\,.
\end{aligned}$$

Now, the integral of the second term

$$\begin{aligned}
&\int \int_{D(\beta_n,\rho_n)} L_n(\beta_n, z; y) 1_{F_n^c(z,\epsilon)}\, \mathrm{d}\Pi_0(z) \mathrm{d}P_{\beta_0,z_0}(y) \\
&= \int_{D(\beta_n,\rho_n)} \int L_n(\beta_n, z; y) 1_{F_n^c(z,\epsilon)}\, \mathrm{d}P_{\beta_0,z_0}(y) \mathrm{d}\Pi_0(z).
\end{aligned}$$

using Fubini's theorem. Using the Fatou's lemma with Assumption 3 as a domination condition, we have

$$\begin{aligned}
&\limsup_{n\to\infty} \int_{D(\beta_n,\rho_n)} \int L_n(\beta_n, z; y) 1_{F_n^c(z,\epsilon)}\, \mathrm{d}P_{\beta_0,z_0}(y) \mathrm{d}\Pi_0(z) \\
&\le \int \limsup_{n\to\infty} 1_{D(\beta_n,\rho_n)} \int L_n(\beta_n, z; y) 1_{F_n^c(z,\epsilon)}\, \mathrm{d}P_{\beta_0,z_0}(y) \mathrm{d}\Pi_0(z) \\
&= \int \limsup_{n\to\infty} 1_{z=\hat{z}(\hat{\beta}_n)} \int L_n(\beta_n, z; y) 1_{F_n^c(z,\epsilon)}\, \mathrm{d}P_{\beta_0,z_0}(y) \mathrm{d}\Pi_0(z) \\
&\overset{(a)}{=} 0,
\end{aligned}$$

where (a) is using $F_n^c\{\hat{z}(\hat{\beta}_n), \epsilon\} \to \varnothing$. Hence,

$$\int_{D(\beta_n,\rho_n)} L_n(\beta_n, z; y)\, \mathrm{d}\Pi_0(z) = \int_{D(\beta_n,\rho_n)} L_n(\beta_n, z; y) 1_{F_n(z,\epsilon)}\, \mathrm{d}\Pi_0(z) + o_{P_{\beta_0,z_0}}(1).$$

Let $A_n(z,\epsilon) = \{|r_n(h_n,z)| \le \epsilon\}$. Then

$$\begin{aligned}
&\int_{D(\beta_n,\rho_n)} L_n(\beta_n,z;y) 1_{F_n(z,\epsilon)} \,\mathrm{d}\Pi_0(z) \\
&= \int_{D(\beta_n,\rho_n)} L_n(\beta_n,z;y) 1_{F_n(z,\epsilon)} 1_{A_n(z,\epsilon)} \,\mathrm{d}\Pi_0(z) + \int_{D(\beta_n,\rho_n)} L_n(\beta_n,z;y) 1_{F_n(z,\epsilon)} 1_{A_n^c(z,\epsilon)} \,\mathrm{d}\Pi_0(z) \,.
\end{aligned}$$

Similarly, the integral of the second term

$$\begin{aligned}
&\int \int_{D(\beta_n,\rho_n)} L_n(\beta_n,z;y) 1_{F_n(z,\epsilon)} 1_{A_n^c(z,\epsilon)} \,\mathrm{d}\Pi_0(z) \mathrm{d}P_{\beta_0,z_0}(y) \\
&= \int_{D(\beta_n,\rho_n)} \int L_n(\beta_n,z;y) 1_{F_n(z,\epsilon)} 1_{A_n^c(z,\epsilon)} \,\mathrm{d}P_{\beta_0,z_0}(y) \mathrm{d}\Pi_0(z).
\end{aligned}$$

using Fubini's theorem. Using the Fatou's lemma with Assumption 3 as a domination condition, we have

$$\begin{aligned}
&\limsup_{n\to\infty} \int_{D(\beta_n,\rho_n)} \int L_n(\beta_n,z;y) 1_{F_n(z,\epsilon)} 1_{A_n^c(z,\epsilon)} \,\mathrm{d}P_{\beta_0,z_0}(y) \mathrm{d}\Pi_0(z) \\
&\le \int \limsup_{n\to\infty} 1_{D(\beta_n,\rho_n)} \int L_n(\beta_n,z;y) 1_{F_n(z,\epsilon)} 1_{A_n^c(z,\epsilon)} \,\mathrm{d}P_{\beta_0,z_0}(y) \mathrm{d}\Pi_0(z) \\
&\overset{(a)}{\le} \int \limsup_{n\to\infty} \int L_n(\beta_n,z;y) 1_{A_n^c(z,\epsilon)} \,\mathrm{d}P_{\beta_0,z_0}(y) \mathrm{d}\Pi_0(z) \\
&\overset{(b)}{=} 0,
\end{aligned}$$

where (a) is using $1_{F_n(z,\epsilon)} \le 1$ and $1_{D(\beta_n,\rho_n)} \le 1$, and (b) is using $r_n(h_n,z) = o_{P_{\beta_0,z_0}}(1)$. Hence,

$$\int_{D(\beta_n,\rho_n)} L_n(\beta_n,z;y) \,\mathrm{d}\Pi_0(z) = \int_{D(\beta_n,\rho_n)} L_n(\beta_n,z;y) 1_{F_n(z,\epsilon)} 1_{A_n(z,\epsilon)} \,\mathrm{d}\Pi_0(z) + o_{P_{\beta_0,z_0}}(1). \quad \text{(C-1)}$$

Let N large enough such that $\rho_n < \delta$ for all $n > N$. For all z satisfying $d_H\{z, \hat{z}(\beta_n); \beta_n\} < \rho_n$, we have $l_n(\beta_n,z) = l_n\{\hat{\beta}_n, \hat{z}(\hat{\beta}_n)\} - \frac{1}{2} h_n^{\mathrm{T}} H_n(z) h_n + r_n(h_n,z)$ by Assumption 2. Therefore,

$$\begin{aligned}
&\int_{D(\beta_n,\rho_n)} L_n(\beta_n,z;y) 1_{F_n(z,\epsilon)} 1_{A_n(z,\epsilon)} \,\mathrm{d}\Pi_0(z) \\
&= \int_{D(\beta_n,\rho_n)} \exp\left\{ l_n\{\hat{\beta}_n, \hat{z}(\hat{\beta}_n)\} - \frac{1}{2} h_n^{\mathrm{T}} H_n(z) h_n + r_n(h_n,z) \right\} 1_{F_n(z,\epsilon)} 1_{A_n(z,\epsilon)} \,\mathrm{d}\Pi_0(z).
\end{aligned}$$

For all $y \in F_n(z,\epsilon) \cap A_n(z,\epsilon)$, we have

$$-\frac{1}{2}h_n^{\mathrm{T}}H_0h_n - 2\epsilon \le -\frac{1}{2}h_n^{\mathrm{T}}H_n(z)h_n + r_n(h_n,z) \le -\frac{1}{2}h_n^{\mathrm{T}}H_0h_n + 2\epsilon,$$

so we have the lower and upper bounds

$$\begin{aligned}
&\int_{D(\beta_n,\rho_n)} \exp\left\{l_n\{\hat{\beta}_n,\hat{z}(\hat{\beta}_n)\} - \frac{1}{2}h_n^{\mathrm{T}}H_0h_n - 2\epsilon\right\} 1_{F_n(z,\epsilon)}1_{A_n(z,\epsilon)}\,\mathrm{d}\Pi_0(z)\\
&\le \int_{D(\beta_n,\rho_n)} L_n(\beta_n,z;y)1_{F_n(z,\epsilon)}1_{A_n(z,\epsilon)}\,\mathrm{d}\Pi_0(z)\\
&\le \int_{D(\beta_n,\rho_n)} \exp\left\{l_n\{\hat{\beta}_n,\hat{z}(\hat{\beta}_n)\} - \frac{1}{2}h_n^{\mathrm{T}}H_0h_n + 2\epsilon\right\} 1_{F_n(z,\epsilon)}1_{A_n(z,\epsilon)}\,\mathrm{d}\Pi_0(z).
\end{aligned}$$

Especially, when $h_n = 0$ and $\beta_n = \hat{\beta}_n$, we have $l_n(\hat{\beta}_n,z) = l_n[\hat{\beta}_n,\hat{z}(\hat{\beta}_n)] + r_n(h_n,z)$. Hence,

$$\begin{aligned}
&\int_{D(\beta_n,\rho_n)} \exp(l_n\{\hat{\beta}_n,\hat{z}(\hat{\beta}_n)\} - \epsilon) 1_{F_n(z,\epsilon)}1_{A_n(z,\epsilon)}\,\mathrm{d}\Pi_0(z)\\
&\le \int_{D(\beta_n,\rho_n)} L_n(\hat{\beta}_n,z;y)1_{F_n(z,\epsilon)}1_{A_n(z,\epsilon)}\,\mathrm{d}\Pi_0(z)\\
&\le \int_{D(\beta_n,\rho_n)} \exp(l_n\{\hat{\beta}_n,\hat{z}(\hat{\beta}_n)\} + \epsilon) 1_{F_n(z,\epsilon)}1_{A_n(z,\epsilon)}\,\mathrm{d}\Pi_0(z).
\end{aligned}$$

