Bridged Posterior: Optimization, Profile Likelihood and a New Approach of Generalized Bayes

Cheng Zeng [*†]   Eleni Dilma [*‡]   Jason Xu [§]   Leo L Duan [¶‖]

January 18, 2024

Abstract

Optimization techniques such as dual ascent, alternating direction method of multipliers, and majorization-minimization are widely used in high-dimensional applications. These methods are often able to efficiently deliver point estimates on statistically useful constrained spaces, such as those satisfying low rank, low cardinality or combinatorial structure. To quantify uncertainty around point estimates, a popular generalized Bayes solution known as Gibbs posterior exponentiates the negative loss function to form a posterior density. Despite successful theoretical justifications, Gibbs posteriors are supported in a high-dimensional space, and do not fully inherit the computational efficiency or clean constraint formulations from optimization. In this article, we are motivated by the discovery that a large class of penalized profile likelihoods, which partially maximize over a subset of parameters, enjoy equivalence to another generative model for the data. This leads us to explore a new generalized Bayes approach, viewing the likelihood as a function jointly specified based on latent variables that are conditionally deterministic given the data and parameters, linked via solving an optimization sub-problem. This new likelihood can be understood as a special case of the data augmented likelihood where the latent variable is used to model the dependency among the data. Marginally, the latent variable given the data remains stochastic, and is characterized by its posterior distribution. Therefore, this framework coined "bridged posterior" conforms to the Bayesian methodology. Besides providing a new, valid generative model, a positively surprising theoretical finding is that asymptotically under mild conditions, the $\sqrt{n}$-adjusted posterior distribution of the parameter under the bridged posterior converges to the same normal distribution as that of the canonical posterior marginalized over non-deterministic latent variables. That is, our result formally dispels a long-time

---

[*]Joint first authors

[†]Department of Statistics, University of Florida, U.S.A. czeng1@ufl.edu

[‡]Department of Statistics, University of Florida, U.S.A. edilma@ufl.edu

[§]Department of Statistical Science, Duke University, U.S.A. jason.q.xu@duke.edu

[¶]Department of Statistics, University of Florida, U.S.A. li.duan@ufl.edu

[‖]Corresponding author.

belief that partial optimization of latent variable might lead to under-estimation of parameter uncertainty. We demonstrate the practical advantages of our approach in computation and constraint-accommodation for models in both parametric and semi-parametric settings, such as maximum-margin classification, latent normal model, and harmonization of multiple networks.

KEY WORDS: Constrained conditional probability; Data augmentation; Duality; Profile likelihood; Parametric and semi-parametric Bernstein–von Mises.

1 Introduction

The Generalized Bayes approach is becoming increasingly popular due to its potential advantages in model simplicity and robustness. A generalized posterior can be specified based on partial information from the data, or via a loss function that characterizes an inferential summary of the data. There is a well-established literature on partial information settings including methods based on composite likelihood [Lindsay, 1988, Varin et al., 2011], partial likelihood [Sinha et al., 2003, Dunson and Taylor, 2005], pairwise likelihood [Jensen and Künsch, 1994], and others. Recently, there has been a burgeoning interest in loss-based Bayesian models, including works involving classification loss [Polson and Scott, 2011], distance-based losses [Duan and Dunson, 2021, Rigon et al., 2023, Natarajan et al., 2023]. Loss-based generalized Bayes models typically use a probability distribution called the "Gibbs posterior" [Jiang and Tanner, 2008], taking the form:

$$\Pi(\theta \mid y) \propto \exp[-g(\theta, y)],$$

where $g(\theta, y)$ is some loss function with y the data and θ the parameter.

There is a vast generalized Bayes literature in using Gibbs posterior for explicit model weighting, with g chosen according to some utility function such as predictive accuracy [Lavine et al., 2021, Tallman and West, 2023], scoring rule [Gneiting and Raftery, 2007, Dawid and Musio, 2015], fairness metrics [Chakraborty et al., 2024] or summary statistics-based divergence [Frazier and Drovandi, 2021, Frazier et al., 2022]. With connections to these methods, our focus is on the cases when one wants to adopt a certain loss g from the optimization literature for statistical modeling, while needing to quantify uncertainty beyond point estimates.

In optimization, the point estimate $\hat{\theta} = \arg\min_\theta g(\theta, y)$ can often be efficiently computed using an iterative algorithm, even under a wide range of constraints. For example, the convex clustering algorithm [Tan and Witten, 2015, Chi and Lange, 2015] uses $g(\theta, y) = (1/2) \sum_{i=1}^n \|y_i - \theta_i\|_2^2 + \lambda \sum_{(i,j):i<j} \|\theta_i - \theta_j\|_2$ for data $y_i \in \mathbb{R}^p$, location parameter $\theta_i \in \mathbb{R}^p$, and tuning parameter $\lambda > 0$. This can be understood as a relaxation of hierarchical clustering; the penalty term encourages most of the L_2-norms $\|\hat{\theta}_i - \hat{\theta}_j\|_2$ to be zero, encouraging a small number of unique $\hat{\theta}_i$, promoting cluster structure. The estimate

$\hat{\theta}$ can be easily obtained using convex, continuous optimization. A popular combinatorial alternative to clustering makes use of the k-means loss [MacQueen, 1967] $g[(c_1, \ldots, c_n), y] = \sum_{k=1}^{K} \sum_{(i<j):c_i=c_j=k} \|\vec{y}_i - \vec{y}_j\|_2^2/n_k$, with $c_i \in \{1, \ldots, K\}$ where the discrete cluster assignment label $c_i = c_j$ if $\theta_i = \theta_j$. Here note $n_k = \sum_i 1(c_i = k)$ and that $\hat{\theta}_i = \sum_{i:c_i=k} y_i/n_k$. Rigon et al. [2023] use this loss toward forming a Gibbs posterior for quantifying the uncertainty of c_i, which shows model robustness to the distributional asymmetry.

Under the exponential negative transform, the Gibbs posterior distribution concentrates near the posterior mode. This induces variability around the point estimate and in turn, enables uncertainty quantification. How to interpret this uncertainty is not immediately obvious, and one may question the authenticity of inferential procedures such as hypotheses tests or intervals based on this posterior that may not derive from a generative likelihood model. There are several recent works that provide sound justification for the large n regime. First, Gibbs posteriors admit a coherent update scheme for θ toward minimizing the expected loss $\int g(\theta, y)\mathcal{F}(\mathrm{d}y; \theta)$, where $\mathcal{F}$ denotes the true data generating distribution [Bissiri et al., 2016]. Second, if the Gibbs posterior density is proportional to a composite likelihood, such as the conditional density using some insufficient statistics [Lewis et al., 2021] derived from a full likelihood model $F(y; \theta_0)$, then the Gibbs posterior of θ concentrates toward θ_0 and enjoys asymptotic normality under mild conditions [Miller, 2021].

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

The second methodological gap relates to the modeling front: continuity of the Gibbs posterior distribution often yields a mismatch to constraint conditions on $\hat{\theta}$ except on a set of measure zero. To illustrate, consider the Bayesian lasso [Park and Casella, 2008], which can be viewed as the Gibbs posterior under the lasso loss. Though this promotes a

3

sparse estimate $\hat{\theta}$, note that θ is not sparse almost everywhere. Similarly, under a Gibbs posterior approach to support vector machines, the maximum-margin hyperplane has zero posterior measure, and studies from this view focus on improving point estimation [Polson and Scott, 2011]. Beyond this incongruence between the distribution on $\hat{\theta}$ and the samples from $\Pi(\theta \mid y)$, invariance to changes in $g(\theta, y)$ presents another consideration. To obtain an estimate $\hat{\theta}$ residing in a constrained or low dimensional space, it is common practice in optimization to employ an alternative $\tilde{g}(\theta, y)$ that has superior computational properties. For example, $\tilde{g}$ can be convex, unconstrained, or non-combinatorial, under the condition that $\arg\min_\theta \tilde{g}(\theta, y) = \arg\min_\theta g(\theta, y)$ — that is, the two distinct loss functions touch at the minima. This invariance at the optimum is routinely exploited in methods such as variable splitting, Lagrange multiplier approaches, proximal methods, majorization-minimization, and convex relaxations. However, the Gibbs posterior does not enjoy such an invariance, as the distribution $\Pi(\theta \mid y)$ changes whenever g changes.

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

It is natural to ask whether such an approach is consistent with Bayesian methodology in the sense that there exists a valid generative model corresponding to a likelihood that depends only on λ. This article answers this question affirmatively. We begin with a set of profile likelihoods that partially maximize a joint likelihood $L(y; z, \lambda)$ over z, showing that each corresponds to another common likelihood where the data are modeled dependently. We then establish the theoretical result that under mild conditions, the $\sqrt{n}$-adjusted posterior distribution of the parameter under our framework converges asymptotically to the same normal limit as canonical posteriors marginalized over non-deterministic latent variables.

This work is closely related to prior work by Polson and Scott [2016], which discovers a hierarchical duality: the scale mixture of univariate exponential or location-scale mixture of normal is proportional to another (potentially intractable) density maximized over a univariate latent variable. This perspective inspires efficient new algorithms for producing point estimates. Despite some similarities in the univariate setting, our method applies generally to multivariate problems and to settings where the latent variables may exhibit dependence. In other related work, Lee et al. [2005] interpret the profile likelihood as resulting from an empirical prior. A key difference is that our proposed framework $L(y; z, \lambda)$ can lead to a fully Bayesian method, where the latent variable z characterizes the latent

4

dependency among the data.

2 Method

2.1 Augmented Likelihood with Conditional Optimization

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

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

in which we refer to $\lambda \in \mathbb{R}^d$ as parameter, and $z \in \mathbb{R}^p$ as latent variable. Here $\Pi_{\mathcal{L}}$ denotes the marginal latent variable distribution for z. Since z could be associated with a continuous, discrete, or degenerate distribution, we use the integration with respect to a probability measure notation $\int f(z) \mu(\mathrm{d}z)$, in which $\mathrm{d}z$ refers to $z \sim \mu$ is the one that we integrate over. The joint distribution $L(y, z; \lambda)$ is also known as an "augmented likelihood" [Tanner and Wong, 1987]. Examples abound in statistics: for instance, augmented likelihoods are used in characterizing dependence among discrete y via a correlated normal latent variable z [Wolfinger, 1993, Rue et al., 2009], or model-based clustering on grouping data y via a latent discrete label z [Blei et al., 2003, Fraley and Raftery, 2002].

We now consider a special case when given y and λ:

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

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

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

$$\begin{aligned} &z \sim \Pi_{\mathcal{L}}(z; \lambda); \\ &y \mid z, \lambda \sim L(y \in \mathcal{Y}_{\lambda, z} \mid z, \lambda), \text{ where } \mathcal{Y}_{\lambda, z} = \{y : \min_{\zeta} g(\zeta, y; \lambda) = g(z, y; \lambda)\}. \end{aligned} \tag{3}$$

That is, y is generated under the constraint given by z.

Remark 1. To clarify, the latent z may have varying dimension p and $\Pi_{\mathcal{L}}$ according to the sample size n.

For concreteness, we present two illustrative examples based on the profile likelihood. Profile likelihoods have a frequentist origin, motivated by the convenience of testing or

constructing confidence intervals for a parameter of interest λ, in the presence of other parameters ζ, often called "nuisance parameters". There is a long-standing debate on whether using a profile likelihood leads to a coherent Bayes' procedure [Lee et al., 2005, Cheng and Kosorok, 2009, Evans, 2016, Maclaren, 2018]. Using the above, we can now view the profile likelihood as a special case of (2), taking $g(\zeta, y; \lambda) = -\log L(y, \zeta; \lambda)$.

