Gibbs Posterior Distributions: New Theory and Applications

by

Nicholas Aaron Syring
B.S., Illinois State University, Normal, IL, 2009
M.S., Northern Illinois University, Dekalb, IL, 2013

Thesis submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Mathematics
in the Graduate College of the
University of Illinois at Chicago, 2017

Chicago, Illinois

Defense Committee:
Dr. Min Yang, Chair and Advisor, MSCS
Dr. Ryan Martin, North Carolina State University
Dr. Cheng Ouyang, MSCS
Dr. Jing Wang, MSCS
Dr. Jie Yang, MSCS

ACKNOWLEDGMENTS

I owe Dr. Ryan Martin a debt of gratitude for welcoming me into the PhD program at the University of Illinois at Chicago. From my first semester as a graduate student there he became not just an academic advisor but a trusted mentor as well. During our collaboration Dr. Martin gave graciously of his time, spending many hours with me at the blackboard hashing out ideas and mathematical arguments that would later form several chapters of this dissertation. Although the depth and breadth of my statistical knowledge was far surpassed by his, Dr. Martin worked very patiently and diligently with me, and always treated me as his intellectual equal. While most of his peers will regard him as an accomplished researcher, I will always consider Dr. Martin to be my greatest teacher.

Pursuing a graduate degree is challenging and rewarding, but also stressful at times. I have been lucky to have a very loving and supportive family. I thank my parents, Lori and Randal, and my sister and brother-in-law, Alison and Roy, for their constant support and encouragement; I needed it.

I thank my professors and committee members Drs. Cheng Ouyang, Min Yang, Jing Wang, and Jie Yang both for their instruction during my graduate studies and their efforts to improve this dissertation. I also thank Dr. Samad Hedayat for his service as leader of our statistics family at UIC and for his mentorship.

I am grateful for the help and encouragement of the faculty and staff at MSCS who were particularly involved in advising me during my studies: Maureen Madden, the Associate Direc-

ACKNOWLEDGMENTS (Continued)

tor of Graduate Studies; Dr. Ramin Takloo-Bighash and Dr. Alex Furman who both served as Director of Graduate Studies; and Dr. Brooke Shipley, our MSCS Department Head, who led an orientation semester course for new graduate students during my first semester.

Finally, I wish to acknowledge two sources whose financial support benefited me during my work on this dissertation: the U. S. Army Research Offices, through Award #W911NF-15-1-0154, and Google's Summer of Code program.

TABLE OF CONTENTS

CHAPTER PAGE

TABLE OF CONTENTS (Continued)

LIST OF TABLES

LIST OF FIGURES

FIGURE **PAGE**

LIST OF FIGURES (Continued)

SUMMARY

Bayesian inference is, by far, the most well-known statistical method for updating beliefs about a population feature of interest in light of new data. Current beliefs, characterized by a probability distribution called a prior, are updated by combining with data, which is modeled as a random draw from another probability distribution. The Bayesian framework, therefore, depends heavily on the choices of model distributions for prior and data, and it is the latter that is of particular concern in this dissertation. Often, as will be shown in various examples, it is particularly difficult to make a good choice of data model: a bad choice may lead to misspecification and inconsistency of the posterior distribution, or may introduce nuisance parameters, increasing computational burden and complicating the choice of prior. Some particular statistical problems that may give Bayesians pause are classification and quantile regression. In these two problems a mathematical function called a loss function serves as the natural connection between the data and the population feature. Statistical inference based on loss functions can avoid having to specify a probability model for the data and parameter, which may be incorrect. Bayes' Theorem cannot reconcile a posterior update using anything other than a probability model for data, so alternative methods are needed, besides Bayes, in order to take advantage of loss functions in these types of problems.

Gibbs posteriors, like Bayes posteriors, incorporate prior information and new data via an updating formula. However, the Gibbs posterior does not require modeling the data with a probability model as in Bayes; rather, data and parameter may be linked by a more general function, like the loss functions mentioned above. The Gibbs approach offers many potential benefits including robustness when the data distribution is not known and a natural avoidance of nuisance parameters, but Gibbs posteriors are not common throughout statistics literature. In an effort to raise awareness of Gibbs posteriors, this dissertation both develops new theoretical foundations and presents numerous examples highlighting the usefulness of Gibbs posteriors in statistical applications.

Two new asymptotic results for Gibbs posteriors are contributed. The main conclusion of the first result is that Gibbs posteriors have similar asymptotic behavior to a class of statistical estimators called M-estimators in a wide range of problems. The main advantage of the Gibbs posterior, then, is its ability to incorporate prior information. The second result extends results for Bayesian posteriors to Gibbs posteriors in a statistics problems where the population feature of interest is a set with a smooth boundary.

Additionally, two main applications are considered, one in medical statistics and one in image analysis. The first application concerns the minimum clinically important difference (MCID), a parameter designed to indicate whether the effect of a medical treatment is practically significant. Modeling for the purpose of inference on the MCID is non-trivial, and concerns about bias from a misspecified parametric model or inefficiency from a nonparametric model motivate using the alternative Gibbs approach, which balances robustness and efficiency. The second application concerns the detection of an image boundary when the image pixels are observed with noise. Likelihood-based methods for the image boundary require modeling the pixel intensities inside and outside the image boundary, even though these are typically of no practical interest. However, a Gibbs posterior can be defined directly on the image boundary parameter, thereby avoiding this issue.

Finally, the Gibbs posterior comes with a scale parameter, also referred to as a learning rate, which mainly affects its finite sample performance. Current research directions do not agree on how to select the learning rate. This dissertation presents a new method, called Gibbs posterior calibration (GPC), to select the learning rate so that Gibbs posterior credible regions are approximately calibrated to their nominal frequency coverage probabilities. Simulation results demonstrate that the proposed algorithm yields highly efficient credible regions in a variety of applications when compared to existing methods.

CHAPTER 1

STATISTICAL PRELIMINARIES

This chapter reviews several fundamental topics in statistics. Besides setting the notations to be used throughout this dissertation, this chapter will begin to relate Gibbs posteriors to more traditional statistical methods.

1.1 Setup of a statistical inference problem

A statistical inference problem begins with a population of individuals under study and research questions about one or more features of the population. A data analyst possesses a sample of data from the population denoted by $x^n = (x_1, x_2, ..., x_n)$ where each data point $x_i \in \mathbb{X}$ for $i = 1, ..., n$. The goal is to use the sampled data to learn about the population and ultimately give informed answers to the research questions. As an example, imagine a demographer studying the US population. The demographer wants to know the median age of new parents in 2016. The population under study is all US persons who had their first child in 2016 and the population feature of interest is the median age.

Data will always be assumed to be sampled in some random manner from the population, so the data can rightly be viewed as a realization from a probability distribution; i.e. the data, prior to observation, is expressed as a random variable $\mathcal{X}^n \sim P^n$ where P^n is the joint distribution of $\mathcal{X}^n = (\mathcal{X}_1, \mathcal{X}_2, ..., \mathcal{X}_n)$. In the case of a simple random sample, each $\mathcal{X}_i \overset{\text{i.i.d.}}{\sim} P$

for $i = 1, 2, ..., n$ where i.i.d. stands for independent and identically distributed, and P^n is the n-fold convolution of P.