We combine the above two group of inequalities to induce

$$\begin{aligned}
&\exp\left\{-\frac{1}{2}h_n^{\mathrm{T}}H_0h_n - 3\epsilon\right\} \int_{D(\beta_n,\rho_n)} L_n(\hat{\beta}_n,z;y)1_{F_n(z,\epsilon)}1_{A_n(z,\epsilon)}\,\mathrm{d}\Pi_0(z)\\
&\le \int_{D(\beta_n,\rho_n)} L_n(\beta_n,z;y)1_{F_n(z,\epsilon)}1_{A_n(z,\epsilon)}\,\mathrm{d}\Pi_0(z)\\
&\le \exp\left\{-\frac{1}{2}h_n^{\mathrm{T}}H_0h_n + 3\epsilon\right\} \int_{D(\beta_n,\rho_n)} L_n(\hat{\beta}_n,z;y)1_{F_n(z,\epsilon)}1_{A_n(z,\epsilon)}\,\mathrm{d}\Pi_0(z).
\end{aligned}$$

Hence,

$$\begin{aligned}
&-\frac{1}{2}h_n^{\mathrm{T}}H_0h_n - 3\epsilon + \log\int_{D(\beta_n,\rho_n)} L_n(\hat{\beta}_n,z;y)1_{F_n(z,\epsilon)}1_{A_n(z,\epsilon)}\,\mathrm{d}\Pi_0(z)\\
&\le \log\int_{D(\beta_n,\rho_n)} L_n(\beta_n,z;y)1_{F_n(z,\epsilon)}1_{A_n(z,\epsilon)}\,\mathrm{d}\Pi_0(z)\\
&\le -\frac{1}{2}h_n^{\mathrm{T}}H_0h_n + 3\epsilon + \log\int_{D(\beta_n,\rho_n)} L_n(\hat{\beta}_n,z;y)1_{F_n(z,\epsilon)}1_{A_n(z,\epsilon)}\,\mathrm{d}\Pi_0(z).
\end{aligned}$$

By (C-1), we have

$$
\begin{aligned}
&-\frac{1}{2}h_n^{\mathrm{T}}H_0h_n - 3\epsilon + \log\int_{D(\beta_n,\rho_n)} L_n(\hat{\beta}_n, z; y)\,\mathrm{d}\Pi_0(z) + o_{P_{\beta_0,z_0}}(1)\\
&\le \log\int_{D(\beta_n,\rho_n)} L_n(\beta_n, z; y)\,\mathrm{d}\Pi_0(z)\\
&\le -\frac{1}{2}h_n^{\mathrm{T}}H_0h_n + 3\epsilon + \log\int_{D(\beta_n,\rho_n)} L_n(\hat{\beta}_n, z; y)\,\mathrm{d}\Pi_0(z) + o_{P_{\beta_0,z_0}}(1).
\end{aligned}
$$

Hence, we have

$$
\log\int_{D(\beta_n,\rho_n)} L_n(\beta_n, z; y)\,\mathrm{d}\Pi_0(z) = \log\int_{D(\beta_n,\rho_n)} L_n(\hat{\beta}_n, z; y)\,\mathrm{d}\Pi_0(z) - \frac{1}{2}h_n^{\mathrm{T}}H_0h_n + o_{P_{\beta_0,z_0}}(1).
$$

This suffices to prove Lemma 1. □

Proof of Theorem 2. We have $\pi_n(\beta) = \exp\{s_n(\beta)\}\pi_0(\beta)/m_n$ where $m_n = \int_{\mathbb{R}^d}\exp\{s_n(\beta)\}\pi_0(\beta)\,\mathrm{d}\beta$, and $q_n(x) = \pi_n(\hat{\beta}_n + x/\sqrt{n})n^{-d/2}$. Let

$$
g_n(x) = q_n(x)\exp\{-s_n(\hat{\beta}_n)\}n^{d/2}m_n = \exp\{s_n(\hat{\beta}_n + x/\sqrt{n}) - s_n(\hat{\beta}_n)\}\pi_0(\hat{\beta}_n + x/\sqrt{n})
$$

and define $g_0(x) = \exp\{-x^{\mathrm{T}}H_0x/2\}\pi_0(\beta_0)$. We first show that $\int_B |g_n(x) - g_0(x)|\,\mathrm{d}x \xrightarrow[n\to\infty]{P_{\beta_0,z_0}} 0$, where $B = \{x : \|x\| \le M\}$ for any fix positive number M. Since π_0 is continuous at β_0, we choose sufficiently small ϵ such that $\pi_0(\beta) \le 2\pi_0(\beta_0)$ for all $\beta \in B_\epsilon(\beta_0)$.

Since $s_n(\hat{\beta}_n + x/\sqrt{n}) - s_n(\hat{\beta}_n) = -\frac{1}{2}x^{\mathrm{T}}H_0x + o_{P_{\beta_0,z_0}}(1)$, and π_0 is continuous at β_0 and $\hat{\beta}_n + x/\sqrt{n} \to \beta_0$, we have $g_n(x) \to g_0(x)$ pointwise with probability converge to 1. Consider n sufficiently large such that the term $o_{P_{\beta_0,z_0}}(1) < \alpha/2$ where α is less than the smallest eigenvalue of H_0. We denote $A_0 = H_0 - \alpha I$, and define

$$
h_0(x) = \exp(-x^{\mathrm{T}}A_0x/2)2\pi_0(\beta_0).
$$

When $\|x\| \le M$, for n large enough we have $\|(\hat{\beta}_n + x/\sqrt{n}) - \beta_0\| < \epsilon$. By the choice of ϵ, we have $\pi_0(\hat{\beta}_n + x/\sqrt{n}) \le 2\pi_0(\beta_0)$. Hence $g_n(x) \le h_0(x)$ with probability converge to 1 for n sufficiently

large. Since g_n, g_0, h_0 are integrable, by the dominated convergence theorem (the version for convergence in probability), we have $\int_B |g_n(x) - g_0(x)| \, \mathrm{d}x \xrightarrow[n\to\infty]{P_{\beta_0,z_0}} 0$ and $\int_B g_n(x) \, \mathrm{d}x \xrightarrow[n\to\infty]{P_{\beta_0,z_0}} \int_B g_0(x) \, \mathrm{d}x$.

Let $a_n = 1/\int_B g_n(x) \, \mathrm{d}x$ and $a_0 = 1/\int_B g_0(x) \, \mathrm{d}x$. Then $a_n \to a_0$ in P_{β_0,z_0}-probability, and thus

$$\begin{aligned}
&\int_B \left| \frac{q_n(x)}{\int_B q_n} - \mathcal{N}^B\left(x \mid 0, H_0^{-1}\right) \right| \mathrm{d}x = \int_B |a_n g_n(x) - a_0 g_0(x)| \, \mathrm{d}x \\
&\leq \int_B |a_n g_n(x) - a_n g_0(x)| \, \mathrm{d}x + \int_B |a_n g_0(x) - a_0 g_0(x)| \, \mathrm{d}x \\
&\leq |a_n| \int_B |g_n(x) - g_0(x)| \, \mathrm{d}x + |a_n - a_0| \int_B |g_0(x)| \, \mathrm{d}x \xrightarrow[n\to\infty]{P_{\beta_0,z_0}} 0.
\end{aligned}$$

Then we can choose slowly enough increasing $M_n \to \infty$ such that

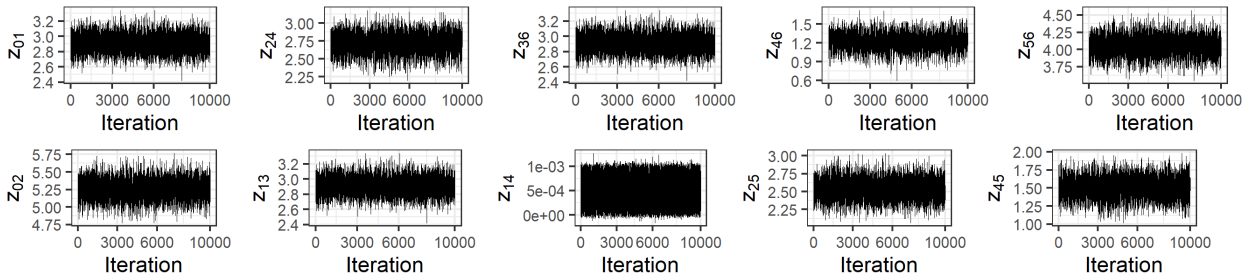
$$\int_{B_n} \left| \frac{q_n(x)}{\int_{B_n} q_n} - \mathcal{N}^{B_n}\left(x \mid 0, H_0^{-1}\right) \right| dx \xrightarrow[n\to\infty]{P_{\beta_0,z_0}} 0$$

where $B_n = \{x : \|x\| \leq M_n\}$. By Bickel and Kleijn (2012, Lemma 6.1), $\Pi(B_n^c) = \int_{B_n^c} q_n(x) \, dx \xrightarrow[n\to\infty]{P_{\beta_0,z_0}} 0$. Since $\|\Pi - \Pi^B\| \leq 2\Pi(B^c)$, this proves $\mathrm{d_{TV}}\left\{q_n, \mathcal{N}\left(0, H_0^{-1}\right)\right\} \xrightarrow[n\to\infty]{P_{\beta_0,z_0}} 0$, and the concentration follows from Miller (2021, Lemma 28). □

