Bayesian Computation for High-Dimensional Continuous & Sparse Count Data

by

Ye Wang

Department of Statistical Science
Duke University

Date: ____________________

Approved:

____
David B. Dunson, Supervisor

____
Colin Rundel

____
Rebecca C. Steorts

____
Rong Ge

Dissertation submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in the Department of Statistical Science
in the Graduate School of Duke University
2018

Abstract

Bayesian Computation for High-Dimensional Continuous & Sparse Count Data

by

Ye Wang

Department of Statistical Science
Duke University

Date: ______________________
Approved:

____
David B. Dunson, Supervisor

____
Colin Rundel

____
Rebecca C. Steorts

____
Rong Ge

An abstract of a dissertation submitted in partial fulfillment of the requirements for
the degree of Doctor of Philosophy in the Department of Statistical Science
in the Graduate School of Duke University
2018

Abstract

Probabilistic modeling of multidimensional data is a common problem in practice. When data are continuous, one common approach is to suppose that the observed data are close to a lower-dimensional smooth manifold. There are a rich variety of manifold learning methods available, which allow mapping of data points to the manifold. However, there is a clear lack of probabilistic methods that allow learning of the manifold along with the generative distribution of the observed data. The best attempt is the Gaussian process latent variable model (GP-LVM), but identifiability issues lead to poor performance. We solve these issues by proposing a novel Coulomb repulsive process (Corp) for locations of points on the manifold, inspired by physical models of electrostatic interactions among particles. Combining this process with a GP prior for the mapping function yields a novel electrostatic GP (electroGP) process.

Another popular approach is to suppose that the observed data are closed to one or a union of lower-dimensional linear subspaces. However, popular methods such as probabilistic principal component analysis scale poorly computationally. We introduce a novel empirical Bayesian method that we term geometric density estimation (GEODE), which assumes the data is centered near a low-dimensional linear subspace. We show that, with mild assumptions on the prior, the subspace spanned by the principal axes of the data maximizes the posterior mode. Hence, leveraged on the geometric information of the data, GEODE easily scales to massive dimensional problems. It is also capable of learning the intrinsic dimension via a novel shrinkage prior. Finally we mix GEODE across a dyadic clustering tree to account for nonlinear cases.

When data are discrete, a common strategy is to define a generalized linear model (GLM) for each variable, with dependence in the different variables induced through including multivariate latent variables in the GLMs. The Bayesian inference for these models usually rely on data augmented Markov chain Monte Carlo (DA-MCMC) method, which has a provable slow mixing rate when the data is imbalanced. For more scalable inference, we proposes *Bayesian mosaic*, a parallelizable composite posterior, for scalable Bayesian inference on a broad class of multivariate discrete data models. Sampling is embarrassingly parallel since *Bayesian mosaic* is a multiplication of component posteriors that can be independently sampled from. Analogous to composite likelihood methods, these component posteriors are based on univariate or bivariate marginal densities. Utilizing the fact that the score functions of these densities are unbiased, we show that *Bayesian mosaic* is consistent and asymptotically normal under mild conditions. Since the evaluation of univariate or bivariate marginal densities can rely on numerical integration, sampling from *Bayesian mosaic* bypasses the traditional data augmented Markov chain Monte Carlo (DA-MCMC) method, which has a provably slow mixing rate when data are imbalanced. Moreover, we show that sampling from *Bayesian mosaic* has better scalability to large sample size than DA-MCMC. The performance of the proposed methods and models will be demonstrated via both simulation studies and real world applications.

To my family.

Contents

List of Tables

List of Figures

List of Abbreviations and Symbols

Symbols

$\mathbb{R}$	Set of all real numbers
$\mathbb{N}_0$	Set of all natural numbers
$\mathbb{N}_1$	Set of all non-negative integers
$\mathbb{E}$	The expectation operator
P	The probability symbo
∇	The gradient operator

Abbreviations

MCMC	Markov chaine Monte Carlo
DA-MCMC	Data-augmented MCMC
MH	Metropolis Hastings
MFA	Mixtures of factor analyzers (Chen et al., 2010)
GP-LVM	Gaussian process latent variable model (Lawrence, 2005)

Acknowledgements

During the past six years here at Duke, I fortunately met a lot of incredible people, making my graduate student life an invaluable memory. I would like to express my deep gratefulness to all these people. First of all, I want to thank my advisor, Dr. David Dunson, for being a great mentor. David is a great statistician with superior creativeness and excellent insights. I learned a lot from him during my Ph.D. study. He has been really supportive of my designing my own research projects. This helped me find my true research interests and taught me how to work efficiently independently.

I thank Dr. Charles Maxwell Becker and Dr. Jerry Reiter for all the supports and help during my master study. They were my advisors for the master's degree in statistical and economical modeling. They both have been really supportive of my pursuing a PhD degree and given me many good advices.

I would also like to thank my other PhD thesis and Master thesis committee members, Dr. Colin Rundel, Dr. Rebecca Steorts, Dr. Rong Ge and Dr. Surya Tokdar for all your constructive suggestions and discussions on my research. I want to thank Dr. Mike West for all your advices while I was applying to this PhD program. I also thank Dr. Youngdeok Hwang and Dr. Yasuo Amemiya for the guidance on my summer project at IBM Watson Research center, Dr. Andrew Clayton for my summer project at Amazon and Tobias Wooldridge for my summer project at Facebook.

I would like to thank everyone in the department for their invaluable help and all my friends for making my Ph.D. life cheerful. Finally, I want to express my deepest thankfulness to my parents. I thank everything they have done for me.

1

Introduction

High-dimensional data are ubiquitously common in real-world applications. High dimensional continous data examples include images, sound signals and sensor data while discrete data examples include text, advertisement click data and network relationship data, just naming a few. This paper proposes two new models for learning the density of high dimensional continuous data and a novel composite posterior method for scalable Bayesian inference on high dimensional discrete data.

When data are continuous, there is broad interest in learning and exploiting the lower-dimensional structure. A canonical case is when the low dimensional structure corresponds to a p-dimensional smooth Riemannian manifold $\mathcal{M}$ embedded in the d-dimensional ambient space $\mathcal{Y}$ of the observed data $\boldsymbol{y}$. Assuming that the observed data are close to $\mathcal{M}$, it becomes of substantial interest to learn $\mathcal{M}$ along with the mapping μ from $\mathcal{M} \to \mathcal{Y}$. This allows better data visualization and for one to exploit the lower-dimensional structure to combat the curse of dimensionality in developing efficient algorithms for a variety of tasks.

The current literature on *manifold learning* focuses on estimating the coordinates $\boldsymbol{x} \in \mathcal{M}$ corresponding to $\boldsymbol{y}$ by optimization, finding $\boldsymbol{x}$'s on the manifold $\mathcal{M}$ that preserve distances between the corresponding $\boldsymbol{y}$'s in $\mathcal{Y}$. There are many such methods, including Isomap (Tenenbaum et al., 2000), locally-linear embedding (Roweis and Saul, 2000) and Laplacian eigenmaps (Belkin and Niyogi, 2002). Such methods have seen broad use, but have some clear limitations relative to *probabilistic manifold learning* approaches, which allow explicit learning of $\mathcal{M}$, the mapping μ and the distribution of $\boldsymbol{y}$.

There has been some considerable focus on probabilistic models, which would seem to allow learning of $\mathcal{M}$ and μ. Two notable examples are mixtures of factor analyzers (Chen et al., 2010; Wang et al., 2014) and Gaussian process latent variable models (Lawrence, 2005). Such approaches are useful in exploiting lower-dimensional structure in estimating the distribution of $\boldsymbol{y}$, but unfortunately have critical problems in terms of reliable estimation of the manifold and mapping function. MFA is not smooth in approximating the manifold with a collage of lower dimensional hyperplanes, and hence we focus further discussion on GP-LVM. Similar problems occur for MFA and other probabilistic manifold learning methods.

In general form for the ith data vector, GP-LVM lets $\boldsymbol{y}_i = \mu(\boldsymbol{x}_i) + \boldsymbol{\epsilon}_i$, with μ assigned a Gaussian process prior, $\boldsymbol{x}_i$ generated from a pre-specified Gaussian or uniform distribution over a p-dimensional space, and the residual $\boldsymbol{\epsilon}_i$ drawn from a d-dimensional Gaussian centered on zero with diagonal or spherical covariance. While this model seems appropriate to manifold learning, identifiability problems lead to extremely poor performance in estimating $\mathcal{M}$ and μ. To give an intuition for the root cause of the problem, consider the case in which $\boldsymbol{x}_i$ are drawn independently from a uniform distribution over $[0,1]^p$. The model is so flexible that we could fit the training data $\boldsymbol{y}_i$, for $i = 1, \ldots, n$, just as well if we did not use the entire hypercube but just placed all the $\boldsymbol{x}_i$ values in a small subset of $[0,1]^p$. The uniform prior will not discourage this tendency to not spread out the latent coordinates, which unfortunately has disasterous consequences illustrated in our experiments.

The structure of the model is just too flexible, and further constraints are needed. Replacing the uniform with a standard Gaussian does not solve the problem.

To make the problem more tractable, we focus on the case in which $\mathcal{M}$ is a one-dimensional smooth compact manifold. Assume $\boldsymbol{y}_i = \boldsymbol{\mu}(x_i) + \boldsymbol{\epsilon}_i$, with $\boldsymbol{\epsilon}_i$ Gaussian noise, and $\boldsymbol{\mu} : (0, 1) \mapsto \mathcal{M}$ a smooth mapping such that $\mu_j(\cdot) \in C^\infty$ for $j = 1, \ldots, d$, where $\boldsymbol{\mu}(x) = (\mu_1(x), \ldots, \mu_d(x))$. We focus on finding a good estimate of $\boldsymbol{\mu}$, and hence the manifold, via a probabilistic learning framework. We refer to this problem as probabilistic curve learning (PCL) motivated by the principal curve literature (Hastie and Stuetzle, 1989). PCL differs substantially from the principal curve learning problem, which seeks to estimate a non-linear curve through the data, which may be very different from the true manifold.

Our first proposed approach builds on GP-LVM; in particular, our primary innovation is to generate the latent coordinates $\boldsymbol{x}_i$ from a novel repulsive process. There is an interesting literature on repulsive point process modeling ranging from various Matern processes (Rao et al., 2017) to the determinantal point process (DPP) (Hough et al., 2009). In our very different context, these processes lead to unnecessary complexity — computationally and otherwise — and we propose a new *Coulomb repulsive process* (Corp) motivated by Coulomb's law of electrostatic interaction between electrically charged particles. Using Corp for the latent positions has the effect of strongly favoring spread out locations on the manifold, effectively solving the identifiability problem mentioned above for the GP-LVM. We refer to the GP with Corp on the latent positions as an electrostatic GP (electroGP).

Assuming the intrinsic dimension is one dimensional can be too restrictive for many applications. Let $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{iD})^\top$, for $i = 1, \ldots, N$, be a sample from an unknown distribution having support in a subset of $\Re^D$. We are interested in estimating its density when D is large, and the data have a low-dimensional structure with intrinsic dimension p such that $p \ll D$. Kernel methods work well in low dimensions, but face challenges in scaling up to large D settings. Moreover, careful tuning of bandwidth is needed since the choice of bandwidth fundamentally impacts performance (Liu et al., 2007). Bayesian nonparametric models (Escobar and West, 1995; Rasmussen, 1999) provide an alternative approach for density estimation, specifying priors for the bandwidth parameters allowing adaptive estimation without cross-validation (Shen et al., 2013). However, inference is prohibitively costly.

To combat the curse of dimensionality, it is popular to assume that the data concentrates near a low-dimensional linear subspace. Principal component analysis (PCA) is a ubiquitous technique building upon such assumption. The p principal axes $\boldsymbol{w}_j$, $j \in \{1, \ldots, p\}$, are a set of orthonormal axes whose span captures the maximal amount of variability. It can be shown that these axes are given by the p leading right singular vectors of the demeaned observation matrix $\boldsymbol{Y}$, where $\boldsymbol{Y}$ is a $N \times D$ matrix with each row being a demeaned data vector. Tipping and Bishop (1999b) generalized PCA within a density estimation framework and introduced the probabilistic PCA (PPCA). PPCA can be viewed as a special case of the factor analyzer model (FA) which does not assume isotropic error. Carvalho et al. (2008) and Bhattacharya and Dunson (2011) (among many others) have successfully applied FA under the Bayesian paradigm while additionally assuming sparsity. However, FA involves complex computation that does not scale well, hence is not considered here. The PPCA model can be written as

$$\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{C}), \tag{1.1}$$

where $\boldsymbol{C} = \boldsymbol{W}\boldsymbol{\Sigma}\boldsymbol{W}^\top + \sigma^2\boldsymbol{I}$, $\boldsymbol{W}$ is a orthogonal $D \times p$ matrix and $\boldsymbol{\Sigma} = \operatorname{diag}(\alpha_1^2, \ldots, \alpha_p^2)$ is a diagonal $p \times p$ matrix with positive diagonal elements. It can be shown that the principal axes $\boldsymbol{w}_j$ give the column vectors of the MLE of $\boldsymbol{W}$, and the MLE for σ^2 is a function of the $D - p$ discarded singular values. Despite the fact that the MLE solution resides neatly in a simple singular value decomposition (SVD) on $\boldsymbol{Y}$, its computational cost being $\mathcal{O}(ND^2 + D^3)$ makes standard SVD not scalable to massive dimensional problems. Roweis (1998) pointed out that using expectation maximization (EM) could be computationally cheaper in such cases. However, it is not clear how fast the convergence rate of EM decays in p. Moreover, the performance of PPCA is sensitive to the choice of p, as will be illustrated in § 3.2 of this paper. In practice, p is always picked by cross validation, whose computational cost is prohibitive when D or N is large.

When only the leading d singular values and vectors are required, randomized SVD (Rokhlin et al., 2009) is able to reduce the computational cost to $\mathcal{O}(NDd)$ and approximate the exact solution with small error. However, this is not yet a solution to model (1.1) since the MLE of σ^2 requires all D singular values. Leveraging on this fast SVD technique, we propose a novel Bayesian model that we term geometric density estimation (GEODE). The proposed model learns the geometry of the data vectors by cheaply obtaining the corresponding d principal axes, and then restrict itself in the subspace spanned by these axes. Hence it easily scales to massive dimensional problems ($D \approx 10^6$ in our simulation). We also generalize the model to a mixture of GEODE (mGEODE) to account for nonlinear cases via a dyadic clustering tree and illustrate its performance via real world image data.

When data are continuous, a common strategy is to define a generalized linear model (GLM) for each variable, with dependence in the different variables induced through including multivariate latent variables in the GLMs. Alternatively, discrete data can be directly linked to the latent variables via some link functions. A popular choice for the latent variable distribution is the multivariate Gaussian due to simplicity in modeling the dependence structure. For instance, multivariate Poisson regression with the underlying intercepts modeled jointly as a Gaussian has been widely used in accident analysis (Ma et al., 2008; El-Basyouny et al., 2014). Canale and Dunson (2011) proposed a multivariate count model that handles both overdispersion and under-dispersion. This model uses a rounding function to directly link the multivariate count data to a latent Gaussian. We term these models as multivariate latent Gaussian models. Unfortunately, despite their great flexibility, the usage of this class of models is limited by the computationally challenging model fitting.

The challenge is due partially to the fact that likelihood functions marginalizing out the latent variables lack analytic forms. Hence, Bayesian inference is usually done via data augmented Markov chain Monte Carlo (DA-MCMC) algorithms that sample both the latent variables and the model parameters from their joint posterior. However, it is well known that posterior dependence between the latent variables and the model parameters can substantially slow down the mixing rate of the Markov chain. In fact, Johndrow et al. (2016) has shown that the mixing rate can be so slow that the DA-MCMC sampler cannot generate any reliable posterior samples when the data are severely imbalanced (e.g., excessive zeros in count data).

One possible solution is to bypass sampling entirely using one of the following two strategies. The first is the integrated nested Laplace approximation (INLA), which is designed for latent Gaussian models that have a small number of parameters remaining after marginalizing out the latent variables (Rue et al., 2009). Although INLA has had excellent performance in specialized settings, in many applications, there are moderate to large numbers of population parameters, ruling out such approaches. Another strategy is the so-called variational approximations (Attias, 2000; Jaakkola and Jordan, 2000), which introduce an approximate posterior with a factorized form. One then optimizes the parameters of this approximate posterior to minimize its Kullback-Leibler divergence from the exact posterior. However, in general one has no idea how accurate this approximation is and additionally it is well known that it often substantially underestimates the true posterior covariance.

We propose *Bayesian mosaic*, which is a surrogate posterior derived by multiplying a collection of component posteriors. Unlike variational approximations, where one has to choose the variational class and optimize its parameters, the construction of *Bayesian mosaic* is automatically determined by the data distribution. It is related to the composite likelihood approach (Cox and Reid, 2004) with its component posteriors being based on univariate and bivariate marginal distributions. However, *Bayesian mosaic* is different from Bayesian composite likelihood methods (Pauli et al., 2011) in that it has an easy-to-sample multiplicative form while a posterior density induced by a composite likelihood does not. Utilizing that these marginal densities have unbiased score functions, we have shown that *Bayesian mosaic* is consistent and asymptotically normal under mild conditions. It is applicable to a class of *mosaic-type* data distributions that covers and is much broader than the class of multivariate latent Gaussian models mentioned earlier.

We also propose an efficient parallel sampling strategy utilizing the posterior dependence structure induced by the multiplicative form. This parallelization is substantially different from standard parallel MCMC algorithms which are based on partitions of the dataset (Wang and Dunson, 2013; Scott et al., 2016). The sparse dependence structure of *Bayesian mosaic* allows us to directly sample from each component posterior independently. Moreover, we have shown that the asymptotic per-iteration computational complexity of sampling from *Bayesian mosaic* is linear in the cardinality of the observed data, which is in general much smaller than the sample size. On the other hand, the per-iteration computational complexity of DA-MCMC is linear in the sample size.

The remainder of the dissertation is organized as follows. Chapter 2 presents the electrostatic Gaussian process model for learning smooth curve. Chapter 3 presents the geometric density estimator for learning lower-dimensional structures. Chapter 4 presents *Bayesian mosaic* for fitting multivariate discrete data models. A conclusion will be given in Chapter 5.

2

Electrostatic Gaussian Process

2.1 Coulomb repulsive process

2.1.1 Formulation

Definition 2.1.1. *A univariate process is a Coulomb repulsive process (Corp) if and only if for every finite set of indices $t_1, \ldots, t_k$ in the index set $\mathbb{N}_+$,*

$$\begin{aligned} &X_{t_1} \sim \mathit{unif}(0,1), \\ &p(X_{t_i}|X_{t_1}, \ldots, X_{t_{i-1}}) \propto \Pi_{j=1}^{i-1} \sin^{2r}\left(\pi X_{t_i} - \pi X_{t_j}\right) \mathbb{1}_{X_{t_i} \in [0,1]},\ i > 1, \end{aligned} \tag{2.1}$$

where $r > 0$ is the repulsive parameter. The process is denoted as $X_t \sim \mathit{Corp}(r)$.

The process is named by its analogy in electrostatic physics where by Coulomb law, two electrostatic positive charges will repel each other by a force proportional to the reciprocal of their square distance. Letting $d(x,y) = \sin|\pi x - \pi y|$, the above conditional probability of X_{t_i} given X_{t_j} is proportional to $d^{2r}(X_{t_i}, X_{t_j})$, shrinking the probability exponentially fast as two states get closer to each other. Note that the periodicity of the sine function eliminates the edges of $[0,1]$, making the electrostatic energy field homogeneous everywhere on $[0,1]$.

Several observations related to Kolmogorov extension theorem can be made immediately, ensuring Corp to be well defined. Firstly, the conditional density defined in (2.1) is positive and integrable, since X_t's are constrained in a compact interval, and $\sin^{2r}(\cdot)$ is positive and bounded. Hence, the finite distributions are well defined.

Secondly, the joint finite p.d.f. for $X_{t_1}, \ldots, X_{t_k}$ can be derived as

$$p(X_{t_1}, \ldots, X_{t_k}) \propto \Pi_{i>j} \sin^{2r}\left(\pi X_{t_i} - \pi X_{t_j}\right). \tag{2.2}$$

As can be easily seen, any permutation of $t_1, \ldots, t_k$ will result in the same joint finite distribution, hence this finite distribution is exchangeable.

Thirdly, it can be easily checked that for any finite set of indices $t_1, \ldots, t_{k+m}$,

$$p(X_{t_1}, \ldots, X_{t_k}) = \int_0^1 \cdots \int_0^1 p(X_{t_1}, \ldots, X_{t_k}, X_{t_{k+1}}, \ldots, X_{t_{k+m}}) \mathrm{d}X_{t_{k+1}} \ldots \mathrm{d}X_{t_{k+m}},$$

by observing that

$$p(X_{t_1}, \ldots, X_{t_k}, X_{t_{k+1}}, \ldots, X_{t_{k+m}}) = p(X_{t_1}, \ldots, X_{t_k}) \Pi_{j=1}^m p(X_{t_{k+j}} | X_{t_1}, \ldots, X_{t_{k+j-1}}).$$

2.1.2 Properties

Assuming X_t, $t \in \mathbb{N}_+$ is a realization from Corp, then the following lemmas hold.

Lemma 2.1.2. *For any $n \in \mathbb{N}_+$, any $1 \leqslant i < n$ and any $\epsilon > 0$, we have*

$$p(X_n \in \mathcal{B}(X_i, \epsilon) | X_1, \ldots, X_{n-1}) < \frac{2\pi^2 \epsilon^{2r+1}}{2r+1}$$

where $\mathcal{B}(X_i, \epsilon) = \{X \in (0,1) : d(X - X_i) < \epsilon\}$.

Lemma 2.1.3. *For any $n \in \mathbb{N}_+$, the p.d.f. (2.2) of $X_1, \ldots, X_n$ (due to the exchangeability, we can assume $X_1 < X_2 < \cdots < X_n$ without loss of generality) is maximized when and only when*

$$d(X_i - X_{i-1}) = \sin\left(\frac{\pi}{n+1}\right) \text{ for all } 2 \leqslant i \leqslant n.$$

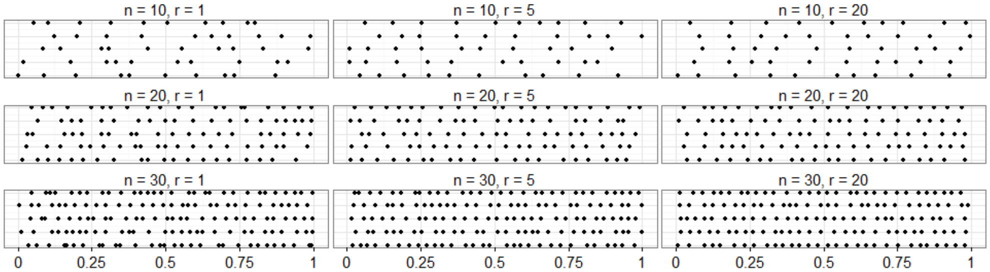

FIGURE 2.1: Simulations from Corp.

According to Lemma 2.1.2 and Lemma 2.1.3, Corp will nudge the x's to be spread out within $[0,1]$, and penalizes the case when two x's get too close. Figure 2.1 presents some simulations from Corp, where each facet consists of 5 rows, with each row representing an 1-dimensional scatterplot of a random realization of Corp under certain n and r. This nudge becomes stronger as the sample size n grows, or as the repulsive parameter r grows. The properties of Corp makes it ideal for strongly favoring spread out latent positions across the manifold, avoiding the gaps and clustering in small regions that plague GP-LVM-type methods. The proofs for the lemmas and a simulation algorithm based on rejection sampling can be found in the supplement.

2.1.3 Multivariate Corp

Definition 2.1.4. *A p-dimensional multivariate process is a Coulomb repulsive process if and only if for every finite set of indices $t_1, \ldots, t_k$ in the index set $\mathbb{N}_+$,*

$$X_{m,t_1} \sim \textit{unif}(0,1),\ \textit{for } m = 1, \ldots, p$$

$$p(\boldsymbol{X}_{t_i}|\boldsymbol{X}_{t_1}, \ldots, \boldsymbol{X}_{t_{i-1}}) \propto \Pi_{j=1}^{i-1}\left[\sum_{m=1}^{p+1}(Y_{m,t_i} - Y_{m,t_j})^2\right]^r \mathbb{1}_{X_{t_i}\in(0,1)},\ i > 1$$

where the p-dimensional spherical coordinates $\boldsymbol{X}_t$'s have been converted into the $(p+1)$-dimensional Cartesian coordinates $\boldsymbol{Y}_t$:

$$
\begin{aligned}
Y_{1,t} &= \cos(2\pi X_{1,t}) \\
Y_{2,t} &= \sin(2\pi X_{1,t})\cos(2\pi X_{2,t}) \\
&\vdots \\
Y_{p,t} &= \sin(2\pi X_{1,t})\sin(2\pi X_{2,t})\ldots\sin(2\pi X_{p-1,t})\cos(2\pi X_{p,t}) \\
Y_{p+1,t} &= \sin(2\pi X_{1,t})\sin(2\pi X_{2,t})\ldots\sin(2\pi X_{p-1,t})\sin(2\pi X_{p,t}).
\end{aligned}
$$

The multivariate Corp maps the hyper-cubic $(0,1)^p$ through a spherical coordinate system to a unit hyper-ball in $\Re^{p+1}$. The repulsion is then defined as the reciprocal of the square Euclidean distances between these mapped points in $\Re^{p+1}$. Based on this construction of multivariate Corp, a straightfoward generalization of the electroGP model to a p-dimensional manifold could be made, where $p > 1$.

2.2 Electrostatic Gaussian Process

2.2.1 Formulation and Model Fitting

In this section, we propose the electrostatic Gaussian process (electroGP) model. Assuming n d-dimensional data vectors $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ are observed, the model is given by

$$
\begin{aligned}
y_{i,j} &= \mu_j(x_i) + \epsilon_{i,j}, \quad \epsilon_{i,j} \sim \mathcal{N}(0, \sigma_j^2), \\
x_i &\sim \text{Corp}(r), \quad i = 1, \ldots, n, \\
\mu_j &\sim \mathcal{GP}(0, K^j), \quad j = 1, \ldots, d,
\end{aligned}
\tag{2.3}
$$

where $\boldsymbol{y}_i = (y_{i,1}, \ldots, y_{i,d})$ for $i = 1, \ldots, n$ and $\mathcal{GP}(0, K^j)$ denotes a Gaussian process prior with covariance function $K^j(x,y) = \phi_j \exp\big\{ - \alpha_j (x-y)^2 \big\}$.

Letting $\Theta = (\sigma_1^2, \alpha_1, \phi_1, \ldots, \sigma_d^2, \alpha_d, \phi_d)$ denote the model hyperparameters, model (2.3) could be fitted by maximizing the joint posterior distribution of $\boldsymbol{x} = (x_1, \ldots . x_n)$

and Θ,

$$(\hat{\boldsymbol{x}}, \hat{\Theta}) = \arg\max_{\boldsymbol{x}.\Theta} p(\boldsymbol{x}|\boldsymbol{y}_{1:n}, \Theta, r), \tag{2.4}$$

where the repulsive parameter r is fixed and can be tuned using cross validation. Based on our experience, setting $r = 1$ always yields good results, and hence is used as a default across this paper. For the simplicity of notations, r is excluded in the remainder. The above optimization problem can be rewritten as

$$(\hat{\boldsymbol{x}}, \hat{\Theta}) = \arg\max_{\boldsymbol{x}.\Theta} l(\boldsymbol{y}_{1:n}|\boldsymbol{x}, \Theta) + \log\big[\pi(\boldsymbol{x})\big],$$

where $l(\cdot)$ denotes the log likelihood function and $\pi(\cdot)$ denotes the finite dimensional pdf of Corp. Hence the Corp prior can also be viewed as a repulsive constraint in the optimization problem.

It can be easily checked that $\log\big[\pi(x_i = x_j)\big] = -\infty$, for any i and j. Starting at initial values x_0, the optimizer will converge to a local solution that maintains the same order as the initial x_0's. We refer to this as the *self-truncation property.* We find that conditionally on the starting order, the optimization algorithm converges rapidly and yields stable results. Although the x's are not identifiable, since the target function (2.4) is invariant under rotation, a unique solution does exist conditionally on the specified order.

Self-truncation raises the necessity of finding good initial values, or at least a good initial ordering for x's. Fortunately, in our experience, simply applying any standard manifold learning algorithm to estimate x_0 in a manner that preserves distances in $\mathcal{Y}$ yields good performance. We find very similar results using LLE, Isomap and eigenmap, but focus on LLE in all our implementations. Our algorithm can be summarized as follows.

1. Learn the one dimensional coordinate $\boldsymbol{x}_0$ by your favorite distance-preserving manifold learning algorithm and rescale $\boldsymbol{x}_0$ into $(0, 1)$;

2. Solve $\Theta_0 = \arg\max_\Theta p(\boldsymbol{y}_{1:n}|\boldsymbol{x}_0, \Theta, r)$ using scaled conjugate gradient descent (SCG);

3. Using SCG, setting $\boldsymbol{x}_0$ and Θ_0 to be the initial values, solve $\hat{\boldsymbol{x}}$ and $\hat{\Theta}$ w.r.t. (2.4).

2.2.2 Posterior Mean Curve and Uncertainty Bands

In this subsection, we describe how to obtain a point estimate of the curve $\boldsymbol{\mu}$ and how to characterize its uncertainty under electroGP. Such point and interval estimation is as of yet unsolved in the literature, and is of critical importance. In particular, it is difficult to interpret a single point estimate without some quantification of how uncertain that estimate is. We use the posterior mean curve $\hat{\boldsymbol{\mu}} = \mathbb{E}(\boldsymbol{\mu}|\hat{\boldsymbol{x}}, \boldsymbol{y}_{1:n}, \hat{\Theta})$ as the Bayes optimal estimator under squared error loss. As a curve, $\hat{\boldsymbol{\mu}}$ has infinite dimensions. Hence, in order to store and visualize it, we discretize $[0, 1]$ to obtain n_μ equally-spaced grid points $x_i^\mu = \frac{i-1}{n_\mu - 1}$ for $i = 1, \ldots, n_\mu$. Using basic multivariate Gaussian theory, the following expectation is easy to compute.

$$\big(\hat{\boldsymbol{\mu}}(x_1^\mu), \ldots, \hat{\boldsymbol{\mu}}(x_{n_\mu}^\mu)\big) = \mathbb{E}\big(\boldsymbol{\mu}(x_1^\mu), \ldots, \boldsymbol{\mu}(x_{n_\mu}^\mu)|\hat{\boldsymbol{x}}, \boldsymbol{y}_{1:n}, \hat{\Theta}\big).$$

Then $\hat{\boldsymbol{\mu}}$ is approximated by linear interpolation using $\big\{x_i^\mu, \hat{\boldsymbol{\mu}}(x_i^\mu)\big\}_{i=1}^{n_\mu}$. For ease of notation, we use $\hat{\boldsymbol{\mu}}$ to denote this interpolated piecewise linear curve later on. Figure 2.2 provides visualization of three simulation experiments where the data (triangles) are simulated from a bivariate Gaussian (**left**), a rotated parabola with Gaussian noises (**middle**) and a spiral with Gaussian noises (**right**). The dotted shading denotes the 95% posterior predictive uncertainty band of (y_1, y_2) under electroGP. The black curve denotes the posterior mean curve under electroGP and the red curve denotes the P-curve. The three dashed curves denote three realizations from GP-LVM. The middle panel shows a zoom-in region and the full figure is shown in the embedded box. All the mean curves (black solid) were obtained using the above method.

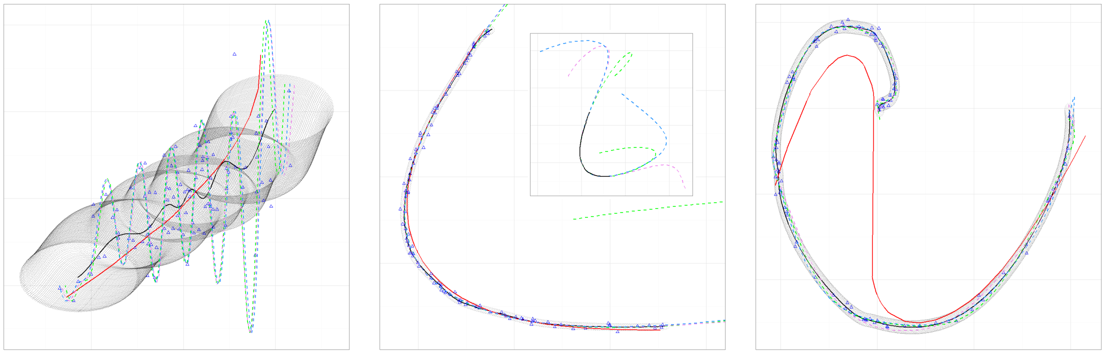

FIGURE 2.2: Visualization of three simulation experiments.

Estimating an uncertainty region including data points with η probability is much more challenging. We addressed this problem by the following heuristic algorithm.

Step 1. Draw x_i^*'s from Unif(0,1) independently for $i = 1, \dots, n_1$;

Step 2. Sample the corresponding $\boldsymbol{y}_i^*$ from the posterior predictive distribution conditional on these latent coordinates $p(\boldsymbol{y}_1^*, \dots, \boldsymbol{y}_{n_1}^* | x_{1:n_1}^*, \hat{\boldsymbol{x}}, \boldsymbol{y}_{1:n}, \hat{\Theta})$;

Step 3. Repeat steps 1-2 n_2 times, collecting all $n_1 \times n_2$ samples $\boldsymbol{y}^*$'s;

Step 4. Find the shortest distances from these $\boldsymbol{y}^*$'s to the posterior mean curve $\hat{\boldsymbol{\mu}}$, and find the η-quantile of these distances denoted by ρ;

Step 5. Moving a radius-ρ ball through the entire curve $\hat{\boldsymbol{\mu}}([0,1])$, the envelope of the moving trace defines the η% uncertainty band.

Note that step 4 can be easily solved since $\hat{\boldsymbol{\mu}}$ is a piecewise linear curve. Examples can be found in Figure 2.2, where the 95% uncertainty bands (dotted shading) were found using the above algorithm.

2.2.3 Simulation

In this subsection, we compare the performance of electroGP with GP-LVM and principal curves (P-curve) in simple simulation experiments. 100 data points were sampled from each of the following three 2-dimensional distributions: a Gaussian distribution, a rotated parabola with Gaussian noises and a spiral with Gaussian noises. ElectroGP and GP-LVM were fitted using the same initial values obtained from LLE, and the P-Curve was fitted using the `princurve` package in R.

Figure 2.2 presents the zoom-in of the spiral case 3 (**left**) and the corresponding coordinate function, $\mu_2(x)$, of electroGP (**middle**) and GP-LVM (**right**). The gray shading denotes the heatmap of the posterior distribution of (x, y_2) and the black curve denotes the posterior mean. The dotted shading represents a 95% posterior predictive uncertainty band for a new data point $\boldsymbol{y}_{n+1}$ under the electroGP model. This illustrates that electroGP obtains an excellent fit to the data, provides a good characterization of uncertainty, and accurately captures the concentration near a 1d manifold embedded in two dimensions. The P-curve is plotted in red. The extremely poor representation of P-curve is as expected based on our experience in fitting principal curve in a wide variety of cases; the behavior is highly unstable. In the first two cases, the P-Curve corresponds to a smooth curve through the center of the data, but for the more complex manifold in the third case, the P-Curve is an extremely poor representation. This tendency to cut across large regions of near zero data density for highly curved manifolds is common for P-Curve.

For GP-LVM, we show three random realizations (dashed) from the posterior in each case. It is clear the results are completely unreliable, with the tendency being to place part of the curve through where the data have high density, while also erratically adding extra outside the range of the data. The GP-LVM model does not appropriately penalize such extra parts, and the very poor performance shown in the top right of Figure 2.2 is not unusual. We find that electroGP in general performs dramatically better than competitors. More simulation results can be found in the supplement. To better illustrate the results for the spiral case 3, we zoom in and present some further comparisons of GP-LVM and electroGP in Figure 2.3.

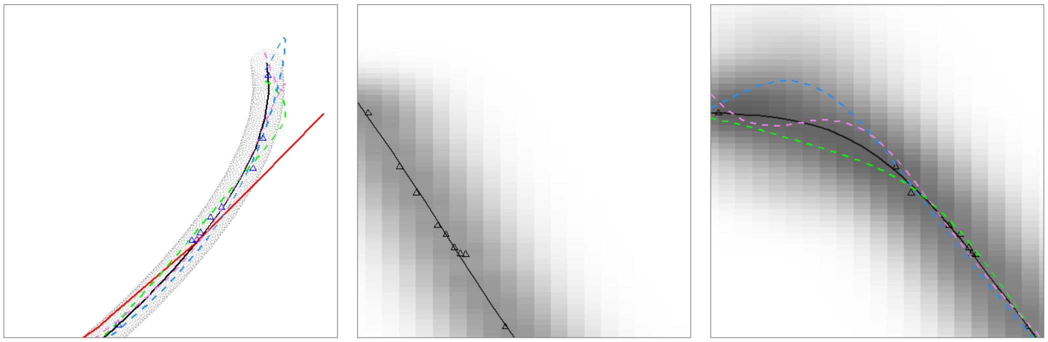

FIGURE 2.3: The zoom-in of the functional posterior.

As can be seen the right panel, optimizing x's without any constraint results in "holes" on $[0, 1]$. The trajectories of the Gaussian process over these holes will become arbitrary, as illustrated by the three realizations. This arbitrariness will be further projected into the input space $\mathcal{Y}$, resulting in the erratic curve observed in the left panel. Failing to have well spread out x's over $[0, 1]$ not only causes trouble in learning the curve, but also makes the posterior predictive distribution of $\boldsymbol{y}_{n+1}$ overly diffuse near these holes, e.g., the large gray shading area in the right panel. The middle panel shows that electroGP fills in these holes by softly constraining the latent coordinates x's to spread out while still allowing the flexibility of moving them around to find a smooth curve snaking through them.

2.2.4 Prediction

Broad prediction problems can be formulated as the following missing data problem. Assume m new data $\boldsymbol{z}_i$, for $i = 1, \ldots, m$, are partially observed and the missing entries are to be filled in. Letting $\boldsymbol{z}_i^O$ denote the observed data vector and $\boldsymbol{z}_i^M$ denote the missing part, the conditional distribution of the missing data is given by

$$\begin{aligned} & p(\boldsymbol{z}_{1:m}^M | \boldsymbol{z}_{1:m}^O, \hat{\boldsymbol{x}}, \boldsymbol{y}_{1:n}, \hat{\Theta}) \\ = & \int_{x_1^z} \cdots \int_{x_m^z} p(\boldsymbol{z}_{1:m}^M | x_{1:m}^z, \hat{\boldsymbol{x}}, \boldsymbol{y}_{1:n}, \hat{\Theta}) \times p(x_{1:m}^z | \boldsymbol{z}_{1:m}^O, \hat{\boldsymbol{x}}, \boldsymbol{y}_{1:n}, \hat{\Theta}) \mathrm{d}x_1^z \cdots \mathrm{d}x_m^z, \end{aligned}$$

where x_i^z is the corresponding latent coordinate of $\boldsymbol{z}_i$, for $i = 1, \dots, n$. However, dealing with $(x_1^z, \dots, x_m^z)$ jointly is intractable due to the high non-linearity of the Gaussian process, which motivates the following approximation,

$$p(x_{1:m}^z | \boldsymbol{z}_{1:m}^O, \hat{\boldsymbol{x}}, \boldsymbol{y}_{1:n}, \hat{\Theta}) \approx \Pi_{i=1}^m p(x_i^z | \boldsymbol{z}_i^O, \hat{\boldsymbol{x}}, \boldsymbol{y}_{1:n}, \hat{\Theta}).$$

The approximation assumes $(x_1^z, \dots, x_m^z)$ to be conditionally independent. This assumption is more accurate if $\hat{\boldsymbol{x}}$ is well spread out on $(0, 1)$, as is favored by Corp.