Example 1. Consider linear regression with $y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n \times d}$, $\lambda \in \mathbb{R}^d$, $z > 0$, and some constant $v > 0$:

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

$$z = \arg\min_\zeta [-\log L(y, \zeta; \lambda)].$$

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

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

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

$$L(y; \lambda) \propto \left(1 + \frac{\|y - X\lambda\|^2}{(v + 2)\frac{v}{v+2}}\right)^{-(n+v+2)/2}, \tag{4}$$

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

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

$$L(y, z; \lambda) \propto \exp\left(-\frac{\|y - Cz\|^2}{2\sigma^2}\right) \exp\left(-\frac{1}{2} z^\top G^{-1} z\right);$$

$$z = \arg\min_\zeta \{-\log L(y, \zeta; \lambda)\}.$$

The first line has the same form as a likelihood with z regularized by multivariate normal distribution $\mathrm{N}(0, G)$. Here, minimization yields $z = (C^\top C/\sigma^2 + G^{-1})^{-1} C^\top y/\sigma^2 = [G - GC^\top (I\sigma^2 + CGC^\top)^{-1} CG] C^\top y/\sigma^2$. Therefore, we have:

$$\Pi_{\mathcal{L}}(z; \lambda) \propto \exp\left(-\frac{1}{2} z^\top [GC^\top (I\sigma^2 + CGC^\top)^{-1} CG]^{-1} z\right);$$

$$L(y \mid z, \lambda) \propto \exp\{-\frac{\|y - Cz\|^2}{2\sigma^2}\} \mathbb{1}\{C^\top y - (C^\top C + \sigma^2 G^{-1})z = 0\}.$$

6

In this example, the indicator puts an affine equality constraint on y. On the other hand, the marginal of y is

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

which corresponds to a multivariate normal $N(0, I\sigma^2 + CGC^\top)$.

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

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

2.2 Bridged Posterior Distributions and Posterior Propriety

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

$$\Pi(\lambda \mid y) = \frac{\int L(y, \mathrm{d}z; \lambda)\pi_0(\lambda)}{\int \int L(y, \mathrm{d}z; \lambda)\pi_0(\mathrm{d}\lambda)}, \quad \text{subject to } z = \arg\min_\zeta g(\zeta, y; \lambda). \tag{5}$$

When z is the unique minimizer, we may remove the first integration from both the numerator and denominator, replacing $\mathrm{d}z$ by z. The above distribution can be viewed as obeying an *equality constraint*, which acts as a bridge between a probabilistic model and an optimization problem. Therefore, we refer to (5) as a "bridged posterior". To clarify, the above formulation encompasses the setting where $\lambda = (\lambda_A, \lambda_B)$, where only the first part λ_A influences the minimization of g, and λ_B corresponds to the other parameters.

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

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

$$L(y, z; \lambda) = \exp\{-h(y, \lambda)\} \exp\{-\min_\zeta g(\zeta, y; \lambda)\}, \tag{6}$$

7

where $z = \arg\min_\zeta g(\zeta, y; \lambda)$. In the optimization literature, $\min_\zeta g(\zeta, y; \lambda)$ is often referred to as the *primal problem*, and z as the primal solution. A useful technique is to instead solve an associated dual problem $\sup_\alpha g^\dagger(\alpha, y; \lambda)$, where $\alpha \in \mathbb{R}^q$ is known as the dual variable. For example, the Fenchel dual for convex g is based on the conjugate function $g^\dagger(\alpha, y; \lambda) := \sup_\zeta \{\alpha^\top \zeta - g(\zeta, y; \lambda)\}$, and the Lagrangian dual for α under constraints $\tilde{c}(\alpha) \leq \vec{0}$, with $\tilde{c}(\alpha) \in \mathbb{R}^q$ and where inequality holds pointwise, is $g^\dagger(\alpha, y; \lambda) := \inf_{\zeta \in \mathbb{R}^p} g(\zeta, y; \lambda) + \alpha^\top \tilde{c}(\zeta)$, where the dual variable $\alpha \geq \vec{0}$. In many cases, the dual problem may be easier to solve, and with helpful techniques such as variable splitting (discussed further below in the context of latent normal models), one can often obtain $g^\dagger(\alpha, y; \lambda)$ in a closed form, even when $\sup_\alpha g^\dagger(\alpha, y; \lambda)$ may not have a closed form.

Duality theory provides a simple way to produce a useful bound:

Theorem 1. *For a likelihood taking the form of* (6), *consider* $\inf_\zeta g(\zeta, y; \lambda)$ *as the primal problem, and* $\sup_\alpha g^\dagger(\alpha, y; \lambda)$ *as the dual problem with E the feasible region of α. Assume that weak duality holds: that is,* $\sup_\alpha g^\dagger(\alpha, y; \lambda) \leq \inf_\zeta g(\zeta, y; \lambda)$ *for any λ in the domain, given y. If there exists $\tilde{\alpha} \in E$ such that $\int \exp[-h(y, \lambda)] \exp[-g^\dagger(\tilde{\alpha}, y; \lambda)] \pi_0(d\lambda) < \infty$, then $\int \Pi(d\lambda \mid y) < \infty$.*

Remark 3. The proof follows rather trivially from the fact that $g^\dagger(\tilde{\alpha}, y; \lambda) \leq \sup_\alpha g^\dagger(\alpha, y; \lambda)$. On the other hand, this result leads to a very useful method for checking integrability —we do not have to solve for the optimal dual variable $\hat{\alpha}$ at which $\sup g^\dagger(\alpha, y; \lambda)$ is attained. Instead, we just need to find any $\tilde{\alpha} \in E$ that makes the product integrable. Moreover, the criteria for weak duality are straightforward to check: for Fenchel duals, g needs to be convex, and for Lagrangian duals, g can be convex or non-convex.

We now illustrate the application of the theorem via a working example.

Example 3 (Latent normal model and latent quadratic exponential model)**.** We consider a modification of the canonical latent normal model that takes a full likelihood denoted by $\tilde{L}$:

$$\tilde{L}(y, \zeta; \lambda) \propto \exp\left\{ -\frac{1}{2} \zeta^\top Q^{-1}(\lambda; x) \zeta \right\} \prod_{i=1}^n v(y_i \mid \zeta_i), \tag{7}$$

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

We now minimize $g(\zeta, y; \lambda) = -\log \tilde{L}(y, \zeta; \lambda)$ over $\zeta \in \mathbb{R}^n$ to induce a conditionally

deterministic z, with profile likelihood:

$$L(y, z; \lambda) \propto \exp\left\{-\frac{1}{2}z^\top Q^{-1}(\lambda; x)z\right\}\left\{\prod_{i=1}^{n} v(y_i \mid z_i)\right\},$$

$$z = \arg\min_\zeta\{-\log L(y, \zeta; \lambda)\}.$$

(8)

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

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

where we use $Q = Q(\lambda; x)$ for notational ease, which leads to

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

whenever all $(\alpha_i + y_i) \in (0, 1)$; otherwise the infimum takes $-\infty$. Therefore, we have the dual function:

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

(9)

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

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

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

Due to some similarity of the above form to the quadratic exponential model in [McCullagh, 1994], we refer to (8) as a "latent quadratic exponential model". It is not hard to see that the above is an integrable upper bound for y, since $y_i \in \{0, 1\}$, and $(\alpha_i + y_i)$ and $1 - (\alpha_i + y_i)$ are both bounded from above for any given α. To find an appropriate prior for λ, note that the second part does not involve λ at any fixed α. Hence it suffices to find a prior such that for a feasible $\tilde{\alpha}$:

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

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

9

Remark 4. For the above example, strong duality holds $\sup_\alpha g^\dagger(\alpha, y; \lambda) = \inf_\zeta g(\zeta, y; \lambda)$. Therefore, we can use the dual ascent algorithm to find $\hat\alpha = \arg\max_{\alpha:(\alpha+y)\in(0,1)^n} g^\dagger(\alpha, y; \lambda)$, and then set $z = -Q\hat\alpha$. Note that neither the dual function (9) nor its gradient with respect to α requires the inversion Q^{-1}, an $O(n^3)$ operation, so that optimization can be carried out very efficiently. At the same time, $L(y, z; \lambda)$ can be evaluated quickly since $z^\top Q^{-1} z = \hat\alpha^\top Q\hat\alpha$. In comparison, the latent normal model would involve matrix inversion and decomposition for sampling latent ζ. We defer the numerical experiments to Section 5.

2.3 Predictive Distribution

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

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

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

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

suggesting that we can simulate y_{n+j} sequentially for $j = 1\ldots k$. When y_{n+j} is in a low-dimensional (often one-dimensional) space, we can employ a simple algorithm such as rejection sampling.

Note that *all* elements in $z = \hat{z}(y_{1:(n+k)}, \lambda) \in \mathbb{R}^p$ may vary according to $y_{1:(n+k)}$; we emphasize this by using the notation $\hat{z}(y_{1:(n+k)}, \lambda)$.

One should carefully distinguish from those conditionally independent latent variables, such as predicting y_{n+1} based on $L(y_{n+1}, z_{n+1} \mid y_{1:n}, z_{1:n}, \lambda)$ when $z_{1:n}$ were conditionally independent from y_{n+1}. Here, we do not have conditional independence due to the optimization-based equality constraint. This implies that when there is no closed-form solution for z, there is an additional burden to compute $\hat{z}(y_{1:(n+j)}, \lambda)$ wherever j increments to $j + 1$. Fortunately, for the optimization problem

$$\hat{z}(y_{1:(n+j+1)}, \lambda) = \arg\min_\zeta g(\zeta, y_{1:(n+j+1)}; \lambda),$$

we can initialize ζ at the last optimal when predicting y_{n+j}, $\hat{z}(y_{1:(n+j)}, \lambda)$, then it usually takes a few iterations of optimization steps to converge to $\hat{z}(y_{1:(n+j+1)}, \lambda)$. For advanced problems, there is a large literature on online optimization algorithms [Bubeck, 2011] that one could efficiently obtain sequential updates.