Broadly speaking, there are parametric and distribution-free (sometimes called nonparametric) approaches to statistics. In parametric statistics, it is assumed that the data is sampled from among a family of probability distributions $\{P_\theta : \theta \in \Theta\}$ indexed by the parameter θ. Typically, this probability model is assumed to be correctly specified, meaning that there is a unique point $\theta^\star \in \Theta$ corresponding to the unknown true value of the population feature and $\mathcal{X} \sim P_{\theta^\star}$, so that the model holds for the unique, true parameter value. Continuing the above example, suppose the demographer has a simple random sample of n individuals who became new parents in 2016. One possible parametric model says $\mathcal{X}_i \overset{\text{i.i.d.}}{\sim} \Phi(\theta, 4)$ where $\Phi(\theta, \sigma)$ denotes the normal distribution with mean θ and standard deviation σ. In this case, θ is also the median, so questions about the median age of new parents are equivalent to questions about the mean of this normal model distribution. Distribution-free methods (see, for example, (104)) do not assume a particular form of the sampling distribution P. Sections 1.2 and 1.3 discuss methods from both parametric and distribution-free statistics for the given example in more detail.

Statistical inference, commonly, refers to estimation and hypothesis testing for the population feature. Estimation encompasses data-dependent guesses, which may be a points or sets, for the value of the population feature. Hypothesis tests evaluate the truthfulness of assertions about the population feature, such as "the population median age lies between 26 and 31". Introductions to statistical inference can be found in (10), (49), and many other texts. In this dissertation, inference will most often mean point and set estimates. A point estimate will be denoted $\widehat{\theta}_n \in \Theta$ when the population feature is represented by a parameter $\theta \in \Theta$. A set estimate $A(\mathcal{X}^n) \subset \Theta$ will have a nominal coverage probability $\alpha \in (0,1)$. A set estimate will be called calibrated if $P^n(\theta^\star \in A(\mathcal{X}^n)) \geq \alpha$, where P^n denotes the sampling distribution of data $\mathcal{X}^n$ and $\theta^\star$, again, denotes the true value of the population feature.

1.2 Maximum likelihood estimation

Likelihood inference is an approach in parametric statistics. For a model $\{P_\theta : \theta \in \Theta\}$, the likelihood function is defined

$$L(\theta|\mathcal{X}^n = (x_1, ..., x_n)) := f(x_1, ..., x_n|\theta), \tag{1.2.1}$$

which is just the joint density of P_θ, evaluated at the observed data, and viewed as a function of the parameter θ varying over Θ. Often, the loglikelihood, $l(\theta|\mathcal{X}^n = (x_1, ..., x_n)) := \log L(\theta|\mathcal{X}^n = (x_1, ..., x_n))$ is of interest because it has the same maximum as the likelihood function and is easier to compute for large data sets. When the data is a random sample from $P_{\theta^\star}$, the loglikelihood may be written $l(\theta|\mathcal{X}^n = (x_1, ..., x_n)) = \sum_{i=1}^n l(\theta|x_i) = \sum_{i=1}^n f(x_i|\theta)$. A well-known result states that if the model is identifiable, meaning $P_\theta \neq P_{\theta^\star}$ for all $\theta \neq \theta^\star$, then the expectation of the loglikelihood function under the true distribution, $P_{\theta^\star}$, is maximized uniquely at $\theta^\star$, i.e.

$$\theta^\star = \arg\max_{\theta\in\Theta} E_{P_{\theta^\star}}\left(\sum_{i=1}^n l(\theta|x_i)\right);$$

see, for example, Chapter 5 in (99). Denote the above expectation by $R(\theta) := E_{P_{\theta^\star}}\left(\sum_{i=1}^{n} l(\theta|x_i)\right)$ and write its empirical analog as $R_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} l(\theta|x_i)$. The sequence of estimators maximizing functions $R_n(\theta)$ are called maximum likelihood estimators, and can be written mathematically as

$$\hat{\theta}_n = \arg\max_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^{n} l(\theta|x_i).$$

The following well-known result says that, under some conditions, the sequence of maximum likelihood estimators is consistent for $\theta^\star$. Consistency simply means that the sequence of estimators converges in probability to the true parameter value, but in a statistical sense consistency ensures that the data analyst learns more about the parameter as more data is collected (eventually knowing the true parameter exactly, in the limit, as the number of sampled data points grows to infinity). The following conditions are sufficient for consistency of maximum likelihood estimators. Equation 1.2.2 stipulates uniform convergence of the functions $R_n(\theta)$ to $R(\theta)$. Equation 1.2.3 says that the point of maximum of $R(\theta)$ is "well-separated"; that is, the maximum is unique and there is no sequence of points with function values converging to the maximum.

Proposition 1.2.1 *Suppose that for some distance function* $d(\cdot,\cdot) : \Theta \times \Theta \mapsto \mathbb{R}^+$ *and any* $\epsilon > 0$

$$\sup_{\theta\in\Theta} |R_n(\theta) - R(\theta)| \overset{i.p.}{\to} 0, \tag{1.2.2}$$

$$\sup_{\theta: d(\theta,\theta^\star)\geq\epsilon} R(\theta) < R(\theta^\star). \tag{1.2.3}$$

Then, any sequence of estimators satisfying $\mathsf{R}_n(\hat{\theta}_n) \geq \mathsf{R}_n(\theta^\star) - o_{\mathsf{P}_{\theta^\star}(1)}$ *converges in* $\mathsf{P}_{\theta^\star}$*—probability to* $\theta^\star$*.*

For a proof of Proposition 1.2.1 see Theorem 5.7 in (99).

Consider using likelihood inferences to obtain a point estimate of the median age of new parents as in the example discussed in Section 1.1. The analyst chooses the normal model $\{\Phi(\theta, 4) : \theta \in \mathbb{R}\}$, the set of normal distributions with mean θ and standard deviation 4. The mean and median are equal for the normal distribution, and the maximum likelihood estimator is easily shown to be the sample mean $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n \mathcal{X}_i$, which, under $\mathsf{P}_{\theta^\star} := \Phi(\theta^\star, 4)$ has distribution $\Phi(\theta^\star, \frac{4}{\sqrt{n}})$. Since the asymptotic normal distribution is centered at $\theta^\star$ and its variance vanishes, it is clear that the sequence of maximum likelihood estimators $\bar{X}_n$ converges in probability to $\theta^\star$. In this case, consistency is easy to establish, but checking the conditions of Proposition 1.2.2 is usually far from trivial and may require using concepts from empirical processes; see Chapter 19 in (99).

Stronger asymptotic results are available for maximum likelihood estimators. Specifically, it can be shown that the sequence of maximum likelihood estimators converge in distribution to a normal distribution centered, at least approximately, at $\theta^\star$ and with standard deviation on the order $n^{-1/2}$; see, for instance, Theorem 5.23 in (99). From such results, it is straightforward to produce asymptotically calibrated set estimators of $\theta^\star$.

It may happen that the probability model is misspecified, i.e. $\mathcal{X} \sim \mathsf{P}$ for some distribution $\mathsf{P} \notin \{\mathsf{P}_\theta : \theta \in \Theta\}$. Several authors have studied the properties of the so–called quasilikelihood derived by using the model $\mathcal{X} \sim \mathsf{P}_\theta$ when the model does not hold. The main idea is that, under some technical conditions, the quasi-maximum likelihood estimates converge asymptotically to some θ minimizing the Kullback-Leibler divergence between P_θ and P, defined by

$$D_{KL}(P, Q) = \int_{\mathbb{X}} \log(\frac{dP_\theta}{dP}) dP_\theta$$

where $\frac{dP_\theta}{dP}$ is the Radon-Nikodym derivative of P_θ with respect to P. A good technical study of these ideas is presented in (106). Some connections to misspecified models will be made in Section 5.3.1 in regards to Gibbs posteriors.