The univariate distribution $p(x_i^z | \boldsymbol{x}_i^O, \boldsymbol{y}_{1:n}, \hat{\boldsymbol{u}}, \hat{\Theta})$, though still intractable, is much easier to deal with. Depending on the purpose of the application, either a Metropolis Hasting algorithm could be adopted to sample from the predictive distribution, or a optimization method could be used to find the MAP of x^z's. The details of both algorithms can be found in the supplement.

2.3 Experiments

Video-inpainting 200 consecutive frames (of size 76 × 101 with RGB color) Weinberger and Saul (2006) were collected from a video of a teapot rotating 180°. Clearly these images roughly lie on a curve. 190 of the frames were assumed to be fully observed in the natural time order of the video, while the other 10 frames were given without any ordering information. Moreover, half of the pixels of these 10 frames were missing. The electroGP was fitted based on the other 190 images and was used to reconstruct the broken frames and impute the reconstructed frames into the whole frame series with the correct order. The reconstruction results are presented in Figure 2.4, where the **Left Panel** presents three randomly selected reconstructions using electroGP compared with those using Bayesian GP-LVM; the **Right Panel** presents another three reconstructions from electroGP, with the first row presenting the original images, the second row presenting the observed images and the third row presenting the reconstructions. As can be seen, the reconstructed images are

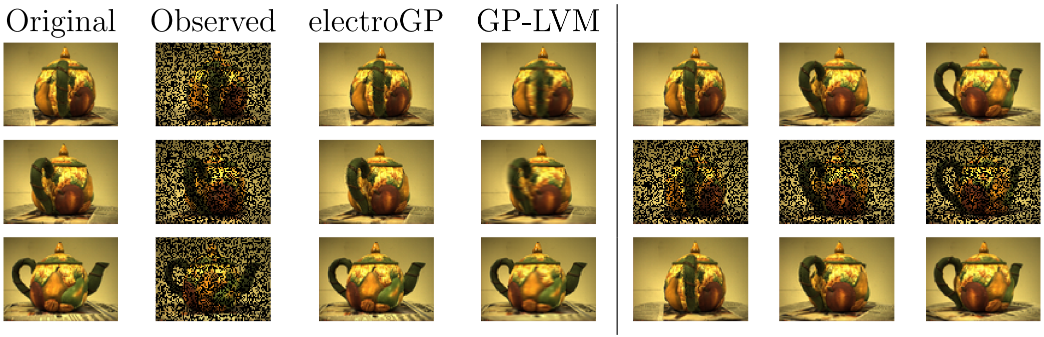

FIGURE 2.4: Reconstruction of teapot images.

almost indistinguishable from the original ones. Note that these 10 frames were also correctly imputed into the video with respect to their latent position x's. ElectroGP was compared with Bayesian GP-LVM Titsias and Lawrence (2010) with the latent dimension set to 1. The reconstruction mean square error (MSE) using electroGP is 70.62, compared to 450.75 using GP-LVM. The comparison is also presented in Figure 2.4. It can be seen that electroGP outperforms Bayesian GP-LVM in high-resolution precision (e.g., how well they reconstructed the handle of the teapot) since it obtains a much tighter and more precise estimate of the manifold.

Super-resolution & Denoising 100 consecutive frames (of size 100×100 with gray color) were collected from a video of a shrinking shockwave. Frame 51 to 55 were assumed completely missing and the other 95 frames were observed with the original time order with strong white noises. The shockwave is homogeneous in all directions from the center; hence, the frames roughly lie on a curve. The electroGP was applied for two tasks: 1. Frame denoising; 2. Improving resolution by interpolating frames in between the existing frames. Note that the second task is hard since there are 5 consecutive frames missing and they can be interpolated only if the electroGP correctly learns the underlying manifold.

The denoising performance was compared with non-local mean filter (NLM) Buades et al. (2005) and isotropic diffusion (IsD) Perona and Malik (1990). The interpolation performance was compared with linear interpolation (LI). The com-

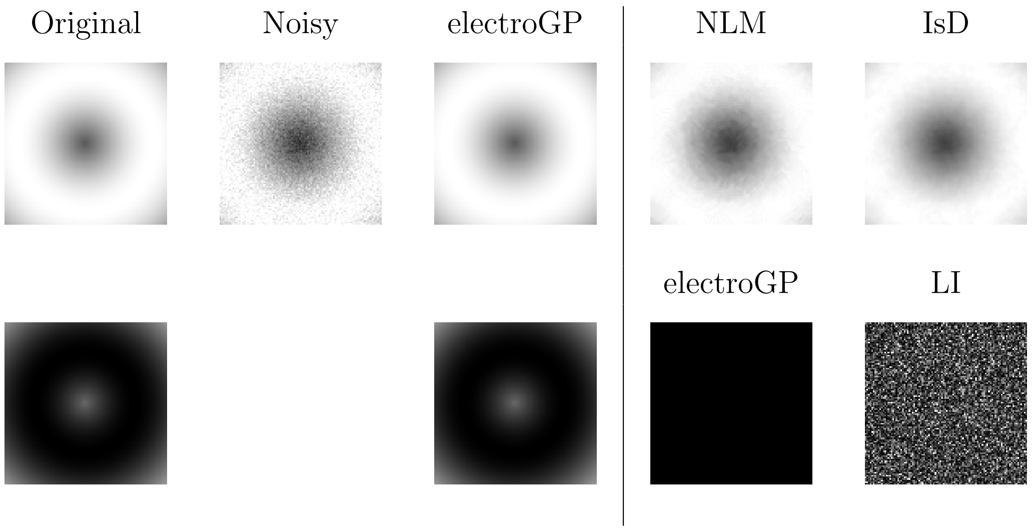

FIGURE 2.5: Performance comparison on the shrinking shockwave images.

parison is presented in Figure 2.5. From left to right on row 1 are the original 95th frame, its noisy observation, its denoised result by electroGP, NLM and IsD; From left to right on row 2 are the original 53th frame, its regeneration by electroGP, the residual image (10 times of the absolute error between the imputation and the original) of electroGP and LI. The blank area denotes its missing observation. As can be clearly seen, electroGP greatly outperforms other methods since it correctly learned this one-dimensional manifold. To be specific, the denoising MSE using electroGP is only 1.8×10^{-3}, comparing to 63.37 using NLM and 61.79 using IsD. The MSE of reconstructing the entirely missing frame 53 using electroGP is 2×10^{-5} compared to 13 using LI. An online video of the super-resolution result using electroGP can be found in this link[1]. The frame per second (fps) of the generated video under electroGP was tripled compared to the original one. Though over two thirds of the frames are pure generations from electroGP, this new video flows quite smoothly. Another noticeable thing is that the 5 missing frames were perfectly regenerated by electroGP.

---

[1] https://youtu.be/N1BG220J5Js This online video contains no information regarding the authors.

3

Scalable Multiscale Density Estimation

3.1 Geometric Density Estimation

In this section, GEODE is being built piece by piece. The correctness of the model is first justified via a theorem, and a shrinkage prior is then specified to enable GEODE automatically learn the intrinsic dimension p. Finally an efficient Gibbs sampler is designed for posterior computation.

3.1.1 Model Formulation

In GEODE, the data likelihood is assumed to be the same as that of model (1.1), with priors specified for the noise variance σ^2 and the diagonal matrix $\boldsymbol{\Sigma}$. The model is as follow

$$\begin{aligned} &\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{C}) \\ &\sigma^2 \sim \mathrm{IG}(a_\sigma, b_\sigma),\ \boldsymbol{\Sigma} \sim \Pi, \end{aligned} \tag{3.1}$$

where $\mathrm{IG}(a_\sigma, b_\sigma)$ denotes a inverse Gamma distribution with shape parameter a_σ and rate parameter b_σ, and Π is some prior distribution for $\boldsymbol{W}$. The corresponding log-posterior is as follow

$$\mathcal{L} = -\frac{N}{2}\big\{D\ln(2\pi) + \ln|\boldsymbol{C}| + \mathrm{tr}(\boldsymbol{C}^{-1}\boldsymbol{S}) - \ln(\Pi(\boldsymbol{\Sigma})) + \frac{a_\sigma + 1}{N}\ln(\sigma^{-2}) - \frac{b_\sigma}{N}\sigma^{-2}\big\},$$

where $\boldsymbol{S} = \frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{y}_i - \boldsymbol{\mu})(\boldsymbol{y}_i - \boldsymbol{\mu})^\top$ is the sample covariance matrix.

3.1.2 Empirical Bayes Solution

$\boldsymbol{\mu}$ and $\boldsymbol{W}$ carries the geometric information of the data vectors since they uniquely define a linear subspace in $\Re^D$ with $\boldsymbol{\mu}$ being the origin. GEODE treats them as model hyperparameters due to computational concern. Hence from a empirical Bayesian perspective, we will first learn them by a single pass through the data and fix them at the learned values afterwards. To be specific, we learn $\boldsymbol{\mu}$ and $\boldsymbol{W}$ by solving the following optimization

$$(\boldsymbol{\mu}, \boldsymbol{W}) = \arg\max_{\boldsymbol{\mu}, \boldsymbol{W}}\big[\max_{\sigma^2, \boldsymbol{\Sigma}}\mathcal{L}(\boldsymbol{\mu}, \boldsymbol{W}, \sigma^2, \boldsymbol{\Sigma})\big]. \tag{3.2}$$

Intuitively, the most “likely” choice for $\boldsymbol{\mu}$ and $\boldsymbol{W}$ are those that maximize the posterior mode. The following theorem shows that a closed form solution to (A.1) exists and can be obtained via a single SVD through the data. The proof is reported in the supplementary material.

Theorem 3.1.1. *Suppose*

- Π *has support on all* $d \times d$ *positive diagonal matrices.*
- *The shape parameter* a_σ *satisfies*

$$\frac{a_\sigma + 1}{N} < \frac{D - p}{\lambda_D}\Big[\sum_{j=p-1}^{D-1}\lambda_j/(D-p) - \lambda_D\Big]. \tag{3.3}$$

Then

$$\boldsymbol{\mu} = \bar{\boldsymbol{y}},\ \boldsymbol{W} = \boldsymbol{U}_p$$

solves (A.1), where $\lambda_1, \ldots, \lambda_D$ *are the eigenvalues of* $\boldsymbol{S}$ *in a descending order and the* p *column vectors in the* $D \times p$ *matrix* $\boldsymbol{U}_p$ *are the* p *leading eigenvectors of* $\boldsymbol{S}$*.*

Interestingly, the above result does not require any specific distributional form for Π. In practice, condition (3.3) should also be easily met when N is large. It can be easily checked that the eigenvectors of $\boldsymbol{S}$ are the right singular vectors of $\boldsymbol{Y}$, i.e., $\boldsymbol{w}_j$ for $j = 1, \ldots, d$. Hence Theorem A.1.1 shows that the span of the p leading principal axes is the "Bayesian optimal" p-dimensional linear subspace in $\Re^D$ under model (3.1), no matter how you specified your priors.

The theorem raises a practical method for finding $\boldsymbol{\mu}$ and $\boldsymbol{W}$, which is summarized as follows

- $\boldsymbol{\mu} = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{y}_i$.
- obtain $\boldsymbol{w}_j$, for $j = 1, \ldots, d$ via applying the fast rank-d SVD (Rokhlin et al., 2009) on $\boldsymbol{Y}$.

This will be the first step in our method. Note that here we obtain d principal axes instead of p. This is because we do not know the true p hence we start from a conservative guess d such that $d \geqslant p$. In future subsections we will see how one can define a shrinkage prior on $\boldsymbol{\Sigma}$ and derive a adaptive Gibbs sampler to automatically learn p.

3.1.3 Bayesian Learning of Intrinsic Dimension

Equipped with $\boldsymbol{\mu}$ and $\boldsymbol{W}$ and with another pass through the data, we can obtain for all $i = 1, \ldots, N$ sufficient statistics $A_i = (\boldsymbol{y}_i - \boldsymbol{\mu})^\top (\boldsymbol{y}_i - \boldsymbol{\mu})$ and $\boldsymbol{Z}_i = \boldsymbol{W}^\top (\boldsymbol{y}_i - \boldsymbol{\mu})$, with $Z_i^{(j)}$ denoting the j th element of $\boldsymbol{Z}_i$. We then apply a random variable transformation $u_j = (1 + \sigma^{-2}\alpha_j^2)^{-1}$, for $j = 1, \ldots, d$. With basic algebra, the likelihood of GEODE can then be written as

$$f(\boldsymbol{y}_i)\propto(\sigma^2)^{-D/2}\prod_{j=1}^{d}u_j^{1/2}\exp\Big\{-\frac{1}{2}\sigma^{-2}\times\big[A_i-\sum_{j=1}^{d}(1-u_j)(Z_i^{(j)})^2\big]\Big\}.$$

Detailed derivation is reported in the supplementary material.

Learning p, we rely on the following geometric intuition. It is easy to check that $\boldsymbol{w}_j^\top\boldsymbol{y}\sim\mathcal{N}(\boldsymbol{w}_j^\top\boldsymbol{\mu},\alpha_j^2+\sigma^2)$. Hence α_j^2 is the signal variance along the direction $\boldsymbol{w}_j$, which should be decreasing for $j=1,\ldots,p$ and be zero for $j=p+1,\ldots,d$. This motivates us to penalize α_j^2 by increasingly shrinking them towards zero with respect to j, which is equivalent to shrinking u_j increasingly for larger j. To accomplish this adaptive shrinkage, we propose a multiplicative exponential process prior that adapts the prior of Bhattacharya and Dunson (2011). Letting $\delta_j=\prod_{k=1}^{j}\tau_k$, the prior is given for $j=1,\ldots,d$ as follow

$$u_j\sim\mathrm{Ga}_{(0,1)}(\delta_j+1,1)$$
$$\tau_j\sim\mathrm{Exp}_{[1,\infty)}(a_\tau)$$

where $\mathrm{Ga}_{(0,1)}(\delta_j+1,1)$ denotes a Gamma distribution with shape parameter δ_j+1 and rate parameter 1 truncated within $(0,1)$ and $\mathrm{Exp}_{[1,\infty)}(a_\tau)$ denotes a Exponential distribution with parameter a_τ truncated within $[1,\infty)$. δ_j and τ_j are the global and the local shrinkage parameter for $\boldsymbol{w}_j$ respectively. Since $\tau_j\geqslant 1$ for $j=1,\ldots,d$, $\delta_j=\prod_{k=1}^{j}\tau_k$ is increasing with respect to j. As a result, u_j is stochastically approaching one since the truncated gamma density concentrates around one as δ_j increases.

3.1.4 Posterior Computation

In practice, conservative choice for d is implemented in order to ensure $d\geqslant p$, adding a burden to both computation and storage. We avoid this by automatically deleting redundant principal axes, and hence decreasing d, as computation proceeds. To this end we adopt an adaptive Gibbs sampler similar to that developed by Bhattacharya and Dunson (2011). The adaptive Gibbs sampler randomly deletes redundant axes at t th iteration according to probability $p(t) = \exp(c_0 + c_1 t)$. The values of c_0 and c_1 are chosen to ensure frequent adaption at the beginning of the chain and an exponentially fast decay in frequency after that. In practice, we fix $c_0 = -1$, $c_1 = -0.005$, $a_\sigma = 2$, $b_\sigma = 2$, $a_\tau = 0.05$ and $\text{tol} = 10^{-2}$ as default, where tol is a prespecified threshold. This set of default values is validated through our simulations.

The algorithm fitting GEODE can then be summarized as follows

Step 1 (preprocessing): Compute $\boldsymbol{\mu}$ and $\boldsymbol{W}$ as described in § 3.1.2 and compute sufficient statistics A_i and $\boldsymbol{Z}_i$ for $i = 1, \ldots, N$.

Step 2 (Gibbs sampler): Set $\mathcal{R} = \{1, \ldots, d\}$ and $\mathcal{D} = \varnothing$. Iterate until the desired posterior sample size:

1. Update u_j for all $j \in \mathcal{R}$ according to $\text{Ga}_{(0,1)}\big(\hat{a}_j, \hat{b}_j\big)$, where $\hat{a}_j = \prod_{k<=j,k\in\mathcal{R}} \tau_k + N/2$ and $\hat{b}_j = 1 + \frac{1}{2}\sigma^{-2}\sum_{i=1}^{n}(\boldsymbol{Z}_i^{(j)})^2$.

2. Update τ_j for all $j \in \mathcal{R}$ according to $\text{Exp}_{[1,\infty)}\big(\hat{\lambda}_j\big)$, where $\hat{\lambda}_j = a_\tau - \ln(\prod_{k>j-1,j\in\mathcal{R}} u_k)$

3. Update σ^{-2} according to $\text{Ga}\big(\hat{c}, \hat{d}\big)$, where $\hat{c} = a_\sigma + DN/2$, $\hat{d} = \frac{1}{2}\sum_{i=1}^{N}\big[A_i - \sum_{j\in\mathcal{R}}(1-u_j)(\boldsymbol{Z}_i^{(j)})^2\big] + b_\sigma$.

4. Compute $p(t) = \exp(c_0 + c_1 t)$, generate g from Uniform$(0,1)$. If $g > p(t)$, go back to step 2 until the desired iteration number.

5. Move all $j \in \mathcal{R}$ such that $r_j^t = \big(\alpha_j^t\big)^2 / \max_{j\in\mathcal{R}}\big(\alpha_j^t\big)^2 < tol$ from $\mathcal{R}$ to $\mathcal{D}$. If no such j exists, then move the smallest j from $\mathcal{D}$ to $\mathcal{R}$.

The derivation of all the conditional posteriors can be found in the supplementary material. In the proposed algorithm, the preprocessing part only involves two pass through the data with a computational cost linear in D and the cost of the Gibbs sampler is independent of D. This makes it easily scale to massive dimensional problems. The superior computational performance of GEODE is illustrated in the next section via simulations and a detailed discussion on the computational cost is reported in § 3.4.

3.1.5 Missing Data Imputation

Bayesian models better utilize the partially observed data by probabilistically imputing the missing features based on its conditional posterior distribution. Moreover, prediction can also be viewed as a missing data imputation problem. We propose several scalable missing data strategies for GEODE and discuss the appropriateness of these strategies in different missing data scenarios.

Notations $\boldsymbol{y}_M$ and $\boldsymbol{y}_O$ are introduced as the missing part and the observed part of $\boldsymbol{y}$ respectively. Similarly, slightly abusing the notations, let $\boldsymbol{\mu}_M$ and $\boldsymbol{W}_M$ denote the missing parts of $\boldsymbol{\mu}$ and $\boldsymbol{W}$, and let $\boldsymbol{\mu}_O$ and $\boldsymbol{W}_O$ denote the observed parts. The following proposition enables efficient sampling from the conditional posterior distribution $p(\boldsymbol{y}_M|\boldsymbol{y}_O, \boldsymbol{\Theta})$, where $\boldsymbol{\Theta}$ denotes all the unknown parameters in the model.

Proposition 3.1.2. *Introduce augmented data $\boldsymbol{\eta} \in \Re^d$ such that $(\boldsymbol{y}|\boldsymbol{\eta}, \boldsymbol{\Theta}) \sim \mathcal{N}(\boldsymbol{\mu} + \boldsymbol{W}\boldsymbol{\eta}, \sigma^2 \boldsymbol{I})$ and $(\boldsymbol{\eta}|\boldsymbol{\Theta} \sim \mathcal{N}(0, \boldsymbol{\Sigma})$. Then we have the conditional distribution with $\boldsymbol{\eta}$ marginalized out equal $(\boldsymbol{y}|\boldsymbol{\Theta}) \sim \mathcal{N}(\boldsymbol{\mu} + \boldsymbol{W}\boldsymbol{\Sigma}\boldsymbol{W}^\top, \sigma^2\boldsymbol{I})$. Furthermore, we have*

$$\boldsymbol{\eta}|\boldsymbol{y}_O, \boldsymbol{\Theta} \sim \mathcal{N}(\hat{\boldsymbol{\mu}}_\eta, \hat{\boldsymbol{C}}_\eta),$$
$$\boldsymbol{y}_M|\boldsymbol{\eta}, \boldsymbol{y}_O, \boldsymbol{\Theta} \sim \mathcal{N}(\boldsymbol{\mu}_M + \boldsymbol{W}_M\boldsymbol{\eta}_i, \sigma^2\boldsymbol{I}),$$

where $\hat{\boldsymbol{C}}_\eta = \left(\boldsymbol{\Sigma}\boldsymbol{W}_O^\top\boldsymbol{W}_O/\sigma^2 + \boldsymbol{I}\right)^{-1}\boldsymbol{\Sigma}$ and $\hat{\boldsymbol{\mu}}_\eta = \hat{\boldsymbol{C}}_\eta\boldsymbol{W}_O^\top(\boldsymbol{y}_O - \boldsymbol{\mu}_O)/\sigma^2$.

Corollary 3.1.3. *For any $\boldsymbol{\Theta}$, the followings are true*

$$\boldsymbol{y}_O|\boldsymbol{\Theta}, \boldsymbol{\mu}_O, \boldsymbol{W}_O \sim \mathcal{N}(\boldsymbol{\mu}_O, \boldsymbol{W}_O\boldsymbol{\Sigma}\boldsymbol{W}_O^\top + \sigma^2\boldsymbol{I}),$$
$$\boldsymbol{y}_M|\boldsymbol{y}_O, \boldsymbol{\Theta}, \boldsymbol{\mu}_O, \boldsymbol{W}_O \sim \mathcal{N}(\hat{\boldsymbol{\mu}}_M, \hat{\boldsymbol{C}}_M), \tag{3.4}$$

where $\hat{\boldsymbol{\mu}}_M = \boldsymbol{\mu}_M + \boldsymbol{W}_M\hat{\boldsymbol{\mu}}_\eta$ and $\hat{\boldsymbol{C}}_M = \boldsymbol{W}_M\hat{\boldsymbol{C}}_\eta\boldsymbol{W}_M^\top + \sigma^2\boldsymbol{I}$.

Proofs are reported in the supplementary material.

Let D_M denote the number of missing features in $\boldsymbol{y}$. Equipped with Proposition 3.1.2 and Corolary 3.1.3, we propose the following three imputation strategies

- Small D_M: sample from (3.4), which requires a Cholesky factorization of $\hat{\boldsymbol{C}}_M$ with a complexity cubic in D_M. We do not recommend the data augmentation technique since it has the potential to harm the mixing of the Gibbs sampler.
- Moderately large D_M: sample via the data augmentation technique provided in Proposition 3.1.2. This allows us to sample with a complexity linear in D_M. In practice we recommend to run a longer Markov chain to compensate the potential worse mixing.
- Large D_M: sample via the data augmentation in the first few steps of the Gibbs sampler and later on fix the value of $\boldsymbol{y}_M$ to its last update. When D_M is large we cannot afford to run a Gibbs sampler with each step having a complexity linear in D_M.

In practice, we treat $D_M \leqslant 50$ as small, $50 < D_M \leqslant 1000$ as moderately large and $D_M > 1000$ as large.

3.2 Simulation Studies

In this section, we compare GEODE to its counterpart PPCA in terms of accuracy, robustness and computational efficiency via simulations. All experiments are conducted in matlab version 2015a on a OS X laptop with a double 3.1 GHz Intel(R) Core(TM) i7 processor. PPCA is fitted using matlab inbuilt function `ppca` under the statistics and machine learning toolbox. This function implements an EM algorithm for PPCA, which is computational friendlier and handles missing data (Roweis,

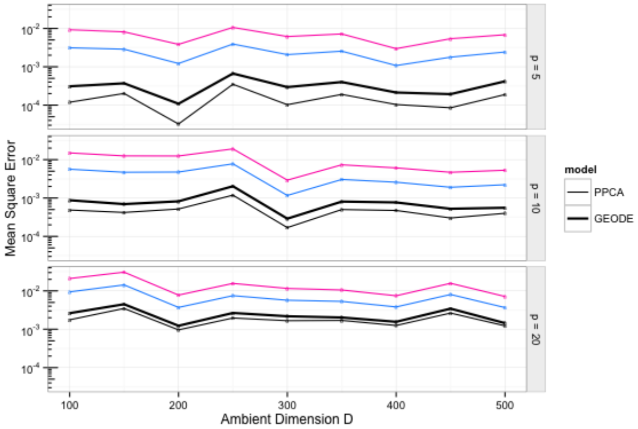

FIGURE 3.1: Comparison of MSE.

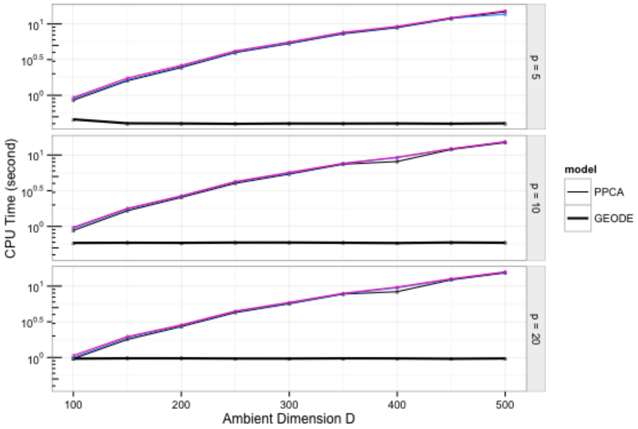

FIGURE 3.2: Comparison of CPU time.

1998; Ilin and Raiko, 2010). All results reported are obtained by averaging over 10 replicated experiments.

Moderately large D without missing data: In this first simulation study we let D vary from 100 to 500 and fix N to be 500. We test both methods on three different intrinsic dimensions, i.e., $p \in \{5, 10, 20\}$. To test the robustness of PPCA to the choices of d, we let d taking values from $\{p, p+5, p+10\}$. We fix $d = 30$ for GEODE since it can automatically learn p. We evaluate the performance of both models in terms of their mean square errors (MSE) in estimating σ^2. The comparison of

MSE of estimating σ^2 under different D and p is presented in Figure 3.1. Results for PPCA are color coded with black denoting $d = p$, blue denoting $d = p + 5$ and purple denoting $d = p + 10$. The thin black lines denote the MSE of PPCA with a correct guess of p, i.e., $d = p$. Though they seem to be consistently better than GEODE, the difference decreases as p increases. However, when incorrectly choosing the d, which always happens in practice, the performance of PPCA (denoted by the colored lines) drops dramatically and is much worse than that of GEODE. The comparison of CPU time of estimating σ^2 under different D and p is reported in Figure 3.2. Results for PPCA are color coded with black denoting $d = p$, blue denoting $d = p + 5$ and purple denoting $d = p + 10$. It seems that the computational cost of PPCA grows exponentially fast in D, while the cost of GEODE grows much slower. In fact, GEODE is so computational efficient that the cost of the preprocessing step is dominated by the Gibbs sampler. This explains why in Figure 3.2 the cost of GEODE seems not grow in D. To check how well GEODE learns the true intrinsic dimension p, we calculate the average probability of j being inside $\mathcal{R}$ for $j = 1, \ldots, d$. These probabilities are visualized in Figure 3.3. It can be easily seen that GEODE is able to provide a very tight guess of p. Hence we can conclude from the simulations that GEODE performs almost as well but slightly worse than the best PPCA can achieve, but with a much smaller computational cost. Moreover, GEODE automatically learns p starting from any crude guess d, while the performance of PPCA is very sensitive to the choice of d.

Moderately large D with missing data: Though the EM algorithm of PPCA offers a straightforward way to impute missing data, it turned out that the existence of a tiny proportion of missingness will explode its computational cost. In this simulation study we fix $N = 500$, and randomly select 25 observations with 5 features missing. We fit PPCA with $d = p$ and run simulations for $D = 100$ and $D = 200$. In estimating σ^2, GEODE generate almost as good results as PPCA, with a CPU time

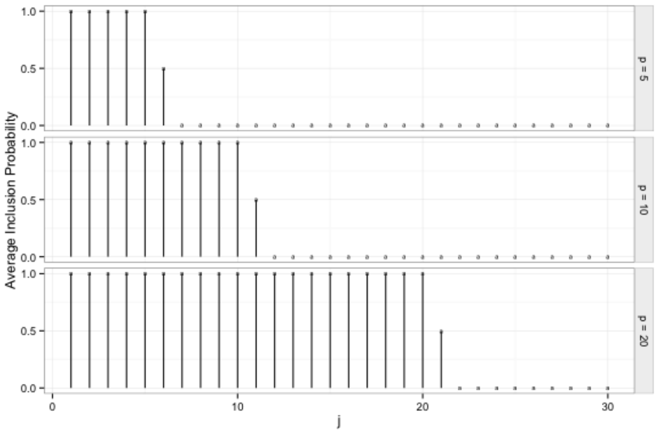

Figure 3.3: Inclusion probabilities for j under different p averaged across all D.

less than 4 seconds for both cases. However, the CPU time of fitting PPCA is 177 seconds for $D = 100$ and 578 seconds for $D = 200$.

Massive D: To evaluate the scalability of GEODE to massive dimensional problem, we redo the previous two experiments on GEODE with D varying from 10^5 to 10^6. Note that $D = 10^6$ is the largest that we can test on the computer due to the storage limit (with $N = 500$ and $D = 10^6$, a single $\boldsymbol{Y}$ takes more than 3 GB storage). For a illustration purpose, we fix $p = 5$. The MSE in estimating σ^2 (top panel) and the CPU time (bottom panel) are reported in Figure 3.4. It is obvious that GEODE remains computationally feasible even when $D = 10^6$, while providing very good performance.

3.3 Mixture of GEODE

Mixture of PPCA (Tipping and Bishop, 1999a) extends PPCA to be able to characterize non-Gaussian data. However, it inherits the computational drawbacks of PPCA. Mixture of factor analyzers model (MFA) frees the isotropic error constrain of mixture of PPCA, but the corresponding EM algorithm (Ghahramani and Hinton, 1996) suffers similar computational bottleneck. Bayesian MFA is a straightforward

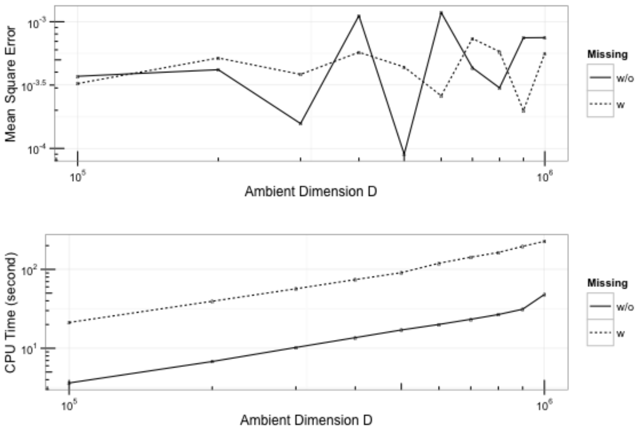

Figure 3.4: MSE and the CPU time of GEODE.

Bayesian implementation in small dimensional problems (Diebolt and Robert, 1994; Richardson and Green, 1997), but faces problems in scaling beyond a few 100 dimensions. Inspired by the empirical Bayes idea of GEODE in linear cases, we propose to learn and fix a multiscale set of potential principal component spaces in a first step. In the second step, we mix across these potential principal component spaces with respect to their likelihood in a Bayesian paradigm. The ability to learn the intrinsic dimension p's (we allow different spaces having different dimensions) and the ability to characterize the uncertainty are both inherited from the linear cases.

3.3.1 Multiscale Principal Axes

To aquire a multiscale set of principal directions, we first adopt METIS (Karypis and Kumar, 1998) to obtain a dyadic clustering tree of the dataset. This is partly motivated by the compressive sensing technique developed by Allard et al. (2012), which efficiently compresses the data by locally finding the best linear subspaces to approximate these dyadic clusters. A dyadic clustering tree of the data $\{\boldsymbol{y}_i\}_{i=1}^N$ is defined as follows

Definition 3.3.1. *With $s = 0, \dots, L$ denoting the scale index and $h = 1, \dots, 2^s$ denoting the node index within scale s, a level-L* ***dyadic clustering tree*** *of $\{\boldsymbol{y}_i\}_{i=1}^N$*

is a family of index sets $\mathcal{A}_{sh} \subseteq \{1, \dots, N\}$ *such that*

- *for every s,* $\bigcup_{h=1}^{2^s} \mathcal{A}_{sh} = \{1, \dots, N\}$*;*
- *for* $s \leqslant s'$ *and* $1 \leqslant h' \leqslant 2^{s'}$*, either* $\mathcal{A}_{s'h'} \subseteq \mathcal{A}_{sh}$ *or* $\mathcal{A}_{s'h'} \cap \mathcal{A}_{sh} = \varnothing$*;*
- *for* $s < s'$ *and* $1 \leqslant h' \leqslant 2^{s'}$*, there exists a unique* $h = 1, 2, \dots, 2^s$ *such that* $\mathcal{A}_{s'h'} \subseteq \mathcal{A}_{sh}$*.*

The tree is denoted by $\{\mathcal{A}_{sh}\}_L$*.*

METIS generates the tree-structure clustering by partitioning a weighted graph constructed from the data. Following the suggestion by Allard et al. (2012), we add an edge between each data point and its k nearest neighbors and set the weight between any $\boldsymbol{y}_i$ and $\boldsymbol{y}_j$ to be $e^{-\|\boldsymbol{y}_i - \boldsymbol{y}_j\|_2^2/\delta}$. δ is chosen adaptively at each point $\boldsymbol{y}_i$ as the distance between $\boldsymbol{y}_i$ and its $\lfloor k/2 \rfloor$ nearest neighbor. In practice we fix k to be 30 and constrain the leaf size $|\mathcal{A}_{Lh}|$ to be greater than 10, for $h = 1, \dots, 2^L$. The depth of the tree L depends on the sample size N and is automatically decided by METIS. Performance of METIS are illustrated in the supplementary material through multiple simulations.

Equiped with a leve-L dyadic clustering tree $\{\mathcal{A}_{sh}\}_L$, the corresponding multiscale principal axes is defined as follows

Definition 3.3.2. *The* ***multiscale principal axes*** *of* $\{\boldsymbol{y}_i\}_{i=1}^N$ *with respect to a leve-L dyadic clustering tree* $\{\mathcal{A}_{sh}\}_L$ *is defined as a family of centroids* $\boldsymbol{mu}_{sh} \in \Re^D$ *and a family of othorgonal matrices* $\boldsymbol{W}_{sh} \in \Re^{D \times d}$ *such that for all s and h*

- $\boldsymbol{\mu}_{sh} = \frac{1}{|\mathcal{A}_{sh}|} \sum_{i \in \mathcal{A}_{sh}} \boldsymbol{y}_i$*;*
- $\boldsymbol{W}_{sh} = \boldsymbol{U}_{sh}$*, where the column vectors of* $\boldsymbol{U}_{sh}$ *are the leading d right singular vectors of* $\boldsymbol{Y}_{sh}$ *and* $\boldsymbol{Y}_{sh}$ *is a* $|\mathcal{A}_{sh}| \times D$ *matrix with each row representing a single demeaned observation from* $\mathcal{A}_{sh}$*.*

The multiscale principal axes is denoted as $\{\boldsymbol{\mu}_{sh}, \boldsymbol{W}_{sh}\}_L$.

In practice we use fast rank-d SVD to compute $\boldsymbol{W}_{sh}$.

By Theorem (A.1.1), the principal space defined by $\boldsymbol{\mu}_{sh}$ and $\boldsymbol{W}_{sh}$ is the "optimal" in the sense that the posterior mode of local GOEDE fitted on the subset $\mathcal{A}_{sh}$ of the data is maximized. Then, intuitively, a probabilistic mixture accross these locally "optimal" spaces will give a good nonlinear estimation of the high diemensional data distribution. We learn a multiscale principal axes instead of a single scale one to allow the model adapting to different local smoothness by mixing across fine scales and coarse scales.

3.3.2 Model Formulation

Equiped with the multiscale principal axes $\{\boldsymbol{\mu}_{sh}, \boldsymbol{W}_{sh}\}_L$, the mixture of GEODE model (mGEODE) is given by

$$\boldsymbol{y} \sim \sum_{sh} \pi_{sh} \mathcal{N}(\boldsymbol{\mu}_{sh}, \boldsymbol{C}_{sh}) \tag{3.5}$$

where $\boldsymbol{C}_{sh} = \boldsymbol{W}_{sh} \boldsymbol{\Sigma}_{sh} \boldsymbol{W}_{sh}^\top + \sigma_s^2 \boldsymbol{I}$ and $\boldsymbol{\Sigma}_{sh}$ is a $d \times d$ positive diagonal matrix, for $s = 0, \ldots, L$ and $h = 1, \ldots, 2^s$. For all s and h, $\boldsymbol{\Sigma}_{sh}$ and σ_s^2 are given the same prior distribution as in GEODE. We assume isotropic error variance σ_s^2 for each scale s to enable clusters from the same scale share information between each other.

We then finish the formulation of mGEODE by choosing a prior for the multiscale mixing weights $\pi_{s,h}$. This prior should be structured to allow adaptive learning of the appropriate tradeoff between coarse and fine scales. Heavily favoring coarse scales may lead to reduced variance but also high bias if the coarse scale approximation is not accurate. High weights on fine scales may lead to low bias but high variance due to limited sample size in each fine resolution component. With this motivation, Canale and Dunson (2014) proposed a multiresolution stick-breaking process generalizing usual “flat” stick-breaking (Sethuraman, 1994). In particular, let

$$S_{sh} \sim \mathrm{Be}(1, a_S),\ R_{sh} \sim \mathrm{Be}(b_R, b_R) \tag{3.6}$$

with S_{sh} denoting the probability that the observation stops at node (s, h) of a binary tree and R_{sh} denoting the probability that the observation moves down to the right from node (s, h) conditioning on not stopping at node (s, h). Hence

$$\pi_{sh} = S_{sh} \prod_{r<s} (1 - S_{r\ g_{shr}}) T_{shr} \tag{3.7}$$

where $g_{shr} = \lceil h/2^{s-r} \rceil$ denotes the ancestors of node (s, h) at scale r, $T_{shr} = R_{r\ g_{shr}}$ if node $(r + 1, g_{sh(r+1)})$ is the right daughter of node$(r + 1, g_{shr})$, otherwise $T_{shr} = 1 - R_{r\ g_{shr}}$. Canale and Dunson (2014) showed that $\sum_{s=0}^{\infty} \sum_{h=1}^{2^s} \pi_{sh} = 1$ almost surely for any $a_S, b_R > 0$. This result makes the defined weights a proper set of multiscale mixing weights. As a_S increases, finer scales are favored, resulting in a highly non-Gaussian density.

In practice, we only consider a truncated finite-depth multiscale mixture with depth being L. Let $\{\tilde{\pi}_{sh}\}_{s \leqslant L}$ denote the truncated weights, which are identical to $\{\pi_{sh}\}$ except that the stopping probabilities at scale L are set to be equal to one to ensure $\sum_{s=1}^{L} \sum_{h=1}^{2^s} \tilde{\pi}_{sh} = 1$.

3.3.3 Posterior Computation

The posterior sampling for the mGEODE is almost identical to the GEODE, except for the newly introduced membership variables (s_i, h_i). The conditional posterior of these variables is given by

$$p(s_i = s, h_i = h) \propto \pi_{sh} \mathcal{N}_D(\boldsymbol{\mu}_{sh}, \boldsymbol{W}_{sh} \boldsymbol{\Sigma}_{sh} \boldsymbol{W}_{sh}^\top + \sigma_s^2 \boldsymbol{I}).$$

The conditional posteriors of S_{sh} and R_{sh} are given by

$$S_{sh} \sim \mathrm{Beta}(1 + n_{sh}, a_S + v_{sh} - n_{sh}),$$

$$R_{sh} \sim \mathrm{Beta}(b_R + r_{sh}, b_R + v_{sh} - n_{sh} - r_{sh}),$$

where v_{sh} is the number of observations passing through node (s, h), n_{sh} is the number of observations stopping at node (s, h), and r_{sh} is the number of observations that continue to the right after passing through node (s, h).