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

Example 4 (Bayesian Maximum Margin Classifier for Partially Labeled Data). Consider the following likelihood that extends the support vector machine [Cortes and Vapnik, 1995], for n labeled data $(x_i, y_i) \in \mathbb{R}^{\tilde{p}} \times \{-1, 1\}$ and k unlabeled predictor $x_j \in \mathbb{R}^{\tilde{p}}$:

$$L\big[\{y_i\}_{i=1}^n, z = (z_w, z_b); \lambda, \{x_i\}_{i=1}^{n+k}\big] \propto \sum_{\{y_{n+j}\}_{j=1}^k \in \{-1,1\}^k} \exp\Big\{ -\frac{1}{2}\lambda\|z_w\|_2^2 - \sum_{i=1}^{n+k} h(z, y_i; \lambda, x_i)\Big\},$$

$$\text{subject to } z = \arg\min_\zeta \frac{1}{2}\lambda\|\zeta_w\|_2^2 + \sum_{i=1}^{n+k} h\{(\zeta_w, \zeta_b), y_i; \lambda, x_i\},$$

$$h(\zeta, y_i; \lambda, x_i) = \max\Big\{1 - y_i(\zeta_w^\top x_i + \zeta_b), 0\Big\},$$

$$(10)$$

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

For comparison, if we use a Gibbs posterior with likelihood in the form of (10), except without the equality constraint in the second line and with z replaced by ζ, then we can see the following:

$$P(y_i \mid \zeta, \lambda, x_i) \propto \exp\{-h(\zeta, y_i; \lambda, x_i)\}$$

independently for $i = n+1, \ldots, n+k$. Therefore, the distribution based on the Gibbs posterior has $\tilde{L}\big[\{y_i\}_{i=1}^n, \zeta; \lambda, \{x_i\}_{i=1}^{n+k}\big] = \tilde{L}\big[\{y_i\}_{i=1}^n, \zeta; \lambda, \{x_i\}_{i=1}^n\big]$ via marginalization, ending up not using any information from the observed $(x_{n+1}, \ldots, x_{n+k})$.
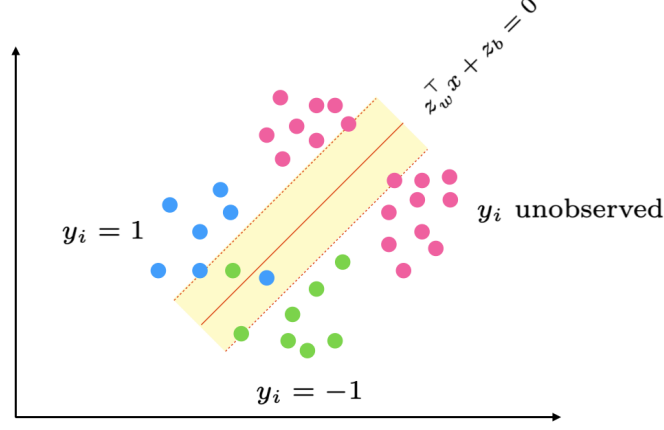
Figure 1: Intuition on how the Bayesian maximum margin classifier (a bridged posterior based on conditional minimization of the hinge loss) incorporates information from both the labeled (y_i observed) and unlabeled data (y_i unobserved). The posterior puts a high probability on a decision boundary with a small misclassification error among observed data (blue and green points), while trying to avoid having the decision band $\{x : -1 < \zeta_w^\top x + \zeta_b < 1\}$ cover the unlabeled predictors x_j (magenta points).

For a bridged posterior with a conditionally optimal $z = \hat{z}(\{y_i\}_{i=1}^{n+k}, \lambda; \{x_i\}_{i=1}^{n+k})$, although we do not have a closed-form marginal form in (10), we know from the Lagrange dual method with multiplier $\alpha \in \mathbb{R}^{n+k}$ [Chang and Lin, 2011], the decision hyperplane $\mathcal{C}_z = \{x : z_w^\top x + z_b = 0\}$ satisfies

$$z_w = \sum_{i=1}^{n+k} (\alpha_i y_i) x_i,$$

where $0 \le \alpha_i \le \lambda^{-1}$, and there are only few $\alpha_i \ne 0$, referred to as the "support vectors", for which $(\zeta_w^\top x_i + \zeta_b) y_i \le 1$. Therefore, we can see that regardless of the values of $\{y_i\}_{i=n+1}^{n+k}$, the decision boundary can be influenced by the unlabeled predictor $\{x_i\}_{i=n+1}^{n+k}$. Intuitively, the bridged posterior assigns higher probability to a hyperplane $\mathcal{C}_z$ that has a small misclassification error among observed data, while avoiding unlabeled predictors x_i in the band $\{-1 < \zeta_w^\top x + \zeta_b < 1\}$. We use Figure 1 to illustrate the intuition.

3   Posterior Computation

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

3.1 Metropolis-Hastings with Conditional Optimization

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

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

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

- Draw proposal $\lambda^* \sim G(\cdot; \lambda)$

- Run optimization subroutine to find $z^* = \arg\min_\zeta g(\zeta, y; \lambda^*)$.

- Set $\lambda^{(t+1)} \leftarrow \lambda^*$, $z^{(t+1)} \leftarrow z^*$ with probability:

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

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

In this article, for algorithmic simplicity, we take λ as unconstrained in $\mathbb{R}^d$ under appropriate reparametrization (such as the softplus transform for positive scalars $\tilde{\lambda}_1 = \log[1+\exp(\lambda_1)] > 0$). We use $G(\cdot; \lambda^{(t)})$ as $\text{Uniform}(\lambda^{(t)} - s, \lambda^{(t)} + s)$, where $s \in \mathbb{R}_{\geq 0}^d$ is a tuning parameter that controls the step size in each dimension. When running MCMC for each of the examples presented, we make use of an adaptation period to tune s so that the empirical acceptance rate is close to 0.23, after which we fix s and collect Markov chain samples. This exhibits excellent mixing performance empirically.

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

Instead of uniform random walk proposals G, more informative proposals such as the Metropolis-adjusted Langevin algorithm (MALA) or Hamiltonian Monte Carlo (HMC) may yield better performance. These algorithms become especially advantageous compared to random walk methods in terms of mixing as the dimension d increases.

In this subsection, we first show that gradients (or sub-gradients) are readily available in cases when a profile likelihood is used, and then we discuss its use in the MALA algorithm. Under the bridged posterior, the lack of closed forms for z presents a potential challenge to these methods, leading to intractable gradients or subgradients with respect to z. However,

for those based on the profile likelihood, this issue can be bypassed entirely. Consider the posterior deriving from (6),

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

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

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

This is due to the first-order optimality condition $\partial g(\zeta, y; \lambda)/\partial\zeta = 0$ at $\zeta = z$. By applying the chain rule, we see that $[(\partial g(\zeta, y; \lambda)/\partial\zeta)\big|_{\zeta=z}][(\partial z/\partial\zeta)] = 0$.

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

We use $\tilde{\nabla} \log \Pi(\lambda \mid y)$ to denote a subgradient evaluated at point λ. For reversibility, in the case when there is more than one subgradient at λ, we impose a constraint that $\tilde{\nabla} \log \Pi(\lambda \mid y)$ chooses one of the subgradients in a pre-determined way. This constraint is implicitly satisfied in most computing software, for example, most packages will output $\tilde{\nabla}|\lambda_1|_1 = 0$ when $\lambda_1 = 0$, even though any value $[-1, 1]$ is a subgradient. We now describe the MALA algorithm with preconditioning.

- Draw proposal $\lambda^* \sim \mathrm{N}\big[\cdot; \lambda^{(t)} + \tau M \tilde{\nabla} \log \Pi(\lambda^{(t)} \mid y), 2\tau M\big]$.

- Run optimization algorithm to find $z^* = \arg\min_\zeta g(\zeta, y; \lambda^*)$.

- Set $\lambda^{(t+1)} \leftarrow \lambda^*$, $z^{(t+1)} \leftarrow z^*$ with probability:

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

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

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

4 Asymptotic Theory

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

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

In a bridged posterior, since we make z to be conditionally determined given (y, λ), it may appear intuitive to believe that the posterior of λ would characterize a lower amount of uncertainty (such as having smaller marginal variances) compared to its integrated posterior counterpart. Surprisingly, we find this belief to be false in the asymptotic region — the bridged posterior of λ enjoys the same BvM results with covariance H_0^{-1}.

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

In the following, we first focus on the BvM result for general bridged posterior which may or may not based on a profile likelihood. Due to the broad distribution family that we consider, we rely on relatively strong conditions including differentiability of likelihood and parametric setting, in order to establish BvM results. Next, we relax the differentiability requirement and extend our scope to the semi-parametric setting. Due to the more challenging conditions under the semi-parametric setting, for a tractable analysis, we will restrict our focus to the sub-class of bridged posteriors involving a profile likelihood.

In both settings, we consider λ in the parameter space $\Theta \subset \mathbb{R}^d$ and that there is a fixed ground-truth λ_0, and the prior density $\pi_0(\lambda)$ to be continuous at λ_0 with $\pi_0(\lambda_0) > 0$. We use $\| \cdot \|$ as the Euclidean–Frobenius norm, and $B_r(\lambda_0) = \{\lambda \in \mathbb{R}^d : \|\lambda - \lambda_0\| < r\}$ as a ball of radius r.

4.1 General Bridged Posterior under Parametric Setting

For a real-valued function $\alpha(x)$ defined on $\mathbb{R}^d$, we denote first, second and third derivatives by $\alpha'(x) \in \mathbb{R}^d$, $\alpha''(x) \in \mathbb{R}^{d \times d}$ and $\alpha'''(x) \in \mathbb{R}^{d \times d \times d}$, respectively. For vector-valued function

15

$\alpha(x) = \{\alpha_1(x), \ldots, \alpha_m(x)\}$, we again use notations $\alpha'(x)$, $\alpha''(x)$ and $\alpha'''(x)$ to denote the derivatives, except that they are tensors of one larger order. We say a sequence of functions α_n uniformly bounded on E if the set $\{\|\alpha_n(x)\| : x \in E, n \in \mathbb{N}\}$ is bounded. We use $\xrightarrow[n\to\infty]{a.s.[y_{1:n}]}$ for almost sure convergence, $\xrightarrow[n\to\infty]{P_{\lambda_0,\zeta_0}}$ for convergence in probability.

To ease the notation, we define

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

where the CDLV $\hat{z}_n(\lambda) := \arg\min_\zeta g_n(\zeta, y_{1:n}; \lambda)$. Let E be an open and bounded subset of Θ such that $\lambda_0 \in E$. We now state the BvM result on the general bridged posterior for parametric models where $\zeta \in \mathbb{R}^p$ has a fixed and finite dimension.

Theorem 2. *Suppose the following conditions hold: (A1) l_n has continuous third derivatives on $E \times \hat{z}_n(E)$, $\hat{z}_n$ has continuous third derivatives on E, l_n''' is uniformly bounded on $E \times \hat{z}_n(E)$, and $\hat{z}_n'''$ is uniformly bounded on E, a.s.$[y_{1:n}]$; (A2) a.s.$[y_{1:n}]$, $\hat{z}_n \to \hat{z}_*$ on Θ for some function $\hat{z}_*$, $l_n \to l_*$ for some function l_*; (A3) l_* has positive definite $l_*''\{\lambda_0, \hat{z}_*(\lambda_0)\}$ and satisfies $\frac{\partial l_*(\lambda_0, \zeta)}{\partial \zeta}|_{\zeta = \hat{z}_*(\lambda_0)} = 0$; (A4) and for some compact $K \subseteq E$ with λ_0 in the interior of K,*

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

and

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

Then there is a sequence $\lambda_n \to \lambda_0$ such that $\hat{l}_n'(\lambda_n) = 0$ for all n large enough, $\hat{l}_n(\lambda_n) \to \hat{l}_*(\lambda_0)$ where $\hat{l}_*(\lambda) = l_*\{\lambda, \hat{z}_*(\lambda)\}$. Further, letting q_n be the density of $\sqrt{n}(\lambda - \lambda_n)$ when $\lambda \sim \Pi_n(\lambda \mid y)$, and $\mathcal{N}$ the normal density, we have the total variational distance

$$d_{TV}\{q_n, \mathcal{N}(0, H_0^{-1})\} \xrightarrow[n\to\infty]{a.s.[y_{1:n}]} 0 \text{ with } H_0 = \hat{l}_*''(\lambda_0).$$

Remark 5. Conditions (A1-A2) are often imposed to enable a second-order Taylor expansion around λ_n [Miller, 2021]; (A3) focuses on the cases when $\lambda = \lambda_0$ and gives the local second-order optimal condition of $l_*(\lambda_0, \zeta)$ at $\zeta = \hat{z}_*(\lambda_0)$, where $\hat{z}_*(\lambda_0)$ is produced as the minimizer of another loss function g; (A4) ensures the dominance of l_* at $(\lambda_0, \hat{z}_*(\lambda_0))$ over all possible (λ, ζ) in the described neighborhood, including those points with $\lambda \neq \lambda_0$.

As mentioned above, the result above shows that fixing ζ to z does not impact the asymptotic variance of λ. On the other hand, since z is finite-dimensional and differentiable on E, we can use delta method to find out the asymptotic covariance of z. For bridged posterior using profile likelihood, we do find lower uncertainty in $\Pi(z \mid y)$ under a bridged posterior compared to $\Pi(\zeta \mid y)$ under an integrated one, as formalized below.