1.3 M-estimation

In maximum likelihood estimation the probability model for the data determines the function to maximize for estimation purposes. Sometimes, a given statistics problem may suggest an estimating function independently of the probability model. When the minimizer (or maximizer) of a data-dependent function other than a likelihood is used to estimate a parameter, the estimator is called an M-estimator. Consider again the problem described in Section 1.1 which considers estimating the population median age of new parents from a random sample $\mathcal{X}^n = (\mathcal{X}_1, \mathcal{X}_2, ..., \mathcal{X}_n)$ (see also Example 5.11 in (99)). The function $\frac{1}{n} |\sum_{i=1}^n sign(\mathcal{X}_i - \theta)|$, where $sign(x)$ is 1 for $x \geq 0$ and -1 otherwise, is minimized at the sample median $\hat{\theta}_n$. The function $R(\theta)$, as in Proposition 1.2.1, is $E_{P_{\theta^\star}}(|sign(\mathcal{X} - \theta)|) = |P_{\theta^\star}(\mathcal{X} > \theta) - P_{\theta^\star}(\mathcal{X} < \theta)|$, which is clearly minimized at the population median, $\theta^\star$. The uniform convergence in $P_{\theta^\star}$—probability of $|R_n(\theta) - R(\theta)|$ over Θ in Proposition 1.2.1 can be shown using the Glivenko-Cantelli Theorem; see Theorem 19.1 in (99).

Using Proposition 1.2.1, it can be shown that the sample median is consistent for the population median. Further results, like asymptotic normality, that apply to maximum likelihood estimators can likewise be applied to M-estimators; see, for instance, Theorem 5.23 in (99). Moreover, the consistency of the M-estimator is not dependent upon choosing a correctly-specified probability model for the data.

M-estimation is an important concept in the context of Gibbs posteriors. As discussed below in Sections 1.4 and 2.2, Bayesian posteriors depend on likelihood functions while Gibbs posteriors depend analogously on functions like $R_n(\theta)$ used in M-estimation. Therefore, like maximum likelihood estimators, Bayesian posteriors are sensitive to the specified probability model for the data, while Gibbs posteriors, like M-estimators, are not.

1.4 Review of Bayesian statistics

This section provides a brief review of Bayesian statistics, its formulation, interpretation, and advantages and potential shortcomings. An excellent overview of Bayesian statistics is given in (26).

1.4.1 Formulation of Bayesian statistics

Given a statistical inference problem, likelihood inference or M-estimation methods typically produce point and set estimates for $\theta^\star$. Bayesian statistics, on the other hand, produces a probability distribution characterizing $\theta^\star$. A sort of axiom of Bayesian statistics states that all unknown quantities are represented as random variables with a probability distribution. So, before data is even collected, a Bayesian summarizes all available information about the population feature with a probability distribution called the prior, denoted $\theta \sim \Pi$, with π

denoting the density of Π provided it exists. Besides the prior, a Bayesian needs to specify a probability model of the data $\{P_\theta : \theta \in \Theta\}$, as in likelihood inference. Given these two ingredients, Bayes Theorem states how to combine the information in the prior with that in the data in order to produce a posterior distribution for $\theta^\star$ (see Equation 2.1 in (26)),

$$f(\theta|x_1, ..., x_n) = \frac{L(\theta|\mathcal{X}^n = (x_1, ..., x_n))\pi(\theta)}{\int_\Theta L(\theta'|\mathcal{X}^n = (x_1, ..., x_n))\pi(\theta')d\theta'}. \tag{1.4.1}$$

A less restrictive definition of the Bayesian posterior probability of a Π—measurable set A when the prior may not have a density can be written,

$$\Pi_n(A|x_1, x_2, ..., x_n) = \int_A \frac{L(\theta'|\mathcal{X}^n = (x_1, ..., x_n))\Pi(d\theta')}{\int_\Theta L(\theta'|\mathcal{X}^n = (x_1, ..., x_n))\Pi(d\theta')}; \tag{1.4.2}$$

see, for example, Section 1.3 in (27).

Consider a Bayesian approach to inference on the median age of new parents, as in Sections 1.1-1.3 above. The data analyst again uses the normal model $\{\Phi(\theta, \sigma = 4) : \theta \in \mathbb{R}\}$ as in Section 1.2, and chooses a normal prior distribution $\Pi := \Phi(\mu_0 = 27, \sigma_0 = 2)$. From Equation 1.4.1 it is easy to see the posterior distribution is normal with mean $(\sigma^2\mu_0/n + \sigma_0^2\bar{x}_n)/(\sigma^2/n + \sigma_0^2)$ and variance $\sigma^2\sigma_0^2/(\sigma^2 + n\sigma_0^2)$, where $\bar{x}_n$ denotes the mean of the observed values $x_1, x_2, ..., x_n$.

1.4.2 Interpretation of the prior distribution

A point of frequent confusion (and sometimes contention) concerns the prior distribution; particularly, "What is the interpretation of the prior/posterior distribution?" and "How is the prior chosen?".

The prior encompasses all available information about the population feature of interest before observing any new data. The process of choosing the prior is referred to as prior elicitation, and the manner in which this is carried out affects both the interpretation of the prior and the posterior. A so-called subjective prior is one chosen by the data analyst perhaps with input from subject-matter experts, and examples of eliciting subjective priors are given in Section 5.4 of (26). The ability of Bayesian statistics to incorporate information about the population prior to sampling data is a potential advantage. Estimation by maximum likelihood or M-estimation offers no automatic way of incorporating such information. However, the notion of subjectivity does impact the interpretation of probability statements obtained from the Bayesian posterior. In short, these probability statements should be interpreted as "personal", i.e. as belonging to the data analyst who made the choice of prior, since another data analyst handling the same problem may make a different choice of prior, equally reasonable, and thus arrive at a different posterior distribution. This subjective, or personal view of probability stands in contrast to the frequentist or objectivist view of probability, in which probabilities reflect long-run averages like the number of tails in a sequence of coin tosses. A personal view of probability is discussed in (81) and (18).

On the other hand, the parameter may be too complicated to elicit a prior using real prior information or, worse, in some cases no prior information exists. How, then, to determine the prior distribution? Many authors have studied the concept of default priors (sometimes referred to as objective, non-informative, or reference priors as in (5)), distributions meant to reflect an absence of prior information; see also Chapter 5 in (26) and (29). Despite the potential of default priors to alleviate the difficulties associated with prior elicitation, there are several drawbacks to these types of priors. One serious problem is that most, if not all, default priors are improper, that is, they are not probability distributions because they are not integrable. In that case, it is not clear how to interpret posterior probabilities since the probability calculus rests upon the prior distribution; recall Equation 1.4.1 above. A critical view of default priors is taken in (23) where it is shown in some examples that the resulting posteriors have poor asymptotic properties, which are explained briefly in the next section. Another objection to default priors is that they cannot truly represent absence of prior information. There is no totally non-informative prior; see (70) Section 2.4.

The prior distribution is a shared feature of Bayesian and Gibbs models. The issues mentioned here concerning prior elicitation and interpretation apply equally in the context of Gibbs posteriors.

1.4.3 Evaluation of a Bayesian posterior

This section addresses the question “Is a particular Bayesian posterior ‘good’ for a particular statistical inference problem?” The answer to this question depends heavily upon the level of subjectivity of the analysis. Given the prior and likelihood, deriving the Bayesian posterior is simply a matter of applying the rules of probability, i.e. Bayes Theorem; see Equation 1.4.2 above. So, if the data analyst believes strongly in the chosen prior and probability model, no further evaluation is needed to judge the appropriateness of the Bayesian posterior. When there is a collection of posteriors under consideration (stemming from different models, different priors, or both), data analysts may compare posteriors using Bayes factors, described in (46) and (75), for example.