The rest steps of the Gibbs sampler of the mGEODE are neglected for succinctness since it is very similar to the GEODE. Details can be found in the supplementary material.

3.4 Computational Aspects

GEODE Letting T denote the total number of Gibbs sampler interations, the computational cost can be split as follows.

Construction of principal axes: The complexity of fast rank-d SVD is $\mathcal{O}(NDd)$.

Comstruction of sufficient statistics: The complexity of computing $A_i = \tilde{\boldsymbol{y}}_i^\top \tilde{\boldsymbol{y}}_i$ for all i is $\mathcal{O}(ND)$ and the complexity of computing $\boldsymbol{Z}_i = \boldsymbol{W}^\top \tilde{\boldsymbol{y}}_i$ for all i is $\mathcal{O}(NDd)$. Hence the overall complexity is $\mathcal{O}(NDd)$.

Gibbs sampler: The complexity of the sampler is dominated by updating σ^2 and updating $\boldsymbol{u}$, whose complexities are both $\mathcal{O}(NTd)$.

Hence the overall complexity of the GEODE is $\mathcal{O}(NDd + NTd)$.

mGEODE Letting K denote the number of nearest neighbours in constructing the weighted graph, the computational cost can be split as follows.

Construction of weighted graph: The complexity of ANN in finding K nearest neighbours is $\mathcal{O}(DN \log N)$ (Arya et al., 1998). The complexity of computing the weights for the graph is $O(KND)$.

Graph partition: The complexity of METIS in partition the data into a level-L dyadic clustering tree is $\mathcal{O}(KN \log N)$.

Construction of multiscale principal axes: For each node (s, h), the complexity of applying the fast rank-d SVD is $\mathcal{O}(|\mathcal{A}_{sh}|Dd)$. We have $|\mathcal{A}_{sh}| = \mathcal{O}(2^{-s}N)$ and there are 2^L such $\mathcal{A}_{sh}$'s. Summing over all of them with $L < \log_2 N$ we obtain a tatal cost of $\mathcal{O}(N \log NDd)$.

Construction of sufficient statistics: Same to the linear cases, for each node (s, h), the complexity is $\mathcal{O}(|\mathcal{A}_{sh}|Dd)$. Similar to deriving the compelxity for multiple principal axes, the complexity for constructing the sufficient statistics is $\mathcal{O}(N \log NDd)$.

Gibbs sampler: The complexity of the sampler is dominated by updating (s_i, h_i) for all i and updating $\boldsymbol{u}_{sh}$ for all nodes, whose complexities are both $\mathcal{O}(NT2^Ld)$.

Hence the overall complexity of the mGEODE is $\mathcal{O}(N \log NDd + NT2^Ld)$.

Moreover, the Gibbs samplers in both linear and nonlinear cases are converging very fast with superb mixing. Hence in practice we fix the T to be 1000, with the number of burn-ins fixed to be 500. No thinning is needed. The posterior diagnostic results for the simulation studies in linear cases can be found in the supplementary materials.

3.5 Application

mGEODE is demonstrated first in a multivariate response regression application and then in a supervised classification problem. In both applications, $d = 20$. Increasing d moderately had essentially no impact on the results.

3.5.1 Image Inpainting

The Frey faces data (Roweis et al., 2002) contains 1965 20×28 video frames of a single face with different expressions. Conducting the same experiment as done by Titsias and Lawrence (2010), the data set is randomly split into 1000 training images and 965 testing images with a random half of the pixels missing. mGEODE was trained for less than 2 minutes, and reconstruction (prediction) of all 965 testing images was done in less than 10 minutes. The mean absolute reconstruction error of mGEODE is 7.04, which outperforms the error of 7.40 reported by Titsias and

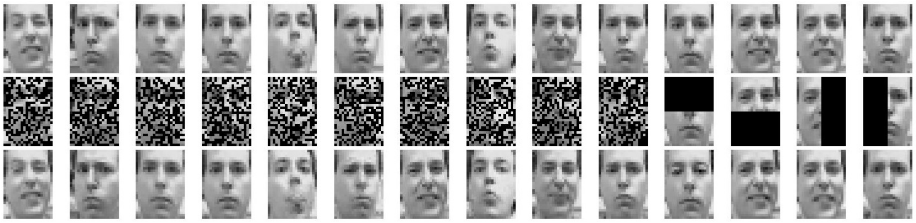
FIGURE 3.5: Performance of mGEODE on image inpainting.

Lawrence (2010). 10 randomly selected reconstructions are shown on the left in Figure 3.5, with 4 manually designed missingness cases shown on the right. The first row shows the original images, second row shows the images with pixels missing, and the third row shows the reconstructed images. mGEODE also outperforms the results shown by Adams et al. (2010) by looking at their visualized results. It is also noted that Adams et al. (2010) reported a few hours of computational time in reconstructing 100 images based on 1865 training images.

3.5.2 Digit Classification

mGEODE was used as a probabilistic classifier for the MNIST handwritten data, which contains 70000 28 × 28 grey scale handwritten digits images. First, one mGEODE was trained for each of the 10 digits over a total of 60000 training data for around 90 minutes. Then within each iteration of the Gibbs sampler, the 10 mGEODE's worked in a Naive Bayes way and generated a likely class. The "voting" process took 7 minutes for 10000 testing images and the mode of these votes were computed as the classification results. The classification error was 2.32%.

4

Parallelizable Composite Posterior

4.1 Bayesian Mosaic

We start by introducing notation that will be used throughout the paper. Before presenting the formal definition, we will first motivate the proposed method by introducing a class of multivariate latent Gaussian models and describing computational issues DA-MCMC algorithms encounter in fitting these models. After defining *Bayesian mosaic*, we present a sampling algorithm and a post-processing method to handle parameter constraints. We end the section by generalizing *Bayesian mosaic* to dependent data.

4.1.1 Notations

We represent vectors by lower case letters and matrices by capital letters, both in a boldface font. Unless otherwise stated, all vectors will be column vectors. We use $\mathbb{R}$ to denote the set of all real numbers, $\mathbb{N}_0$ the nonnegative integers, $\mathbb{N}_1$ the positive integers and $\|\cdot\|$ the Euclidean norm. For $d \in \mathbb{N}_1$, $\boldsymbol{\theta} \in \mathbb{R}^d$ and $\delta > 0$, we define a radius-δ ball of $\boldsymbol{\theta}$ at $\boldsymbol{\theta}_0$ as $\mathcal{B}_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0, \delta) = \{\boldsymbol{\theta} : \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| < \delta\}$. For succinctness, we denote the multiple integral of a multivariate function $g(\boldsymbol{y})$ as

$$\int g(\boldsymbol{y})\mathrm{d}\boldsymbol{y} = \int \cdots \int g(y_1, \ldots, y_p)\mathrm{d}y_1 \cdots \mathrm{d}y_p.$$

We will always use f to denote a density function and ℓ to denote a log-density function. The density function will be presented in a conditional style, e.g., $f(\boldsymbol{y}|\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ are the model parameters. Given that $\boldsymbol{y}$ follows some distribution $P_{\boldsymbol{\theta}}$ with a density function $f(\boldsymbol{y}|\boldsymbol{\theta})$, we use $\mathbb{E}_{\boldsymbol{\theta}} g(\boldsymbol{y})$ to denote the expectation of $g(\boldsymbol{y})$. More specifically,

$$\mathbb{E}_{\boldsymbol{\theta}} g(\boldsymbol{y}) = \int g(\boldsymbol{y}) f(\boldsymbol{y}|\boldsymbol{\theta})\mathrm{d}\boldsymbol{y}.$$

We use $\phi(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ to denote the density function of a Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$, and use $\Phi\left(U|\boldsymbol{\mu}, \boldsymbol{\Sigma}\right)$ to denote its cdf function:

$$\begin{aligned} \phi(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) =& |2\pi\boldsymbol{\Sigma}|^{-1/2} \exp\big\{ -(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})/2 \big\}, \\ \Phi\left(U|\boldsymbol{\mu}, \boldsymbol{\Sigma}\right) =& \int_U \phi(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})\mathrm{d}\boldsymbol{x}. \end{aligned}$$

For better representation of the higher-order remainder of the Taylor expansion for multivariate functions, we adopt the notations of Folland (2005). For any $d \in \mathbb{N}_1$, a d-dimensional *multi-index* $\boldsymbol{\alpha}$ for $\boldsymbol{x} = (x_1, \ldots, x_d)^\top \in \mathbb{R}^d$ is defined as a d-tuple of nonnegative integers, i.e., $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_d)$, where $\alpha_j \in \mathbb{N}_0$ for $j = 1, \ldots, d$. We further define

$$|\boldsymbol{\alpha}| = \sum_{j=1}^d \alpha_j, \quad \boldsymbol{\alpha}! = \prod_{j=1}^d \alpha_j!, \quad \boldsymbol{x}^{\boldsymbol{\alpha}} = \prod_{j=1}^d x_j^{\alpha_j}, \quad \partial^{\boldsymbol{\alpha}} g(\boldsymbol{x}) = \frac{\partial^{|\boldsymbol{\alpha}|} g(\boldsymbol{x})}{\partial x_1^{\alpha_1} \cdots \partial x_d^{\alpha_d}}.$$

We will also use the following vector calculus notation to ease our representation of the gradient vector and the Hessian matrix. Considering two vectors $\boldsymbol{\eta} = \left(\eta_1, \ldots, \eta_{d_\eta}\right)^\top$, $\boldsymbol{\zeta} = \left(\zeta_1, \ldots, \zeta_{d_\zeta}\right)^\top$ and some function $g(\boldsymbol{\eta}, \boldsymbol{\zeta})$, we denote the gradient of $f(\boldsymbol{\eta}, \boldsymbol{\zeta})$ w.r.t. $\boldsymbol{\eta}$ as

$$\nabla_{\boldsymbol{\eta}} g(\boldsymbol{\eta}, \boldsymbol{\zeta}) = \left[\frac{\partial g(\boldsymbol{\eta}, \boldsymbol{\zeta})}{\partial \eta_1}, \ldots, \frac{\partial g(\boldsymbol{\eta}, \boldsymbol{\zeta})}{\partial \eta_{d_\eta}}\right]^\top,$$

and the gradient of $g(\boldsymbol{\eta}, \boldsymbol{\zeta})$ w.r.t. $\boldsymbol{\eta}$ and $\boldsymbol{\zeta}$ as

$$\nabla_{\boldsymbol{\eta},\boldsymbol{\zeta}} g(\boldsymbol{\eta}, \boldsymbol{\zeta}) = \left[\frac{\partial g(\boldsymbol{\eta}, \boldsymbol{\zeta})}{\partial \eta_1}, \ldots, \frac{\partial g(\boldsymbol{\eta}, \boldsymbol{\zeta})}{\partial \eta_{d_\eta}}, \frac{\partial g(\boldsymbol{\eta}, \boldsymbol{\zeta})}{\partial \zeta_1}, \ldots, \frac{\partial g(\boldsymbol{\eta}, \boldsymbol{\zeta})}{\partial \zeta_{d_\zeta}} \right]^\top .$$

We further define

$$\nabla_{\boldsymbol{\zeta}} \nabla_{\boldsymbol{\eta}} g(\boldsymbol{\eta}, \boldsymbol{\zeta}) = \begin{bmatrix} \frac{\partial^2 g(\boldsymbol{\eta},\boldsymbol{\zeta})}{\partial \eta_1 \partial \zeta_1} & \cdots & \frac{\partial^2 g(\boldsymbol{\eta},\boldsymbol{\zeta})}{\partial \eta_1 \partial \zeta_{d_\zeta}} \\ \vdots & & \vdots \\ \frac{\partial^2 g(\boldsymbol{\eta},\boldsymbol{\zeta})}{\partial \eta_{d_\eta} \partial \zeta_1} & \cdots & \frac{\partial^2 g(\boldsymbol{\eta},\boldsymbol{\zeta})}{\partial \eta_{d_\eta} \partial \zeta_{d_\zeta}} \end{bmatrix} .$$

We suppress $\nabla_{\boldsymbol{\eta}} \nabla_{\boldsymbol{\eta}}$ to $\nabla^2_{\boldsymbol{\eta}}$. Note that $\nabla^2_{\boldsymbol{\eta},\boldsymbol{\zeta}} g(\boldsymbol{\eta}, \boldsymbol{\zeta})$ is the Hessian of $g(\boldsymbol{\eta}, \boldsymbol{\zeta})$.

Consider $p \in \mathbb{N}_1$ and a sequence indexed by two subscripts $\{x_{st}\}$ where $1 \leqslant s < t \leqslant p$. Whenever we write $x_{12}, \ldots, x_{(p-1)p}$, we mean

$$x_{12}, \ldots, x_{1p}, x_{23}, \ldots, x_{2p}, \ldots, x_{(p-2)(p-1)}, x_{(p-2)p}, x_{(p-1)p}$$

where the elements are ordered in a row-major manner.

4.1.2 Multivariate Latent Gaussian Model

Suppose $p \in \mathbb{N}_1$, $n \in \mathbb{N}_1$ and $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ are i.i.d. p-dimensional observations from some multivariate latent Gaussian model. Letting $\boldsymbol{y} = (y_1, \ldots, y_p)^\top \in \mathcal{Y} \subseteq \mathbb{R}^p$ and introducing $\boldsymbol{x} = (x_1, \ldots, x_p)^\top \in \mathbb{R}^p$, the density function of the model can be written as

$$f(\boldsymbol{y}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \int \prod_{j=1}^{p} h_j(y_j|x_j) \phi(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \mathrm{d}\boldsymbol{x}, \tag{4.1}$$

where h_j's are univariate density functions, $\boldsymbol{\Sigma} = \{\sigma_{st}\}$ is a $p \times p$ positive definite matrix and $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)^\top \in \mathbb{R}^p$. We refer to h_j's as link densities.

The integral in (4.1) usually does not have an analytical solution, and accurate numerical integration is infeasible even for moderately large p. Hence, fully Bayesian inference is usually based on a DA-MCMC algorithm, where $\boldsymbol{x}$ is augmented and sampled together with the model parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$.

If we let h_j's be discrete data densities, then (4.1) provides a rich class of multivariate discrete data models. However, real world discrete datasets are often severely imbalanced. Taking online advertising as an example, the click through rate of a certain link is usually very close to 0. Supposing that one wants to fit a logistic regression to predict the probability of a certain user's clicking the link, only a tiny faction of the responses will be 1.

Unfortunately, DA-MCMC has a provably slow mixing rate in imbalanced discrete data problems. Considering an intercept-only probit model and assuming that the data are infinite imbalanced, Johndrow et al. (2016) have shown that the step size of DA-MCMC is roughly $O(\frac{1}{\sqrt{n}})$, while the width of the high probability region of the posterior is roughly $O(\frac{1}{\log n})$. For large n, the step size will become much smaller than the width of the high probability bulk, causing extreme slow mixing. Moreover, this mismatch will become worse as n grows, and presents huge practical problems in broad settings.

Another drawback of DA-MCMC is its poor scalability to large sample size. The per-iteration computational complexity is at least $O(n)$ due to the need to sample an augmented $\boldsymbol{x}_i$ for each $\boldsymbol{y}_i$. Moreover, the number of model parameters in (4.1) increases quadratically as the data dimensionality grows.

One way to bypass DA-MCMC is to evaluate the integral in (4.1) directly via deterministic numerical integration methods. Unfortunately, these methods are only feasible for small p. A potential solution is to approximate Bayesian inference by using a composite likelihood whose individual components are low-dimensional conditional or marginal densities that can be numerically evaluated (Pauli et al., 2011). *Bayesian mosaic* is partially motivated by this idea.

Suppose $\boldsymbol{y} = (y_1, \ldots, y_p)^\top$ follows the multivariate latent Gaussian model whose density function is defined in (4.1). One can easily prove the following:

i) For $j = 1, \ldots, p$, the univariate marginal density for y_j is

$$f_{jj}(y_j|\mu_j, \sigma_{jj}) = \int h_j(y_j|x_j)\phi(x_j|\mu_j, \sigma_{jj})\mathrm{d}x_j,$$

ii) For $1 \leqslant s < t \leqslant p$, the bivariate marginal density for y_s and y_t is

$$\begin{aligned} & f_{st}(y_s, y_t|\sigma_{st}, \mu_s, \sigma_{ss}, \mu_t, \sigma_{tt}) \\ = & \int h_s(y_s|x_1)h_t(y_t|x_2)\phi\left(\left[\begin{smallmatrix} x_1 \\ x_2 \end{smallmatrix}\right]\middle|\left[\begin{smallmatrix} \mu_s \\ \mu_t \end{smallmatrix}\right], \left[\begin{smallmatrix} \sigma_{ss} & \sigma_{st} \\ \sigma_{ts} & \sigma_{tt} \end{smallmatrix}\right]\right)\mathrm{d}\left[\begin{smallmatrix} x_1 \\ x_2 \end{smallmatrix}\right]. \end{aligned}$$

There is a rich literature on numerical integration methods for univariate and bivariate functions. Hence f_{jj}'s and f_{st}'s can be efficiently evaluated. In fact, many composite likelihood methods have been using these lower-dimensional densities as individual components due to the fact that they are computationally easier to work with (Cox and Reid, 2004).

Consider the following composite log-likelihood which consists of only univariate marginal densities:

$$Q_n(\boldsymbol{\mu}, \sigma_{11}, \ldots, \sigma_{pp}) = \sum_{i}^{n}\sum_{j=1}^{p} \ell_{jj}(y_{ij}|\mu_j, \sigma_{jj}),$$

where $\ell_{jj}(y_j|\mu_j, \sigma_{jj}) = \log f_{jj}(y_j|\mu_j, \sigma_{jj})$ for $j = 1, \ldots, p$. One can construct the following posterior distribution:

$$\pi_n^*(\boldsymbol{\mu}, \sigma_{11}, \ldots, \sigma_{pp}) \propto \exp\left\{Q_n(\boldsymbol{\mu}, \sigma_{11}, \ldots, \sigma_{pp})\right\}\pi(\boldsymbol{\mu}, \sigma_{11}, \ldots, \sigma_{pp}),$$

given prior $\pi(\boldsymbol{\mu}, \sigma_{11}, \ldots, \sigma_{pp})$. If we assume prior independence, so that $\pi(\boldsymbol{\mu}, \sigma_{11}, \ldots, \sigma_{pp}) = \prod_{j=1}^{p} \pi_{jj}(\mu_j, \sigma_{jj})$, and let

$$\pi_{n,jj}^*(\mu_j, \sigma_{jj}) \propto \exp\left\{\sum_{i}^{n} \ell_{jj}(y_{ij}|\mu_j, \sigma_{jj})\right\}\pi_{jj}(\mu_j, \sigma_{jj}),$$

it can be shown that

$$\pi_n^*(\boldsymbol{\mu}, \sigma_{11}, \ldots, \sigma_{pp}) = \prod_{j=1}^{p} \pi_{n,jj}^*(\mu_j, \sigma_{jj}). \tag{4.2}$$

We have constructed a surrogate posterior distribution for $\boldsymbol{\mu}$ and σ_{jj}'s. These are the parameters that characterize the univariate marginal distributions of the data. The factorized form of the composite likelihood and prior independence induces posterior independence in (μ_j, σ_{jj})'s. Therefore, sampling from (4.2) can be split into independently sampling from each $\pi_{n,jj}^*(\mu_j, \sigma_{jj})$.

To complete our surrogate posterior distribution, we need some conditional distribution of σ_{st}'s given $\boldsymbol{\mu}$ and σ_{jj}'s. Consider the following composite log-likelihood:

$$L_n(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{i}^{n} \sum_{s<t}^{p} \ell_{st}(y_{is}, y_{it} | \sigma_{st}, \mu_s, \sigma_{ss}, \mu_t, \sigma_{tt}),$$

where $\ell_{st}(y_s, y_t | \sigma_{st}, \mu_s, \sigma_{ss}, \mu_t, \sigma_{tt}) = \log f_{st}(y_s, y_t | \sigma_{st}, \mu_s, \sigma_{ss}, \mu_t, \sigma_{tt})$ for $1 \leqslant s < t \leqslant p$. This time we will assume prior conditional independence, i.e., the prior density takes the following factorized form:

$$\pi(\sigma_{12}, \ldots, \sigma_{(p-1)p} | \boldsymbol{\mu}, \sigma_{11}, \ldots, \sigma_{pp}) = \prod_{1 \leqslant s<t \leqslant p}^{p} \pi_{st}(\sigma_{st} | \mu_s, \sigma_{ss}, \mu_t, \sigma_{tt}).$$

Letting

$$\begin{aligned} &\pi_{n,st}^*(\sigma_{st} | \mu_s, \sigma_{ss}, \mu_t, \sigma_{tt}) \\ &\propto \exp\left\{ \sum_{i}^{n} \ell_{st}(y_{is}, y_{it} | \sigma_{st}, \mu_s, \sigma_{ss}, \mu_t, \sigma_{tt}) \right\} \pi_{st}(\sigma_{st} | \mu_s, \sigma_{ss}, \mu_t, \sigma_{tt}), \end{aligned}$$

we can construct the following conditional posterior density:

$$\pi_n^*(\sigma_{12}, \ldots, \sigma_{(p-1)p} | \boldsymbol{\mu}, \sigma_{11}, \ldots, \sigma_{pp}) = \prod_{1 \leqslant s<t \leqslant p}^{p} \pi_{n,st}^*(\sigma_{st} | \mu_s, \sigma_{ss}, \mu_t, \sigma_{tt}). \tag{4.3}$$

Similarly, we have posterior conditional independence in σ_{st}'s given $\boldsymbol{\mu}$ and σ_{jj}'s. Therefore, sampling from (4.3) can be split into independently sampling from each $\pi^*_{n,st}(\sigma_{st}|\mu_s, \sigma_{ss}, \mu_t, \sigma_{tt})$.

Combining (4.2) and (4.3) we construct the following surrogate posterior density:

$$\pi^*_n(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \pi^*_n(\sigma_{12}, \ldots, \sigma_{(p-1)p}|\boldsymbol{\mu}, \sigma_{11}, \ldots, \sigma_{pp})\pi^*_n(\boldsymbol{\mu}, \sigma_{11}, \ldots, \sigma_{pp}).$$

To summarize, we have proposed a surrogate posterior distribution which is a multiplication of component posteriors. These component posteriors are based on either univariate or bivariate marginal densities. Sampling from this posterior can be done via a composite sampling strategy that contains two steps. In the first step, we sample the parameters that characterize the univariate marginal densities ($\boldsymbol{\mu}$ and σ_{jj}'s). In the second step, we plug the previous samples into the conditional densities and sample those parameters that characterize the pairwise relationship (σ_{st}'s). The computation of both steps can be easily parallelized due to the sparse posterior dependence structure. We term $\pi^*_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ as a *Bayesian mosaic* posterior under model (4.1). A formal definition will follow.

4.1.3 Definition of Bayesian Mosaic

It can be seen that the independence structure in (4.2) relies on the fact that univariate marginal distributions do not share parameters, and that the conditional independence structure in (4.3) requires that the parameters characterizing the pairwise relationships (σ_{st}'s) only appear in one bivariate marginal distribution. We term the class of data distributions that satisfy the above conditions as *mosaic-type*. Below is a formal definition.

Definition 4.1.1. *Suppose* $p \in \mathbb{N}_1$ *and* $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ *are i.i.d.* p*-dimensional data vectors from distribution* $P_{\boldsymbol{\theta}}$ *with density function* $f(\boldsymbol{y}|\boldsymbol{\theta})$. *Let* $\boldsymbol{\theta}_{st}$, $1 \leqslant s \leqslant t \leqslant p$, *be* *non-overlapping sub-vectors of* $\boldsymbol{\theta}$ *such that*

$$\boldsymbol{\theta} = \left[\boldsymbol{\theta}_{12}^\top, \ldots, \boldsymbol{\theta}_{(p-1)p}^\top, \boldsymbol{\theta}_{11}^\top, \ldots, \boldsymbol{\theta}_{pp}^\top\right]^\top,$$

then the data distribution $P_{\boldsymbol{\theta}}$ *is mosaic-type if there exists a collection of density functions* f_{st} *for* $1 \leqslant s \leqslant t \leqslant p$ *such that*

i) for $j = 1, \ldots, p$*, the density of the univariate marginal distribution for dimension* j *is*

$$f_{jj}(y_j|\boldsymbol{\theta}_{jj}).$$

ii) for $1 \leqslant t < s \leqslant p$*, the density of the bivariate marginal data distribution for dimension* s *and* t *is*

$$f_{st}(y_s, y_t|\boldsymbol{\theta}_{st}, \boldsymbol{\theta}_{ss}, \boldsymbol{\theta}_{tt}).$$

We term $\boldsymbol{\theta}_{jj}$'s as *knots* since they are shared among multiple bivariate marginal distributions. We term $\boldsymbol{\theta}_{st}$'s as *tiles* since they only appear in one bivariate marginal distribution. In the multivariate latent Gaussian example,

$$\boldsymbol{\theta}_{jj} = \left[\begin{smallmatrix} \mu_j \\ \sigma_{jj} \end{smallmatrix}\right], \quad \boldsymbol{\theta}_{st} = \sigma_{st}.$$

We will show that Definition 4.1.1 provides a rich class of models later in §4.2.1. Although we require $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ to be independent for now, to ease our analysis of asymptotic properties, in practice this requirement can be relaxed. We provide a more general definition in §4.1.6.

Before defining *Bayesian mosaic*, we will first introduce some notation. For $j = 1, \ldots, p$, define

$$\ell_{jj}(\boldsymbol{\theta}_{jj}, y_j) = \log f_{jj}(y_j|\boldsymbol{\theta}_{jj}), \quad Q_{n,j}(\boldsymbol{\theta}_{jj}) = \sum_{i=1}^{n} \ell_{jj}(\boldsymbol{\theta}_{jj}, y_{ij}).$$

For $1 \leqslant s < t \leqslant p$, define

$$\ell_{st}(\boldsymbol{\theta}_{st}, \boldsymbol{\theta}_{ss}, \boldsymbol{\theta}_{tt}, y_s, y_t) = \log f_{st}(y_s, y_t | \boldsymbol{\theta}_{st}, \boldsymbol{\theta}_{ss}, \boldsymbol{\theta}_{tt}),$$

$$L_{n,st}(\boldsymbol{\theta}_{st}, \boldsymbol{\theta}_{ss}, \boldsymbol{\theta}_{tt}) = \sum_{i=1}^{n} \ell_{st}(\boldsymbol{\theta}_{st}, \boldsymbol{\theta}_{ss}, \boldsymbol{\theta}_{tt}, y_{is}, y_{it}).$$

The formal definition of *Bayesian mosaic* is given below.

Definition 4.1.2. *Under the setup of Definition 4.1.1 and considering prior densities* $\pi_{jj}(\boldsymbol{\theta}_{jj})$ *for* $j = 1, \ldots, p$ *and* $\pi_{st}(\boldsymbol{\theta}_{st} | \boldsymbol{\theta}_{ss}, \boldsymbol{\theta}_{tt})$ *for* $1 \leqslant t < s \leqslant p$, *we introduce the following:*

i) For $j = 1, \ldots, p$, *the knot marginal for* $\boldsymbol{\theta}_{jj}$ *is*

$$\kappa_{n,j}(\boldsymbol{\theta}_{jj}) \propto e^{Q_{n,j}(\boldsymbol{\theta}_{jj})} \pi_{jj}(\boldsymbol{\theta}_{jj}).$$

ii) For $1 \leqslant t < s \leqslant p$, *the tile conditional for* $\boldsymbol{\theta}_{st}$ *given* $\boldsymbol{\theta}_{ss}$ *and* $\boldsymbol{\theta}_{tt}$ *is*

$$\tau_{n,st}(\boldsymbol{\theta}_{st} | \boldsymbol{\theta}_{ss}, \boldsymbol{\theta}_{tt}) \propto e^{L_{n,st}(\boldsymbol{\theta}_{st}, \boldsymbol{\theta}_{ss}, \boldsymbol{\theta}_{tt})} \pi_{st}(\boldsymbol{\theta}_{st} | \boldsymbol{\theta}_{ss}, \boldsymbol{\theta}_{tt}).$$

Then we call

$$\tilde{\pi}(\boldsymbol{\theta}) = \prod_{j}^{p} \kappa_{n,j}(\boldsymbol{\theta}_{jj}) \prod_{s<t}^{p} \tau_{n,st}(\boldsymbol{\theta}_{st} | \boldsymbol{\theta}_{ss}, \boldsymbol{\theta}_{tt}) \tag{4.4}$$

a Bayesian mosaic posterior under model $P_{\boldsymbol{\theta}}$.

4.1.4 Sampling Bayesian Mosaic

It is easily seen from (4.4) that the *knots* are marginally independent and the *tiles* are conditionally independent given the *knots*. This sparse dependence structure of *Bayesian mosaic* can be represented by a directed acyclic graph (DAG), as demonstrated in Figure 4.1. Utilizing this structure, we propose a simple parallel sampling strategy which is summarized in Algorithm 4.1, where $M \in \mathbb{N}_1$ denotes the total number of posterior samples to be collected.

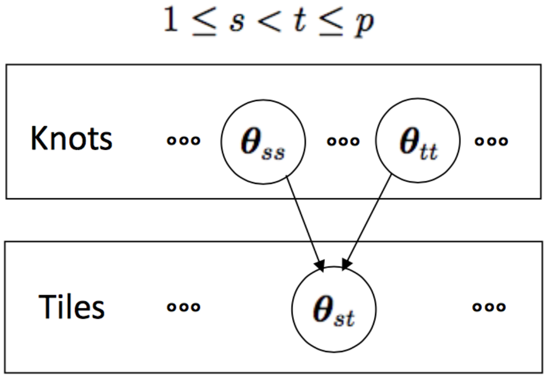
$1 \leq s < t \leq p$
Knots
$\boldsymbol{\theta}_{ss}$
$\boldsymbol{\theta}_{tt}$
Tiles
$\boldsymbol{\theta}_{st}$

FIGURE 4.1: DAG representation of a *Bayesian mosaic*.

```
Step 1 # on each of the knot marginals in parallel
for $j = 1, \dots, p$ do
    sample $\boldsymbol{\theta}_{jj}^1, \dots, \boldsymbol{\theta}_{jj}^M$ from $\kappa_{n,j}(\boldsymbol{\theta}_{jj})$
end
Step 2 # on each of the tile conditionals in parallel
for $s = 2, \dots, p$ do
    for $t = 1, \dots, s-1$ do
        sample $\boldsymbol{\theta}_{st}^m$ from $\tau_{n,st}(\boldsymbol{\theta}_{st}|\boldsymbol{\theta}_{ss}^m, \boldsymbol{\theta}_{tt}^m)$, for $m = 1, \dots, M$
    end
end
```

Algorithm 4.1: Parallel sampler for *Bayesian mosaic*.

Usually, the *knot marginals* and the *tile conditionals* can not be directly sampled from. We propose to sample the *knot marginals* via Metropolis-Hastings (MH) algorithms with the $Q_n(\boldsymbol{\theta}_{jj})$'s being evaluated via numerical integration. Sampling from the *tile conditionals* is harder, since the conditional distribution $\tau_{n,st}(\boldsymbol{\theta}_{st}|\boldsymbol{\theta}_{ss}^m, \boldsymbol{\theta}_{tt}^m)$ changes w.r.t. the values of $\boldsymbol{\theta}_{ss}^m$ and $\boldsymbol{\theta}_{tt}^m$. We propose the following three options:

i) Suppose that the MH sampler on $\tau_{n,st}(\boldsymbol{\theta}_{st}|\boldsymbol{\theta}_{ss}^m, \boldsymbol{\theta}_{tt}^m)$ converges rapidly, then for each m, one can run the sampler for a fixed small number of steps and use the last draw as the sample.

ii) Suppose that $\tau_{n,st}(\boldsymbol{\theta}_{st}|\boldsymbol{\theta}_{ss}^m, \boldsymbol{\theta}_{tt}^m)$ is easy to optimize w.r.t. $\boldsymbol{\theta}_{st}$, then one can compute the mode and the maximum density value. Then one can either do rejection sampling using the maximum density value or obtain the Hessian matrix at the mode and approximate the density by its Laplace approximation.

iii) One can directly plug in the posterior means of the *knots* into the *tile conditionals* so that they remain the same across iterations. Simply substitute $\tau_{n,st}(\boldsymbol{\theta}_{st}|\boldsymbol{\theta}_{ss}^m, \boldsymbol{\theta}_{tt}^m)$ in the second step of Algorithm 4.1 with $\tau_{n,st}\left(\boldsymbol{\theta}_{st}\middle|\frac{1}{M}\sum_{m=1}^{M}\boldsymbol{\theta}_{ss}^m, \frac{1}{M}\sum_{m=1}^{M}\boldsymbol{\theta}_{tt}^m\right)$.

The third option is the fallback plan when the first two are unavailable. Note that when applying the third option, instead of sampling from the *Bayesian mosaic*, one actually samples from the following approximation:

$$\prod_{j}^{p} \kappa_{n,j}(\boldsymbol{\theta}_{jj}) \prod_{s<t}^{p} \tau_{n,st}(\boldsymbol{\theta}_{st}|\boldsymbol{\theta}_{ss}^*, \boldsymbol{\theta}_{tt}^*), \tag{4.5}$$

where $\boldsymbol{\theta}_{jj}^* = \int \kappa_{n,j}(\boldsymbol{\theta}_{jj})\mathrm{d}\boldsymbol{\theta}_{jj}$ is the posterior mean of $\boldsymbol{\theta}_{jj}$, for $j = 1, \dots, p$. In §4.2.3 we will show that (4.5) is still consistent and asymptotically normal in a slightly weaker sense, but will underestimate the uncertainty compared to the exact *Bayesian mosaic*.

4.1.5 Handling Parameter Constraints

In some cases, the model parameters $\boldsymbol{\theta}$ live in a constrained space $\mathcal{T}$. However, the samples from *Bayesian mosaic* do not necessarily also live in this space. For instance, in (4.1), the samples of $\boldsymbol{\Sigma}$ from *Bayesian mosaic* are not guaranteed to be positive definite. One can easily see this from the fact that the off-diagonal elements σ_{st}'s are conditionally independent given the diagonal elements σ_{jj}'s.

To address this, we propose to project the samples from *Bayesian mosaic* back to the constrained space w.r.t. the Euclidean distance. Specifically, we solve the following optimization problem for each sample $\boldsymbol{\theta}^m$:

$$\tilde{\boldsymbol{\theta}}^m = \arg\min_{\tilde{\boldsymbol{\theta}}^m \in \mathcal{T}} \|\tilde{\boldsymbol{\theta}}^m - \boldsymbol{\theta}^m\|. \tag{4.6}$$

We term $\tilde{\boldsymbol{\theta}}^m$'s as the corrected samples from *Bayesian mosaic*. For many structured constrained parameter spaces $\mathcal{T}$, (4.6) has an analytical solution. For instance, if $\mathcal{T}$ is the cone of positive definite matrices, then $\tilde{\boldsymbol{\theta}}^m$ can be obtained via an eigenvalue decomposition of $\boldsymbol{\theta}^m$.

In §4.2.2, we will prove that the probability mass of *Bayesian mosaic* asymptotically concentrates within a small neighbourhood of the "true" value $\boldsymbol{\theta}_0$. This implies that when n is large for many constraints, the majority of the samples should automatically live inside $\mathcal{T}$ and we only need to correct the rest. Hence, this correcting step should have minimal impact on the overall performance.

4.1.6 Generalization

In this subsection, we will extend *Bayesian mosaic* to dependent data. We first provide a more general definition of *mosaic-type* data distributions.

Definition 4.1.3. *Suppose $p \in \mathbb{N}_1$ and $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ are p-dimensional data vectors jointly from distribution $P_{\boldsymbol{\theta}}$ with a joint density function $f(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n|\boldsymbol{\theta})$. Let $\boldsymbol{\theta}_{st}$, $1 \leqslant t \leqslant s \leqslant p$, be non-overlapping sub-vectors of $\boldsymbol{\theta}$ such that*

$$\boldsymbol{\theta} = \left[\boldsymbol{\theta}_{12}^\top, \ldots, \boldsymbol{\theta}_{(p-1)p}^\top, \boldsymbol{\theta}_{11}^\top, \ldots, \boldsymbol{\theta}_{pp}^\top\right]^\top,$$

then the data distribution $P_{\boldsymbol{\theta}}$ is mosaic-type if there exists a collection of density functions f_{st}, $1 \leqslant s \leqslant t \leqslant p$ such that

i) for $j = 1, \ldots, p$, the density of the univariate marginal distribution for dimension j is

$$f_{jj}(y_{1j}, \ldots, y_{nj}|\boldsymbol{\theta}_{jj}).$$

ii) for $1 \leqslant t < s \leqslant p$, the density of the bivariate marginal data distribution for dimension s and t is

$$f_{st}(y_{1s}, \ldots, y_{ns}, y_{1t}, \ldots, y_{nt}|\boldsymbol{\theta}_{st}, \boldsymbol{\theta}_{ss}, \boldsymbol{\theta}_{tt}).$$

For $j = 1, \dots, p$, we redefine $Q_{n,j}(\boldsymbol{\theta}_{jj})$ as

$$Q_{n,j}(\boldsymbol{\theta}_{jj}) = \log f_{jj}(y_{1j}, \dots, y_{nj} | \boldsymbol{\theta}_{jj}).$$

For $1 \leqslant s < t \leqslant p$, we redefine $L_{n,st}(\boldsymbol{\theta}_{st}, \boldsymbol{\theta}_{ss}, \boldsymbol{\theta}_{tt})$ as

$$L_{n,st}(\boldsymbol{\theta}_{st}, \boldsymbol{\theta}_{ss}, \boldsymbol{\theta}_{tt}) = \log f_{st}(y_{1s}, \dots, y_{ns}, y_{1t}, \dots, y_{nt} | \boldsymbol{\theta}_{st}, \boldsymbol{\theta}_{ss}, \boldsymbol{\theta}_{tt}).$$

Then *Bayesian mosaic* still follows Definition 4.1.2.

Under this generalization, one can include random effects and still be able to use *Bayesian mosaic*. We will give an example of such a model in §4.3.3, where we include random temporal effects.

4.2 Theoretical Analysis

In this section we will first demonstrate that the *mosaic-type* class contains a rich collection of models. We will then provide regularity conditions and prove under these conditions that *Bayesian mosaic* is consistent and asymptotically normal. Moreover, we will analyze the asymptotic distribution of the *tiles* conditional on the posterior means of the *knots*. Finally we will build a connection between the sampling computational complexity and the cardinality of the data. We prove the main result and defer other proofs to the supplement.

4.2.1 Richness of the Mosaic-type Distribution Class

To evaluate how widely *Bayesian mosaic* can be applied in practice, it is crucial to understand how rich the *mosaic-type* distribution class is. The following lemma provides one simple rule to construct new *mosaic-type* distributions from any existing *mosaic-type* distributions. With the help of this rule, one can build models for any type of data with the dependence induced by latent variables with some underlying *mosaic-type* distribution.

Lemma 4.2.1. *Suppose that $P_{\boldsymbol{\psi}}$ is some data distribution with density function $f_0(\boldsymbol{x}|\boldsymbol{\psi})$ and consider another data distribution $P_{\boldsymbol{\mu},\boldsymbol{\psi}}$ with the following density function:*

$$f(\boldsymbol{y}|\boldsymbol{\mu},\boldsymbol{\psi}) = \int f_0(\boldsymbol{x}|\boldsymbol{\psi}) \prod_{j=1}^{p} g_j(y_j|x_j,\boldsymbol{\mu}_j)\,d\boldsymbol{x}, \tag{4.7}$$

where $\boldsymbol{\mu} = \left(\boldsymbol{\mu}_1^\top,\ldots,\boldsymbol{\mu}_p^\top\right)^\top$ and g_j's are proper density functions. If $P_{\boldsymbol{\psi}}$ is mosaic-type, so is $P_{\boldsymbol{\mu},\boldsymbol{\psi}}$.