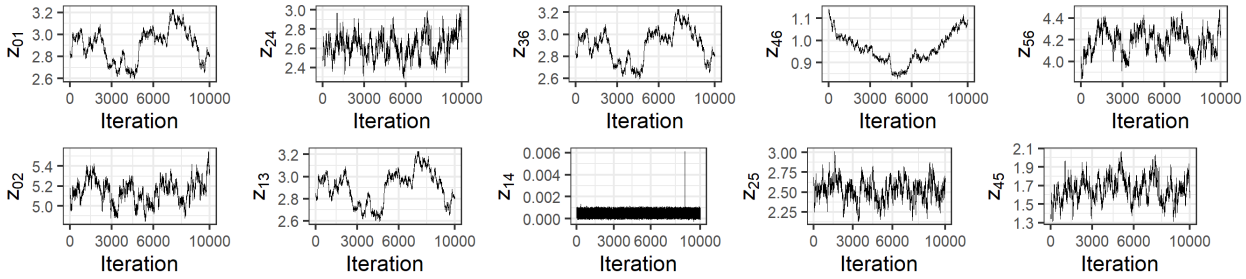
C.2 Additional Simulation results

C.2.1 Flow network

Figure C-1 and C-2 show the traces of the Markov chains of the posterior samples of z and β in Section 5.

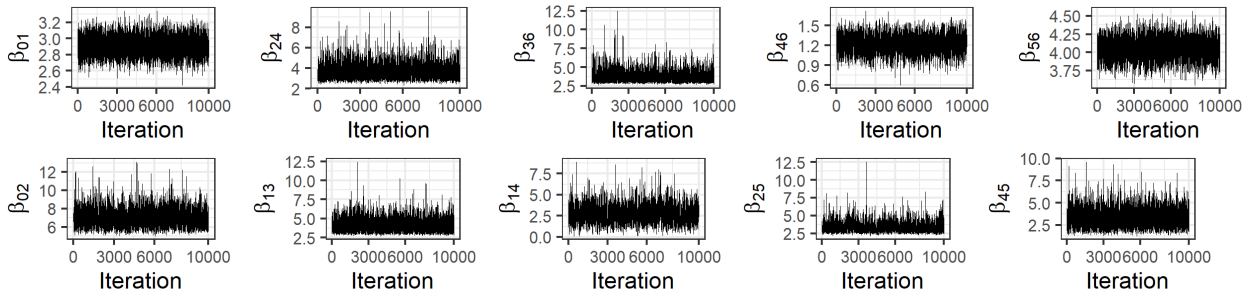


(a) Using our suggested inverse mass matrix

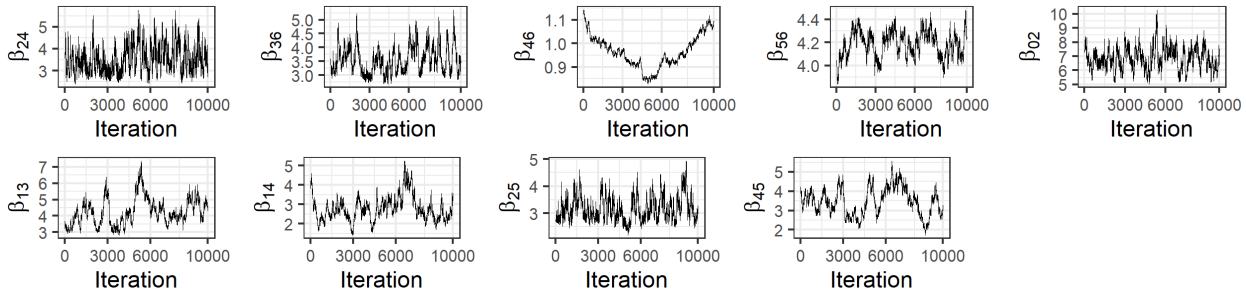


(b) Using the default adapted matrix as the inverse mass

Figure C-1. Traceplots of Markov chain for sampling the posterior of z using different choices of inverse mass matrix.



(a) Using our suggested inverse mass matrix



(b) Using the default adapted matrix as the inverse mass

Figure C-2. Traceplots of Markov chains for sampling the posterior of β using different choices of inverse mass matrix.

C.2.2 Latent quadratic model

We use another numerical experiment on latent quadratic model to illustrate the advantage of the dual form. Consider the canonical latent normal model with the likelihood kernel:

$$g(y;\beta,z)=\exp\left\{-\frac{1}{2}z^{\mathrm{T}}Q^{-1}(\beta;x)z\right\}\prod_{i=1}^{n}v(y_i\mid z_i),$$

where $v(y_i \mid z_i)$ is a log-concave conditional density of y_i, and $y_1,\ldots,y_n$ are assumed to be conditionally independent given z. The covariance structure of the latent variable z is parameterized by $Q(\beta;x)$, where $Q(\beta;x)_{i,j}=\tau\exp\{-\|x_i-x_j\|^2/(2b)\}$ with $x_i\in\mathbb{R}^d$ representing observed predictors or spatial locations. The parameter $\beta=(\tau,b)\in\mathbb{R}^2$ controls the scale of dependence in the covariance structure. As a concrete example, we consider binary observations y_i from Bernoulli distribution under logistic link $v(y_i\mid z_i)=\exp(y_iz_i)/\{1+\exp(z_i)\}$. This formulation corresponds to a latent Gaussian process model for binary classification, where z acts as an underlying continuous latent variable governing the probability of success.

To create a gradient-bridged posterior, we seek to minimize $-\log g(y;\beta,z)$ over $z\in\mathbb{R}^n$. We define the optimization function $h(\beta,z;y)=-\log g(y;\beta,z)$. However, direct posterior computation involves the inversion $Q^{-1}(\beta;x)$, which is computationally expensive with a complexity of $O(n^3)$. Hence, we use the dual form to address this inefficiency.

Since $-\log g(y;\beta,z)$ can be decomposed into the sum of a quadratic function and a convex function, we use the variable splitting technique by introducing the constraint $u=z$. Introducing the Lagrange multiplier $\alpha\in\mathbb{R}^n$, the Lagrangian dual function takes the form:

$$\begin{aligned}h^{\dagger}(\beta,\alpha;y)&=\inf_{z,u}\ \frac{1}{2}z^{\mathrm{T}}Q^{-1}(\beta;x)z+\alpha^{\mathrm{T}}(z-u)+\sum_{i=1}^{n}\left[-y_iu_i+\log\{1+\exp(u_i)\}\right]\\&=-\frac{1}{2}\alpha^{\mathrm{T}}Q(\beta;x)\alpha-\sum_{i=1}^{n}\left\{(\alpha_i+y_i)\log(\alpha_i+y_i)+(1-\alpha_i-y_i)\log(1-\alpha_i-y_i)\right\}\end{aligned}$$

subject to $\alpha_i+y_i\in(0,1)$ $(i=1,\ldots,n)$. Strong duality holds in this setting, ensuring that $\sup_\alpha h^{\dagger}(\beta,\alpha;y)=\inf_z h(\beta,z;y)$, where the optimal solutions satisfy $\hat{z}=-Q(\beta;x)\hat{\alpha}$ with

$\hat{z} = \arg\inf_z h(\beta, z; y)$ and $\hat{\alpha} = \arg\sup_\alpha h^\dagger(\beta, \alpha; y)$. This allows us to reparameterize $z = -Q(\beta; x)\alpha$, and then use $-h^\dagger$ as the gradient-bridge function, leading to the gradient-bridged posterior:

$$L(y; \beta, \alpha) \propto g\{y; \beta, -Q(\beta; x)\alpha\} \exp\left\{-\lambda \|\nabla_\alpha h^\dagger(\beta, \alpha; y)\|_2^2\right\},$$

where we can explicitly calculate

$$\nabla_\alpha h^\dagger(\beta, \alpha; y) = -Q(\beta; x)\alpha - \log(\alpha + y) + \log(1 - \alpha - y).$$

This guarantees that the conditional maximum likelihood estimator is $\alpha = \hat{\alpha}$, which coincides with $\hat{z} = -Q(\beta; x)\hat{\alpha}$.

A key advantage of this approach is that neither the gradient-bridge function nor its gradient with respect to α requires the inversion Q^{-1}, significantly improving the posterior computational efficiency. Additionally, evaluating $g(y; \beta, z)$ remains quick since $z^{\mathrm{T}} Q^{-1}(\beta; x) z = \alpha^{\mathrm{T}} Q(\beta; x)\alpha$.

To simulate data for benchmarking, we generate 1000 random locations $x_1, \ldots, x_{1000} \sim \text{Uniform}(-6, 6)$. The ground truth latent curve $\tilde{z}$ is defined using a cubic spline interpolation with 20 control points, which are evenly spaced along $[-6, 6]$, with their corresponding z-values sampled from $\text{Uniform}(-3, 3)$. The spline is evaluated at each x_i to produce $\tilde{z}_i$. Finally, binary observations are generated as $y_i \sim \text{Bernoulli}\left[1/\{1 + \exp(-\tilde{z}_i)\}\right]$.

