ON THE POSTERIOR CONSISTENCY AND BERNSTEIN-VON MISES PHENOMENON FOR THE DIACONIS-YLVISAKER PRIOR

Xin Jin

A Dissertation

Submitted to the Graduate College of Bowling Green
State University in partial fulfillment of
the requirements for the degree of

DOCTOR OF PHILOSOPHY

April 2024

Committee:

Riddhi Pratim Ghosh, Committee Chair

John Boman,
Graduate Faculty Representative

John Chen

Junfeng Shang

This content is unable to be made fully accessible because it would result in a fundamental alteration of the document.

ABSTRACT

Riddhi Pratim Ghosh, Committee Chair

We investigate the asymptotic behavior of the posterior distribution of the canonical parameter within the exponential family when the dimension of the parameter space grows with the sample size, specifically focusing on the Diaconis-Ylvisaker prior. This prior is notable as it acts as a conjugate prior for the exponential family. Our analysis establishes that, under mild conditions on both the true parameter value $\boldsymbol{\theta}_0$ and the hyperparameters of the prior, the distance between the posterior distribution and a normal distribution, centered at the maximum likelihood estimator with a variance equal to the inverse of the Fisher information matrix, approaches zero in the expected total variation distance norm. Our Bernstein-von Mises theorem requires only that the dimension of the parameter space d grows linearly with the sample size n, with the condition $d = o(n)$. In the process, we derive a concentration inequality for the quadratic form of the maximum likelihood estimator, circumventing the need for specific assumptions such as sub-Gaussianity. To illustrate our findings, we offer a specific application to the Multinomial-Dirichlet model, extending our analysis to deal with density estimation and Normal mean estimation problems.

ACKNOWLEDGMENTS

I would like to express my deepest gratitude to my advisor, Dr. Ghosh, for his unwavering support, invaluable guidance, and continuous encouragement throughout this journey. His expertise, patience, and insightful feedback have been instrumental in shaping this work.

I am also indebted to my committee members, Dr. Shang, Dr. Chen, and Dr. Boman, for their valuable contributions, constructive criticism, and insightful suggestions. Their expertise and diverse perspectives have greatly enriched the quality of this research.

Moreover, I extend my sincere appreciation to Dr. Anirban Bhattacharya in the Department of Statistics at Texas A&M University for many useful discussions.

Furthermore, I extend my heartfelt appreciation to my family for their unwavering support, encouragement, and understanding. Their love, encouragement, and sacrifices have been a constant source of strength and motivation for me.

I would also like to acknowledge Bowling Green State University for providing the resources and environment conducive to research and learning.

Finally, I am grateful to all those who have supported me in various ways throughout this endeavor. Your encouragement, assistance, and belief in me have been invaluable.

TABLE OF CONTENTS

Page

LIST OF FIGURES

Figure Page

LIST OF TABLES

Table Page

CHAPTER 1 INTRODUCTION

1.1 Basics of Bayesian Statistics

In Bayesian inference, to learn about the unknown parameter $\boldsymbol{\theta}$ given data $\boldsymbol{x}$, we utilize a model $f(\boldsymbol{x}|\boldsymbol{\theta})$ called the sampling distribution or data distribution, along with an appropriate prior distribution for $\boldsymbol{\theta}$, to obtain the posterior distribution. The prior distribution captures our uncertainty regarding $\boldsymbol{\theta}$ before observing the data. The choice of prior distribution can significantly impact the resulting posterior distribution and inference. It may reflect an integration of our subjective beliefs and knowledge about the parameter, constituting a subjective prior, or it may be a conventional prior representing minimal or no information, termed an objective prior.

Given the model and the prior, Bayesian inference determines the conditional probability density of $\boldsymbol{\theta}$ given $\boldsymbol{X} = \boldsymbol{x}$ using Bayes' theorem,

$$\pi(\boldsymbol{\theta}|\boldsymbol{x}) = \frac{\pi(\boldsymbol{\theta}) f(\boldsymbol{x}|\boldsymbol{\theta})}{\int_{\Theta} \pi(\boldsymbol{\theta}') f(\boldsymbol{x}|\boldsymbol{\theta}') d\boldsymbol{\theta}'}, \tag{1.1.1}$$

where $f(\boldsymbol{x}|\boldsymbol{\theta})$ denote the conditional density of $\boldsymbol{X}$ given $\boldsymbol{\theta}$, and $\pi(\boldsymbol{\theta})$ denote the prior density function. The numerator represents the joint density of $\boldsymbol{\theta}$ and $\boldsymbol{X}$, combining information from both the observed data and prior beliefs, while the denominator represents the marginal density of $\boldsymbol{X}$, containing solely information from the observed data. Here, the symbol $\boldsymbol{\theta}$ represents both a random variable and one of its realizations. Note that Equation (1.1.1) is defined for the continuous parameters. In cases where the parameter $\boldsymbol{\theta}$ is discrete, the integral in the denominator of Equation (1.1.1) is substituted with a sum. Given that the denominator remains independent of $\boldsymbol{\theta}$ for fixed $\boldsymbol{x}$, it can be regarded as a constant. Thus, an alternative representation of Equation (1.1.1) is

$$\pi(\boldsymbol{\theta}|\boldsymbol{x}) \propto \pi(\boldsymbol{\theta}) f(\boldsymbol{x}|\boldsymbol{\theta}). \tag{1.1.2}$$

The conditional density $\pi(\boldsymbol{\theta}|\boldsymbol{x})$ of $\boldsymbol{\theta}$ given $\boldsymbol{X} = \boldsymbol{x}$ is known as the posterior density, represent-

ing our uncertainty about the parameter $\boldsymbol{\theta}$ in view of the observed data. The prior beliefs $\pi(\boldsymbol{\theta})$ regarding the parameter is updated to $\pi(\boldsymbol{\theta}|\boldsymbol{x})$ by incorporating the insights gained from the data.

Unlike a Frequentist statistician, who typically reports properties of estimators such as unbiasedness, consistency, and efficiency, a Bayesian statistician may choose to either simply report the posterior distribution or provide summary descriptive statistics associated with it. As an illustration, the posterior mean for a real-valued parameter $\boldsymbol{\theta}$ might be presented

$$E(\boldsymbol{\theta}|\boldsymbol{x}) = \int_{-\infty}^{\infty} \boldsymbol{\theta}\pi(\boldsymbol{\theta}|\boldsymbol{x})d\boldsymbol{\theta},$$

and the posterior variance

$$\begin{aligned} \text{Var}(\boldsymbol{\theta}|\boldsymbol{x}) &= E\left\{(\boldsymbol{\theta} - E(\boldsymbol{\theta}|\boldsymbol{x}))^2|\boldsymbol{x}\right\} \\ &= \int_{-\infty}^{\infty} (\boldsymbol{\theta} - E(\boldsymbol{\theta}|\boldsymbol{x}))^2 \pi(\boldsymbol{\theta}|\boldsymbol{x})d\boldsymbol{\theta}. \end{aligned}$$

Alternatively, the posterior standard deviation could be presented as another measure of the dispersion or spread of the parameter, given a set of data points. Moreover, the posterior distribution offers a powerful and flexible framework for tackling complex problems in statistics. In estimation problems, the posterior distribution serves as a foundation for deriving point estimates (e.g., posterior mean, median) and interval estimates (e.g., credible sets) of the parameters of interest. In Bayesian hypothesis testing, the posterior distribution provides a natural framework for comparing competing hypotheses by evaluating their posterior probabilities, or odds.

1.2 Bayesian Inference

Bayesian inference, recognized for its robust and adaptable statistical framework, is renowned for its inherent flexibility in updating beliefs, offering predictions, and making inferences. This flexibility arises from its integration of multiple levels of randomness, encompassing both prior knowledge and observed data. Adopting a recursive perspective, the previous posterior distribution can serve as the updated prior distribution, amalgamating with new observed data to yield an up-to-date posterior distribution. Its origins can be traced back to the foundational work of Reverend

Thomas Bayes, who established the framework for this methodology. Grounded in probability theory, Bayesian inference offers a structured approach for evaluating unknowns in light of the data at hand. This approach enables the integration of information from diverse sources while comprehensively addressing all reasonable sources of uncertainty in inferential summaries. Its appealing methodological elegance has drawn attention from practitioners in a variety of fields.

Important components of Bayesian inference include the posterior distribution, which represents refined beliefs after observed data has been assimilated; the likelihood function, which measures the probability of observing data under various parameter settings; and the prior distribution, which acts as a store for prior knowledge or assumptions prior to data observation. Bayesian methods, extensively used in artificial intelligence, machine learning, and statistics, offer a cohesive and accessible framework for quantifying uncertainty, estimating parameters, and making decisions. Due to its adaptability and interpretability, Bayesian inference is a valuable tool for solving complicated problems in a variety of fields. It gives practitioners the confidence and comprehension needed to navigate intricate environments effectively.

1.2.1 Example: Estimating the Normal Mean with a Normal Prior

An example of inference concerning the parameter μ in a normal distribution $\mathcal{N}(\mu, \sigma^2)$, which is defined by mean μ and variance σ^2, is used to demonstrate these ideas (Ghosh, Delampady, and Samanta, 2006). The dataset consists of $x_1, x_2, \cdots, x_n$ observations from this distribution that are independent and identically distributed (i.i.d.). A normal distribution with an appropriate mean and variance, η and τ^2, respectively, is a mathematically convenient and reasonably flexible prior distribution for μ. A higher prior variance τ^2 indicates greater uncertainty about the true value of the parameter before observing any data, while a lower prior variance τ^2 suggests more confidence or precision in the prior beliefs about the parameter. A method to calibrate τ^2 by comparing it with σ^2 was proposed by Jeffreys (1961). For example, setting the prior variance τ^2 to be the data variance of size m, σ^2/m, indicates that the information concerning η is about equal to the information in m observations. The posterior density, represented by a normal distribution

expression, can be shown

$$\pi(\mu|\boldsymbol{x}) = \frac{1}{\sqrt{2\pi}\frac{\tau\sigma}{\sqrt{n\tau^2+\sigma^2}}} \exp\left\{ -\frac{n\tau^2+\sigma^2}{2\tau^2\sigma^2}\left(\mu - \frac{\sigma^2\eta + n\tau^2\bar{x}}{n\tau^2+\sigma^2}\right)^2\right\}.$$

Thus, the posterior mean is

$$E(\mu|\boldsymbol{x}) = \frac{\sigma^2\eta + n\tau^2\bar{x}}{n\tau^2+\sigma^2}, \tag{1.2.1}$$

and the posterior variance is

$$\text{Var}(\mu|\boldsymbol{x}) = \frac{\tau^2\sigma^2}{n\tau^2+\sigma^2}. \tag{1.2.2}$$

The variability decreases from σ^2 to $\tau^2\sigma^2/(n\tau^2+\sigma^2)$ when μ changes from the prior estimate η to a weighted average of the prior estimate and the sample mean $\bar{x}$. A large τ^2 or substantial data presence indicates inadequate prior information. In such instances, the posterior mean closely approximates the maximum likelihood estimator (MLE) $\bar{x}$.

Later, by comparing the prior $\pi(\mu)$ and the posterior $\pi(\mu|\boldsymbol{x})$, we will look into how we might quantify the knowledge obtained from the data. Both the prior and the data have an impact on the posterior distribution. The impact of the data tends to outweigh the influence of the prior as we get more and more data.

1.2.2 Example: Estimating the Probability of Success in Bernoulli Trials with a Beta Prior

Bernoulli trials consist of a sequence of independent experiments or observations, each resulting in either "success" or "failure". The objective is to estimate an unknown population proportion of "success" from the outcomes of a sequence of independent and identically distributed Bernoulli trials. The data can be summarized by the total number of successes in the n trials.

Suppose we select a random sample of n children, denoted as $x_1, \ldots, x_n$, where each child

can be categorized as either 0 or 1. For each $i \in \{1, \ldots, n\}$, we define X_i as follows,

$$X_i = \begin{cases} 1 & \text{if the } i\text{th child in the sample has a food allergy.} \\ 0 & \text{otherwise} \end{cases}$$

Then, the random variables X_i's are from $B(1, p)$, i.e., Bernoulli random variables with probability of having a food allergy p. Hence, the number of children in the sample $\boldsymbol{x} = (x_1, \ldots, x_n)^\top$ who have a food allergy follows a binomial sampling model,

$$\pi(\boldsymbol{x}|p) = \binom{n}{\sum_{i=1}^n x_i} p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i}, \quad 0 \leq p \leq 1, \tag{1.2.3}$$

and

$$\pi(\boldsymbol{x}|p) \propto p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i}.$$

The choice of the prior is flexible. We can opt for a non-informative prior, where the prior distribution for p is uniform on the interval $[0, 1]$. Alternatively, we can explore a family of priors for p that facilitate the computation of the posterior. For simplicity at this stage, let's begin with a conjugate prior. Let $\pi(p)$ denote the prior distribution of p, then

$$\pi(p) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}, \quad 0 \leq p \leq 1; \alpha > 0, \beta > 0. \tag{1.2.4}$$

We refer to this distribution as a Beta distribution with hyperparameters α and β, denoted as Beta(α, β). Removing the constant term in Equation (1.2.4), we have

$$\pi(p) \propto p^{\alpha-1}(1-p)^{\beta-1}.$$

The prior mean and variance are given by

$$E(p) = \frac{\alpha}{\alpha + \beta},$$

$$\text{Var}(p) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}.$$

Keep prior density $\pi(p)$ in line with the sampling distribution Equation (1.2.3), $\pi(p)$ is equivalent to $\alpha - 1$ prior successes and $\beta - 1$ prior failures. Using Bayes' theorem, we can obtain the posterior density as

$$\pi(p|\boldsymbol{X} = \boldsymbol{x}) \propto p^{\alpha+\sum_{i=1}^n x_i - 1}(1-p)^{\beta+(n-\sum_{i=1}^n x_i)-1}, \quad 0 \leq p \leq 1; \alpha > 0, \beta > 0. \tag{1.2.5}$$

The posterior distribution has a Beta density, as can be seen by comparing it with Equation (1.2.4), replacing $\beta + (n - \sum_{i=1}^n x_i)$ with β, $\alpha + \sum_{i=1}^n x_i$ with α. Note that the exact expression for the posterior can be obtained by adding a constant term $C(\boldsymbol{x})$ to Equation (1.2.5), where

$$C(\boldsymbol{x}) = \frac{\Gamma(\alpha+\beta+n)}{\Gamma(\alpha+\sum_{i=1}^n x_i)\Gamma(\beta+n-\sum_{i=1}^n x_i)}.$$

The variance and mean of the posterior distribution can be computed

$$\begin{aligned} E(p|\boldsymbol{x}) &= \frac{\alpha + \sum_{i=1}^n x_i}{\alpha+\beta+n}, \\ \text{Var}(p|\boldsymbol{x}) &= \frac{(\alpha+\sum_{i=1}^n x_i)/(\beta+n-\sum_{i=1}^n x_i)}{(\alpha+\beta+n)^2(\alpha+\beta+n+1)}. \end{aligned} \tag{1.2.6}$$

As previously noted, a Bayesian statistician has two options for reporting: either they report the posterior only, see Equation (1.2.5), or they report the posterior mean and variance, see Equation (1.2.6), which gives information about the location and dispersion of the posterior distribution. It is important to note that the posterior variance is modest if n is big, and the posterior mean is roughly equivalent to the maximum likelihood estimator, $\hat{p} = \sum_{i=1}^n x_i/n = \bar{x}$. As a result, the posterior distribution is centered at p for large n. This discovery supports the earlier idea that the influence of the prior diminishes with an increasing amount of data.

The posterior mean can be expressed by a weighted average of the prior mean and the

maximum likelihood estimator

$$\frac{(\alpha+\beta)}{(\alpha+\beta+n)}\frac{\alpha}{\alpha+\beta}+\frac{n}{(\alpha+\beta+n)}\bar{x}.$$

This implies that always the posterior mean lies between the sample proportion, $\bar{x}$, and the prior mean, $\alpha/(\alpha+\beta)$. Again, the prior and the data both have significance, and $(\alpha+\beta)$ and n, respectively, indicate the relative importance of the two information sources.

Let's suppose our interest, based on the existing data, lies in predicting the posterior outcome of a single new trial, rather than forecasting another set of n new trials. In this example, we aim to predict the probability of a child having a food allergy in a new $(n+1)$-th draw. This prediction underscores a fundamental issue in scientific research. Employing the same estimate as previously determined, the posterior mean $E(p|\boldsymbol{x})$, makes intuitive sense. The following explanation includes some commonly used priors and their corresponding values of the posterior mean.

Let x_{i+1} denote the result of a new trial. Using $\alpha=\beta=1$ as the uniform prior, we obtain a posterior prediction

$$\begin{aligned}\Pr(x_{i+1}=1|\boldsymbol{x}) &= \int_0^1 \Pr(x_{i+1}=1|p,\boldsymbol{x})\pi(p|\boldsymbol{x})dp\\ &= \int_0^1 p\pi(p|\boldsymbol{x})dp\\ &= E(p|\boldsymbol{x})\\ &= \frac{\sum_{i=1}^n x_i+1}{n+2}.\end{aligned}$$

For further details on the calculation, refer to Gelman, Carlin, Stern, and Rubin (2004). Although Laplace and Bayes favored this prior, it is not as widely used as it formerly was. As an alternative, we derive the Jeffreys prior for $\alpha=\beta=1/2$, with a posterior mean of $(\sum_{i=1}^n x_i+1/2)/(n+1)$. This prior is frequently applied, particularly when dealing with one-dimensional parameters. As Bernardo (1979) points out, it is also a reference prior. The use of reference priors

is widespread. A Beta density that has $\alpha = 0, \beta = 0$ integrates to infinity, which makes it an improper prior. However, unless two extreme cases occur, where $\sum_{i=1}^{n} x_i = 0$ or n, the posterior remains valid when calculating a posterior density using the Bayesian method. In this instance, the posterior mean precisely corresponds to the maximum likelihood estimator.

Objective priors, also known as non-informative priors, are a type of prior distribution used in Bayesian statistics. Unlike subjective priors, which are chosen based on prior knowledge or beliefs about the parameters being estimated, objective priors are designed to be minimally informative and to reflect a lack of prior information about the parameters. Objective priors are typically chosen to satisfy certain desirable properties, such as being invariant under reparameterization, being minimally informative in terms of influence on the posterior distribution, or being invariant under transformations of the parameter space.

From the discussed examples, we learned that although objective priors are usually improper, they can provide proper posteriors in order to be useful. Objective priors are particularly useful when little or no prior information is available about the parameters being estimated, or when it is desirable to minimize the influence of the prior on the posterior distribution.

Assume that the problem is represented by the production of both defective and non-defective products in a factory that makes switches, where the functional switches are represented by black and the defective ones by red. Engineers might have some prior knowledge in this case. They could be able to determine the most likely value of p, which could be $\alpha/(\alpha + \beta)$, the prior mean. Two formulae to calculate α and β would be provided if one also knows the prior variability. In this case, Jeffreys prior, characterized by having a significant portion of its probability mass concentrated at both extreme endpoints, might be appropriate. This is particularly true when the process typically operates at a high level of quality, which corresponds to small values of the parameter p. However, there are instances when the process deviates from this expected behavior, leading to higher values of p indicating lack of control. The peak of the prior distribution near $p = 1$ could signify frequent occurrences of such lack of control situations or could represent a pessimistic prior belief aimed at anticipating and mitigating potential disasters.

It is noteworthy advantages of objective priors, such as the uniform, Jeffreys, and reference priors. These priors, even with small sample sizes n, tend to yield posterior means that closely approximate the maximum likelihood estimator. Additionally, in situations where the maximum likelihood estimator suggests $\hat{p} = 0$, which is unlikely in most cases, the classical statistical analysis based on frequency does not make sense. In contrast, objective Bayes estimates adjust significantly towards $p = 1/2$, representing a state of complete ignorance. This adjustment, known as shrinkage, results in more plausible point estimates and more reliable confidence intervals.

1.2.3 Example: Hypothesis Testing for the Normal Mean with a Normal Prior

Assuming for convenience that σ^2 is known, let $X_1, X_2, \ldots, X_n$ be independent and identically distributed random variables from $\mathcal{N}(\mu, \sigma^2)$. In the example discussed in Ghosh et al. (2006), μ denotes the expected drop in blood pressure brought on by a novel medication. The hypothesis to test is $H_0 : \mu \leq \mu_0$ vs $H_a : \mu > \mu_0$, where μ_0 denotes the degree of efficacy of a standard medication that is now available for purchase.

Let $\pi(\mu)$ denote the prior distribution. Determine the posterior density $\pi(\mu|\boldsymbol{X})$ first. Next, find out

$$\int_{-\infty}^{\mu_0} \pi(\mu|\boldsymbol{X}) d\mu = \Pr\{H_0|\boldsymbol{X}\},$$

and

$$\int_{\mu_0}^{\infty} \pi(\mu|\boldsymbol{X}) d\mu = 1 - \Pr\{H_0|\boldsymbol{X}\} = \Pr\{H_1|\boldsymbol{X}\}.$$

If one of the two hypotheses is much more likely than the other, one might choose that hypothesis or just report the values.

Assuming that the prior for μ has a normal distribution with mean η and variance τ^2, we show some computations. With mean and variance provided by Equation (1.2.1) and Equa-

tion (1.2.2), the posterior for μ is also normal. If after that

$$\pi(\mu \leq \mu_0 | \boldsymbol{X}) = \Phi(z) \text{ and } \pi(\mu > \mu_0 | \boldsymbol{X}) = 1 - \Phi(z),$$

where Φ denote the standard normal distribution function and

$$z = \frac{\mu_0 - (\eta/\tau^2 + n/\sigma^2 \bar{X})/(1/\tau^2 + n/\sigma^2)}{\sqrt{(\sigma^2\tau^2/n)/(\sigma^2/n + \tau^2)}}.$$

Setting τ^2 to approach infinity is a common method that has the same result as assuming a uniform prior

$$\pi(\mu) = c, \qquad -\infty < \mu < \infty.$$

Any of these could yield

$$z = (\mu_0 - \bar{X})\frac{\sqrt{n}}{\sigma}.$$

Assume that if the posterior probability of H_0 is less than 0.05, we intend to reject the null hypothesis. As a result, we reject H_0 if

$$\mu_0 - \bar{X} \leq -1.64\frac{\sigma}{n} \text{ or } \bar{X} \geq \mu_0 + 1.64\frac{\sigma}{n}.$$

At a significance level of $\alpha = 0.05$, this decision rule perfectly is in line with the results of the conventional test for this problem in Frequentist statistics.

In the broad field of research, our goal has been to test the sharp null hypothesis $H_0 : \mu = \mu_0$ against $H_a : \mu \neq \mu_0$. Dealing with this type of hypothesis testing would require selecting a different prior, as the prior we would have chosen would assign zero probability to H_0. In such cases, Bayes factors often deviate significantly from those of classical results.

Johnson and Rossell (2010) proposes non-local prior densities in Bayesian hypothesis tests.

Non-local prior densities impact the accumulation of evidence by providing a more balanced approach towards true null and true alternative hypotheses. Traditional Bayesian hypothesis tests using local alternative priors tend to accumulate evidence much more rapidly in favor of true alternative models as the sample size increases. This asymmetry results in a linear increase in the logarithm of the Bayes factor in favor of the true alternative hypothesis, while evidence for the true null hypothesis accumulates at a slower rate.

On the other hand, non-local prior densities assign non-negligible probability to regions of the parameter space consistent with null hypotheses, leading to exponential accumulation of evidence in favor of true alternative hypotheses and sublinear accumulation of evidence in favor of true null hypotheses. This balanced approach allows for a more equitable evaluation of both hypotheses, addressing the issue of asymmetry in evidence accumulation seen in tests using local alternative priors.

An example regarding the test of a normal mean is provided to contrast the performance of local and non-local alternative priors (Johnson and Rossell, 2010). Consider independent and identically distributed data from a normal distribution with mean parameter θ and unit variance. The null hypothesis being tested is $H_0 : \theta = 0$ against various alternative hypotheses. Specifically, the alternative hypotheses considered in the example are:

$$H_1^a : \pi(\theta) = \mathcal{N}(\theta; 0, 2)$$
$$H_1^b : \pi(\theta) = \text{Cauchy}(\theta)$$
$$H_1^c : \pi(\theta) \propto (\theta^2)^{-1} \exp(-0.318/\theta^2)$$
$$H_1^d : \pi(\theta) \propto \theta^2 n(\theta; 0, 0.159)$$

These alternative hypotheses are defined using non-local densities, with H_1^a corresponding to an intrinsic prior, H_1^b following Jeffreys's recommendation, and H_1^c and H_1^d utilizing inverse moment priors. The parameters of the inverse moment prior in H_1^c were specified to match the tails of the Cauchy prior in H_1^b. This example demonstrates how different types of non-local alternative priors can be used in the context of testing a normal mean, showcasing the impact of these priors on the accumulation of evidence in Bayesian hypothesis testing.

1.3 Bayesian Decision Theory

According to Bayesian theory, the conditional law of $\boldsymbol{X}$ given a random variable $\boldsymbol{\theta}$ is defined as the distribution $P_{\boldsymbol{\theta}}$ of a random variable $\boldsymbol{X}$ under a parameter $\boldsymbol{\theta}$. The conditional distribution of $\boldsymbol{\theta}$ given $\boldsymbol{X}$ is called the posterior distribution, and the distribution of the random parameter $\boldsymbol{\theta}$ is called the prior distribution. The posterior density is determined by Bayes' theorem if $\boldsymbol{\theta}$ has a density of π and $P_{\boldsymbol{\theta}}$ admits a density of $p_{\boldsymbol{\theta}}$ in relation to dominant measures,

$$\pi(\boldsymbol{\theta}|\boldsymbol{X}=\boldsymbol{x}) = \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x})\pi(\boldsymbol{\theta})}{\int p_{\boldsymbol{\theta}}(\boldsymbol{x})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}}.$$

It's essentially saying that even if the prior distribution π doesn't strictly meet the criteria of being a probability density function, e.g., it might not integrate to 1 over its domain, the Bayesian theorem can still be applied to update our beliefs and derive a posterior probability distribution. Such improper priors refer to prior distributions having infinite mass or where the integral diverges or does not exist.