One could choose $P_{\boldsymbol{\psi}}$ to be the multivariate Gaussian distribution, and $g_j(y_j|x_j,\boldsymbol{\mu}_j)$'s to be any univariate density. This implies that the *mosaic-type* model class contains the multivariate latent Gaussian models. Note that besides Gaussian, $P_{\boldsymbol{\psi}}$ could also be a Dirichlet or a multinomial distribution. It is easy to check that both distributions are *mosaic-type*. Moreover, one can construct arbitrarily complex models by repeatedly applying Lemma 4.2.1.

4.2.2 Posterior Consistency & Asymptotic Normality

We start our analysis in a simpler yet more general setup. Consider $p \in \mathbb{N}_1$, $d \in \mathbb{N}_1$ and $\boldsymbol{y}_1,\ldots,\boldsymbol{y}_n$ which are i.i.d. p-dimensional observations from distribution $P_{\boldsymbol{\theta}}$ possessing a density $f(\boldsymbol{y}|\boldsymbol{\theta})$ where $\boldsymbol{\theta} \in \mathcal{T} \subset \mathbb{R}^d$. We fix $\boldsymbol{\theta}_0$ to be the "true value" of the parameters and require that $\boldsymbol{\theta}_0$ is an interior point of $\mathcal{T}$. Consider two non-overlapping sub-vectors of $\boldsymbol{\theta}$, $\boldsymbol{\eta}$ and $\boldsymbol{\zeta}$, where $\boldsymbol{\eta}$ is d_η-dimensional and $\boldsymbol{\zeta}$ is d_ζ-dimensional. Let $\boldsymbol{\eta}_0$ and $\boldsymbol{\zeta}_0$ be the corresponding "true values". Consider two pseudo density functions $f_1(\boldsymbol{y}|\boldsymbol{\zeta})$ and $f_2(\boldsymbol{y}|\boldsymbol{\eta},\boldsymbol{\zeta})$, which do not have to integrate to one. In order for proper Bayesian inference, the following regularity conditions need to hold for both functions. To avoid redundancy, we only define these conditions for $f_1(\boldsymbol{y}|\boldsymbol{\zeta})$.

Condition 1. *The support set $\{\boldsymbol{y} : f_1(\boldsymbol{y}|\boldsymbol{\zeta}) > 0\}$ is the same for all $\boldsymbol{\zeta}$.*

Condition 2. *Consider* $\ell_1(\boldsymbol{\zeta}, \boldsymbol{y}) = \log f_1(\boldsymbol{y}|\boldsymbol{\zeta})$*.* $\ell_1(\boldsymbol{\zeta}, \boldsymbol{y})$ *is thrice differentiable with respect to* $\boldsymbol{\zeta}$ *in a neighborhood* $\mathcal{B}_{\boldsymbol{\zeta}}(\boldsymbol{\zeta}_0, \delta)$*. The expectations* $\mathbb{E}_{\boldsymbol{\theta}_0} \nabla_{\boldsymbol{\zeta}} \ell_1(\boldsymbol{\zeta}, \boldsymbol{y})$ *and* $\mathbb{E}_{\boldsymbol{\theta}_0} \nabla^2_{\boldsymbol{\zeta}} \ell_1(\boldsymbol{\zeta}, \boldsymbol{y})$ *are both finite and for any* multi-index $\boldsymbol{\alpha}$ *for* $\boldsymbol{\zeta}$ *such that* $|\boldsymbol{\alpha}| = 3$*, we have*

$$\sup_{\boldsymbol{\zeta} \in \mathcal{B}_{\boldsymbol{\zeta}}(\boldsymbol{\zeta}_0, \delta)} |\partial^{\boldsymbol{\alpha}} \ell_1(\boldsymbol{\zeta}, \boldsymbol{y})| \leqslant M_{\boldsymbol{\alpha}}(\boldsymbol{y}),$$

and $\mathbb{E}_{\boldsymbol{\theta}_0} M_{\boldsymbol{\alpha}}(\boldsymbol{y}) < \infty$*.*

Condition 3. *Consider* $\ell_1(\boldsymbol{\zeta}, \boldsymbol{y}) = \log f_1(\boldsymbol{y}|\boldsymbol{\zeta})$*. Then* $\mathbb{E}_{\boldsymbol{\theta}_0} \nabla_{\boldsymbol{\zeta}} \ell_1(\boldsymbol{\zeta}, \boldsymbol{y})|_{\boldsymbol{\zeta}=\boldsymbol{\zeta}_0} = \mathbf{0}$ *and*

$$\mathbb{E}_{\boldsymbol{\theta}_0} \nabla^2_{\boldsymbol{\zeta}} \ell_1(\boldsymbol{\zeta}, \boldsymbol{y})|_{\boldsymbol{\zeta}=\boldsymbol{\zeta}_0} = -\, \mathbb{E}_{\boldsymbol{\theta}_0} \big[\nabla_{\boldsymbol{\zeta}} \ell_1(\boldsymbol{\zeta}, \boldsymbol{y})\big]\big[\nabla_{\boldsymbol{\zeta}} \ell_1(\boldsymbol{\zeta}, \boldsymbol{y})\big]^{\top}|_{\boldsymbol{\zeta}=\boldsymbol{\zeta}_0}$$

Also, the Fisher information $-\mathbb{E}_{\boldsymbol{\theta}_0} \nabla^2_{\boldsymbol{\zeta}} \ell_1(\boldsymbol{\zeta}, \boldsymbol{y})|_{\boldsymbol{\zeta}=\boldsymbol{\zeta}_0}$ *is positive definite.*

Condition 4. *Consider* $Q_n(\boldsymbol{\zeta}) = \sum_{i=1}^n \log f_1(\boldsymbol{y}_i|\boldsymbol{\zeta})$*. For any* $\delta > 0$*,* $\exists \epsilon > 0$ *such that with* $P_{\boldsymbol{\theta}_0}$*-probability one*

$$\sup_{\boldsymbol{\zeta} \notin \mathcal{B}_{\boldsymbol{\zeta}}(\boldsymbol{\zeta}_0, \delta)} \frac{1}{n} \big[Q_n(\boldsymbol{\zeta}) - Q_n(\boldsymbol{\zeta}_0)\big] < -\epsilon$$

for all sufficiently large n*.*

Condition 5. *Consider* $Q_n(\boldsymbol{\zeta}) = \sum_{i=1}^n \log f_1(\boldsymbol{y}_i|\boldsymbol{\zeta})$ *and* $\tilde{\boldsymbol{\zeta}}_n = \arg\max_{\boldsymbol{\zeta}} Q_n(\boldsymbol{\zeta})$*.* $\tilde{\boldsymbol{\zeta}}_n$ *is consistent at* $\boldsymbol{\zeta}_0$*, i.e.,* $\lim_{n \to \infty} \tilde{\boldsymbol{\zeta}}_n = \boldsymbol{\zeta}_0$ *with* $P_{\boldsymbol{\theta}_0}$*-probability one.*

For ease of presentation, we introduce some notation. We define

$$\ell_1(\boldsymbol{\zeta}, \boldsymbol{y}) = \log f_1(\boldsymbol{y}|\boldsymbol{\zeta}), \quad Q_n(\boldsymbol{\zeta}) = \sum_{i=1}^n \ell_1(\boldsymbol{\zeta}, \boldsymbol{y}_i),$$

$$\ell_2(\boldsymbol{\eta}, \boldsymbol{\zeta}, \boldsymbol{y}) = \log f_2(\boldsymbol{y}|\boldsymbol{\eta}, \boldsymbol{\zeta}), \quad L_n(\boldsymbol{\eta}, \boldsymbol{\zeta}) = \sum_{i=1}^n \ell_2(\boldsymbol{\eta}, \boldsymbol{\zeta}, \boldsymbol{y}_i),$$

$$\left[\begin{smallmatrix} \hat{\boldsymbol{\eta}}_n \\ \hat{\boldsymbol{\zeta}}_n \end{smallmatrix}\right] = \arg\max_{\boldsymbol{\eta}, \boldsymbol{\zeta}} L_n(\boldsymbol{\eta}, \boldsymbol{\zeta}), \quad \tilde{\boldsymbol{\zeta}}_n = \arg\max_{\boldsymbol{\zeta}} Q_n(\boldsymbol{\zeta}),$$

and

$$\tilde{\boldsymbol{I}}_0 = -\mathbb{E}_{\boldsymbol{\theta}_0} \nabla^2_{\boldsymbol{\zeta}} \ell_1(\boldsymbol{\zeta}, \boldsymbol{y})|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}, \quad \boldsymbol{I}_0 = -\mathbb{E}_{\boldsymbol{\theta}_0} \nabla^2_{\boldsymbol{\eta},\boldsymbol{\zeta}} \ell_2(\boldsymbol{\eta}, \boldsymbol{\zeta}, \boldsymbol{y})|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}.$$

Given a prior density $\pi(\boldsymbol{\zeta})$, consider the following posterior density of $\boldsymbol{\zeta}$:

$$\kappa_n\big(\boldsymbol{\zeta}\big) \propto \exp\{Q_n(\boldsymbol{\zeta})\}\pi(\boldsymbol{\zeta}).$$

Introducing $\boldsymbol{\omega} = \sqrt{n}(\boldsymbol{\zeta} - \tilde{\boldsymbol{\zeta}}_n)$, the posterior density of $\boldsymbol{\omega}$ is

$$\pi^*_{n,1}(\boldsymbol{\omega}) = \frac{1}{\sqrt{n}}\kappa_n\bigg(\frac{\boldsymbol{\omega}}{\sqrt{n}} + \tilde{\boldsymbol{\zeta}}_n\bigg).$$

The following lemmas state that $\pi^*_{n,1}(\boldsymbol{\omega})$ is asymptotically normal under the specified conditions. This lemma is basically the multivariate version of Theorem 4.2 in Ghosh et al. (2007), hence the proof will be omitted.

Lemma 4.2.2. *Suppose that conditions 1-5 hold for $f_1(\boldsymbol{y}|\boldsymbol{\zeta})$, and the prior density $\pi(\boldsymbol{\zeta})$ is continuous and positive at $\boldsymbol{\zeta}_0$, then with $P_{\boldsymbol{\theta}_0}$-probability one*

$$\lim_{n\to\infty} \int \big|\pi^*_{n,1}(\boldsymbol{\omega}) - \phi\big(\boldsymbol{\omega}\big|\mathbf{0}, \tilde{\boldsymbol{I}}_0^{-1}\big)\big| d\boldsymbol{\omega} = 0. \tag{4.8}$$

With a prior density $\pi(\boldsymbol{\eta}|\boldsymbol{\zeta})$, consider the following conditional posterior density of $\boldsymbol{\eta}$ given $\boldsymbol{\zeta}$:

$$\tau_n(\boldsymbol{\eta}|\boldsymbol{\zeta}) \propto \exp\big\{L_n\big(\boldsymbol{\eta}, \boldsymbol{\zeta}\big) - L_n\big(\hat{\boldsymbol{\eta}}_n, \hat{\boldsymbol{\zeta}}_n\big)\big\}\pi(\boldsymbol{\eta}|\boldsymbol{\zeta}).$$

Introducing $\boldsymbol{t} = \sqrt{n}\big(\boldsymbol{\eta} - \hat{\boldsymbol{\eta}}_n\big)$ and $\boldsymbol{r} = \sqrt{n}\big(\boldsymbol{\zeta} - \hat{\boldsymbol{\zeta}}_n\big)$, the conditional posterior density of $\boldsymbol{t}$ given $\boldsymbol{r}$ can be written as

$$\begin{aligned}\pi^*_{n,2}(\boldsymbol{t}|\boldsymbol{r}) =& a_n^{-1}(\boldsymbol{r}) \exp\big\{L_n\big(\hat{\boldsymbol{\eta}}_n + \boldsymbol{t}/\sqrt{n}, \hat{\boldsymbol{\zeta}}_n + \boldsymbol{r}/\sqrt{n}\big) - L_n\big(\hat{\boldsymbol{\eta}}_n, \hat{\boldsymbol{\zeta}}_n\big)\big\} \\ & \times \pi\big(\hat{\boldsymbol{\eta}}_n + \boldsymbol{t}/\sqrt{n}\big|\hat{\boldsymbol{\zeta}}_n + \boldsymbol{r}/\sqrt{n}\big)\end{aligned}$$

with $a_n(\boldsymbol{r})$ being the normalizing constant. We define

$$\begin{aligned}\boldsymbol{I}_0^{11} &= -\mathbb{E}_{\boldsymbol{\theta}_0} \nabla^2_{\boldsymbol{\eta}} \ell_2(\boldsymbol{\eta}, \boldsymbol{\zeta}, \boldsymbol{y})|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}, \quad \boldsymbol{I}_0^{12} = -\mathbb{E}_{\boldsymbol{\theta}_0} \nabla_{\boldsymbol{\zeta}} \nabla_{\boldsymbol{\eta}} \ell_2(\boldsymbol{\eta}, \boldsymbol{\zeta}, \boldsymbol{y})|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}, \\ \boldsymbol{I}_0^{21} &= -\mathbb{E}_{\boldsymbol{\theta}_0} \nabla_{\boldsymbol{\eta}} \nabla_{\boldsymbol{\zeta}} \ell_2(\boldsymbol{\eta}, \boldsymbol{\zeta}, \boldsymbol{y})|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}, \quad \boldsymbol{I}_0^{22} = -\mathbb{E}_{\boldsymbol{\theta}_0} \nabla^2_{\boldsymbol{\zeta}} \ell_2(\boldsymbol{\eta}, \boldsymbol{\zeta}, \boldsymbol{y})|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}.\end{aligned}$$

It is easily seen that $\boldsymbol{I}_0 = \left[\begin{smallmatrix}\boldsymbol{I}_0^{11} & \boldsymbol{I}_0^{12} \\ \boldsymbol{I}_0^{21} & \boldsymbol{I}_0^{22}\end{smallmatrix}\right]$.

Theorem 4.2.3. *Suppose that the conditions for Lemma 4.2.2 hold, conditions 1-5 hold for* $f_2(\boldsymbol{y}|\boldsymbol{\eta}, \boldsymbol{\zeta})$ *and that the prior density* $\pi(\boldsymbol{\eta}|\boldsymbol{\zeta})$ *is continuous and positive at* $\left[\begin{smallmatrix}\boldsymbol{\eta}_0\\ \boldsymbol{\zeta}_0\end{smallmatrix}\right]$*, then with* $P_{\boldsymbol{\theta}_0}$*-probability one*

$$\lim_{n\to\infty} \int\int \left| \pi_n^*(\boldsymbol{t}, \boldsymbol{r}) - \phi\left(\boldsymbol{t}\middle| -\left(\boldsymbol{I}_0^{11}\right)^{-1} \boldsymbol{I}_0^{12} \boldsymbol{r}, \left(\boldsymbol{I}_0^{11}\right)^{-1}\right) \phi\left(\boldsymbol{r}\middle|\boldsymbol{\mu}_n, \tilde{\boldsymbol{I}}_0^{-1}\right)\right| d\boldsymbol{r}\, d\boldsymbol{t} = 0, \qquad (4.9)$$

where $\pi_n^*(\boldsymbol{t}, \boldsymbol{r}) = \pi_{n,2}^*(\boldsymbol{t}|\boldsymbol{r})\pi_{n,1}^*(\boldsymbol{r})$ *is the joint posterior density of* $\boldsymbol{t}$ *and* $\boldsymbol{r}$ *and* $\boldsymbol{\mu}_n = \sqrt{n}\big(\tilde{\boldsymbol{\zeta}}_n - \hat{\boldsymbol{\zeta}}_n\big)$.

Corollary 4.2.4. *Under the same setup in Theorem 4.2.3,* $\tau_n\left(\boldsymbol{\eta}|\boldsymbol{\zeta}\right)\kappa_n\left(\boldsymbol{\zeta}\right)$ *is consistent at* $\boldsymbol{\eta}_0$ *and* $\boldsymbol{\zeta}_0$.

We can directly apply Theorem 4.2.3 and Corollary 4.2.4 to analyze the asymptotic properties of *Bayesian mosaic*. All we need is to let $f_1(\boldsymbol{y}|\boldsymbol{\zeta})$ be the multiplication of the densities of the univariate marginal distributions,

$$f_1(\boldsymbol{y}|\boldsymbol{\zeta}) = \prod_{j=1}^{p} f_{jj}(y_j|\boldsymbol{\theta}_{jj}),$$

$f_2(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{\zeta})$ be the multiplication of the densities of the bivariate marginal distributions,

$$f_2(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{\zeta}) = \prod_{1\leqslant s<t\leqslant p} f_{st}(y_s, y_t|\boldsymbol{\theta}_{st}, \boldsymbol{\theta}_{ss}, \boldsymbol{\theta}_{tt}),$$

$\pi(\boldsymbol{\zeta})$ be the multiplication of the prior densities for *knots*,

$$\pi(\boldsymbol{\zeta}) = \prod_{j=1}^{p} \pi_{jj}(\boldsymbol{\theta}_{jj}),$$

and $\pi(\boldsymbol{\eta}|\boldsymbol{\zeta})$ be the multiplication of the conditional prior densities for *tiles*,

$$\pi(\boldsymbol{\eta}|\boldsymbol{\zeta}) = \prod_{1\leqslant s<t\leqslant p} \pi_{st}(\boldsymbol{\theta}_{st}|\boldsymbol{\theta}_{ss}, \boldsymbol{\theta}_{tt}).$$

We immediately have

$$\ell_1(\boldsymbol{\zeta}, \boldsymbol{y}) = \sum_{j=1}^{p} \ell_{jj}(y_j|\boldsymbol{\theta}_{jj}), \quad \ell_2(\boldsymbol{\eta}, \boldsymbol{\zeta}, \boldsymbol{y}) = \sum_{1\leqslant s<t\leqslant p} \ell_{st}(y_s, y_t|\boldsymbol{\theta}_{st}, \boldsymbol{\theta}_{ss}, \boldsymbol{\theta}_{tt}).$$

It can be shown that

$$\begin{aligned}
\kappa_n(\boldsymbol{\zeta}) &\propto \exp\left\{\sum_{i=1}^{n} \ell_1(\boldsymbol{\zeta}, \boldsymbol{y}_i)\right\}\pi(\boldsymbol{\zeta}) \\
&= \exp\left\{\sum_{j=1}^{p}\sum_{i=1}^{n} \ell_{jj}(y_{ij}|\boldsymbol{\theta}_{jj})\right\}\prod_{j=1}^{p}\pi_{jj}(\boldsymbol{\theta}_{jj}) \\
&= \prod_{j=1}^{p}\exp\left\{\sum_{i=1}^{n} \ell_{jj}(y_{ij}|\boldsymbol{\theta}_{jj})\right\}\pi_{jj}(\boldsymbol{\theta}_{jj}) \\
&= \prod_{j=1}^{p}\kappa_{n,jj}(\boldsymbol{\theta}_{jj}).
\end{aligned}$$

Similarly we can show that

$$\tau_n(\boldsymbol{\eta}|\boldsymbol{\zeta}) = \prod_{s<t}^{p} \tau_{n,st}(\boldsymbol{\theta}_{st}|\boldsymbol{\theta}_{ss}, \boldsymbol{\theta}_{tt}).$$

Therefore $\tau_n(\boldsymbol{\eta}|\boldsymbol{\zeta})\kappa_n(\boldsymbol{\zeta})$ is exactly the *Bayesian mosaic*, specifically,

$$\tau_n(\boldsymbol{\eta}|\boldsymbol{\zeta})\kappa_n(\boldsymbol{\zeta}) = \prod_{s<t}^{p} \tau_{n,st}(\boldsymbol{\theta}_{st}|\boldsymbol{\theta}_{ss}, \boldsymbol{\theta}_{tt}) \prod_{j=1}^{p} \kappa_{n,jj}(\boldsymbol{\theta}_{jj}) = \tilde{\pi}_n(\boldsymbol{\theta}),$$

where $\boldsymbol{\theta} = \left[\begin{smallmatrix}\boldsymbol{\eta}\\ \boldsymbol{\zeta}\end{smallmatrix}\right]$. Consequently, Theorem 4.2.3 and Corollary 4.2.4 can be used directly to analyze the asymptotic properties of *Bayesian mosaic.* The following lemma provides sufficient conditions for the regularity conditions for Theorem 4.2.3 to hold. The proof of this lemma is straightforward and hence is omitted.

Lemma 4.2.5. *Suppose that for $j = 1, \ldots, p$, $f_{jj}(\boldsymbol{y}|\boldsymbol{\eta}_{jj}) = f_{jj}(y_j|\boldsymbol{\eta}_{jj})$ satisfies conditions 1-5, and for $1 \leqslant s < t \leqslant p$, $f_{st}(\boldsymbol{y}|\boldsymbol{\eta}_{st}, \boldsymbol{\zeta}_{st}) = f_{st}(\boldsymbol{y}|\boldsymbol{\eta}_{st}, \boldsymbol{\eta}_{ss}, \boldsymbol{\eta}_{t})$ satisfies conditions 1-5. Then conditions 1-5 also hold for both $f_1(\boldsymbol{y}|\boldsymbol{\zeta})$ and $f_2(\boldsymbol{y}|\boldsymbol{\eta}, \boldsymbol{\zeta})$.*

Recall that $\boldsymbol{t} = \sqrt{n}\big(\boldsymbol{\eta} - \hat{\boldsymbol{\eta}}_n\big)$ and $\boldsymbol{r} = \sqrt{n}\big(\boldsymbol{\zeta} - \hat{\boldsymbol{\zeta}}_n\big)$, then the *Bayesian mosaic* of $\boldsymbol{t}$ and $\boldsymbol{r}$ can be written as $\tilde{\pi}_n^*(\boldsymbol{t}, \boldsymbol{r}) = \frac{1}{\sqrt{n}}\tilde{\pi}_n(\hat{\boldsymbol{\eta}}_n + \boldsymbol{t}/\sqrt{n}, \hat{\boldsymbol{\zeta}}_n + \boldsymbol{r}/\sqrt{n})$. Applying Lemma 4.2.5 and Theorem 4.2.3, we know that if the requirements of Lemma 4.2.5 are satisfied, with $P_{\boldsymbol{\theta}_0}$-probability one

$$\lim_{n\to\infty} \int\int \left|\pi_n^*(\boldsymbol{t}, \boldsymbol{r}) - \phi\left(\boldsymbol{t}\middle| -\left(\boldsymbol{I}_0^{11}\right)^{-1}\boldsymbol{I}_0^{12}\boldsymbol{r}, \left(\boldsymbol{I}_0^{11}\right)^{-1}\right)\phi\left(\boldsymbol{r}\middle|\boldsymbol{\mu}_n, \tilde{\boldsymbol{I}}_0^{-1}\right)\right| \mathrm{d}\boldsymbol{r}\mathrm{d}\boldsymbol{t} = 0,$$

where $\tilde{\boldsymbol{I}}_0 = -\mathbb{E}_{\boldsymbol{\theta}_0}\nabla_{\boldsymbol{\zeta}}^2\ell_1(\boldsymbol{\zeta}, \boldsymbol{y})|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}$, $\boldsymbol{I}_0^{11} = -\mathbb{E}_{\boldsymbol{\theta}_0}\nabla_{\boldsymbol{\eta}}^2\ell_2(\boldsymbol{\eta}, \boldsymbol{\zeta}, \boldsymbol{y})|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}$ and $\boldsymbol{I}_0^{12} = -\mathbb{E}_{\boldsymbol{\theta}_0}\nabla_{\boldsymbol{\zeta}}\nabla_{\boldsymbol{\eta}}\ell_2(\boldsymbol{\eta}, \boldsymbol{\zeta}, \boldsymbol{y})|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}$.

To gain more insight on what the asymptotic covariance of $\tilde{\pi}_n^*(\boldsymbol{t}, \boldsymbol{r})$ is, we will look at $\boldsymbol{I}_0^{11}$, $\boldsymbol{I}_0^{12}$ and $\tilde{\boldsymbol{I}}_0$ in more detail. For $j = 1, \dots, p$, we define

$$\boldsymbol{\Sigma}_{jj} = \mathbb{E}_{\boldsymbol{\theta}_0}\nabla_{\boldsymbol{\theta}_{jj}}^2\ell_{jj}(\boldsymbol{\theta}_{jj}, y_j)|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}.$$

Since $\ell_1(\boldsymbol{\zeta}, \boldsymbol{y}) = \sum_{j=1}^p \ell_{jj}(y_j|\boldsymbol{\theta}_{jj})$ and $\boldsymbol{\zeta} = (\boldsymbol{\theta}_{11}, \dots, \boldsymbol{\theta}_{pp})^\top$, it is easy to see that

$$\tilde{\boldsymbol{I}}_0 = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_{22} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \boldsymbol{\Sigma}_{pp} \end{bmatrix}.$$

For $1 \leqslant s < t \leqslant p$, we define

$$\boldsymbol{\Sigma}_{st} = \mathbb{E}_{\boldsymbol{\theta}_0}\nabla_{\boldsymbol{\theta}_{st}}^2\ell_{st}(\boldsymbol{\theta}_{st}, \boldsymbol{\theta}_{ss}, \boldsymbol{\theta}_{tt}, y_s, y_t)|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}.$$

Since $\ell_2(\boldsymbol{\eta}, \boldsymbol{\zeta}, \boldsymbol{y}) = \sum_{1\leqslant s<t\leqslant p} \ell_{st}(y_s, y_t|\boldsymbol{\theta}_{st}, \boldsymbol{\theta}_{ss}, \boldsymbol{\theta}_{tt})$ and $\boldsymbol{\eta} = \big[\boldsymbol{\theta}_{12}, \dots, \boldsymbol{\theta}_{(p-1)p}\big]^\top$, it is easy to see that

$$\boldsymbol{I}_0^{11} = \begin{bmatrix} \boldsymbol{\Sigma}_{12} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_{13} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \boldsymbol{\Sigma}_{(p-1)p} \end{bmatrix}.$$

Note that the $\boldsymbol{\Sigma}_{st}$'s are ordered in a row-major manner on the diagonal of $\boldsymbol{\Sigma}_0$. For $1 \leqslant s < t \leqslant p$ and $j = 1, \dots, p$, we define

$$\boldsymbol{\Sigma}_{st,j} = \mathbb{E}_{\boldsymbol{\theta}_0}\nabla_{\boldsymbol{\theta}_{jj}}\nabla_{\boldsymbol{\theta}_{st}}\ell_{st}(\boldsymbol{\theta}_{st}, \boldsymbol{\theta}_{ss}, \boldsymbol{\theta}_{tt}, y_s, y_t)|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}.$$

It can be shown that

$$\boldsymbol{I}_0^{12} = \begin{bmatrix} \boldsymbol{\Sigma}_{12,1} & \boldsymbol{\Sigma}_{12,2} & \cdots & \boldsymbol{\Sigma}_{12,p} \\ \boldsymbol{\Sigma}_{13,1} & \boldsymbol{\Sigma}_{13,2} & \cdots & \boldsymbol{\Sigma}_{13,p} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{\Sigma}_{(p-1)p,1} & \boldsymbol{\Sigma}_{(p-1)p,2} & \cdots & \boldsymbol{\Sigma}_{(p-1)p,p} \end{bmatrix},$$

where $\boldsymbol{\Sigma}_{st,j}$'s are ordered in a row-major manner within their column for $j = 1\ldots,p$. Note that $\boldsymbol{I}_0^{12}$ is sparse since $\boldsymbol{\Sigma}_{st,j} = \boldsymbol{0}$ if $j \neq s$ and $j \neq t$.

$\tilde{\boldsymbol{I}}_0^{-1}$ is the marginal variance for $\boldsymbol{r}$. It is block diagonal due to the posterior independence of the *knots*. It can be seen that each block is the Fisher information induced by a univariate marginal data distribution. $(\boldsymbol{I}_0^{11})^{-1}$ is the conditional variance for $\boldsymbol{t}$. It is also block diagonal due to the conditional independence of the *tiles* given the *knots*. Each block is the Fisher information induced by a bivariate marginal data distribution. $\boldsymbol{I}_0^{12}$ characterizes the connection between *knots* and *tiles*. Its sparsity is due to the fact that $\boldsymbol{\theta}_{st}$ given $\boldsymbol{\theta}_{ss}$ and $\boldsymbol{\theta}_{tt}$ is conditionally independent of other *knots*.

4.2.3 Asymptotic Properties of the Posterior Mean

Under the same setup of Lemma 4.2.2, define $\boldsymbol{\zeta}_n^*$ as the posterior mean w.r.t. $\kappa_n(\boldsymbol{\zeta})$, i.e., $\boldsymbol{\zeta}_n^* = \int \boldsymbol{\zeta}\kappa_n(\boldsymbol{\zeta})\mathrm{d}\boldsymbol{\zeta}$. We can prove the following lemma.

Lemma 4.2.6. *Suppose the conditions for Lemma 4.2.2 hold and that the prior* $\pi(\boldsymbol{\zeta})$ *has a finite expectation, then* $\lim_{n\to\infty} \sqrt{n}(\boldsymbol{\zeta}_n^* - \tilde{\boldsymbol{\zeta}}_n) = 0$ *with* $P_{\boldsymbol{\theta}_0}$*-probability one.*

Lemma 4.2.6 is a multivariate version of Theorem 4.3 in Ghosh et al. (2007). It states that the posterior mean is approximately the same as the MLE when n is large.

Now we will investigate sampling from *tile conditionals* by directly plugging in the posterior mean of the *knots*. Under the same setup of Theorem 4.2.3, if we plug $\boldsymbol{\zeta}_n^*$

into the conditional density $\tau_n(\boldsymbol{\eta}|\boldsymbol{\zeta})$, we will get the following posterior distribution

$$\tau_n(\boldsymbol{\eta}|\boldsymbol{\zeta}_n^*)\kappa_n(\boldsymbol{\zeta}),$$

which is different from the exact *Bayesian mosaic* posterior. Recalling that $\boldsymbol{t} = \sqrt{n}(\boldsymbol{\eta} - \hat{\boldsymbol{\eta}}_n)$ and letting $\pi_{n,3}^*(\boldsymbol{t}) = \frac{1}{\sqrt{n}}\tau_n(\hat{\boldsymbol{\eta}}_n + \boldsymbol{t}/\sqrt{n}|\boldsymbol{\zeta}_n^*)$, the following theorem states that $\pi_{n,3}^*(\boldsymbol{t})$ is also asymptotic normal in a slightly weaker sense.

Theorem 4.2.7. *Suppose that the conditions for Theorem 4.2.3 hold that the prior* $\pi(\boldsymbol{\zeta})$ *has a finite expectation , then*

$$\int \left|\pi_{n,3}^*(\boldsymbol{t}) - \phi\left(\boldsymbol{t}\middle|-\left(\boldsymbol{I}_0^{11}\right)^{-1}\boldsymbol{I}_0^{12}\boldsymbol{\mu}_n, \left(\boldsymbol{I}_0^{11}\right)^{-1}\right)\right| d\boldsymbol{t} \overset{P_{\boldsymbol{\theta}_0}}{\to} 0. \tag{4.10}$$

Note that the integral in (4.10) converges to zero in probability, which is slightly weaker than the almost surely convergence in Theorem 4.2.3. Moreover, Theorem 4.2.3 implies that

$$\lim_{n\to\infty}\int\left|\int \pi_n^*(\boldsymbol{t},\boldsymbol{r})\mathrm{d}\boldsymbol{r} - \phi\left(\boldsymbol{t}\middle|-\left(\boldsymbol{I}_0^{11}\right)^{-1}\boldsymbol{I}_0^{12}\boldsymbol{\mu}_n, \left(\boldsymbol{I}_0^{11}\right)^{-1} + \boldsymbol{\Lambda}_0\right)\right|\mathrm{d}\boldsymbol{t} = 0,$$

where

$$\boldsymbol{\Lambda}_0 = \boldsymbol{I}_0^{12}\left(\boldsymbol{I}_0^{11}\right)^{-1}\tilde{\boldsymbol{I}}_0\left(\boldsymbol{I}_0^{11}\right)^{-1}\boldsymbol{I}_0^{21}$$

is positive semi-definite. This indicates that plugging in the posterior mean leads to some under-estimation of uncertainty as expected.

4.2.4 Asymptotic Bound on Computational Complexity

We finish this section by investigating the per-iteration computational complexity of sampling from *Bayesian mosaic* when the data are discrete. We start by analyzing sampling from the *knot marginals*. Recall

$$\kappa_j(\boldsymbol{\theta}_{jj}) \propto e^{\sum_{i=1}^n \ell_{jj}(\boldsymbol{\theta}_{jj}, y_{ij})}\pi_{jj}(\boldsymbol{\theta}_{jj}). \tag{4.11}$$

For discrete data, we assume the cardinality of $\{y_{1j}, \ldots, y_{nj}\}$ is $K \in \mathbb{N}_1$ and that $y_{i_1j}, \ldots, y_{i_Kj}$ are K unique values of $y_{1j}, \ldots, y_{nj}$. For $k = 1, \ldots, K$, we define $n_k = \sum_{i=1}^{n} \mathbb{1}\{y_{ij} = y_{i_kj}\}$. Then (4.11) can be written as

$$\kappa_j(\boldsymbol{\theta}_{jj}) \propto e^{\sum_{k=1}^{K} n_k \ell_{jj}(\boldsymbol{\theta}_{jj}, y_{i_kj})} \pi_{jj}(\boldsymbol{\theta}_{jj}).$$

Clearly the per-iteration computational complexity of sampling from $\kappa_{n,j}(\boldsymbol{\theta}_{jj})$ is dominated by evaluating $\sum_{k=1}^{K} n_k \ell_{jj}(\boldsymbol{\theta}_{jj}, y_{i_kj})$, which scales linearly with K. It is easily seen that K is bounded by $\max_{1 \leqslant i \leqslant n} y_{ij} - \min_{1 \leqslant i \leqslant n} y_{ij}$. For simplicity, we assume that the data only take positive values, which implies that K is upper-bounded by $\max_{1 \leqslant i \leqslant n} y_{ij}$, whose asymptotic distribution is studied in extreme value theory. Since this asymptotic distribution is model specific, we use the rounded multivariate Gaussian model of Canale and Dunson (2011) as an illustration. This model is a special case of the multivariate latent Gaussian model defined in (4.1) with $h_j(y_j|x_j) = \mathbb{1}\{x_j > 0\}\lceil x_j \rceil$. Basically $h_j(y_j|x_j)$ is a rounding function that rounds x_j to the smallest integer larger than it while mapping all x_j below 0 to 0.

Lemma 4.2.8. *Consider model (4.1) with* $h_j(y_j|x_j) = \mathbb{1}\{x_j > 0\}\lceil x_j \rceil$*, for* $j = 1, \ldots, p$*, we have that* $\forall \delta > 0$*,* $\exists N \in \mathbb{N}_1$ *such that* $\forall n > N$*,*

$$pr\left[\max_{1 \leqslant i \leqslant n} y_{ij} < \sqrt{\sigma_{jj}}\left(\frac{2\delta}{\sqrt{\log n}} + \sqrt{2\log n}\right) + \mu_j + 1\right] > e^{-\exp(-\delta/2)}. \tag{4.12}$$

Intuitively, (4.12) implies that K is at most $O(\sqrt{\log n})$ with high probability. Similarly, one can show that the computational complexity of evaluating the data likelihood of any bivariate marginal distribution is at most $O(\log n)$.

To summarize, we have shown that the per-iteration computational complexity is linear in the cardinality of the discrete observations. This cardinality can be bounded by the data maxima; hence its asymptotic distribution can be analyzed using standard extreme value theory. We have shown that the per-iteration complexity is at most $O(\log n)$ with high probability for the rounded multivariate Gaussian model.

4.3 Experiments

The performance of *Bayesian Mosaic* will be illustrated via two simulation studies and a citation network application. The first simulation study demonstrates the superiority of *Bayesian Mosaic* over DA-MCMC for imbalanced count data. The second simulation study demonstrates that *Bayesian Mosaic* achieves similar accuracy with a provably more scalable computational complexity for large balanced count data. *Bayesian mosaic* is also applied to a citation count dataset to infer the overlapping structure of a group of researchers' interests.

All experiments are conducted in R on a machine with 12 3.50 GHz Intel(R) Xeon(R) CPU E5-1650 v3 processors. All results are based on 100 replicate experiments.

4.3.1 Multivariate log-Gaussian Mixture of Poisson

In the first simulation study, we considered a special case of multivariate latent Gaussian models with $h_j(y_j|x_j)$ being the density function of a Poisson distribution whose rate parameter equals e^{x_j}. We generated 100 datasets for each unique data dimensionality p in $\{3, 5, 7\}$. We fixed the sample size to be 10000. For each synthetic dataset we randomly generated $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ from some distribution so that the simulated dataset has an excessive amount of zeros. More specifically, for $j = 1, \dots, p$, we generated μ_j from Unif$(-4, -3)$ and σ_{jj} from Unif$(0.5, 1)$. We randomly generated a correlation matrix from the standard LKJ distribution Lewandowski et al. (2009) and then combined this correlation matrix with $\sigma_{11}, \dots, \sigma_{pp}$ into $\boldsymbol{\Sigma}$. Roughly 90% of the simulated data entries are zeros.

We used weakly-informative priors in both simulation studies. Specifically, for $j = 1, \dots, p$,

$$\pi_{jj}(\mu_j, \sigma_{jj}) \propto \sigma_{jj}^{-1/2} \mathbb{1}\left\{|\mu_j| < A, 0 < \sigma_{jj} < B\right\},$$

Table 4.1: Mean Square Error Comparison.[1]

		$p = 3$	$p = 5$	$p = 7$
Bayesian Mosaic	ρ	6.79 (9.76)	5.78 (8.14)	5.62 (7.92)
	s	5.9 (9.54)	5.86 (9.13)	5.86 (8.63)
	μ	1.74 (2.39)	1.74 (2.71)	1.7 (2.52)
DA-MCMC	ρ	8.41 (15.3)	9.93 (14.8)	10 (14.4)
	s	150 (1892)	303 (3020)	443 (3744)
	μ	54.9 (345)	123 (522)	126 (571)

where $A > 0$ and $B > 0$. For $1 \leqslant s < t \leqslant p$,

$$\pi_{jj}(\sigma_{st}|\mu_s, \sigma_{ss}, \mu_t, \sigma_{tt}) \propto \mathbb{1}\left\{|\sigma_{st}| < \sqrt{\sigma_{ss}\sigma_{tt}}\right\}.$$

The propriety of the posterior is guaranteed since the support of the prior is compact. Moreover, for sufficiently large A and B, the posterior becomes insensitive to the choice of A and B (Gelman et al., 2006). We let $A = 100$ and $B = 10$. We used a similar prior specification in citation count application.

Normal independent MH sampler was implemented for sampling the *knot marginals* for 200 iterations with the first 100 as burn-in. We then approximated *tile conditionals* via Laplace approximation and drew 100 samples of the *tiles* from the resulting conditional Gaussian distribution given the previous draws of the *knots*. On average, a single run with the computation distributed to 11 parallel workers took 90 seconds for $p = 3$, 126 seconds for $p = 5$ and 227 seconds for $p = 7$. As a comparison, we ran DA-MCMC sampler for the 5 times the amount of time with the computation tasks within each iteration distributed to 11 parallel workers as well. In both simulation studies, we gave DA-MCMC an unfair advantage by initializing the parameter values at the true values.

We first compared accuracies of estimating the model parameters w.r.t. square error loss. Average MSE within each group are presented in Table 4.1, where the number in the parenthesis is the standard error. It can be clearly seen that the

---

[1] All numbers have been multiplied by 100.

Table 4.2: Empirical Coverage Comparison

		$p = 3$	$p = 5$	$p = 7$
Bayesian Mosaic	ρ	95%	93.9%	93%
	s	93.7%	93.4%	94.1%
	μ	93%	92.6%	93.7%
DA-MCMC	ρ	64%	56.8%	59.3%
	s	41.7%	32.4%	31.1%
	μ	37%	24.8%	24.9%

estimates based on *Bayesian mosaic* outperforms those based on DA-MCMC samples in terms of square error loss.