We fit the gradient-bridged posterior, the bridged posterior and the Gibbs posterior to the simulated data. For all models, we assign independent $\text{Ga}^{-1}(2, 0.1)$ priors on τ and b. We use no-u-turn sampler for the gradient-bridged posterior and the Gibbs posterior, and use random walk Metropolis for the bridged posterior.

We run each MCMC algorithm for $10,000$ iterations and discard the first $4,000$ as burn-ins, and apply thinning at 20. Figure C-3 shows the traces of the first 3 elements of w, and the autocorrelation functions for all elements of w. The mixing performance is very good. Figure C-4 shows the mixing of the parameters b and τ.

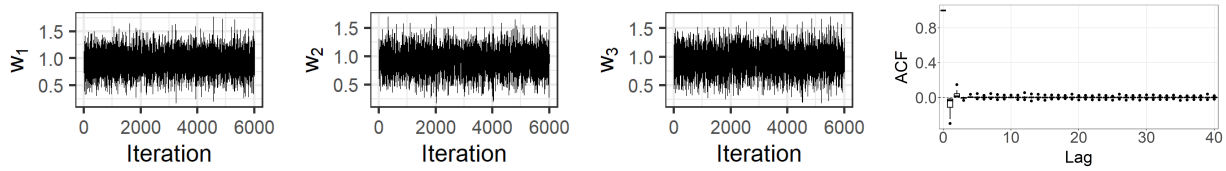


Figure C-3. The trace of the Markov chain of posterior samples for the first 3 components of w, and the autocorrelation functions for all elements of w.

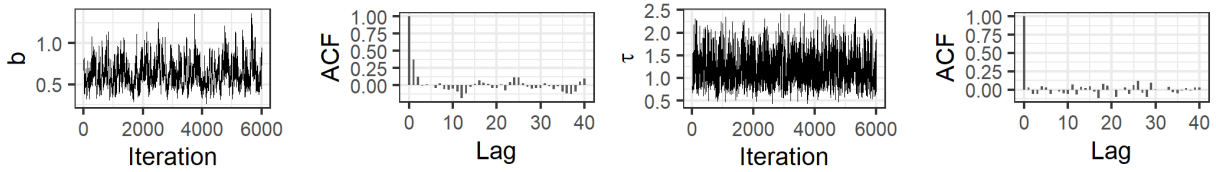


Figure C-4. Traces and autocorrelation functions for the posterior samples of b and τ.

We compare the posterior distributions of parameters (τ, b). As can be seen in Fig. C-5, these distributions show a similar range of τ and b. Since the distributions of τ or b arise from different models, their distributions are not expected to align precisely.

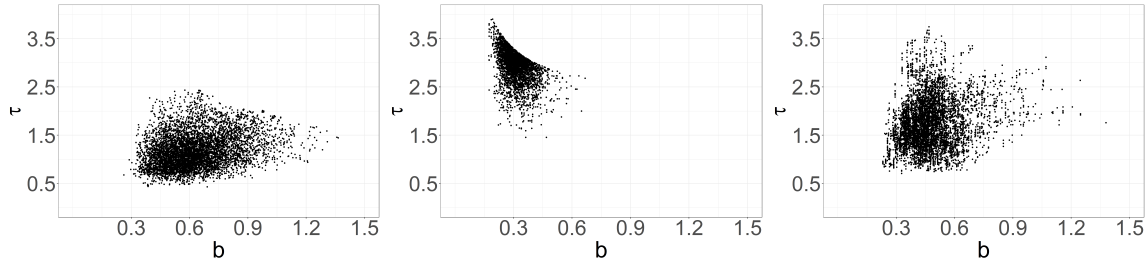


Figure C-5. The posterior distributions of the covariance kernel parameters τ and b for the gradient-bridged posterior (left), the bridged posterior (middle) and the Gibbs posterior (right).

C.3 Data preprocessing for single-cell data integration

The preprocessing of the panc8 dataset is conducted by first grouping the data based on sequencing technology. Within each group, gene expression levels are standardized to ensure comparability. Then, highly variable genes are identified to capture the most informative features, followed by the application of principal component analysis. The number of principal components to retain is determined using a scree plot, with five components selected as the optimal choice.

To standardize sample sizes across all technology groups, the group with the smallest number of samples is identified, and denoted as X^*_{raw}. Cell types within X^*_{raw} containing 20 or

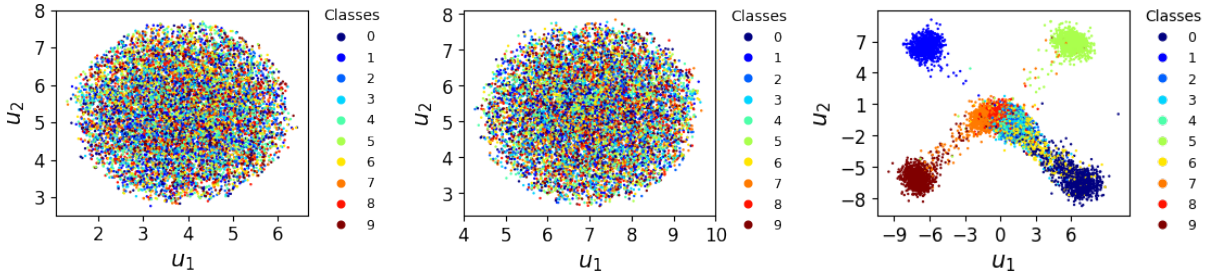
fewer samples are removed to prevent the influence of small sample sizes on integration. For the remaining groups, downsampling is performed using stratified sampling to ensure that the sample counts for each cell type matched those in X^*_{raw}. If the available sample count for a given cell type in a group is insufficient to meet the target, additional samples are randomly drawn from other cell types within the group to maintain the overall sample size. This procedure ensured consistency in sample sizes across all groups while preserving the distribution of cell types.

APPENDIX D
APPENDIX FOR CHAPTER 5: MNIST HANDWRITTEN DIGIT IMAGES

In this chapter, additional experiment results are shown using the MNIST data.

The MNIST data is used to illustrate the CDE when x is discrete. This dataset contains $70,000$ processed images of handwritten digits, each having 28×28 pixels. The authors use $60,000$ for training purposes and the remaining for testing. Each image y_i is associated with a discrete label x_i recording the ground-truth digit from 0 to 9.

The structure of this dataset is very similar to FashionMNIST (Xiao et al., 2017) which is used in the main paper, and the same models as in Section 3.1 is used. The latent $z_{P,i}$ is chosen to be $z^{(3)}_{1:16,1,1}$. Unsurprisingly, the results corresponding to this dataset is quite similar to the ones from the FashionMNIST dataset.



(a) Glow (b) Glow-Mix (c) AP-CDE

Figure D-1. Latent representations estimated by three models applied on the MNIST training set. For Glow and Glow-Mix models, since the latent are in high dimensions, the UMAP is used to reduce the dimensions to 2. Neither of them produces interpretable latent presentations, whereas AP-CDE model does the latent that can be easily separated into groups.

Figure D-1 plots the latent representations produced by Glow, Glow-Mix and AP-CDE. For AP-CDE, the latent $z^{(3)}_{1:2,1,1}$ is used as the $z_{P,i}$ for a better illustration in two dimensional reduction. For Glow and Glow-Mix, the UMAP (McInnes et al., 2018) is used to reduce the dimension and plot its output in 2D. For AP-CDE, the latent $z_{P,i}$ is provided. As expected, the latent variable produced by Glow follows a simple spherical Gaussian, and thus is not interpretable. The Glow-Mix does not produce a meaningful result either. Using AP-CDE and the supervising label

information form x_i, the authors obtain a good separation of the ten digits based on the low-dimensional representation $z_{P,i} \in \mathbb{R}^2$ (shown in Figure D-1(c)).