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

4.2 Bridged Posterior using Profile Likelihood under Semi-parametric Setting

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

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

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

To facilitate analysis under these challenges, we use the "approximately least-favorable submodel" technique. We refer the readers to the book of Kosorok [2008] for the detailed explanation. For this section to be self-contained, we review the important definitions that are involved as the building blocks for establishing BvM results.

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

$$l_n\{t, \tilde{\zeta}_t(\lambda, \zeta)\} \text{ is twice differentiable in } t \in \Theta, \qquad \tilde{\zeta}_{t=\lambda}(\lambda, \zeta) = \zeta. \tag{11}$$

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

Efficient score and Fisher information Conventionally, the λ-score function of the full likelihood is $\dot{l}_n(\lambda, \zeta) = \frac{\partial l_n(\lambda, \zeta)}{\partial \lambda}$. Consider a direction $\delta \in \widetilde{\mathcal{H}}$ (another Hilbert space) such that a path $\{\zeta_\gamma^\delta \in \mathcal{H}\}_{\gamma \in \mathbb{R}^d}$ with $\zeta_\gamma^\delta \to \zeta_0$ as $\gamma \to \lambda_0$. We can now define the generalized ζ-score function at $\zeta = \zeta_0$ in the direction of δ by $A_{\lambda_0,\zeta_0}^n \delta := \left. \dfrac{\partial l_n(\lambda_0, \zeta_\gamma^\delta)}{\partial \gamma}\right|_{\gamma=\lambda_0}$, where $A_{\lambda_0,\zeta_0}^n : \widetilde{\mathcal{H}} \mapsto L_2^d(P_{\lambda_0,\zeta_0})$ is a map, and $L_2^d(P_{\lambda_0,\zeta_0})$ is the space of d-dimensional vector-valued functions $[\alpha_1(y), \ldots, \alpha_d(y)]$ where each $\alpha_i(y)$ is L_2-integrable on $y \sim P_{\lambda_0,\zeta_0}$.

The "efficient score function" for λ at (λ_0, ζ_0) is defined by

$$\widetilde{\mathscr{Q}l}_n(\lambda_0, \zeta_0) := \dot{l}_n(\lambda_0, \zeta_0) - \mathscr{P}\dot{l}_n(\lambda_0, \zeta_0),$$

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

The "efficient Fisher information" at (λ_0, ζ_0) is defined as

$$\tilde{I}_0 := \mathbb{E}_{\lambda_0,\zeta_0}\{\widetilde{\mathscr{Q}l}_n(\lambda_0, \zeta_0)\widetilde{\mathscr{Q}l}_n(\lambda_0, \zeta_0)^\top\}.$$

Equivalently, $\mathscr{P}\dot{l}_n(\lambda_0, \zeta_0)$ is the projection of the score function for λ_0 onto the closed linear space spanned by the set $\{A_{\lambda_0,\zeta_0}^n \delta\}_{\delta \in \widetilde{\mathcal{H}}}$. In the appendix, we provide an illustration of the above via the Cox regression model.

Least favorable model To connect the two topics above, if (11) further satisfies

$$\left.\frac{\partial \tilde{l}_n(t, \lambda_0, \zeta_0)}{\partial t}\right|_{t=\lambda_0} = \mathscr{2i}_n(\lambda_0, \zeta_0),$$

then we have the submodel $\tilde{l}_n(t, \lambda, \zeta)$ "least favorable" at $t = \lambda_0$: since among all the submodels $\tilde{l}_n(t, \lambda_0, \zeta_0)$, this submodel has the smallest Fisher information on each dimension of $t \in \Theta \subset \mathbb{R}^d$ by the definition of $\mathscr{2i}_n(\lambda_0, \zeta_0)$. Since our goal focus on the asymptotic region, we only need the least favorable model condition to hold in a limiting sense. This leads to the "approximately least favorable model".

With these ingredients, we are ready to derive our results. We first show the locally asymptotically normal (LAN) for the profile $\hat{l}_n(\lambda)$.

Lemma 1. *Suppose that (A1) there exists a neighborhood $V \subset \Theta \times \Theta \times \mathcal{H}$ containing $(\lambda_0, \lambda_0, \zeta_0)$ such that*

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

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

where the P_{λ_0,ζ_0} and $\mathbb{E}_{\lambda_0,\zeta_0}$ are defined with respect to the ground-truth distribution of $y_{1:n}$. Further, (A2) $\hat{z}_n(\lambda)$ converges to ζ_0 when $\lambda \to \lambda_0$ and $n \to \infty$, and (A3) there exists a neighborhood $U \subset \Theta$ containing λ_0 such that

$$\left.\mathbb{E}_{\lambda_0,\zeta_0} \frac{\partial \tilde{l}_n\{t, \lambda, \hat{z}_n(\lambda)\}}{\partial t}\right|_{t=\lambda_0} = o_{P_{\lambda_0,\zeta_0}}(1)\left(\|\lambda - \lambda_0\| + n^{-1/2}\right) \tag{12}$$

holds for all $\lambda \in U$. Here, $o_{P_{\lambda_0,\zeta_0}}(1)$ refers to a term that converges to 0 in P_{λ_0,ζ_0} as $n \to \infty$. Then there exists a neighborhood $B_\epsilon(\lambda_0)$ for some $\epsilon > 0$ such that

$$\hat{l}_n(\lambda) - \hat{l}_n(\lambda_0) = (\lambda - \lambda_0)^\top h_n - \frac{1}{2}(\lambda - \lambda_0)^\top H_0(\lambda - \lambda_0) + o_{P_{\lambda_0,\zeta_0}}(1)\left\{\left(\|\lambda - \lambda_0\| + n^{-1/2}\right)^2\right\} \tag{13}$$

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

Remark 6. A1 is the approximately least favorable submodel condition. A similar result for independent and identically distributed (iid) data given (ζ_0, λ_0) has been previously shown in Murphy and Van der Vaart [2000, Theorem 1]. However, a difference here is that our result is general and holds regardless of if $(y_{1:n} \mid \zeta_0, \lambda_0)$ are iid or non-iid, hence may be of independent interest outside the context of establishing BvM result.

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

Theorem 3. *Assume* (13) *holds with positive definite* H_0. *Suppose that (A1) the maximum likelihood estimator* $\hat{\lambda}_n$ *exists and converges to* λ_0 *when* $n \to \infty$; *(A2) for any* $\epsilon > 0$, *there exists* $\delta > 0$ *such that*

$$P_{\lambda_0,\zeta_0}\left[\inf_{\|\lambda - \hat{\lambda}_n\| \geq \epsilon}\left\{\hat{l}_n(\hat{\lambda}_n) - \hat{l}_n(\lambda)\right\} \geq \delta\right] \xrightarrow[n \to \infty]{} 1. \tag{14}$$

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

$$\int_{B_\epsilon(\lambda_0)} \pi_n(\lambda)\,\mathrm{d}\lambda \xrightarrow[n \to \infty]{P_{\lambda_0,\zeta_0}} 1 \text{ for all } \epsilon > 0, \tag{15}$$

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

Remark 7. In (A2), we assume the profile likelihood outside a neighborhood of the MLE to be strictly less than the maximum profile likelihood. Specifically, there is a gap such that the posterior distribution will put negligible mass outside that neighborhood.

4.3 Comparison with Existing BvM Results on Semi-parametric Models

There is a rich theory literature on BvM results on semi-parametric models, so it is interesting to compare our results with them and clarify our contribution.

Overall, the existing results can be divided into two categories. The first category is similar to our focused setting, where they obtain the posterior under a profile likelihood. Lee et al. [2005] showed that $\mathbb{E}_{\lambda \sim \Pi(\lambda|y)}g\{\sqrt{n}(\lambda - \lambda_n)\}$ converges to $\mathbb{E}_{u \sim \mathcal{N}(0,\tilde{I}_0^{-1})}g(u)$ in probability assuming a Taylor expansion form for iid data. Their condition is similar to (13), and on the other hand, they do not give the result of the posterior density converging to normal density in total variation. Cheng and Kosorok [2008] showed a BvM result for the posterior induced from profile likelihood for iid probability model, under the assumption that the third derivatives exist. Compared to their result, ours is general in the sense that it is applicable to non-iid data and under potential non-differentiability.

The second category of BvM results relate to canonical Bayesian methodology involving integrated posterior $\Pi(\lambda \mid y) = \int \Pi(\lambda, \mathrm{d}\zeta \mid y)\,\zeta$ over a non-deterministic ζ. Bickel and Kleijn [2012] proved a BvM result for marginal posterior distribution of λ using the LAN property for the marginal likelihood of λ, which has similar form with (13). On the other hand, they additionally assume that the marginal posterior probability of λ inside the neighborhood $B_{M_n/\sqrt{n}}(\lambda_0)$ converges to 1 for every $M_n \to \infty$. This condition is similar to but arguably stronger than (14). Castillo and Rousseau [2015] proved a BvM result on a functional of the parameters, under an essentially necessary no-bias condition which is related to both the

likelihood and the prior specification of (λ, ζ). To be informative to the interested readers, we provide a detailed summary of their conditions in the supplementary material.

Before we conclude this section, it is of particular interest to make the connection between our approach (based on partial optimization) and the second category (based on marginalizing). It is very encouraging to see the asymptotic equivalence of the two $\sqrt{n}$-adjusted posterior distributions of λ. Potentially, it allows one to eliminate the computational and interpretation problem created by high-dimensional ζ, by instead solving an optimization sub-problem and using canonical MCMC on the remaining part for uncertainty estimation.

5 Numerical Experiments
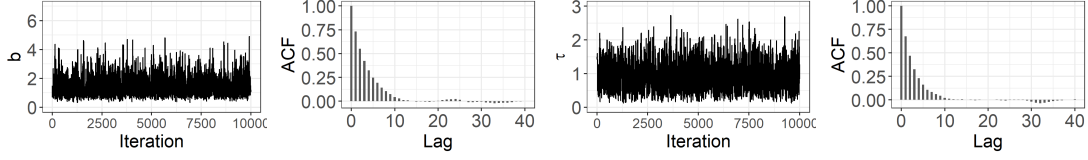
5.1 Latent Quadratic Exponential Model

We conduct simulation experiments comparing a latent normal model and a latent quadratic exponential model (based on Example 3). To simulate data for benchmarking, we generate random locations $x_1, \ldots, x_{1000} \sim \text{Uniform}(-6, 6)$, and ground-truth mean from a latent curve $\tilde{z}_i = \cos(x_i)$. At each x_i, we generate a binary observation $y_i \sim \text{Bernoulli}(1/[1 + \exp\{-\tilde{z}_i\}])$.

We fit the latent quadratic exponential model (8) and the latent normal model in (7) to the simulated data. For both models, we assign half-normal $\text{N}_+(0, 1)$ prior on τ and Inverse-Gamma$(2, 5)$ prior on b. We use random walk Metropolis for the latent quadratic exponential model, and data augmentation Gibbs sampler for the latent normal model (detail provided in the supplementary materials).