We evaluated *Bayesian mosaic*'s performance in quantifying the uncertainty through the empirical coverage (EC) of credible intervals. Average EC's are presented in Table 4.2. The empirical coverages of *Bayesian mosaic* are close to 95%, indicating good uncertainty quantification, whereas the empirical coverages based on DA-MCMC are terribly off.

4.3.2 Rounded Multivariate Gaussian

In the second simulation study, we considered the rounded multivariate Gaussian model (Canale and Dunson, 2011) given in §4.2.4. We fixed the sample size to be 10000, data dimensionality $p = 4$ and generated 100 datasets. For each synthetic dataset we randomly generated $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ from some distribution so that the simulated data are well balanced (majority of the data entries are non-zero). More specifically, for $j = 1, \ldots, p$, we generated μ_j from Unif$(4, 5)$ and σ_{jj} from Unif$(1, 1.5)$. The analysis was done exactly as in §4.3.1.

The average MSE and EC are summarized in Table 4.3. DA-MCMC seems to do slightly better than *Bayesian mosaic*. But the difference in performance is marginal. Due to the limitation of computation power for DA-MCMC, we did not do experiments with larger sample size n. According to our discussion in §4.2.4, the per-iteration computational complexity of *Bayesian mosaic* is roughly $O(\log n)$ while

Table 4.3: Performance Comparison

		MSE[2]	EC
Bayesian Mosaic	ρ	1.07 (1.59)	90.2%
	s	3.23 (4.33)	94%
	μ	1.45 (1.89)	91.5%
DA-MCMC	ρ	0.91 (1.37)	94.2%
	s	3.29 (4.34)	95.3%
	μ	1.45 (1.91)	91.3%

that of DA-MCMC is $O(n)$. This implies that *Bayesian mosaic* should be favored in large sample size applications even if data are well balanced.

4.3.3 Citation Network Application

In this study, we considered a real-world citation network dataset (Tang et al., 2008) that contains papers and citation relationships from a computer science bibliography website called DBLP. Our goal is to study the overlapping structure of a group of researchers' interests. Intuitively, two researchers who have many research interests in common tend to be cited together more frequently. Meanwhile, we also want to see how is the research impact of these researchers varying in time. We hand-picked 11 active researchers[3] in the machine learning community.

In processing the database, we focused on the machine learning literature and removed irrelevant papers. When counting the number of citations, we ignored papers co-authored by multiple researchers in our hand-picked group. The final dataset contains roughly 80000 11-dimensional observations with each one being the number of citations of a certain paper go to each of the 11 researchers. We used i as the index for papers and j as the index for the researchers. Letting t_i be the year paper i was published and n_{jt} be the total number of publications of researcher j up to

---

[2] All numbers have been multiplied by 10^4.

[3] Michael Jordan, Robert Brunner, Yann LeCun, Andrew McCallum, Chih-Jen Lin, Christopher Bishop, Yoshua Bengio, David Blei, Padhraic Smyth, Richard Sutton, Guillermo Sapiro

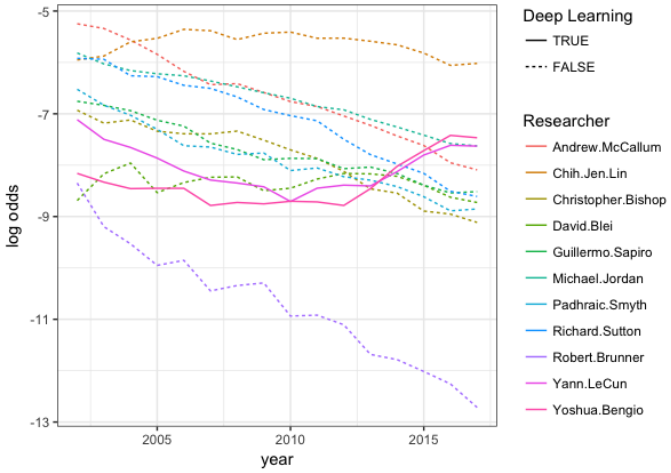

FIGURE 4.2: Visualizing the posterior mean of $\boldsymbol{\mu}_t$'s by researchers.

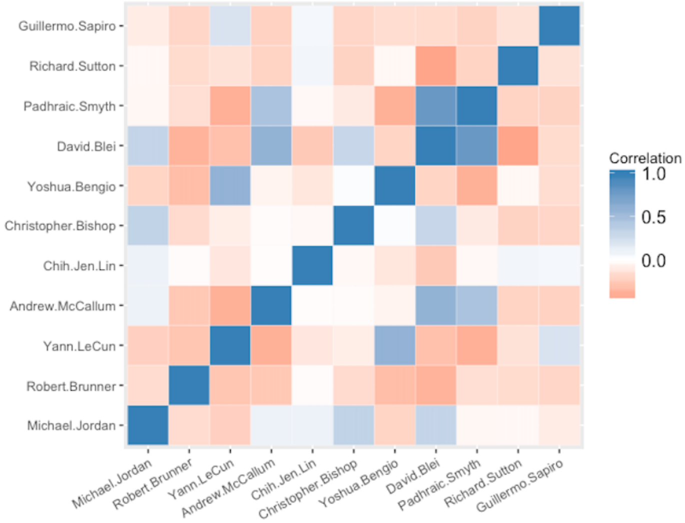

FIGURE 4.3: Visualizing the correlation matrix induced by $\boldsymbol{\Sigma}$.

year t, we used the following model

$$y_{ij} \overset{ind}{\sim} \text{Binomial}\left(n_{jt_i}, \text{logit}^{-1}(x_{ij})\right) \text{ for } j = 1, \ldots, p,$$

$$\boldsymbol{x}_i \overset{ind}{\sim} N(\boldsymbol{\mu}_{t_i}, \boldsymbol{\Sigma}), \quad \boldsymbol{\mu}_t \overset{iid}{\sim} N(\boldsymbol{\mu}_0, \boldsymbol{D}),$$

where $\boldsymbol{x}_i = \left(x_{i1}, \ldots, x_{ip}\right)^\top$, $p = 11$ and $\boldsymbol{D}$ is a diagonal matrix with the diagonal elements being positive. The model parameters are $\boldsymbol{\mu}_0$, $\boldsymbol{\Sigma}$ and $\boldsymbol{D}$ whereas $\boldsymbol{\mu}_t$'s are random effects.

After integrating out $\boldsymbol{\mu}_t$'s and $\boldsymbol{x}_i$'s, $\boldsymbol{y}_i$'s are no longer independent. It is easy to check that the above model is *mosaic-type* in the generalized *Bayesian mosaic* framework of §4.1.6. Normal random walk MH sampler was implemented for sampling the *knot marginals* for 40000 iterations with the first 20000 as burn-in and thinning the rest into 500 samples. In sampling the *tiles*, we used the plug-in approach discussed in §4.1.4 and sampled from the resulted *tile conditionals* via MH. For each *tile*, we ran the MH sampler 10000 iterations with the first 5000 as burn-in and thinned the rest into 500 final samples. The entire sampling process took around 5 hours with the jobs distributed to 11 parallel workers.

Figure 4.2 visualizes the posterior mean of the random effects $\boldsymbol{\mu}_t$'s by different researchers. Intuitively, $\boldsymbol{\mu}_t$ is a vector of the average log-odds of a single paper citing these researchers. Interestingly, while most of the researchers' log-odds of being cited is decreasing, the only two exceptions are both working on deep learning.

We also computed the posterior mean of $\boldsymbol{\Sigma}$ (after correction). The induced correlation matrix is visualized via a heatmap in Figure 4.3. Clearly, some researchers are more likely to be cited together compared to the others, indicating their strong overlapping research interests. For instance, Yann Lecun and Yoshua Bengio have a stronger correlation since they are both studying deep learning. There are researchers whose research interests seem to overlap with many others, e.g., David Blei. Also, there are researchers whose research interests seem to be unique in this selected group, e.g., Richard Sutton.

5

Conclusion

Manifold learning has dramatic importance in many applications where high-dimensional data are collected with unknown low dimensional manifold structure. While most of the methods focus on finding lower dimensional summaries or characterizing the joint distribution of the data, there is (to our knowledge) no reliable method for probabilistic learning of the manifold. This turns out to be a daunting problem due to major issues with identifiability leading to unstable and generally poor performance for current probabilistic non-linear dimensionality reduction methods. It is not obvious how to incorporate appropriate geometric constraints to ensure identifiability of the manifold without also enforcing overly-restrictive assumptions about its form. We tackled this problem in the one-dimensional manifold (curve) case and built a novel electrostatic Gaussian process model based on the general framework of GP-LVM by introducing a novel Coulomb repulsive process. Both simulations and real world data experiments showed excellent performance of the proposed model in accurately estimating the manifold while characterizing uncertainty. Indeed, performance gains relative to competitors were dramatic. The proposed electroGP is shown to be applicable to many learning problems including video-inpainting, super-resolution and video-denoising. There are many interesting areas for future study including the development of efficient algorithms for applying the model for multidimensional manifolds, while learning the dimension.

In many applications, high-dimensional data with unknown joint distribution are collected. Despite the dramatic importance of learning the joint distribution of such data, few probabilistic methods that scale well to high-dimension and provide an adequate characterization of uncertainty are available. Bayesian nonparametric methods based on mixtures of multivariate Gaussian kernels are widely used, but face major bottlenecks in scaling to higher dimensions. To tackle this problem, we proposed an empirical Bayes density estimator combining manifold learning and Bayesian nonparametric density estimation. One of the building blocks of our method focuses on single Gaussian factor decomposition in which variables are linearly related, showing excellent performance in scaling computationally and in generalization error, while providing a valid characterization of uncertainty in predictions. The other building block is a multiscale mixture generalization, which accommodates unknown density, nonlinear relationships and nonlinear subspaces. This approach showed excellent performance in inferring the subspace dimension, estimating the subspace, and characterizing the joint density of the data in the ambient space. The proposed methods are broadly applicable to many learning problems including regression or classification with missing features.

While developing *Bayesian mosaic*, we are focusing on the multivariate latent Gaussian models, which offer great flexibility in modeling high-dimensional discrete data. Unfortunately, their use is limited by the computationally challenging model fitting. We tackle this problem by proposing a surrogate composite posterior termed as *Bayesian mosaic* whose sampling can be easily parallelized. *Bayesian mosaic* is consistent and asymptotically normal under mild conditions. It showed excellent performance in terms of not only computation but also parameter estimation accuracy and uncertainty quantification. A generalization is discussed to handle dependent data.

Appendix A

Appendix to Chapter 3

A.1 Proof of Theorem 1

The log-posterior of GEODE, up to a additive constant, is as follow

$$\mathcal{L} = -\frac{N}{2}\big\{\ln|\boldsymbol{C}| + \mathrm{tr}(\boldsymbol{C}^{-1}\boldsymbol{S}) -$$

$$\tilde{a}\ln(\sigma^2) + \tilde{b}\sigma^{-2}\big\},$$

where $\tilde{a} = \frac{a+1}{2N}$ and $\tilde{b} = \frac{b}{2N}$ Given the empirical Bayes problem

$$(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{W}}) = \arg\max_{\boldsymbol{\mu},\boldsymbol{W}} \big[\max_{\sigma^2,\boldsymbol{\Sigma}} \mathcal{L}(\boldsymbol{\mu}, \boldsymbol{W}, \sigma^2, \boldsymbol{\Sigma})\big]. \tag{A.1}$$

Theorem A.1.1. *Suppose*

Condition 1:

$$d < rank(\boldsymbol{S})$$

Condition 2:

$$\tilde{a}\lambda_D \leqslant \sum_{j=d+1}^{D} \lambda_j - (D-d)\lambda_D \tag{A.2}$$

Condition 3: For all k such that $k \leqslant d+1$,

$$\tilde{b} < \max\left\{(D-d)\lambda_k - \sum_{j=k}^{k+D-d-1}\lambda_j, 0\right\} \tag{A.3}$$

Then

$$\hat{\boldsymbol{\mu}} = \bar{\boldsymbol{y}},\ \hat{\boldsymbol{W}} = \boldsymbol{U}_p$$

solves (A.1), where $\lambda_1, \dots, \lambda_D$ are the eigenvalues of $\boldsymbol{S}$ in a descending order and the p column vectors in the $D \times p$ matrix $\boldsymbol{U}_p$ are the p leading eigenvectors of $\boldsymbol{S}$.

Proof. The proof can be split into two parts.

Part 1: $\hat{\boldsymbol{\mu}} = \bar{\boldsymbol{y}}$ and $\hat{\boldsymbol{W}} = \boldsymbol{U}_p$ is the stationary point of $\mathcal{L}$.

From standard matrix differentiation results:

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\mu}} = -N\left\{\sum_{i=1}^{N}\boldsymbol{C}^{-1}(\boldsymbol{\mu} - \boldsymbol{y}_i)\right\}$$

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{W}} = -\frac{N}{2}\left\{2\boldsymbol{C}^{-1}\boldsymbol{W}\boldsymbol{\Sigma} - 2\boldsymbol{C}^{-1}\boldsymbol{S}\boldsymbol{C}^{-1}\boldsymbol{W}\boldsymbol{\Sigma}\right\}$$

And solving for $\frac{\partial \mathcal{L}}{\partial \boldsymbol{\mu}} = 0$ gives $\boldsymbol{\mu} = \bar{\boldsymbol{y}}$. Solving for $\frac{\partial \mathcal{L}}{\partial \boldsymbol{W}} = 0$ gives

$$\boldsymbol{C}^{-1}\boldsymbol{S}\boldsymbol{C}^{-1}\boldsymbol{W}\boldsymbol{\Sigma} = \boldsymbol{C}^{-1}\boldsymbol{W}\boldsymbol{\Sigma}$$

$$\Leftrightarrow \boldsymbol{S}\boldsymbol{C}^{-1}\boldsymbol{W} = \boldsymbol{W} \tag{A.4}$$

Neither of the two trivial solutions to (A.4), $\boldsymbol{W} = \boldsymbol{0}$ and $\boldsymbol{C} = \boldsymbol{S}$ maximizes $\mathcal{L}$ hence will not be discussed. The left solution corresponds to a $\boldsymbol{W}$ such that $\boldsymbol{W} \neq \boldsymbol{0}$ and $\boldsymbol{C} \neq \boldsymbol{S}$. With the fact that column vectors of $\boldsymbol{W}$ are orthonormal, we have

$$\boldsymbol{S}\boldsymbol{C}^{-1}\boldsymbol{W} = \boldsymbol{W}$$

$$\Leftrightarrow \boldsymbol{S}\big[\sigma^{-2}\boldsymbol{I} - \sigma^{-2}\boldsymbol{W}(\sigma^2\boldsymbol{I} + \boldsymbol{\Sigma})^{-1}\boldsymbol{\Sigma}\boldsymbol{W}^\top\big]\boldsymbol{W} = \boldsymbol{W}$$

$$\Leftrightarrow \boldsymbol{S}\boldsymbol{W} = (\sigma^2\boldsymbol{I} + \boldsymbol{\Sigma})\boldsymbol{W} \tag{A.5}$$

Equation (A.5) implies that each column of $\boldsymbol{W}$ must be an eigenvector of $\boldsymbol{S}$, with corresponding eigenvalues $\gamma_j = \sigma^2 + \alpha_j^2$. Note that this also implies that $\sigma^2 \leqslant \gamma_j$ for $j = 1, \ldots, d$.

Now we check if $\frac{\partial \mathcal{L}}{\partial \alpha_j^2}|_{\alpha_j^2=\gamma_j-\sigma^2} = 0$, which will complete the proof that $\hat{\boldsymbol{\mu}} = \bar{\boldsymbol{y}}$ and $\hat{\boldsymbol{W}} = \boldsymbol{U}_p$ is the stationary point of $\mathcal{L}$.

We substitute stationary point of $\boldsymbol{W}$ into $\mathcal{L}$ to give

$$\begin{aligned} \mathcal{L} = -\frac{N}{2}\bigg\{(D-d)\ln(\sigma^2) + \sum_{j=1}^{d}\ln(\sigma^2+\alpha_j^2) + \frac{1}{\sigma^2}\sum_{j=1}^{D}\gamma_j \\ -\frac{1}{\sigma^2}\sum_{j=1}^{d}\frac{\gamma_j\alpha_j^2}{\sigma^2+\alpha_j^2} + \tilde{a}\ln(\sigma^2) + \tilde{b}\sigma^{-2}\bigg\}. \end{aligned} \tag{A.6}$$

From (A.6) one can easily check $\frac{\partial \mathcal{L}}{\partial \alpha_j^2}|_{\alpha_j^2=\gamma_j-\sigma^2} = 0$, for $j = 1, \ldots, d$.

Part 2: Show $\hat{\boldsymbol{\mu}} = \bar{\boldsymbol{y}}$ and $\hat{\boldsymbol{W}} = \boldsymbol{U}_p$ maximizes $\mathcal{L}$.

Matrix $\boldsymbol{W}$ may contain any of the eigenvectors of $\boldsymbol{S}$. To figure out when is $\mathcal{L}$ maximized, we substitute stationary point of $\boldsymbol{W}$ and $\boldsymbol{\Sigma}$ into $\mathcal{L}$ to give

$$\begin{aligned} \mathcal{L} = -\frac{N}{2}\bigg\{\sum_{j=1}^{d}\ln(\gamma_j) + \frac{1}{\sigma^2}\sum_{j=d+1}^{D}\gamma_j + (D-d)\ln\sigma^2 + d \\ + \tilde{a}\ln\sigma^2 + \tilde{b}\frac{1}{\sigma^2}\bigg\}, \end{aligned} \tag{A.7}$$

where $\gamma_1, \ldots, \gamma_d$ are the eigenvalues corresponding to the eigenvectors 'retained' in $\boldsymbol{W}$ and $\gamma_{d+1}, \ldots, \gamma_D$ are those 'discarded'. Here we slightly abuse notations: we use $\lambda_1, \ldots, \lambda_D$ as the eigenvalues of $\boldsymbol{S}$ ordered descendingly. We use $\gamma_1, \ldots, \gamma_D$ also as the eigenvalues of $\boldsymbol{S}$ but with the first d corresponding to the stationary point $\boldsymbol{W}$. Note that γ_j's are not necessarily ordered, at least we do not know yet.

Maximizing (A.7) w.r.t. σ^2 gives

$$\sigma^2 = \frac{1}{D-d+\tilde{a}}\bigg(\sum_{j=d+1}^{D}\gamma_j + \tilde{b}\bigg) > 0. \tag{A.8}$$

When $\lambda_D > 0$, with (A.2), it is easy to check that

$$\frac{1}{D-d+\tilde{a}}\left(\sum_{j=d}^{D-1}\lambda_j+\tilde{b}\right) > \lambda_D$$

Since $\sigma^2 \leqslant \gamma_j$, for $j = 1, \ldots, d$, we know immediately that λ_D has to be discarded. If $\lambda_D = 0$, it is obvious that it also has to be discarded since $\sigma^2 > 0$. Note that Condition 1 ensures the existence of b_σ that satisfies condition 3. Condition 3 ensures the the existence of the stationary point to $\mathcal{L}$ since

$$\frac{1}{D-d+\tilde{a}}\left(\sum_{j=d+1}^{D}\lambda_j+\tilde{b}\right) < \lambda_d,$$

which means that we at least have one stationary point solution.

Substituting σ^2 w.r.t. (A.8) gives

$$\begin{aligned}\mathcal{L} = -\frac{N}{2}\Bigg\{&\sum_{j=1}^{d}\ln(\gamma_j)+D+\tilde{a}+\\ &(D-d+\tilde{a})\ln\left[\frac{1}{D-d+\tilde{a}}\left(\sum_{j=d+1}^{D}\gamma_j+\tilde{b}\right)\right]\Bigg\}.\end{aligned} \tag{A.9}$$

When all eigenvalues are non-zero, with the fact that the sum of all the eigenvalues is a constant, maximizing (A.9) is equivalent to minimizing the following quantity

$$E = \ln\left[\frac{1}{D-d+\tilde{a}}\left(\sum_{j=d+1}^{D}\gamma_j+\tilde{b}\right)\right] - \frac{1}{D-d+\tilde{a}}\left(\sum_{j=d+1}^{D}\ln(\gamma_j)+\tilde{b}\right) \tag{A.10}$$

When there are zero eigenvalues, we simply ignore these zero eigenvalues and restrict us only to the non-zeros ones and very statement in this proof hold.

The first order derivative of E is given by

$$\frac{\partial E}{\partial \gamma_j} = \frac{1}{\sum_{j=d+1}^{D}\gamma_j+\tilde{b}} - \frac{1}{(D-d+\tilde{a})\gamma_j} \tag{A.11}$$

Suppose $\gamma_{d+1}, \ldots, \gamma_D$ minimizes (A.10). Without loss of generality, we assume $\gamma_{d+1} \geqslant \ldots \geqslant \gamma_D$. From the previous discussion we know that $\gamma_D = \lambda_D$. We want to show that $\gamma_{d+1}, \ldots, \gamma_D$ *have to be adjacent within the spectrum of the ordered eigenvalues of* $\boldsymbol{S}$. Define

$$c = \begin{cases} \sum_{j=d+2}^{D-1} \gamma_j, & \text{if } D-d \geqslant 2 \\ 0, & \text{if } D-d = 1 \end{cases}$$

We don't need to discuss the case $D-d=1$. Since if that is the case then $\gamma_1, \ldots, \gamma_d$ must be the d leading eigenvalues because λ_D has to be discarded.

From (A.11) we immediately have

$$\begin{aligned} \frac{\partial E}{\partial \gamma_{d+1}} =& \frac{1}{\gamma_{d+1} + \gamma_D + c + \tilde{b}} - \frac{1}{(D-d+\tilde{a})\gamma_{d+1}} \\ \frac{\partial E}{\partial \gamma_D} =& \frac{1}{\gamma_D + \gamma_{d+1} + c + \tilde{b}} - \frac{1}{(D-d+\tilde{a})\gamma_D} \end{aligned}$$

From (A.3) it is easy to check that

$$\gamma_D + \gamma_{d+1} + c + \tilde{b} < (D-d)\lambda_{d+1}.$$

hence $\frac{\partial E}{\partial \gamma_{d+1}} > 0$. Similarly, with condition (A.2) one can also check that $\frac{\partial E}{\partial \gamma_D} < 0$. It can also be checked that for any λ_{d+1} such that $\lambda_{d+1} > \frac{\lambda_D + c + \tilde{b}}{D-d+\tilde{a}-1}$, $\frac{\partial E}{\partial \gamma_{d+1}} > 0$ holds. And for any λ_D such that $\lambda_D < \frac{\lambda_{d+1} + c + \tilde{b}}{D-d+\tilde{a}-1}$, $\frac{\partial E}{\partial \gamma_D} < 0$ holds. Moreover, $\frac{\lambda_{d+1} + c + \tilde{b}}{D-d+\tilde{a}-1} > \frac{\lambda_D + c + \tilde{b}}{D-d+\tilde{a}-1}$. Hence $\gamma_{d+1}, \ldots, \gamma_D$ *have to be adjacent within the spectrum of the ordered eigenvalues of* $\boldsymbol{S}$. Otherwise, if there is a γ in between such that $\gamma_D < \gamma < \gamma_{d+1}$ then either $\frac{\partial E}{\partial \gamma_D}|_{\gamma_D = \gamma} < 0$ or $\frac{\partial E}{\partial \gamma_{d+1}}|_{\gamma_{d+1} = \gamma} > 0$ must be true. Hence either replacing γ_D or γ_{d+1} with γ will further decrease $\mathcal{L}$, which contradicts the assumption that $\gamma_{d+1}, \ldots, \gamma_D$ minimizes $\mathcal{L}$.

Coupled with the fact that $\gamma_D = \lambda_D$, we have shown that the $D-d$ smallest eigenvalues minimizes (A.10) hence $\mathcal{L}$ is maximized if $\gamma_1, \ldots, \gamma_d$ are the d leading eigenvalues of $\boldsymbol{S}$. Hence we have

$$\hat{\boldsymbol{\mu}} = \bar{\boldsymbol{y}}$$

and substituting $\boldsymbol{\mu} = \bar{\boldsymbol{y}}$ into $\boldsymbol{S}$ we have $\boldsymbol{S} = \boldsymbol{Y}\boldsymbol{Y}^\top$ hence we have

$$\hat{\boldsymbol{W}} = \boldsymbol{U}_d$$

□

A.2 Formulation

To illustrate the binary clustering tree, a 4–level binary clustering tree of a synthetic parabola point cloud obtained using GMRA can be found in Figure A.1, which visualizes a 4 level binary tree decomposition of a parabola using METIS, with the black rectangular denoting the second level cells, the red denoting the third level cells and the green denoting the leaf cells.

The likelihood function of GEODE can be written as

$$\begin{aligned} f_{s,h}(\boldsymbol{y}_i) \propto (\sigma_s^2)^{-D/2} \prod_{m=1}^{d} u_{s,h,m}^{1/2} \times \exp\Bigg\{ -\frac{1}{2}\sigma_s^{-2} \\ \Big[A_{s,h,i} - \sum_{m=1}^{d}(1-u_{s,h,m})(Z_{s,h,i}^{(m)})^2\Big]\Bigg\}, \end{aligned} \tag{A1}$$

which can be derived using the following two propositions.

Proposition A.2.1. *$\boldsymbol{\Sigma} = diag(\alpha_1^2, \dots, \alpha_d^2)$ is a $d \times d$ matrix with all diagonal entries larger than 0, $\boldsymbol{\Phi}$ is a $D \times d$ orthonormal matrix, we have,*

$$(\sigma^2 \boldsymbol{I} + \boldsymbol{\Phi}\boldsymbol{\Sigma}\boldsymbol{\Phi}^T) = \sigma^{-2}\boldsymbol{I} - \sigma^{-4}\boldsymbol{\Phi}\tilde{\boldsymbol{\Sigma}}\boldsymbol{\Phi}^T,$$

where $\tilde{\boldsymbol{\Sigma}} = diag(\frac{\alpha_1^2}{1+\sigma^{-2}\alpha_1^2}, \frac{\alpha_2^2}{1+\sigma^{-2}\alpha_2^2}, \dots, \frac{\alpha_d^2}{1+\sigma^{-2}\alpha_d^2})$.

Proof. By the orthonormality of the dictionary, we have $\boldsymbol{\Phi}^T\boldsymbol{\Phi} = \boldsymbol{I}_d$. And by the matrix inversion formula,

$$\begin{aligned}(\sigma^2 I + \boldsymbol{\Phi}\boldsymbol{\Sigma}\boldsymbol{\Phi}^T)^{-1} &= \sigma^{-2}\boldsymbol{I} - \sigma^{-4}\boldsymbol{\Phi}(\boldsymbol{I} + \sigma^{-2}\boldsymbol{\Sigma}\boldsymbol{\Phi}^T\boldsymbol{\Phi})^{-1}\boldsymbol{\Sigma}\boldsymbol{\Phi}^T \\ &= \sigma^{-2}\boldsymbol{I} - \sigma^{-4}\boldsymbol{\Phi}(\boldsymbol{I} + \sigma^{-2}\boldsymbol{\Sigma})^{-1}\boldsymbol{\Sigma}\boldsymbol{\Phi}^T \\ &= \sigma^{-2}\boldsymbol{I} - \sigma^{-4}\boldsymbol{\Phi}\tilde{\boldsymbol{\Sigma}}\boldsymbol{\Phi}^T\end{aligned}$$

□

Proposition A.2.2. *Under the same setting of Proposition A.2.1, we have*

$$|\sigma^2\boldsymbol{I} + \boldsymbol{\Phi}\boldsymbol{\Sigma}\boldsymbol{\Phi}^T|^{-1/2} = (\sigma^2)^{-D/2}\prod_{m=1}^{d}(\frac{1}{1+\sigma^{-2}\alpha_m^2})^{1/2}.$$

Proof. By Theorem Schur's formula,

$$\begin{aligned}|\sigma^2\boldsymbol{I} + \boldsymbol{\Phi}\boldsymbol{\Sigma}\boldsymbol{\Phi}^T|^{-1/2} &= (\sigma^2)^{-D/2}|\boldsymbol{I}_D + \sigma^{-2}\boldsymbol{\Phi}\boldsymbol{\Sigma}\boldsymbol{\Phi}^T|^{-1/2} \\ &= (\sigma^2)^{-D/2}|\boldsymbol{I}_d + \sigma^{-2}\boldsymbol{\Sigma}^{1/2}\boldsymbol{\Phi}^T\boldsymbol{\Phi}\boldsymbol{\Sigma}^{1/2}|^{-1/2} \\ &= (\sigma^2)^{-D/2}|\boldsymbol{I}_d + \sigma^{-2}\boldsymbol{\Sigma}| \\ &= (\sigma^2)^{-D/2}\prod_{m=1}^{d}(\frac{1}{1+\sigma^{-2}\alpha_m^2})^{1/2}\end{aligned}$$

□

Theorem A.2.3. *Assume* $\boldsymbol{\Omega}_{s,h} = \boldsymbol{\Psi}\boldsymbol{\Sigma}_{s,h}\boldsymbol{\Psi}^T + \sigma_s^2\boldsymbol{I}$ *where* $\boldsymbol{\Psi}$ *is a orthonormal* $D \times D$ *matrix and* $\boldsymbol{\Sigma}_{s,h}$ *is a* $D \times D$ *positive diagonal matrix. The distributions of* $\boldsymbol{\Sigma}_{s,h}$ *and* σ_s^2 *are defined in (6) and (7) in the submitted paper. Let* $\boldsymbol{\Psi}^d$ *denote the first* d *columns of* $\boldsymbol{\Psi}$, $\boldsymbol{\Sigma}_{s,h}^d = \mathrm{diag}(\alpha_{s,h,1}^2, \ldots, \alpha_{s,h,d}^2)$ *and let* $\boldsymbol{\Omega}_{s,h}^d = \boldsymbol{\Psi}^d\boldsymbol{\Sigma}_{s,h}^d(\boldsymbol{\Psi}^d)^T + \sigma_s^2\boldsymbol{I}$. *Then for any* $\epsilon > 0$,

$$Pr\{d_\infty(\boldsymbol{\Omega}_{s,h}, \boldsymbol{\Omega}_{s,h}^d) > \epsilon\} < \frac{6ba^d}{\epsilon(1-a)}$$

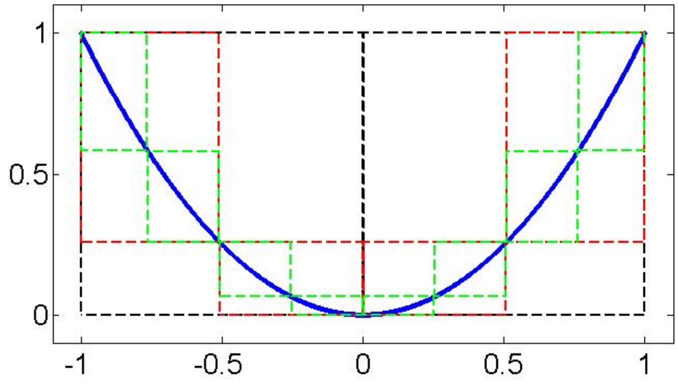

FIGURE A.1: An example of a binary clustering tree.

for $d > 2\log\{b/\epsilon(1-a)\}/\log(1/a)$, *where* $d_\infty(\mathbf{\Omega}_{s,h}, \mathbf{\Omega}_{s,h}^d)$ *is defined as* $\|\mathbf{\Omega}_{s,h} - \mathbf{\Omega}_{s,h}^d\|_\infty$. $\|A\|_\infty$ *calculates the maximum absolute row sum of the matrix* A, $b = E(\sigma_s^2)$ *and* $a = E(\frac{1}{\tau_{s,h,1}})$.

Proof. With a slight abuse of notation, we write $u_{s,h,k}$ as u and let $A = \prod_{m=1}^K \tau_{s,h,m}$. Let $\triangle_d = \mathbf{\Psi}\mathbf{\Sigma}_{s,h}\mathbf{\Psi}^T - \mathbf{\Psi}^d\mathbf{\Sigma}_{s,h}^d(\mathbf{\Psi}^d)^T$, $\triangle_d = \{a_{i.j}\}$ and $\mathbf{\Psi} = \{\psi_{i,j}\}$. Clearly, $d_\infty(\mathbf{\Omega}_{s,h}, \mathbf{\Omega}_{s,h}^d) = \max_{1\leqslant i,j\leqslant D} |a_{i,j}^d|$, and $a_{i,j}^d = \sum_{k=d+1}^D \alpha_k^2\psi_{i,k}\psi_{j,k}$. By Cauchy-Schwartz inequality,

$$|\sum_{k=d+1}^{D} \alpha_k^2\psi_{i,k}\psi_{j,k}| \leqslant \max_{1\leqslant m\leqslant D}(\sum_{k=H+1}^{D} \alpha_k^2\psi_{m,k}^2).$$

Since $\mathbf{\Psi}$ is orthonormal, we have $\psi_{i,j}^2 \leqslant 1$ for any i and j. Hence

$$d_\infty(\mathbf{\Omega}_{s,h}, \mathbf{\Omega}_{s,h}^d) \leqslant \sum_{k=d+1}^{D} \alpha_k^2.$$

For a fixed $\epsilon > 0$, by Chebyshev's inequalities

$$\begin{aligned} p\{d_\infty(\boldsymbol{\Omega}_{s,h}, \boldsymbol{\Omega}^d_{s,h}) \leqslant \epsilon\} &\geqslant p\left\{\sum_{k=d+1}^{D} \alpha_k^2 \leqslant \epsilon\right\} \\ &= E\left\{p(\sum_{k=d+1}^{D} \alpha_k^2 \leqslant \epsilon|\tau)\right\} \\ &= 1 - E\left\{p(\sum_{k=d+1}^{D} \alpha_k^2 > \epsilon|\tau)\right\} \\ &\geqslant 1 - E\left\{\frac{E(\sum_{k=d+1}^{D} \alpha_k^2|\tau)}{\epsilon}\right\}. \end{aligned}$$

By design we have $u \sim \mathrm{Ga}_{(0,1)}(A+1, 1)$ and u and σ_s^2 are conditionally independent, hence

$$E[(\frac{1}{u} - 1)\sigma_s^2|\tau] = E[(\frac{1}{u} - 1)|\tau]E(\sigma_s^2).$$

Then we have

$$\begin{aligned} E[(\frac{1}{u} - 1)|\tau] &= \frac{\int_0^1 (1/u - 1)\frac{u^A}{\Gamma(A+1)} e^{-u}\mathrm{d}u}{\int_0^1 \frac{u^A}{\Gamma(A+1)} e^{-u}\mathrm{d}u} = \frac{\int_0^1 1/u \times u^A e^{-u}\mathrm{d}u}{\int_0^1 u^A e^{-u}\mathrm{d}u} - 1 \\ &= \frac{\int_0^1 u^{A-1} e^{-u}\mathrm{d}u}{\int_0^1 u^A e^{-u}\mathrm{d}u} - 1 = \frac{\frac{1}{A} u^A e^{-u}|_0^1 + \int_0^1 \frac{1}{A} u^A e^{-u}\mathrm{d}u}{\int_0^1 u^A e^{-u}\mathrm{d}u} - 1 \\ &= \frac{e^{-1}}{A\int_0^1 u^A e^{-u}\mathrm{d}u} - 1 + \frac{1}{A}. \end{aligned}$$

Let $\gamma(s,x) = \int_0^x t^{s-1}e^{-t}\mathrm{d}t$ be the lower incomplete Gamma function. Note that,

$$\begin{aligned}
A\gamma(A+1,1) &= \frac{A}{A+1}u^{A+1}e^{-u}|_0^1 + \frac{A}{A+1}\gamma(A+2,1) \\
&= \frac{A}{A+1}e^{-1} + \frac{A}{A+1}\left[\frac{1}{A+2}e^{-1} + \frac{1}{A+2}\gamma(A+3,1)\right] \\
&= \lim_{K\to\infty}\left\{\sum_{k=1}^{K}\frac{\Gamma(A+1)^2}{\Gamma(A)\Gamma(A+k+1)}e^{-1} + A\Gamma(A+1)F(1;A+K,1)\right\} \\
&= \sum_{k=1}^{\infty}\frac{\Gamma(A+1)^2}{\Gamma(A)\Gamma(A+k+1)}e^{-1} \\
&= \sum_{k=1}^{\infty}\frac{A}{(A+1)(A+2)\dots(A+k)}e^{-1}
\end{aligned}$$

where $F(x;a,b)$ is the cdf of $\mathrm{Ga}(a,b)$ and $\lim_{a=\infty} F(1;a,1) = 0$. Furthermore we have

$$\sum_{k=1}^{\infty}\frac{\Gamma(A+1)^2}{\Gamma(A)\Gamma(A+k+1)} \quad = \textstyle\sum_{k=1}^{\infty}\frac{A}{(A+1)(A+2)\dots(A+k)} \quad \geqslant 1/2,$$

and

$$1 - \sum_{k=1}^{\infty}\frac{\Gamma(A+1)^2}{\Gamma(A)\Gamma(A+k+1)} \quad \leqslant 1 - \tfrac{A}{A+1} \quad \leqslant \frac{1}{A},$$

thus we have

$$\begin{aligned}
\frac{e^{-1}}{A\int_0^1 u_{s,h,k}^{A}e^{-u_h}\mathrm{d}u_{s,h,k}} - 1 + \frac{1}{A} &= \frac{1}{\sum_{k=1}^{\infty}\frac{\Gamma(A+1)^2}{\Gamma(A)\Gamma(A+k+1)}} - 1 + \frac{1}{A} \\
&= \frac{1 - \sum_{k=1}^{\infty}\frac{\Gamma(A+1)^2}{\Gamma(A)\Gamma(A+k+1)}}{\sum_{k=1}^{\infty}\frac{\Gamma(A+1)^2}{\Gamma(A)\Gamma(A+k+1)}} + \frac{1}{A} \\
&\leqslant \frac{1/A}{1/2} + \frac{1}{A} \\
&= \frac{3}{A}.
\end{aligned}$$

Hence $E[(\frac{1}{u}-1)|\tau] \leqslant 3/(\prod_{m=1}^{k}\tau_{s,h,m})$. Based on this inequality, we have

$$\sum_{k=d+1}^{D} E\Big\{E[(\frac{1}{u}-1)\sigma_s^2|\tau]\Big\} \leqslant \textstyle\sum_{k=d+1}^{D} E\bigg(\frac{3}{\prod_{m=1}^{k}\tau_{s,h,m}}\bigg)E(\sigma_s^2)$$

$$= \textstyle\sum_{k=d+1}^{D} 3ba^k \leqslant \frac{3ba^d}{1-a}$$

where $b = E(\sigma_s^2)$ and a $= E(\frac{1}{\tau_{s,h,1}})$. Note that $\tau_{s,h,m} \sim \mathrm{Exp}_{[1,\infty)}(\lambda)$, thus $a < 1$. By Fubini's theorem, $E\Big\{E(\sum_{k=H+1}^{\infty}\alpha_k^2|\tau)\Big\} = \sum_{k=d+1}^{\infty} E\Big\{E[(\frac{1}{u_{s,h,k}}-1)\sigma_s^2|\tau]\Big\}$. Now use inequality $(1-x/2) > \exp(-x)$ if $0 < x \leqslant 1.5$ to get

$$p\{d_\infty(\mathbf{\Omega}_{s,h}, \mathbf{\Omega}_{s,h}^d) \leqslant \epsilon\} \geqslant \exp\{\frac{-6ba^d}{\epsilon(1-a)}\}$$

if $d > 2\log\{b/\epsilon(1-a)\}/\log(1/a)$. Hence,

$$p\{d_\infty(\mathbf{\Omega}_{s,h}, \mathbf{\Omega}_{s,h}^d) > \epsilon\} \leqslant 1 - \exp\{\frac{-6ba^d}{\epsilon(1-a)}\} \leqslant \frac{6ba^d}{\epsilon(1-a)},$$

since $6ba^d/\{\epsilon(1-a)\} < 1$. □

Theorem A.2.4. *Let*

$$f^L(\boldsymbol{y}_i) = \sum_{s=1}^{L}\sum_{h=1}^{2^s} \tilde{\pi}_{s,h}\mathcal{N}_D(\boldsymbol{y}_i; \boldsymbol{\mu}_{s,h}, \mathbf{\Phi}_{s,h}\mathbf{\Sigma}_{s,h}\mathbf{\Phi}_{s,h}^T + \sigma_s^2\boldsymbol{I})$$

denote the approximation at scale L, let $P(B) = \int_B f(\boldsymbol{y}_i)dy$ *and* $P^L(B) = \int_B f^L(\boldsymbol{y}_i)dy$*, for all* $B \subset \Re^D$ *denote the probability measures corresponding to density* $f(\boldsymbol{y}_i)$ *and* $f^L(\boldsymbol{y}_i)$*. Then we have,*

$$d_{TV}(P_L, P) < \bigg(\frac{a_S}{1+a_S}\bigg)^L,$$

where $d_{TV}(P_L, P)$ *denotes the total variation distance between* $P_L(B)$ *and* $P(B)$*.*