One could consider the main goal of a Bayesian analysis to be the computation of the posterior distribution. As an alternative, one may use the posterior distribution in an attempt to make a point estimator for the parameter $\boldsymbol{\theta}$. For this task, the posterior mean

$$E(\boldsymbol{\theta}|\boldsymbol{X}) = \int \boldsymbol{\theta}\pi(\boldsymbol{\theta}|\boldsymbol{X}=\boldsymbol{x})d\boldsymbol{\theta}$$

is commonly used; however, other location estimators, including the posterior median and mode, are also valid.

A loss function may be used to guide the choice of "best" point estimator. Refer to Van der Vaart (2000), the definition of the Bayes risk of an estimator $\boldsymbol{T}^*$ in relation to the loss function L is $E_{\boldsymbol{\theta}|\boldsymbol{X}}[L(\boldsymbol{T}^* - \boldsymbol{\theta})]$.

In this case, the conditional risk $E\left[L(\boldsymbol{T}^* - \boldsymbol{\theta})|\boldsymbol{\theta}\right]$ in Bayesian notation is the same as the expectation of $E[L(\boldsymbol{T}^* - \boldsymbol{\theta})]$, which in the classical framework reflects the risk function of $\boldsymbol{T}$.

The estimator $\boldsymbol{T}$ that minimizes the Bayes risk is the corresponding Bayes estimator. Given that $E\left[E\left(L(\boldsymbol{T}^* - \boldsymbol{\theta})|\boldsymbol{X}\right)\right]$ is one way to describe the Bayes estimator, for any fixed $\boldsymbol{x}$, the value $\boldsymbol{T}^* = \boldsymbol{T}(\boldsymbol{x})$ minimizes the posterior risk

$$E\left[L(\boldsymbol{T}^* - \boldsymbol{\theta})|\boldsymbol{X} = \boldsymbol{x}\right] = \frac{\int L(\boldsymbol{T}^* - \boldsymbol{\theta}) p_{\boldsymbol{\theta}}(\boldsymbol{x}) \pi(\boldsymbol{\theta}) d\boldsymbol{\theta}}{\int p_{\boldsymbol{\theta}}(\boldsymbol{x}) \pi(\boldsymbol{\theta}) d\boldsymbol{\theta}}.$$

Reducing this expression could once more be a well-defined issue, even at earlier densities that had an infinite total mass. The posterior mean $E(\boldsymbol{\theta}|\boldsymbol{X})$ is the solution $\boldsymbol{T}^*$ for the loss function $L(y) = \|y\|_2^2$. The posterior median provides a solution for the absolute loss $L(y) = |y|$.

Other Bayesian point estimators include the posterior mode, which, with a uniform prior density, converges to the maximum likelihood estimator. The maximum probability estimator is an additional technique that finds the center of the smallest ball that contains at least half of the posterior mass.

1.3.1 Example: Bayes Estimators for the Normal Mean with a Normal Prior

Consider $X_1, \ldots, X_n$ are iid with distribution $N(\theta, 1)$, and the prior $\pi(\theta) \sim N(0, 1)$. The likelihood function is

$$L(\theta) = f(\boldsymbol{x}|\theta) = \left(\frac{1}{\sqrt{2\pi}}\right)^n \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(x_i - \theta)^2\right\},$$

and the prior is

$$\pi(\theta) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\theta^2\right\}.$$

By Equation (1.1.2), the posterior can be computed by

$$\begin{aligned}
p(\theta|\boldsymbol{x}) &\propto f(\boldsymbol{x}|\theta)p(\theta) \\
&\propto \exp\left\{-\frac{1}{2}\left[\sum_{i=1}^{n} x_i^2 - 2n\bar{x}\theta + n\theta^2 + \theta^2\right]\right\} \\
&\propto \exp\left\{-\frac{1}{2}\left[(n+1)\theta^2 - 2n\bar{x}\theta\right]\right\} \\
&= \exp\left\{-\frac{n+1}{2}\left[\theta^2 - 2\frac{n\bar{x}}{n+1}\theta\right]\right\} \\
&\propto \exp\left\{-\frac{n+1}{2}\left[\theta^2 - 2\frac{n\bar{x}}{n+1}\theta + \left(\frac{n\bar{x}}{n+1}\right)^2\right]\right\} \\
&= \exp\left\{-\frac{n+1}{2}\left[\left(\theta - \frac{n\bar{x}}{n+1}\right)^2\right]\right\}.
\end{aligned}$$

Therefore, the posterior follows a normal distribution as shown below

$$\pi(\theta|\boldsymbol{x}) \sim N\left(\frac{n\bar{x}}{n+1}, \frac{1}{n+1}\right).$$

We are interested in the Bayes estimator for θ under squared error loss. This Bayes estimator for θ corresponds to the posterior mean of θ under this loss function. That is,

$$T^* = E_{\theta|\boldsymbol{x}}(\theta) = \frac{n\bar{x}}{n+1}.$$

Suppose the parameter of interest is θ^2. In this case, the Bayes estimator for θ^2 under squared error loss is the posterior mean of θ^2. That is,

$$T^* = E_{\theta|\boldsymbol{x}}(\theta^2) = \text{Var}_{\theta|\boldsymbol{x}}(\theta) + \left[E_{\theta|\boldsymbol{x}}(\theta)\right]^2 = \frac{1}{n+1} + \left(\frac{n\bar{x}}{n+1}\right)^2.$$

Under absolute error loss, the Bayes estimator for θ is the posterior median of θ. Given the symmetric nature of the posterior distribution, the posterior median coincides with the posterior

mean. Therefore, the Bayes estimator in this case is

$$T^* = E_{\theta|\boldsymbol{x}}(\theta) = \frac{n\bar{x}}{n+1}.$$

1.3.2 Convergence of Bayes Estimators

In general, all these estimators are asymptotically equivalent if the underlying experiments converge to a Gaussian location experiment in a reasonable sense. In this instance, the observation comprises a random sample of size n drawn from a density $p_{\boldsymbol{\theta}}$, which is smoothly dependent on a $\boldsymbol{\theta}$ Euclidean parameter. As a result, the density $p_{\boldsymbol{\theta}}$ takes on a product form, and the posterior density, given a prior Lebesgue density π, follows the form as described in (Van der Vaart, 2000),

$$p(\boldsymbol{\theta}|\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n) = \frac{\prod_{i=1}^{n} p_{\boldsymbol{\theta}}(\boldsymbol{X}_i)\pi(\boldsymbol{\theta})}{\int \prod_{i=1}^{n} p_{\boldsymbol{\theta}}(\boldsymbol{X}_i)\pi(\boldsymbol{\theta})d\boldsymbol{\theta}}.$$

akin to Equation (1.1.1). As the sample size n grows indefinitely, the distribution corresponding to this measure typically tends to converge to a measure that is concentrated at the true parameter value $\boldsymbol{\theta}_0$. Bayesian estimators are usually consistent in this setting. In order to investigate a more complex limit, we first normalize the parameter as usual and look at the sequence of posterior distributions of $\sqrt{n}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)$, whose densities are given by

$$p_{\sqrt{n}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)|\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n}(\boldsymbol{h}) = \frac{\prod_{i=1}^{n} p_{\boldsymbol{\theta}_0+\boldsymbol{h}/\sqrt{n}}(\boldsymbol{X}_i)\pi(\boldsymbol{\theta}_0 + \boldsymbol{h}/\sqrt{n})}{\int \prod_{i=1}^{n} p_{\boldsymbol{\theta}_0+\boldsymbol{h}/\sqrt{n}}(\boldsymbol{X}_i)\pi(\boldsymbol{\theta}_0 + \boldsymbol{h}/\sqrt{n})d\boldsymbol{h}}.$$

If π is the continuous prior density, then $\pi(\boldsymbol{\theta}_0 + \boldsymbol{h}/\sqrt{n})$ behaves like the constant $\pi(\boldsymbol{\theta}_0)$ as n grows, and π can be dropped from the expression for the posterior density. Local asymptotic normality is demonstrated by the sequence of models $\left(P_{\boldsymbol{\theta}_0+\boldsymbol{h}/\sqrt{n}} : \boldsymbol{h} \in \mathbb{R}^d\right)$ with densities $p_{\boldsymbol{\theta}}$ that

show appropriate smoothness in parameter space. This means that the likelihood ratio satisfies

$$\boldsymbol{h} \to \prod_{i=1}^{n} p_{\boldsymbol{\theta}_0+\boldsymbol{h}/\sqrt{n}}/p_{\boldsymbol{\theta}_0}(\boldsymbol{X}_i).$$

They behave like the likelihood ratio process of the typical experiment $\left(\mathcal{N}(\boldsymbol{h}, \boldsymbol{I}_{\boldsymbol{\theta}_0}^{-1}) : \boldsymbol{h} \in \mathbb{R}^d\right)$ in the asymptotic regime. As a result, we predict that the statement above is asymptotically equivalent in distribution to

$$\frac{d\mathcal{N}(\boldsymbol{h}, \boldsymbol{I}_{\boldsymbol{\theta}_0}^{-1})(\boldsymbol{X}^*)}{\displaystyle\int d\mathcal{N}(\boldsymbol{h}, \boldsymbol{I}_{\boldsymbol{\theta}_0}^{-1})(\boldsymbol{X}^*)d\boldsymbol{h}} = d\mathcal{N}(\boldsymbol{X}^*, \boldsymbol{I}_{\boldsymbol{\theta}_0}^{-1})(\boldsymbol{h}), \tag{1.3.1}$$

where the normal distribution density is denoted by the expression $d\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. In terms of the improper Lebesgue prior distribution, the expression previously given exactly matches the posterior density for the experiment $\left(\mathcal{N}(\boldsymbol{h}, \boldsymbol{I}_{\boldsymbol{\theta}_0}^{-1}) : \boldsymbol{h} \in \mathbb{R}^d\right)$. This distribution is normal, as indicated by the formula on the right in Equation (1.3.1), which has a mean of $\boldsymbol{X}^*$ and a covariance matrix of $\boldsymbol{I}_{\boldsymbol{\theta}_0}^{-1}$.

According to this heuristic argument, for the true parameter $\boldsymbol{\theta}_0$, the posterior distribution of the Gaussian limit experiment respect to the Lebesgue prior is predicted to converge to the posterior distribution of $\sqrt{n}(\boldsymbol{\theta}_n - \boldsymbol{\theta}_0)$. In contrast, $\boldsymbol{X}^*$ follows the $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_{\boldsymbol{\theta}_0}^{-1})$-distribution, and the latter is equivalent to the $\mathcal{N}(\boldsymbol{X}^*, \boldsymbol{I}_{\boldsymbol{\theta}_0}^{-1})$-distribution. Here, convergence is characterized in terms of stochastic processes and probability measures; the details of the heuristics do not have to be precisely defined at this point. However, the convergence should include the assumption that, for well-behaved Euclidean-valued functionals applied to the posterior laws, there is a typical convergence in distribution.

Because of this, it is expected that a sequence of Bayes point estimators, which are effectively location functionals applied to the posterior distributions, will converge to the appropriate Bayes point estimator in the limit experiment. The majority of location estimators that are judged reasonable correspond to the centers of symmetry of symmetric distributions, such as the normal distribution. As a result, $\boldsymbol{X}^*$ is the Bayes point estimator in the limit. We expect the distribution of

Bayes point estimators to converge to the random vector $\boldsymbol{X}^*$, or to the $\mathcal{N}(\mathbf{0}, \boldsymbol{I}_{\boldsymbol{\theta}_0}^{-1})$-distribution under $\boldsymbol{\theta}_0$. Specifically, assuming regularity criteria, they are anticipated to be asymptotically efficient and asymptotically comparable to maximum likelihood estimators.

This result has an interesting side effect: the limit distribution of a sequence of Bayes estimators is invariant with respect to the prior measure that is used. Apparently, the observed data gradually modifies one's prior ideas as the number of observations rises. This statement mostly depends on the assumption that the previous distribution has a positive, smooth density in the neighborhood of the true parameter value. In the absence of this attribute, the previously stated conclusion is invalid. The sequence of posterior distributions of $\boldsymbol{\theta}$, for example, cannot even be consistent if one strictly follows a fixed discrete distribution that gives $\boldsymbol{\theta}_0$ zero probability mass.

We focus on the locally asymptotically normal situation, although the heuristic argument holds for convergence scenarios beyond Gaussian location experiments. In particular, we consider that the observations consist of a random sample $x_1, \ldots, x_n$ from a distribution $P_{\boldsymbol{\theta}}$, which admits a density $p_{\boldsymbol{\theta}}$ with respect to a measure ν on a measurable space $(\mathcal{X}, \mathcal{A})$. The true parameter $\boldsymbol{\theta}_0$ is assumed to be an interior point of a measurable subset $\boldsymbol{\Theta}$ of $\mathbb{R}^d$, to which the parameter $\boldsymbol{\theta}$ is assumed to belong. The mappings $(\boldsymbol{\theta}, \boldsymbol{x}) \to p_{\boldsymbol{\theta}}(\boldsymbol{x})$ are also conjectured to be jointly measurable.

1.4 Advantages of Being a Bayesian

The Bayes' theorem, a fundamental concept outlining the reassessment of uncertainty in response to new data, lies at the heart of Bayesian inference. This iterative process generates a posterior probability distribution that embodies updated beliefs conditioned on the observed data, integrating prior knowledge with empirical evidence. Through successive iterations, Bayesian inference develops a rational evolution of information, enhancing understanding continuously as new data is incorporated.

The Bayesian approach provides a well-defined remedy for common problems in statistical inference. In high-dimensional data analysis, Bayesian methods offer advantages such as regularization through prior distributions, which help prevent overfitting and improve model robustness. Additionally, Bayesian techniques facilitate uncertainty quantification, allowing for the incorpora-

tion of uncertainty in both the data and the model parameters. This is particularly beneficial when dealing with limited or noisy data. Furthermore, Bayesian approaches enable the integration of prior knowledge, expert opinions, and observed data to form a coherent framework for decision making. By explicitly modeling uncertainty and updating beliefs based on new evidence, Bayesian methods provide a principled way to make decisions under uncertainty.

When there is access to subjective data, it is possible to elicit a subjective prior, which makes it easier to include expert knowledge in the analysis. As an alternative, objective priors can frequently be selected to offer a consistent method for inference. It is wise to assess how resilient certain components of the posterior distribution are to slight changes in the prior specification, regardless of the choice of the prior.

Wald's Minimax Theorem (Wald, 1949), states that under certain conditions, the minimax decision rule is optimal in decision making. The minimax rule aims to minimize the maximum possible loss, assuming that the decision maker faces a worst-case scenario. In Bayesian decision theory, this theorem can be interpreted in terms of Bayes risk, which is the expected loss under a given decision rule and a prior distribution.

A number of axiom sets can be used to develop the Bayesian framework, which provides a sound basis for statistical inference. Moreover, the subjective Bayesian method resolves some paradoxes or principles violations related to classical statistics. These undesirable characteristics result from the dependence of classical statistics on evaluations like risk functions or confidence coefficients, which are derived by integrating over the entire sample space and may produce illogical results when specific data are available, or measures like p-values, which can be difficult to interpret. These paradoxes sometimes have a strong presence. Even while the objective Bayesian approach helps to alleviate some of these problems, there are still some rules that it does not adhere to.

Generally speaking, Bayesians emphasize the importance of real-world validation whenever feasible (Ghosh et al., 2006). Essentially, Bayesians advocate for the validation of statistical models and predictions through real-world observations or experiments. However, in cases where

direct real-world validation is not possible, Bayesians may resort to using conceptual constructs derived from Frequentist approaches to represent plausible scenarios. Despite being based on different statistical philosophies, these conceptual constructs serve as proxies for real-world situations. In such instances, Bayesians may seek validation by comparing the outcomes or predictions derived from their Bayesian models with those derived from the Frequentist constructs.

This approach allows Bayesians to assess the performance and reliability of their models in contexts where direct real-world validation is challenging or impractical. Compared to conventional approaches, Morris (1983) and Ghosh (2021) demonstrate the effective application of the parametric empirical Bayes methodology. Cross-validation is discussed in Morris (1983). The Bayesian approach to model selection is validated in Hoeting, Madigan, Raftery, and Volinsky (1999). Most Bayesian publications offer validation for novel methods. In this study, we examine the validation of proposed theorems in the Multinomial-Dirichlet model in Chapter 4.

Furthermore, Bayesian methods are easy to interpret for the general public and can be understood by individuals without a background in statistics. In many practical cases, clients interpret interval estimates provided by statisticians as Bayesian intervals, meaning they view them as probability statements regarding the likely values of unknown quantities based on the evidence in the data.

Notwithstanding these advantages, the Bayesian paradigm has primarily gained traction and broad adoption in recent times, especially in the last fifteen years. A major force behind this explosion has been the incredible progress in computing techniques, most notably the widespread application of Markov Chain Monte Carlo (MCMC) methods. These developments have made it possible to compute posterior distributions efficiently even in high-dimensional parameter spaces, which makes Bayesian analysis useful in a wide range of real-world situations. A fundamental study on sampling-based techniques (Gelfand and Smith, 1990) marked the beginning of these revolutionary breakthroughs.

1.5 Bayesian vs. Frequentist in Normal Mean Estimation

Consider a set of independent random samples $x_1, \ldots, x_n$ that have the same distribution. These samples come from a normal distribution, i.e., $\boldsymbol{x} = (x_1, \ldots, x_n)^\top \sim \mathcal{N}(\mu, \sigma^2)$, where the population mean is μ and the population variance is known to be σ^2. The maximum likelihood estimator $\bar{x}$ is commonly used in Frequentist statistics to estimate the population mean μ.

Bayesian statistics, on the other hand, offers an alternative viewpoint. Suppose, for the purposes of this discussion, that the prior distribution was a normal distribution with variance τ^2 and mean η, or $\mu \sim \mathcal{N}(\eta, \tau^2)$. In this approach, the maximum likelihood estimator $\bar{x}$ and the previous mean η are weighted to get the posterior mean,

$$\hat{\mu}|\boldsymbol{x} = \frac{\sigma^2}{n\tau^2 + \sigma^2} \cdot \eta + \frac{n\tau^2}{n\tau^2 + \sigma^2} \cdot \bar{x}.$$

This combination is commonly termed as the convex combination of the prior mean and the maximum likelihood estimator. It underscores the significance of both prior information and observed data, where the relative importance is influenced by factors such as the prior variance, data variance, and the sample size. As the prior variance τ^2 goes to infinity, that is considering a flat prior, which is also known as non-informative prior, the coefficient associated with the prior mean goes to 0 and the coefficient associated with the maximum likelihood estimator goes to 1. In this case, the posterior mean converges to the MLE. As the sample size n tends to infinity, the coefficient associated with the prior mean tends to 0, while the coefficient associated with the maximum likelihood estimator tends to 1. Consequently, the posterior mean converges towards the maximum likelihood estimator. This implies that with substantial amounts of data, the influence of the prior diminishes, and the data takes precedence in the estimation process.

1.6 Bayesian vs. Frequentist in Consistency

In Frequentist statistics, an estimator $\hat{\boldsymbol{\theta}}_n$ obtained from n independent and identically distributed random samples of $\boldsymbol{\theta}_0$ is regarded as consistent if, for any $\epsilon > 0$,

$$\Pr\left(\left|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\right| > \epsilon\right) \to 0 \text{ as } n \to \infty.$$

An estimator $\hat{\boldsymbol{\theta}}_n$ of the true parameter is deemed consistent if, for every $\epsilon > 0$, the probability of the set lying outside a neighborhood centered around the true parameter tends towards 0 as the sample size n approaches infinity. In other words, as the number of data points used increases indefinitely, the resulting sequence of estimates converges in probability to $\boldsymbol{\theta}_0$. This indicates that, with an increasing sample size, the distribution of the estimator gradually converges towards the true parameter.

Consider an illustrative scenario where our objective is to estimate the population mean of a normal distribution $\mathcal{N}(\theta, 1)$ with a known variance of 1. The maximum likelihood estimator for θ is denoted by the sample mean $\bar{x}$. In our analysis, we manipulate the sample size, choosing $n = 5, 10$, and 50, while fixing the true parameter $\theta_0 = 1$. Remarkably, as we increase the sample size, the estimators become progressively more tightly clustered around the true value θ_0. This phenomenon is visually depicted in Figure 1.1, illustrating that a larger proportion of the probability mass is contained within the same neighborhood as the sample size grows.

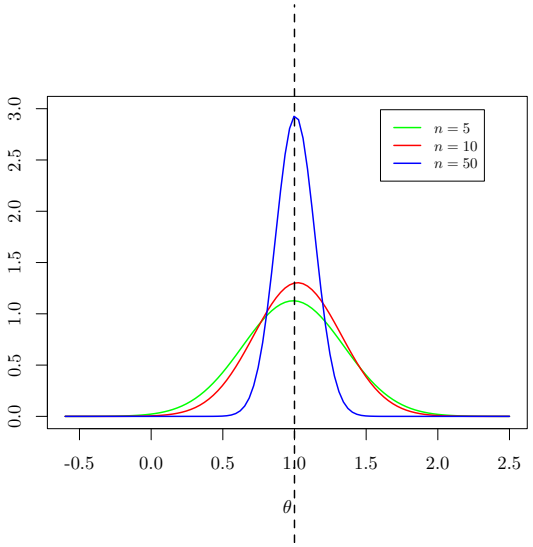


Figure 1.1 Density curves of the estimators for the mean of a normal distribution are illustrated for sample sizes $n = 5, 10$, and 50. The true parameter value $\theta_0 = 1$ is indicated by a vertical dashed line.

Assume that a series of data with density $f(\boldsymbol{x}|\boldsymbol{\theta}_0)$ was produced as independent, identically distributed random variables. As further information from the data is gathered, our initial understanding of $\boldsymbol{\theta}$ is gradually transformed into the posterior distribution. As the sample size increases, this updated knowledge about $\boldsymbol{\theta}$, as represented by its posterior distribution, should ideally become more concentrated around the true parameter value $\boldsymbol{\theta}_0$. This characteristic is an asymptotic phenomenon and is called the consistency of the posterior distribution at $\boldsymbol{\theta}_0$. In brief, it suggests that with more data, our uncertainty about the true parameter value diminishes, and our estimate becomes more accurate, eventually converging towards the true value.

Let $x_1, \ldots, x_n$ denote the observations at the nth stage, abbreviated as $\boldsymbol{x}_n$, with density $f(\boldsymbol{x}_n|\boldsymbol{\theta})$, where $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subset \mathbb{R}^d$. Let $\pi(\boldsymbol{\theta})$ be a prior density, $\pi(\boldsymbol{\theta}|\boldsymbol{x}_n)$ the posterior density defined in Equation (1.1.1), and $\Pi(\cdot|\boldsymbol{x}_n)$ the corresponding posterior distribution. In Bayesian statistics, $\pi(\boldsymbol{\theta}|\boldsymbol{x}_n)$ is consistent if and only if for every open neighborhood U of $\boldsymbol{\theta}_0$,

$$\pi(U^c|\boldsymbol{x}_n) \to 0 \text{ as } n \to \infty.$$

The consistency of $\pi(\boldsymbol{\theta}|\boldsymbol{x}_n)$ is characterized by the property that, for every open neighborhood U

of $\boldsymbol{\theta}_0$, the probability $\pi(U^c|\boldsymbol{x}_n)$ tends to 0 as $n \to \infty$, almost surely under the distribution specified by $\boldsymbol{\theta}_0$. Similarly, the posterior is deemed consistent if, for each open neighborhood U containing the true parameter, the posterior probability of the complement of U approaches 0 as the sample size n increases indefinitely. This indicates that, as the sample size grows, the posterior distribution converges to the true distribution governing the data generation process.

The concept originated with Laplace, who proved that if $x_1, \ldots, x_n$ are independent, identically distributed Bernoulli random variables and $\Pr(x_i = 1) = \theta$, and $\pi(\theta)$ is a continuous, positive prior density on the interval (0, 1), then for all θ_0 in the interval (0, 1), the posterior distribution is consistent. This idea supplements Bernoulli's weak law of big numbers, which functions as the first fundamental rule. It is referred to as the second basic law of large numbers by Von Mises (1981). Freedman (1963, 1965), and Diaconis and Freedman (1986) have highlighted the importance of posterior consistency.

According to the definition of convergence in distribution, the observation that $\Pi(\cdot|\boldsymbol{x}_n)$ converges, with probability 1 under $\boldsymbol{\theta}_0$, to the distribution concentrated at $\boldsymbol{\theta}_0$, is equal to the consistency of $\Pi(\cdot|\boldsymbol{x}_n)$ at $\boldsymbol{\theta}_0$. More generally, under some plausible conditions, the consistency of the posterior distribution holds for situations involving a parameter of finite dimensions. In particular, for a real-valued parameter $\boldsymbol{\theta}$, $\text{Var}(\boldsymbol{\theta}|\boldsymbol{x}_n) \to 0$ and $E(\boldsymbol{\theta}|\boldsymbol{x}_n) \to \boldsymbol{\theta}_0$ with a probability of one under $\boldsymbol{\theta}_0$ are required to be shown to be consistent at $\boldsymbol{\theta}_0$. The application of Chebyshev's inequality can demonstrate this.

Think about the case of Bernoulli-Beta that is discussed in Section 1.2.2. Let $x_1, \ldots, x_n$ be independent and identically distributed Bernoulli observations with $\Pr(x_i = 1) = \theta$ for all $i = 1, \ldots, n$. Assume that for θ, we use a Beta(α, β) prior density. Next, for any given $x_1, \ldots, x_n$, the posterior density of θ takes the form of a Beta$(\alpha + \sum_{i=1}^{n} x_i, \beta + n - \sum_{i=1}^{n} x_i)$, as described in Equation (1.2.4). The rule of large numbers implies that $\sum_{i=1}^{n} x_i/n \to \theta_0$ with a probability of 1

under P_{θ_0}, and

$$E(\theta|x_1, \ldots, x_n) \to \theta_0$$

$$\text{Var}(\theta|x_1, \ldots, x_n) \to 0$$

with a probability of 1 under θ_0. As such, the posterior distribution of θ is consistent with the result described in the previous paragraph.