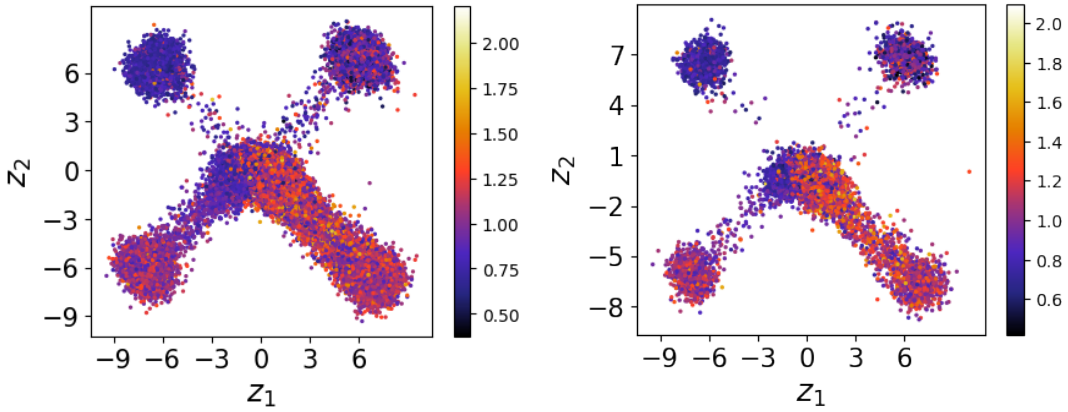
Under the working model using $z^{(3)}_{1:16,1,1}$ as $z_{P,i}$, for the testing set, the authors obtain $z_{P,i}$ and predict the label via the trained logistic regression model [using $\arg\max_{k\in\{0,\ldots,9\}} f_{x|z_P}(k \mid T^P_{\hat{\theta}}(y_i))$] to obtain classification accuracy of 98.5%. Empirically it is shown that the low-dimensional representation $z_{P,i}$ contains almost all the information to separate the different classes. As described in Section 3.1, the authors employ the ResNet101 to classify the artificially generated images from the proposed AP-CDE model with fixed $z_{P,i}$'s, and find that 96.14% of them are still classified to the same class label as the y_i's. Hence, it is concluded that the $z_{N,i}$ largely corresponds to within-class variation, whereas $z_{P,i}$ captures between-class variation.

Table D-1. The averages of the bits per dimension on the training and the testing sets of MNIST data.

Models	Glow	Glow-Mix	Glow-CDE	Glow-NCDE	AP-CDE
Training set	1.01	1.02	1.03	1.23	1.01
Testing set	1.01	1.03	1.04	1.28	1.01

As seen in Figure D-1(c), x_i tends to make each group of z_i's (each group corresponding to a digit) more concentrated to their respective center, which leads us to compare the marginal densities $f_y(y_i)$ with the ones from competing methods. Table D-1 depicts the average of the log-densities on the training and the testing sets for all models. Clearly, for this dataset, AP-CDE, likely due to a better group-wise concentration, produces overall higher or not lower marginal densities compared to its competitors. Note that bits per dimension is the negative log density over the number of dimensions.

Further, the latent $z_{P,i}$ is colored using the magnitude of the bits per dimension (Figure D-2). In the result, those points with relatively high BPM values (low density values) tend to correspond to images of low quality or higher ambiguity regarding the digit class. The easier classified classes have relatively higher densities. To show this, the authors plot a few sampled digits in Figure D-3 and sort each row by the density value in increasing order. It can be seen that the images on the left tend to be harder to assign to a class, compared to the ones on the right.



(a) Training set.

(b) Testing set.

Figure D-2. The first two dimensions of the latent variables from AP-CDE, colored by the estimated densities in the scale of bits per dimension.

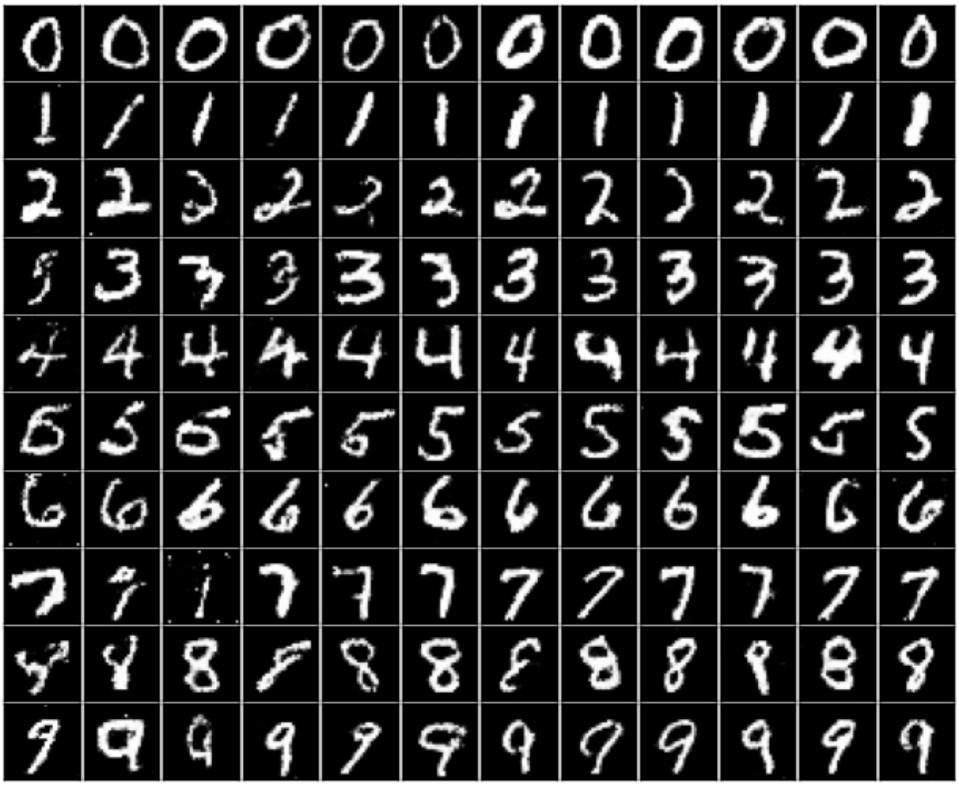


Figure D-3. Sample images from the MNIST data, with each row sorted in the increasing order of estimated densities.

LIST OF REFERENCES

Luca Ambrogioni, Umut Güçlü, Marcel AJ van Gerven, and Eric Maris. The kernel mixture network: A nonparametric method for conditional density estimation of continuous random variables. *arXiv preprint arXiv:1705.07111*, 2017.

Filippo Ascolani, Antonio Lijoi, Giovanni Rebaudo, and Giacomo Zanella. Clustering consistency with Dirichlet process mixtures. *Biometrika*, 110(2):551–558, 2023.

Christian Aßmann, Jens Boysen-Hogrefe, and Markus Pape. Bayesian analysis of static and dynamic factor models: An ex-post approach towards the rotation problem. *Journal of Econometrics*, 192(1):190–206, 2016.

Lachlan Astfalck, Deborshee Sen, Sayan Patra, Edward Cripps, and David Dunson. Posterior projection for inference in constrained spaces. *arXiv:1812.05741v5*, 2024.

Alejandra Avalos-Pacheco, David Rossell, and Richard S Savage. Heterogeneous large datasets integration using Bayesian factor regression. *Bayesian Analysis*, 17(1):33–66, 2022.

Sohail Bahmani and Justin Romberg. A flexible convex relaxation for phase retrieval. *Electronic Journal of Statistics*, 11(2):5254–5281, 2017.

Pau Batlle, Pratik Patil, Michael Stanley, Houman Owhadi, and Mikael Kuusela. Optimization-based frequentist confidence intervals for functionals in constrained inverse problems: Resolving the Burrus conjecture. *arXiv:2310.02461v3*, 2023.

Mokhtar S Bazaraa, John J Jarvis, and Hanif D Sherali. *Linear programming and network flows*. John Wiley & Sons, 2011.

Jens Behrmann, Will Grathwohl, Ricky TQ Chen, David Duvenaud, and Jörn-Henrik Jacobsen. Invertible residual networks. In *International Conference on Machine Learning*, pages 573–582. PMLR, 2019.

Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(6):1373–1396, 2003.

Alexandre Belloni and Victor Chernozhukov. On the computational complexity of MCMC-based estimators in large samples. *The Annals of Statistics*, 37(4):2011–2055, 2009.

James O Berger, Brunero Liseo, and Robert L Wolpert. Integrated likelihood methods for eliminating nuisance parameters. *Statistical Science*, 14(1):23–25, 1999.

Indrabati Bhattacharya and Ryan Martin. Gibbs posterior inference on multivariate quantiles. *Journal of Statistical Planning and Inference*, 218:106–121, 2022.

P. J. Bickel and B. J. K. Kleijn. The semiparametric Bernstein–von Mises theorem. *The Annals of Statistics*, 40(1):206–237, 2012.

Pier Giovanni Bissiri, Chris C Holmes, and Stephen G Walker. A general framework for updating belief distributions. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 78(5):1103–1130, 2016.

David Blackwell and James B MacQueen. Ferguson distributions via Pólya urn schemes. *The Annals of Statistics*, 1(2):353–355, 1973.

David M Blei, Andrew Y Ng, and Michael I Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

Zdravko I Botev, Joseph F Grotowski, and Dirk P Kroese. Kernel density estimation via diffusion. *The Annals of Statistics*, 38(5):2916–2957, 2010.

Diana Cai, Trevor Campbell, and Tamara Broderick. Finite mixture models do not reliably learn the number of components. In *International Conference on Machine Learning*, pages 1158–1169. PMLR, 2021.

Ismaël Castillo and Judith Rousseau. A Bernstein-–von Mises theorem for smooth functionals in semiparametric models. *The Annals of Statistics*, 43(6):2353–2383, 2015.

Ismaël Castillo, Johannes Schmidt-Hieber, and Aad Van der Vaart. Bayesian Linear Regression with Sparse Priors. *The Annals of Statistics*, 43(5):1986–2018, 2015.

Abhisek Chakraborty, Anirban Bhattacharya, and Debdeep Pati. A gibbs posterior framework for fair clustering. *Entropy*, 26(1):63, 2024.