Besides comparing different Bayesian posteriors for a given data set, it can be helpful to study the asymptotic properties of a Bayesian posterior with respect to a hypothetical sample, as the sample size increases to infinity. A good introduction to Bayesian asymptotics is given in (26), with a more advanced treatment in (27) and (99) Chapter 10. There are three main types of results studied: consistency, convergence rates, and convergence in distribution. A sequence of Bayesian posteriors Π_n is said to be consistent for the true parameter $\theta^\star$ if for every $\Pi-$measureable set A containing $\theta^\star$, $\Pi_n(A) \overset{\text{i.p.}}{\to} 1$ with respect to the distribution $P_{\theta^\star}$. This definition of consistency is similar to consistency of maximum likelihood estimators (and M-estimators) discussed above, and says that a sequence of Bayesian posteriors learns the true value of the population feature as data accumulates. But, consistency does not say anything about how much data is needed for a given level of accuracy. To answer that question, a notion of convergence (or concentration) rates is needed. Suppose $d(\cdot, \cdot)$ is a distance measure on $\Theta \times \Theta$. Let A_n be a sequence of $\Pi-$measureable sets such that $A_n = \{\theta \in \Theta : d(\theta, \theta^\star) \leq \delta_n\}$ for a real-number sequence $\delta_n \to 0$. If the sequence of posterior probabilities satisfies $\Pi_n(A_n) \overset{\text{i.p.}}{\to} 1$ with respect to the distribution $P_{\theta^\star}$ for any sequence A_n as above, then the sequence of posteriors is said to converge to $\theta^\star$ (or concentrate on neighborhoods of $\theta^\star$) at rate δ_n with respect to the distance d. In order to know the precise asymptotic variance of the posterior distribution and construct tests and credible intervals with accurate coverage probabilities it is usually necessary to determine the asymptotic distribution of the posterior. Posterior convergence to a normal distribution can sometimes be shown, and this type of result is called a Bernstein-von Mises Theorem; see, for instance, Section 10.2 in (99).

Recall the Bayesian posterior for the median in Section 1.1. Inspection of the posterior mean and variance formulas reveals the posterior is consistent, converges at rate $n^{-1/2}$, and that the posterior credible intervals $(\Pi_{n,\alpha/2}, \Pi_{n,1-\alpha/2})$, where $\Pi_{n,t}$ is such that $\int_{-\infty}^{\Pi_{n,t}} d(\Pi_n(s))ds = t$, are approximately $100\alpha\%$—calibrated.

The same types of asymptotic analyses can be studied in the context of Gibbs posteriors, although some of the proof techniques differ. Chapter 4 of this dissertation describes some existing asymptotic results for Gibbs posteriors and presents two new convergence rate results.

1.5 Illustrative comparison of estimation methods

This section ties together many topics in Chapter 1 with a brief simulation example of inference on the population median age of new parents discussed throughout the chapter. Two simulations were conducted by taking 1000 i.i.d. samples of size 1000 from two different populations. The first population is characterized by a normal distribution with mean 27 and standard deviation 4 while the second population is characterized by a Gamma distribution with shape parameter 36 and scale parameter 0.75. Both distributions have mean 27, but the normal distribution has median 27 while the Gamma distribution has median approximately

26.75. For each simulated sample, the values of the maximum likelihood estimate, M-estimate, and Bayesian posterior mean were recorded using the methods described in Sections 1.2-1.4. The results of the simulation experiment are illustrated in Figure 1.5.1. In both sets of simulations, the maximum likelihood estimates and the Bayesian posteriors are based on the same normal model with mean 27 and standard deviation 4. However, that model is only correctly specified in the first set of simulations; the model is misspecified in the second set due to the fact that the data were generated by the Gamma distribution. The maximum likelihood estimates agree with the M-estimates in the set of simulations with the correctly-specified normal model, but the maximum likelihood estimates appear to be off target in the second set of simulations where they are based on the wrong model. The M-estimates, however, are robust to the underlying data distribution; they seem to have concentrated near the correct answer in both sets of simulations. Since the Bayesian posterior is built from the likelihood, it inherits the bad behavior of the maximum likelihood estimates when the model is wrong. It would be desirable to combine the robustness of the M-estimation methods with the ability of Bayesian methods to incorporate prior information. Gibbs posteriors, introduced in Chapter 2, accomplish just that.

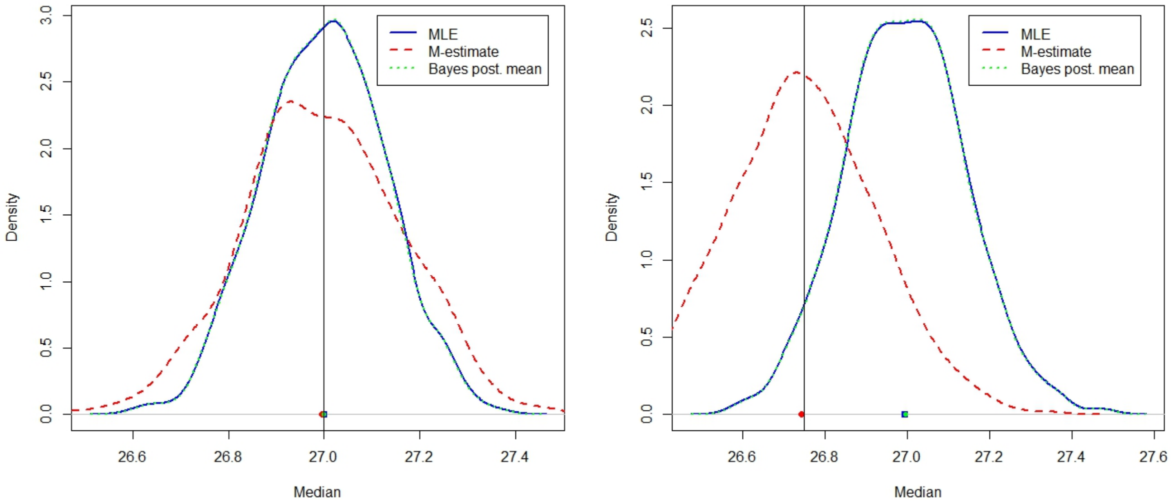


Figure 1.5.1. Density estimates of the 1000 median estimates from each method: maximum likelihood, M-estimation, and the Bayesian posterior mean. Left: simulation results from normal distribution. Right: simulation results from Gamma distribution.

CHAPTER 2

INTRODUCTION TO GIBBS POSTERIORS

2.1 Motivating the Gibbs posterior distribution

This section is intended to persuade the reader that an alternative to standard, probability model–centric Bayesian inference is needed and that Gibbs posteriors can provide that alternative. Many of the following ideas are present in (6) and (93), and these issues roughly categorize into modeling, prior specification, and computation.

To provide a context for discussing these issues, consider a linear quantile regression model where a conditional quantile is modeled as a function of covariates. In particular, for data $\mathcal{X} = (Y, X)$ and fixed $\tau \in (0, 1)$, interest is in the τ^{th} quantile of the response $Y \in \mathbb{R}$, given the covariates $X \in \mathbb{R}^{p+1}$, expressed as

$$Q_\tau(Y \mid X) = X^\top \theta, \tag{2.1.1}$$

where dimension $p + 1$ represents an intercept and p covariates. In this formula, the vector θ depends on τ but, for notational simplicity, this dependence is omitted.