We use the Bernoulli model to generate $n = 5, 10$, and 50 data points with a true parameter $\theta_0 = 0.4$. In other words, the probability of success is 0.4. We choose a prior distribution Beta$(4, 6)$, leading to a posterior distribution of Beta$(4 + \sum_{i=1}^{n} x_i, 6 + n - \sum_{i=1}^{n} x_i)$ according to Equation (1.2.4). As depicted in the scenario illustrated in Figure 1.2, with an increase in the sample size, the posterior distribution becomes increasingly concentrated around the true value, which, in this instance, is 0.4.

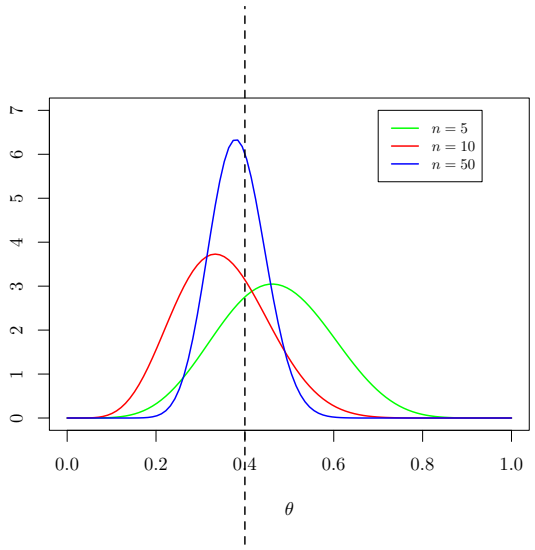


Figure 1.2 Density curves of the posterior distribution for a prior distribution Beta$(4, 6)$, are depicted for sample sizes $n = 5, 10$, and 50. The true parameter value $\theta_0 = 0.4$ is indicated by a vertical dashed line.

Another popular example is multinomial model with a Dirichlet prior. Let $\boldsymbol{X} \sim$ Multinomial$(n, \boldsymbol{\theta})$ where $\boldsymbol{\theta} = (\theta_0, \theta_1, \ldots, \theta_d)^\top$ be a $(d + 1)$-dimensional parameter $(d > 0)$. The multinomial model with a Dirichlet prior is a generalization of the Bernoulli model and Beta prior of the previous

example. The Dirichlet distribution for $d+1$ outcomes is the exponential family distribution on the $(d+1)$ dimensional probability is given by

$$\pi(\boldsymbol{\theta}) = \frac{\Gamma(\sum_{j=0}^{d} \alpha_j)}{\prod_{j=0}^{d} \Gamma(\alpha_j)} \prod_{j=0}^{d} \theta_j^{\alpha_j - 1},$$

where $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \ldots, \alpha_d)^\top$ is a non-negative vector of scaling coefficients, which are the parameters of the prior. For the multinomial model with $d+1$ outcomes, each observation of dimension $d+1$ has $\sum_{j=0}^{d} x_j = n$. The probability mass function for $\boldsymbol{X}|\boldsymbol{\theta}$ is defined as

$$f(\boldsymbol{x}|\boldsymbol{\theta}) = \frac{\Gamma(\sum_{j=0}^{d} x_j + 1)}{\prod_{j=0}^{d} \Gamma(x_j + 1)} \prod_{j=0}^{d} \theta_j^{x_j}.$$

Then the posterior satisfies

$$\pi(\boldsymbol{\theta}|\boldsymbol{x}) \propto \prod_{j=0}^{d} \theta_j^{x_j + \alpha_j - 1}.$$

We see that the posterior is also a Dirichlet distribution: Dirichlet$(\boldsymbol{\alpha} + \boldsymbol{x})$.

The posterior mean of a multinomial with Dirichlet prior is

$$E(\boldsymbol{\theta}|\boldsymbol{x}) = \left(\frac{\alpha_0 + x_0}{\sum_{j=0}^{d} \alpha_j + n}, \ldots, \frac{\alpha_d + x_d}{\sum_{j=0}^{d} \alpha_j + n} \right)^\top.$$

The posterior mean can be viewed as smoothing out the maximum likelihood estimate by allocating some additional probability mass to low frequency observations.

An illustration of this example is shown in Figures 1.3 to 1.5. We adopt a prior Dirichlet$(6, 6, 6)$. The ternary contour plot for the prior is shown below

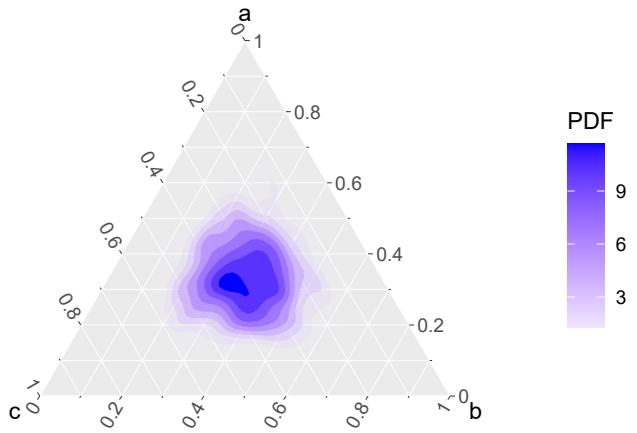


Figure 1.3 Contour plot of the prior distribution: Dirichlet$(6, 6, 6)$.

We use the multinomial model to generate data points with parameter $\boldsymbol{\theta} = (0.2, 0.3, 0.5)^\top$. The contours of the likelihood and posterior with $n = 30$ observed data are shown in Figure 1.4.

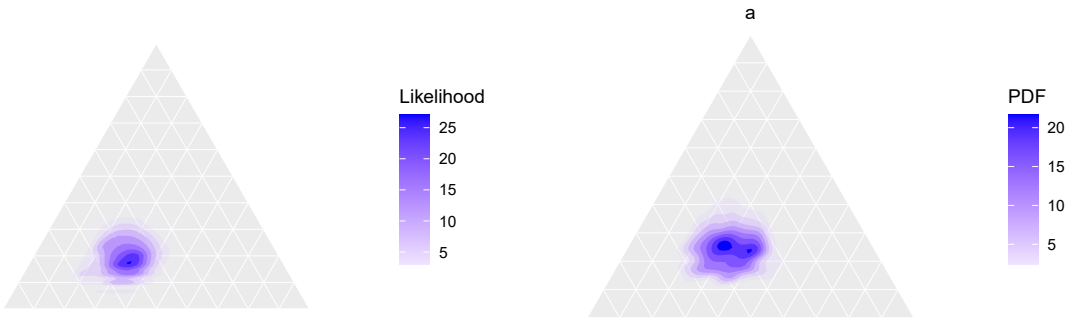


Figure 1.4 Contour plots of likelihood for Multinomial$(30, [0.2, 0.3, 0.5]^\top)$ and posterior distribution with Dirichlet prior (6, 6, 6).

As a comparison, we also provide the contour of the posterior with $n = 300$ observed data in Figure 1.5. From this experiment, we see that when the number of observed data is small, the posterior is affected by both the prior and the likelihood; when the number of observed data is large, the posterior is mainly dominated by the likelihood.

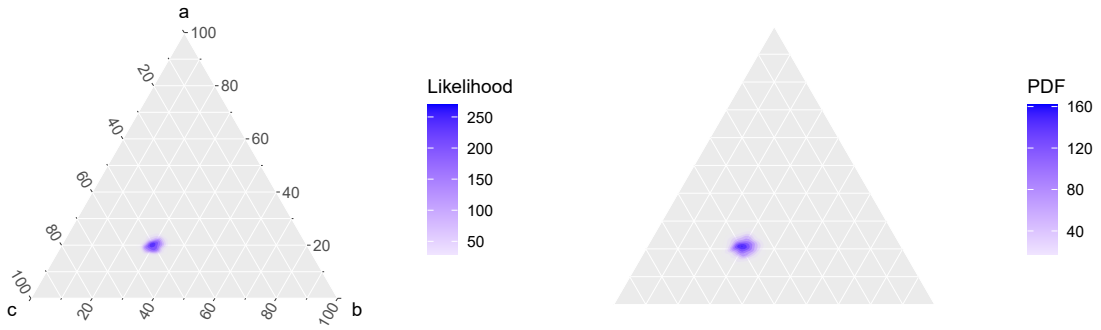


Figure 1.5 Contour plots of likelihood for Multinomial$(300, [0.2, 0.3, 0.5]^\top)$ and posterior distribution with Dirichlet prior (6, 6, 6).

In the previous two examples, the prior is a Dirichlet distribution and the posterior is also a Dirichlet. When a posterior distribution belongs to the same family of distributions as the prior, we say that the prior is conjugate with respect to the model. Conjugate priors offer several advantages in Bayesian inference. Firstly, they provide analytical simplicity by yielding closed-form solutions for the posterior distribution, streamlining computational processes. This simplicity enhances computational efficiency, particularly for large datasets, as it reduces the time and resources needed for inference. Additionally, the interpretability of conjugate priors facilitates intuitive understanding and communication of results, as the posterior distribution belongs to the same family as the prior. Moreover, conjugate priors are valuable in teaching and learning Bayesian concepts due to their straightforward nature, aiding in the comprehension of fundamental principles. They also offer insight into prior selection, allowing practitioners to incorporate prior beliefs or knowledge effectively. Furthermore, conjugate priors enable sensitivity analysis by varying prior parameters, providing insights into the robustness of conclusions. It will be discussed further in Chapter 3.

The robustness of posterior inference with respect to prior selection is a significant consistency finding. Assume that $x_1, \ldots, x_n$ are independent observations with the same distribution. To ensure that the resulting posterior distributions, $\Pi_1(\cdot|\boldsymbol{x}_n)$ and $\Pi_2(\cdot|\boldsymbol{x}_n)$, are consistent at $\boldsymbol{\theta}_0$, let π_1 and π_2 indicate two prior densities, both of which are positive and continuous at $\boldsymbol{\theta}_0$, an interior

point of $\boldsymbol{\Theta}$. At that point, under $\boldsymbol{\theta}_0$, with probability 1,

$$\int_{\Theta} |\pi_1(\boldsymbol{\theta}|\boldsymbol{x}_n) - \pi_2(\boldsymbol{\theta}|\boldsymbol{x}_n)| \, d\boldsymbol{\theta} \to 0,$$

or equivalently, for any measurable set $A \in \boldsymbol{\Theta}$,

$$\sup_{A} |\Pi_1(A|\boldsymbol{x}_n) - \Pi_2(A|\boldsymbol{x}_n)| \to 0.$$

As a result, almost identical posterior distributions result from varying choices of prior distributions. A formal demonstration of this result is provided in Ghosh, Ghosal, and Samanta (1994).

CHAPTER 2 BERNSTEIN-VON MISES THEOREM

2.1 Convergence in Total Variation Distance

In the realm of probability theory, the total variation distance serves as a metric for quantifying the disparity between two probability distributions. Convergence in total variation signifies that as a sequence of random variables or stochastic processes progresses, the distribution associated with these variables or processes gradually approaches a specified target distribution. This convergence is characterized by the total variation distance between the evolving distribution and the target distribution diminishing to zero.

This concept of convergence is frequently employed to characterize the behavior of random variables or stochastic processes as they approach a particular limiting behavior or distribution. It represents a more stringent form of convergence compared to convergence in probability or convergence in distribution.

For two probability measures P and Q defined on $(\mathcal{X}, \mathcal{A})$. Suppose that ν is a σ-finite measure on $(\mathcal{X}, \mathcal{A})$ satisfying $P \ll \nu$ and $Q \ll \nu$. The total variation distance is defined as (Tsybakov, 2008)

$$\left\|P - Q\right\|_{TV} = \sup_{A \in \mathcal{A}} \left|P(A) - Q(A)\right|. \tag{2.1.1}$$

By Scheffé's lemma,

$$\|P - Q\|_{TV} = \frac{1}{2} \int \left|p(\boldsymbol{\theta}) - q(\boldsymbol{\theta})\right| d\nu(\boldsymbol{\theta}),$$

where p and q are densities of P and Q. Consider an example of the total variation distance between two high-dimensional normal distributions $\mathcal{N}(\mu_1, \Sigma_1)$ and $\mathcal{N}(\mu_2, \Sigma_2)$. Then refer to the Proposition 2.1 (Devroye, Mehrabian, and Reddad, 2023), if Σ_1 and Σ_2 are positive definite, we

have

$$\left\|\mathcal{N}(\mu_1,\Sigma_1)-\mathcal{N}(\mu_2,\Sigma_2)\right\|_{TV}\leq\frac{1}{2}\sqrt{\operatorname{tr}(\Sigma_1^{-1}\Sigma_2-I)+(\mu_1-\mu_2)^\top\Sigma_1^{-1}(\mu_1-\mu_2)-\log\det(\Sigma_2\Sigma_1^{-1})}.$$

This inequality establishes an upper bound for the total variation distance suing the Hellinger distance, which will be discussed on later with more detail.

A sequence of random variables converges in total variation to a variable X if

$$\sup_B|\Pr(X_n\in B)-\Pr(X\in B)|\to 0,$$

where the supremum is taken over all measurable sets B. According to the Portmanteau lemma, this form of convergence is stronger than convergence in distribution. Not only is it necessary for the sequence $\Pr(X_n\in B)$ to converge for every Borel set B, but the convergence must also be uniform across all B. A straightforward condition for convergence in total variation is the pointwise convergence of densities. If X_n and X possess densities p_n and p relative to a measure ν, then

$$\sup_B|\Pr(X_n\in B)-\Pr(X\in B)|=\frac{1}{2}\int|p_n-p|\,d\nu.$$

Hence, convergence in total variation can be established using convergence theorems for integrals from measure theory.

The Kullback-Leibler divergence serves as a metric to quantify the dissimilarity between two probability distributions. For a pair of probability measures P and Q defined on a common probability space, the Kullback-Leibler divergence from Q to P is expressed as

$$D_{\mathrm{KL}}(P||Q)=\int\log\left(\frac{dP}{dQ}\right)dP,$$

where dP/dQ denotes the Radon-Nikodym derivative. It measures the information gain or loss

incurred when P is approximated by Q.

The Kullback-Leibler divergence quantifies the average disparity in log-likelihood ratios between corresponding events, offering insight into the overall information gain or loss when one distribution is approximated by another. Convergence in Kullback-Leibler divergence signifies a reduction in the relative entropy between the distributions, indicating a convergence in their informational content.

The Hellinger distance between P and Q is defined as follows

$$H(P,Q) = \left(\int (\sqrt{p} - \sqrt{q})^2 d\nu \right)^{1/2}.$$

Hellinger distance is sensitive to differences in the shape and spread of distributions.

In contrast, total variation distance focuses on the maximal difference in probabilities assigned to identical events by two distributions, serving as a measure of the most significant potential discrepancy between them. Convergence in total variation implies that the distributions approach each other closely in terms of their greatest possible difference.

The relationship between total variation distance and Kullback-Leibler divergence is elucidated by Pinsker's inequality, first proposed in Pinsker (1964), expressed as

$$\left\| P - Q \right\|_{TV} \leq \sqrt{\frac{1}{2} D_{KL}(P \| Q)}.$$

Le Cam's inequalities in Cam (1960) shows the link between total variation distance and Hellinger distance,

$$\frac{1}{2} H^2(P,Q) \leq \left\| P - Q \right\|_{TV} \leq H(P,Q) \sqrt{1 - \frac{H^2(P,Q)}{4}}.$$

2.2 Bayesian vs. Frequentist in Asymptotic Normality

In statistical inference, Bayesian and Frequentist methodologies represent two distinct frameworks employed for drawing conclusions about population parameters. Asymptotic nor-

mality, often examined in the context of large sample sizes, holds significance in both approaches.

In Frequentist statistics, the central limit theorem stands as a cornerstone concept concerning asymptotic normality. It asserts that given a sufficiently large sample from any distribution with finite mean and variance, the distribution of the sample mean will tend towards a normal distribution. Frequentist methodologies commonly utilize large-sample approximations, relying on the asymptotic normality of estimators to construct confidence intervals and perform hypothesis tests effectively.

In contrast, Bayesian statistics places less emphasis on the notion of asymptotic normality relative to Frequentist statistics. Nonetheless, in practical applications, the posterior distribution for certain parameters may exhibit asymptotic normality under specific conditions. Bayesian approaches center on characterizing the entire posterior distribution rather than solely focusing on point estimates. Through MCMC techniques, prevalent in Bayesian analysis, posterior samples can be obtained, with the posterior distribution converging towards normality as the sample size increases.

2.2.1 Central Limit Theorem

During the interwar period, modern probability theory emerged as a distinct mathematical subdiscipline, characterized by the development of foundational concepts, fundamental theorems, and methodological frameworks. This evolution was marked by the synthesis of various subfields, including axiomatics encompassing elements of measure theory, robust laws of large numbers, stochastic processes, and limit theorems governing the distributions of sums of random variables. Initially, these subfields were loosely connected under the overarching label of "probability". Among these, the domain of limit theorems stood out as it had made notable contributions during the 18th and 19th centuries, assuming a pivotal role in the transition from classical to modern probability theory. Consider a sequence of random samples denoted as $x_1, \ldots, x_n$ drawn from a population with a finite expected value $E(x_i) = \mu < \infty$ and a finite variance

$0 < \text{Var}(x_i) = \sigma^2 < \infty$. The random variable z_n, defined as

$$z_n = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{x_1 + x_2 + \ldots + x_n - n\mu}{\sqrt{n}\sigma}$$

converges in distribution to the standard normal random variable as $n \to \infty$, meaning that

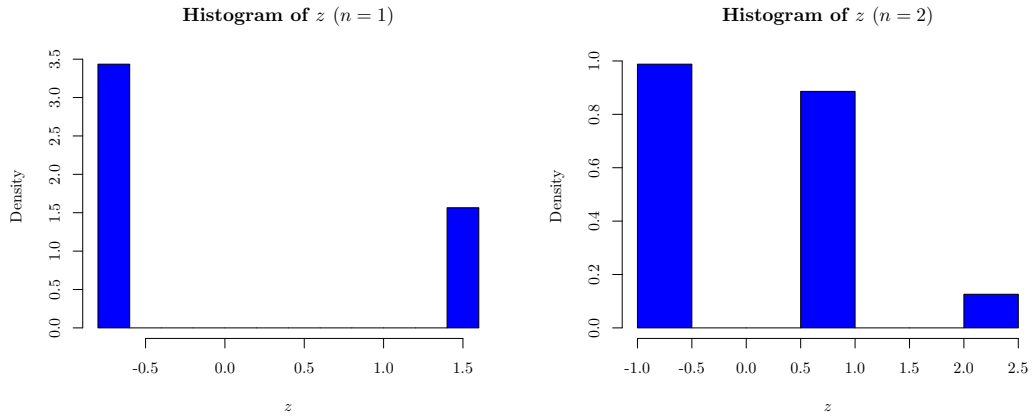
$$\lim_{n\to\infty} P(z_n \leq x) = \Phi(x)$$

holds for all $x \in \mathbb{R}$, where $\Phi(x)$ represents the standard normal cumulative distribution function.

An intriguing aspect of the central limit theorem is its independence from the underlying distribution of the random variables x_i's. The theorem holds regardless of the x_i's have a discrete, continuous, or mixed distribution. In order to better understand the central limit theorem, let us examine a few examples. Assume that a Bernoulli distribution with parameter p dominates the x_i's values. Following that, $\text{Var}(x_i) = p(1-p)$ and $E(x_i) = p$. In addition, the sum $y_n = x_1 + x_2 + \cdots + x_n$ is represented by $y_n \sim \text{Binomial}(n, p)$, which is a Binomial distribution with parameters n and p. Consequently,

$$z_n = \frac{y_n - np}{\sqrt{np(1-p)}},$$

where $y_n \sim \text{Binomial}(n, p)$.



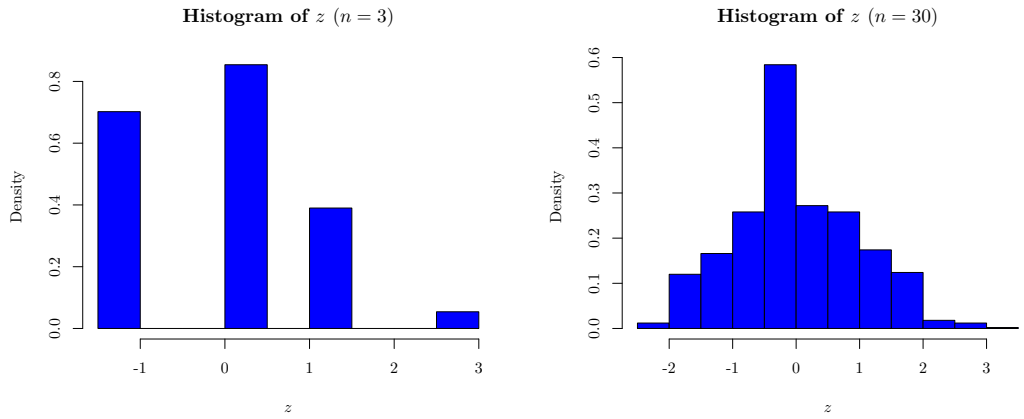


Figure 2.1 Histograms of z_n when $n = 1, 2, 3$ and 30, and $p = 0.3$

The probability mass function of z_n for a range of n values is shown in Figure 2.1. As n increases, the shape of the probability mass function progressively approaches a normal probability density function curve. Interestingly, since z_n is a discrete random variable, its probability mass function is what matters instead than its probability density function. As a result, the central limit theorem states that z_n's cumulative distribution function (also known as the CDF) converges to the conventional normal CDF. However, the image helps to visualize the convergence to a normal distribution because of their conceptual closeness.

Now, let us examine an alternative situation in which the x_i's are selected from a Uniform$(0, 1)$ distribution. Here, $E(x_i) = 1/2$ and Var$(x_i) = 1/12$ are the values we have.

$$z_n = \frac{x_1 + x_2 + \cdots + x_n - n/2}{\sqrt{n/12}}.$$

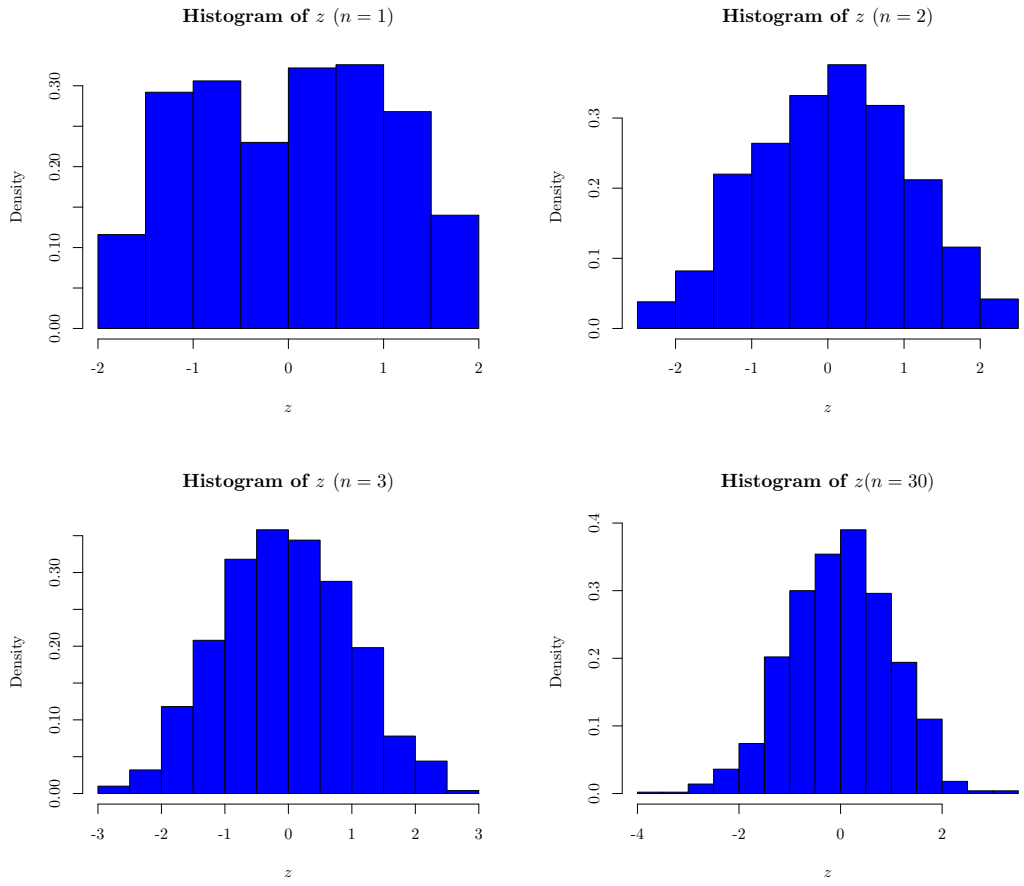


Figure 2.2 Histograms of z_n when $n = 1, 2, 3$ and 30

When the x_i's are drawn from a Uniform$(0, 1)$ distribution, the probability density function of z_n for a range of values of n is shown in Figure 2.2. As n increases, the shape of the probability density function gradually gets close to a normal probability density function curve.

A direct analysis of the sum $y_n = x_1 + x_2 + \cdots + x_n$ would have been possible. Why then should we adjust it first and say that z_n becomes roughly normal after normalization? This makes sense since as n gets closer to infinity, both the variance and mean of z_n, represented by $\text{Var}(z_n) = n\sigma^2$ and $E(z_n) = nE(x_i)$, respectively, tend to infinity. In order to correct this, we normalize z_n such that $E(z_n) = 0$ and $\text{Var}(z_n) = 1$, the mean and variance, respectively, are finite. On the other hand, scaling and shifting can be used to derive the cumulative distribution function of z_n from that of y_n for any fixed n. As a result, the forms of the two cumulative distribution functions are comparable.