We run each of MCMC algorithms for $10,000$ iterations and discard the first $2,000$ as burn-ins. The latent quadratic exponential model takes about 8.37 minutes, and the latent normal model takes about 11.87 minutes on a 12-core laptop. Figure 2 compares the mixing of MCMC algorithms for those two models. Clearly, the latent quadratic exponential model enjoys much better mixing while taking less time. We compute the effective sample size (ESS) per time unit (10 seconds wall time). The ESSs for parameter (b, τ) in the latent quadratic exponential model are 9.99% and 12.91%, and the ones in the latent normal model are and 0.79% and 0.09%.
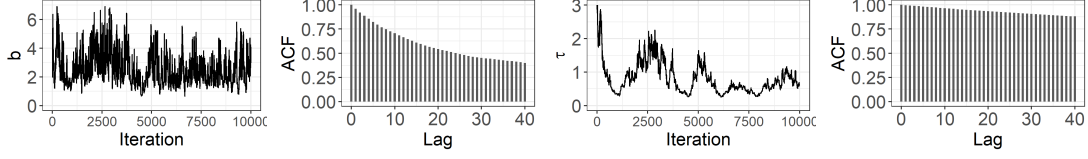
Traceplot of b Autocorrelation of b Traceplot of τ Autocorrelation of τ

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

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

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

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

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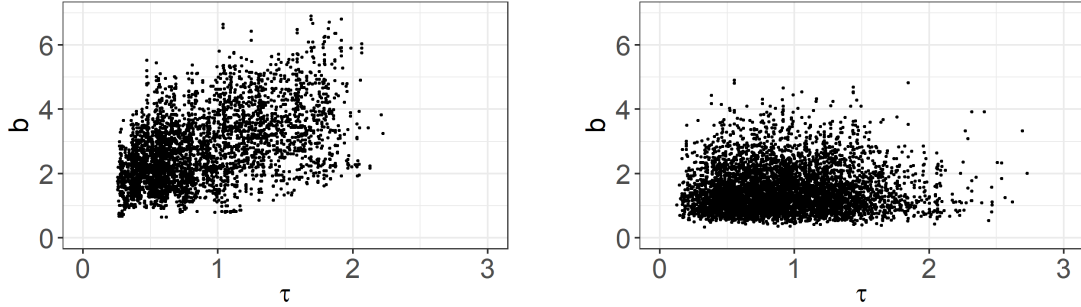
(a) Posterior distribution of (b, τ) from latent normal model.

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

Figure 3: The estimated posterior distributions of the covariance kernel parameters from the latent normal model and the latent quadratic exponential model.

5.2 Bayesian Maximum Margin Classifier

To illustrate the strength of the Bayesian maximum margin classifier (based on Example 4), in terms of uncertainty quantification and borrowing information from unlabeled data, we apply it in a data application of heart failure-related death prediction. The dataset consists of 299 total patients who had a previous occurrence of heart failure, the outcome is whether each patient died during a follow-up care period between April and December 2015, at the Faisalabad Institute of Cardiology and at the Allied Hospital in Faisalabad (Punjab, Pakistan). There are 194 men and 105 women between the ages of 40 and 95 years old. For each patient, we have 12 clinical features, and the death/surviving event as y_i.
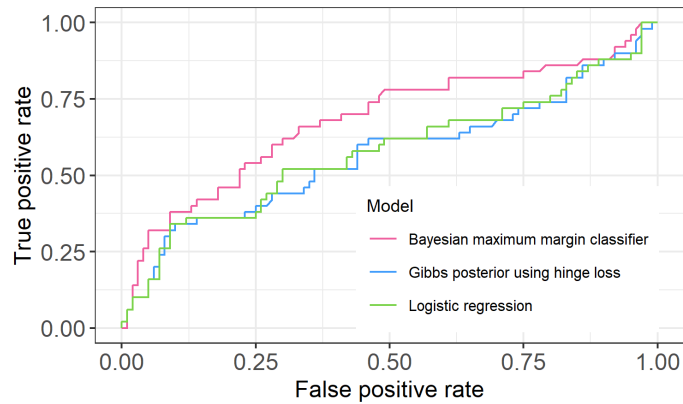

Figure 4: The prediction receiver operating characteristic curves from the three models.

We mask the outcomes of randomly chosen 97 men and 52 women (corresponding to close to 50% missing labels), and fit the data to (i) Bayesian maximum margin classifier model, (ii) Gibbs posterior model using hinge loss, (iii) logistic regression. We use prior $\lambda \sim \text{Gamma}(3, 2)$ for models (i) and (ii), $\zeta_w \sim \text{N}(0, 3^2 I)$ and $\zeta_b \sim \text{N}(0, 3^2)$ for models (ii) and (iii). For each model, we run the MCMC algorithm for $1,500$ iterations and discard the first 500 as burn-in. In each iteration, we make a binary prediction on each unlabeled x_j, take the average over the iteration as a posterior estimate for prediction $P(y_j = 1 \mid y_{1:n})$ for $j = n + 1, \ldots, n + k$. Comparison between each prediction probability with true y_j produces the prediction receiver operating characteristic curves. For binary point estimate, we threshold the probability at 0.5, and compute the classification accuracy. Figure 4 shows that the difference between the Gibbs posterior using hinge loss and logistic regression is barely noticeable. on the other hand, the Bayesian maximum margin classifier is clearly more advantageous in producing higher area under the curve (AUC) as well as higher classification accuracy, as listed in Table 1.

To show that the performance gain is largely due to borrowing information from the unlabeled part, we fit a support vector machine on the labeled part only, and predict on the unlabeled part. The classification accuracy falls to the similar level as in the other two Bayesian models when thresholding probability at 0.5.

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

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

In addition to the ROC curves, Figure 5 shows the other uncertainty estimates, in how the posterior prediction $P(y_j = 1)$ changes in the distance of x_j to the posterior mean of decision boundary hyperplane $\{x : z_w^\top x + z_b = 0\}$, and how the posterior sample of the decision boundary varies around the posterior mean in terms of angle between z_w and $\bar{z}_w$.
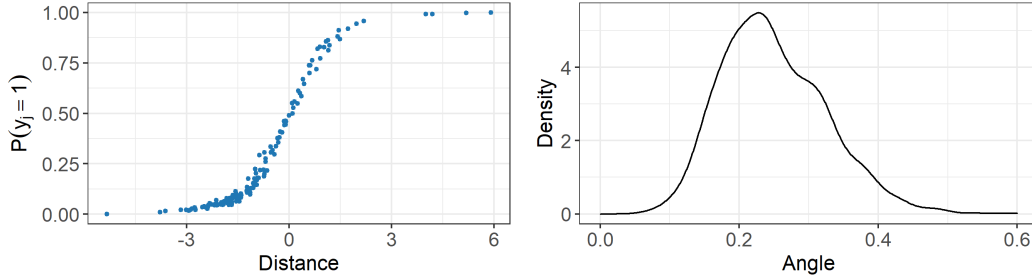
(a) Posterior prediction $P(y_j = 1)$ versus (b) Posterior distribution of the absolute an-
distance to the posterior mean of decision gle (in radians) between z_w and the posterior
boundary hyperplane. mean $\bar{z}_w$.

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

6 Data Application: Harmonization of Functional Connectivity Graphs

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

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

Figure 6, Panel (a) plots the pairwise distances between the observed $\mathcal{L}^{(s)}$'s, using three boxplots: the one among subjects within the diseased group, the one within the

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

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

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

Note that treating $Z^{(s)}$ here as a latent variable is especially appealing — if we viewed it as a parameter in $\mathbb{R}^{n\times n}$, it would involve a high modeling and computational cost. The question boils down to a meaningful choice of $\tilde{\lambda}_s$. Due to the heterogeneity of $\mathcal{L}^{(s)}$, the same value of $\tilde{\lambda}_s = \tilde{\lambda}^{(s')}$ may yield quite different $Z^{(s)}$ and $Z^{(s')}$. As a result, for the purpose of data harmonization, instead of assigning an independent prior or equal value on $\tilde{\lambda}_s$, we assign a dependent likelihood based on the pairwise distances among $Z^{(s)}$:

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

The second line is a pairwise kernel via the average of total squared geodesic distances of each $Z^{(s)}$ with respect to other $Z^{(s')}$, so that it borrows information across subjects to reduce the heterogeneity. It is important to clarify that we do not use the group information in the above model, hence it can serve as a data harmonization tool, even in the absence of group labels.
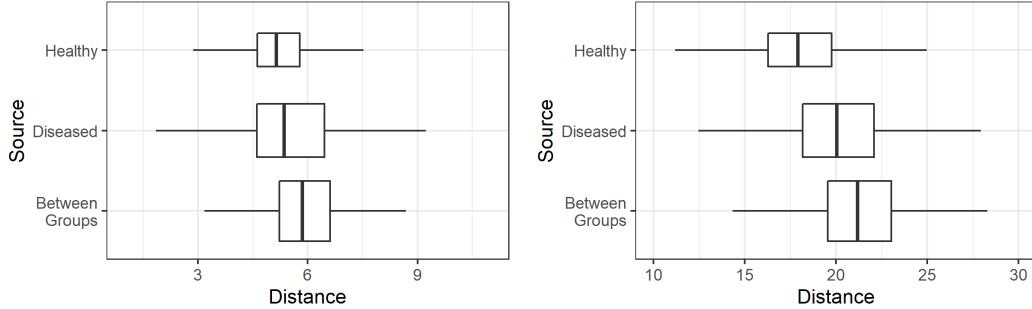
(a) Boxplot of pairwise distances of the observed Laplacian matrices.

(b) Boxplot of pairwise distances of the low-rank smoothed Laplacian matrices using different $\tilde{\lambda}_s$ for each subject.
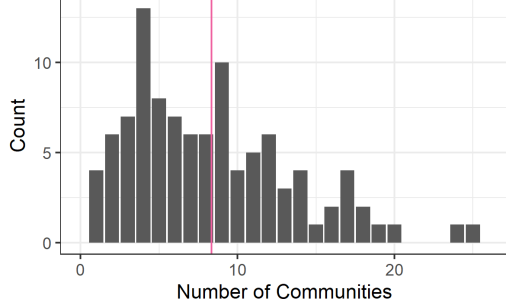
Figure 6: Boxplots of the pairwise distances among the observed Laplacian matrices, and that among the smoothed Laplacian matrices.

To calculate $Z^{(s)}$ at different $\tilde{\lambda}_s$, we use the alternating direction method of multipliers (ADMM) algorithm (see Appendix E). To facilitate computation, for each $\tilde{\lambda}_s$, we assign a discrete uniform equally spread over 10 values in $(0, 5]$, so that we could precompute all the possible values of $Z^{(s)}$, as well as their associated pairwise geodesic distances before running MCMC. For the priors for other parameters, we use inverse-gamma$(2, 1)$ for σ^2 and inverse-gamma$(2, 1)$ for τ. We run MCMC for $10,000$ iterations and discard the first $2,000$ as burn-in. The computation takes about 46.5 minutes on a 12-core laptop.

Using the smoothed Laplacian $Z^{(s)}$, we calculate posterior mean of the the distance matrix $\{\text{dist}(Z^{(s)}, Z^{(s')})\}_{\text{all } s, s'}$ and show re-calculated boxplots of geodesic distances in Figure 6 Panel (b). Clearly, between the low-rank smoothed $Z^{(s)}$, the healthy group now has much lower pairwise distances than the diseased group; and the diseased group has slightly lower pairwise distances compared to the between-group. We compute the Kolmogorov–Smirnov (KS) statistical metric between the empirical distribution of geodesic distances. When we switch from using raw $\mathcal{L}^{(s)}$ to smoothed $Z^{(s)}$, the KS metric between the diseased and healthy increases from 0.143 to 0.299.
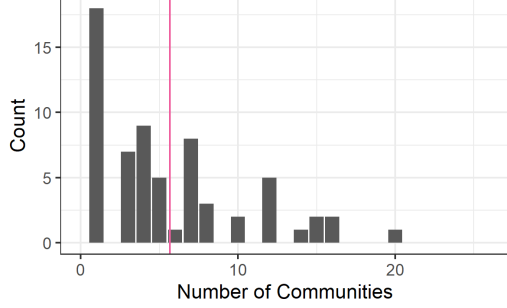
(a) Healthy group (64 subjects).　　　(b) Diseased group (102 subjects).