Proof. The total variation distance is given by

$$
\begin{aligned}
d_{TV}(P_L, P) &= \sup_{B \in \Re^D} |P^L(B) - P(B)| \\
&= \sup_{B \in \Re^D} |\sum_{h=1}^{2^L} \tilde{\pi}_{s,h} N(B; \boldsymbol{\mu}_{s,h}, \boldsymbol{\Phi}_{s,h}\boldsymbol{\Sigma}_{s,h}\boldsymbol{\Phi}_{s,h}^T + \sigma_s^2 \boldsymbol{I}) - ... \\
&\quad \sum_{s=L}^{\infty} \sum_{h=1}^{2^s} \pi_{s,h} N(B; \boldsymbol{\mu}_{s,h}, \boldsymbol{\Phi}_{s,h}\boldsymbol{\Sigma}_{s,h}\boldsymbol{\Phi}_{s,h}^T + \sigma_s^2 \boldsymbol{I})| \\
&\leqslant \max\{\sum_{h=1}^{2^L} \tilde{\pi}_{s,h}, \sum_{s=L}^{\infty} \sum_{h=1}^{2^s} \pi_{s,h}\} \\
&= \max\left\{2^L (\frac{a_S}{1+a_S})^{L-1} \frac{1}{1+a_S} 2^{-L}, \sum_{s=L}^{\infty} 2^s \frac{1}{1+a_S} (\frac{a_S}{2+2a_S})^s\right\} \\
&= \sum_{s=L}^{\infty} \frac{1}{1+a_S} (\frac{a_S}{1+a_S})^s \\
&= (\frac{a_S}{1+a_S})^L
\end{aligned}
$$

□

A.3 Posterior Conditional Derivation

Based on the likelihood function (A1), the derivation of conditional posterior of σ_s^{-2} is given by

$$
\begin{aligned}
p(\sigma_s^{-2}|-) &\propto (\sigma_s^{-2})^{a_\sigma - 1} \exp(-b_\sigma \sigma_s^{-2}) \prod_{y_i \in C_s} (\sigma_s^2)^{-D/2} \\
&\quad \exp\left\{-\frac{1}{2}\sigma_s^{-2}(A_{s,h,i} - \sum_{j=1}^{d}(1-u_{s,h,j})(Z_{s,h,i}^{(j)})^2)\right\} \\
&\propto (\sigma_s^{-2})^{Dn_s/2 + a_\sigma - 1} \\
&\quad \exp\left\{-\sigma_s^{-2}[\frac{1}{2}\sum_{y_i \in C_s}(A_{s,h,i} - \sum_{j=1}^{d}(1-u_{s,h,j})(Z_{s,h,i}^{(j)})^2) + b_\sigma]\right\}.
\end{aligned}
$$

The derivation of conditional posterior of $u_{s,h,m}$ is given by

$$\begin{aligned} p(u_{s,h,m}|-) &\propto \prod_{y_i \in C_{s,h}} u_{s,h,m}^{1/2} \exp\Big\{ -\frac{1}{2}\sigma_s^{-2} u_{s,h,m} (Z_{s,h,i}^{(m)})^2 \Big\} \\ & \quad u_{s,h,m}^{\prod_{j=1}^{m} \tau_{s,h,j} - 1} \exp\{-u_{s,h,m}\} I_{(0,1)} \\ &\propto u_{m,s,h}^{\prod_{j=1}^{m} \tau_{s,h,j} + n_{s,h}/2 - 1} \\ & \quad \exp\Big\{ -[1 + \frac{1}{2}\sigma_s^{-2} \sum_{y_i \in C_{s,h}} (Z_{s,h,i}^{(m)})^2] u_{s,h,m} \Big\} I_{(0,1)}. \end{aligned}$$

The derivation of conditional posterior of $\tau_{s,h,m}$ is given by

$$\begin{aligned} p(\tau_{s,h,m}|-) &\propto (\prod_{j>m-1} u_{j,s,h})^{\tau_{s,h,j}} \exp\{-a_\tau \tau_{s,h,m}\} I_{[1,\infty)} \\ &\propto \exp\Big\{ -[a_\tau - ln(\prod_{j>m-1} u_{s,h,j})] \tau_{s,h,m} \Big\} \end{aligned}$$

A.4 Missing Data Imputation

Proposition A.4.1. *For node (s,h), introduce augmented data $\boldsymbol{\eta}_i$ such that $(\boldsymbol{y}_i|\boldsymbol{\eta}_i, \boldsymbol{\Theta}, s_i = s, h_i = h) \sim \mathcal{N}_D(\boldsymbol{\mu}_{s,h} + \boldsymbol{\Phi}_{s,h}\boldsymbol{\eta}_i, \sigma_s^2 \boldsymbol{I}_D)$ and $(\boldsymbol{\eta}_i|\boldsymbol{\Theta}, s_i = s, h_i = h) \sim \mathcal{N}_d(0, \boldsymbol{\Sigma}_{s,h})$, for $i = 1, \ldots, n$. Then we have the conditional distribution with $\boldsymbol{\eta}_i$ marginalized out equal $(\boldsymbol{y}_i|\boldsymbol{\Theta}, s_i = s, h_i = h) \sim \mathcal{N}_D(\boldsymbol{\mu}_{s,h} + \boldsymbol{\Phi}_{s,h}\boldsymbol{\Sigma}_{s,h}\boldsymbol{\Phi}_{s,h}^T, \sigma_s^2 \boldsymbol{I}_D)$. Furthermore, conditional on $s_i = s$ and $h_i = h$ we have*

$$\boldsymbol{\eta}_i|\boldsymbol{y}_O, \boldsymbol{\Theta} \sim \mathcal{N}_d(\hat{\boldsymbol{\mu}}_\eta, \hat{\boldsymbol{\Sigma}}_\eta), \qquad \boldsymbol{y}_M|\boldsymbol{\eta}_i, \boldsymbol{y}_O, \boldsymbol{\Theta} \sim \mathcal{N}_{m_i}(\boldsymbol{\mu}_M + \boldsymbol{\Phi}_M \boldsymbol{\eta}_i, \sigma_s^2 I_{m_i}),$$

where $\hat{\boldsymbol{\Sigma}}_\eta = \big(\boldsymbol{\Sigma}_{s,h}\boldsymbol{\Phi}_O^T\boldsymbol{\Phi}_O/\sigma_s^2 + I\big)^{-1}\boldsymbol{\Sigma}_{s,h}$ and $\hat{\boldsymbol{\mu}}_\eta = \hat{\boldsymbol{\Sigma}}_\eta \boldsymbol{\Phi}_O^T(\boldsymbol{y}_O - \boldsymbol{\mu}_O)/\sigma_s^2$,

Proof. The proposition can be easily proved using Bayes rule. The joint density of

$(\boldsymbol{y}_O, \boldsymbol{y}_M, \boldsymbol{\eta}_i | \boldsymbol{\Theta})$ is given by

$$\begin{aligned} p(\boldsymbol{y}_O, \boldsymbol{y}_M, \boldsymbol{\eta}_i | \boldsymbol{\Theta}, s_i = s, h_i = h) &\propto \exp\left\{ -\frac{\|\boldsymbol{y}_i - \boldsymbol{\Phi}\boldsymbol{\eta}_i - \boldsymbol{\mu}\|_2}{2\sigma_s^2} - \frac{\boldsymbol{\eta}_i^T \boldsymbol{\Sigma}_{s,h}^{-1} \boldsymbol{\eta}_i}{2} \right\} \\ &\propto \exp\left\{ -\frac{\|\boldsymbol{y}_M - \boldsymbol{\Phi}_M\boldsymbol{\eta}_i - \boldsymbol{\mu}_M\|_2}{2\sigma_s^2} \right. \\ &\quad \left. -\frac{\|\boldsymbol{y}_O - \boldsymbol{\Phi}_O\boldsymbol{\eta}_i - \boldsymbol{\mu}_O\|_2}{2\sigma_s^2} - \frac{\boldsymbol{\eta}_i^T \boldsymbol{\Sigma}_{s,h}^{-1} \boldsymbol{\eta}_i}{2} \right\}. \end{aligned}$$

Hence the conditional density $(\boldsymbol{y}_M | \boldsymbol{\eta}_i, \boldsymbol{y}_O, \boldsymbol{\Theta}, s_i = s, h_i = h)$ is given by

$$p(\boldsymbol{y}_M | \boldsymbol{\eta}_i, \boldsymbol{y}_O, \boldsymbol{\Theta}, s_i = s, h_i = h) \propto \exp\left\{ -\frac{\|\boldsymbol{y}_M - \boldsymbol{\Phi}_M\boldsymbol{\eta}_i - \boldsymbol{\mu}_M\|_2}{2\sigma_s^2} \right\}.$$

The marginal conditional density $(\boldsymbol{\eta}_i | \boldsymbol{y}_O, \boldsymbol{\Theta}, s_i = s, h_i = h)$ is given by

$$\begin{aligned} \mathrm{p}(\boldsymbol{\eta}_i | \boldsymbol{y}_i^O, \boldsymbol{\Theta}, s_i = s, h_i = h) &\propto \int \mathrm{p}(\boldsymbol{y}_M, \boldsymbol{\eta}_i | \boldsymbol{y}_O) \mathrm{d}\boldsymbol{y}_M \\ &\propto \exp\left\{ \frac{\|\boldsymbol{y}_O - \boldsymbol{\Phi}_O\boldsymbol{\eta}_i - \boldsymbol{\mu}_O\|_2}{2\sigma_s^2} - \frac{\boldsymbol{\eta}_i^T \boldsymbol{\Sigma}_{s,h}^{-1} \boldsymbol{\eta}_i}{2} \right\}. \end{aligned}$$

□

To finish the missing data imputation algorithm, the conditional posterior distribution of the membership variable (s_i, h_i) of partially observed subject i, $p(s_i, h_i | \boldsymbol{y}_O, \boldsymbol{\Theta})$ is needed. $\boldsymbol{y}_M$ has been marginalized out to reduce the sample autocorrelation, and the distribution is given by

$$\begin{aligned} \mathrm{p}(s_i, h_i | \boldsymbol{y}_O, \boldsymbol{\Theta}) &\propto \int p(\boldsymbol{y}_M, \boldsymbol{y}_O, \boldsymbol{\Theta}, s_i, h_i) \mathrm{d}\boldsymbol{y}_M \\ &\propto \int \pi_{s_i,h_i} \mathcal{N}_D(\boldsymbol{y}_i; \boldsymbol{\mu}_{s_i,h_i}, \boldsymbol{\Phi}_{s_i,h_i} \boldsymbol{\Sigma}_{s_i,h_i} \boldsymbol{\Phi}_{s_i,h_i}^T + \sigma_{s_i}^2 \mathbf{I}) \mathrm{d}\boldsymbol{y}_M. \end{aligned}$$

With a slight abuse of notation, we write $\boldsymbol{\Phi}$ as $\boldsymbol{\Phi}_{s_i,h_i}$, $\boldsymbol{\Sigma}$ as $\boldsymbol{\Sigma}_{s_i,h_i}$, σ^2 denote $\sigma_{s_i}^2$ and

$\boldsymbol{\mu}$ as $\boldsymbol{\mu}_{s_i,h_i}$. By properties of multicariate Gaussian, we have

$$\int \mathcal{N}_D(\boldsymbol{y}_i; \boldsymbol{\mu}, \boldsymbol{\Phi}\boldsymbol{\Sigma}\boldsymbol{\Phi}^T + \sigma^2\mathrm{I})\mathrm{d}\boldsymbol{y}_M = \mathcal{N}_{D-m_i}(\boldsymbol{y}_O; \boldsymbol{\mu}_O, \boldsymbol{\Phi}_O\boldsymbol{\Sigma}\boldsymbol{\Phi}_O^T + \sigma^2\boldsymbol{I}).$$

Hence we have $p(s_i, h_i|\boldsymbol{y}_O, \boldsymbol{\Theta}) \propto \mathcal{N}_{D-m_i}(\boldsymbol{y}_O; \boldsymbol{\mu}_O, \boldsymbol{\Phi}_O\boldsymbol{\Sigma}\boldsymbol{\Phi}_O^T + \sigma^2\boldsymbol{I})$. Directly computing this value includes inverting a $(D - m_i) \times (D - m_i)$ matrix, which is computational intractable when $D - m_i$ is large. With basic linear algebra, we have

$$\left|\boldsymbol{\Phi}_O\boldsymbol{\Sigma}\boldsymbol{\Phi}_O^T + \sigma^2\boldsymbol{I}\right| = (\sigma^2)^{D-m_i}\left|\boldsymbol{I} + \boldsymbol{\Sigma}\boldsymbol{\Phi}_O^T\boldsymbol{\Phi}_O/\sigma^2\right|,$$

$$\left(\boldsymbol{\Phi}_O\boldsymbol{\Sigma}\boldsymbol{\Phi}_O^T + \sigma^2\boldsymbol{I}\right)^{-1} = \frac{\boldsymbol{I}}{\sigma^2} - \frac{\boldsymbol{\Phi}_O\left(\boldsymbol{I} + \boldsymbol{\Sigma}\boldsymbol{\Phi}^T\boldsymbol{\Phi}_O/\sigma^2\right)^{-1}\boldsymbol{\Sigma}\boldsymbol{\Phi}_O^T}{\sigma^4}.$$

Hence we have

$$\begin{aligned}&\mathcal{N}_{D-m_i}(\boldsymbol{y}_O; \boldsymbol{\mu}_O, \boldsymbol{\Phi}_O\boldsymbol{\Sigma}\boldsymbol{\Phi}_O^T + \sigma^2\boldsymbol{I})\\ =&(2\pi\sigma^2)^{-(D-m_i)/2}\left|\boldsymbol{I} + \boldsymbol{\Sigma}\boldsymbol{A}/\sigma^2\right|^{-1/2}\\ &\times \exp\left\{-\frac{B_i}{2\sigma^2} + \frac{\boldsymbol{C}_i^T\left(\boldsymbol{\Sigma}^{-1} + \boldsymbol{A}/\sigma^2\right)^{-1}\boldsymbol{C}_i}{2\sigma^4}\right\}\end{aligned} \tag{A2}$$

where $\boldsymbol{A} = \boldsymbol{\Phi}_O^T\boldsymbol{\Phi}_O$, $B_i = \|\boldsymbol{y}_O - \boldsymbol{\mu}_O\|_2$ and $\boldsymbol{C}_i = \boldsymbol{\Phi}_O^T(\boldsymbol{y}_O - \boldsymbol{\mu}_O)$. Note that $\boldsymbol{A}$, B_i and $\boldsymbol{C}_i$ can be computed before the MCMC algorithm with a computational cost being $O\big((D - m_i)d\big)$. Within the MCMC, the cost to compute (A2) is only $O(d^3)$.

Nonlinear GEODE Conditional on the membership $\big(s_i, h_i\big)$, the imputational strategies of nonlinear GEODE are exactly the same as those of the linear GEODE. Hence we only discuss the conditional posterior distribution of the membership variable given a partially observed $\boldsymbol{y}_O$, which is given as follow

$$\begin{aligned}&p(s_i = s, h_i = h|\boldsymbol{y}_O, \boldsymbol{\Theta}, \{\boldsymbol{\mu}_{sh}, \boldsymbol{W}_{sh}\}_L)\\ \propto&\pi_{s,h}\phi(\boldsymbol{y}_O; \boldsymbol{\mu}_{sh}, \boldsymbol{W}_{sh}\boldsymbol{\Sigma}_{sh}\boldsymbol{W}_{sh}^\top + \sigma_s^2\boldsymbol{I})\end{aligned}$$

where $\phi(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the density function of a multivariate Gaussian with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$.

A.5 Simulation Studies

In the missing data imputation simulatoin study, we simulated 100 independent samples of size $n = 600$ from different scenarios as follows.

> **Scenario 1-6:** Data $\boldsymbol{y}_i$, for $i = 1, \ldots, 600$, were generated from $\mathcal{N}_D(0, \boldsymbol{\Lambda}\boldsymbol{\Lambda}^T + \sigma^2\boldsymbol{I})$. $\boldsymbol{\Lambda}$ is a $D \times p$ matrix with each entry generated from $\mathcal{N}(0, 25)$ and $10\sigma^2$ was generated from $\chi_{(1)}$. This scenario includes different cases where $p \in \{10, 50\}$, $D \in \{5000, 10000, 15000\}$ and with or without a 20% missing data. We fixed the upper bound to $d = 100$.
>
> **Scenario 7-9:** 3–D data $\boldsymbol{\eta}_i$, for $i = 1, \ldots, 600$, were generated on the Swissroll with Gaussian noise distributed as $\mathcal{N}(0, 2.5 \times 10^{-5})$ along each dimension. Data $\boldsymbol{y}_i$, for $i = 1, \ldots, 600$, were obtained by $\boldsymbol{y}_i = \boldsymbol{\Lambda}\boldsymbol{\eta}_i$ where $\boldsymbol{\Lambda}$ were generated in the same way as in Scenario 1. This scenario includes different cases where $D \in \{5000, 10000, 15000\}$ and with or without a 20% missing data. We fixed the upper bound to $d = 10$.

The average inclusion probabilities of each presetted dimensions were computed in the following way. Let $\mathcal{R}^t_{s,h}$ denotes the set of retained column indices of node (s, h) at the tth iteration, and let (s^t_i, h^t_i) denote the node index of the ith observation at the tth iteration. Then the inclusion probability of dimension $j = 1, 2, \ldots, 10$ in scenario 2 is given by

$$p_j^{inclu} = \frac{1}{n_{adapt} \times N} \sum_{t:\ adapt} \sum_{i=1}^{N} I_{(j \in \mathcal{R}^t_{s^t_i, h^t_i})}$$

where n_{adapt} denotes the number of adaptation steps during the MCMC collection interval.

Appendix B

Appendix to Chapter 4

Whenever we write $\lim_{n\to\infty} a_n = a_0$, we mean the limit holds with $P_{\boldsymbol{\theta}_0}$-probability one. We will omit the phrase "with $P_{\boldsymbol{\theta}_0}$-probability one" for succinctness.

B.0.1 Proof of Lemma 4.2.1

Since $P_{\boldsymbol{\psi}}$ is *mosaic-type*, from Definition 4.1.1, we have $\boldsymbol{\psi}_{st}$, $1 \leqslant t \leqslant s \leqslant p$ such that

$$\boldsymbol{\psi} = \left[\boldsymbol{\psi}_{12}^\top, \ldots, \boldsymbol{\psi}_{(p-1)p}^\top, \boldsymbol{\psi}_{11}^\top, \ldots, \boldsymbol{\psi}_{pp}^\top\right]^\top.$$

For $j = 1, \ldots, p$, the density of the univariate marginal distribution of $P_{\boldsymbol{\psi}}$ is $f_{0,jj}(x_j|\boldsymbol{\psi}_{jj})$. And for $1 \leqslant t < s \leqslant p$, the density of the bivariate marginal data distribution of $P_{\boldsymbol{\psi}}$ is $f_{0,st}(x_s, x_t|\boldsymbol{\psi}_{st}, \boldsymbol{\psi}_{ss}, \boldsymbol{\psi}_{tt})$. Introduce

$$\boldsymbol{\theta} = \left[\boldsymbol{\psi}_{12}^\top, \ldots, \boldsymbol{\psi}_{(p-1)p}^\top, \boldsymbol{\psi}_{11}^\top, \boldsymbol{\mu}_1^\top, \ldots, \boldsymbol{\psi}_{pp}^\top, \boldsymbol{\mu}_p^\top\right]^\top,$$

and let $\boldsymbol{\theta}_{st} = \boldsymbol{\psi}_{st}$ for $1 \leqslant s < t \leqslant p$ and $\boldsymbol{\theta}_{jj} = \left(\boldsymbol{\psi}_{jj}^\top, \boldsymbol{\mu}_j^\top\right)^\top$ for $j = 1, \ldots, p$. From (4.2.1), it can be shown that

$$
\begin{aligned}
&\int \cdots \int f(\boldsymbol{y}|\boldsymbol{\mu}, \boldsymbol{\psi}) \mathrm{d}y_2 \cdots \mathrm{d}y_p \\
=&\int f_0(\boldsymbol{x}|\boldsymbol{\psi}) g_1(y_1|x_1, \boldsymbol{\mu}_1) \prod_{j=2}^{p} \left[\int g_j(y_j|x_j, \boldsymbol{\mu}_j) \mathrm{d}y_j \right] \mathrm{d}\boldsymbol{x} \\
=&\int f_{0,11}(x_1|\boldsymbol{\psi}_{11}) g_1(y_1|x_1, \boldsymbol{\mu}_1) \mathrm{d}x_1 \\
=&f_{11}(y_1|\boldsymbol{\theta}_{11}).
\end{aligned}
$$

Similarly, one can show that there exists a collection of density functions $\{f_{jj}(y_j|\boldsymbol{\theta}_{jj})\}$ such that for $j = 2, \ldots, p$, the density of the univariate marginal distribution of $P_{\boldsymbol{\mu},\boldsymbol{\psi}}$ is $f_{jj}(y_j|\boldsymbol{\theta}_{jj})$. And that there exists a collection of density functions $\{f_{st}(y_s, y_t|\boldsymbol{\theta}_{st}, \boldsymbol{\theta}_{ss}, \boldsymbol{\theta}_{tt})\}$ such that for $1 \leqslant t < s \leqslant p$, the density of the bivariate marginal data distribution of $P_{\boldsymbol{\mu},\boldsymbol{\psi}}$ is $f_{st}(y_s, y_t|\boldsymbol{\theta}_{st}, \boldsymbol{\theta}_{ss}, \boldsymbol{\theta}_{tt})$. Hence $P_{\boldsymbol{\mu},\boldsymbol{\psi}}$ is also *Mosaic-type*.

B.0.2 Taylor Expansions & Upper Bounds

We will find the limit and derive an upper bound for the Taylor expansion of $L_n(\boldsymbol{\eta}, \boldsymbol{\zeta})$, which will be used in later proofs.

Lemma B.0.1. *Letting $\boldsymbol{\alpha}$ be a d-dimensional* multi-index *for $\boldsymbol{x}$, consider $M_{\boldsymbol{\alpha}} < \infty$ for all $\boldsymbol{\alpha}$ such that $|\boldsymbol{\alpha}| = 3$. Then for any positive definite matrix $\boldsymbol{\Lambda}$, we can find $\delta > 0$ such that when $|\boldsymbol{x}| < \sqrt{n}\delta$,*

$$
\left| \sum_{|\boldsymbol{\alpha}|=3} M_{\boldsymbol{\alpha}} \frac{\boldsymbol{x}^{\boldsymbol{\alpha}}}{\sqrt{n}} \right| < \frac{1}{2} \boldsymbol{x}^\top \boldsymbol{\Lambda} \boldsymbol{x}.
$$

Proof. It is easily seen that

$$
\left| \sum_{|\boldsymbol{\alpha}|=3} M_{\boldsymbol{\alpha}} \frac{\boldsymbol{x}^{\boldsymbol{\alpha}}}{\sqrt{n}} \right| < \sum_{|\boldsymbol{\alpha}|=3} M_{\boldsymbol{\alpha}} \left| \frac{\boldsymbol{x}^{\boldsymbol{\alpha}}}{\sqrt{n}} \right|.
$$

Consider $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_d)$ where $|\boldsymbol{\alpha}| = 3$, and suppose α_{j_1} is the first non-zero index. Since $|\boldsymbol{x}| < \sqrt{n}\delta$, we have $\frac{|x_{j_1}|}{\sqrt{n}} < \delta$. Assume that α_{j_2} and α_{j_3} are the other two non-zero indices, note that we allow $j_2 = j_3$. We have

$$\left|\frac{\boldsymbol{x}^{\boldsymbol{\alpha}}}{\sqrt{n}}\right| < \delta |x_{j_2} x_{j_3}| \leqslant \frac{\delta}{2}\big(x_{j_2}^2 + x_{j_3}^2\big).$$

Doing this for all $\boldsymbol{\alpha}$ such that $|\boldsymbol{\alpha}| = 3$, it can be shown that

$$\sum_{|\boldsymbol{\alpha}|=3} M_{\boldsymbol{\alpha}} \left|\frac{\boldsymbol{x}^{\boldsymbol{\alpha}}}{\sqrt{n}}\right| < \frac{\delta}{2}\boldsymbol{x}^\top \boldsymbol{\Psi} \boldsymbol{x},$$

where $\boldsymbol{\Psi}$ is some diagonal matrix with all diagonal elements being positive. Since $\boldsymbol{\Lambda}$ is positive definite, we can always find $\delta > 0$ so that $\boldsymbol{\Lambda} - \delta\boldsymbol{\Psi}$ is also positive definite, which implies that for any $\boldsymbol{x} \in \mathbb{R}^d$,

$$\frac{1}{2}\boldsymbol{x}^\top \boldsymbol{\Lambda} \boldsymbol{x} - \frac{\delta}{2}\boldsymbol{x}^\top \boldsymbol{\Psi} \boldsymbol{x} = \frac{1}{2}\boldsymbol{x}^\top \big(\boldsymbol{\Lambda} - \delta\boldsymbol{\Psi}\big)\boldsymbol{x} > 0.$$

□

Recall that $\boldsymbol{t} = \sqrt{n}\big(\boldsymbol{\eta} - \hat{\boldsymbol{\eta}}_n\big)$ and $\boldsymbol{r} = \sqrt{n}\big(\boldsymbol{\zeta} - \hat{\boldsymbol{\zeta}}_n\big)$. Letting $\boldsymbol{\alpha}$ be a d_η-dimensional *multi-index* for $\boldsymbol{t}$ and $\boldsymbol{\beta}$ be a d_ζ-dimensional *multi-index* for $\boldsymbol{r}$, we expand the Taylor series for $L_n(\hat{\boldsymbol{\eta}}_n + \boldsymbol{t}/\sqrt{n}, \hat{\boldsymbol{\zeta}}_n + \boldsymbol{r}/\sqrt{n}) - L_n(\hat{\boldsymbol{\eta}}_n, \hat{\boldsymbol{\zeta}}_n)$ and get

$$\begin{aligned} & L_n(\hat{\boldsymbol{\eta}}_n + \boldsymbol{t}/\sqrt{n}, \hat{\boldsymbol{\zeta}}_n + \boldsymbol{r}/\sqrt{n}) - L_n(\hat{\boldsymbol{\eta}}_n, \hat{\boldsymbol{\zeta}}_n) \\ =& \frac{1}{2}\big[\begin{smallmatrix}\boldsymbol{t}\\ \boldsymbol{r}\end{smallmatrix}\big]^\top \frac{1}{n} L_n^{(2)}(\hat{\boldsymbol{\eta}}_n, \hat{\boldsymbol{\zeta}}_n)\big[\begin{smallmatrix}\boldsymbol{t}\\ \boldsymbol{r}\end{smallmatrix}\big] + R_n\left(\boldsymbol{t}, \boldsymbol{r}\right), \end{aligned} \tag{B.1}$$

where $\boldsymbol{\eta}_n^{'}$ is between $\hat{\boldsymbol{\eta}}_n + \boldsymbol{t}/\sqrt{n}$ and $\hat{\boldsymbol{\eta}}_n$, $\boldsymbol{\zeta}_n^{'}$ is between $\hat{\boldsymbol{\zeta}}_n + \boldsymbol{r}/\sqrt{n}$ and $\hat{\boldsymbol{\zeta}}_n$ and

$$R_n\left(\boldsymbol{t}, \boldsymbol{r}\right) = \sum_{|\boldsymbol{\alpha}|+|\boldsymbol{\beta}|=3} \frac{1}{n\boldsymbol{\alpha}!}\partial^{\boldsymbol{\alpha}}\partial^{\boldsymbol{\beta}} L_n(\boldsymbol{\eta}_n^{'}, \boldsymbol{\zeta}_n^{'})\frac{\boldsymbol{t}^{\boldsymbol{\alpha}}\boldsymbol{r}^{\boldsymbol{\beta}}}{\sqrt{n}}.$$

Letting $\hat{\boldsymbol{I}}_n = -\frac{1}{n}L_n^{(2)}(\hat{\boldsymbol{\eta}}_n, \hat{\boldsymbol{\lambda}}_n)$ and $\hat{\boldsymbol{I}}_n = \big[\begin{smallmatrix}\hat{\boldsymbol{I}}_n^{11} & \hat{\boldsymbol{I}}_n^{12}\\ \hat{\boldsymbol{I}}_n^{21} & \hat{\boldsymbol{I}}_n^{22}\end{smallmatrix}\big]$, (B.1) can be written as

$$\begin{aligned} & L_n(\hat{\boldsymbol{\eta}}_n + \boldsymbol{t}/\sqrt{n}, \hat{\boldsymbol{\zeta}}_n + \boldsymbol{r}/\sqrt{n}) - L_n(\hat{\boldsymbol{\eta}}_n, \hat{\boldsymbol{\zeta}}_n) \\ =& -\frac{1}{2}\boldsymbol{t}^\top \hat{\boldsymbol{I}}_n^{11}\boldsymbol{t} - \frac{1}{2}\boldsymbol{r}^\top \hat{\boldsymbol{I}}_n^{22}\boldsymbol{r} - \boldsymbol{t}^\top \hat{\boldsymbol{I}}_n^{12}\boldsymbol{r} + R_n(\boldsymbol{t}, \boldsymbol{r}). \end{aligned} \tag{B.2}$$

Corollary B.0.2. *If conditions 1-5 hold for* $f_2(\boldsymbol{y}|\boldsymbol{\eta},\boldsymbol{\zeta})$*, then* $\exists\delta > 0$ *such that for all* $\boldsymbol{t}$ *and* $\boldsymbol{r}$ *satisfying* $\left\|\left[\begin{smallmatrix}\boldsymbol{t}/\sqrt{n}\\ \boldsymbol{r}/\sqrt{n}\end{smallmatrix}\right]\right\| < \delta$*, the followings are true:*

i) For any fixed $\boldsymbol{t}$ *and* $\boldsymbol{r}$*,* $\lim_{n\to\infty} R_n(\boldsymbol{t},\boldsymbol{r}) = 0$*.*

ii) For any positive definite matrix $\boldsymbol{\Lambda}$*,* $\exists N \in \mathbb{N}_1$ *such that* $\forall n > N$*,*

$$\left|R_n(\boldsymbol{t},\boldsymbol{r})\right| < \left[\begin{smallmatrix}\boldsymbol{t}\\ \boldsymbol{r}\end{smallmatrix}\right]^\top \boldsymbol{\Lambda}\left[\begin{smallmatrix}\boldsymbol{t}\\ \boldsymbol{r}\end{smallmatrix}\right]. \tag{B.3}$$

Proof. From condition 2, for some $\delta^{'} > 0$ we have

$$\sup_{\boldsymbol{z}\in\mathcal{B}(\boldsymbol{z}_0,\delta^{'})} |\partial^{\boldsymbol{\alpha}}\ell_2(\boldsymbol{\eta},\boldsymbol{\zeta},\boldsymbol{y})| \leqslant M_{\boldsymbol{\alpha},2}(\boldsymbol{y}),$$

and $\mathbb{E}_{\boldsymbol{\theta}_0}M_{\boldsymbol{\alpha},2}(\boldsymbol{y}) < \infty$. Letting $\delta < \delta^{'}$, we have

$$\left|\sum_{|\boldsymbol{\alpha}|+|\boldsymbol{\beta}|=3}\frac{1}{n\boldsymbol{\alpha}!}\partial^{\boldsymbol{\alpha}}\partial^{\boldsymbol{\beta}}L_n(\boldsymbol{\eta}_n^{'},\boldsymbol{\zeta}_n^{'})\frac{\boldsymbol{t}^{\boldsymbol{\alpha}}\boldsymbol{r}^{\boldsymbol{\beta}}}{\sqrt{n}}\right| < \sum_{|\boldsymbol{\alpha}|+|\boldsymbol{\beta}|=3}\frac{1}{n\boldsymbol{\alpha}!}\sum_{i=1}^{n}M_{\boldsymbol{\alpha},2}(\boldsymbol{y}_i)\frac{\boldsymbol{t}^{\boldsymbol{\alpha}}\boldsymbol{r}^{\boldsymbol{\beta}}}{\sqrt{n}}.$$

From strong law of large numbers (SLLN) we know that $\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}M_{\boldsymbol{\alpha},2}(\boldsymbol{y}_i) = \mathbb{E}_{\boldsymbol{\theta}_0}M_{\boldsymbol{\alpha},2}(\boldsymbol{y})$, hence $\exists N_1 \in \mathbb{N}_1$ such that $\forall n > N_1$,

$$\left|\sum_{|\boldsymbol{\alpha}|+|\boldsymbol{\beta}|=3}\frac{1}{n\boldsymbol{\alpha}!}\partial^{\boldsymbol{\alpha}}\partial^{\boldsymbol{\beta}}L_n(\boldsymbol{\eta}_n^{'},\boldsymbol{\zeta}_n^{'})\frac{\boldsymbol{t}^{\boldsymbol{\alpha}}\boldsymbol{r}^{\boldsymbol{\beta}}}{\sqrt{n}}\right| < 2\sum_{|\boldsymbol{\alpha}|+|\boldsymbol{\beta}|=3}\frac{1}{\boldsymbol{\alpha}!}\mathbb{E}_{\boldsymbol{\theta}_0}M_{\boldsymbol{\alpha},2}(\boldsymbol{y})\frac{\boldsymbol{t}^{\boldsymbol{\alpha}}\boldsymbol{r}^{\boldsymbol{\beta}}}{\sqrt{n}}.$$

Apparently, for fixed $\boldsymbol{t}$ and $\boldsymbol{r}$,

$$\lim_{n\to\infty}|R_n(\boldsymbol{t},\boldsymbol{r})| < \lim_{n\to\infty} 2\sum_{|\boldsymbol{\alpha}|+|\boldsymbol{\beta}|=3}\frac{1}{\boldsymbol{\alpha}!}\mathbb{E}_{\boldsymbol{\theta}_0}M_{\boldsymbol{\alpha},2}(\boldsymbol{y})\frac{\boldsymbol{t}^{\boldsymbol{\alpha}}\boldsymbol{r}^{\boldsymbol{\beta}}}{\sqrt{n}} = 0.$$

Therefore $\lim_{n\to\infty}|R_n(\boldsymbol{t},\boldsymbol{r})| = 0$.

Applying Lemma B.0.1, we could find $\delta < \delta^{'}$ and $N_2 \in \mathbb{N}_1$ such that $\forall n > \max\{N_1, N_2\}$,

$$\left|2\sum_{|\boldsymbol{\alpha}|+|\boldsymbol{\beta}|=3}\frac{1}{\boldsymbol{\alpha}!}\mathbb{E}_{\boldsymbol{\theta}_0}M_{\boldsymbol{\alpha},2}(\boldsymbol{y})\frac{\boldsymbol{t}^{\boldsymbol{\alpha}}\boldsymbol{r}^{\boldsymbol{\beta}}}{\sqrt{n}}\right| < \left[\begin{smallmatrix}\boldsymbol{t}\\ \boldsymbol{r}\end{smallmatrix}\right]^\top \boldsymbol{\Lambda}\left[\begin{smallmatrix}\boldsymbol{t}\\ \boldsymbol{r}\end{smallmatrix}\right].$$

Letting $N = \max\{N_1, N_2\}$ finishes the proof. □

B.0.3 Proof of Lemma B.0.3

For $\delta > 0$, we define $A_{n,1}(\delta) = \{\boldsymbol{r} : \|\boldsymbol{r}\| < \sqrt{n}\delta\}$, $A_{n,2}(\delta) = \{\boldsymbol{r} : \|\boldsymbol{r}\| > \sqrt{n}\delta\}$, $B_{n,1}(\delta) = \{\boldsymbol{t} : \|\boldsymbol{t}\| < \sqrt{n}\delta\}$ and $B_{n,2}(\delta) = \{\boldsymbol{t} : \|\boldsymbol{t}\| > \sqrt{n}\delta\}$. We first provide a lemma that will be used in our later proof of the main result.

Lemma B.0.3. *Suppose that conditions 1-5 hold for* $f_2(\boldsymbol{y}|\boldsymbol{\eta}, \boldsymbol{\zeta})$ *and the prior density* $\pi(\boldsymbol{\eta}|\boldsymbol{\zeta})$ *is continuous and positive at* $\left[\begin{smallmatrix}\boldsymbol{\eta}_0 \\ \boldsymbol{\zeta}_0\end{smallmatrix}\right]$, *then* $\exists \delta_t > 0$, $\delta_r > 0$ *and* $N \in \mathbb{N}_1$ *such that the followings are true:*

- *For any fixed* $\boldsymbol{t}$ *and* $\boldsymbol{r}$, *with* $P_{\boldsymbol{\theta}_0}$*-probability one*

$$\begin{aligned}&\lim_{n=\infty} \pi^*_{n,2}(\boldsymbol{t}|\boldsymbol{r})\mathbb{1}\{\boldsymbol{r} \in A_{n,1}(\delta_r), \boldsymbol{t} \in B_{n,1}(\delta_t)\} \\ =&\phi\left(\boldsymbol{t}\middle|-\left(\boldsymbol{I}_0^{11}\right)^{-1}\boldsymbol{I}_0^{12}\boldsymbol{r}, \left(\boldsymbol{I}_0^{11}\right)^{-1}\right).\end{aligned} \tag{B.4}$$

- $\exists \epsilon > 0$ *and* $c(\delta_r, \delta_t) > 0$ *such that*

$$\pi^*_{n,2}(\boldsymbol{t}|\boldsymbol{r})\mathbb{1}\{\boldsymbol{r} \in A_{n,1}(\delta_r), \boldsymbol{t} \in B_{n,2}(\delta_t)\} < \frac{\left|4\pi\boldsymbol{I}_0^{11}\right|^{-1/2}}{c(\delta_r, \delta_t)} \exp\{-n\epsilon\}. \tag{B.5}$$

Proof. The proof consists of the following four steps.

Step 1 In this step we will find the limit of the normalizing constant. The constant is

$$a_n(\boldsymbol{r}) = \int g_n(\boldsymbol{t}, \boldsymbol{r})\mathrm{d}\boldsymbol{t}, \tag{B.6}$$

where

$$\begin{aligned}g_n(\boldsymbol{t}, \boldsymbol{r}) =& \exp\left\{L_n\left(\hat{\boldsymbol{\eta}}_n + \boldsymbol{t}/\sqrt{n}, \hat{\boldsymbol{\zeta}}_n + \boldsymbol{r}/\sqrt{n}\right) - L_n\left(\hat{\boldsymbol{\eta}}_n, \hat{\boldsymbol{\zeta}}_n\right)\right\} \\ &\times \pi\left(\hat{\boldsymbol{\eta}}_n + \boldsymbol{t}/\sqrt{n}\middle|\hat{\boldsymbol{\zeta}}_n + \boldsymbol{r}/\sqrt{n}\right)\mathbb{1}\{\boldsymbol{r} \in A_{n,1}(\delta_r)\}.\end{aligned}$$

We can find the limit of $a_n(\boldsymbol{r})$ by finding the limits of $\int g_n(\boldsymbol{t}, \boldsymbol{r})\mathbb{1}\{\boldsymbol{t} \in B_{n,1}(\delta_t)\}\mathrm{d}\boldsymbol{t}$ and of $\int g_n(\boldsymbol{t}, \boldsymbol{r})\mathbb{1}\{\boldsymbol{t} \in B_{n,2}(\delta_t)\}\mathrm{d}\boldsymbol{t}$, since $a_n(\boldsymbol{r})$ is the sum of the these two integrals. We start with the first one.