Model specification is difficult for quantile regression. The model setup in Equation 2.1.1 does not imply much of anything about the underlying probability model, and families of probability distributions often used in modeling are not parametrized in terms of quantiles. Moreover, as illustrated by the example of inference on a population median in Chapter 1, using a misspecified model may lead to inconsistency of the resulting Bayesian posterior. The most common approach to quantile regression is to avoid specifying the conditional distribution of Y given X and instead use M-estimation; see, for instance, (53). The loss function used in M-estimation is

$$R_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} |(Y_i - X_i^\top\theta)(\tau - I_{Y_i - X_i^\top\theta<0})| \tag{2.1.2}$$

for observations $\mathcal{X}_i = (Y_i, X_i)$, $i = 1, \ldots, n$, and estimators minimizing Equation 2.1.2 are shown to be consistent in (53) and Section 4.2.3. M-estimation provides a method for obtaining a consistent estimator for θ, but does not account for prior information. If the data analyst hopes to use prior information, they likely are only familiar with Bayesian methods for accomplishing this, which require a probability model for the conditional distribution of Y given X. A sort of default model that has been considered by several authors (e.g. (111; 90; 89)) is the (misspecified) asymmetric Laplace likelihood. The form of this loglikelihood is equivalent to Equation 2.1.2, and the consistency of M-estimators minimizing Equation 2.1.2 is the basis for this choice of likelihood. As shown in Section 4.2.3, a Gibbs posterior can be constructed using Equation 2.1.2.

Rather than using a "default" misspecified model, suppose the data analyst strongly believes in a normal model. For example, $Y_i \overset{\text{ind.}}{\sim} \Phi(X^\top\theta - z_\tau\sigma, \sigma)$ where z_τ is the τ^{th} quantile of the standard normal distribution so that each Y_i is independent with the same standard deviation but a different mean. This model is over-parametrized; the interest parameter θ enters into the mean of the normal model, but a nuisance parameter σ enters into both the mean and standard deviation. Since σ is a nuisance parameter not appearing in the problem setup in Equation 2.1.1, it is unlikely the data analyst has any prior information about σ. This is a common problem for Bayesian posteriors; it is necessary to introduce nuisance parameters in order to write down a likelihood, but it may be difficult to translate prior information about the parameter of interest into a higher dimension. More examples of this issue can be found in Section 5.3 and (94) for the MCID problem and in (61) where a Bayesian posterior in an imaging problem is over-parametrized.

Besides having to assign a prior distribution to each nuisance parameter, it is also necessary to carry out posterior computations on a space of dimension often larger than the dimension of the parameter of interest. This means Bayesian posteriors usually involve at least as much computation as a Gibbs posterior for a given problem, and often more, since the Gibbs posterior is defined on the parameter space of the interest parameter.

The growing body of work on Gibbs posteriors (see Chapter 3) suggests on its own that an alternative to Bayesian posteriors has value. It is hoped that this dissertation will provide a broad overview of the state of research on Gibbs models, contribute significantly to their theoretical basis, and provide several examples highlighting their use.

2.2 Defining the Gibbs posterior

This dissertation considers statistical inference problems with the following ingredients: data, which, prior to collection, is represented by a random variable $\mathcal{X} \in \mathbb{X}$; a data-generating distribution denoted $\mathcal{X} \sim \mathsf{P}$; a parameter of interest, about which inferences are to be made, is denoted by θ with parameter space Θ; a prior distribution Π on Θ, usually with a density π; and a function linking data to parameter, denoted by $\ell(\mathcal{X}, \theta)$. In particular, the linking function $\ell(\mathcal{X}, \theta)$ is typically assumed to be consistent, i.e. the true parameter value $\theta^\star$ satisfies $\theta^\star =$

$\arg\min_{\theta\in\Theta} \mathsf{E}_P[\ell(\mathcal{X},\theta)]$. The expectation $\mathsf{E}_P[\ell(\mathcal{X},\theta)]$ is called the risk function and is denoted $R(\theta)$; its empirical analogue is $R_n(\theta) = \frac{1}{n}\sum_{i=1}^n \ell(\mathcal{X}_i,\theta)$ for an i.i.d. sample $\mathcal{X}^n = (\mathcal{X}_1, ..., \mathcal{X}_n)$.

Given the above ingredients, the posterior probability of a $\Pi-$measureable set A is denoted $\Pi_n(A)$ and is calculated

$$\Pi_n(A) = \frac{\int_A \exp(-\omega R_n(\theta))\Pi(d\theta)}{\int_\Theta \exp(-\omega R_n(\theta))\Pi(d\theta)} \tag{2.2.1}$$

for a positive scale factor ω. This general definition includes many important special cases:

1. Bayesian posterior – when $-R_n(\theta)$ is a true loglikelihood and $\omega = 1$ Equation 2.2.1 is simply the usual Bayesian posterior.

2. Bayes with learning rate – when $-R_n(\theta)$ is a working loglikelihood and $\omega < 1$ Equation 2.2.1 is a weighted Bayesian posterior, placing more faith in the prior and less in the model than a default Bayesian posterior. The scale parameter ω is called the learning rate, and $\omega < 1$ is typically used when the posited probability model is in doubt; see (6) and (38).

3. Gibbs posterior – when $R_n(\theta)$ is the empirical risk function for some loss function $\ell(\mathcal{X},\theta)$ like those used in M-estimation, the general posterior in (Equation 2.2.1) is a Gibbs posterior.

Remark 1. To make it more clear how $\ell(\mathcal{X},\theta)$ may represent a loss function, consider a linear regression problem with response and predictor data $\mathcal{X} = (Y, X) \in \mathbb{R}\times\mathbb{R}^p$, and vector parameter $\theta \in \mathbb{R}^p$. Then, a reasonable loss function linking data and parameter is squared-error loss, $\ell(\mathcal{X},\theta) = (Y - X^\top\theta)^2$, since the minimizer of the corresponding empirical risk $R_n(\theta)$ for an i.i.d. sample $\mathcal{X}^n$ is the least squares estimator. Note that if one assumes Gaussian errors for the regression model, the Bayes posterior with a learning rate is equivalent to the Gibbs posterior based on the squared-error loss.

Remark 2. There are several available methods for selecting the learning rate ω, and a full discussion of existing methods is provided in Section 3.3 with new contributions in Chapter 7. In (38) selection of the learning rate is discussed in the context of Bayes posteriors, (34) introduce a method for selecting the learning rate in order to improve prediction, and (93) (see also Chapter 7) use a computational method that selects the learning rate in order to calibrate posterior credible sets. In particular, for Gibbs posteriors, the learning rate ω controls the variance of the posterior distribution and the volume of its credible sets.

2.3 Deriving the Gibbs posterior

Bayes Theorem provides justification for calculating Bayesian posterior probabilities using Equation 2.2.1 when $-R_n(\theta)$ is a true loglikelihood, but why does this formula make sense to define Gibbs posteriors when the linking function is a loss function? Several authors have given derivations of the above definition for the Gibbs posterior. Below are two constructions of Gibbs posteriors supporting this definition; the first is based on the idea of coherence, and appears in (6).