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

Healthy subject 1.　　　Healthy subject 2.

Top-left 64×64: healthy subjects, bottom-right 102×102: diseased subjects.

(b) Posterior mean of the pairwise geodesic distance among the smoothed Laplacians.

Diseased subject 1.　　　Diseased subject 2.
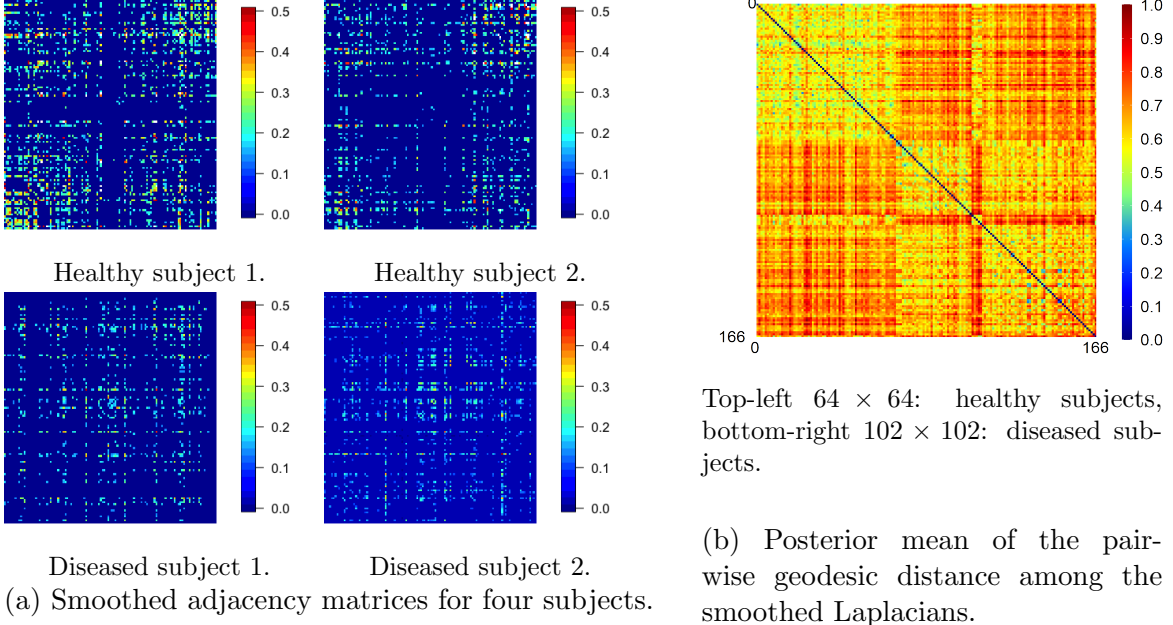(a) Smoothed adjacency matrices for four subjects.

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

By calculating the number of zero eigenvalues of the $Z^{(s)}$, we find $K^{(s)}$ as the number of communities for each subject. Figure 7 shows histograms of $K^{(s)}$ evaluated each subject's

27

posterior mean $\tilde{\lambda}^s$. The average number of communities for the healthy subjects is 5.77 while it is 8.41 for the diseased subjects. This is consistent with the known fact that a diseased brain tends to be more fragmented than a healthy one, due to the disruptions caused by Alzheimer's disease.

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

7 Discussion

In this article, we use optimization as a modeling tool to form a class of augmented likelihoods. These likelihoods enjoy generative interpretation, with the use of latent variable z and constrained distribution of y given z, hence can be used in the Bayesian framework for uncertainty quantification. We demonstrate both computational and modeling advantages over the related Gibbs posterior alternatives.

In our presented examples, we focus on well-behaved loss functions whose optimal solution is unique and can be obtained efficiently with high numerical accuracy. On the other hand, there can be complications that merit future study as extension/generalization of bridged posterior. First, when a loss function has multiple local optima, the converging point z of the algorithm depends on the choice of initialization $\tilde{z}$. One possibility is to assign a probability distribution over $\tilde{z}$, and treat the optimization algorithm as some mapping that produces another distribution for z. Second, often some popular optimization algorithms (such as stochastic gradient descent) may only produce the approximately optimal solutions, hence $\Pi(z \mid y, \lambda)$ may not be a point mass but carry some uncertainty due to numerical errors or stopping criterion. It would be interesting to explore the intersection between bridged posterior and Bayesian probabilistic numerical methodology [Cockayne et al., 2019]. Third, as its name implies, the bridged posterior methodology aims to bridge the gap between the optimization and Bayesian literature. This task is becoming especially important under the popularization of large models (those under heavy parameterization, such as deep neural networks), and the vital need for uncertainty quantification driven by the goal of statistical inference. For those models, rather than viewing all the parameters within as random, it is more practical to conditionally optimize a subset of parameters, while treating the rest as random — this solution may be justified within the bridged posterior framework.

References

Alexandre Belloni and Victor Chernozhukov. On the computational complexity of MCMC-based estimators in large samples. *The Annals of Statistics*, 37(4):2011–2055, 2009. ISSN 00905364, 21688966. URL http://www.jstor.org/stable/30243694.

James O Berger, Brunero Liseo, and Robert L Wolpert. Integrated likelihood methods for eliminating nuisance parameters. *Statistical Science*, pages 1–22, 1999.

P. J. Bickel and B. J. K. Kleijn. The semiparametric Bernstein–von Mises theorem. *The Annals of Statistics*, 40(1):206 – 237, 2012.

Pier Giovanni Bissiri, Chris C Holmes, and Stephen G Walker. A general framework for updating belief distributions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5):1103–1130, 2016.

David M Blei, Andrew Y Ng, and Michael I Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

Sébastien Bubeck. Introduction to online optimization. *Lecture notes*, 2:1–86, 2011.

Ismaël Castillo and Judith Rousseau. A Bernstein–von Mises theorem for smooth functionals in semiparametric models. *The Annals of Statistics*, 43(6):2353 – 2383, 2015.

Abhisek Chakraborty, Anirban Bhattacharya, and Debdeep Pati. A Gibbs posterior framework for fair clustering. *Entropy*, 26(1):63, 2024.

Chih-Chung Chang and Chih-Jen Lin. LIBSVM: a library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, 2(3):1–27, 2011.

O. Chapelle, B. Scholkopf, and A. Zien. *Semi-supervised learning*. Adaptive Computation and Machine Learning series. MIT Press, 2010. ISBN 9780262514125. URL https://books.google.com/books?id=TtWMEAAAQBAJ.

Guang Cheng and Michael R. Kosorok. Higher order semiparametric frequentist inference with the profile sampler. *The Annals of Statistics*, 36(4):1786–1818, 2008.

Guang Cheng and Michael R Kosorok. The penalized profile sampler. *Journal of Multivariate Analysis*, 100(3):345–362, 2009.

Eric C Chi and Kenneth Lange. Splitting methods for convex clustering. *Journal of Computational and Graphical Statistics*, 24(4):994–1013, 2015.

Jon Cockayne, Chris J Oates, Timothy John Sullivan, and Mark Girolami. Bayesian probabilistic numerical methods. *SIAM review*, 61(4):756–789, 2019.

Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20: 273–297, 1995.

A Philip Dawid and Monica Musio. Bayesian model selection based on proper scoring rules. *Bayesian analysis*, 10(2):479–499, 2015.

Leo L Duan and David B Dunson. Bayesian distance clustering. *The Journal of Machine Learning Research*, 22(1):10228–10254, 2021.

David B Dunson and Jack A Taylor. Approximate Bayesian inference for quantiles. *Journal of Nonparametric Statistics*, 17(3):385–400, 2005.

Michael Evans. Measuring statistical evidence using relative belief. *Computational and Structural Biotechnology Journal*, 14:91–96, 2016.

Chris Fraley and Adrian E Raftery. Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, 97(458):611–631, 2002.

David T Frazier and Christopher Drovandi. Robust approximate Bayesian inference with synthetic likelihood. *Journal of Computational and Graphical Statistics*, 30(4):958–976, 2021.

David T Frazier, David J Nott, Christopher Drovandi, and Robert Kohn. Bayesian inference using synthetic likelihood: asymptotics and adjustments. *Journal of the American Statistical Association*, pages 1–12, 2022.

Satyajit Ghosh, Kshitij Khare, and George Michailidis. Strong selection consistency of Bayesian vector autoregressive models based on a pseudo-likelihood approach. *The Annals of Statistics*, 49(3):1267–1299, 2021.

Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

Jens Ledet Jensen and Hans R Künsch. On asymptotic normality of pseudo likelihood estimates for pairwise interaction processes. *Annals of the Institute of Statistical Mathematics*, 46(3):475–486, 1994.

Wenxin Jiang and Martin A Tanner. Gibbs posterior for variable selection in high-dimensional classification and data mining. *The Annals of Statistics*, 36(5):2207–2231, 2008.

James E Johndrow, Aaron Smith, Natesh Pillai, and David B Dunson. MCMC for imbalanced categorical data. *Journal of the American Statistical Association*, 114(527): 1394–1403, 2019.

Kshitij Khare, Sang-Yun Oh, and Bala Rajaratnam. A convex pseudolikelihood framework for high dimensional partial correlation estimation with convergence guarantees. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, pages 803–825, 2015.

Youngseok Kim and Chao Gao. Bayesian model selection with graph structured sparsity. *The Journal of Machine Learning Research*, 21(1):4394–4454, 2020.

Michael R Kosorok. *Introduction to empirical processes and semiparametric inference*, volume 61. Springer, 2008.

Isaac Lavine, Michael Lindon, and Mike West. Adaptive variable selection for sequential prediction in multivariate dynamic models. *Bayesian Analysis*, 16(4):1059–1083, 2021.

Bee Leng Lee, Michael R Kosorok, and Jason P Fine. The profile sampler. *Journal of the American Statistical Association*, 100(471):960–969, 2005.

John R Lewis, Steven N MacEachern, and Yoonkyung Lee. Bayesian restricted likelihood methods: Conditioning on insufficient statistics in Bayesian regression (with discussion). *Bayesian Analysis*, 16(4):1393–1462, 2021.

Lek-Heng Lim, Rodolphe Sepulchre, and Ke Ye. Geometric distance between positive definite matrices of different dimensions. *IEEE Transactions on Information Theory*, 65(9): 5401–5405, 2019.

Bruce G Lindsay. Composite likelihood methods. *Contemporary mathematics*, 80(1):221–239, 1988.

Oliver J Maclaren. Is profile likelihood a true likelihood? an argument in favor. *arXiv Preprint arXiv:1801.04369*, 2018.

James MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297. Oakland, CA, USA, 1967.

Peter McCullagh. Exponential mixtures and quadratic exponential families. *Biometrika*, 81(4):721–729, 1994.

Jeffrey W Miller. Asymptotic normality, concentration, and coverage of generalized posteriors. *Journal of Machine Learning Research*, 22:168–1, 2021.

Susan A Murphy and Aad W Van der Vaart. On profile likelihood. *Journal of the American Statistical Association*, 95(450):449–465, 2000.

Abhinav Natarajan, Maria De Iorio, Andreas Heinecke, Emanuel Mayer, and Simon Glenn. Cohesion and repulsion in Bayesian distance clustering. *Journal of the American Statistical Association*, pages 1–11, 2023.

Trevor Park and George Casella. The Bayesian Lasso. *Journal of the American Statistical Association*, 103(482):681–686, 2008.

Nicholas G Polson and James G Scott. Mixtures, envelopes and hierarchical duality. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(4):701–727, 2016.