The central limit theorem is of ultimate importance due to its relevance in numerous practical scenarios, where a particular random variable of interest arises as the summation of a large number of independent random variables. This theorem provides a solid foundation for justifying the use of the normal distribution in such cases. Such random variables are pervasive across various disciplines, highlighting their broad applicability. For instance, laboratory measurement errors are frequently modeled as normal random variables. Similarly, in communication and signal processing, Gaussian noise is commonly employed as a model for noise. Additionally, in finance, percentage changes in asset prices are sometimes represented by normal random variables. Moreover, when conducting random sampling from a population to extract statistical insights, the resulting quantity is often viewed as a normal random variable.

When many independent, identically distributed random variables are added together, the central limit theorem provides a large computing advantage. Take, for example, a case where the total of a thousand independent and identically distributed random variables is of interest. Finding the distribution of this total directly may out to be extremely difficult, if not impossible. But if we know the mean and variance of each individual x_i, we can quickly determine the distribution by applying the central limit theorem.

A frequently asked question concerns the suitability of the normal approximation and the necessary sample size n. The distribution properties of the x_i's usually determine the response. However, a widely used heuristic indicates that the normal approximation is typically very accurate if n is more than or equal to 30.

2.2.2 Bernstein-von Mises Theorem

The normal distribution is widely used in large sample Bayesian methods to approximate the posterior distribution of $\boldsymbol{\theta}$. When n is large enough, the posterior distribution tends to normality under certain regularity requirements as sample size increases. This allows for an efficient approximation by a suitable normal distribution. The posterior distribution is more tightly packed in a smaller area around the posterior mode as n increases. This posterior distribution mode can be represented by the notation $\tilde{\boldsymbol{\theta}}_n$. A Taylor expansion of $\log \pi(\boldsymbol{\theta}|\boldsymbol{X}_n)$ at $\tilde{\boldsymbol{\theta}}_n$ yields, under appropriate

regularity conditions

$$\begin{aligned}\log \pi(\boldsymbol{\theta}|\boldsymbol{X}_n) &= \log \pi(\tilde{\boldsymbol{\theta}}_n|\boldsymbol{X}_n) + (\boldsymbol{\theta}-\tilde{\boldsymbol{\theta}}_n)' \frac{\partial}{\partial \boldsymbol{\theta}} \log \pi(\boldsymbol{\theta}|\boldsymbol{X}_n)|_{\tilde{\boldsymbol{\theta}}_n} - \frac{1}{2}(\boldsymbol{\theta}-\tilde{\boldsymbol{\theta}}_n)' \tilde{\boldsymbol{I}}_n (\boldsymbol{\theta}-\tilde{\boldsymbol{\theta}}_n) + \cdots \\ &\approx \log \pi(\tilde{\boldsymbol{\theta}}_n|\boldsymbol{X}_n) - \frac{1}{2}(\boldsymbol{\theta}-\tilde{\boldsymbol{\theta}}_n)' \tilde{\boldsymbol{I}}_n (\boldsymbol{\theta}-\tilde{\boldsymbol{\theta}}_n),\end{aligned} \tag{2.2.1}$$

where $\tilde{\boldsymbol{I}}_n$ is a $d \times d$ matrix defined as

$$\tilde{\boldsymbol{I}}_n = \left(-\frac{\partial^2}{\partial \theta_i \partial \theta_j} \log \pi(\tilde{\boldsymbol{\theta}}_n|\boldsymbol{X}_n) \right) \Bigg|_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}}_n}.$$

The local curvature of the log posterior density at the posterior mode $\tilde{\boldsymbol{\theta}}_n$ is characterized by the matrix incorporating second derivatives, which is called the generalized observed Fisher information matrix. The expression is made simpler by the first derivative term, which vanishes at the mode $\tilde{\boldsymbol{\theta}}_n$. Furthermore, because of $\boldsymbol{\theta}$'s close proximity to the mode, higher-order derivative terms asymptotically become insignificant as $\boldsymbol{\theta}$ gets closer to $\tilde{\boldsymbol{\theta}}_n$. As a function of $\boldsymbol{\theta}$, the posterior $\pi(\tilde{\boldsymbol{\theta}}_n|\boldsymbol{X}_n)$ can be approximated by a density proportional to $\exp\left[-\frac{1}{2}(\boldsymbol{\theta}-\tilde{\boldsymbol{\theta}}_n)' \tilde{\boldsymbol{I}}_n (\boldsymbol{\theta}-\tilde{\boldsymbol{\theta}}_n)\right]$. This is because the first term in Equation (2.2.1) is independent of θ. The distribution is similar to a $\mathcal{N}_d(\tilde{\boldsymbol{\theta}}_n, \tilde{\boldsymbol{I}}_n^{-1})$ distribution, in which d denotes the dimension of $\boldsymbol{\theta}$.

The posterior density $\pi(\boldsymbol{\theta}|\boldsymbol{X}_n)$ approximates the likelihood $f(\boldsymbol{X}_n|\boldsymbol{\theta})$ when the posterior distribution becomes extremely concentrated in a narrow neighborhood around the posterior mode $\tilde{\boldsymbol{\theta}}_n$, where the prior density $\pi(\boldsymbol{\theta})$ remains almost constant. Thus, in the previously mentioned setting, we can replace $\tilde{\boldsymbol{\theta}}_n$ with the maximum likelihood estimate $\hat{\boldsymbol{\theta}}_n$ and $\tilde{\boldsymbol{I}}_n^{-1}$ with the observed Fisher information matrix

$$\boldsymbol{I}_n = \left(-\frac{\partial^2}{\partial \theta_i \partial \theta_j} \log f(\boldsymbol{X}_n|\boldsymbol{\theta}) \right) \Bigg|_{\hat{\boldsymbol{\theta}}_n}.$$

This means that the posterior distribution of $\boldsymbol{\theta}$ is roughly $\mathcal{N}_d(\hat{\boldsymbol{\theta}}_n, \boldsymbol{I}_n^{-1})$.

Consequently, we arrive at the following deduction: Assume that X_n represents the set

of i.i.d. observations, where $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subset \mathbb{R}^d$. Let $X_1, X_2, \ldots, X_n$ be these observations. Assume that, according to Equation (1.1.1), the posterior density is $\pi(\boldsymbol{X}_n)$ and the prior density is $\pi(\boldsymbol{\theta})$. Let $\tilde{\boldsymbol{\theta}}_n$ be the posterior mode and $\hat{\boldsymbol{\theta}}_n$ be the maximum likelihood estimator. Also, consider the previously mentioned forms of the Fisher information matrix, denoted as $\tilde{\boldsymbol{I}}_n$ and $\boldsymbol{I}_n$, evaluated at the posterior mode and maximum likelihood estimator, respectively. Then, for large n, any of the normal distributions, $\mathcal{N}_d(\tilde{\boldsymbol{\theta}}_n, \tilde{\boldsymbol{I}}_n^{-1})$ or $\mathcal{N}_d(\hat{\boldsymbol{\theta}}_n, \boldsymbol{I}_n^{-1})$, can approximate the posterior distribution of $\boldsymbol{\theta}$, given appropriate regularity conditions.

In particular, the true data-generating model approaches $\mathcal{N}_d(\mathbf{0}, I)$ with probability 1 under the posterior distribution of $\boldsymbol{I}^{1/2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_n)$, conditioned on $\boldsymbol{X}_n$. Here, I denotes the identity matrix of size p. With repeated sampling, the distribution of $\boldsymbol{I}^{1/2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_n)$, given $\boldsymbol{\theta}$, also tends to $\mathcal{N}_d(\mathbf{0}, I)$, according to this convergence, which is consistent with conclusions drawn from classical statistical theory.

The classical Bernstein-von Mises theorem, often referred to as the Bayesian central limit theorem, states the asymptotic behavior of the posterior distribution in Bayesian statistics.

Theorem 2.1. *Under certain regularity conditions,*

$$\mathbf{E}\left\|\mathcal{N}\left(\hat{\boldsymbol{\theta}}_n, \boldsymbol{I}_n^{-1}\right) - \pi(\boldsymbol{\theta}|\boldsymbol{X}_n)\right\|_{TV} \to 0 \text{ as } n \to \infty.$$

The asymptotic posterior normality in Theorem 2.1 is in terms of the convergence mode of expected total variation distance with respect to the posterior distribution. In simpler terms, the theorem suggests that when we have a large amount of data, the posterior distribution becomes approximately normal, centered around the maximum likelihood estimate of the parameter, and with a variance that reflects the uncertainty in our parameter estimation process.

This theorem has profound implications in Bayesian inference, as it allows practitioners to make probabilistic statements about the parameters of interest based on their observed data, leveraging the asymptotic normality of the posterior distribution.

2.3 Literature Review on the Bernstein-von Mises Theorem

Under specific assumptions on the statistical models and the prior, it is well known that increasing the sample size washes away the influence of the prior distribution, leaving the likelihood function as the sole determinant. As a result, Bayesian approaches approximately agree with likelihood-based frequentist approaches as the sample size grows. This result has been verified in regular smooth parametric models, such as linear regression models, through the so-called Bernstein-von Mises theorem. The Bernstein-von Mises theorem states that under adequate conditions on the prior, the posterior distribution asymptotically converges to a Gaussian distribution with the mean, maximum likelihood estimator, and variance, the inverse of the observed Fisher information matrix. The asymptotic posterior normality allows us to construct approximate credible sets for $\boldsymbol{\theta}$, especially when sampling from the posterior distribution is challenging. Benefiting from the alignment between frequentist and Bayesian approaches, these credible sets can act as valid frequentist confidence intervals Giné and Nickl (2021).

The Bernstein-von Mises theorem has been broadly studied in the growing dimension (Bontemps, 2011; Boucheron and Gassiat, 2009; Ghosal, 1999, 2000; Ghosal, Ghosh, and Van Der Vaart, 2000; Johnstone, 2010), semiparametric (Bickel and Kleijn, 2012; Castillo, 2012; Shen, 2002; Rivoirard and Rousseau, 2012) and nonparametric (Castillo and Nickl, 2013, 2014; Leahu, 2011; Ray, 2017) frameworks. Inevitably, there is substantial overlap among the scopes of these frameworks. Le Cam and Yang (2000) offer a rigorous proof for the Bernstein-von Mises theorem, which holds true under the assumption of a parametric i.i.d. scenario. Van der Vaart (2000) provides explicit proof of the parametric Bernstein-von Mises theorem under remarkably weak conditions that the differentiability in quadratic mean and the existence of a sequence of uniformly consistent tests. Similarly, according to Bickel and Kleijn (2012), the semiparametric Bernstein-von Mises theorem requires an additional condition on a parametric convergence rate under the premise of general conditions, such as differentiability. They investigate the effectiveness of Bayesian point estimators utilizing Hájek's convolution theorem and extend these findings to the estimation of linear coefficients in partial linear regression scenarios with a Gaussian prior. This theorem, under

specific conditions, demonstrates that the sequence of marginal posteriors converges to a normal distribution, offering valuable insights into semiparametric estimation. Their study delves into the convergence of Bayesian point estimators within the context of semiparametric estimation problems, establishing the necessary conditions for the application of the Bernstein-von Mises theorem in such settings. In addition to exploring the conditions of asymptotic posterior normality, some preceding works develop the semiparametric Bernstein-von Mises theorem in several situations. Rivoirard and Rousseau (2012) present the posterior consistency of linear functionals of the density within the framework of infinite-dimensional exponential families. Their work investigates the asymptotic posterior distribution of these linear functionals and establishes conditions for a semiparametric version of the Bernstein-von Mises theorem. The study sheds light on both positive and negative phenomena that may emerge during the analysis of Bernstein-von Mises results, with a specific focus on the challenges and insights posed by infinite-dimensional exponential families. Additionally, they underscore the significance of defining a change of parameter and examines the influence of different types of priors on the theorem's applicability. In Castillo (2012), two semiparametric Bernstein-von Mises theorems with Gaussian process priors are established, contingent upon whether the efficient information aligns with the information in the associated parametric model. His work delves into the realm of Bayesian estimation, particularly emphasizing the semiparametric Bernstein-von Mises theorem. The focus is on estimating the parameters (θ, f), where θ represents the parameter of interest, and f is an infinite-dimensional nuisance parameter. The study explores the incorporation of Gaussian process priors and provides application scenarios for the theorems, including instances such as estimating the center of symmetry in Gaussian white noise.

As highlighted by Cox (1993) and Freedman (1999), certain nonparametric priors, seemingly natural and innocuous, may lead to posterior inconsistency. In the realm of Bayesian inference for nonparametric regression models, Cox (1993) investigates an observation model where the response variable is a smooth function of a covariate, characterized by unknown parameters and Gaussian prior distributions. The study rigorously examines estimation errors, providing asymp-

totic posterior and sampling distributional approximations. It further explores topics such as the coverage probability of posterior probability regions and continuous-time signal estimation problems. Emphasizing the versatility and accuracy of Bayesian methods in addressing statistical inference problems, the author underscores their role in providing flexible solutions. Freedman (1999) contributes to the discourse by discussing the Bernstein-von Mises theorem within the context of infinite-dimensional parameters. He illuminates the distinctions between Bayesian and frequentist approaches in statistical modeling, underscoring the significance of employing smooth, finite-dimensional models for precise estimation. His research delves into the implications of the theorem for confidence intervals and coverage properties in statistical inference, offering insights into the challenges and considerations inherent in the infinite-dimensional scenario.

However, the extension of the Bernstein-von Mises theorem to growing or increasing dimension settings has received significant attention in the past two decades. To achieve asymptotic posterior normality in increasing-dimensional linear regression models, Ghosal (1999) suggests imposing the constraint on the parameter dimension, specifically that "$d_n^4 \ln(d_n)/n$ is small". In his work, he delves into the asymptotic normality of posterior distributions within the context of high-dimensional linear models. The study is centered on investigating the consistency and asymptotic behavior of posterior distributions as the parameter dimension experiences growth. Key outcomes of the research include establishing conditions for consistency and asymptotic normality, exploring implications for statistical inference in high-dimensional settings, and examining the use of prior distributions to achieve desired properties. Ghosal (2000) subsequently establishes that the growth rate on parameter dimension, specifically "$d_n^3 \ln(d_n)/n$ is small" leads to the asymptotic convergence of the posterior distribution of the natural parameter for an exponential family to a Gaussian distribution. This work delves into the consistency and asymptotic normality of posterior distributions in the context of exponential families as the number of parameters approaches infinity. The study explores conditions regarding the growth of the parameter dimension for the posterior distributions to concentrate around the true parameter. Additionally, it touches upon the approximation of posterior distributions with normal distributions, especially for exponential families with a large

number of parameters.

The reference priors examined by Clarke and Ghosal (2010) are grounded in independent and identically distributed data within an exponential family, while the entropy estimation investigated by Boucheron and Gassiat (2009) is specifically applied to families of discrete distributions. Clarke and Ghosal explore reference priors in exponential families with increasing dimensions, scrutinizing the asymptotic properties of the posterior distribution and delving into the Shannon mutual information. The authors concentrate on identifying optimal rates of parameter growth to ensure asymptotic normality, emphasizing the significance of the expected Kullback-Leibler distance in exponential families. Additionally, they showcase how Jeffreys' prior can be derived as the reference prior by optimizing certain terms in the asymptotic expansion. In a related vein, Boucheron and Gassiat introduce a Bernstein-von Mises Theorem for discrete probability distributions, with a specific focus on the asymptotic normality of the posterior distribution as the model dimension grows with the sample size. This theorem carries implications for Bayesian estimators of Shannon and Rényi entropies, shedding light on the convergence properties of the posterior distribution to a Gaussian distribution.

Within a nonparametric framework, Bontemps (2011) demonstrates that the convergence rate on parameter dimension, specifically "$d_n \ln(d_n)/n$ is small" holds in Gaussian linear regression models with an increasing number of regressors, providing clarification on an earlier proposition by Ghosal (1999). Bontemps examines Bernstein-von Mises Theorems for Gaussian regression with a growing number of regressors. The study delves into the asymptotic normality of the posterior distribution in Gaussian linear regression models as the number of regressors expands with the sample size. It applies these theorems to the Gaussian sequence model and regression of functions in Sobolev and C^α classes, emphasizing the crucial aspect of adaptivity for Bayesian estimators of functionals across diverse applications.

In the case of i.i.d. data, the results of the Bernstein-von Mises theorem are applicable to any smooth parametric family, provided the condition "d_n^3/n is small" as discussed in Spokoiny (2013). This work delves into the Bernstein-von Mises theorem within the context of expanding

parameter dimensions, addressing challenges such as model misspecification and small sample sizes. It builds upon classical results, including the Fisher and Wilks Theorems, extending insights into non-asymptotic frameworks and Bayesian procedures. The study offers explicit bounds and expansions, providing a clearer understanding of the behavior of estimators and excess functions. Overall, it makes a significant contribution to advancing our comprehension of statistical theory in complex settings.

Some positive Bernstein-von Mises results in the Gaussian white noise model, Gaussian nonparametric regression, and i.i.d. sampling model in the same spirit are obtained in (Castillo and Nickl, 2013, 2014). They dig into nonparametric Bernstein-von Mises theorems within the context of the Gaussian white noise model. The study illustrates how these theorems validate the use of Bayesian methods as efficient frequentist inference procedures for various nonparametric problems. The document encompasses the construction of Bayesian credible sets with precise frequentist coverage levels and shrinking L^2-diameter. It explores applications to linear and nonlinear functionals, credible bands for auto-convolutions, and considers nonconjugate product priors defined on orthonormal bases of L^2. The results underscore the robust performance of Bayesian methods in nonparametric settings. In 2014, Castillo and Nick extend their exploration of the Bernstein-von Mises phenomenon to nonparametric Bayesian procedures, with a specific focus on Gaussian nonparametric regression and i.i.d. sampling models. The study introduces multiscale spaces for defining nonparametric priors and posteriors, emphasizing the alignment of posterior-based inference with efficient frequentist procedures. Insights into the application of Bernstein-von Mises theorems in nonparametric settings are provided, along with practical implications, including applications to Donsker- and Kolmogorov-Smirnov theorems for random posterior cumulative distribution functions.

A substantial portion of existing research on the nonparametric Bernstein-von Mises theorem revolves around addressing constraints on the parameter dimension to ensure posterior consistency. The primary challenge lies in meeting the rigorous condition imposed on the growth rate of the parameter dimension, limiting the broader applicability of the nonparametric Bernstein-von

Mises theorem. Classical results by (Portnoy, 1984, 1985, 1988) establish certain constraints on the parameter dimension to ensure consistency and asymptotic normality of the M-estimator in regression models. The extension of Schwartz's theorem has facilitated the derivation of asymptotic posterior normality in various setups, leading to a relaxation of restrictions on the parameter dimension to some extent. This is evident in works such as (Ghosal, 1999, 2000) concerning linear regression and exponential family models, as well as Bontemps (2011) in the context of regression models. Despite these advancements, the current growth rate of the parameter dimension still poses limitations on the applicability of the nonparametric Bernstein-von Mises theorem, considering computational demands and algorithmic criteria. This limitation prompts the exploration of further relaxation in the growth rate of the parameter dimension, particularly in exponential family models, through the lens of a nonparametric Bernstein-von Mises theorem. In this context, we choose the Diaconis-Ylvisaker prior due to its conjugate properties. The Diaconis-Ylvisaker prior belongs to a family of conjugate priors designed for the natural parameter of an exponential family. This work aims to establish sufficient conditions under which the asymptotic normality of the posterior distribution, utilizing the Diaconis-Ylvisaker prior, becomes achievable, thereby making the increasing-dimensional Bernstein-von Mises result attainable.

The primary focus of our endeavor in this work is to ascertain an efficient growth rate for the parameter dimension, offering enhanced insights into the underlying problem while remaining practically applicable across a spectrum of diverse applications. This pursuit aims to contribute a nuanced understanding of the parameter dimension's expansion, shedding light on optimal rates that balance computational feasibility and theoretical efficacy.

A significant facet of our work involves extending and refining the existing literature by establishing asymptotic posterior normality with the Diaconis-Ylvisaker prior under the assumption that "d_n/n is small" (Jin, Bhattacharya, and Ghosh, 2024). This extension serves as a valuable enhancement to the current body of knowledge, providing a more encompassing perspective on the behavior of the posterior distribution under varying conditions. By incorporating the Diaconis-Ylvisaker prior, we aim to offer a comprehensive and nuanced understanding of the asymptotic

behavior in scenarios where the ratio of the parameter dimension to the sample size remains small. Furthermore, this works makes a noteworthy technical contribution by developing a general result that bounds the tail probability of the quadratic form of the maximum likelihood estimator. Importantly, this contribution is achieved without relying on the sub-Gaussianity assumption, thereby broadening the applicability and robustness of our findings. The exploration of tail probabilities offers valuable insights into the behavior of the maximum likelihood estimator, contributing to a more comprehensive understanding of its statistical properties. This endeavor adds a layer of sophistication to the existing methodologies, providing researchers and practitioners with a versatile tool for robust statistical inference in various settings.

The rest of this work is organized as follows: Section 3.1 provides some notations used in this research. The exponential families and their conjugate priors are discussed in Section 3.2 along with an example of the Multinomial-Dirichlet model. In Section 3.3, we derive the conditions for the asymptotic posterior normality with the Diaconis-Ylvisaker prior. By conducting a simulation outlined in Chapter 4, we confirm the validity of these conditions within the framework of the Multinomial-Dirichlet model. Our proposed theorem is evaluated and demonstrated to exhibit superior convergence rates compared to preceding theorems in Section 4.1. Besides that, we present the practical applications of this theorem in Bayesian density estimation in Section 4.2 and the estimation of the mean of an infinite-dimensional normal distribution in Section 4.3.

CHAPTER 3 EXPONENTIAL FAMILY MODEL AND DIACONIS-YLVISAKER PRIOR

3.1 Preliminaries

In this section, we introduce some terminologies and notations that are used throughout the work.

For a vector $\boldsymbol{x} = (x_1, \ldots, x_d)^\top$, $\|\boldsymbol{x}\|_2$ denotes its Euclidean norm, $\left(\sum_{j=1}^d x_j^2\right)^{1/2}$. For a square matrix $\boldsymbol{A}$, $\|\boldsymbol{A}\|_2$ denotes its operator norm defined by $\sup\left\{\|\boldsymbol{A}\boldsymbol{x}\|_2 : \boldsymbol{x}^\top\boldsymbol{x} = 1,\, \boldsymbol{x} \in \mathbb{R}^d\right\}$. The spectral norm of a tensor $\mathcal{B} \in \mathbb{R}^{d\times d\times d}$ is defined as

$$\|\mathcal{B}\| = \sup\left\{\langle \mathcal{B}, \boldsymbol{x}\otimes\boldsymbol{y}\otimes\boldsymbol{z}\rangle : \boldsymbol{x}^\top\boldsymbol{x} = \boldsymbol{y}^\top\boldsymbol{y} = \boldsymbol{z}^\top\boldsymbol{z} = 1,\, \boldsymbol{x},\boldsymbol{y},\boldsymbol{z} \in \mathbb{R}^d\right\}.$$

Additional properties of the spectral norm of a third-order tensor can be found in Qi and Hu (2019) (see also Appendix .3). For two sequences of real numbers $\{a_n\}$ and $\{b_n\}$, we write $a_n = o(b_n)$ if $a_n/b_n \to 0$ as $n \to \infty$. Throughout the thesis, $c_1, c_2, \ldots$ are generally used to denote constants whose values might change from one line to another but are independent of everything else.

3.2 Exponential Families and Conjugate Priors

Referring to Diaconis and Ylvisaker (1979), let ν be a fixed σ-finite measure on the Borel sets $\mathcal{B}_{\mathbb{R}^d}$, and let $\mathcal{F}$ be the interior of the convex hull of the support set of ν. Assume that $\mathcal{F}$ is a nonempty open set in $\mathbb{R}^d$. For $\boldsymbol{\theta} \in \mathbb{R}^d$, define $\Psi(\boldsymbol{\theta}) = \ln \int_{\mathcal{F}} \exp\left\{\boldsymbol{x}^\top\boldsymbol{\theta}\right\} d\nu(\boldsymbol{x})$ and let $\boldsymbol{\Theta} = \left\{\boldsymbol{\theta} \in \mathbb{R}^d | \Psi(\boldsymbol{\theta}) < \infty\right\}$. Assume that the natural parameter space $\boldsymbol{\Theta}$ is a nonempty open set in $\mathbb{R}^d$. The exponential family is defined by

$$dP(\boldsymbol{x}|\boldsymbol{\theta}) = \exp\left\{\boldsymbol{x}^\top\boldsymbol{\theta} - \Psi(\boldsymbol{\theta})\right\} d\nu(\boldsymbol{x}), \quad \boldsymbol{\theta} \in \boldsymbol{\Theta}. \tag{3.2.1}$$

Given an independent sample $\boldsymbol{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^\top$ from Equation (3.2.1), the density takes the form

$$f(\boldsymbol{x}|\boldsymbol{\theta}) = \exp\left\{n\bar{\boldsymbol{x}}^\top\boldsymbol{\theta} - n\Psi(\boldsymbol{\theta})\right\}, \tag{3.2.2}$$

where $\bar{\boldsymbol{x}}$ is the sample mean; see also Ghosal (2000). In this model, the natural parameter $\boldsymbol{\theta}$ is a d_n-dimensional vector where d_n is allowed to grow with the sample size n. The family of conjugate priors for the parameter $\boldsymbol{\theta}$ of the regular exponential family, referred to as Diaconis-Ylvisaker priors is given by

$$\pi(\boldsymbol{\theta}; n_0, \boldsymbol{s}_0) = \exp\left\{\boldsymbol{s}_0^\top \boldsymbol{\theta} - n_0 \Psi(\boldsymbol{\theta})\right\} h(n_0, \boldsymbol{s}_0), \quad n_0 \in \mathbb{R}, \boldsymbol{s}_0 \in \mathbb{R}^{d_n}. \tag{3.2.3}$$

Then the posterior distribution belongs to the same family Equation (3.2.3), with parameters $n_0 + n$ and $\boldsymbol{s}_0 + n\bar{\boldsymbol{x}}$ (Johndrow and Bhattacharya, 2018).