The goal is to produce a posterior distribution, generally, a probability measure ν on the space Θ, given data $\mathcal{X}^n$, empirical risk function $R_n(\theta)$, and prior distribution Π. A reasonable strategy is to construct a loss function $L(\nu; \Pi, \mathcal{X}^n)$ on the space of probability distributions on Θ and then minimize this loss in order to choose which posterior to use for inference. Arguing from independence of data and prior, (6) identify the basic form $L(\nu;\Pi,\mathcal{X}^n) = h_1(\nu,\mathcal{X}^n)+h_2(\nu,\Pi)$ so that the loss function decomposes into loss to data and loss to prior. Further, (6) show that the loss to data must be the expected loss $h_1(\nu,\mathcal{X}^n) = \int R_n(\theta)\nu(d\theta)$ while the loss to prior must be the Kullback-Leibler divergence $h_2(\nu,\Pi) = D_{KL}(\nu,\pi) = \int \log\{\nu(d\theta)/\Pi(d\theta)\}\nu(d\theta)$. The choice of Kullback-Leibler divergence requires some additional explanation, and (6) use a *coherence* argument to justify this choice. Roughly, the idea is the following. First, divide the n data points into the first $1 \leq m < n$ and the remaining $n-m$. Using the first m data points, $\mathcal{X}^m$, choose the posterior $\hat{\nu}_m$ minimizing

$$L(\nu;\Pi,\mathcal{X}^m) = \sum_{i=1}^{m} h_1(\nu,\mathcal{X}_i) + h_2(\nu,\Pi).$$

For the remaining data, $\mathcal{X}^{n-m}$, $\hat{\nu}_m$ serves as the updated prior distribution so that

$$L(\nu;\Pi,\mathcal{X}^n) = \sum_{i=m+1}^{n} h_1(\nu,\mathcal{X}_i) + h_2(\nu,\hat{\nu}_m).$$

The coherence property states that the loss function $L(\nu;\Pi,\mathcal{X}^n)$ should remain the same no matter the value of m. In other words, one should not arrive at different posterior distributions by updating one's beliefs using all the data at once compared to updating first using part of the data and next using the remainder. For this property to hold, and the updating of beliefs to be internally consistent, (6) show the only choice is to take the loss to prior to be the Kullback-Leibler divergence. It follows that the minimizer of $L(\nu;\pi,x)$ is given by Equation 2.2.1.

The above construction of the Gibbs posterior in (6) is motivated by the need for a coherent method of updating beliefs in the absence of a probability model, P, for the data. Several other authors have motivated the Gibbs posterior from the point of view of classification or prediction. Probably approximately correct Bayes (often called PAC-Bayes) methods seek classifiers or predictors minimizing a bound on the so-called the posterior averaged risk. Bounds of this type are studied in (84), (72), (12), (113), and (114) and elsewhere with an extensive overview given in (11). Although the motivation is substantially different, the posterior distribution minimizing the PAC-Bayes bound is generally the Gibbs posterior; see Equations 4 and 5 in (114) and Corollary 5.1 in (12). Below is a brief summary of the derivation of the Gibbs posterior based on minimizing the posterior averaged risk.

Again, the goal is to produce a posterior distribution on Θ, call it Π_n. Suppose the data analyst wants the best posterior distribution in terms of predicting the future observation, referred to as $\tilde{\mathcal{X}}$, based on observations $\mathcal{X}^n$ and linking function $\ell(\mathcal{X}, \theta)$. One way to define the "best" posterior, in terms of prediction, is to consider the posterior minimizing the posterior averaged risk (or generalization error) $\mathsf{E}_{\Pi_n}(\mathsf{E}_{\tilde{\mathcal{X}}}\ell(\tilde{\mathcal{X}}, \theta))$. The posterior averaged risk is complicated and deserves an explanation. When the linking function is a loss function, the inner expectation becomes the average loss for a new observation as a function of θ. The outer expectation averages over θ according to the posterior distribution. Sometimes, a third expectation is taken, this time with respect to the original data, $\mathcal{X}^n$, that the posterior is based upon. The idea is that a posterior minimizing this quantity will fit new data well in addition to the data it is based upon, on average. It is straightforward to bound the posterior averaged risk using (reverse) Jensens's inequality,

$$\mathsf{E}_{\Pi_n}(\mathsf{E}_{\tilde{\mathcal{X}}}\ell(\tilde{\mathcal{X}},\theta)) \leq \mathsf{E}_{\Pi_n}(\ln \mathsf{E}_{\tilde{\mathcal{X}}} e^{-\ell(\tilde{\mathcal{X}},\theta)}).$$

In (114) Theorem 2.1, the above quantity is bounded using the Kullback-Leibler divergence,

$$\mathsf{E}_{\Pi_n}(\ln \mathsf{E}_{\tilde{\mathcal{X}}} e^{-\ell(\tilde{\mathcal{X}},\theta)}) \leq \mathsf{E}_{\Pi_n}(\ell(\mathcal{X}^n,\theta)) + \mathsf{D}_{\mathsf{KL}}(\Pi_n,\Pi)$$

where Π denotes the prior. The posterior Π_n minimizing this bound on the posterior averaged risk is the Gibbs posterior given in Equation 2.2.1.

2.4 Contributions of this dissertation

This dissertation provides contributions to three areas of study concerning Gibbs posterior distributions. First, Chapter 4 reviews asymptotic theory for Gibbs posteriors, including two general theorems new in this dissertation, which are applied in later sections. Next, Chapters 5 and 6 present two detailed applications of Gibbs posteriors to problems in medical statistics and image analysis. Finally, Chapter 7 contributes an algorithm for determining the scale factor, ω, in the Gibbs posterior; see Equation 2.2.1. But, before presenting new contributions, Chapter 3 summarizes some previous work on applying Gibbs posteriors to statistics problems.

CHAPTER 3

GIBBS POSTERIORS IN PREVIOUS WORKS

3.1 Gibbs posteriors for econometric models

In (14), the authors investigate the use of Gibbs posteriors for several models used in economics. The Gibbs posteriors they develop are attractive compared to existing methods for three reasons: the asymptotic Gibbs posteriors are normal and admit approximately calibrated set estimates, computation of the Gibbs posteriors using MCMC is often easier than computation of alternative estimators which rely on optimization, and the Gibbs posteriors do not require full probability models of the data. A review of some of the author's asymptotic results is given in Section 4.4. In this section, one example application from (14) is described in detail.

Censored data is common in applications. Right censoring occurs when values cannot be observed above a certain threshold. Maybe the most well-known example of right censoring is in survival data where the survival time of an individual cannot be known past the end of the experiment's duration. Left censoring, when the value of an observation cannot be measured under a certain threshold can also occur. In (14), the authors consider a median regression model with left censoring similar to the following,

$$Y^\star = \beta_0 + X^\top \beta + \epsilon, \tag{3.1.1}$$

$$X \sim N(0, I_3),\ \ \epsilon \sim X_1^2 N(0, 1),\ \ Y = \max(0, Y^\star), \tag{3.1.2}$$

where I_3 denotes the 3×3 identity matrix. The response variable $Y^\star$ is left censored so that only Y is observed. A Gibbs posterior may be constructed for this model using the empirical risk function $R_n(\theta = (\beta_0, \beta)) = \sum_{i=1}^{n} |Y_i - \max(0, \beta_0 + X^\top \beta)|$. In (14), this Gibbs posterior is compared to an optimization approach using iterated linear programming due to (8). The conclusion is that the Gibbs posterior performs better than the linear programming method in this example in terms of bias and mean squared error of the parameter estimates. Moreover, the linear programming converged to a local minimum away from the true parameter value roughly $5 - 10\%$ of the time.