Nicholas G Polson and Steven L Scott. Data augmentation for support vector machines. *Bayesian Analysis*, 6(1):1–23, 2011.

Nicholas G Polson, James G Scott, and Jesse Windle. Bayesian inference for logistic models using Pólya–Gamma latent variables. *Journal of the American Statistical Association*, 108(504):1339–1349, 2013.

Tommaso Rigon, Amy H Herring, and David B Dunson. A generalized Bayes framework for probabilistic clustering. *Biometrika*, page asad004, 01 2023.

Gareth O Roberts and Jeffrey S Rosenthal. Optimal scaling for various Metropolis-Hastings algorithms. *Statistical Science*, 16(4):351–367, 2001.

R Tyrrell Rockafellar and Roger J-B Wets. *Variational analysis*, volume 317. Springer Science & Business Media, 2009.

Håvard Rue, Sara Martino, and Nicolas Chopin. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(2):319–392, 2009.

Thomas A Severini. Integrated likelihood functions for non-Bayesian inference. *Biometrika*, 94(3):529–542, 2007.

Debajyoti Sinha, Joseph G Ibrahim, and Ming-Hui Chen. A Bayesian justification of Cox's partial likelihood. *Biometrika*, 90(3):629–641, 2003.

Nicholas Syring and Ryan Martin. Robust and rate-optimal Gibbs posterior inference on the boundary of a noisy image. *The Annals of Statistics*, 48(3):1498 – 1513, 2020. doi: 10.1214/19-AOS1856. URL https://doi.org/10.1214/19-AOS1856.

Nicholas Syring and Ryan Martin. Gibbs posterior concentration rates under sub-exponential type losses. *Bernoulli*, 29(2):1080–1108, 2023.

Emily Tallman and Mike West. Bayesian predictive decision synthesis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, page qkad109, 10 2023. ISSN 1369-7412. doi: 10.1093/jrsssb/qkad109. URL https://doi.org/10.1093/jrsssb/qkad109.

Kean Ming Tan and Daniela Witten. Statistical properties of convex clustering. *Electronic Journal of Statistics*, 9(2):2324, 2015.

Rong Tang and Yun Yang. On the computational complexity of Metropolis-adjusted Langevin algorithms for Bayesian posterior sampling. *arXiv preprint arXiv:2206.06491*, 2022.

Martin A Tanner and Wing Hung Wong. The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, 82(398):528–540, 1987.

Cristiano Varin, Nancy Reid, and David Firth. An overview of composite likelihood methods. *Statistica Sinica*, pages 5–42, 2011.

Steven Winter, Omar Melikechi, and David B Dunson. Sequential Gibbs posteriors with applications to principal component analysis. *arXiv preprint arXiv:2310.12882*, 2023.

Russ Wolfinger. Covariance structure selection in general mixed models. *Communications in Statistics-Simulation and Computation*, 22(4):1079–1106, 1993.

Jun Yang, Gareth O Roberts, and Jeffrey S Rosenthal. Optimal scaling of Metropolis algorithms on general target distributions. *arXiv preprint arXiv:1904.12157*, 3, 2019.

A Proofs

The proof of Theorem 2 uses a theorem from Miller [2021]. We provide the complete statement of that theorem as following.

Theorem 4. *Miller [2021, Theorem 5] Let $\Theta \subseteq \mathbb{R}^D$. Let $E \subseteq \Theta$ be open (in $\mathbb{R}^D$) and bounded. Fix $\theta_0 \in E$ and let $\pi : \Theta \to \mathbb{R}$ be a probability density with respect to Lebesgue measure such that π is continuous at θ_0 and $\pi(\theta_0) > 0$. Let $f_n : \Theta \to \mathbb{R}$ have continuous third derivatives on E. Suppose $f_n \to f$ pointwise for some $f : \Theta \to \mathbb{R}, f''(\theta_0)$ is positive definite, and (f_n''') is uniformly bounded on E. If either of the following two assumptions is satisfied:*

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

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

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

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

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

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

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

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

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

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

$$\hat{l}''_*(\lambda_0) = \left.\frac{\partial^2 l_*(\lambda_0, \zeta)}{\partial \lambda^2}\right|_{\zeta=\hat{z}_*(\lambda_0)} + 2\left.\frac{\partial^2 l_*(\lambda_0, \zeta)}{\partial \lambda \partial \zeta}\right|_{\zeta=\hat{z}_*(\lambda_0)} \hat{z}'_*(\lambda_0)$$

$$+ [\hat{z}'_*(\lambda_0)]^\top \left.\frac{\partial^2 l_*(\lambda_0, \zeta)}{\partial \zeta^2}\right|_{\zeta=\hat{z}_*(\lambda_0)} \hat{z}'_*(\lambda_0) + \left.\frac{\partial l_*(\lambda_0, \zeta)}{\partial \zeta}\right|_{\zeta=\hat{z}_*(\lambda_0)} \hat{z}''_*(\lambda_0)$$

$$= \begin{bmatrix} I_d & [\hat{z}'_*(\lambda_0)]^\top \end{bmatrix} l''_*(\lambda_0, \hat{z}_*(\lambda_0)) \begin{bmatrix} I_d \\ \hat{z}'_*(\lambda_0) \end{bmatrix}$$

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

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

34

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

$$\hat{l}'''_n(\lambda) = \frac{\partial^3 l_n(\lambda,\zeta)}{\partial\lambda^3}\bigg|_{\zeta=\hat{z}_n(\lambda)} + 3\frac{\partial^3 l_n(\lambda,\zeta)}{\partial\lambda^2\partial\zeta}\bigg|_{\zeta=\hat{z}_n(\lambda)} \hat{z}'_n(\lambda)$$

$$+ 3[\hat{z}'_n(\lambda)]^\top \frac{\partial^3 l_n(\lambda,\zeta)}{\partial\lambda\partial\zeta^2}\bigg|_{\zeta=\hat{z}_n(\lambda)} \hat{z}'_n(\lambda) + [\hat{z}'_n(\lambda)]^\top \frac{\partial^3 l_n(\lambda,\zeta)}{\partial\zeta^3}\bigg|_{\zeta=\hat{z}_n(\lambda)} [\hat{z}'_n(\lambda)]^2$$

$$+ 3[\hat{z}'_n(\lambda)]^\top \frac{\partial^2 l_n(\lambda,\zeta)}{\partial\zeta^2}\bigg|_{\zeta=\hat{z}_n(\lambda)} \hat{z}''_n(\lambda) + 3\frac{\partial^2 l_n(\lambda,\zeta)}{\partial\lambda\partial\zeta}\bigg|_{\zeta=\hat{z}_n(\lambda)} \hat{z}''_n(\lambda) + \frac{\partial l_n(\lambda,\zeta)}{\partial\zeta}\bigg|_{\zeta=\hat{z}_n(\lambda)} \hat{z}'''_n(\lambda).$$

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

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

Proof. **of Corollary 1** Now we consider the posterior distribution of CDLV $\hat{z}_n(\lambda)$. Since the transformation $\hat{z}_n$ is not between two Euclidean spaces with equal dimensions, it is hard to calculate the density function of $\hat{z}_n(\lambda)$ when $\lambda \sim \Pi_n(\lambda \mid y)$. However, we can use the delta method to find the asymptotic distribution of $\sqrt{n}$-adjusted $\hat{z}_n(\lambda)$. Since convergence in total variation implies convergence in distribution (weak convergence), the random vector $\sqrt{n}(\lambda - \lambda_n) \rightsquigarrow N(0, H_0^{-1})$. Using delta method, we can prove $\sqrt{n}\{\hat{z}_n(\lambda) - \hat{z}_n(\lambda_n)\} \rightsquigarrow N(0, H_z^{-1})$ where $H_z^{-1} = \hat{z}'_*(\lambda_0)H_0^{-1}\hat{z}'_*(\lambda_0)^\top$.

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

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

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

Hence

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

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

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

35

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

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

If we treat both λ and ζ as interest parameters in the likelihood $L(y, \zeta; \lambda)$ with some joint prior, then Miller [2021] proves $\sqrt{n}([\lambda\ \zeta]^\top - [\lambda_n\ \zeta_n]^\top) \rightsquigarrow N(0, \tilde{H}_0^{-1})$ for some sequences λ_n and ζ_n when $(\lambda, \zeta) \sim \Pi(\lambda, \zeta \mid y)$, where $\tilde{H}_0 = l''_*(\lambda_0, \zeta_0)$ and ζ_0 is a fixed ground-truth of ζ. Marginally, the asymptotic variance of $\sqrt{n}(\zeta - \zeta_n)$ is the ζ-block of $\tilde{H}_0^{-1}$, which is equal to $(l_{*0,\zeta\zeta} - l_{*0,\zeta\lambda}l_{*0,\lambda\lambda}^{-1}l_{*0,\lambda\zeta})^{-1} = l_{*0,\zeta\zeta}^{-1} + l_{*0,\zeta\zeta}^{-1}l_{*0,\zeta\lambda}(l_{*0,\lambda\lambda} - l_{*0,\lambda\zeta}l_{*0,\zeta\zeta}^{-1}l_{*0,\zeta\lambda})^{-1}l_{*0,\lambda\zeta}l_{*0,\zeta\zeta}^{-1}$. Hence, the asymptotic variance of the j-th element of $\sqrt{n}(\zeta - \zeta_n)$ is strictly greater than the one of the j-th element of $\sqrt{n}\{\hat{z}_n(\lambda) - \hat{z}_n(\lambda_n)\}$. $\qquad\square$

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

$$
\begin{aligned}
\hat{l}_n(\lambda) - \hat{l}_n(\lambda_0) &= l_n(\lambda, \hat{z}_\lambda) - l_n(\lambda_0, \hat{z}_{\lambda_0}) \\
&\geq l_n\{\lambda, \tilde{\zeta}_\lambda(\lambda_0, \hat{z}_{\lambda_0})\} - l_n\{\lambda_0, \tilde{\zeta}_{\lambda_0}(\lambda_0, \hat{z}_{\lambda_0})\} \\
&= \tilde{l}_n(\lambda, \lambda_0, \hat{z}_{\lambda_0}) - \tilde{l}_n(\lambda_0, \lambda_0, \hat{z}_{\lambda_0})
\end{aligned}
\tag{16}
$$

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

$$
\begin{aligned}
\hat{l}_n(\lambda) - \hat{l}_n(\lambda_0) &= l_n(\lambda, \hat{z}_\lambda) - l_n(\lambda_0, \hat{z}_{\lambda_0}) \\
&\leq l_n\{(\lambda, \tilde{\zeta}_\lambda(\lambda, \hat{z}_\lambda)\} - l_n\{\lambda_0, \tilde{\zeta}_{\lambda_0}(\lambda, \hat{z}_\lambda)\} \\
&= \tilde{l}_n(\lambda, \lambda, \hat{z}_\lambda) - \tilde{l}_n(\lambda_0, \lambda, \hat{z}_\lambda)
\end{aligned}
\tag{17}
$$

by using $\tilde{\zeta}_\lambda(\lambda, \zeta) = \zeta$ for the first term. Both (16) and (17) are differences between function l_n evaluating at λ and the one at λ_0 while keeping the other arguments unchanged. They have the Taylor expansion of the function $\tilde{l}_n$ with respect to the first argument,

$$
(\lambda - \lambda_0)^\top \frac{\partial \tilde{l}_n(t, \psi, \hat{z}_\psi)}{\partial t}\bigg|_{t=\lambda_0} + \frac{1}{2}(\lambda - \lambda_0)^\top \frac{\partial^2 \tilde{l}_n(t, \psi, \hat{z}_\psi)}{\partial t^2}\bigg|_{t=\tilde{t}}(\lambda - \lambda_0)
\tag{18}
$$

where $\tilde{t}$ is somewhere between λ and λ_0, and ψ can be λ or λ_0. By the assumption 1, the second term is equal to $-(\lambda - \lambda_0)^\top H_0(\lambda - \lambda_0)/2 + o_{P_{\lambda_0,\varsigma_0}}(1)(\|\lambda - \lambda_0\|^2)$. By the assumption 2, the first term is equal to

$$
(\lambda - \lambda_0)^\top h_n + (\lambda - \lambda_0)^\top \mathbb{E}_{\lambda_0,\varsigma_0} \frac{\partial \tilde{l}_n(t, \lambda, \hat{z}_\lambda)}{\partial t}\bigg|_{t=\lambda_0} + o_{P_{\lambda_0,\varsigma_0}}(1)\big(\|\lambda - \lambda_0\|n^{-1/2}\big).
$$