3.2.1 Multinomial-Dirichlet Model

As an example, we consider one member of the exponential family, the multinomial distribution with $(d_n + 1)$ cells. Suppose that $\boldsymbol{x} = (x_0, x_1, \ldots, x_{d_n})^\top$ is a multinomial sample from n trials, with cell probabilities $\boldsymbol{p} = (p_0, p_1, \ldots, p_{d_n})^\top$, where $\sum_{j=0}^{d_n} x_j = n$ and $\sum_{j=0}^{d_n} p_j = 1$. The probability mass function is given by

$$f(\boldsymbol{x}|\boldsymbol{p}) = \frac{\Gamma(\sum_{j=0}^{d_n} x_j + 1)}{\prod_{j=0}^{d_n} \Gamma(x_j + 1)} \prod_{j=0}^{d_n} p_j^{x_j}.$$

The Dirichlet distribution is a conjugate prior for the multinomial distribution. Let $\boldsymbol{p}$ have the Dirichlet prior with density function

$$\pi(\boldsymbol{p}|\boldsymbol{\alpha}) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{j=0}^{d_n} p_j^{\alpha_j - 1},$$

where $B(\boldsymbol{\alpha}) = \prod_{j=0}^{d_n} \Gamma(\alpha_j) / \Gamma(\sum_{j=0}^{d_n} \alpha_j)$. Denoting the canonical parameter by $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_{d_n})^\top$,

where $\theta_j = \ln\Big[p_j/\big(1-\sum_{k=1}^{d_n} p_k\big)\Big]$, $j \in \{1,\ldots,d_n\}$. Then the joint distribution of $\boldsymbol{x}$ is

$$
\begin{aligned}
f(\boldsymbol{x}|\boldsymbol{\theta}) &\propto \prod_{j=0}^{d_n} p_j^{x_j} \\
&= \exp\Big\{\sum_{j=0}^{d_n} x_j \log p_j\Big\} \\
&= \exp\Big\{\sum_{j=1}^{d_n} x_j \log p_j + \Big(n - \sum_{j=1}^{d_n} x_j\Big)\log\Big(1-\sum_{j=1}^{d_n} p_j\Big)\Big\} \\
&= \exp\Big\{\sum_{j=1}^{d_n} x_j \log\Big(\frac{p_j}{1-\sum_{j=1}^{d_n} p_j}\Big) + n\log\Big(1-\sum_{j=1}^{d_n} p_j\Big)\Big\} \\
&= \exp\Big\{\sum_{j=1}^{d_n} x_j\theta_j + n\log\Big(1-\sum_{j=1}^{d_n} p_j\Big)\Big\} \\
&= \exp\Big\{\sum_{j=1}^{d_n} x_j\theta_j - n\log\Big[\sum_{j=1}^{d_n}\exp(\theta_j)+1\Big]\Big\}.
\end{aligned}
$$

The Dirichlet prior density can be expressed as follows:

$$
\begin{aligned}
f(\boldsymbol{\theta}|\boldsymbol{\alpha}) &= \frac{1}{B(\boldsymbol{\alpha})}\prod_{j=0}^{d_n} p_j^{\alpha_j-1} \\
&= \frac{1}{B(\boldsymbol{\alpha})}\exp\Big\{\sum_{j=0}^{d_n}(\alpha_j-1)\log p_j\Big\} \\
&= \frac{1}{B(\boldsymbol{\alpha})}\exp\Big\{\sum_{j=1}^{d_n}(\alpha_j-1)\log p_j + (\alpha_0-1)\log\Big(1-\sum_{j=1}^{d_n} p_j\Big)\Big\} \\
&= \frac{1}{B(\boldsymbol{\alpha})}\exp\Big\{\sum_{j=1}^{d_n}(\alpha_j-1)\log p_j + \Big(\sum_{j=0}^{d_n}\alpha_j - d_n - 1 - \sum_{j=1}^{d_n}(\alpha_j-1)\Big)\log\Big(1-\sum_{j=1}^{d_n} p_j\Big)\Big\} \\
&= \frac{1}{B(\boldsymbol{\alpha})}\exp\Big\{\sum_{j=1}^{d_n}(\alpha_j-1)\log\Big(\frac{p_j}{1-\sum_{j=1}^{d_n} p_j}\Big) + \Big(\sum_{j=0}^{d_n}\alpha_j - d_n - 1\Big)\log\Big(1-\sum_{j=1}^{d_n} p_j\Big)\Big\} \\
&= \frac{1}{B(\boldsymbol{\alpha})}\exp\Big\{\sum_{j=1}^{d_n}(\alpha_j-1)\theta_j - \Big(\sum_{j=0}^{d_n}\alpha_j - d_n - 1\Big)\Psi(\boldsymbol{\theta})\Big\}.
\end{aligned}
$$

The multinomial model and the Dirichlet prior render to Equation (3.2.2) and Equation (3.2.3)

with $\Psi(\boldsymbol{\theta}) = \ln\Big[\sum_{j=1}^{d_n} \exp(\theta_j) + 1\Big]$, hyper-parameters $n_0 = \sum_{j=0}^{d_n} \alpha_j - d_n - 1$, $s_{0j} = \alpha_j - 1$ for $j \in \{1, \ldots, d_n\}$, and the normalizing constant $h(n_0, \boldsymbol{s}_0) = 1/B(\boldsymbol{\alpha})$.

3.3 Bernstein-von Mises Theorem for the Diaconis-Ylvisaker Prior

In this section, we develop the Bernstein-von Mises theorem with the Diaconis-Ylvisaker prior under sufficient conditions. Those conditions call for the prior to concentrate its mass on a moderately-sized neighborhood of the true parameter $\boldsymbol{\theta}_0$. Moreover, the prior is required to be sufficiently flat such that the prior density fraction of any two arbitrary local parameters approaches 1 as the sample size grows. The key idea of the proof is along the lines with Bontemps (2011); Boucheron and Gassiat (2009); Van der Vaart (2000), see Appendix .2.

3.3.1 Theorem

The Bernstein-von Mises theorem manifests sufficient conditions on the prior under which the posterior distribution converges to a normal distribution centered at the maximum likelihood estimator with variance, the inverse of the observed Fisher information matrix.

Let $\boldsymbol{\theta}_0$ be the true parameter. Then the Fisher information matrix is equal to $\Psi''(\boldsymbol{\theta}_0)$ Ghosal (2000). Let $\boldsymbol{U}$ be a square root of $\Psi''(\boldsymbol{\theta}_0)$, i.e., $\boldsymbol{U}\boldsymbol{U}^\top = \Psi''(\boldsymbol{\theta}_0)$. For $R > 0$, define the ellipsoid

$$\varepsilon_{\boldsymbol{\theta_0},\boldsymbol{U}}(R) = \left\{\boldsymbol{\theta} \in \mathbb{R}^{d_n} : n(\boldsymbol{\theta} - \boldsymbol{\theta}_0)^\top \Psi''(\boldsymbol{\theta}_0)(\boldsymbol{\theta} - \boldsymbol{\theta}_0) \leq R\right\}.$$

Theorem 3.1. *Suppose that the following conditions, referring to them by $C1$, $C2$, and $C3$, respectively, hold*

$C1$. $d_n = o(R_n)$ *and* $\sup_{\boldsymbol{\theta} \in \varepsilon^c_{\boldsymbol{\theta}_0,\boldsymbol{U}}(R_n/4)} \|\Psi'''(\boldsymbol{\theta})\| \to 0$ *as* $R_n \to \infty$.

$C2$. $n_0 = o(n/R_n)$ *and* $\sqrt{1/n_0}\,\|\boldsymbol{U}^{-1}\left(\boldsymbol{s}_0 - n_0\Psi'\left(\boldsymbol{\theta}_0\right)\right)\|_2$ *is bounded.*

$C3$. $R_n = o(n)$.

Then

$$E\left\|\mathcal{N}\left(\hat{\boldsymbol{\theta}}, \left(n\Psi''(\boldsymbol{\theta}_0)\right)^{-1}\right) - \pi(\boldsymbol{\theta}|\boldsymbol{x})\right\|_{TV} \to 0 \text{ as } n \to \infty.$$

In the theorem, $\hat{\boldsymbol{\theta}}$ is the maximum likelihood estimator of $\boldsymbol{\theta}$ and $E[\cdot]$ refers to the expecta-

tion under $P(\cdot|\boldsymbol{\theta}_0)$. The above conditions illustrate mild constraints on the prior such that a high-dimensional Bernstein-von Mises theorem result follows. By condition $C1$, we require that the prior concentrates on a neighborhood of $\boldsymbol{\theta}_0$. With condition $C2$, the prior needs to be flat enough in this neighborhood. Through condition $C3$, we ensure a moderate size of this neighborhood.

$C1$ implies that d_n becomes insignificant relative to R_n as n approaches infinity and the spectral norm of the second derivative of the mean vector approaches zero as the parameter approaches the true parameter. $C2$ restricts the standardized distance between two arbitrary parameters to be bounded. $C3$ requires that R_n becomes negligible relative to n.

In this work, $\Psi'(\boldsymbol{\theta}_0)$ is the true mean vector and the maximum likelihood estimator $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ uniquely satisfies $\Psi'(\hat{\boldsymbol{\theta}}) = \bar{\boldsymbol{x}}$. Let $\mathcal{N}$ be the normal distribution $\mathcal{N}\left(\hat{\boldsymbol{\theta}}, (n\Psi''(\boldsymbol{\theta}_0))^{-1}\right)$, and let $\mathcal{N}^{R_n}$ be the normal distribution $\mathcal{N}$ restricted and renormalized to the ellipsoid $\varepsilon_{\boldsymbol{\theta}_0,\boldsymbol{U}}(R_n)$. Similarly, π stands for the posterior distribution $\pi(\boldsymbol{\theta}|\boldsymbol{x})$, and π^{R_n} stands for the truncated posterior distribution. Thus, we have

$$E\left\|\mathcal{N}\left(\hat{\boldsymbol{\theta}}, (n\Psi''(\boldsymbol{\theta}_0))^{-1}\right) - \pi(\boldsymbol{\theta}|\boldsymbol{x})\right\|_{TV} = E\left\|\mathcal{N} - \mathcal{N}^{R_n} + \mathcal{N}^{R_n} - \pi^{R_n} + \pi^{R_n} - \pi\right\|_{TV}.$$

We prove the Bernstein-von Mises theorem by splitting the above expected total variation distance expression into three terms,

$$\begin{aligned} &E\left\|\mathcal{N}\left(\hat{\boldsymbol{\theta}}, (n\Psi''(\boldsymbol{\theta}_0))^{-1}\right) - \pi(\boldsymbol{\theta}|\boldsymbol{x})\right\|_{TV} \\ &\quad\leq E\left\|\mathcal{N} - \mathcal{N}^{R_n}\right\|_{TV} + E\|\mathcal{N}^{R_n} - \pi^{R_n}\|_{TV} + E\left\|\pi^{R_n} - \pi\right\|_{TV}. \end{aligned}$$

By noting that

$$E\left\|\mathcal{N}\left(\hat{\boldsymbol{\theta}}, (n\Psi''(\boldsymbol{\theta}_0))^{-1}\right) - \pi(\boldsymbol{\theta}|\boldsymbol{x})\right\|_{TV} \to 0,$$

if there exists a sequence $\{R_n\}$ such that

$$E\left\|\mathcal{N}-\mathcal{N}^{R_n}\right\|_{TV}\to 0,\ E\left\|\mathcal{N}^{R_n}-\pi^{R_n}\right\|_{TV}\to 0,\ \text{and } E\left\|\pi^{R_n}-\pi\right\|_{TV}\to 0 \text{ as } n\to\infty.$$

We refer to these three terms as T_1, T_2, and T_3 in order. To prove T_1, we use Cirelson's inequality Bontemps (2011) and concentration inequality which is shown by Lemma 3.2. We illustrate T_2 by Lemma 3.3 following from the Cauchy-Schwarz inequality. Based on Lemma 3.2, we propose Lemma 3.4 to show T_3. The proofs of T_1, T_2, and T_3 are separated into Section 3.3.2, Section 3.3.3, and Section 3.3.4, respectively where one needs to put conditions on prior such as concentration and flatness, and on the size of the neighborhood of the true parameter $\boldsymbol{\theta}_0$.

3.3.2 Prior Concentration

The magnitude of the eigenvalues in the Fisher information matrix reflects the extent to which the data captures the parameter $\boldsymbol{\theta}$. We have $\sup_{\boldsymbol{\theta}\in\varepsilon^c_{\boldsymbol{\theta}_0,\boldsymbol{U}}(R_n/4)}\|\Psi'''(\boldsymbol{\theta})\|$ bounded by the largest eigenvalue of the Fisher information matrix; see more details in Section 4.1 for the Multinomial-Dirichlet example. The condition states that, as the neighborhood size grows, the largest eigenvalue of the Fisher information matrix for parameters outside this neighborhood becomes smaller. In other words, the data carry less information about the parameter, far from the true value, as the neighborhood expands. This indicates that the prior mass becomes asymptotically negligible for parameters that fall into the complement of the expanding neighborhood. That is to say, we demand that the prior assign the majority of its mass to the neighborhood of $\boldsymbol{\theta}_0$. Later on, a supplementary requirement for such neighborhoods will be introduced by Lemma 3.4.

Lemma 3.1. *If* $d_n = o(R_n)$ *and* $\sup_{\boldsymbol{\theta}\in\varepsilon^c_{\boldsymbol{\theta}_0,\boldsymbol{U}}(R_n/4)}\|\Psi'''(\boldsymbol{\theta})\|\to 0$ *as* $R_n\to\infty$*, then*

$$E\left\|\mathcal{N}\left(\hat{\boldsymbol{\theta}},(n\Psi''(\boldsymbol{\theta}_0))^{-1}\right)-\mathcal{N}^{R_n}\left(\hat{\boldsymbol{\theta}},(n\Psi''(\boldsymbol{\theta}_0))^{-1}\right)\right\|_{TV}\to 0.$$

Proof of Lemma 3.1: For any measurable set $B \subseteq \mathbb{R}^{d_n}$,

$$\left\|\mathcal{N}(B) - \mathcal{N}^{R_n}(B)\right\|_{TV} = \left|\left(\mathcal{N}(B|\varepsilon^c_{\boldsymbol{\theta}_0,\boldsymbol{U}}(R_n)) - \mathcal{N}(B|\varepsilon_{\boldsymbol{\theta}_0,\boldsymbol{U}}(R_n))\right) \cdot \mathcal{N}\left(\varepsilon^c_{\boldsymbol{\theta}_0,\boldsymbol{U}}(R_n)\right)\right|.$$

When $B \subseteq \varepsilon^c_{\boldsymbol{\theta}_0,\boldsymbol{U}}(R_n)$,

$$\begin{aligned}
&\sup_B \left|\mathcal{N}\left(B|\varepsilon^c_{\boldsymbol{\theta}_0,\boldsymbol{U}}(R_n)\right) - \mathcal{N}\left(B|\varepsilon_{\boldsymbol{\theta}_0,\boldsymbol{U}}(R_n)\right)\right| \\
&= \left|\mathcal{N}\left(\varepsilon^c_{\boldsymbol{\theta}_0,\boldsymbol{U}}(R_n)|\varepsilon^c_{\boldsymbol{\theta}_0,\boldsymbol{U}}(R_n)\right) - \mathcal{N}\left(\varepsilon^c_{\boldsymbol{\theta}_0,\boldsymbol{U}}(R_n)|\varepsilon_{\boldsymbol{\theta}_0,\boldsymbol{U}}(R_n)\right)\right| \\
&= |1-0| \\
&= 1.
\end{aligned}$$

Similarly for $B \subseteq \varepsilon_{\boldsymbol{\theta}_0,\boldsymbol{U}}(R_n)$. In all other cases, $\sup_B \left|\mathcal{N}\left(B|\varepsilon^c_{\boldsymbol{\theta}_0,\boldsymbol{U}}(R_n)\right) - \mathcal{N}\left(B|\varepsilon_{\boldsymbol{\theta}_0,\boldsymbol{U}}(R_n)\right)\right| < 1$. So $\sup_{B\subseteq\mathbb{R}^{d_n}} \left|\mathcal{N}\left(B|\varepsilon^c_{\boldsymbol{\theta}_0,\boldsymbol{U}}(R_n)\right) - \mathcal{N}\left(B|\varepsilon_{\boldsymbol{\theta}_0,\boldsymbol{U}}(R_n)\right)\right| = 1$. Thus,

$$\begin{aligned}
\left\|\mathcal{N} - \mathcal{N}^{R_n}\right\|_{TV} &\leq \mathcal{N}\left(\varepsilon^c_{\boldsymbol{\theta}_0,\boldsymbol{U}}(R_n)\right) \cdot \sup_{B\subseteq\mathbb{R}^{d_n}} \left|\mathcal{N}\left(B|\varepsilon^c_{\boldsymbol{\theta}_0,\boldsymbol{U}}(R_n)\right) - \mathcal{N}\left(B|\varepsilon_{\boldsymbol{\theta}_0,\boldsymbol{U}}(R_n)\right)\right| \\
&= \mathcal{N}\left(\varepsilon^c_{\boldsymbol{\theta}_0,\boldsymbol{U}}(R_n)\right).
\end{aligned}$$

Let $\mathcal{N}^0$ be the normal distribution $\mathcal{N}\left(\boldsymbol{\theta}_0, (n\Psi''(\boldsymbol{\theta}_0))^{-1}\right)$. Then

$$\left\|\mathcal{N} - \mathcal{N}^{R_n}\right\|_{TV} \leq \mathcal{N}^0\left(\varepsilon^c_{\boldsymbol{\theta}_0,\boldsymbol{U}}(R_n/4)\right) + I(n(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^\top \Psi''(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) > R_n/4).$$

Let $T \sim \chi^2_{d_n}$. Taking the expectation of total variation distance, we have

$$E\left\|\mathcal{N} - \mathcal{N}^{R_n}\right\|_{TV} \leq \Pr\left(T > R_n/4\right) + \Pr\left(n(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^\top \Psi''(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) > R_n/4\right). \quad (3.3.1)$$

Since $T \sim \chi^2_{d_n}$, $\sqrt{T} \sim \chi_{d_n}$. So $E\sqrt{T} \leq \sqrt{d_n}$ and Var$(\sqrt{T}) \leq 1$. By Cirelson's inequality, If

$d_n = o(R_n)$, for n large enough,

$$\Pr\left(\sqrt{T} > \sqrt{d_n} + (\sqrt{R_n} - 2\sqrt{d_n})/2\right) \leq \exp\left\{-R_n/8\right\}.$$

Hence, the first term on the right-hand side of the inequality Equation (3.3.1) converges 0 as $n \to \infty$. We will show that the second term on the right-hand side of the inequality Equation (3.3.1) also converges 0 as $n \to \infty$ by the following Lemma 3.2. It is instructive to note that our Lemma 3.1 is reminiscent of Proposition 3.11 of Boucheron and Gassiat (2009) and Lemma 5 of Bontemps (2011). While for regression model in Bontemps (2011), it is relatively simple to bound the second term in Equation (3.3.1) using the χ^2 distribution, it runs into difficulty in our case. Therefore, to bound the second term we develop a non-trivial concentration inequality of the maximum likelihood estimator in Lemma 3.2 which goes beyond the standard assumptions such as sub-Gaussianity. □

Lemma 3.2. *If* $\sup_{\boldsymbol{\theta} \in \varepsilon^c_{\boldsymbol{\theta}_0, U}(R_n/4)} \|\Psi'''(\boldsymbol{\theta})\| \to 0$ *as* $R_n \to \infty$*, then*

$$\Pr\left(n(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^\top \Psi''(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) > R_n/4\right) \to 0, \;\; n \to \infty.$$

The proof of Lemma 3.2 is deferred to Appendix .1. Depending on the prior concentration condition, a concentration inequality of the quadratic form of the maximum likelihood estimator is established. Hence,

$$E\left\|\mathcal{N}\left(\hat{\boldsymbol{\theta}}, (n\Psi''(\boldsymbol{\theta}_0))^{-1}\right) - \mathcal{N}^{R_n}\left(\hat{\boldsymbol{\theta}}, (n\Psi''(\boldsymbol{\theta}_0))^{-1}\right)\right\|_{TV} \to 0.$$

We have $\boldsymbol{\theta}$ If $d_n = o(R_n)$, then as $n \to \infty$,

$$E\left\|\mathcal{N} - \mathcal{N}^{R_n}\right\|_{TV} \leq \exp\left\{-\frac{R_n}{8}\right\}.$$

Because the second probability term is bounded up by $4d_n/R_n$, T_1 converges faster than

$2\exp\{-R_n/8\}$ as shown in Bontemps (2011).

Since the largest eigenvalue of $\Psi''(\boldsymbol{\theta}_0)$ is of the order d_n^{-2}, $\sup_{\boldsymbol{\theta}\in\varepsilon_{\boldsymbol{\theta}_0,\boldsymbol{U}}(R_n/4)} \lambda_{\max}\big(\Psi''(\boldsymbol{\theta})\big)$ is of the order d_n^{-2}.

3.3.3 Prior Flatness

Lemma 3.3 plays a crucial role in guiding the shape of the prior distribution used in the analysis. It imposes a condition that the prior should be approximately uniformly distributed in the neighborhood of $\boldsymbol{\theta}_0$, the true parameter of interest. In order to show the right-hand side of Equation (3.3.2) converges towards 0, the prior density fraction between any two arbitrary local parameters is expected to approach 1 (see also Condition 3.4 of Boucheron and Gassiat (2009) and Condition 1 of Bontemps (2011)), indicating that the prior mass is spread out uniformly through the condition of bounded $\sqrt{1/n_0}\left\|\boldsymbol{U}^{-1}\left(\boldsymbol{s}_0 - n_0\Psi'\left(\boldsymbol{\theta}_0\right)\right)\right\|_2$. By employing this specific form of an improper local prior, the posterior distribution becomes highly influenced by the likelihood of the observed data. This kind of assumption is quite common in the literature dealing with the concentration of the posterior distribution. Subsequently, Taylor's approximation comes into play, making the exponential family density and the normal density almost identical under condition $n_0 = o(n/R_n)$. As the sample size n increases, the term T_2 in the theorem tends to 0.