To complement this analysis, another simulation of the model in Equation 3.1.1 was conducted to compare the Gibbs posterior with a Bayesian posterior. The Bayesian posterior is based on a normal likelihood,

$$L(\theta = (\beta_0, \beta, \sigma) \mid (Y, X)) = \prod_{i=1}^{k} \Phi\left(\frac{0 - (\beta_0 + X_i^\top \beta)}{\sigma}\right) \prod_{i=(k+1)}^{n} \phi\left(\frac{Y_i - (\beta_0 + X_i^\top \beta)}{\sigma}\right)$$

where the first k observations are left censored and the last $N - k$ observations are fully observed, $\Phi(\cdot)$ denotes the normal distribution function and $\phi(\cdot)$ denotes the normal density function. This Bayesian likelihood is misspecified for the model in Equation 3.1.1 because it does not correctly model the heteroscedastic variance of the error term $\epsilon \sim X_1^2 N(0, 1)$. So, in addition to the model in Equation 3.1.1, a Gibbs posterior and a correctly specified Bayesian posterior are compared using a modified model with $\epsilon \sim N(0, 1)$, so that observations have homoscedastic variances.

In 3.1 below, results are given for 100 simulations of both the homoscedastic and heteroscedastic variance models, with sample sizes $n = 400$ and $n = 1600$ for $(\beta_0, \beta) = (1, 3, 3, 3)$, and using 1000 posterior samples. This model results in about 40% of responses being censored. The mean squared errors are compared for the Gibbs and Bayesian posteriors and the conclusion is that the Gibbs posterior has lower mean squared error in both models. Not only does the Gibbs posterior give apparently more accurate point estimates, it does so with a simpler model. The Bayesian model includes an extra nuisance parameter, σ.

		Mean Squared Errors	
Sample Size	Posterior	Homoscedastic	Heteroscedastic
400	Gibbs	0.10	0.11
	Bayes	0.20	0.36
1600	Gibbs	0.08	0.09
	Bayes	0.20	0.27

TABLE 3.1.1

MEAN SQUARED ERRORS FOR GIBBS AND BAYESIAN POSTERIORS FOR CENSORED MEDIAN REGRESSION MODELS.

Other interesting applications covered in (14) include instrumental variables regression and time series applications for stock prices.

3.2 Variable selection in binary regression models

In (44), the authors study Gibbs posteriors for variable selection in high-dimensional binary regressions. The data is $\mathcal{X}^n = ((Y_1, X_1), (Y_2, X_2), ..., (Y_n, X_n))$ where (Y_i, X_i) is a response predictor pair. The responses Y_i are binary, i.e. $Y_i \in \{0, 1\}$, while the predictors are high dimensional vectors $X_i \in \mathbb{R}^p$ where p may be much larger than n. The goal is to select a small subset of predictors that are useful for predicting future responses.

To predict an unobserved response Y using predictors X, the authors of (44) consider linear classifiers $I(X^\top \beta > 0)$, where $I(\cdot)$ is the indicator function, with corresponding risk function $R(\beta) = P(Y \neq I(X^\top \beta > 0))$.

Bayesian approaches to the variable selection problem require a probability model for the data. One example model is logistic regression, where the likelihood has the form

$$L(\beta \mid \mathcal{X}^n) = \prod_{i=1}^{n} \left(\frac{\exp(X_i^\top \beta)}{1 + \exp(X_i^\top \beta)} \right)^{Y_i} \left(1 - \frac{\exp(X_i^\top \beta)}{1 + \exp(X_i^\top \beta)} \right)^{1-Y_i},$$

which is a Bernoulli likelihood with success probability expressed using the logistic function. The prior distribution is typically chosen to enforce sparsity of the parameter vector β, by putting high probability on coefficients having value exactly 0. Then, variable selection is accomplished by removing all predictors with zero coefficients.

Recall from Section 2.2 that one motivation for deriving the Gibbs posterior is to define the posterior distribution with the minimal risk, on average, with respect to a linking function; in this case the linking function is $\ell(\mathcal{X}, \beta) := I(X^\top \beta > 0)$. The Bayesian posterior, on the other hand, may have poor predictive performance with respect to this linking function and corresponding risk function. The authors of (44) give the following example. Suppose that the predictor is a scalar and that $P(X = \pm 1) = \lambda$ and $P(X = 0) = 1 - 2\lambda$ for some $\lambda \in (0, 0.25)$.

Also, let $P(Y = 1 \mid x) = 1 - P(Y = 0 \mid x) = I(x \neq 0)$. The distribution P, so described, is not a member of the logistic family defined above, so the logistic model is misspecified in this case. As noted in (44), the asymptotic Bayesian posterior predictive probability $\Pi_n(Y = 1 \mid x) = 2\lambda$ when based on the above logistic model. By construction, the resulting linear classifier $I(\Pi_n(Y = 1 \mid x) > 0.5) = I(x^\top \beta > 0) = 0$ always predicts zero. The misclassification error using the Bayesian posterior is 2λ. However, the risk function $R(\beta)$ has minimum λ, which can be obtained, for example, by a logistic model with $P(Y = 1 \mid x) = \exp(x-0.7)/[1+\exp(x-0.7)]$ with corresponding linear classifier $I(x - 0.7 > 0)$. So, interestingly enough, the optimal posterior with respect to the misclassification error of predicting the next response corresponds to a Bayesian posterior based on the logistic model even though this model is misspecified, but the asymptotic Bayesian posterior does not converge to the risk-optimal posterior.

As argued in (44) this poor risk behavior of the Bayesian posterior, especially in misspecified models, provides motivation to look for an alternative variable selection approach. Since the Gibbs posterior can be constructed from the empirical risk function of interest, it is reasonable to expect it should have good performance with respect to that risk function. The authors suggest two different empirical risk functions for the Gibbs posterior

$$(i)\ R_n(\beta) := \frac{-1}{n} \sum_{i=1}^{n} \log(I(x_i^\top \beta > 0)e^{y_i - 1} + [1 - I(x_i^\top \beta > 0)]e^{-y_i}) \tag{3.2.1}$$

$$(ii)\ R_n(\beta) := \frac{-1}{n} \sum_{i=1}^{n} \log(\Phi_i e^{y_i - 1} + [1 - \Phi_i]e^{-y_i}), \tag{3.2.2}$$

where $\Phi_i = \Phi(\sigma_n^{-1} x_i^\top \beta)$ and $\Phi(\cdot)$ denotes the standard normal distribution function. While choice (i) is the sample version of the risk function $R(\beta)$, (44) recommends using choice (ii) because it is smooth in the parameter β and is close to choice (i) for small σ_n.

In addition to the empirical risk function, a prior distribution on β is needed to construct a Gibbs posterior. Roughly, the prior distribution used in (44) is a hierarchical normal-binary prior. A Bernoulli distribution models which indices of β correspond to non-zero coefficients, while a normal distribution models the size of the non-zero coefficients. Based on the empirical risk in Equation 3.2.2 and the normal-binary prior, the authors show that the Gibbs posterior is consistent for the parameter $\beta^\star := \arg\inf_\beta R(\beta)$ for the risk of a linear classifer, $R(\beta)$, under certain conditions on the sparsity of $\beta^\star$ and the parameter space.

3.3 Selecting the Gibbs posterior scale parameter ω

The purpose of this section is to review recently proposed methods for the determination of the scale parameter ω present in the Gibbs posterior; see Equation 2.2.1. Although many authors, including references mentioned above, acknowledge the scale parameter in their formulations and applications of Gibbs posteriors, few authors have provided systematic, non-problem-specific approaches towards its determination.