Combining with (12) and $\|\lambda - \lambda_0\|n^{-1/2} \leq \big(\|\lambda - \lambda_0\| + n^{-1/2}\big)^2$, the first term of (18) becomes $(\lambda - \lambda_0)^\top h_n + o_{P_{\lambda_0,\varsigma_0}}(1)\big\{\big(\|\lambda - \lambda_0\| + n^{-1/2}\big)^2\big\}$, and hence (13) is proved. $\qquad\square$

36

Proof. **of Theorem 3** We have $\pi_n(\lambda) = \exp\{n\hat{l}_n(\lambda)\}\pi_0(\lambda)/m_n$ where $m_n = \int_{\mathbb{R}^d} \exp\{n\hat{l}_n(\lambda)\}\pi_0(\lambda)\,\mathrm{d}\lambda$, and $q_n(x) = \pi_n(\lambda_n + x/\sqrt{n})n^{-d/2}$. Let

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

and define $g_0(x) = \exp\{-x^\top H_0 x/2\}\pi_0(\lambda_0)$. We first show that $\int_{\mathbb{R}^d} |g_n(x) - g_0(x)|\,\mathrm{d}x \xrightarrow[n\to\infty]{P_{\lambda_0,\varsigma_0}} 0$. Denote the ϵ chosen from Lemma 1 as ϵ_0. Since π_0 is continuous at λ_0, we choose sufficiently small $\epsilon \in (0, \epsilon_0/2)$ such that $\pi_0(\lambda) \leq 2\pi_0(\lambda_0)$ for all $\lambda \in B_{2\epsilon}(\lambda_0)$. Let δ be the number from (14).

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

$$\hat{l}_n(\lambda) - \hat{l}_n(\lambda_n) = -\frac{1}{2}(\lambda - \lambda_n)^\top H_0(\lambda - \lambda_n) + o_{P_{\lambda_0,\varsigma_0}}(1)\{(\|\lambda - \lambda_n\| + n^{-1/2})^2\}.$$

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

$$n\{\hat{l}_n(\lambda_n + x/\sqrt{n}) - \hat{l}_n(\lambda_n)\} = -\frac{1}{2}x^\top H_0 x + o_{P_{\lambda_0,\varsigma_0}}(1)(\|x\| + 1)^2.$$

Combining with π_0 is continuous at λ_0 and $\lambda_n + x/\sqrt{n} \to \lambda_0$, we have $g_n(x) \to g_0(x)$ pointwise with probability converge to 1. Consider $n > 1/\epsilon^2$ sufficiently large such that the term $o_{P_{\lambda_0,\varsigma_0}}(1) < \alpha/4$ where α is less than the smallest eigenvalue of H_0. We denote $A_0 = H_0 - \alpha I$, and define

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

When $\|x\| < \epsilon\sqrt{n}$, for n large enough we have $\|(\lambda_n + x/\sqrt{n}) - \lambda_0\| < \|\lambda_n - \lambda_0\| + \epsilon < 2\epsilon$. By the choice of ϵ, we have $\pi_0(\lambda_n + x/\sqrt{n}) \leq 2\pi_0(\lambda_0)$. Since $(\|x\| + 1)^2 \leq 2\|x\|^2 + 2$, we have $o_{P_{\lambda_0,\varsigma_0}}(1)(\|x\| + 1)^2 \leq \alpha(\|x\|^2 + 1)/2 = \alpha x^\top x/2 + \alpha/2$. Hence $g_n(x) \leq h_n(x)$ with probability converge to 1 for n sufficiently large, combining with (14) when $\|x\| \geq \epsilon\sqrt{n}$.

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

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

The second term is less than $\int_{\mathbb{R}^d} e^{-n\delta/2}\pi_0(\lambda_n + x/\sqrt{n})\,\mathrm{d}x = e^{-n\delta/2}\int_{\mathbb{R}^d}\pi_0(\lambda)n^{d/2}\,\mathrm{d}\lambda = e^{-n\delta/2}n^{d/2} \to 0$, while the first term monotonically converges to $\int_{\mathbb{R}^d} h_0(x)\,\mathrm{d}x$. Since g_n, g_0, h_n, h_0 are integrable, by the generalized dominated convergence theorem (the version for convergence in probability), we have $\int_{\mathbb{R}^d} |g_n(x) - g_0(x)|\,\mathrm{d}x \xrightarrow[n\to\infty]{P_{\lambda_0,\varsigma_0}} 0$ and $\int_{\mathbb{R}^d} g_n(x)\,\mathrm{d}x \xrightarrow[n\to\infty]{P_{\lambda_0,\varsigma_0}} \int_{\mathbb{R}^d} g_0(x)\,\mathrm{d}x$.

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

$$\int_{\mathbb{R}^d} \left| q_n(x) - \mathcal{N}\left(x \mid 0, H_0^{-1}\right) \right| \mathrm{d}x = \int_{\mathbb{R}^d} |a_n g_n(x) - a_0 g_0(x)|\,\mathrm{d}x$$

$$\leq \int_{\mathbb{R}^d} |a_n g_n(x) - a_n g_0(x)|\,\mathrm{d}x + \int_{\mathbb{R}^d} |a_n g_0(x) - a_0 g_0(x)|\,\mathrm{d}x$$

$$\leq |a_n| \int_{\mathbb{R}^d} |g_n(x) - g_0(x)|\,\mathrm{d}x + |a_n - a_0| \int_{\mathbb{R}^d} |g_0(x)|\,\mathrm{d}x \xrightarrow[n\to\infty]{P_{\lambda_0,\zeta_0}} 0.$$

This proves $\mathrm{d}_{\mathrm{TV}}\{q_n, \mathcal{N}(0, H_0^{-1})\} \xrightarrow[n\to\infty]{P_{\lambda_0,\zeta_0}} 0$, and (15) follows from [Miller, 2021, Lemma 28]. $\qquad\square$

B   An illustration of least favorable submodel

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

Example 5 (Cox regression model without censoring). Consider the density function of the survival time $(y_1, \ldots, y_n)$ where $y_i \in \mathbb{R}_+ : \mathbb{R} \cap [0, \infty]$ with covariates $(x_1, \ldots, x_n)$ where $x_i \in \mathbb{R}$:

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

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

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

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

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

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

38

C   More Details of Proofs of Existing BvM Results

In Bickel and Kleijn [2012], to get the LAN property for the marginal likelihood of λ, they first assume that a neighborhood of z has enough prior mass, and the parameter space of z has bounded Hellinger metric entropy (covering number). They assume that for z outside a neighborhood of the fixed z_0, the Hellinger distances between the likelihood at $\lambda_0 + h_n/\sqrt{n}$ and the one at λ_0 are uniformly (with respect to z) infinitesimal for every bounded h_n. Finally, assuming that the least favorable submodel exists, that is a submodel $l_n(\lambda, z_\lambda)$ with parameters (λ, z_λ) satisfying $\mathcal{Q}\dot{l}_n(\lambda, z_\lambda) = \dfrac{\partial l_n(\lambda, z_\lambda)}{\partial \lambda}$ for all λ in a neighborhood of λ_0, the conditional posterior distribution of z can be shown to concentrate around the parameter of the least favorable submodel z_λ, that is the probability that the Hellinger distance between z and z_λ is greater than a positive number converges to zero. This leads to the marginal LAN property assuming that the full likelihood has LAN property in the direction of λ when the z is perturbed around the least favorable submodel, that is $z = z_\lambda + \zeta$ for all ζ in a neighborhood of 0.

In Castillo and Rousseau [2015], let h be the least favorable direction satisfying $\mathscr{P}\dot{l}_n(\lambda_0, z_0) = A_{\lambda_0, z_0} h$. Let $\lambda_t = \lambda - t\tilde{I}_0^{-1}/\sqrt{n}$, and $z_t = z + th\tilde{I}_0^{-1}/\sqrt{n}$. Assume over the direction of h, the log full likelihood has the LAN property with a remainder term, and the difference between the remainder terms at (λ, z) and at (λ_t, z_t) converges to zero. Further, suppose that the ratio between the integral of the likelihood under the prior at (λ_t, z_t) and the one at (λ, z) converge to 1 for all (λ, z). They proved the BvM theorem where the mean will be λ_0 plus the first order term of the LAN expansion.

D   Data augmentation for Latent Normal Model

For the latent normal model with binary observations, we follow Polson et al. [2013] and use the following data augmentation:

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

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

E   ADMM for Optimization Problem in Data Application

To solve the optimization problem

$$\min_{\tilde{L}} \frac{1}{2}\|\mathcal{L} - \tilde{L}\|_F^2 + \lambda\|\tilde{L}\|_* \qquad \text{subject to } \tilde{L}\mathbb{1} = \vec{0}, \tilde{L}_{ij} \leq 0 \text{ for } i \neq j, \tilde{L} = \tilde{L}^T,$$

we use ADMM under constraints and log-barrier:

$$\min_{\tilde{L},Z} \frac{1}{2}\|\mathcal{L} - \tilde{L}\|_F^2 + \rho \sum_{(i,j):i\neq j} \{-\log(-\tilde{L}_{i,j})\} + \lambda\|Z\|_* + \frac{\eta}{2}\|\tilde{L} - Z + W\|_F^2$$

$$\text{subject to } \tilde{L}\vec{1} = \vec{0}, \tilde{L} = \tilde{L}^\top, \tilde{L}_{ij} \leq 0 \text{ for } i \neq j$$

where $W = W^T$ is the Lagrangian multiplier. The ADMM algorithm iterates the following steps:

1. Constrained gradient descent for $\tilde{L}$: set $\tilde{L}$ to be

$$\arg\min \frac{1}{2}\|\mathcal{L} - \tilde{L}\|_F^2 + \rho \sum_{(i,j):i\neq j} \{-\log(-\tilde{L}_{i,j})\} + \frac{\eta}{2}\|\tilde{L} - Z + W\|_F^2$$

$$\text{subject to } \tilde{L}\vec{1} = \vec{0}, \tilde{L} = \tilde{L}^\top.$$

The constraints are easy to satisfy, by restricting the free parameters to $\{\tilde{L}_{i,j}\}_{i>j}$, and setting $\tilde{L}_{i,i} = -\sum_{j<i}\tilde{L}_{i,j} - \sum_{j>i}\tilde{L}_{j,i}$.

2. Minimizing over Z: set $Z = S_{\lambda/\eta}(\tilde{L} + W)$ where $S_{\lambda/\eta}(X) = \sum_{i=1}^n (\sigma_i - \lambda/\eta)_+ u_i v_i^T$, and $X = U\text{diag}(\sigma_i)V$ is the singular value decomposition. The solution of this step satisfies the conditions of symmetry and the rows add to zero.

3. Updating W: set W to be $\tilde{L} - Z + W$.