Lemma 3.3. *If* $n_0 = o(n/R_n)$ *and* $\sqrt{1/n_0}\left\|\boldsymbol{U}^{-1}\left(\boldsymbol{s}_0 - n_0\Psi'\left(\boldsymbol{\theta}_0\right)\right)\right\|_2$ *is bounded, then*

$$E\left\|\pi^{R_n}(\boldsymbol{\theta}|\boldsymbol{x}) - \mathcal{N}^{R_n}\left(\hat{\boldsymbol{\theta}}, \left(n\Psi''(\boldsymbol{\theta}_0)\right)^{-1}\right)\right\|_{TV} \to 0, \;\; n \to \infty.$$

Proof of Lemma 3.3: The total variation distance between two arbitrary probability measures L

and K can be expressed in the form $\|L-K\|_{TV} = 2\int(1-l/k)^+ dK$. So

$$\begin{aligned}
&\frac{1}{2}\left\|\pi^{R_n}(\boldsymbol{\theta}|\boldsymbol{x}) - \mathcal{N}^{R_n}\left(\hat{\boldsymbol{\theta}}, (n\Psi''(\boldsymbol{\theta}_0))^{-1}\right)\right\|_{TV} \\
&= \int\left(1 - \frac{d\mathcal{N}^{R_n}(\boldsymbol{\theta})}{d\pi^{R_n}(\boldsymbol{\theta}|\boldsymbol{x})}\right)^+ d\pi^{R_n}(\boldsymbol{\theta}|\boldsymbol{x}) \\
&= \int\left(1 - \frac{d\mathcal{N}^{R_n}(\boldsymbol{\theta})}{\frac{\mathbf{1}_{\varepsilon_{\boldsymbol{\theta}_0,\boldsymbol{U}^{(R_n)}}}(\boldsymbol{\theta})\pi(\boldsymbol{\theta})f(\boldsymbol{x}|\boldsymbol{\theta})}{\int_{\varepsilon_{\boldsymbol{\theta}_0,\boldsymbol{U}^{(R_n)}}}\pi(\boldsymbol{\tau})f(\boldsymbol{x}|\boldsymbol{\tau})d\boldsymbol{\tau}}}\right)^+ d\pi^{R_n}(\boldsymbol{\theta}|\boldsymbol{x}) \\
&\leq \int\int\left(1 - \frac{\pi(\boldsymbol{\tau})f(\boldsymbol{x}|\boldsymbol{\tau})d\mathcal{N}^{R_n}(\boldsymbol{\theta})}{\pi(\boldsymbol{\theta})f(\boldsymbol{x}|\boldsymbol{\theta})d\mathcal{N}^{R_n}(\boldsymbol{\tau})}\right)^+ d\mathcal{N}^{R_n}(\boldsymbol{\tau})d\pi^{R_n}(\boldsymbol{\theta}|\boldsymbol{x}).
\end{aligned} \tag{3.3.2}$$

The integrand can be expanded as

$$\begin{aligned}
1 - \exp\Big\{&\boldsymbol{s}_0^\top(\boldsymbol{\tau}-\boldsymbol{\theta}) - n_0[\Psi(\boldsymbol{\tau}) - \Psi(\boldsymbol{\theta})] + n\bar{\boldsymbol{x}}^\top(\boldsymbol{\tau}-\boldsymbol{\theta}) - n[\Psi(\boldsymbol{\tau}) - \Psi(\boldsymbol{\theta})] \\
&+ \frac{n}{2}[(\boldsymbol{\tau}-\hat{\boldsymbol{\tau}})^\top\Psi''(\boldsymbol{\theta}_0)(\boldsymbol{\tau}-\hat{\boldsymbol{\tau}}) - (\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})^\top\Psi''(\boldsymbol{\theta}_0)(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})]\Big\}.
\end{aligned}$$

We rescale the parameters $\boldsymbol{\tau}$ and $\boldsymbol{\theta}$ to the local parameters $\boldsymbol{h} = \sqrt{n}(\boldsymbol{\tau}-\boldsymbol{\theta}_0)$ and $\boldsymbol{g} = \sqrt{n}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)$, respectively. Then the integrand can be written as

$$\begin{aligned}
1 - \exp\Big\{&\frac{1}{\sqrt{n}}(\boldsymbol{h}-\boldsymbol{g})^\top\boldsymbol{s}_0 - n_0\big[\Psi(\boldsymbol{\theta}_0 + \frac{1}{\sqrt{n}}\boldsymbol{h}) - \Psi(\boldsymbol{\theta}_0 + \frac{1}{\sqrt{n}}\boldsymbol{g})\big] \\
&+ \sqrt{n}(\boldsymbol{h}-\boldsymbol{g})^\top\bar{\boldsymbol{x}} - n\big[\Psi(\boldsymbol{\theta}_0 + \frac{1}{\sqrt{n}}\boldsymbol{h}) - \Psi(\boldsymbol{\theta}_0 + \frac{1}{\sqrt{n}}\boldsymbol{g})\big] \\
&+ \frac{n}{2}[(\boldsymbol{\theta}_0 + \frac{1}{\sqrt{n}}\boldsymbol{h} - \hat{\boldsymbol{\tau}})^\top\Psi''(\boldsymbol{\theta}_0)(\boldsymbol{\theta}_0 + \frac{1}{\sqrt{n}}\boldsymbol{h} - \hat{\boldsymbol{\tau}}) \\
&- (\boldsymbol{\theta}_0 + \frac{1}{\sqrt{n}}\boldsymbol{g} - \hat{\boldsymbol{\theta}})^\top\Psi''(\boldsymbol{\theta}_0)(\boldsymbol{\theta}_0 + \frac{1}{\sqrt{n}}\boldsymbol{g} - \hat{\boldsymbol{\theta}})]\Big\}.
\end{aligned}$$

By the second-order Taylor's approximation,

$$\Psi(\boldsymbol{\theta}_0 + \frac{1}{\sqrt{n}}\boldsymbol{h}) - \Psi(\boldsymbol{\theta}_0 + \frac{1}{\sqrt{n}}\boldsymbol{g}) \approx \frac{1}{\sqrt{n}}(\boldsymbol{h}-\boldsymbol{g})^\top\Psi'(\boldsymbol{\theta}_0) + \frac{1}{2n}[\boldsymbol{h}^\top\Psi''(\boldsymbol{\theta}_0)\boldsymbol{h} - \boldsymbol{g}^\top\Psi''(\boldsymbol{\theta}_0)\boldsymbol{g}].$$

Then the integrand can be rewritten as

$$1-\exp\Bigg\{\frac{1}{\sqrt{n}}(\boldsymbol{h}-\boldsymbol{g})^\top[\boldsymbol{s}_0-n_0\Psi'(\boldsymbol{\theta}_0)+n\bar{\boldsymbol{x}}-n\Psi'(\boldsymbol{\theta}_0)]-\frac{n_0+n}{2n}[\boldsymbol{h}^\top\Psi''(\boldsymbol{\theta}_0)\boldsymbol{h}-\boldsymbol{g}^\top\Psi''(\boldsymbol{\theta}_0)\boldsymbol{g}]$$
$$+\frac{n}{2}[(\hat{\boldsymbol{\tau}}-\boldsymbol{\theta}_0)^\top\Psi''(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\tau}}-\boldsymbol{\theta}_0)-(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)^\top\Psi''(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)]$$
$$+\frac{1}{2}[\boldsymbol{h}^\top\Psi''(\boldsymbol{\theta}_0)\boldsymbol{h}-\boldsymbol{g}^\top\Psi''(\boldsymbol{\theta}_0)\boldsymbol{g}]-\sqrt{n}[\boldsymbol{h}^\top\Psi''(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\tau}}-\boldsymbol{\theta}_0)-\boldsymbol{g}^\top\Psi''(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)]\Bigg\}.$$

Then, by the first-order Taylor's approximation,

$$n\bar{\boldsymbol{x}}-n\Psi'(\boldsymbol{\theta}_0)\approx n\Psi''(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0).$$

Thus, we have

$$1-\exp\Bigg\{\frac{1}{\sqrt{n}}(\boldsymbol{h}-\boldsymbol{g})^\top[\boldsymbol{s}_0-n_0\Psi'(\boldsymbol{\theta}_0)+n\Psi''(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)]-\frac{n_0+n}{2n}[\boldsymbol{h}^\top\Psi''(\boldsymbol{\theta}_0)\boldsymbol{h}-\boldsymbol{g}^\top\Psi''(\boldsymbol{\theta}_0)\boldsymbol{g}]$$
$$+\frac{n}{2}[(\hat{\boldsymbol{\tau}}-\boldsymbol{\theta}_0)^\top\Psi''(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\tau}}-\boldsymbol{\theta}_0)-(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)^\top\Psi''(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)]+$$
$$\frac{1}{2}[\boldsymbol{h}^\top\Psi''(\boldsymbol{\theta}_0)\boldsymbol{h}-\boldsymbol{g}^\top\Psi''(\boldsymbol{\theta}_0)\boldsymbol{g}]-\sqrt{n}[\boldsymbol{h}^\top\Psi''(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\tau}}-\boldsymbol{\theta}_0)-\boldsymbol{g}^\top\Psi''(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)]\Bigg\}.$$

Further simplifying it,

$$1-\exp\Bigg\{\frac{1}{\sqrt{n}}(\boldsymbol{h}-\boldsymbol{g})^\top[\boldsymbol{s}_0-n_0\Psi'(\boldsymbol{\theta}_0]-\frac{n_0}{2n}[\boldsymbol{h}^\top\Psi''(\boldsymbol{\theta}_0)\boldsymbol{h}-\boldsymbol{g}^\top\Psi''(\boldsymbol{\theta}_0)\boldsymbol{g}]$$
$$+\frac{n}{2}[(\hat{\boldsymbol{\tau}}-\boldsymbol{\theta}_0)^\top\Psi''(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\tau}}-\boldsymbol{\theta}_0)-(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)^\top\Psi''(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)]$$
$$+\sqrt{n}\boldsymbol{h}^\top\Psi''(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}}-\hat{\boldsymbol{\tau}})\Bigg\}.$$

Note that $n[(\hat{\boldsymbol{\tau}}-\boldsymbol{\theta}_0)^\top\Psi''(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\tau}}-\boldsymbol{\theta}_0)-(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)^\top\Psi''(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)]\to 0$ and $\sqrt{n}\boldsymbol{h}^\top\Psi''(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}}-\hat{\boldsymbol{\tau}})\to 0$ as $n\to\infty$. So we have

$$1-\exp\left\{\frac{1}{\sqrt{n}}(\boldsymbol{h}-\boldsymbol{g})^\top\left(\boldsymbol{s}_0-n_0\Psi'\left(\boldsymbol{\theta}_0\right)\right)-\frac{n_0}{2n}\left(\boldsymbol{h}^\top\Psi''(\boldsymbol{\theta}_0)\boldsymbol{h}-\boldsymbol{g}^\top\Psi''(\boldsymbol{\theta}_0)\boldsymbol{g}\right)\right\}.$$

Note that

$$\sup_{\|\boldsymbol{U}\boldsymbol{h}\|_2^2 \leq R_n} \frac{n_0}{2n} \boldsymbol{h}^\top \Psi''(\boldsymbol{\theta}_0)\boldsymbol{h} = \frac{n_0 R_n}{2n}.$$

If $n_0 = o(n/R_n)$, then $(n_0/2n)\boldsymbol{h}^\top \Psi''(\boldsymbol{\theta}_0)\boldsymbol{h} \to 0$ by squeeze law. Thus,

$$(n_0/2n)\left(\boldsymbol{h}^\top \Psi''\left(\boldsymbol{\theta}_0\right)\boldsymbol{h} - \boldsymbol{g}^\top \Psi''\left(\boldsymbol{\theta}_0\right)\boldsymbol{g}\right) \to 0 \text{ as } n \to \infty.$$

By Cauchy-Schwarz inequality

$$\begin{aligned}\frac{1}{\sqrt{n}}(\boldsymbol{h}-\boldsymbol{g})^\top \left(\boldsymbol{s}_0 - n_0\Psi'\left(\boldsymbol{\theta}_0\right)\right) &= \frac{1}{\sqrt{n}}(\boldsymbol{h}-\boldsymbol{g})^\top \boldsymbol{U}\boldsymbol{U}^{-1}\left(\boldsymbol{s}_0 - n_0\Psi'\left(\boldsymbol{\theta}_0\right)\right) \\ &\leq \sqrt{\frac{n_0}{n}}\left\|\boldsymbol{U}(\boldsymbol{h}-\boldsymbol{g})\right\|_2 \cdot \frac{1}{\sqrt{n_0}}\left\|\boldsymbol{U}^{-1}\left(\boldsymbol{s}_0 - n_0\Psi'\left(\boldsymbol{\theta}_0\right)\right)\right\|_2.\end{aligned}$$

Since $\sqrt{n_0/n}\left\|\boldsymbol{U}(\boldsymbol{h}-\boldsymbol{g})\right\|_2 \to 0$ as $n \to \infty$, if $\sqrt{1/n_0}\left\|\boldsymbol{U}^{-1}\left(\boldsymbol{s}_0 - n_0\Psi'\left(\boldsymbol{\theta}_0\right)\right)\right\|_2$ is bounded, then

$$\inf_{\|\boldsymbol{U}\boldsymbol{h}\|_2^2 \leq R_n, \|\boldsymbol{U}\boldsymbol{g}\|_2^2 \leq R_n} \frac{1}{\sqrt{n}}(\boldsymbol{h}-\boldsymbol{g})^\top \left(\boldsymbol{s}_0 - n_0\Psi'(\boldsymbol{\theta}_0)\right) \to 0 \text{ as } n \to \infty.$$

Therefore,

$$E\left\|\pi^{R_n}(\boldsymbol{\theta}|\boldsymbol{x}) - \mathcal{N}^{R_n}\left(\hat{\boldsymbol{\theta}}, (n\Psi''(\boldsymbol{\theta}_0))^{-1}\right)\right\|_{TV} \to 0 \text{ as } n \to \infty.$$

□

3.3.4 Moderately-sized Neighborhood of $\boldsymbol{\theta_0}$

By combining $d_n = o(R_n)$ in $C1$ and $R_n = o(n)$ in $C3$, we request a moderately-sized neighborhood of $\boldsymbol{\theta}_0$. The radius of the ellipsoid R_n grows faster than the parameter dimension but slower than the sample size. Such a neighborhood can capture sufficient posterior mass to claim posterior consistency.

Lemma 3.4. *If $R_n = o(n)$, then*

$$E\left\|\pi(\boldsymbol{\theta}|\boldsymbol{x}) - \pi^{R_n}(\boldsymbol{\theta}|\boldsymbol{x})\right\|_{TV} \to 0, \quad \textit{as } n \to \infty.$$

Proof of Lemma 3.4:

$$
\begin{aligned}
&\mathbf{E}\|\pi - \pi^{R_n}\|_{TV} = \mathbf{E}\big[\pi\big(\varepsilon^c_{\boldsymbol{\theta}_0,\boldsymbol{U}}(R_n)\big)\big] \\
&= \int_{\boldsymbol{\theta}\in\epsilon^c_{\boldsymbol{\theta}_0,\boldsymbol{U}}(R_n)} \exp\left\{(\boldsymbol{s}_0 + n\bar{\boldsymbol{x}})^\top\boldsymbol{\theta} - (n_0+n)\Psi(\boldsymbol{\theta})\right\}h(n_0,\boldsymbol{s}_0)d\boldsymbol{\theta} \\
&= \int_{\|\boldsymbol{U}\boldsymbol{h}\|_2^2>R_n} \exp\left\{(\boldsymbol{s}_0 + n\bar{\boldsymbol{x}})^\top(\boldsymbol{\theta}_0 + \frac{1}{\sqrt{n}}\boldsymbol{h}) - (n_0+n)\Psi(\boldsymbol{\theta}_0 + \frac{1}{\sqrt{n}}\boldsymbol{h})\right\}h(n_0,\boldsymbol{s}_0)d\boldsymbol{h} \\
&= 1 - \int_{\|\boldsymbol{U}\boldsymbol{h}\|_2^2\le R_n} \exp\left\{(\boldsymbol{s}_0 + n\bar{\boldsymbol{x}})^\top(\boldsymbol{\theta}_0 + \frac{1}{\sqrt{n}}\boldsymbol{h}) - (n_0+n)\Psi(\boldsymbol{\theta}_0 + \frac{1}{\sqrt{n}}\boldsymbol{h})\right\}h(n_0,\boldsymbol{s}_0)d\boldsymbol{h} \\
&= 1 - \int_{\|\boldsymbol{U}\boldsymbol{h}\|_2^2\le R_n} \exp\Bigg\{(\boldsymbol{s}_0 + n\bar{\boldsymbol{x}})^\top(\boldsymbol{\theta}_0 + \frac{1}{\sqrt{n}}\boldsymbol{h}) \\
&\qquad - (n_0+n)\left[\Psi(\boldsymbol{\theta}_0) + \frac{1}{\sqrt{n}}\boldsymbol{h}^\top\Psi'(\boldsymbol{\theta}_0) + \frac{1}{2n}\boldsymbol{h}^\top\Psi''(\boldsymbol{\theta}_0)\boldsymbol{h}\right]\Bigg\}h(n_0,\boldsymbol{s}_0)d\boldsymbol{h} \\
&= 1 - \int_{\|\boldsymbol{U}\boldsymbol{h}\|_2^2\le R_n} \exp\Bigg\{(\boldsymbol{s}_0 + n\bar{\boldsymbol{x}})^\top\boldsymbol{\theta}_0 - (n_0+n)\Psi(\boldsymbol{\theta}_0) \\
&\qquad + \frac{1}{\sqrt{n}}\boldsymbol{h}^\top\big(\boldsymbol{s}_0 + n\bar{\boldsymbol{x}} - n_0\Psi'(\boldsymbol{\theta}_0) - n\Psi'(\boldsymbol{\theta}_0)\big) - \frac{n_0+n}{2n}\boldsymbol{h}^\top\Psi''(\boldsymbol{\theta}_0)\boldsymbol{h}\Bigg\}h(n_0,\boldsymbol{s}_0)d\boldsymbol{h} \\
&= 1 - \int_{\|\boldsymbol{U}\boldsymbol{h}\|_2^2\le R_n} \exp\Bigg\{(\boldsymbol{s}_0 + n\bar{\boldsymbol{x}})^\top\boldsymbol{\theta}_0 - (n_0+n)\Psi(\boldsymbol{\theta}_0) \\
&\qquad + \frac{1}{\sqrt{n}}\boldsymbol{h}^\top\big(\boldsymbol{s}_0 - n_0\Psi'(\boldsymbol{\theta}_0) + n\Psi''(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)\big) - \frac{n_0+n}{2n}\boldsymbol{h}^\top\Psi''(\boldsymbol{\theta}_0)\boldsymbol{h}\Bigg\}h(n_0,\boldsymbol{s}_0)d\boldsymbol{h} \\
&= 1 - \exp\left\{(\boldsymbol{s}_0 + n\bar{\boldsymbol{x}})^\top\boldsymbol{\theta}_0 - (n_0+n)\Psi(\boldsymbol{\theta}_0)\right\}h(n_0,\boldsymbol{s}_0) \\
&\qquad \int_{\|\boldsymbol{U}\boldsymbol{h}\|_2^2\le R_n} \exp\left\{\frac{1}{\sqrt{n}}\boldsymbol{h}^\top\big(\boldsymbol{s}_0 - n_0\Psi'(\boldsymbol{\theta}_0) + n\Psi''(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)\big) - \frac{n_0+n}{2n}\boldsymbol{h}^\top\Psi''(\boldsymbol{\theta}_0)\boldsymbol{h}\right\}d\boldsymbol{h} \\
&\le 1 - \exp\left\{(\boldsymbol{s}_0 + n\bar{\boldsymbol{x}})^\top\boldsymbol{\theta}_0 - (n_0+n)\Psi(\boldsymbol{\theta}_0)\right\}h(n_0,\boldsymbol{s}_0) \\
&\qquad \int_{\|\boldsymbol{U}\boldsymbol{h}\|_2^2\le R_n} \exp\left\{\frac{1}{\sqrt{n}}\boldsymbol{h}^\top\big(\boldsymbol{s}_0 - n_0\Psi'(\boldsymbol{\theta}_0) + n\Psi''(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)\big) - \frac{(n_0+n)R_n}{2n}\right\}d\boldsymbol{h} \\
&= 1 - \exp\left\{(\boldsymbol{s}_0 + n\bar{\boldsymbol{x}})^\top\boldsymbol{\theta}_0 - (n_0+n)\Psi(\boldsymbol{\theta}_0)\right\}h(n_0,\boldsymbol{s}_0) \\
&\qquad \int_{\|\boldsymbol{U}\boldsymbol{h}\|_2^2\le R_n} \exp\left\{\frac{1}{\sqrt{n}}\boldsymbol{h}^\top\big(\boldsymbol{s}_0 - n_0\Psi'(\boldsymbol{\theta}_0)\big) + \sqrt{n}\boldsymbol{h}^\top\Psi''(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) - \frac{(n_0+n)R_n}{2n}\right\}d\boldsymbol{h}.
\end{aligned}
$$

Note that the posterior density evaluated at the true parameter $\boldsymbol{\theta}_0$,

$$\exp\left\{(\boldsymbol{s}_0 + n\bar{\boldsymbol{x}})^\top\boldsymbol{\theta}_0 - (n_0+n)\Psi(\boldsymbol{\theta}_0)\right\}h(n_0,\boldsymbol{s}_0) \to 1 \text{ as } n\to\infty.$$

By Cauchy-Schwarz inequality,

$$\begin{aligned}\frac{1}{\sqrt{n}}\boldsymbol{h}^\top\left(\boldsymbol{s}_0 - n_0\Psi'\left(\boldsymbol{\theta}_0\right)\right) &= \frac{1}{\sqrt{n}}\boldsymbol{h}^\top\boldsymbol{U}\boldsymbol{U}^{-1}\left(\boldsymbol{s}_0 - n_0\Psi'\left(\boldsymbol{\theta}_0\right)\right)\\ &\leq \frac{1}{\sqrt{n}}\left\|\boldsymbol{h}^\top\boldsymbol{U}\right\|_2 \cdot \left\|\boldsymbol{U}^{-1}\left(\boldsymbol{s}_0 - n_0\Psi'\left(\boldsymbol{\theta}_0\right)\right)\right\|_2\\ &\leq \sqrt{\frac{R_n}{n}} \cdot \left\|\boldsymbol{U}^{-1}\left(\boldsymbol{s}_0 - n_0\Psi'\left(\boldsymbol{\theta}_0\right)\right)\right\|_2\\ &\leq \sqrt{\frac{R_n}{n}}|c_2 - n_0|.\end{aligned}$$

If $R_n = o(n)$, then

$$\frac{1}{\sqrt{n}}\boldsymbol{h}^\top\left(\boldsymbol{s}_0 - n_0\Psi'\left(\boldsymbol{\theta}_0\right)\right) \to 0 \text{ as } n \to \infty.$$

Since $n_0 = o(n/R_n)$,

$$\frac{(n_0+n)R_n}{2n} \to \frac{R_n}{2} \text{ as } n \to \infty.$$

Notice that

$$\begin{aligned}&\Pr\left(\sup_{\|\boldsymbol{U}\boldsymbol{h}\|_2^2 \leq R_n} \sqrt{n}\boldsymbol{h}^\top\Psi''(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = \frac{R_n}{2}\right)\\ &= \Pr\left(\sup_{\|\boldsymbol{U}\boldsymbol{h}\|_2^2 \leq R_n} \sqrt{n}\boldsymbol{h}^\top\boldsymbol{U}\boldsymbol{U}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = \frac{R_n}{2}\right)\\ &= \Pr\left(\sup_{\|\boldsymbol{U}\boldsymbol{h}\|_2^2 \leq R_n} \left\|\boldsymbol{h}^\top\boldsymbol{U}\right\|_2 \cdot \left\|\sqrt{n}\boldsymbol{U}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)\right\|_2 = \frac{R_n}{2}\right)\\ &= \Pr\left(\sup_{\boldsymbol{\theta} \in \epsilon_{\boldsymbol{\theta}_0, \boldsymbol{U}}(R_n)} \left\|\sqrt{n}\boldsymbol{U}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)\right\|_2 = \frac{\sqrt{R_n}}{2}\right).\end{aligned}$$

From Lemma 3.2, we have $\Pr\left(n(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^\top\Psi''(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) > R_n/4\right) \to 0$ as $n \to \infty$. Thus,

$$\Pr\left(\sup_{\boldsymbol{\theta} \in \epsilon_{\boldsymbol{\theta}_0, \boldsymbol{U}}(R_n)} \left\|\sqrt{n}\boldsymbol{U}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)\right\|_2 = \frac{\sqrt{R_n}}{2}\right) \to 1 \text{ as } n \to \infty.$$

□

Hence,

$$E\left\|\pi(\boldsymbol{\theta}|\boldsymbol{x})-\pi^{R_n}(\boldsymbol{\theta}|\boldsymbol{x})\right\|_{TV}\leq 1-\exp\left\{\frac{\sqrt{R_n}|c_2-n_0|}{\sqrt{n}}-\frac{n_0R_n}{2n}\right\}.$$

If $n_0=o(n/R_n)$, then as $n\to\infty$,

$$E\left\|\pi(\boldsymbol{\theta}|\boldsymbol{x})-\pi^{R_n}(\boldsymbol{\theta}|\boldsymbol{x})\right\|_{TV}\leq 1-\exp\left\{\frac{\sqrt{R_n}|c_2|}{\sqrt{n}}\right\}.$$

It is worth noting that Bontemps (2011) shows that T_3 converges to 0 at a rate of $\sqrt{r_n}/\sqrt{2\pi}$ for a sequence of positive numbers $\{r_n\}$ with $r_n=o(1)$ and $-\ln(r_n)=o(R_n/d_n)$, whereas we show that T_3 converges exponentially at a rate of $\sqrt{R_n}/\sqrt{n}$. Under our circumstance, T_3 converges somewhat faster. Taking into account the trade-off between the rate of convergence and the growth rate of parameter dimension, we can still acquire posterior concentration by assuming $d_n=o(n)$ rather than $d_n\ln(d_n)=o(n)$ Bontemps (2011).

From Theorem 3.1 one can establish the consistency of the posterior distribution in the following corollary. For more details and proof, we refer to Corollary 2.1 of Ghosal (2000).

Corollary 3.1. *(Posterior consistency) Under the condition of the Theorem 3.1, there is a positive real number c such that the posterior probability of*

$$\left\{\boldsymbol{\theta}:\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|_2\leq\sqrt{cd_n\left\|\left(\Psi''(\boldsymbol{\theta}_0)\right)^{-1}\right\|_2/n}\right\}$$

converges to 1 in probability.

Proof. To demonstrate the corollary, utilize the fact that $\Pr n^{-1/2}\left|\boldsymbol{U}^{-1}\xi\right|_2<\delta$ approximates, according to Theorem 3.1, the posterior probability of $\{\boldsymbol{\theta}:|\boldsymbol{\theta}-\boldsymbol{\theta}_0|\,2<\delta\}$, where ξ follows a normal distribution $\mathcal{N}\left(\boldsymbol{\Delta},\boldsymbol{I}d_n\right)$ and $\Delta=\sqrt{n}\boldsymbol{U}^{-1}(\bar{\boldsymbol{x}}-\boldsymbol{\mu})$. Hence, it suffices to establish that the latter

converges to 1 in probability. Now

$$\begin{aligned} n^{-1/2} \left\| \boldsymbol{U}^{-1} \boldsymbol{\Delta} \right\|_2 &\leq \|\boldsymbol{\Delta}\|_2 \sqrt{\left\| \left(\Psi''(\boldsymbol{\theta}_0)\right)^{-1} \right\|_2 / n} \\ &= O_{d_n}(\sqrt{d_n \left\| \left(\Psi''(\boldsymbol{\theta}_0)\right)^{-1} \right\|_2 / n}) \\ &= o_{d_n}(1), \end{aligned}$$

so that the probability tends to 1 over the entire set,

$$\begin{aligned} \Pr\left(n^{-1/2} \left\| \boldsymbol{U}^{-1} \xi \right\|_2 \geq \delta\right) &\leq \Pr\left(n^{-1/2} \left\| \boldsymbol{U}^{-1} \right\|_2 \|\xi - \boldsymbol{\Delta}\|_2 \geq \frac{\delta}{2}\right) \\ &= \Pr\left(n^{-1} \left\| \left(\Psi''(\boldsymbol{\theta}_0)\right)^{-1} \right\|_2 W \geq \frac{\delta^2}{4}\right), \end{aligned} \tag{3.3.3}$$

where W has a central chi-square distribution with d_n degrees of freedom. As

$$E\left(n^{-1} \left\| \left(\Psi''(\boldsymbol{\theta}_0)\right)^{-1} \right\|_2 W\right) = d_n \left\| \left(\Psi''(\boldsymbol{\theta}_0)\right)^{-1} \right\|_2 / n \to 0,$$

Equation (3.3.3) tends to 0. With a sufficient constant c, end up with the bound $\Pr\left(W > c d_n / 4\right)$ on Equation (3.3.3). The posterior consistency result now follows. □

CHAPTER 4 APPLICATIONS

By demonstrating the Bernstein-von Mises theorem through a simulation study in the Multinomial-Dirichlet model and discussing its applications in Bayesian density estimation and the estimation of the mean of an infinite-dimensional normal distribution, we provide further insight into the practical relevance and versatility of this theorem, particularly when dealing with high-dimensional setup where $d_n = o(n)$.

4.1 Application to the Multinomial-Dirichlet Model

We validate the conditions of Lemma 3.2 and Lemma 3.3 in the multinomial model with the Dirichlet prior case. Meanwhile, the condition of Lemma 3.4 is free of validation since merely the existence of such $\{R_n\}$ is needed.