Moumita Chakraborty and Subhashis Ghosal. Rates and coverage for monotone densities using projection-posterior. *Bernoulli*, 28(2):1093–1119, 2022.

Saptarshi Chakraborty and Jason Xu. Biconvex clustering. *Journal of Computational and Graphical Statistics*, 32(4):1524–1536, 2023.

Noirrit Kiran Chandra, Antonio Canale, and David B Dunson. Escaping the curse of dimensionality in Bayesian model-based clustering. *Journal of machine learning research*, 24 (144):1–42, 2023.

Noirrit Kiran Chandra, David B Dunson, and Jason Xu. Inferring covariance structure from multiple data sources via subspace factor analysis. *Journal of the American Statistical Association (in press)*, pages 1–15, 2024.

Chih-Chung Chang and Chih-Jen Lin. LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):1–27, 2011.

O. Chapelle, B. Scholkopf, and A. Zien. *Semi-supervised learning*. Adaptive Computation and Machine Learning series. MIT Press, 2010.

Ricky TQ Chen, Jens Behrmann, David K Duvenaud, and Jörn-Henrik Jacobsen. Residual flows for invertible generative modeling. *Advances in Neural Information Processing Systems*, 32, 2019.

Guang Cheng and Michael R Kosorok. Higher order semiparametric frequentist inference with the profile sampler. *The Annals of Statistics*, 36(4):1786–1818, 2008.

Guang Cheng and Michael R Kosorok. The penalized profile sampler. *Journal of Multivariate Analysis*, 100(3):345–362, 2009.

Eric C Chi and Kenneth Lange. Splitting methods for convex clustering. *Journal of Computational and Graphical Statistics*, 24(4):994–1013, 2015.

Hugh A Chipman, Edward I George, and Robert E McCulloch. Bart: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1):266–298, 2010.

Jon Cockayne, Chris J Oates, Timothy John Sullivan, and Mark Girolami. Bayesian probabilistic numerical methods. *SIAM Review*, 61(4):756–789, 2019.

R Dennis Cook and Liqiang Ni. Sufficient dimension reduction via inverse regression: A minimum discrepancy approach. *Journal of the American Statistical Association*, 100(470): 410–428, 2005.

Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20:273–297, 1995.

David L Davies and Donald W Bouldin. A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-1(2):224–227, 1979.

A Philip Dawid and Monica Musio. Bayesian model selection based on proper scoring rules. *Bayesian Analysis*, 10(2):479–499, 2015.

Roberta De Vito, Ruggero Bellio, Lorenzo Trippa, and Giovanni Parmigiani. Bayesian multistudy factor analysis for high-throughput biological data. *The Annals of Applied Statistics*, 15(4): 1723–1741, 2021.

Mingzhou Ding, Yonghong Chen, and Steven L Bressler. Granger causality: basic theory and application to neuroscience. In *Handbook of Time Series Analysis: Recent Theoretical Developments and Applications*, pages 437–460. John Wiley & Sons, 2006.

Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real NVP. *arXiv Preprint arXiv:1605.08803*, 2016.

Ian L Dryden and Kanti V Mardia. *Statistical shape analysis: with applications in R*. John Wiley & Sons, 2016.

Leo L Duan and David B Dunson. Bayesian distance clustering. *The Journal of Machine Learning Research*, 22(224):1–27, 2021.

Leo L Duan, Alexander L Young, Akihiko Nishimura, and David B Dunson. Bayesian constraint relaxation. *Biometrika*, 107(1):191–204, 2020.

David B Dunson and Ju-Hyun Park. Kernel stick-breaking processes. *Biometrika*, 95(2):307–323, 2008.

David B Dunson and Jack A Taylor. Approximate Bayesian inference for quantiles. *Journal of Nonparametric Statistics*, 17(3):385–400, 2005.

Michael Evans. Measuring statistical evidence using relative belief. *Computational and Structural Biotechnology Journal*, 14:91–96, 2016.

Chris Fraley and Adrian E Raftery. Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, 97(458):611–631, 2002.

David T Frazier and Christopher Drovandi. Robust approximate Bayesian inference with synthetic likelihood. *Journal of Computational and Graphical Statistics*, 30(4):958–976, 2021.

David T Frazier, David J Nott, Christopher Drovandi, and Robert Kohn. Bayesian inference using synthetic likelihood: asymptotics and adjustments. *Journal of the American Statistical Association*, 118(544):2821–2832, 2023.

Ashutosh Garg and Dan Roth. Understanding probabilistic classifiers. In *European Conference on Machine Learning*, pages 179–191. Springer, 2001.

Alan E Gelfand. Gibbs sampling. *Journal of the American Statistical Association*, 95(452): 1300–1304, 2000.

Alan E Gelfand, Adrian FM Smith, and Tai-Ming Lee. Bayesian analysis of constrained parameter and truncated data problems using Gibbs sampling. *Journal of the American Statistical Association*, 87(418):523–532, 1992.

Andrew Gelman and Donald B Rubin. Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4):457–472, 1992.

Mathieu Germain, Karol Gregor, Iain Murray, and Hugo Larochelle. Made: Masked autoencoder for distribution estimation. In *International Conference on Machine Learning*, pages 881–889. PMLR, 2015.

John Geweke and Michael Keane. Smoothly mixing regressions. *Journal of Econometrics*, 138 (1):252–290, 2007.

Satyajit Ghosh, Kshitij Khare, and George Michailidis. Strong selection consistency of Bayesian vector autoregressive models based on a pseudo-likelihood approach. *The Annals of Statistics*, 49(3):1267–1299, 2021.

Mark Girolami and Ben Calderhead. Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 73(2): 123–214, 2011.

Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

Colin Goodall. Procrustes methods in the statistical analysis of shape. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 53(2):285–321, 1991.

John C Gower. Generalized Procrustes analysis. *Psychometrika*, 40(1):33–51, 1975.

Will Grathwohl, Ricky TQ Chen, Jesse Bettencourt, Ilya Sutskever, and David Duvenaud. Ffjord: Free-form continuous dynamics for scalable reversible generative models. *arXiv Preprint arXiv:1810.01367*, 2018.

Peter Grünwald and Thijs Van Ommen. Inconsistency of Bayesian inference for misspecified linear models, and a proposal for repairing it. *Bayesian Analysis*, 12(4):1069–1103, 2017.

Benjamin D Haeffele, Chong You, and René Vidal. A critique of self-expressive deep subspace clustering. *arXiv Preprint arXiv:2010.03697*, 2020.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Matthew Heiner, Athanasios Kottas, and Stephan Munch. Structured priors for sparse probability vectors with application to model selection in markov chains. *Statistics and Computing*, 29(5): 1077–1093, 2019.

Qiang Heng, Hua Zhou, and Eric C Chi. Bayesian trend filtering via proximal Markov chain Monte Carlo. *Journal of Computational and Graphical Statistics*, 32(3):938–949, 2023.

Peter Hoff and Alexander Franks. *rstiefel: Random orthonormal matrix generation and optimization on the Stiefel manifold*, 2021. URL https://CRAN.R-project.org/package=rstiefel. R package version 1.0.1.

Peter D. Hoff. Simulation of the matrix Bingham–von Mises–Fisher distribution, with applications to multivariate and relational data. *Journal of Computational and Graphical Statistics*, 18(2):438–456, 2009.

Matthew D Hoffman, Andrew Gelman, et al. The No-U-Turn Sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15(1): 1593–1623, 2014.

Ziyi Huang, Henry Lam, and Haofeng Zhang. Evaluating aleatoric uncertainty via conditional generative models. *arXiv Preprint arXiv:2206.04287*, 2022.

Hemant Ishwaran and Lancelot F. James. Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, 96(453):161–173, 2001.

Pavel Izmailov, Polina Kirichenko, Marc Finzi, and Andrew Gordon Wilson. Semi-supervised learning with normalizing flows. In *International Conference on Machine Learning*, pages 4615–4630. PMLR, 2020.

Pierre E Jacob, Lawrence M Murray, Chris C Holmes, and Christian P Robert. Better together? Statistical learning in models made of modules. *arXiv:1708.08719*, 2017.

Ali Jadbabaie, Alexander Rakhlin, Shahin Shahrampour, and Karthik Sridharan. Online optimization: Competing with dynamic comparators. In *International Conference on Artificial Intelligence and Statistics*, pages 398–406. PMLR, 2015.

Sonia Jain and Radford M. Neal. Splitting and merging components of a nonconjugate Dirichlet process mixture model. *Bayesian Analysis*, 2(3):445–472, 2007.

Jens Ledet Jensen and Hans R Künsch. On asymptotic normality of pseudo likelihood estimates for pairwise interaction processes. *Annals of the Institute of Statistical Mathematics*, 46(3): 475–486, 1994.

Wenxin Jiang and Martin A Tanner. Hierarchical mixtures-of-experts for exponential family regression models: Approximation and maximum likelihood estimation. *The Annals of Statistics*, 27(3):987–1011, 1999.

Wenxin Jiang and Martin A Tanner. Gibbs posterior for variable selection in high-dimensional classification and data mining. *The Annals of Statistics*, 36(5):2207–2231, 2008.