The Taylor expansion of $L_n\big(\hat{\boldsymbol{\eta}}_n + \boldsymbol{t}/\sqrt{n}, \hat{\boldsymbol{\zeta}}_n + \boldsymbol{r}/\sqrt{n}\big) - L_n\big(\hat{\boldsymbol{\eta}}_n, \hat{\boldsymbol{\zeta}}_n\big)$ is given in (B.2). Applying Corollary B.0.2, and since $\big\|\big[\begin{smallmatrix}\boldsymbol{t}\\ \boldsymbol{r}\end{smallmatrix}\big]\big\| \leqslant \|\boldsymbol{r}\| + \|\boldsymbol{t}\|$, we could find $\delta_t > 0$ and $N_{11} \in \mathbb{N}_1$ such that $\forall \boldsymbol{t} \in B_{n,1}(\delta_t)$, $\forall \boldsymbol{r} \in A_{n,1}(\delta_r)$ and $\forall n > N_{11}$,

$$\big|R_n(\boldsymbol{t}, \boldsymbol{r})\big| < \frac{1}{8}\big[\begin{smallmatrix}\boldsymbol{t}\\ \boldsymbol{r}\end{smallmatrix}\big]^\top \boldsymbol{I}_0 \big[\begin{smallmatrix}\boldsymbol{t}\\ \boldsymbol{r}\end{smallmatrix}\big]. \tag{B.7}$$

From condition 5, we know that $\lim_{n\to\infty} \hat{\boldsymbol{\eta}}_n = \boldsymbol{\eta}_0$ and $\lim_{n\to\infty} \hat{\boldsymbol{\zeta}}_n = \boldsymbol{\zeta}_0$. Applying condition 3, we can show that $\lim_{n\to\infty} \hat{\boldsymbol{I}}_n = \boldsymbol{I}_0$. Moreover, for any fixed $\boldsymbol{t}$ and $\boldsymbol{r}$, we know that $\lim_{n\to\infty} R_n(\boldsymbol{t}, \boldsymbol{r}) = 0$ (Corollary B.0.2). Therefore,

$$\begin{aligned}
&\lim_{n\to\infty} \exp\big\{L_n(\hat{\boldsymbol{\eta}}_n + \boldsymbol{t}/\sqrt{n}, \hat{\boldsymbol{\zeta}}_n + \boldsymbol{r}/\sqrt{n}) - L_n(\hat{\boldsymbol{\eta}}_n, \hat{\boldsymbol{\zeta}}_n)\big\} \\
&= \exp\left\{ -\frac{1}{2}\big[\begin{smallmatrix}\boldsymbol{t}\\ \boldsymbol{r}\end{smallmatrix}\big]^\top \boldsymbol{I}_0 \big[\begin{smallmatrix}\boldsymbol{t}\\ \boldsymbol{r}\end{smallmatrix}\big]\right\} = \exp\left\{-\frac{1}{2}\boldsymbol{t}^\top \boldsymbol{I}_0^{11}\boldsymbol{t} - \frac{1}{2}\boldsymbol{r}^\top \boldsymbol{I}_0^{22}\boldsymbol{r} - \boldsymbol{t}^\top \boldsymbol{I}_0^{12}\boldsymbol{r}\right\},
\end{aligned} \tag{B.8}$$

where $\boldsymbol{I}_0 = \big[\begin{smallmatrix}\boldsymbol{I}_0^{11} & \boldsymbol{I}_0^{12}\\ \boldsymbol{I}_0^{21} & \boldsymbol{I}_0^{22}\end{smallmatrix}\big]$. Moreover, since $\pi(\boldsymbol{\eta}|\boldsymbol{\zeta})$ is positive and continuous at $\boldsymbol{\eta} = \boldsymbol{\eta}_0$ and $\boldsymbol{\zeta} = \boldsymbol{\zeta}_0$,

$$\begin{aligned}
&\lim_{n\to\infty} g_n(\boldsymbol{t}, \boldsymbol{r})\mathbb{1}\{\boldsymbol{t} \in B_{n,1}(\delta_t)\} \\
&= \exp\left\{ -\frac{1}{2}\boldsymbol{t}^\top \boldsymbol{I}_0^{11}\boldsymbol{t} - \frac{1}{2}\boldsymbol{r}^\top \boldsymbol{I}_0^{22}\boldsymbol{r} - \boldsymbol{t}^\top \boldsymbol{I}_0^{12}\boldsymbol{r}\right\}\pi(\boldsymbol{\eta}_0|\boldsymbol{\zeta}_0).
\end{aligned} \tag{B.9}$$

From (B.8), we could find $N_{12} \in \mathbb{N}_1$ such that $\forall n > N_{12}$,

$$\exp\big\{L_n(\hat{\boldsymbol{\eta}}_n + \boldsymbol{t}/\sqrt{n}, \hat{\boldsymbol{\zeta}}_n + \boldsymbol{r}/\sqrt{n}) - L_n(\hat{\boldsymbol{\eta}}_n, \hat{\boldsymbol{\zeta}}_n)\big\} < \exp\left\{ -\frac{1}{4}\big[\begin{smallmatrix}\boldsymbol{t}\\ \boldsymbol{r}\end{smallmatrix}\big]^\top \boldsymbol{I}_0 \big[\begin{smallmatrix}\boldsymbol{t}\\ \boldsymbol{r}\end{smallmatrix}\big]\right\} \tag{B.10}$$

Let $N_1 = \max\{N_{11}, N_{12}\}$. Combining (B.7) and (B.10) we have, $\forall \boldsymbol{t} \in B_{n,1}(\delta_t)$, $\forall \boldsymbol{r} \in A_{n,1}(\delta_r)$ and $\forall n > N_1$,

$$\exp\big\{L_n(\hat{\boldsymbol{\eta}}_n + \boldsymbol{t}/\sqrt{n}, \hat{\boldsymbol{\zeta}}_n + \boldsymbol{r}/\sqrt{n}) - L_n(\hat{\boldsymbol{\eta}}_n, \hat{\boldsymbol{\zeta}}_n)\big\} < \exp\left\{ -\frac{1}{8}\big[\begin{smallmatrix}\boldsymbol{t}\\ \boldsymbol{r}\end{smallmatrix}\big]^\top \boldsymbol{I}_0 \big[\begin{smallmatrix}\boldsymbol{t}\\ \boldsymbol{r}\end{smallmatrix}\big]\right\}$$

Let $b(\delta_t, \delta_r) = \sup_{\|\boldsymbol{\eta}-\boldsymbol{\eta}_0\|<2\delta_t, \|\boldsymbol{\zeta}-\boldsymbol{\zeta}_0\|<2\delta_r} \pi(\boldsymbol{\eta}|\boldsymbol{\zeta})$. Given that $\pi(\boldsymbol{\eta}|\boldsymbol{\zeta})$ is positive and continuous at $\boldsymbol{\eta} = \boldsymbol{\eta}_0$ and $\boldsymbol{\zeta} = \boldsymbol{\zeta}_0$, we can choose δ_r and δ_t small enough so that

$b(\delta_t, \delta_r) > 0$. Then $g_n(\boldsymbol{t}, \boldsymbol{r})\mathbb{1}\{\boldsymbol{t} \in B_{n,1}(\delta_t)\}$ is bounded by

$$b(\delta_t, \delta_r)\int_{B_{n,1}(\delta_t)} \exp\Big\{ -\frac{1}{8}\big[\begin{smallmatrix}\boldsymbol{t}\\ \boldsymbol{r}\end{smallmatrix}\big]^\top \boldsymbol{I}_0\big[\begin{smallmatrix}\boldsymbol{t}\\ \boldsymbol{r}\end{smallmatrix}\big]\Big\}\mathrm{d}\boldsymbol{t},$$

which is clearly integrable. Applying DCT,

$$\begin{aligned}
&\lim_{n\to\infty}\int g_n(\boldsymbol{t}, \boldsymbol{r})\mathbb{1}\{\boldsymbol{t} \in B_{n,1}(\delta_t)\}\mathrm{d}\boldsymbol{t}\\
=&\int \exp\Big\{ -\frac{1}{2}\boldsymbol{t}^\top \boldsymbol{I}_0^{11}\boldsymbol{t} - \frac{1}{2}\boldsymbol{r}^\top \boldsymbol{I}_0^{22}\boldsymbol{r} - \boldsymbol{t}^\top \boldsymbol{I}_0^{12}\boldsymbol{r}\Big\}\pi(\boldsymbol{\eta}_0|\boldsymbol{\zeta}_0)\mathrm{d}\boldsymbol{t}\\
=&\exp\Big\{ -\frac{1}{2}\boldsymbol{r}^\top \Big(\boldsymbol{I}_0^{22} - \boldsymbol{I}_0^{21}\big(\boldsymbol{I}_0^{11}\big)^{-1}\boldsymbol{I}_0^{12}\Big)\boldsymbol{r}\Big\}\pi(\boldsymbol{\eta}_0|\boldsymbol{\zeta}_0)|\boldsymbol{I}_0^{11}/2\pi|^{-1/2}
\end{aligned}$$

We complete the this step by finding the limit for $\int g_n(\boldsymbol{t}, \boldsymbol{r})\mathbb{1}\{\boldsymbol{t} \in B_{n,2}(\delta_t)\}\mathrm{d}\boldsymbol{t}$. Similar to Step 3 in the proof of Theorem 4.2.3, we can find $\epsilon > 0$ and $N_2 \in \mathbb{N}_1$ such that $\forall n > N_2$,

$$\frac{1}{n}\Big[L_n(\hat{\boldsymbol{\eta}}_n + \boldsymbol{t}/\sqrt{n}, \hat{\boldsymbol{\zeta}}_n + \boldsymbol{r}/\sqrt{n}) - L_n(\hat{\boldsymbol{\eta}}_n, \hat{\boldsymbol{\zeta}}_n)\Big] < -\epsilon.$$

This implies

$$\begin{aligned}
&\lim_{n\to\infty}\int g_n(\boldsymbol{t}, \boldsymbol{r})\mathbb{1}\{\boldsymbol{t} \in B_{n,2}(\delta_t)\}\mathrm{d}\boldsymbol{t}\\
\leqslant&\lim_{n\to\infty}\exp\big\{-n\epsilon\big\}\int \pi(\hat{\boldsymbol{\eta}}_n + \boldsymbol{t}/\sqrt{n}|\hat{\boldsymbol{\zeta}}_n + \boldsymbol{r}/\sqrt{n})\mathbb{1}\{\boldsymbol{t} \in B_{n,2}(\delta_t)\}\mathrm{d}\boldsymbol{t}\\
\leqslant&\lim_{n\to\infty}\exp\big\{-n\epsilon\big\}\\
=&0.
\end{aligned}$$

Hence we have shown that $\forall \boldsymbol{r} \in A_{n,2}(\delta_r)$,

$$\lim_{n\to\infty} a_n(\boldsymbol{r}) = \exp\Big\{ -\frac{1}{2}\boldsymbol{r}^\top \Big(\boldsymbol{I}_0^{22} - \boldsymbol{I}_0^{21}\big(\boldsymbol{I}_0^{11}\big)^{-1}\boldsymbol{I}_0^{12}\Big)\boldsymbol{r}\Big\}\pi(\boldsymbol{\eta}_0|\boldsymbol{\zeta}_0)|\boldsymbol{I}_0^{11}/2\pi|^{-1/2} \tag{B.11}$$

Step 2 In this step we will find the limit for $\pi_n^*(\boldsymbol{t}|\boldsymbol{r})\mathbb{1}\{\boldsymbol{r} \in A_{n,1}(\delta_r)\}\mathbb{1}\{\boldsymbol{t} \in B_{n,1}(\delta_t)\}$. Since

$$\pi_n^*(\boldsymbol{t}|\boldsymbol{r})\mathbb{1}\{\boldsymbol{r} \in A_{n,1}(\delta_r)\}\mathbb{1}\{\boldsymbol{t} \in B_{n,1}(\delta_t)\} = a_n(\boldsymbol{r})^{-1}g_n(\boldsymbol{t}, \boldsymbol{r})\mathbb{1}\{\boldsymbol{t} \in B_{n,1}(\delta_t)\},$$

combining (B.9) and (B.11) we immediately have

$$\begin{aligned}&\lim_{n\to\infty} \pi_n^*(\boldsymbol{t}|\boldsymbol{r})\mathbb{1}\{\boldsymbol{r} \in A_{n,1}(\delta_r)\}\mathbb{1}\{\boldsymbol{t} \in B_{n,1}(\delta_r)\}\\ =&\lim_{n\to\infty} a_n(\boldsymbol{r}) \lim_{n\to\infty} g_n(\boldsymbol{t},\boldsymbol{r})\mathbb{1}\{\boldsymbol{t} \in B_{n,1}(\delta_t)\}\\ =&|\boldsymbol{I}_0^{11}/2\pi|^{1/2}\exp\{-\frac{1}{2}\Big[\boldsymbol{t}-\big(\boldsymbol{I}_0^{11}\big)^{-1}\boldsymbol{I}_0^{12}\boldsymbol{r}\Big]^\top \boldsymbol{I}_0^{11}\Big[\boldsymbol{t}-\big(\boldsymbol{I}_0^{11}\big)^{-1}\boldsymbol{I}_0^{12}\boldsymbol{r}\Big]\}\\ =&\phi\bigg(\boldsymbol{t}\bigg|-\big(\boldsymbol{I}_0^{11}\big)^{-1}\boldsymbol{I}_0^{12}\boldsymbol{r},\big(\boldsymbol{I}_0^{11}\big)^{-1}\bigg).\end{aligned}$$

Step 3 In this step, we complete the proof by finding a lower bound for $a_n(\boldsymbol{r})$. Applying Corollary B.0.2 one more time, by choosing δ_t and δ_r small enough, $\exists N_{31} \in \mathbb{N}_1$ such that $\forall \boldsymbol{t} \in B_{n,1}(\delta_t)$, $\forall \boldsymbol{r} \in A_{n,1}(\delta_r)$ and $\forall n > N_{21}$,

$$\big|R_n(\boldsymbol{t},\boldsymbol{r})\big| < \frac{1}{4}\big[\begin{smallmatrix}\boldsymbol{t}\\ \boldsymbol{r}\end{smallmatrix}\big]^\top \boldsymbol{I}_0\big[\begin{smallmatrix}\boldsymbol{t}\\ \boldsymbol{r}\end{smallmatrix}\big]. \tag{B.12}$$

Also from (B.8), $\exists N_{32} \in \mathbb{N}_1$ such that $\forall n > N_{22}$,

$$\begin{aligned}&\exp\big\{L_n(\hat{\boldsymbol{\eta}}_n+\boldsymbol{t}/\sqrt{n},\hat{\boldsymbol{\zeta}}_n+\boldsymbol{r}/\sqrt{n})-L_n(\hat{\boldsymbol{\eta}}_n,\hat{\boldsymbol{\zeta}}_n)\big\}\\ >&\exp\bigg\{-\frac{1}{2}\big[\begin{smallmatrix}\boldsymbol{t}\\ \boldsymbol{r}\end{smallmatrix}\big]^\top\big(\boldsymbol{I}_0+\frac{1}{2}\boldsymbol{I}_0\big)\big[\begin{smallmatrix}\boldsymbol{t}\\ \boldsymbol{r}\end{smallmatrix}\big]\bigg\}.\end{aligned}$$

Combining above and (B.12), we immediately have that $\forall \boldsymbol{r} \in A_{n,1}(\delta_r)$, $\forall \boldsymbol{t} \in B_{n,1}(\delta_t)$ and $\forall n > \max\{N_{31}, N_{32}\}$,

$$\exp\big\{L_n(\hat{\boldsymbol{\eta}}_n+\boldsymbol{t}/\sqrt{n},\hat{\boldsymbol{\zeta}}_n+\boldsymbol{r}/\sqrt{n})-L_n(\hat{\boldsymbol{\eta}}_n,\hat{\boldsymbol{\zeta}}_n)\big\} > \exp\bigg\{-\big[\begin{smallmatrix}\boldsymbol{t}\\ \boldsymbol{r}\end{smallmatrix}\big]^\top \boldsymbol{I}_0\big[\begin{smallmatrix}\boldsymbol{t}\\ \boldsymbol{r}\end{smallmatrix}\big]\bigg\}. \tag{B.13}$$

Let $c(\delta_r,\delta_t) = \inf_{\|\boldsymbol{\eta}-\boldsymbol{\eta}_0\|<2\delta_t,\|\boldsymbol{\zeta}-\boldsymbol{\zeta}_0\|<2\delta_r} \pi(\boldsymbol{\eta}|\boldsymbol{\zeta})$. Given that $\pi(\boldsymbol{\eta}|\boldsymbol{\zeta})$ is positive and continuous at $\boldsymbol{\eta} = \boldsymbol{\eta}_0$ and $\boldsymbol{\zeta} = \boldsymbol{\zeta}_0$, we can choose δ_r and δ_t small enough so that $c(\delta_r,\delta_t) > 0$. Since $\lim_{n\to\infty} \hat{\boldsymbol{\eta}}_n = \boldsymbol{\eta}_0$ and $\lim_{n\to\infty} \hat{\boldsymbol{\zeta}}_n = \boldsymbol{\zeta}_0$, there $\exists N_{33} \in \mathbb{N}_1$ such that $\forall \boldsymbol{r} \in A_{n,1}(\delta_r)$ and $\forall \boldsymbol{t} \in B_{n,1}(\delta_t)$, $\pi(\hat{\boldsymbol{\eta}}_n+\boldsymbol{t}/\sqrt{n}|\hat{\boldsymbol{\zeta}}_n+\boldsymbol{r}/\sqrt{n}) > c(\delta_r,\delta_t)$. Letting $N_3 = \max\{N_{31}, N_{32}, N_{33}\}$, together with (B.13), we have shown that $\forall n > N_3$, $a_n(\boldsymbol{r})$

is lower-bounded by

$$
\begin{aligned}
&c(\delta_r, \delta_t) \int \exp\left\{ - \left[\begin{smallmatrix} \boldsymbol{t} \\ \boldsymbol{r} \end{smallmatrix}\right]^\top \boldsymbol{I}_0 \left[\begin{smallmatrix} \boldsymbol{t} \\ \boldsymbol{r} \end{smallmatrix}\right] \right\} \mathrm{d}\boldsymbol{t} \\
=&c(\delta_r, \delta_t) |4\pi \boldsymbol{I}_0^{11}|^{\frac{1}{2}} \exp\left\{ - \boldsymbol{r}^\top \left(\boldsymbol{I}_0^{22} - \boldsymbol{I}_0^{21} \left(\boldsymbol{I}_0^{11}\right)^{-1} \boldsymbol{I}_0^{12} \right) \boldsymbol{r} \right\}. \qquad \text{(B.14)}
\end{aligned}
$$

Step 4 Consider $\delta_t > 0$. Since $\lim_{n\to\infty} \hat{\boldsymbol{\eta}}_n = \boldsymbol{\eta}_0$ (condition 5), for $\forall \boldsymbol{t} \in B_{n,2}(\delta_t)$ and $\forall \boldsymbol{r} \in A_{n,1}(\delta_r)$, $\exists N_1 \in \mathbb{N}_1$ such that $\forall n > N_1$,

$$
\left\| \left[\begin{smallmatrix} \hat{\boldsymbol{\eta}}_n + \boldsymbol{t}/\sqrt{n} - \boldsymbol{\eta}_0 \\ \hat{\boldsymbol{\zeta}}_n + \boldsymbol{t}/\sqrt{n} - \boldsymbol{\zeta}_0 \end{smallmatrix}\right] \right\| > \|\hat{\boldsymbol{\eta}}_n + \boldsymbol{t}/\sqrt{n} - \boldsymbol{\eta}_0\| > \frac{\delta_t}{2}.
$$

Applying condition 4, $\exists \epsilon > 0$ and $N_2 \geqslant N_1$ such that $\forall \boldsymbol{t} \in B_{n,2}(\delta_t)$, $\forall \boldsymbol{r} \in A_{n,1}(\delta_r)$ and $\forall n > N_2$,

$$
\frac{1}{n}\left[L_n(\hat{\boldsymbol{\eta}}_n + \boldsymbol{t}/\sqrt{n}, \hat{\boldsymbol{\zeta}}_n + \boldsymbol{r}/\sqrt{n}) - L_n(\boldsymbol{\eta}_0, \boldsymbol{\zeta}_0) \right] < -3\epsilon. \qquad \text{(B.15)}
$$

From condition 2, $\exists N_3 \in \mathbb{N}_1$ such that $\forall n > N_3$, $\left[\begin{smallmatrix} \hat{\boldsymbol{\eta}}_n \\ \hat{\boldsymbol{\zeta}}_n \end{smallmatrix}\right] \in \mathcal{B}_{\boldsymbol{z}}\left(\boldsymbol{z}_0, \delta\right)$ where $\ell_2(\boldsymbol{\eta}, \boldsymbol{\zeta}, \boldsymbol{y})$ is thrice differentiable with respect to $\boldsymbol{\eta}$ and $\boldsymbol{\zeta}$. Applying mean value theorem, we have $\frac{1}{n}\left[L_n(\hat{\boldsymbol{\eta}}_n, \hat{\boldsymbol{\zeta}}_n) - L_n(\boldsymbol{\eta}_0, \boldsymbol{\zeta}_0) \right] = \frac{1}{n} L_n^{(1)}(\boldsymbol{\eta}_n', \boldsymbol{\zeta}_n')^\top \left[\begin{smallmatrix} \hat{\boldsymbol{\eta}}_n - \boldsymbol{\eta}_0 \\ \hat{\boldsymbol{\zeta}}_n - \boldsymbol{\zeta}_0 \end{smallmatrix}\right]$, where $\boldsymbol{\eta}_n'$ is between $\hat{\boldsymbol{\eta}}_n$ and $\boldsymbol{\eta}_0$ and $\boldsymbol{\zeta}_n'$ is between $\hat{\boldsymbol{\zeta}}_n$ and $\boldsymbol{\zeta}_0$. Using condition 3 and continuous mapping theorem, it can be shown that $\lim_{n\to\infty} \frac{1}{n} L_n^{(1)}(\boldsymbol{\eta}_n', \boldsymbol{\zeta}_n') = \boldsymbol{0}$. Hence, $\exists N_4 > N_3$ such that $\forall n > N_4$, $\frac{1}{n}\left| L_n(\hat{\boldsymbol{\eta}}_n, \hat{\boldsymbol{\zeta}}_n) - L_n(\boldsymbol{\eta}_0, \boldsymbol{\zeta}_0) \right| < \epsilon$. Letting $N = \max\{N_1, \ldots, N_4\}$ and using (B.15), it can be shown that $\forall n > N$,

$$
\frac{1}{n}\left[L_n(\hat{\boldsymbol{\eta}}_n + \boldsymbol{t}/\sqrt{n}, \hat{\boldsymbol{\zeta}}_n + \boldsymbol{r}/\sqrt{n}) - L_n(\hat{\boldsymbol{\eta}}_n, \hat{\boldsymbol{\zeta}}_n) \right] < -2\epsilon. \qquad \text{(B.16)}
$$

Using Lemma B.0.3, we can choose δ_t and δ_r to be small enough so that $a_n(\boldsymbol{r}) \mathbb{1}\{\boldsymbol{r} \in A_{n,1}(\delta_r)\}$ is lower bounded by (B.14). Since $\boldsymbol{I}_0^{22} - \boldsymbol{I}_0^{21} \left(\boldsymbol{I}_0^{11}\right)^{-1} \boldsymbol{I}_0^{12}$ is positive definite, we can always choose δ_r small enough so that

$$
\exp\left\{ -\boldsymbol{r}^\top \left(\boldsymbol{I}_0^{22} - \boldsymbol{I}_0^{21} \left(\boldsymbol{I}_0^{11}\right)^{-1} \boldsymbol{I}_0^{12} \right) \boldsymbol{r} \right\} < \exp\{n\epsilon\}. \qquad \text{(B.17)}
$$

Combining (B.14), (B.16) and (B.17), it can be shown that

$$pi^*_{n,2}(\boldsymbol{t}|\boldsymbol{r})\mathbb{1}\{\boldsymbol{r} \in A_{n,1}(\delta_r), \boldsymbol{t} \in B_{n,1}(\delta_t)\} < \frac{|4\pi \boldsymbol{I}_0^{11}|^{-1/2}}{c(\delta_r, \delta_t)} \exp\{-n\epsilon\}.$$

$\square$

B.0.4 Proof of Theorem 4.2.3

We prove the theorem in the following four steps.

Step 1 Consider $\boldsymbol{r} = \sqrt{n}(\boldsymbol{\zeta} - \hat{\boldsymbol{\zeta}}_n)$ and its posterior density $\pi^*_{n,1}(\boldsymbol{r})$. Applying Lemma 4.2.2 and a simple change of variable, we can show

$$\lim_{n\to\infty} \int |\pi^*_{n,1}(\boldsymbol{r}) - \phi\big(\boldsymbol{r}|\boldsymbol{\mu}_n, \tilde{\boldsymbol{I}}_0^{-1}\big)|\mathrm{d}\boldsymbol{r} = 0. \tag{B.18}$$

It can show that the integral in (4.9) is bounded by

$$\int\int \pi^*_{n,2}(\boldsymbol{t}|\boldsymbol{r})\big|\pi^*_{n,1}(\boldsymbol{r}) - \phi\big(\boldsymbol{r}|\boldsymbol{\mu}_n, \tilde{\boldsymbol{I}}_0^{-1}\big)\big|\mathrm{d}\boldsymbol{r}\mathrm{d}\boldsymbol{t}$$

$$+ \int\int \left|\pi^*_{n,2}(\boldsymbol{t}|\boldsymbol{r}) - \phi\left(\boldsymbol{t}\middle|-\big(\boldsymbol{I}_0^{11}\big)^{-1}\boldsymbol{I}_0^{12}\boldsymbol{r}, \big(\boldsymbol{I}_0^{11}\big)^{-1}\right)\right| \phi\big(\boldsymbol{r}|\boldsymbol{\mu}_n, \tilde{\boldsymbol{I}}_0^{-1}\big)\mathrm{d}\boldsymbol{r}\mathrm{d}\boldsymbol{t},$$

where the first integral equals $\int\big|\pi^*_n(\boldsymbol{r}) - \phi\big(\boldsymbol{r}|\boldsymbol{\mu}_n, \tilde{\boldsymbol{I}}^{-1}\big)\big|\mathrm{d}\boldsymbol{r}$ which goes to zero by (B.18). Hence showing

$$\lim_{n\to\infty}\int\int \left|\pi^*_{n,2}(\boldsymbol{t}|\boldsymbol{r}) - \phi\left(\boldsymbol{t}\middle|-\big(\boldsymbol{I}_0^{11}\big)^{-1}\boldsymbol{I}_0^{12}\boldsymbol{r}, \big(\boldsymbol{I}_0^{11}\big)^{-1}\right)\right| \phi\big(\boldsymbol{r}|\boldsymbol{\mu}_n, \tilde{\boldsymbol{I}}^{-1}\big)\mathrm{d}\boldsymbol{r}\mathrm{d}\boldsymbol{t} = 0 \tag{B.19}$$

would be enough for proving (4.9).

Step 2 For $\delta_r > 0$, the integral in (B.19) can be written as

$$\int\int_{\boldsymbol{r}\in A_{n,1}(\delta_r)} \left|\pi^*_{n,2}(\boldsymbol{t}|\boldsymbol{r}) - \phi\left(\boldsymbol{t}\middle|-\big(\boldsymbol{I}_0^{11}\big)^{-1}\boldsymbol{I}_0^{12}\boldsymbol{r}, \big(\boldsymbol{I}_0^{11}\big)^{-1}\right)\right| \phi\big(\boldsymbol{r}|\boldsymbol{\mu}_n, \tilde{\boldsymbol{I}}^{-1}\big)\mathrm{d}\boldsymbol{r}\mathrm{d}\boldsymbol{t}$$

$$+ \int\int_{\boldsymbol{r}\in A_{n,2}(\delta_r)} \left|\pi^*_{n,2}(\boldsymbol{t}|\boldsymbol{r}) - \phi\left(\boldsymbol{t}\middle|-\big(\boldsymbol{I}_0^{11}\big)^{-1}\boldsymbol{I}_0^{12}\boldsymbol{r}, \big(\boldsymbol{I}_0^{11}\big)^{-1}\right)\right| \phi\big(\boldsymbol{r}|\boldsymbol{\mu}_n, \tilde{\boldsymbol{I}}^{-1}\big)\mathrm{d}\boldsymbol{r}\mathrm{d}\boldsymbol{t},$$

where the second intergal is clearly bounded by

$$\int_{\boldsymbol{r}\in A_{n,2}(\delta_r)}\int\left|\pi^*_{n,2}(\boldsymbol{t}|\boldsymbol{r})-\phi\left(\boldsymbol{t}\middle|-\left(\boldsymbol{I}_0^{11}\right)^{-1}\boldsymbol{I}_0^{12}\boldsymbol{r},\left(\boldsymbol{I}_0^{11}\right)^{-1}\right)\right|\mathrm{d}\boldsymbol{t}\phi\big(\boldsymbol{r}\big|\boldsymbol{\mu}_n,\tilde{\boldsymbol{I}}^{-1}\big)\mathrm{d}\boldsymbol{r}$$

$$\leqslant 2\int_{\boldsymbol{r}\in A_{n,2}(\delta_r)}\phi\big(\boldsymbol{r}\big|\boldsymbol{\mu}_n,\tilde{\boldsymbol{I}}^{-1}\big)\mathrm{d}\boldsymbol{r}$$

Transforming $\boldsymbol{r}$ back to $\boldsymbol{\zeta}$,

$$\int_{\boldsymbol{r}\in A_{n,2}(\delta_r)}\phi\big(\boldsymbol{r}\big|\boldsymbol{\mu}_n,\tilde{\boldsymbol{I}}^{-1}\big)\mathrm{d}\boldsymbol{r}=\Phi\left(\|\boldsymbol{\zeta}-\boldsymbol{\zeta}_0\|>\delta_r\middle|\tilde{\boldsymbol{\zeta}}_n,\tilde{\boldsymbol{I}}_0/n\right).$$

From condition 5, $\lim_{n\to\infty}\hat{\boldsymbol{\zeta}}_n=\boldsymbol{\zeta}_0$ and hence by continuous mapping theorem,

$$\lim_{n\to\infty}\Phi\left(\|\boldsymbol{\zeta}-\boldsymbol{\zeta}_0\|>\delta_r\middle|\tilde{\boldsymbol{\zeta}}_n,\tilde{\boldsymbol{I}}_0/n\right)=\lim_{n\to\infty}\Phi\left(\|\boldsymbol{\zeta}-\boldsymbol{\zeta}_0\|>\delta_r\middle|\boldsymbol{\zeta}_0,\tilde{\boldsymbol{I}}_0/n\right)=0.$$

This implies that showing

$$\int\int_{\boldsymbol{r}\in A_{n,1}(\delta_r)}\left|\pi^*_{n,2}(\boldsymbol{t}|\boldsymbol{r})-\phi\left(\boldsymbol{t}\middle|-\left(\boldsymbol{I}_0^{11}\right)^{-1}\boldsymbol{I}_0^{12}\boldsymbol{r},\left(\boldsymbol{I}_0^{11}\right)^{-1}\right)\right|\phi\big(\boldsymbol{r}\big|\boldsymbol{\mu}_n,\tilde{\boldsymbol{I}}^{-1}\big)\mathrm{d}\boldsymbol{r}\mathrm{d}\boldsymbol{t}\to 0 \tag{B.20}$$

would be enough for proving (B.19).

Step 3 Applying (B.5) in Lemma B.0.3, we have

$$\int_{\boldsymbol{t}\in B_{n,2}(\delta_t)}\int_{\boldsymbol{r}\in A_{n,1}(\delta_r)}\pi^*_{n,2}(\boldsymbol{t}|\boldsymbol{r})\phi\big(\boldsymbol{r}\big|\boldsymbol{\mu}_n,\tilde{\boldsymbol{I}}_0^{-1}\big)\mathrm{d}\boldsymbol{r}\mathrm{d}\boldsymbol{t}<\frac{|4\pi\boldsymbol{I}_0^{11}|^{-1/2}}{c(\delta_r,\delta_t)}\exp\{-n\epsilon\}\to 0.$$

Moreover, it can be easily shown that

$$\int_{\boldsymbol{t}\in B_{n,2}(\delta_t)}\int_{\boldsymbol{r}\in A_{n,1}(\delta_r)}\phi\left(\boldsymbol{t}\middle|-\left(\boldsymbol{I}_0^{11}\right)^{-1}\boldsymbol{I}_0^{12}\boldsymbol{r},\left(\boldsymbol{I}_0^{11}\right)^{-1}\right)\phi\big(\boldsymbol{r}\big|\boldsymbol{\mu}_n,\tilde{\boldsymbol{I}}_0^{-1}\big)\mathrm{d}\boldsymbol{r}\mathrm{d}\boldsymbol{t}\to 0.$$

Hence we have shown that

$$\begin{aligned}\lim_{n\to\infty}\int_{\boldsymbol{t}\in B_{n,2}(\delta_t)}\int_{\boldsymbol{r}\in A_{n,1}(\delta_r)}&\left|\pi^*_{n,2}(\boldsymbol{t}|\boldsymbol{r})-\phi\left(\boldsymbol{t}\middle|-\left(\boldsymbol{I}_0^{11}\right)^{-1}\boldsymbol{I}_0^{12}\boldsymbol{r},\left(\boldsymbol{I}_0^{11}\right)^{-1}\right)\right|\\&\times\phi\big(\boldsymbol{r}\big|\boldsymbol{\mu}_n,\tilde{\boldsymbol{I}}^{-1}\big)\mathrm{d}\boldsymbol{r}\mathrm{d}\boldsymbol{t}=0.\end{aligned}$$

This implies that showing

$$\lim_{n\to\infty}\int_{\boldsymbol{t}\in B_{n,1}(\delta_t)}\int_{\boldsymbol{r}\in A_{n,1}(\delta_r)}\left|\pi^*_{n,2}(\boldsymbol{t}|\boldsymbol{r})-\phi\left(\boldsymbol{t}\middle|-\left(\boldsymbol{I}_0^{11}\right)^{-1}\boldsymbol{I}_0^{12}\boldsymbol{r},\left(\boldsymbol{I}_0^{11}\right)^{-1}\right)\right| \\ \times\phi\big(\boldsymbol{r}\big|\boldsymbol{\mu}_n,\tilde{\boldsymbol{I}}_0^{-1}\big)\mathrm{d}\boldsymbol{r}\mathrm{d}\boldsymbol{t}=0 \tag{B.21}$$

would be enough for proving (B.20).

Step 4 Since $\phi\big(\boldsymbol{r}\big|\boldsymbol{\mu}_n,\tilde{\boldsymbol{I}}_0^{-1}\big)$ is bounded by $|\tilde{\boldsymbol{I}}/2\pi|^{1/2}$, applying (B.4) in Lemma B.0.3 again we have

$$\lim_{n\to\infty}\left|\pi^*_{n,2}(\boldsymbol{t}|\boldsymbol{r})-\phi\left(\boldsymbol{t}\middle|-\left(\boldsymbol{I}_0^{11}\right)^{-1}\boldsymbol{I}_0^{12}\boldsymbol{r},\left(\boldsymbol{I}_0^{11}\right)^{-1}\right)\right| \\ \times\phi\big(\boldsymbol{r}\big|\boldsymbol{\mu}_n,\tilde{\boldsymbol{I}}_0^{-1}\big)\mathbb{1}_{\boldsymbol{t}\in B_{n,1}(\delta_t)}\mathbb{1}_{\boldsymbol{r}\in A_{n,1}(\delta_r)}=0.$$

Moreover,

$$\begin{aligned}&\int_{\boldsymbol{t}\in B_{n,1}(\delta_t)}\int_{\boldsymbol{r}\in A_{n,1}(\delta_r)}\left|\pi^*_{n,2}(\boldsymbol{t}|\boldsymbol{r})-\phi\left(\boldsymbol{t}\middle|-\left(\boldsymbol{I}_0^{11}\right)^{-1}\boldsymbol{I}_0^{12}\boldsymbol{r},\left(\boldsymbol{I}_0^{11}\right)^{-1}\right)\right|\\&\qquad\times\phi\big(\boldsymbol{r}\big|\boldsymbol{\mu}_n,\tilde{\boldsymbol{I}}_0^{-1}\big)\mathrm{d}\boldsymbol{r}\mathrm{d}\boldsymbol{t}\\<&\int_{\boldsymbol{t}\in B_{n,1}(\delta_t)}\int_{\boldsymbol{r}\in A_{n,1}(\delta_r)}\phi\left(\boldsymbol{t}\middle|-\left(\boldsymbol{I}_0^{11}\right)^{-1}\boldsymbol{I}_0^{12}\boldsymbol{r},\left(\boldsymbol{I}_0^{11}\right)^{-1}\right)\phi\big(\boldsymbol{r}\big|\boldsymbol{\mu}_n,\tilde{\boldsymbol{I}}_0^{-1}\big)\mathrm{d}\boldsymbol{r}\mathrm{d}\boldsymbol{t}\\&\qquad+\int_{\boldsymbol{t}\in B_{n,1}(\delta_t)}\int_{\boldsymbol{r}\in A_{n,1}(\delta_r)}\pi^*_{n,2}(\boldsymbol{t}|\boldsymbol{r})\phi\big(\boldsymbol{r}\big|\boldsymbol{\mu}_n,\tilde{\boldsymbol{I}}_0^{-1}\big)\mathrm{d}\boldsymbol{r}\mathrm{d}\boldsymbol{t}\\\leqslant&2.\end{aligned}$$

Therefore by Scheffé's lemma we have shown (B.21).

B.0.5 Proof of Lemma B.0.4

We now provide a lemma that will be needed in our later proof of Corollary 4.2.4. It characterizes the asymptotic difference between $\tilde{\boldsymbol{\zeta}}_n$ and $\hat{\boldsymbol{\zeta}}_n$.

Lemma B.0.4. *Suppose that the conditions for Theorem 4.2.3 hold, then*

$$\sqrt{n}\big(\tilde{\boldsymbol{\zeta}}_n-\hat{\boldsymbol{\zeta}}_n\big)\xrightarrow{d}\mathcal{N}(\boldsymbol{0},\boldsymbol{V}),$$

where $\boldsymbol{V}$ is a positive definite matrix.