Denote the true mean vector $\boldsymbol{p} = (p_1, \ldots, p_{d_n})^\top$ and $\boldsymbol{D} = \text{diag}(p_1, \ldots, p_{d_n})$. Then $\Psi''(\boldsymbol{\theta}_0) = \boldsymbol{D} - \boldsymbol{p}\boldsymbol{p}^\top$. Note that the spectral norm of $\Psi'''(\boldsymbol{\theta})$ equals the largest singular value of $\Psi'''(\boldsymbol{\theta})$, see more details in Appendix .3. Since tensor $\Psi'''(\boldsymbol{\theta})$ is symmetric on $\mathbb{R}^{d_n \times d_n \times d_n}$, the largest singular value of $\Psi'''(\boldsymbol{\theta})$ is the same as the largest eigenvalue of $\Psi''(\boldsymbol{\theta})$. Thus,

$$\sup_{\boldsymbol{\theta} \in \varepsilon^c_{\boldsymbol{\theta}_0, \boldsymbol{U}}(R_n/4)} \|\Psi'''(\boldsymbol{\theta})\| = \sup_{\boldsymbol{\theta} \in \varepsilon^c_{\boldsymbol{\theta}_0, \boldsymbol{U}}(R_n/4)} \lambda_{\max}\left(\Psi''(\boldsymbol{\theta})\right),$$

where $\lambda_{\max}\left(\Psi''(\boldsymbol{\theta})\right)$ is the largest eigenvalue of $\Psi''(\boldsymbol{\theta})$. We notice that

$$\sup_{\boldsymbol{\theta} \in \varepsilon^c_{\boldsymbol{\theta}_0, \boldsymbol{U}}(R_n/4)} \lambda_{\max}\left(\Psi''(\boldsymbol{\theta})\right) \leq \lambda_{\max}\left(\Psi''(\boldsymbol{\theta}_0)\right).$$

According to Watson (1996), without loss of generality, we assume $p_1 < \cdots < p_{d_n}$, then there is an eigenvalue of $\Psi''(\boldsymbol{\theta}_0)$ at 0 and there is one eigenvalue of $\Psi''(\boldsymbol{\theta}_0)$ in each gap between the ordered p_is,

$$p_1 \leq \lambda_1 \leq p_2 \leq \lambda_2 \leq p_3 \leq \cdots \leq \lambda_{d_n - 1} \leq p_{d_n}.$$

In the sparse case where many $p_j = 0$, $\Psi''(\boldsymbol{\theta}_0)$ will possess only a few non-zero eigenvalues, and the largest among them will be bounded by the non-zero values of p_{d_n-1} and p_{d_n}. This implies that $\lambda_{\max}\left(\Psi''(\boldsymbol{\theta}_0)\right)$ might converge to a finite value within the range of p_{d_n-1} and p_{d_n} as d_n increases. However, this scenario does not align with our objectives, as the prior concentration condition is not satisfied. Therefore, we assume that only few $p_j = 0$ for $j \in \{1, \ldots, d_n\}$ as $d_n \to \infty$. Then $p_{d_n} \to 0$ as $d_n \to \infty$, which in turn implies $\lambda_{\max}\left(\Psi''(\boldsymbol{\theta}_0)\right) \to 0$ as $n \to \infty$. Hence,

$$\sup_{\boldsymbol{\theta} \in \varepsilon^c_{\boldsymbol{\theta}_0, \boldsymbol{U}}(R_n/4)} \|\Psi'''(\boldsymbol{\theta})\| \to 0 \text{ as } R_n \to \infty.$$

In such a way, the condition of Lemma 3.2 is validated.

In terms of the condition of Lemma 3.3, let $c_1 = \max\limits_{0 \le j \le d_n} \alpha_j - 1$ be a positive constant. Then

$$\begin{aligned}
\frac{n_0 R_n}{n} &= \frac{\left(\sum_{j=0}^{d_n} \alpha_j - d_n - 1\right) R_n}{n} \\
&\le \frac{(d_n + 1)\left(\max\limits_{0 \le j \le d_n} \alpha_j - 1\right) R_n}{n} \\
&= \frac{c_1 (d_n + 1) R_n}{n}.
\end{aligned}$$

Choose R_n such that $R_n = o(n/d_n)$, then $n_0 = o(n/R_n)$. Select $\boldsymbol{s}_0$ of the form $\boldsymbol{s}_0 = c_2 \Psi'(\boldsymbol{\theta}_0)$, where c_2 is a constant. Then

$$\begin{aligned}
\frac{1}{\sqrt{n_0}} \left\| \boldsymbol{U}^{-1} \left(\boldsymbol{s}_0 - n_0 \Psi'\left(\boldsymbol{\theta}_0\right) \right) \right\|_2 &\le \frac{|c_2 - n_0|}{\sqrt{n_0}} \left\| \boldsymbol{U}^{-1} \Psi'(\boldsymbol{\theta}_0) \right\|_2 \\
&\le \frac{\sqrt{d_n} |c_2 - n_0|}{\sqrt{n_0}} \left(\max_{1 \le j \le d_n} \boldsymbol{\phi}^j(\boldsymbol{\theta}_0) \right).
\end{aligned}$$

where $\boldsymbol{\phi}(\boldsymbol{\theta}_0) = \boldsymbol{U}^{-1} \Psi'(\boldsymbol{\theta}_0)$. Since

$$\boldsymbol{U}^{-1} = \boldsymbol{D}^{-1/2} + \frac{\boldsymbol{D}^{-1} \boldsymbol{p} \boldsymbol{p}^\top \boldsymbol{D}^{-1/2}}{1 - \boldsymbol{p}^\top \boldsymbol{D}^{-1} \boldsymbol{p} + \sqrt{1 - \boldsymbol{p}^\top \boldsymbol{D}^{-1} \boldsymbol{p}}} \quad \text{(Ghosal, 2000)}$$

and $\Psi'(\boldsymbol{\theta}_0) = \boldsymbol{p}$, $\max\limits_{1\leq j\leq d_n} \boldsymbol{\phi}^j(\boldsymbol{\theta}_0)$ is of the order $d_n^{-1/2}$. Therefore, $\sqrt{1/n_0}\left\|\boldsymbol{U}^{-1}\left(\boldsymbol{s}_0 - n_0\Psi'\left(\boldsymbol{\theta}_0\right)\right)\right\|_2$ is bounded up by $\left|c_2 - n_0\right|/\sqrt{n_0}$. In the following section, we demonstrate our result for the Multinomial-Dirichlet model by simulation.

4.1.1 Simulation

The Bernstein-von Mises theorem result is demonstrated with simulations for the Multinomial-Dirichlet model under conditions $C1$, $C2$, and $C3$ in Section 3.3.1. This configuration presupposes that the parameter dimension increases as the sample size grows at a rate satisfying $d_n = o(n)$. To validate the performance of the proposed theorem in terms of convergence rate, we compare it with other three scenarios: $d_n^2 = o(n)$ Portnoy (1988) and $d_n \ln(d_n) = o(n)$ Bontemps (2011); $d_n^3 \ln(d_n) = o(n)$ Ghosal (2000); and $d_n^3 = o(n)$ Spokoiny (2013), in terms of the expected total variation distance between the posterior distribution and normal distribution centered at the maximum likelihood estimator with variance, the inverse of the observed Fisher information matrix. For simplicity, we set $d_n^{1.01} = o(n)$ in our theorem to ensure that d_n grows slower than n as n approaches infinity. In other words, the growth rate of d_n is sublinear compared to n. The other three scenarios can be represented as $d_n^2 = o(n)$, $d_n^3 = o(n)$, and $d_n^4 = o(n)$, respectively.

According to the Multinomial-Dirichlet model in Section 3.2.1, the posterior follows a Dirichlet distribution with the hyper-parameter $\boldsymbol{\alpha} + \boldsymbol{x}$, i.e. $\pi(\boldsymbol{p}|\boldsymbol{\alpha} + \boldsymbol{x})$. Because the maximum likelihood estimator of $\boldsymbol{p}$ is $\hat{\boldsymbol{p}} = \boldsymbol{x}/n$, the estimated mean vector of the normal distribution is $\hat{\boldsymbol{\theta}} = \ln\left(\boldsymbol{x}/(n - \sum_{j=1}^{d_n} x_j)\right)$. Since $\Psi(\boldsymbol{\theta}) = \ln\Big[\sum_{j=1}^{d_n}\exp(\theta_j) + 1\Big]$, the estimated covariance matrix of the target normal distribution is $\left(n\Psi''(\hat{\boldsymbol{\theta}})\right)^{-1}$ satisfying

$$\Psi''(\hat{\theta})_{ij} = \begin{cases} \hat{p}_i(1-\hat{p}_i), \text{ if } i = j, \\ -\hat{p}_i\hat{p}_j, \quad \text{if } i \neq j, \end{cases} \qquad i, j \in \{1, \ldots, d_n\}.$$

In our simulations, the total variation distance is computed by Scheffé's lemma

$$\|P - Q\|_{TV} = \frac{1}{2}\int \left|p(\boldsymbol{\theta}) - q(\boldsymbol{\theta})\right| d\nu(\boldsymbol{\theta}) \quad \text{(Tsybakov, 2008)},$$

where $p(\boldsymbol{\theta})$ is the multivariate normal density function of MVN$\left(\hat{\boldsymbol{\theta}}, (n\Psi''(\hat{\boldsymbol{\theta}}))^{-1}\right)$, $q(\boldsymbol{\theta})$ is the density function of Dirichlet$(\boldsymbol{p}|\boldsymbol{\alpha}+\boldsymbol{x})\,|\partial \boldsymbol{p}/\partial \boldsymbol{\theta}|$, and

$$p_i = \frac{\exp(\theta_i)}{1+\sum_{j=1}^{d_n}\exp(\theta_j)} \text{ for } i \in \{1,\ldots,d_n\}.$$

Note that the Jacobian term $|\partial \boldsymbol{p}/\partial \boldsymbol{\theta}|$ appears in the above expression because we are using built-in function in R to generate samples from the Dirichlet distribution and then using Jacobian to convert it to the density of the canonical parameter $\boldsymbol{\theta}$.

Let $\beta \in \{0.01, 1, 2, 3\}$ be a dimensional factor characterizing the interplay between the sample size (n) and dimension (d_n). For each $d_n \in \{15, 30, 100, 500, 1000\}$, the sample size $n = d_n^{1+\beta}$. The hyperparameter $\boldsymbol{\alpha}$ is chosen to be a vector with all elements set to $1 + d_n^{(d_n-1)\beta/d_n}/(d_n+1)$. In this manner, conditions $C1$, $C2$, and $C3$ concerning the prior can be satisfied. All elements in the hyperparameter vector ensure the concentration of the prior, while elements slightly larger than 1 result in a roughly "flat" prior. The true parameter $\boldsymbol{p} = (p_1, \ldots, p_{d_n})^\top$ is generated from the Dirichlet distribution with concentration parameter $\boldsymbol{\alpha}$.

For the visual representation, as outlined in Monard, Nickl, and Paternain (2021), we plot the marginal densities of θ_1, θ_2, and θ_3, generated by 1000 samples from the posterior distribution $\pi(\boldsymbol{\theta}|\boldsymbol{x})$, superimposed with their corresponding target Gaussian curves. As depicted in Figure 4.1 below, the posterior distributions exhibit bell-shaped curves, closely resembling the target normal curves in the histograms.

The expected total variation distances evaluated using a Monte Carlo method, which approximates the values over 1000 runs with 100 observations in each simulation, are given by

$$E\left\|\mathcal{N}\left(\hat{\boldsymbol{\theta}}, (n\Psi''(\boldsymbol{\theta}_0))^{-1}\right) - \pi(\boldsymbol{\theta}|\boldsymbol{x})\right\|_{TV} \approx \frac{1}{1000}\sum_{r=1}^{1000}\left(\frac{1}{100}\sum_{l=1}^{100}\frac{1}{2}\left|p(\boldsymbol{\theta}_l) - q(\boldsymbol{\theta}_l)\right|\right)_r.$$

The parameter $\boldsymbol{p}$ is generated from Dirichlet$(\boldsymbol{p}|\boldsymbol{\alpha})$, and subsequently, the canonical parameter $\boldsymbol{\theta}$ in the above equation is obtained through variable transformation. The approximated expected total variation distances between the posterior distribution and the approximated normal distribution are

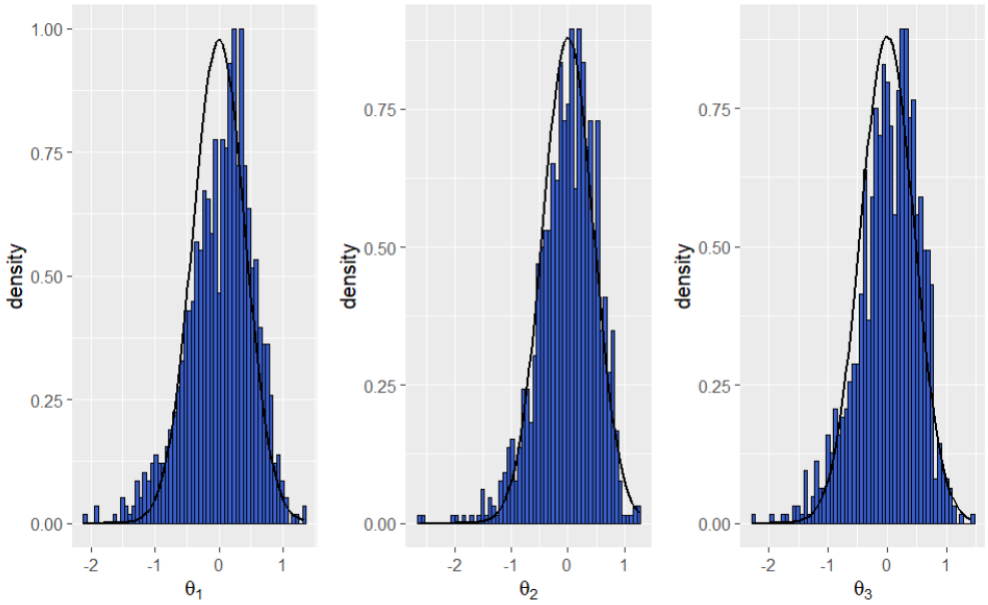


Figure 4.1 Marginal posterior densities of θ_1, θ_2, θ_3 when $d_n = 30$ and $n = 32$

presented in the table below. As the sample size increases, the approximated distances converge to zero in each of the four scenarios. Our simulation results are shown in Table 4.1.

Table 4.1 Approximations of the expected total variation distances $\mathbf{E}\left\|\mathcal{N}\left(\hat{\boldsymbol{\theta}}, (n\Psi''(\boldsymbol{\theta}_0))^{-1}\right) - \pi(\boldsymbol{\theta}|\boldsymbol{x})\right\|_{TV}$ under four scenarios: $n = d_n^{1.01}$; $n = d_n^2$; $n = d_n^3$; $n = d_n^4$ as d_n grows

	$d_n = 15$	$d_n = 30$	$d_n = 100$	$d_n = 500$	$d_n = 1000$
$n = d_n^{1.01}$	8.19e-13	1.03e-26	4.74e-78	0	0
$n = d_n^2$	1.22e-09	1.64e-48	0	0	0
$n = d_n^3$	1.78e-14	3.72e-113	0	0	0
$n = d_n^4$	5.71e-12	1.65e-126	0	0	0

To ensure that the sample size $n = d_n^{1.01}$ is an integer in simulations, we use the ceiling function to round up $d_n^{1.01}$ to the nearest integer. As shown in Table 4.1, when $d_n = 15$, the emergence of asymptotic posterior normality is not yet as apparent as other cases. This is evident from the observed instability in the reduction of approximated expected total variation distances between the posterior distribution and the target Gaussian distribution as the sample size increases. Nevertheless, asymptotic posterior normality is still achieved at a relatively rapid rate when $n =$

$d_n^{1.01}$. Overall, the convergence performance is just as good when using $n = d_n^{1.01}$ as it is for the other three scenarios. This suggests that the result of the Bernstein-von Mises theorem can be attained without imposing a significant constraint on the growth rate of the parameter dimension, such as when $d_n = o(n)$.

4.2 Application to Bayesian Density Estimation

Our result has an interesting connection to a Bayesian density estimation problem. Suppose one wants to estimate a positive Lipschitz continuous density, denoted by f, on the unit interval using a Bayesian method. Suppose $y_1, \ldots, y_n$ is a sample of size n from f. For an integer $d = d_n$ satisfying $d \to \infty$ and $d/n \to 0$, divide the unit interval into $(d+1)$ subintervals $\Delta_0, \Delta_1, \ldots, \Delta_d$ of length $1/(d+1)$. Let $\mathcal{H}_n$ be the set of all histograms on $\{\Delta_0, \Delta_1, \ldots, \Delta_d\}$, and define $p_0, p_1, \ldots, p_d$ to be the probabilities of the subintervals under the density f. For each observation y_i, we can obtain $x_{ij} = I\{y_i \in \Delta_j\}$ for all $j \in \{0, 1, \ldots, d\}$. Therefore, we have a set of i.i.d. multinomial observations $\boldsymbol{x}_i = (x_{i0}, x_{i1}, \ldots, x_{id})^\top$, $i \in 1, \ldots, n$ with $(d+1)$ cells and probabilities $p_0, p_1, \ldots, p_d$. Let

$$f_n(x) = (d+1) \sum_{j=0}^{d} p_j I\{x \in \Delta_j\}$$

be the approximated density. Then $f_n \in \mathcal{H}_n$. Let the true density of f be f_0 and its approximation f_n and cell probabilities p_j's be denoted by f_{0n} and p_{0j}'s, respectively. Suppose the prior has a Dirichlet distribution. The identical model has been verified in Section 4.1, so all three conditions, including $C1$, $C2$, and $C3$, stated in the Theorem 3.1 are satisfied. We refer to Ghosal (2000) to show the consistency of the posterior and achieve the convergence rate. By the definition of f_{0n},

$$\int (f(x) - f_{0n}(x)) \, (f_{0n}(x) - f_0(x)) \, dx = 0.$$

Hence, the error in the estimation of $f_0(x)$ is given by

$$\int \left(f(x) - f_0(x)\right)^2 dx = (d+1)\sum_{j=0}^{d}(p_j - p_{0j})^2 + \int \left(f_{0n}(x) - f_0(x)\right)^2 dx. \tag{4.2.1}$$

The error can be decomposed into two terms as shown in Equation (4.2.1). The first term is related to the discrepancy between the prior and the true distribution of $\boldsymbol{\theta}$, while the second term measures the difference between the estimated density $f_{0n}(x)$ and the true density $f_0(x)$. Because p_j and p_{0j}, $j \in \{0, 1, \ldots, d\}$, are of the order d^{-1}, for a generic positive constant c_3,

$$\begin{aligned} |p_j - p_{0j}| &= \left| \frac{\exp(\theta_j)}{1 + \sum_{k=1}^{d}\exp(\theta_k)} - \frac{\exp(\theta_{0j})}{1 + \sum_{k=1}^{d}\exp(\theta_{0k})} \right| \\ &\leq \sqrt{c_3 d^{-4} \left\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\right\|_2^2}. \end{aligned}$$

Given $\delta > 0$, on $\left\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\right\|_2 > \delta$, for a constant c_4,

$$(d+1)\sum_{j=0}^{d}(p_j - p_{0j})^2 < c_4 d^{-2}\left\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\right\|_2^2.$$

In line with the result presented in Corollary 3.1 and the condition $\left\|\left(\Psi''(\boldsymbol{\theta}_0)\right)^{-1}\right\|_2 = O\left(d^2\right)$, the posterior probability of the set

$$\left\{\boldsymbol{\theta} : (d+1)\sum_{j=0}^{d}(p_j - p_{0j})^2 \leq c_5 d/n\right\}$$

converges to 1 in probability, where c_5 is a sufficiently large constant. It's worth noting that in Equation (4.2.1), the second term

$$\int \left(f_{0n}(x) - f_0(x)\right)^2 dx = O\left(d^{-2}\right).$$

By choosing $d = n^{1-\epsilon}$, where $0 < \epsilon < 1$ so that $d/n \to 0$, the integral $\int \left(f(x) - f_0(x)\right)^2 dx$ converges. Therefore, for a sufficiently large constant c_6, given a random sample of $y_1, \ldots, y_n$,

(i) if $\epsilon \geq 2/3$, then the posterior probability of the set

$$\left\{ f : \int \left(f(x) - f_0(x) \right)^2 dx \leq c_6 n^{-2(1-\epsilon)} \right\}$$

converges to 0 in probability;

(ii) if $\epsilon < 2/3$, then the posterior probability of the set

$$\left\{ f : \int \left(f(x) - f_0(x) \right)^2 dx \leq c_6 n^{-\epsilon} \right\}$$

converges to 0 in probability.

Furthermore, this result can be extended to Hölder classes of order α with an optimal convergence rate of $n^{-\alpha/(2\alpha+1)}$, as explained in Wong and Shen (1995). A Lipschitz function corresponds to the special case where $\alpha = 1$, resulting in $\epsilon = 1/3$ or $5/6$. In general, these findings provide valuable insights into the convergence properties of posterior in Bayesian density estimation.

4.3 Application to the Estimation of the Mean of an Infinite Dimensional Normal Distribution

Assume that we have n i.i.d. random samples $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ from an infinite dimensional normal distribution with mean $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots)^\top$ and covariance as the identity operator on

$$\mathcal{L}_2 = \left\{ (y_1, y_2, \ldots)^\top : \sum_{j=1}^{\infty} y_j^2 < \infty \right\}.$$

Ghosal (2000) noted that although the normal approximation to the posterior distribution of the infinite dimensional parameter does not hold, posterior distribution of a sequence of parametric functions that depend only on $\theta_1, \ldots, \theta_d$ may be approximated using the normal distribution under the condition of "$d^3(\ln d)/n$ is small".

In particular, Ghosal (2000) showed that by assigning a prior on first d components of $\boldsymbol{\theta}$ and the rest to 0, the posterior distribution converges at the rate $n^{-q/(2q+1)}$, which is in line with Pinsker (1980) who provided the minimax rate of convergence is $n^{-q/(2q+1)}$ on the ellipsoid

$\left\{\boldsymbol{\theta} : \sum_{j=1}^{\infty} j^{2q}\theta_j^2 \leq Q\right\}$. Towards this end, Ghosal (2000) considered independent priors on the components of $\boldsymbol{\theta}$ whose logarithms satisfy Lipschitz condition where the Lipschitz constant is tailored to meet the assumptions of the theorem.

Likewise we also consider independent priors on the components of $\boldsymbol{\theta}$ which will lead to the prior

$$\begin{aligned}\pi(\boldsymbol{\theta}; n_0, \boldsymbol{s}_0) &= \exp\left\{\boldsymbol{s}_0^\top \boldsymbol{\theta} - n_0 \boldsymbol{\theta}^\top \boldsymbol{\theta}\right\} \\ &= \prod_{j=1}^{d} \exp\left\{s_{0j}\theta_j - n_0\theta_j^2\right\},\end{aligned}$$

where $\Psi(\boldsymbol{\theta}) = \boldsymbol{\theta}^\top\boldsymbol{\theta}$, $\boldsymbol{\Psi}'(\boldsymbol{\theta}) = 2\boldsymbol{\theta}, \boldsymbol{\Psi}''(\boldsymbol{\theta}) = 2\boldsymbol{I}_d$, and $\boldsymbol{\Psi}'''(\boldsymbol{\theta}) = \mathbf{0}_{d\times d\times d}$. With such a choice, the conditions $C1$ and $C2$ related to Ψ are trivially satisfied. Let $\boldsymbol{\theta}_0$ denote the true mean and $\boldsymbol{\theta}_{0,n} = (\theta_1, \ldots, \theta_d, 0, 0, \ldots)^\top$. Then following the calculations of Ghosal (2000)[p.65], one can get $\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}_{0,n}\|_2 = O\left(d^{-2q}\right)$, and finally achieve the posterior convergence rate to be $n^{-q/(2q+1)}$ with a similar choice of the dimension, $d = n^{1/(2q+1)}$ for which our condition $d = o(n)$ is also satisfied.

The Bernstein-von Mises theorem is a fundamental result in Bayesian statistics that establishes a connection between Bayesian inference and Frequentist asymptotic theory. It provides a way to approximate the posterior distribution with a Gaussian distribution as the sample size increases.

This work presents a high-dimensional Bernstein-von Mises theorem, which delineates the necessary conditions on the prior to attain asymptotic posterior normality. Specifically, it centers on the Diaconis-Ylvisaker prior and presupposes that the problem's dimensionality, denoted as d_n, grows sublinearly with the sample size n, expressed as $d_n = o(n)$. Three modest conditions are delineated, mandating the prior to concentrate and retain a flat profile within a reasonably sized vicinity of the true parameter value $\boldsymbol{\theta}_0$.

The Multinomial-Dirichlet model is a widely used statistical model that describes the distribution of counts across multiple categories. By conducting a simulation study within this model, we can assess the behavior of Bayesian estimators and examine the convergence properties of pos-

terior distributions. The effectiveness of these conditions is exemplified through the Multinomial-Dirichlet model, suggesting that asymptotic posterior normality is attainable when there exists a linear relationship between the parameter dimension d_n and the sample size n. This more lenient condition on the parameter dimension in high-dimensional settings broadens the theorem's applicability.

Furthermore, the practical applications of Bayesian density estimation and the estimation of the mean of an infinite-dimensional normal distribution offer valuable insights into the utility of the Bernstein-von Mises theorem, particularly when the dimensionality of the problem, denoted as d_n, is sublinear in the sample size n, represented by $d_n = o(n)$. This scenario arises when the number of parameters or features grows at a slower rate than the number of observations, which is common in many modern statistical problems.

In Bayesian density estimation, the goal is to estimate the underlying probability density function of a random variable. The Bernstein-von Mises theorem facilitates this task by providing a way to approximate the posterior density with a Gaussian density. This approximation allows for efficient computation and facilitates subsequent inference tasks, such as computing credible intervals or conducting hypothesis tests.