James E Johndrow, Aaron Smith, Natesh Pillai, and David B Dunson. MCMC for imbalanced categorical data. *Journal of the American Statistical Association*, 114(527):1394–1403, 2019.

Maria Kalli, Jim E Griffin, and Stephen G Walker. Slice sampling mixture models. *Statistics and Computing*, 21(1):93–105, 2011.

Kshitij Khare, Sang-Yun Oh, and Bala Rajaratnam. A convex pseudolikelihood framework for high dimensional partial correlation estimation with convergence guarantees. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 77(4):803–825, 2015.

JooSeuk Kim and Clayton D Scott. Robust kernel density estimation. *Journal of Machine Learning Research*, 13(1):2529–2565, 2012.

Youngseok Kim and Chao Gao. Bayesian model selection with graph structured sparsity. *The Journal of Machine Learning Research*, 21(1):4394–4454, 2020.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv Preprint arXiv:1412.6980*, 2014.

Durk P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. *Advances in Neural Information Processing Systems*, 31, 2018.

Michael R Kosorok. *Introduction to empirical processes and semiparametric inference*, volume 61. Springer, 2008.

Yi-Hung Kung, Pei-Sheng Lin, and Cheng-Hsiung Kao. An optimal k-nearest neighbor for density estimation. *Statistics & Probability Letters*, 82(10):1786–1791, 2012.

Shiwei Lan, Vasileios Stathopoulos, Babak Shahbaba, and Mark Girolami. Markov chain Monte Carlo from Lagrangian dynamics. *Journal of Computational and Graphical Statistics*, 24(2): 357–378, 2015.

Alfonso Landeros, Jason Xu, and Kenneth Lange. MM optimization: Proximal distance algorithms, path following, and trust regions. *Proceedings of the National Academy of Sciences*, 120(27):e2303168120, 2023.

Isaac Lavine, Michael Lindon, and Mike West. Adaptive variable selection for sequential prediction in multivariate dynamic models. *Bayesian Analysis*, 16(4):1059–1083, 2021.

Neil Lawrence. Gaussian process latent variable models for visualisation of high dimensional data. *Advances in Neural Information Processing Systems*, 16, 2003.

Bee Leng Lee, Michael R Kosorok, and Jason P Fine. The profile sampler. *Journal of the American Statistical Association*, 100(471):960–969, 2005a.

Kuang-Chih Lee, Jeffrey Ho, and David J Kriegman. Acquiring linear subspaces for face recognition under variable lighting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(5):684–698, 2005b.

Kwangmin Lee, Kyoungjae Lee, and Jaeyong Lee. Post-processed posteriors for banded covariances. *Bayesian Analysis*, 18(3):1017–1040, 2023.

John R Lewis, Steven N MacEachern, and Yoonkyung Lee. Bayesian restricted likelihood methods: Conditioning on insufficient statistics in Bayesian regression (with discussion). *Bayesian Analysis*, 16(4):1393–1462, 2021.

Feng Liang, Sayan Mukherjee, and Mike West. The use of unlabeled data in predictive modeling. *Statistical Science*, 22(2):189–205, 2007.

Lek-Heng Lim, Rodolphe Sepulchre, and Ke Ye. Geometric distance between positive definite matrices of different dimensions. *IEEE Transactions on Information Theory*, 65(9):5401–5405, 2019.

Bruce G Lindsay. Composite likelihood methods. *Contemporary Mathematics*, 80(1):221–239, 1988.

Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.

Rong Ma, Eric D Sun, David Donoho, and James Zou. Principled and interpretable alignability testing and integration of single-cell data. *Proceedings of the National Academy of Sciences*, 121(10):e2313719121, 2024.

YP Mack and Murray Rosenblatt. Multivariate k-nearest neighbor density estimates. *Journal of Multivariate Analysis*, 9(1):1–15, 1979.

Oliver J Maclaren. Is profile likelihood a true likelihood? An argument in favor. *arXiv:1801.04369v4*, 2018.

James MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297, 1967.

Daniel Marcus, John Harwell, Timothy Olsen, Michael Hodge, Matthew Glasser, Fred Prior, Mark Jenkinson, Timothy Laumann, Sandra Curtiss, and David Van Essen. Informatics and data mining tools and strategies for the human connectome project. *Frontiers in Neuroinformatics*, 5:4, 2011.

Ryan Martin and Nicholas Syring. Direct Gibbs posterior inference on risk minimizers: Construction, concentration, and calibration. In *Handbook of Statistics*, volume 47, pages 1–41. Elsevier, 2022.

Peter McCullagh. Exponential mixtures and quadratic exponential families. *Biometrika*, 81(4): 721–729, 1994.

Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv Preprint arXiv:1802.03426*, 2018.

Geoffrey McLachlan and David Peel. Mixtures of factor analyzers. In *International Conference on Machine Learning*. Citeseer, 2000.

Geoffrey J McLachlan, David Peel, and Richard W Bean. Modelling high-dimensional data by mixtures of factor analyzers. *Computational Statistics & Data Analysis*, 41(3-4):379–388, 2003.

Jeffrey W Miller. An elementary derivation of the chinese restaurant process from Sethuraman's stick-breaking process. *Statistics & Probability Letters*, 146:112–117, 2019.

Jeffrey W Miller. Asymptotic normality, concentration, and coverage of generalized posteriors. *Journal of Machine Learning Research*, 22(168):1–53, 2021.

Jeffrey W Miller and Matthew T Harrison. A simple example of Dirichlet process mixture inconsistency for the number of components. *Advances in neural information processing systems*, 26, 2013.

Jeffrey W. Miller and Matthew T. Harrison. Inconsistency of Pitman-Yor process mixtures for the number of components. *Journal of Machine Learning Research*, 15(96):3333–3370, 2014.

Jeffrey W. Miller and Matthew T. Harrison. Mixture models with a prior on the number of components. *Journal of the American Statistical Association*, 113(521):340–356, 2018.

Susan A Murphy and Aad W Van der Vaart. On profile likelihood. *Journal of the American Statistical Association*, 95(450):449–465, 2000.

Eric Nalisnick, Akihiro Matsukawa, Yee Whye Teh, Dilan Gorur, and Balaji Lakshminarayanan. Hybrid models with deep and invertible features. In *International Conference on Machine Learning*, pages 4723–4732. PMLR, 2019.

Abhinav Natarajan, Maria De Iorio, Andreas Heinecke, Emanuel Mayer, and Simon Glenn. Cohesion and repulsion in Bayesian distance clustering. *Journal of the American Statistical Association*, pages 1–11, 2023.

Abhinav Natarajan, Maria De Iorio, Andreas Heinecke, Emanuel Mayer, and Simon Glenn. Cohesion and repulsion in Bayesian distance clustering. *Journal of the American Statistical Association*, 119(546):1374–1384, 2024.

Radford M Neal. MCMC using Hamiltonian dynamics. In *Handbook of Markov Chain Monte Carlo*, pages 113–162. Chapman and Hall/CRC, 2011.

Akihiko Nishimura and David Dunson. Geometrically tempered Hamiltonian Monte Carlo. *arXiv:1604.00872v2*, 2017.

Agostino Nobile. *Bayesian Analysis of Finite Mixture Distributions*. PhD thesis, PhD Thesis. Carnegie Mellon University, Pittsburgh, 1994.

Andriy Norets. Approximation of conditional densities by smooth mixtures of regressions. *The Annals of Statistics*, 38(3):1733–1766, 2010.

David J Nott, Christopher Drovandi, and David T Frazier. Bayesian inference for misspecified generative models. *Annual Review of Statistics and Its Application*, 11, 2023.

Ilsang Ohn and Lizhen Lin. Optimal Bayesian estimation of gaussian mixtures with growing number of components. *Bernoulli*, 29(2):1195–1218, 2023.

George Papamakarios, Theo Pavlakou, and Iain Murray. Masked autoregressive flow for density estimation. *Advances in Neural Information Processing Systems*, 30, 2017.

George Papamakarios, Eric T Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *Journal of Machine Learning Research*, 22(57):1–64, 2021.

Trevor Park and George Casella. The Bayesian lasso. *Journal of the American Statistical Association*, 103(482):681–686, 2008.

Xi Peng, Jiashi Feng, Joey Tianyi Zhou, Yingjie Lei, and Shuicheng Yan. Deep subspace clustering. *IEEE Transactions on Neural Networks and Learning Systems*, 31(12):5509–5521, 2020.

Jim Pitman. Exchangeable and partially exchangeable random partitions. *Probability Theory and Related Fields*, 102(2):145–158, 1995.

Jim Pitman and Marc Yor. The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *Annals of Probability*, 25(2):855–900, 1997.

Nicholas G Polson and James G Scott. Mixtures, envelopes and hierarchical duality. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 78(4):701–727, 2016.

Nicholas G Polson and Steven L Scott. Data augmentation for support vector machines. *Bayesian Analysis*, 6(1):1–23, 2011.

Nicholas G Polson, James G Scott, and Jesse Windle. Bayesian inference for logistic models using Pólya–Gamma latent variables. *Journal of the American Statistical Association*, 108(504): 1339–1349, 2013.