Proof. Expanding Taylor series for $\frac{1}{n}L_n^{(1)}(\hat{\boldsymbol{\eta}}_n, \hat{\boldsymbol{\zeta}}_n)$ we have

$$\mathbf{0} = \frac{1}{n}L_n^{(1)}(\hat{\boldsymbol{\eta}}_n, \hat{\boldsymbol{\zeta}}_n) = \frac{1}{n}L_n^{(1)}(\boldsymbol{\eta}_0, \boldsymbol{\zeta}_0) + \frac{1}{n}L_n^{(2)}(\boldsymbol{\eta}_n^{'}, \boldsymbol{\zeta}_n^{'})\left[\begin{smallmatrix} \hat{\boldsymbol{\eta}}_n - \boldsymbol{\eta}_0 \\ \hat{\boldsymbol{\zeta}}_n - \boldsymbol{\zeta}_0 \end{smallmatrix}\right],$$

where $\boldsymbol{\eta}_n^{'}$ is between $\boldsymbol{\eta}_0$ and $\hat{\boldsymbol{\eta}}_n$ and $\boldsymbol{\zeta}_n^{'}$ is between $\boldsymbol{\zeta}_0$ and $\hat{\boldsymbol{\zeta}}_n$. Letting $\hat{\boldsymbol{I}}_n = \frac{1}{n}L_n^{(2)}(\boldsymbol{\eta}_n^{'}, \boldsymbol{\zeta}_n^{'})$, we have

$$\left[\begin{smallmatrix} \sqrt{n}(\hat{\boldsymbol{\eta}}_n - \boldsymbol{\eta}_0) \\ \sqrt{n}\left(\hat{\boldsymbol{\zeta}}_n - \boldsymbol{\zeta}_0\right) \end{smallmatrix}\right] = -\hat{\boldsymbol{I}}_n^{-1}\sqrt{n}\frac{1}{n}L_n^{(1)}(\boldsymbol{\eta}_0, \boldsymbol{\zeta}_0). \tag{B.22}$$

Since condition 5 holds for $f_2(\boldsymbol{y}|\boldsymbol{\eta}, \boldsymbol{\zeta})$, we know that $\lim_{n\to\infty} \hat{\boldsymbol{\eta}}_n = \boldsymbol{\eta}_0$ and $\lim_{n\to\infty} \hat{\boldsymbol{\zeta}}_n = \boldsymbol{\zeta}_0$. Applying continuous mapping theorem and condition 3, we have that $\lim_{n\to\infty} \hat{\boldsymbol{I}}_n = \boldsymbol{I}_0$. Letting $\boldsymbol{D}_1$ denote a $(d_\eta + d_\zeta)$-dimensional identity matrix, we can rewrite (B.22) as

$$\begin{aligned} \left[\begin{smallmatrix} \sqrt{n}(\hat{\boldsymbol{\eta}}_n - \boldsymbol{\eta}_0) \\ \sqrt{n}\left(\hat{\boldsymbol{\zeta}}_n - \boldsymbol{\zeta}_0\right) \end{smallmatrix}\right] &= -\,\hat{\boldsymbol{I}}_n^{-1}\boldsymbol{I}_0\sqrt{n}\frac{1}{n}\boldsymbol{I}_0^{-1}L_n^{(1)}(\boldsymbol{\eta}_0, \boldsymbol{\zeta}_0) \\ &= -\left(\hat{\boldsymbol{I}}_n^{-1}\boldsymbol{I}_0 - \boldsymbol{D}_1\right)\sqrt{n}\frac{1}{n}\boldsymbol{I}_0^{-1}L_n^{(1)}(\boldsymbol{\eta}_0, \boldsymbol{\zeta}_0) - \sqrt{n}\frac{1}{n}\boldsymbol{I}_0^{-1}L_n^{(1)}(\boldsymbol{\eta}_0, \boldsymbol{\zeta}_0) \end{aligned}$$

Before proceeding, we introduce the following notation. Letting $\boldsymbol{A}$ be any $(d_\eta + d_\zeta)$-dimensional square matrix, we let

$$\boldsymbol{A} = \begin{bmatrix} \text{upper}(\boldsymbol{A}) \\ \text{lower}(\boldsymbol{A}) \end{bmatrix},$$

where upper($\boldsymbol{A}$) is a $d_\eta \times (d_\eta + d_\zeta)$-dimensional matrix and lower($\boldsymbol{A}$) is a $d_\zeta \times (d_\eta + d_\zeta)$ matrix. Using this notation,

$$\begin{aligned} \sqrt{n}\left(\hat{\boldsymbol{\zeta}}_n - \boldsymbol{\zeta}_0\right) = &-\text{lower}\left[\left(\hat{\boldsymbol{I}}_n^{-1}\boldsymbol{I}_0 - \boldsymbol{D}_1\right)\right]\sqrt{n}\frac{1}{n}\boldsymbol{I}_0^{-1}L_n^{(1)}(\boldsymbol{\eta}_0, \boldsymbol{\zeta}_0) \\ &- \text{lower}\left(\boldsymbol{D}_1\right)\sqrt{n}\frac{1}{n}\boldsymbol{I}_0^{-1}L_n^{(1)}(\boldsymbol{\eta}_0, \boldsymbol{\zeta}_0) \end{aligned} \tag{B.23}$$

Expanding Taylor series for $\frac{1}{n}Q_n^{(1)}(\tilde{\boldsymbol{\zeta}}_n)$, we have

$$\mathbf{0} = \frac{1}{n}Q_n^{(1)}(\tilde{\boldsymbol{\zeta}}_n) = \frac{1}{n}Q_n^{(1)}(\boldsymbol{\zeta}_0) + \frac{1}{n}Q_n^{(2)}(\boldsymbol{\zeta}_n^{*})\left(\tilde{\boldsymbol{\zeta}}_n - \boldsymbol{\zeta}_0\right),$$

where $\boldsymbol{\zeta}_n^*$ is between $\boldsymbol{\zeta}_0$ and $\tilde{\boldsymbol{\zeta}}_n$. Letting $\tilde{\boldsymbol{I}}_n = \frac{1}{n} Q_n^{(2)}(\boldsymbol{\zeta}_n^*)$, we have

$$\sqrt{n}\big(\tilde{\boldsymbol{\zeta}}_n - \boldsymbol{\zeta}_0\big) = -\tilde{\boldsymbol{I}}_n^{-1} \sqrt{n} \frac{1}{n} Q_n^{(1)}(\boldsymbol{\zeta}_0). \tag{B.24}$$

Similarly, using the fact that conditions 3, 5 hold for $f_1(\boldsymbol{y}|\boldsymbol{\zeta})$ and the continuous mapping theorem, we have that $\lim_{n\to\infty} \tilde{\boldsymbol{I}}_n = \tilde{\boldsymbol{I}}_0$. Letting $\boldsymbol{D}_2$ denote the d_ζ-dimensional identity matrix, we can rewrite (B.24) as

$$\begin{aligned}
\sqrt{n}\big(\tilde{\boldsymbol{\zeta}}_n - \boldsymbol{\zeta}_0\big) &= -\, \tilde{\boldsymbol{I}}_n^{-1} \tilde{\boldsymbol{I}}_0 \sqrt{n} \frac{1}{n} \tilde{\boldsymbol{I}}_0^{-1} Q_n^{(1)}(\boldsymbol{\zeta}_0) \\
&= -\, \big(\tilde{\boldsymbol{I}}_n^{-1} \tilde{\boldsymbol{I}}_0 - \boldsymbol{D}_2\big) \sqrt{n} \frac{1}{n} \tilde{\boldsymbol{I}}^{-1} Q_n^{(1)}(\boldsymbol{\zeta}_0) - \sqrt{n} \frac{1}{n} \tilde{\boldsymbol{I}}_0^{-1} Q_n^{(1)}(\boldsymbol{\zeta}_0)
\end{aligned}$$

Combining above and (B.23), we have

$$\begin{aligned}
&\sqrt{n}\big(\tilde{\boldsymbol{\zeta}}_n - \hat{\boldsymbol{\zeta}}_n\big) \\
=& -\, \big(\tilde{\boldsymbol{I}}_n^{-1} \tilde{\boldsymbol{I}}_0 - \boldsymbol{D}_2\big) \sqrt{n} \frac{1}{n} \tilde{\boldsymbol{I}}_0^{-1} Q_n^{(1)}(\boldsymbol{\zeta}_0) + \mathrm{lower}\Big[\big(\hat{\boldsymbol{I}}_n^{-1} \boldsymbol{I}_0 - \boldsymbol{D}_1\big)\Big] \sqrt{n} \frac{1}{n} \boldsymbol{I}_0^{-1} L_n^{(1)}(\boldsymbol{\eta}_0, \boldsymbol{\zeta}_0) \\
& - \sqrt{n} \frac{1}{n} \bigg[\tilde{\boldsymbol{I}}_0^{-1} Q_n^{(1)}(\boldsymbol{\zeta}_0) - \mathrm{lower}\big(\boldsymbol{D}_1\big) \boldsymbol{I}_0^{-1} L_n^{(1)}(\boldsymbol{\eta}_0, \boldsymbol{\zeta}_0) \bigg].
\end{aligned}$$

Since condition 3 holds for $f_1(\boldsymbol{y}|\boldsymbol{\zeta})$, we know that $\mathbb{E}_{\boldsymbol{\theta}_0} \nabla_{\boldsymbol{\zeta}} \ell_1(\boldsymbol{\zeta}, \boldsymbol{y})|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} = \boldsymbol{0}$, and hence $\mathrm{Var}_{\boldsymbol{\theta}_0} \nabla_{\boldsymbol{\zeta}} \ell_1(\boldsymbol{\zeta}, \boldsymbol{y})|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} = \tilde{\boldsymbol{I}}_0$. Similarly, since condition 3 holds for $f_2(\boldsymbol{y}|\boldsymbol{\eta}, \boldsymbol{\zeta})$, we know that $\mathbb{E}_{\boldsymbol{\theta}_0} \nabla_{\boldsymbol{\eta},\boldsymbol{\zeta}} \ell_2(\boldsymbol{\eta}, \boldsymbol{\zeta}, \boldsymbol{y})|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} = 0$ and $\mathrm{Var}_{\boldsymbol{\theta}_0} \nabla_{\boldsymbol{\eta},\boldsymbol{\zeta}} \ell_2(\boldsymbol{\eta}, \boldsymbol{\zeta}, \boldsymbol{y})|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} = \boldsymbol{I}_0$. It is easily seen that

$$\mathbb{E}_{\boldsymbol{\theta}_0} \tilde{\boldsymbol{I}}_0^{-1} \nabla_{\boldsymbol{\zeta}} \ell_1(\boldsymbol{\zeta}, \boldsymbol{y})|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} = \boldsymbol{0}, \quad \mathbb{E}_{\boldsymbol{\theta}_0} \mathrm{lower}\big(\boldsymbol{D}_1\big) \boldsymbol{I}_0^{-1} \nabla_{\boldsymbol{\eta},\boldsymbol{\zeta}} \ell_2(\boldsymbol{\eta}, \boldsymbol{\zeta}, \boldsymbol{y})|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} = \boldsymbol{0}$$

and hence

$$\mathbb{E}_{\boldsymbol{\theta}_0} \Big[\tilde{\boldsymbol{I}}_0^{-1} \nabla_{\boldsymbol{\zeta}} \ell_1(\boldsymbol{\zeta}, \boldsymbol{y}) - \mathrm{lower}\big(\boldsymbol{D}_1\big) \boldsymbol{I}_0^{-1} \nabla_{\boldsymbol{\eta},\boldsymbol{\zeta}} \ell_2(\boldsymbol{\eta}, \boldsymbol{\zeta}, \boldsymbol{y}) \Big] \bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} = \boldsymbol{0}.$$

Also, it can be shown that

$$\begin{aligned}
\mathrm{Var}_{\boldsymbol{\theta}_0} \tilde{\boldsymbol{I}}_0^{-1} \nabla_{\boldsymbol{\zeta}} \ell_1(\boldsymbol{\zeta}, \boldsymbol{y})|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} &= \tilde{\boldsymbol{I}}_0^{-1} \tilde{\boldsymbol{I}}_0 \tilde{\boldsymbol{I}}_0^{-1} = \tilde{\boldsymbol{I}}_0^{-1}, \\
\mathrm{Var}_{\boldsymbol{\theta}_0} \boldsymbol{I}_0^{-1} \nabla_{\boldsymbol{\eta},\boldsymbol{\zeta}} \ell_2(\boldsymbol{\eta}, \boldsymbol{\zeta}, \boldsymbol{y})|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} &= \boldsymbol{I}_0^{-1} \boldsymbol{I}_0 \boldsymbol{I}_0^{-1} = \boldsymbol{I}_0^{-1}.
\end{aligned}$$

Applying central limit theorem (CLT), we have

$$\sqrt{n}\frac{1}{n}\tilde{\boldsymbol{I}}^{-1}Q_n^{(1)}(\boldsymbol{\zeta}_0) \xrightarrow{d} \mathcal{N}\big(\mathbf{0}, \tilde{\boldsymbol{I}}^{-1}\big) \tag{B.25}$$

$$\sqrt{n}\frac{1}{n}\boldsymbol{I}_0^{-1}L_n^{(1)}(\boldsymbol{\eta}_0, \boldsymbol{\zeta}_0) \xrightarrow{d} \mathcal{N}\big(\mathbf{0}, \boldsymbol{I}_0^{-1}\big). \tag{B.26}$$

Introducing

$$\boldsymbol{V} = \mathrm{Var}_{\boldsymbol{\theta}_0}\Big[\tilde{\boldsymbol{I}}_0^{-1}\nabla_{\boldsymbol{\zeta}}\ell_1(\boldsymbol{\zeta}, \boldsymbol{y}) - \mathrm{lower}\big(\boldsymbol{D}_1\big)\boldsymbol{I}_0^{-1}\nabla_{\boldsymbol{\eta},\boldsymbol{\zeta}}\ell_2(\boldsymbol{\eta}, \boldsymbol{\zeta}, \boldsymbol{y})\Big]\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}$$

and applying CLT again, we have

$$\sqrt{n}\frac{1}{n}\Big[\tilde{\boldsymbol{I}}_0^{-1}Q_n^{(1)}(\boldsymbol{\zeta}_0) - \mathrm{lower}\big(\boldsymbol{D}_1\big)\boldsymbol{I}_0^{-1}L_n^{(1)}(\boldsymbol{\eta}_0, \boldsymbol{\zeta}_0)\Big] \xrightarrow{d} \mathcal{N}\big(\mathbf{0}, \boldsymbol{V}\big).$$

We have already shown that $\lim_{n\to\infty}\tilde{\boldsymbol{I}}_n = \tilde{\boldsymbol{I}}_0$, which implies that $\lim_{n\to\infty}\tilde{\boldsymbol{I}}_n^{-1}\tilde{\boldsymbol{I}}_0 = \boldsymbol{D}_2$. Similarly, we have shown that $\lim_{n\to\infty}\hat{\boldsymbol{I}}_n = \boldsymbol{I}_0$, which implies $\lim_{n\to\infty}\hat{\boldsymbol{I}}_n^{-1}\boldsymbol{I}_0 = \boldsymbol{D}_1$. Combining (B.25) and (B.26) and applying Slutsky's theorem, we have

$$\big(\tilde{\boldsymbol{I}}_n^{-1}\tilde{\boldsymbol{I}}_0 - \boldsymbol{D}_2\big)\sqrt{n}\frac{1}{n}\tilde{\boldsymbol{I}}^{-1}Q_n^{(1)}(\boldsymbol{\zeta}_0) \xrightarrow{d} \mathbf{0},$$

$$\mathrm{lower}\big[\big(\hat{\boldsymbol{I}}_n^{-1}\boldsymbol{I}_0 - \boldsymbol{D}_1\big)\big]\sqrt{n}\frac{1}{n}\boldsymbol{I}_0^{-1}L_n^{(1)}(\boldsymbol{\eta}_0, \boldsymbol{\zeta}_0) \xrightarrow{d} \mathbf{0}.$$

Since convergence in distribution to a constant implies convergence in probability, we have

$$\big(\tilde{\boldsymbol{I}}_n^{-1}\tilde{\boldsymbol{I}}_0 - \boldsymbol{D}_2\big)\sqrt{n}\frac{1}{n}\tilde{\boldsymbol{I}}^{-1}Q_n^{(1)}(\boldsymbol{\zeta}_0) \xrightarrow{P_{\boldsymbol{\theta}_0}} \mathbf{0},$$

$$\mathrm{lower}\big[\big(\hat{\boldsymbol{I}}_n^{-1}\boldsymbol{I}_0 - \boldsymbol{D}_1\big)\big]\sqrt{n}\frac{1}{n}\boldsymbol{I}_0^{-1}L_n^{(1)}(\boldsymbol{\eta}_0, \boldsymbol{\zeta}_0) \xrightarrow{P_{\boldsymbol{\theta}_0}} \mathbf{0},$$

which implies that

$$\sqrt{n}\big(\tilde{\boldsymbol{\zeta}}_n - \hat{\boldsymbol{\zeta}}_n\big) \xrightarrow{P_{\boldsymbol{\theta}_0}} -\sqrt{n}\frac{1}{n}\Big[\tilde{\boldsymbol{I}}_0^{-1}Q_n^{(1)}(\boldsymbol{\zeta}_0) - \mathrm{lower}\big(\boldsymbol{D}_1\big)\boldsymbol{I}_0^{-1}L_n^{(1)}(\boldsymbol{\eta}_0, \boldsymbol{\zeta}_0)\Big],$$

where the right part has been shown to converge in distribution to $\mathcal{N}(\mathbf{0}, \boldsymbol{V})$. Hence we can conclude that

$$\sqrt{n}\big(\tilde{\boldsymbol{\zeta}}_n - \hat{\boldsymbol{\zeta}}_n\big) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \boldsymbol{V}).$$

□

B.0.6 Proof of Corollary 4.2.4

Applying random variable transformation to (4.9), it can be shown that

$$\lim_{n\to\infty}\int\int |g_n(\boldsymbol{\eta},\boldsymbol{\zeta})|\,\mathrm{d}\boldsymbol{\eta}\mathrm{d}\boldsymbol{\zeta}=0,$$

where

$$\begin{aligned} g_n(\boldsymbol{\eta},\boldsymbol{\zeta}) &= \tau_n\left(\boldsymbol{\eta}|\boldsymbol{\zeta}\right)\kappa_n\left(\boldsymbol{\zeta}\right) \\ &\quad -\phi\left(\boldsymbol{\eta}\middle|\hat{\boldsymbol{\eta}}_n-\left(\boldsymbol{I}_0^{11}\right)^{-1}\boldsymbol{I}_0^{12}\left(\boldsymbol{\zeta}-\hat{\boldsymbol{\zeta}}_n\right),\left(\boldsymbol{I}_0^{11}\right)^{-1}/\sqrt{n}\right)\phi\left(\boldsymbol{\zeta}\middle|\tilde{\boldsymbol{\zeta}}_n,\tilde{\boldsymbol{I}}_0^{-1}/\sqrt{n}\right).\end{aligned}$$

Let $\boldsymbol{z}=\left[\begin{smallmatrix}\boldsymbol{\eta}\\ \boldsymbol{\zeta}\end{smallmatrix}\right]$ and $\boldsymbol{z}_0=\left[\begin{smallmatrix}\boldsymbol{\eta}_0\\ \boldsymbol{\zeta}_0\end{smallmatrix}\right]$. For any neighborhood U of $\boldsymbol{z}_0$, $\exists\delta>0$ such that $B=\mathcal{B}_{\boldsymbol{z}}(\boldsymbol{z}_0,\delta)\in U$. Then it can be shown that

$$\begin{aligned} &\lim_{n\to\infty}\int_U \tau_n\left(\boldsymbol{\eta}|\boldsymbol{\zeta}\right)\kappa_n\left(\boldsymbol{\zeta}\right)\mathrm{d}\boldsymbol{\eta}\mathrm{d}\boldsymbol{\zeta} \\ &=\lim_{n\to\infty}\int_U \phi\left(\boldsymbol{\eta}\middle|\hat{\boldsymbol{\eta}}_n-\left(\boldsymbol{I}_0^{11}\right)^{-1}\boldsymbol{I}_0^{12}\left(\boldsymbol{\zeta}-\hat{\boldsymbol{\zeta}}_n\right),\left(\boldsymbol{I}_0^{11}\right)^{-1}/\sqrt{n}\right)\phi\left(\boldsymbol{\zeta}\middle|\tilde{\boldsymbol{\zeta}}_n,\tilde{\boldsymbol{I}}_0^{-1}/\sqrt{n}\right)\mathrm{d}\boldsymbol{\eta}\mathrm{d}\boldsymbol{\zeta} \\ &\qquad+\lim_{n\to\infty}\int_U |g_n(\boldsymbol{\eta},\boldsymbol{\zeta})|\,\mathrm{d}\boldsymbol{\eta}\mathrm{d}\boldsymbol{\zeta} \\ &=\lim_{n\to\infty}\int_U \phi\left(\boldsymbol{\eta}\middle|\hat{\boldsymbol{\eta}}_n-\left(\boldsymbol{I}_0^{11}\right)^{-1}\boldsymbol{I}_0^{12}\left(\boldsymbol{\zeta}-\hat{\boldsymbol{\zeta}}_n\right),\left(\boldsymbol{I}_0^{11}\right)^{-1}/\sqrt{n}\right)\phi\left(\boldsymbol{\zeta}\middle|\tilde{\boldsymbol{\zeta}}_n,\tilde{\boldsymbol{I}}_0^{-1}/\sqrt{n}\right)\mathrm{d}\boldsymbol{\eta}\mathrm{d}\boldsymbol{\zeta} \\ &\geqslant\lim_{n\to\infty}\int_B \phi\left(\boldsymbol{\eta}\middle|\hat{\boldsymbol{\eta}}_n-\left(\boldsymbol{I}_0^{11}\right)^{-1}\boldsymbol{I}_0^{12}\left(\boldsymbol{\zeta}-\hat{\boldsymbol{\zeta}}_n\right),\left(\boldsymbol{I}_0^{11}\right)^{-1}/\sqrt{n}\right)\phi\left(\boldsymbol{\zeta}\middle|\tilde{\boldsymbol{\zeta}}_n,\tilde{\boldsymbol{I}}_0^{-1}/\sqrt{n}\right)\mathrm{d}\boldsymbol{\eta}\mathrm{d}\boldsymbol{\zeta}\end{aligned}$$

Since condition 5 holds for both $f_1(\boldsymbol{y}|\boldsymbol{\zeta})$ and $f_2(\boldsymbol{y}|\boldsymbol{\eta},\boldsymbol{\zeta})$, we have $\lim_{n\to\infty}\hat{\boldsymbol{\eta}}_n=\boldsymbol{\eta}_0$, $\lim_{n\to\infty}\hat{\boldsymbol{\zeta}}_n=\boldsymbol{\zeta}_0$ and $\lim_{n\to\infty}\tilde{\boldsymbol{\zeta}}_n=\boldsymbol{\zeta}_0$. Hence the above limit goes to 1.

B.0.7 Proof of Theorem 4.2.7

Letting $\boldsymbol{r}_n^*=\sqrt{n}\left(\boldsymbol{\zeta}_n^*-\hat{\boldsymbol{\zeta}}_n\right)$, it can be seen that $\pi_{n,3}^*(\boldsymbol{t})=\pi_{n,2}^*(\boldsymbol{t}|\boldsymbol{r}_n^*)$. Suppose that the conditions for Theorem 4.2.3 hold, then slightly modifying the step 3 and step 4 in the proof of Theorem 4.2.3 we can show that $\exists\delta>0$ such that for $\|\boldsymbol{r}_n^*\|<\sqrt{n}\delta$,

$$\lim_{n\to\infty}\int\left|\pi_{n,3}^*(\boldsymbol{t})-\phi\left(\boldsymbol{t}\middle|-\left(\boldsymbol{I}_0^{11}\right)^{-1}\boldsymbol{I}_0^{12}\boldsymbol{\mu}_n,\left(\boldsymbol{I}_0^{11}\right)^{-1}\right)\right|\mathrm{d}\boldsymbol{t}=0.$$

Define a sequence of events $A_{n,\delta} = \{\boldsymbol{r}_n^* : \|\boldsymbol{r}_n^*\| < \sqrt{n}\delta\}$ and $B_{n,\delta} = \{\boldsymbol{r}_n^* : \|\boldsymbol{r}_n^*\| > \sqrt{n}\delta\}$, then for any $\epsilon > 0$, we have

$$\begin{aligned} & p\left[\int \left|\pi_{n,3}^*(\boldsymbol{t}) - \phi\left(\boldsymbol{t}\middle| -\left(\boldsymbol{I}_0^{11}\right)^{-1}\boldsymbol{I}_0^{12}\boldsymbol{\mu}_n, \left(\boldsymbol{I}_0^{11}\right)^{-1}\right)\right| \mathrm{d}\boldsymbol{t} > \epsilon\right] \\ =& p(A_{n,\delta}) p\left[\int \left|\pi_{n,3}^*(\boldsymbol{t}) - \phi\left(\boldsymbol{t}\middle| -\left(\boldsymbol{I}_0^{11}\right)^{-1}\boldsymbol{I}_0^{12}\boldsymbol{\mu}_n, \left(\boldsymbol{I}_0^{11}\right)^{-1}\right)\right| \mathrm{d}\boldsymbol{t} > \epsilon \middle| A_{n,\delta}\right] \\ & + p(B_{n,\delta}) p\left[\int \left|\pi_{n,3}^*(\boldsymbol{t}) - \phi\left(\boldsymbol{t}\middle| -\left(\boldsymbol{I}_0^{11}\right)^{-1}\boldsymbol{I}_0^{12}\boldsymbol{\mu}_n, \left(\boldsymbol{I}_0^{11}\right)^{-1}\right)\right| \mathrm{d}\boldsymbol{t} > \epsilon \middle| B_{n,\delta}\right] \end{aligned} \tag{B.27}$$

We have already shown that

$$\lim_{n\to\infty} p\left[\int \left|\pi_{n,3}^*(\boldsymbol{t}) - \phi\left(\boldsymbol{t}\middle| -\left(\boldsymbol{I}_0^{11}\right)^{-1}\boldsymbol{I}_0^{12}\boldsymbol{\mu}_n, \left(\boldsymbol{I}_0^{11}\right)^{-1}\right)\right| \mathrm{d}\boldsymbol{t} > \epsilon \middle| A_{n,\delta}\right] = 0.$$

Since $p(A_{n,1}) \leqslant 1$, the first part in (B.27) goes to 0.

Similarly, we also know

$$p\left(\int \left|\pi_{n,3}^*(\boldsymbol{t}) - \phi\left(\boldsymbol{t}\middle| - \left(\boldsymbol{I}_0^{11}\right)^{-1}\boldsymbol{I}_0^{12}\boldsymbol{\mu}_n, \left(\boldsymbol{I}_0^{11}\right)^{-1}\right)\right| d\boldsymbol{t} > \epsilon \middle| A_{n,2}\right) \leqslant 1.$$

From Lemma 4.2.6, we know that $\lim_{n\to\infty} \sqrt{n}\left(\boldsymbol{\zeta}_n^* - \tilde{\boldsymbol{\zeta}}_n\right) = 0$. From Lemma B.0.4, we know that $\sqrt{n}\big(\tilde{\boldsymbol{\zeta}}_n - \hat{\boldsymbol{\zeta}}_n\big) \xrightarrow{d} \mathcal{N}(\boldsymbol{0}, \boldsymbol{V})$. Combining them we can show that $\boldsymbol{r}_n^* \xrightarrow{d} \mathcal{N}(\boldsymbol{0}, \boldsymbol{V})$, hence

$$p(B_{n,\delta}) = p(\|\boldsymbol{r}_n^*\| > \sqrt{n}\delta)$$

also goes to zero. We have shown that the second part in (B.27) goes to 0.

B.0.8 Proof of Lemma 4.2.8

We first provide the following lemma.

Lemma B.0.5. *(David and Nagaraja (1970)) Let Φ be the cdf function of the standard normal distribution, then*

$$\lim_{n\to\infty} \Phi\left(a_n x + b_n\right)^n = e^{-\exp(-x)}, \tag{B.28}$$

where $b_n = \Phi^{-1}\left(1 - \frac{1}{n}\right)$ and $a_n = \frac{1}{n\phi(b_n)}$.

For any j and for fixed μ_j and σ_{jj}, we standardized the x_{ij}'s and introduce $z_i = \frac{x_{ij}-\mu_j}{\sqrt{\sigma_{jj}}}$. Clearly $z_1, \ldots, z_n$ are i.i.d. standard normal random variables. Applying Lemma B.0.5, for any $\delta > 0$,

$$\lim_{n\to\infty} p(\max_{1\leqslant i\leqslant n} z_i < a_n\delta + b_n) = e^{-\exp(-\delta)}.$$

Transforming z_i's back to x_{ij}'s, we get

$$\lim_{n\to\infty} p\left[\max_{1\leqslant i\leqslant n} x_{ij} < \sqrt{\sigma_{jj}}\,(a_n\delta + b_n) + \mu_j\right] = e^{-\exp(-\delta)}.$$

Since $y_{ij} = \mathbb{1}\{x_j > 0\}\lceil x_{ij}\rceil$, it is easily seen that for any $a \geqslant 0$, $\max_{1\leqslant i\leqslant n} x_{ij} < a$ implies $\max_{1\leqslant i\leqslant n} y_{ij} < a + 1$. Hence

$$p\left[\max_{1\leqslant i\leqslant n} y_{ij} < \sqrt{\sigma_{jj}}\,(a_n\delta + b_n) + \mu_j + 1\right] \geqslant p\left[\max_{1\geqslant i\leqslant n} x_{ij} < \sqrt{\sigma_{jj}}\,(a_n\delta + b_n) + \mu_j\right]$$

holds for any n, which implies

$$\lim_{n\to\infty} pr\left[\max_{1\leqslant i\leqslant n} y_{ij} < \sqrt{\sigma_{jj}}\,(a_n\delta + b_n) + \mu_j + 1\right] \geqslant e^{-\exp(-\delta)}.$$

It is easily seen that $b_n \to \infty$. Using Mills ratio, we can show that for any $x > 0$, $\lim_{n\to\infty} \frac{1-\Phi(b_n)}{\phi(b_n)} = \frac{1}{b_n}$. Noting that $1 - \Phi(b_n) = \frac{1}{n}$, we have shown that

$$\lim_{n\to\infty} a_n = \frac{1}{b_n}. \tag{B.29}$$

Integrating by parts, one can easily show the following two bounds:

$$1 - \Phi(x) \leqslant \frac{e^{-x^2/2}}{\sqrt{2\pi x}}, \quad 1 - \Phi(x) \geqslant \frac{e^{-x^2/2}}{\sqrt{2\pi}}\left(\frac{1}{x} - \frac{1}{x^3}\right).$$

Since $1 - \Phi(b_n) = \frac{1}{n}$, it can be shown that $\sqrt{2\log n} \leqslant b_n \leqslant \sqrt{2\log n}$ for sufficiently large n. Coupled with (B.29), we would have $a_n < \frac{2}{\sqrt{\log n}}$. This implies that

$$a_n\delta + b_n < \frac{2\delta}{\sqrt{\log n}} + \sqrt{2\log n},$$

and hence

$$
\begin{aligned}
&pr\left[\max_{1\leqslant i\leqslant n} y_{ij} < \sqrt{\sigma_{jj}}\left(a_n\delta + b_n\right) + \mu_j + 1\right]\\
<&pr\left[\max_{1\leqslant i\leqslant n} y_{ij} < \sqrt{\sigma_{jj}}\left(\frac{2\delta}{\sqrt{\log n}} + \sqrt{2\log n}\right) + \mu_j + 1\right],
\end{aligned}
$$

which completes the proof.

Bibliography

Adams, R. P., Wallach, H. M., and Ghahramani, Z. (2010), "Learning the Structure of Deep Sparse Graphical Models," *Journal of Machine Learning Research: Workshop and Conference Proceedings (AISTATS)*, 9, 1–8.

Allard, W. K., Chen, G., and Maggioni, M. (2012), "Multi-scale geometric methods for data sets II: Geometric multi-resolution analysis," *Applied and Computational Harmonic Analysis*, 32, 435–462.

Arya, S., Mount, D. M., Netanyahu, N. S., Silverman, R., and Wu, A. Y. (1998), "An optimal algorithm for approximate nearest neighbor searching fixed dimensions," *Journal of the ACM (JACM)*, 45, 891–923.

Attias, H. (2000), "A variational baysian framework for graphical models," in *Advances in neural information processing systems*, pp. 209–215.

Belkin, M. and Niyogi, P. (2002), "Laplacian eigenmaps and spectral techniques for embedding and clustering," in *Advances in neural information processing systems*, pp. 585–591.

Bhattacharya, A. and Dunson, D. B. (2011), "Sparse Bayesian infinite factor models," *Biometrika*, 98, 291–306.

Buades, A., Coll, B., and Morel, J.-M. (2005), "A non-local algorithm for image denoising," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 2, pp. 60–65, IEEE.

Canale, A. and Dunson, D. B. (2011), "Bayesian kernel mixtures for counts," *Journal of the American Statistical Association*, 106, 1528–1539.

Canale, A. and Dunson, D. B. (2014), "Multiscale Bernstein polynomials for densities," *arXiv preprint arXiv:1410.0827*.

Carvalho, C. M., Chang, J., Lucas, J. E., Nevins, J. R., Wang, Q., and West, M. (2008), "High-dimensional sparse factor modeling: applications in gene expression genomics," *Journal of the American Statistical Association*, 103, 1438–1456.

Chen, M., Silva, J., Paisley, J., Wang, C., Dunson, D., and Carin, L. (2010), "Compressive sensing on manifolds using a nonparametric mixture of factor analyzers: Algorithm and performance bounds," *IEEE Transactions on Signal Processing*, 58, 6140–6155.

Cox, D. R. and Reid, N. (2004), "A note on pseudolikelihood constructed from marginal densities," *Biometrika*, 91, 729–737.

David, H. A. and Nagaraja, H. N. (1970), *Order statistics*, Wiley Online Library.

Diebolt, J. and Robert, C. P. (1994), "Estimation of finite mixture distributions through Bayesian sampling," *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 56, 363–375.

El-Basyouny, K., Barua, S., and Islam, M. T. (2014), "Investigation of time and weather effects on crash types using full Bayesian multivariate Poisson lognormal models," *Accident Analysis & Prevention*, 73, 91–99.

Escobar, M. D. and West, M. (1995), "Bayesian density estimation and inference using mixtures," *Journal of the American Statistical Association*, 90, 577–588.

Folland, G. (2005), "Higher-order derivatives and Taylor?s formula in several variables," .

Gelman, A. et al. (2006), "Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper)," *Bayesian analysis*, 1, 515–534.

Ghahramani, Z. and Hinton, G. E. (1996), "The EM algorithm for mixtures of factor analyzers," Tech. rep., Technical Report CRG-TR-96-1, University of Toronto.

Ghosh, J. K., Delampady, M., and Samanta, T. (2007), *An introduction to Bayesian analysis: theory and methods*, Springer Science & Business Media.

Hastie, T. and Stuetzle, W. (1989), "Principal curves," *Journal of the American Statistical Association*, 84, 502–516.

Hough, J. B., Krishnapur, M., Peres, Y., et al. (2009), *Zeros of Gaussian analytic functions and determinantal point processes*, vol. 51, American Mathematical Soc.

Ilin, A. and Raiko, T. (2010), "Practical approaches to principal component analysis in the presence of missing values," *The Journal of Machine Learning Research*, 11, 1957–2000.

Jaakkola, T. S. and Jordan, M. I. (2000), "Bayesian parameter estimation via variational methods," *Statistics and Computing*, 10, 25–37.

Johndrow, J. E., Smith, A., Pillai, N., and Dunson, D. B. (2016), "Inefficiency of Data Augmentation for Large Sample Imbalanced Data," *arXiv preprint arXiv:1605.05798*.

Karypis, G. and Kumar, V. (1998), "A fast and high quality multilevel scheme for partitioning irregular graphs," *SIAM Journal on Scientific Computing*, 20, 359–392.

Lawrence, N. D. (2005), "Probabilistic non-linear principal component analysis with Gaussian process latent variable models," *Journal of machine learning research*, 6, 1783–1816.

Lewandowski, D., Kurowicka, D., and Joe, H. (2009), "Generating random correlation matrices based on vines and extended onion method," *Journal of multivariate analysis*, 100, 1989–2001.

Liu, H., Lafferty, J. D., and Wasserman, L. A. (2007), "Sparse nonparametric density estimation in high dimensions using the rodeo," in *International Conference on Artificial Intelligence and Statistics*, pp. 283–290.

Ma, J., Kockelman, K. M., and Damien, P. (2008), "A multivariate Poisson-lognormal regression model for prediction of crash counts by severity, using Bayesian methods," *Accident Analysis & Prevention*, 40, 964–975.

Pauli, F., Racugno, W., and Ventura, L. (2011), "Bayesian composite marginal likelihoods," *Statistica Sinica*, pp. 149–164.

Perona, P. and Malik, J. (1990), "Scale-space and edge detection using anisotropic diffusion," *IEEE Transactions on pattern analysis and machine intelligence*, 12, 629–639.

Rao, V., Adams, R. P., and Dunson, D. D. (2017), "Bayesian inference for Matérn repulsive processes," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79, 877–897.

Rasmussen, C. E. (1999), "The infinite Gaussian mixture model." in *NIPS*, vol. 12, pp. 554–560, MIT; 1998.

Richardson, S. and Green, P. J. (1997), "On Bayesian analysis of mixtures with an unknown number of components (with discussion)," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 59, 731–792.

Rokhlin, V., Szlam, A., and Tygert, M. (2009), "A randomized algorithm for principal component analysis," *SIAM Journal on Matrix Analysis and Applications*, 31, 1100–1124.

Roweis, S. (1998), "EM algorithms for PCA and SPCA," *Advances in neural information processing systems*, pp. 626–632.

Roweis, S. T. and Saul, L. K. (2000), "Nonlinear dimensionality reduction by locally linear embedding," *science*, 290, 2323–2326.

Roweis, S. T., Saul, L. K., and Hinton, G. E. (2002), "Global coordination of local linear models," in *NIPS*, vol. 2, pp. 889–896, MIT; 1998.

Rue, H., Martino, S., and Chopin, N. (2009), "Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations," *Journal of the royal statistical society: Series b (statistical methodology)*, 71, 319–392.

Scott, S. L., Blocker, A. W., Bonassi, F. V., Chipman, H. A., George, E. I., and McCulloch, R. E. (2016), "Bayes and big data: The consensus Monte Carlo algorithm," *International Journal of Management Science and Engineering Management*, 11, 78–88.

Sethuraman, J. (1994), "A constructive definition of Dirichlet priors," *Statistica Sinica*, 4, 639–650.

Shen, W., Tokdar, S. T., and Ghosal, S. (2013), "Adaptive Bayesian multivariate density estimation with Dirichlet mixtures," *Biometrika*, 100, 623–640.

Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., and Su, Z. (2008), "Arnetminer: extraction and mining of academic social networks," in *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 990–998, ACM.

Tenenbaum, J. B., Silva, V. D., and Langford, J. C. (2000), "A global geometric framework for nonlinear dimensionality reduction," *Science*, 290, 2319–2323.

Tipping, M. E. and Bishop, C. M. (1999a), "Mixtures of probabilistic principal component analyzers," *Neural computation*, 11, 443–482.

Tipping, M. E. and Bishop, C. M. (1999b), "Probabilistic principal component analysis," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61, 611–622.

Titsias, M. and Lawrence, N. D. (2010), "Bayesian Gaussian process latent variable model," in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pp. 844–851.

Wang, X. and Dunson, D. B. (2013), "Parallelizing MCMC via Weierstrass sampler," *arXiv preprint arXiv:1312.4605.*

Wang, Y. and Dunson, D. B. (2015), "Probabilistic curve learning: Coulomb repulsion and the electrostatic Gaussian process," in *Advances in Neural Information Processing Systems*, pp. 1738–1746.

Wang, Y., Canale, A., and Dunson, D. B. (2014), "Scalable multiscale density estimation," *arXiv preprint arXiv:1410.7692*.

Wang, Y., Canale, A., and Dunson, D. (2016), "Scalable geometric density estimation," in *Artificial Intelligence and Statistics*, pp. 857–865.

Weinberger, K. Q. and Saul, L. K. (2006), "An introduction to nonlinear dimensionality reduction by maximum variance unfolding," in *AAAI*, vol. 6, pp. 1683–1686.

Biography

Ye Wang was born on Nov 29, 1989 in Dalian, Liaoning, China. He received a B.S. in Mathematics from Harbin Institute of Technology in July 2012, an M.S. in Statical and Economical Modeling from Duke University in 2014, and a Ph.D. in Statistics from Duke University in 2018. He was the winner of the best student award in AISTATS 2016. He also received student travel awards from the Neural Information Processing Systems (NIPS) conference in 2015. His published papers include Wang and Dunson (2015) and Wang et al. (2016).