Similarly, the estimation of the mean of an infinite-dimensional normal distribution is a challenging problem that arises in various fields, including functional data analysis and Bayesian nonparametric statistics. The Bernstein-von Mises theorem enables us to derive asymptotically valid confidence intervals for the mean parameter in this infinite-dimensional setting. By leveraging the Gaussian approximation provided by the theorem, we can make robust and reliable inferences about the unknown mean.

Importantly, the utility of the Bernstein-von Mises theorem under the condition $d_n = o(n)$ highlights its effectiveness in high-dimensional statistical problems. In such scenarios, traditional asymptotic results may fail due to the curse of dimensionality.

The Bernstein-von Mises theorem revolutionizes Bayesian inference by facilitating the approximation of complex posterior distributions with Gaussian densities. This Gaussian approxima-

tion significantly simplifies computational tasks, rendering Bayesian analysis more accessible and efficient. By replacing intricate posterior distributions with Gaussian approximations, the theorem enables practitioners to swiftly compute credible intervals and conduct hypothesis tests, essential components of statistical inference. Moreover, in complex scenarios like estimating the mean of an infinite-dimensional normal distribution prevalent in fields such as functional data analysis and Bayesian nonparametric statistics, the theorem furnishes asymptotically valid confidence intervals for the mean parameter. Relying on the Gaussian approximation provided by the theorem ensures robust and dependable inferences concerning the unknown mean.

To enhance the relevance of the theorem across a wider range of scenarios, forthcoming research endeavors could dig into formulating an extended high-dimensional Bernstein-von Mises theorem tailored specifically for exponential family models incorporating non-conjugate priors. This expansion would facilitate a more comprehensive understanding of the behavior of posterior distributions in complex, high-dimensional settings. Additionally, there is a need to develop more efficient methodologies for managing the intricate interplay between the dimensionality of the problem d_n, and the sample size n. By devising innovative strategies that strike a balance between computational complexity and statistical accuracy, Bayesian statisticians can enhance the theorem's practical utility and applicability in real-world contexts.

In addition, exploring the Bernstein-von Mises theorem within the framework of misspecified models and delineating conditions that ensure the persistence of convergence properties is an area of significant interest. Future research in this realm requires an in-depth examination of several critical aspects. Primarily, there is a necessity for deeper theoretical inquiries to construct more comprehensive frameworks capable of characterizing the behavior of the posterior distribution under diverse forms of model misspecification. This entails understanding how different types and degrees of misspecification affect the convergence properties of the posterior distribution. Additionally, investigating the robustness properties of Bayesian inference methodologies under model misspecification remains pivotal. Bayesian statisticians can examine the circumstances where Bayesian inference retains validity, even when the assumed model diverges from the

true data-generating process. Furthermore, empirical studies and applications across various fields offer valuable insights into the practical implications of model misspecification on Bayesian inference. Such investigations can inform the development of robust Bayesian methodologies better suited to navigate real-world complexities and uncertainties.

BIBLIOGRAPHY

Bernardo, J. M. (1979). Reference posterior distributions for Bayesian inference. *Journal of the Royal Statistical Society Series B: Statistical Methodology 41*(2), 113–128.

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

Bontemps, D. (2011). Bernstein-von Mises theorems for Gaussian regression with increasing number of regressors. *The Annals of Statistics 39*(5), 2557–2584.

Boucheron, S. and E. Gassiat (2009). A Bernstein-von Mises Theorem for discrete probability distributions. *Electronic Journal of Statistics 3*, 114–148.

Cam, L. L. (1960). An approximation theorem for the Poisson binomial distribution. *Pacific Journal of Mathematics 10*(4), 1181–1197.

Castillo, I. (2012). A semiparametric Bernstein-von Mises theorem for Gaussian process priors. *Probability Theory and Related Fields 152*(1), 53–99.

Castillo, I. and R. Nickl (2013). Nonparametric Bernstein-von Mises theorems in Gaussian white noise. *The Annals of Statistics 41*(4), 1999–2028.

Castillo, I. and R. Nickl (2014). On the Bernstein-von Mises phenomenon for nonparametric Bayes procedures. *The Annals of Statistics 42*(5), 1941–1969.

Clarke, B. and S. Ghosal (2010). Reference priors for exponential families with increasing dimension. *Electronic Journal of Statistics 4*, 737–780.

Cox, D. D. (1993). An Analysis of Bayesian Inference for Nonparametric Regression. *The Annals of Statistics*, 903–923.

Devroye, L., A. Mehrabian, and T. Reddad (2023). The total variation distance between high-dimensional Gaussians with the same mean. *arXiv preprint arXiv:1810.08693*.

Diaconis, P. and D. Freedman (1986). On the consistency of Bayes estimates. *The Annals of Statistics*, 1–26.

Diaconis, P. and D. Ylvisaker (1979). Conjugate Priors for Exponential Families. *The Annals of Statistics*, 269–281.

Freedman, D. (1999). Wald Lecture: On the Bernstein-von Mises theorem with infinite-dimensional parameters. *The Annals of Statistics 27*(4), 1119–1141.

Freedman, D. A. (1963). On the asymptotic behavior of Bayes' estimates in the discrete case. *The Annals of Mathematical Statistics 34*(4), 1386–1403.

Freedman, D. A. (1965). On the asymptotic behavior of Bayes' estimates in the discrete case II. *The Annals of Mathematical Statistics 36*(2), 454–456.

Gelfand, A. E. and A. F. Smith (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association 85*(410), 398–409.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin (2004). *Bayesian Data Analysis*. Chapman and Hall/CRC.

Ghosal, S. (1999). Asymptotic normality of posterior distributions in high-dimensional linear models. *Bernoulli*, 315–331.

Ghosal, S. (2000). Asymptotic Normality of Posterior Distributions for Exponential Families when the Number of Parameters Tends to Infinity. *Journal of Multivariate Analysis 74*(1), 49–68.

Ghosal, S., J. K. Ghosh, and A. W. Van Der Vaart (2000). Convergence rates of posterior distributions. *Annals of Statistics*, 500–531.

Ghosh, J. K., M. Delampady, and T. Samanta (2006). *An Introduction to Bayesian Analysis: Theory and Methods*, Volume 725. Springer.

Ghosh, J. K., S. Ghosal, and T. Samanta (1994). Stability and Convergence of the Posterior in Non-Regular Problems. In *Statistical Decision Theory and Related Topics V*, pp. 183–199. Springer.

Ghosh, M. (2021). *Bayesian methods for finite population sampling*. Routledge.

Giné, E. and R. Nickl (2021). *Mathematical Foundations of Infinite-Dimensional Statistical Models*. Cambridge University Press.

Hoeting, J. A., D. Madigan, A. E. Raftery, and C. T. Volinsky (1999). Bayesian model averaging: a tutorial (with comments). *Statistical Science 14*(4), 382–417.

Jeffreys, H. (1961). *Theory of Probability*. Oxford, England: Clarendon Press.

Jin, X., A. Bhattacharya, and R. P. Ghosh (2024). High-dimensional Bernstein-von Mises theorem for the Diaconis-Ylvisaker prior. *Journal of Multivariate Analysis 200*, 105279.

Johndrow, J. and A. Bhattacharya (2018). Optimal Gaussian Approximations to the Posterior for Log-Linear Models with Diaconis-Ylvisaker Priors. *Bayesian Analysis 13*(1), 201–223.

Johnson, V. E. and D. Rossell (2010). On the use of non-local prior densities in bayesian hypothesis tests. *Journal of the Royal Statistical Society Series B: Statistical Methodology 72*(2), 143–170.

Johnstone, I. M. (2010). High dimensional Bernstein-von Mises: simple examples. *Institute of Mathematical Statistics Collections 6*, 87.

Le Cam, L. and G. L. Yang (2000). *Asymptotics in Statistics: Some Basic Concepts*. Springer Science & Business Media.

Leahu, H. (2011). On the Bernstein-von Mises phenomenon in the Gaussian white noise model. *Electronic Journal of Statistics 5*, 373–404.

Monard, F., R. Nickl, and G. P. Paternain (2021). Statistical guarantees for Bayesian uncertainty quantification in nonlinear inverse problems with Gaussian process priors. *The Annals of Statistics 49*(6), 3255–3298.

Morris, C. N. (1983). Parametric empirical bayes inference: theory and applications. *Journal of the American Statistical Association 78*(381), 47–55.

Pinsker, M. (1980). Optimal filtration of square-integrable signals in Gaussian noise. *Problems Inform Transmission 16*(2), 120–133.

Pinsker, M. S. (1964). *Information and Information Stability of Random Variables and Processes*. San Francisco: Holden-Day.

Portnoy, S. (1984). Asymptotic Behavior of M-Estimators of p Regression Parameters when p^2/n is Large. I. Consistency. *The Annals of Statistics*, 1298–1309.

Portnoy, S. (1985). Asymptotic Behavior of M Estimators of p Regression Parameters when p^2/n is Large; II. Normal Approximation. *The Annals of Statistics 13*(4), 1403–1417.

Portnoy, S. (1988). Asymptotic behavior of likelihood methods for exponential families when the number of parameters tends to infinity. *The Annals of Statistics*, 356–366.

Qi, L. and S. Hu (2019). Spectral norm and nuclear norm of a third order tensor. *arXiv preprint arXiv:1909.01529*.

Ray, K. (2017, December). Bernstein-von Mises theorems for adaptive Bayesian nonparametric procedures. *The Annals of Statistics 45*(6).

Rivoirard, V. and J. Rousseau (2012). Bernstein-von Mises theorem for linear functionals of the density. *The Annals of Statistics 40*(3), 1489–1523.

Shen, X. (2002). Asymptotic normality of semiparametric and nonparametric posterior distributions. *Journal of the American Statistical Association 97*(457), 222–235.

Spokoiny, V. (2013). Bernstein-von Mises theorem for growing parameter dimension. *arXiv preprint arXiv:1302.3430*.

Tsybakov, A. (2008). *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer, New York.

Van der Vaart, A. W. (2000). *Asymptotic Statistics*, Volume 3. Cambridge University Press.

Von Mises, R. (1981). *Probability, Statistics, and Truth*. Courier Corporation.

Wald, A. (1949). Statistical Decision Functions. *The Annals of Mathematical Statistics 20*(2), 165–205.

Watson, G. S. (1996). Spectral Decomposition of the Covariance Matrix of a Multinomial. *Journal of the Royal Statistical Society: Series B (Methodological) 58*(1), 289–291.

Wong, W. H. and X. Shen (1995). Probability Inequalities for Likelihood Ratios and Convergence Rates of Sieve MLEs. *The Annals of Statistics*, 339–362.

APPENDIX A PROOFS

.1 Proof of Lemma 3.2

Proof.

$$
\begin{aligned}
&\Pr\left(n(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)^\top \Psi''(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0) > R_n/4\right)\\
&= \Pr\left(\bar{\boldsymbol{x}} \in \left\{\Psi'(\boldsymbol{\theta}) : \boldsymbol{\theta} \in \varepsilon^c_{\boldsymbol{\theta}_0,\boldsymbol{U}}(R_n/4)\right\}\right)\\
&= \Pr\left(\sqrt{n}\boldsymbol{U}^{-1}\left(\bar{\boldsymbol{x}}-\Psi'(\boldsymbol{\theta}_0)\right) \in \left\{\sqrt{n}\boldsymbol{U}^{-1}\left(\Psi'(\boldsymbol{\theta})-\Psi'(\boldsymbol{\theta}_0)\right) : \boldsymbol{\theta} \in \varepsilon^c_{\boldsymbol{\theta}_0,\boldsymbol{U}}(R_n/4)\right\}\right)\\
&= \Pr\left(\sqrt{n}\boldsymbol{U}^{-1}\left(\bar{\boldsymbol{x}}-\Psi'(\boldsymbol{\theta}_0)\right) = \sqrt{n}\boldsymbol{U}^{-1}\left(\Psi'(\boldsymbol{\theta})-\Psi'(\boldsymbol{\theta}_0)\right) \text{ for some } \boldsymbol{\theta} \in \varepsilon^c_{\boldsymbol{\theta}_0,\boldsymbol{U}}(R_n/4)\right).
\end{aligned}
$$

Denote set $A = \left\{\sqrt{n}\boldsymbol{U}^{-1}\left(\bar{\boldsymbol{x}}-\Psi'(\boldsymbol{\theta_0})\right) = \sqrt{n}\boldsymbol{U}^{-1}\left(\Psi'(\boldsymbol{\theta})-\Psi'(\boldsymbol{\theta}_0)\right) \text{ for some } \boldsymbol{\theta} \in \varepsilon^c_{\boldsymbol{\theta}_0,\boldsymbol{U}}(R_n/4)\right\}$. Then

$$
\begin{aligned}
E\left(\left\|\sqrt{n}\boldsymbol{U}^{-1}\left(\bar{\boldsymbol{x}}-\Psi'(\boldsymbol{\theta}_0)\right)\right\|_2^2\right) &\geq E\left(\left\|\sqrt{n}\boldsymbol{U}^{-1}\left(\bar{\boldsymbol{x}}-\Psi'(\boldsymbol{\theta}_0)\right)\right\|_2^2 \mathbf{1}_A\right)\\
&\geq E\left(\inf_{\boldsymbol{\theta}\in\varepsilon^c_{\boldsymbol{\theta}_0,\boldsymbol{U}}(R_n/4)}\left\|\sqrt{n}\boldsymbol{U}^{-1}\left(\Psi'(\boldsymbol{\theta})-\Psi'(\boldsymbol{\theta}_0)\right)\right\|_2^2 \mathbf{1}_A\right)\\
&= E\left(\mathbf{1}_A\right)\inf_{\boldsymbol{\theta}\in\varepsilon^c_{\boldsymbol{\theta}_0,\boldsymbol{U}}(R_n/4)}\left\|\sqrt{n}\boldsymbol{U}^{-1}\left(\Psi'(\boldsymbol{\theta})-\Psi'(\boldsymbol{\theta}_0)\right)\right\|_2^2\\
&= \Pr\left(A\right)\inf_{\boldsymbol{\theta}\in\varepsilon^c_{\boldsymbol{\theta}_0,\boldsymbol{U}}(R_n/4)}\left\|\sqrt{n}\boldsymbol{U}^{-1}\left(\Psi'(\boldsymbol{\theta})-\Psi'(\boldsymbol{\theta}_0)\right)\right\|_2^2.
\end{aligned}
$$

Thus,

$$
\Pr\left(A\right) \leq \frac{E\left(\left\|\sqrt{n}\boldsymbol{U}^{-1}\left(\bar{\boldsymbol{x}}-\Psi'(\boldsymbol{\theta}_0)\right)\right\|_2^2\right)}{\inf_{\boldsymbol{\theta}\in\varepsilon^c_{\boldsymbol{\theta}_0,\boldsymbol{U}}(R_n/4)}\left\|\sqrt{n}\boldsymbol{U}^{-1}\left(\Psi'(\boldsymbol{\theta})-\Psi'(\boldsymbol{\theta}_0)\right)\right\|_2^2}.
$$

By Central Limit Theorem, $\bar{\boldsymbol{x}} \sim N\left(\Psi'(\boldsymbol{\theta}_0), \Psi''(\boldsymbol{\theta}_0)/n\right)$ as $n \to \infty$, so $\left\|\sqrt{n}\boldsymbol{U}^{-1}\left(\bar{\boldsymbol{x}}-\Psi'(\boldsymbol{\theta}_0)\right)\right\|_2^2$ is a Pearson's chi-square statistic. Thus, $E\left(\left\|\sqrt{n}\boldsymbol{U}^{-1}\left(\bar{\boldsymbol{x}}-\Psi'(\boldsymbol{\theta}_0)\right)\right\|_2^2\right) = d_n$. So we have

$$
\frac{E\left(\left\|\sqrt{n}\boldsymbol{U}^{-1}\left(\bar{\boldsymbol{x}}-\Psi'(\boldsymbol{\theta}_0)\right)\right\|_2^2\right)}{\inf_{\boldsymbol{\theta}\in\varepsilon^c_{\boldsymbol{\theta}_0,\boldsymbol{U}}(R_n/4)}\left\|\sqrt{n}\boldsymbol{U}^{-1}\left(\Psi'(\boldsymbol{\theta})-\Psi'(\boldsymbol{\theta}_0)\right)\right\|_2^2} = \frac{d_n}{\inf_{\boldsymbol{\theta}\in\varepsilon^c_{\boldsymbol{\theta}_0,\boldsymbol{U}}(R_n/4)}\left\|\sqrt{n}\boldsymbol{U}^{-1}\left(\Psi'(\boldsymbol{\theta})-\Psi'(\boldsymbol{\theta}_0)\right)\right\|_2^2}.
$$

By the first-order Taylor's approximation, if $\sup_{\boldsymbol{\theta}\in\varepsilon^c_{\boldsymbol{\theta}_0,\boldsymbol{U}}(R_n/4)} \|\Psi'''(\boldsymbol{\theta})\| \to 0$ as $R_n \to \infty$, then

$$\begin{aligned}\frac{d_n}{\inf_{\boldsymbol{\theta}\in\varepsilon^c_{\boldsymbol{\theta}_0,\boldsymbol{U}}(R_n/4)} \left\|\sqrt{n}\boldsymbol{U}^{-1}\left(\Psi'(\boldsymbol{\theta})-\Psi'(\boldsymbol{\theta}_0)\right)\right\|_2^2} &\approx \frac{d_n}{\inf_{\boldsymbol{\theta}\in\varepsilon^c_{\boldsymbol{\theta}_0,\boldsymbol{U}}(R_n/4)} \left\|\sqrt{n}\boldsymbol{U}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)\right\|_2^2} \\ &= \frac{d_n}{\inf_{\boldsymbol{\theta}\in\varepsilon^c_{\boldsymbol{\theta}_0,\boldsymbol{U}}(R_n/4)} n(\boldsymbol{\theta}-\boldsymbol{\theta}_0)^\top\Psi''(\boldsymbol{\theta}_0)(\boldsymbol{\theta}-\boldsymbol{\theta}_0)} \\ &= \frac{d_n}{R_n/4}.\end{aligned}$$

Since $d_n = o(R_n)$, $\frac{d_n}{R_n/4} \to 0$ as $n \to \infty$. Hence,

$$\Pr\left(n(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)^\top\Psi''(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0) > R_n/4\right) \to 0 \text{ as } n \to \infty.$$

□

.2 Key Idea of the Bernstein-von Mises Theorem Proof

This section outlines the key idea behind the proof of the Bernstein-von Mises theorem as presented in Van der Vaart (2000), a fundamental reference that has significantly influenced subsequent versions of the Bernstein-von Mises theorem. For the observation $\boldsymbol{X}$, it possesses an asymptotic equivalent of the "locally sufficient" statistics

$$\Delta_{n,\boldsymbol{\theta}_0} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \boldsymbol{I}_{\boldsymbol{\theta}_0}^{-1}\ell_{\boldsymbol{\theta}_0}(X_i),$$

where $\ell_{\boldsymbol{\theta}}$ represents the score function of the model, i.e., the derivative of the log-likelihood. According to the Bernstein-von Mises theorem, there is a tendency towards 0 for the total variance distance between the posterior distribution of $\sqrt{n}(\boldsymbol{\theta}_n - \boldsymbol{\theta}_0)$ and $\mathcal{N}(\Delta_{n,\boldsymbol{\theta}_0}, \boldsymbol{I}_{\boldsymbol{\theta}_0}^{-1})$. As $\Delta_{n,\boldsymbol{\theta}_0}$ approaches $\boldsymbol{X}$, the posterior distribution of $\sqrt{n}(\boldsymbol{\theta}_n - \boldsymbol{\theta}_0)$ converges to $\mathcal{N}(\boldsymbol{X}, \boldsymbol{I}_{\boldsymbol{\theta}_0}^{-1})$. Therefore, in differentiable parametric models, the heuristic argument implies that the posterior distribution converges to the Gaussian distribution $\mathcal{N}(\boldsymbol{X}, \boldsymbol{I}_{\boldsymbol{\theta}_0}^{-1})$.

The subsequent version of the Bernstein-von Mises theorem has weak requirements on the prior. It assumes the presence of a sequence of uniformly consistent tests for testing $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ against $H_1 : |\boldsymbol{\theta} - \boldsymbol{\theta}_0|_2 \geq \epsilon$, for every $\epsilon > 0$. This is in addition to the need for differentiability in the quadratic mean of the model. Balls centered at the true value $\boldsymbol{\theta}_0$ should be able to separate $\boldsymbol{\theta}_0$ from the complements of the balls. In other words, by introducing a neighborhood around the true value $\boldsymbol{\theta}_0$, the entire parameter space is divided into two sub-spaces: the neighborhood containing $\boldsymbol{\theta}_0$ and its complement, which does not contain $\boldsymbol{\theta}_0$. This separation hypothesis appears rather logical, as the theory suggests the eventual concentration of posterior distributions on balls of radii $R_n/\sqrt{n}$ around $\boldsymbol{\theta}_0$, for every $R_n \to \infty$.

Separation by tests of $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ from $H_1 : |\boldsymbol{\theta} - \boldsymbol{\theta}_0|_2 \geq \epsilon$ for a single $\epsilon > 0$ already implies separation for every $\epsilon > 0$, under the assumptions of continuity and identifiability of the model. Even without the separation condition, the model remains valid, and the inference can be carried out reliably if $\boldsymbol{\Theta}$ is compact and the model is both continuous and identifiable.

Theorem 10.1 (Van der Vaart, 2000) [p.141] Let the experiment $(P_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \boldsymbol{\Theta})$ be differentiable in quadratic mean at $\boldsymbol{\theta}_0$ with nonsingular Fisher information matrix $\boldsymbol{I}_{\theta_0}$, and suppose that for every $\epsilon > 0$ there exists a sequence of tests ϕ_n such that

$$P^n_{\boldsymbol{\theta}}\phi_n \to 0, \qquad \sup_{\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|_2 \geq \epsilon} P^n_{\boldsymbol{\theta}}(1 - \phi_n) \to 0.$$

Furthermore, let the prior measure be absolutely continuous in a neighborhood of $\boldsymbol{\theta}_0$ with a continuous positive density at $\boldsymbol{\theta}_0$. Then the corresponding posterior distributions satisfy

$$\left\| P_{\sqrt{n}(\boldsymbol{\theta}_n - \boldsymbol{\theta}_0)|X_1,\ldots,X_n} - \mathcal{N}(\Delta_{n,\boldsymbol{\theta}_0}, \boldsymbol{I}^{-1}_{\boldsymbol{\theta}}) \right\| \to 0 \text{ in } P^n_{\boldsymbol{\theta}_0}.$$

For details of the proof, interested readers can refer to Van der Vaart (2000), p. 141-143. Instead of working on the entire parameter space, the concept of a neighborhood around the true parameter $\boldsymbol{\theta}_0$ is introduced creatively. This approach allows the proof to be separated into two parts. Firstly, demonstrating that the posterior, truncated by the neighborhood of the true

parameter, converges to the complete posterior. Secondly, showing that the truncated posterior converges to the target normal distribution centered at the maximum likelihood estimator, with variance given by the inverse of the observed Fisher information matrix. These two convergences occur simultaneously for certain neighborhoods with a growing radius. Naturally, the Bernstein-von Mises result follows.

.3 Some Useful Results

Suppose that real number λ is a singular value of a third order tensor $\mathcal{B} = (b_{ijk})$, for all $i \in \{1, \ldots, d_1\}$, $j \in \{1, \ldots, d_2\}$, and $k \in \{1, \ldots, d_3\}$. Then singular vectors $\boldsymbol{x} = (x_1, \ldots, x_{d_1})^\top \in \mathbb{R}^{d_1}$, $\boldsymbol{y} = (y_1, \ldots, y_{d_2})^\top \in \mathbb{R}^{d_2}$, $\boldsymbol{z} = (z_1, \ldots, z_{d_3})^\top \in \mathbb{R}^{d_3}$ satisfy the following equations (Qi and Hu, 2019):

(i) $\sum_{j=1}^{d_2} \sum_{k=1}^{d_3} b_{ijk} y_j z_k = \lambda x_i$ for $i \in \{1, \ldots, d_1\}$;

(ii) $\sum_{i=1}^{d_1} \sum_{k=1}^{d_3} b_{ijk} x_i z_k = \lambda y_j$ for $j \in \{1, \ldots, d_2\}$;

(iii) $\sum_{i=1}^{d_1} \sum_{j=1}^{d_2} b_{ijk} x_i y_j = \lambda z_k$ for $k \in \{1, \ldots, d_3\}$;

(iv) $\boldsymbol{x}^\top \boldsymbol{x} = \boldsymbol{y}^\top \boldsymbol{y} = \boldsymbol{z}^\top \boldsymbol{z} = 1$.

Thus,

$$\sum_{i=1}^{d_1} \sum_{j=1}^{d_2} \sum_{k=1}^{d_3} b_{ijk} x_i y_j z_k = \lambda.$$

Notice that

$$\langle \mathcal{B}, \boldsymbol{x} \otimes \boldsymbol{y} \otimes \boldsymbol{z} \rangle = \sum_{i=1}^{d_1} \sum_{j=1}^{d_2} \sum_{k=1}^{d_3} b_{ijk} x_i y_j z_k.$$

The spectral norm of $\mathcal{B}$,

$$\begin{aligned}\|\mathcal{B}\| &= \sup\left\{\langle \mathcal{B}, \boldsymbol{x}\otimes\boldsymbol{y}\otimes\boldsymbol{z}\rangle : \boldsymbol{x}^\top\boldsymbol{x} = \boldsymbol{y}^\top\boldsymbol{y} = \boldsymbol{z}^\top\boldsymbol{z} = 1\right\} \\ &= \sup \sum_{i=1}^{d_1}\sum_{j=1}^{d_2}\sum_{k=1}^{d_3} b_{ijk}x_i y_j z_k \\ &= \lambda_{\max},\end{aligned}$$

where $\lambda_{\max}$ is the largest singular value of $\mathcal{B}$. Hence, the spectral norm of $\mathcal{B}$ is equal to the largest singular value of $\mathcal{B}$.