Nicholas G Polson, James G Scott, and Brandon T Willard. Proximal algorithms in statistics and machine learning. *Statistical Science*, 30(4):559–581, 2015.

Rick Presman and Jason Xu. Distance-to-set priors and constrained Bayesian inference. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pages 2310–2326. PMLR, 2023.

Lu Ren, Lan Du, Lawrence Carin, and David B Dunson. Logistic stick-breaking process. *Journal of Machine Learning Research*, 12(1), 2011.

Sylvia Richardson and Peter 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.

Tommaso Rigon, Amy H Herring, and David B Dunson. A generalized Bayes framework for probabilistic clustering. *Biometrika*, 110(3):559–578, 2023a.

Tommaso Rigon, Amy H Herring, and David B Dunson. A generalized Bayes framework for probabilistic clustering. *Biometrika*, 110(3):559–578, 2023b.

Gareth O Roberts and Jeffrey S Rosenthal. Optimal scaling for various Metropolis-Hastings algorithms. *Statistical Science*, 16(4):351–367, 2001.

R Tyrrell Rockafellar and Roger J-B Wets. *Variational analysis*, volume 317. Springer Science & Business Media, 2009.

Abel Rodríguez, David B Dunson, and Alan E Gelfand. Latent stick-breaking processes. *Journal of the American Statistical Association*, 105(490):647–659, 2010.

Karl Rohe and Muzhe Zeng. Vintage factor analysis with Varimax performs statistical inference. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(4):1037–1060, 2023.

Karl Rohe, Sourav Chatterjee, and Bin Yu. Spectral clustering and the high-dimensional stochastic blockmodel. *The Annals of Statistics*, 39(4):1878–1915, 2011.

Jonas Rothfuss, Fabio Ferreira, Simon Walther, and Maxim Ulrich. Conditional density estimation with neural networks: Best practices and benchmarks. *arXiv Preprint arXiv:1903.00954*, 2019.

Sam T Roweis and Lawrence K Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000.

Arkaprava Roy, Isaac Lavine, Amy H Herring, and David B Dunson. Perturbed factor analysis: Accounting for group differences in exposure profiles. *The Annals of Applied Statistics*, 15(3): 1386–1404, 2021.

Håvard Rue, Sara Martino, and Nicolas Chopin. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 71(2):319–392, 2009.

Takumi Saikawa, Quan Huu Cap, Satoshi Kagiwada, Hiroyuki Uga, and Hitoshi Iyatomi. Aop: An anti-overfitting pretreatment for practical image-based plant diagnosis. In *2019 IEEE International Conference on Big Data (Big Data)*, pages 5177–5182. IEEE, 2019.

John W Sammon. A nonlinear mapping for data structure analysis. *IEEE Transactions on Computers*, 100(5):401–409, 1969.

Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller. Kernel principal component analysis. In *International Conference on Artificial Neural Networks*, pages 583–588. Springer, 1997.

Erich Schubert, Arthur Zimek, and Hans-Peter Kriegel. Generalized outlier detection with flexible kernel density estimates. In *Proceedings of the 2014 SIAM International Conference on Data Mining*, pages 542–550. SIAM, 2014.

David W Scott. Partial mixture estimation and outlier detection in data and regression. In *Theory and Applications of Recent Robust Methods*, pages 297–306. Springer, 2004.

Jayaram Sethuraman. A constructive definition of the Dirichlet prior. *Statistica Sinica*, 4: 639–650, 1994.

Thomas A Severini. Integrated likelihood functions for non-Bayesian inference. *Biometrika*, 94 (3):529–542, 2007.

Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1):1–48, 2019.

Debajyoti Sinha, Joseph G Ibrahim, and Ming-Hui Chen. A Bayesian justification of Cox's partial likelihood. *Biometrika*, 90(3):629–641, 2003.

Marlon Stoeckius, Christoph Hafemeister, William Stephenson, Brian Houck-Loomis, Pratip K Chattopadhyay, Harold Swerdlow, Rahul Satija, and Peter Smibert. Simultaneous epitope and transcriptome measurement in single cells. *Nature Methods*, 14(9):865–868, 2017.

Liangjun Su and Halbert White. A consistent characteristic function-based test for conditional independence. *Journal of Econometrics*, 141(2):807–834, 2007.

Liangjun Su and Halbert White. A nonparametric hellinger metric test for conditional independence. *Econometric Theory*, 24(4):829–864, 2008.

Nicholas Syring and Ryan Martin. Gibbs posterior concentration rates under sub-exponential type losses. *Bernoulli*, 29(2):1080–1108, 2023.

Emily Tallman and Mike West. Bayesian predictive decision synthesis. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(2):340–363, 2024.

Kean Ming Tan and Daniela Witten. Statistical properties of convex clustering. *Electronic Journal of Statistics*, 9(2):2324, 2015.

Rong Tang and Yun Yang. On the computational complexity of Metropolis-adjusted Langevin algorithms for Bayesian posterior sampling. *arXiv preprint arXiv:2206.06491*, 2022.

Yichuan Tang, Ruslan Salakhutdinov, and Geoffrey Hinton. Deep mixtures of factor analysers. *arXiv Preprint arXiv:1206.4635*, 2012.

Martin A Tanner and Wing Hung Wong. The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, 82(398):528–540, 1987.

George R Terrell and David W Scott. Variable kernel density estimation. *The Annals of Statistics*, pages 1236–1265, 1992.

Michael E Tipping and Christopher M Bishop. Mixtures of probabilistic principal component analyzers. *Neural Computation*, 11(2):443–482, 1999.

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(11), 2008.

David A Van Dyk and Xiao-Li Meng. The art of data augmentation. *Journal of Computational and Graphical Statistics*, 10(1):1–50, 2001.

Cristiano Varin, Nancy Reid, and David Firth. An overview of composite likelihood methods. *Statistica Sinica*, 21:5–42, 2011.

Mattias Villani, Robert Kohn, and Paolo Giordani. Regression density estimation using smooth adaptive gaussian mixtures. *Journal of Econometrics*, 153(2):155–173, 2009.

Stephen G Walker. Sampling the Dirichlet mixture model with slices. *Communications in Statistics—Simulation and Computation*, 36(1):45–54, 2007.

Mike West. Perspectives on constrained forecasting. *Bayesian Analysis*, 19(4):1013–1039, 2024.

Steven Winter, Omar Melikechi, and David B Dunson. Sequential Gibbs posteriors with applications to principal component analysis. *arXiv:2310.12882*, 2023.

Russ Wolfinger. Covariance structure selection in general mixed models. *Communications in Statistics-Simulation and Computation*, 22(4):1079–1106, 1993.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-MNIST: A novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Jason Xu and Kenneth Lange. Power k-means clustering. In *International Conference on Machine Learning*, pages 6921–6931. PMLR, 2019.

Chiao-Yu Yang, Eric Xia, Nhat Ho, and Michael I Jordan. Posterior distribution for the number of clusters in Dirichlet process mixture models. *arXiv preprint arXiv:1905.09959*, 2020a.

Jun Yang, Gareth O Roberts, and Jeffrey S Rosenthal. Optimal scaling of random-walk Metropolis algorithms on general target distributions. *Stochastic Processes and Their Applications*, 130(10):6094–6132, 2020b.

Cheng Zeng, Jeffrey W Miller, and Leo L Duan. Consistent model-based clustering using the quasi-Bernoulli stick-breaking process. *Journal of Machine Learning Research*, 24(153):1–32, 2023.

Cheng Zeng, Eleni Dilma, Jason Xu, and Leo L Duan. The bridged posterior: optimization, profile likelihood and a new approach to generalized Bayes. *arXiv:2403.00968*, 2024a.

Cheng Zeng, George Michailidis, Hitoshi Iyatomi, and Leo L Duan. Normalizing flow to augmented posterior: Conditional density estimation with interpretable dimension reduction for high dimensional data. *International Journal of Computer and Information Engineering*, 18(5): 306–315, 2024b.

Zhiyue Zhang, Kenneth Lange, and Jason Xu. Simple and scalable sparse k-means clustering via feature ranking. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, pages 10148–10160, 2020.

Peng Zheng and Aleksandr Aravkin. Relax-and-split method for nonconvex inverse problems. *Inverse Problems*, 36(9):095013, 2020.

Xinkai Zhou, Qiang Heng, Eric C Chi, and Hua Zhou. Proximal MCMC for Bayesian inference of constrained and regularized estimation. *The American Statistician*, 78(4):379–390, 2024.

Konstantin M Zuev, James L Beck, and Lambros S Katafygiotis. On the optimal scaling of the modified metropolis-hastings algorithm. In *Proceedings of the 11th International Conference on Applications of Statistics and Probability in Civil Engineering*, 2011.

BIOGRAPHICAL SKETCH

Zeng Cheng was born in Zhuzhou, Hunan, China in 1995. He began his undergraduate studies at the University of Science and Technology of China in 2013, and received his B.S. degree in Mathematics and Applied Mathematics in 2017. In 2018, He joined the Department of Statistics at the University of Florida as a graduate student and successfully completed his Ph.D. in Statistics in 2025.