3.3.1 SafeBayes method

In (33), the authors propose choosing the scaling parameter ω in order to minimize a certain loss function they deem the posterior-expected posterior-randomized log-loss, also called the Gibbs loss. This Gibbs loss has the form

$$\sum_{i=1}^{n} \int -\log \ell(\mathcal{X}_i, \theta)\Pi_i(d\theta), \tag{3.3.1}$$

for a linking function $\log \ell(\mathcal{X}_i, \theta)$ and where Π_i denotes the Gibbs posterior based on the first i data, and implicitly depends on ω, as in Equation 2.2.1. Choosing the scaling parameter in this way is referred to as *SafeBayes* by (33).

In several simulation experiments, (33) demonstrate their *SafeBayes* method results in lower prediction errors than standard Bayesian models in a linear regression setting with a misspecified model. When the model was correctly specified, the *SafeBayes* approach was not significantly worse than standard Bayes. Here is an example demonstrating the SafeBayes method.

Consider the multiple linear regression model with heteroscedastic errors,

$$Y_i = \beta_0 + X_i^\top \beta + \sigma\epsilon_i \tag{3.3.2}$$

for response and covariate pairs $\mathcal{X}^n = ((Y_1, X_1), (Y_2, X_2), ..., (Y_n, X_n))$, where $(Y_i, X_i) \in \mathbb{R} \times \mathbb{R}^p$, $\sigma > 0$ is an unknown nuisance parameter, and $\epsilon_i \overset{\text{ind.}}{\sim} N(0, \|X_i\|)$. One reason to consider a Gibbs posterior for multiple linear regression is that the standard model assumes homoscedastic errors, i.e. $\epsilon_i \overset{\text{i.i.d.}}{\sim} N(0,1)$, but this assumption is often violated and the resulting set estimators of β may not be calibrated; see Section 7.4.2.

A Gibbs posterior for β may be constructed using the empirical risk function $R_n(\beta_0, \beta) = \frac{1}{n}\sum_{i=1}^n (Y_i - \beta_0 - X_i^\top \beta)^2$ and, for example, a flat prior $\pi(\cdot) \propto 1$. The following algorithm can be used to approximate the Gibbs loss in Equation 3.3.1,

1. Select a grid of values for the learning rate ω, i.e. $0 \leq \omega_1 < \omega_2 < ... < \omega_k \leq 1$.
2. For i in $1, 2, ..., k$ and for j in $1, 2, ..., n$ sample M times from the Gibbs posterior Π_j based on the first j observations and using ω_k. Denote the M posterior samples as $\beta_{i,j}^1, ..., \beta_{i,j}^M$.
3. Approximate the Gibbs loss in Equation 3.3.1 for each ω_j by the Monte Carlo average

$$\begin{aligned} &\sum_{i=1}^n \frac{1}{M}\sum_{t=1}^M \ell((Y_i, X_i), \beta_{i,j}^t) \\ &= \sum_{i=1}^n \frac{1}{M}\sum_{t=1}^M (Y_i - \beta_{0,i,j}^t - X_i^\top \beta_{i,j}^t)^2. \end{aligned}$$

4. Choose learning rate ω_j with minimum approximate Gibbs loss.

See Section 7.4.2 for simulation results using SafeBayes in the above example along with a comparison with other methods.

3.3.2 A method based on coherence

In (38), as in (6), the authors use a coherence argument to decide the value of the learning rate parameter ω, and their idea meshes nicely with their derivation of the Gibbs posterior in (6) and Section 1.4 above. Their method is somewhat limited, as described below, in that it only applies to the case $R_n(\theta)$ is an actual negative loglikelihood, not just an empirical version of a loss function.

They begin by writing the posterior update in the following way

$$\log \Pi_n(\theta) = -\omega R_n(\theta) + \log \Pi(\theta) + \log Z_\omega,$$

where Z_ω is the normalizing constant. In this section, it is assumed both the posterior and prior have densities and that the prior is proper. From this equation, it is apparent ω can be interpreted as the learning rate, the relative amount by which the posterior depends upon the data, through $R_n(\theta)$, versus the prior. The idea is that the model (here consider $-R_n(\theta)$ to be a loglikelihood, so it is the model), may be uncertain, so it is reasonable to put less than full weight, $\omega = 1$, on the model, and instead less, $\omega < 1$, in order to incorporate model uncertainty.

Then, (38) propose to choose ω such that the prior–to–posterior update results in the same gain in information whether it is assumed the model is true or not. First, suppose the model is true, i.e., the loglikelihood is $-R_n(\theta)$ and $\omega = 1$. Denote the prior expected information gain under this scenario as $I_1(\mathcal{X})$. Next, suppose there is some doubt the data is generated from the given model. Then, the loglikelihood is $-\omega R_n(\theta)$ for some value of ω not necessarily 1. Denote the prior expected information gain to be $I_\omega(\mathcal{X})$. (38) set ω to solve

$$\int I_\omega(x) P(dx) = \int I_1(x) \exp(-R_n(\theta^\star)) dx.$$

The information mentioned above takes the form of the Fisher relative information divergence of posterior from prior,

$$F(\Pi(\cdot), \Pi_n(\cdot)) = \int \Pi(\theta) \left(\frac{\nabla \Pi_n(\cdot)}{\Pi_n(\cdot)} - \frac{\nabla \Pi(\cdot)}{\Pi(\cdot)} \right)^2 d\theta,$$

with derivatives taken with respect to θ. Solving the above, (38) find

$$\omega = \sqrt{\frac{\int \exp(-R_n(\theta^\star)) F(\Pi(\cdot), \Pi_n(\cdot)) dx}{\int F(\Pi(\cdot), \Pi_n(\cdot)) P(dx)}}.$$

Letting $\Delta(x) = F(\Pi(\cdot), \Pi_n(\cdot))$, and estimating $\theta^\star$ with the maximum likelihood estimator, $\hat{\theta}$, and replacing P with the empirical distribution, an empirical solution is found to be

$$\hat{\omega} = \sqrt{\frac{\int \exp(-R_n(\hat{\theta})) \Delta(x) dx}{n^{-1} \sum_{i=1}^{n} \Delta(\mathcal{X}_i)}}.$$

A closed form expression for $\hat{\omega}$ is available for models in the exponential family. When the probability density of the data can be written $f(x;\theta) = c(x) \exp(x\theta - b(\theta))$ for a one-dimensional parameter θ, the learning rate is estimated by

$$\hat{\omega} = \frac{b''(\hat{\theta}) + \int (\bar{X} - b'(\theta))^2 p(\theta) d\theta}{S^2 + \int (\bar{X} - b'(\theta))^2 p(\theta) d\theta}$$

where $\bar{X}$ and S^2 denote the sample mean and sample variance of the data and f', f'' denote first and second derivatives, respectively. For example, for the Poisson model,

$$\hat{\omega} = \frac{[\bar{x}^2 + \bar{x}] \int_{\theta>0} \theta^{-2} p(\theta) d\theta - 2\bar{x} \int_{\theta>0} \theta^{-1} p(\theta) d\theta + 1}{[\bar{x}^2 + S^2] \int_{\theta>0} \theta^{-2} p(\theta) d\theta - 2\bar{x} \int_{\theta>0} \theta^{-1} p(\theta) d\theta + 1}.$$

3.3.3 Unit information loss method

In (6) the authors provide several methods for setting the scale parameter ω. The first method, called unit information loss, is similar to the coherence-based method by two of these authors described above in Section 3.3.2. In this case, before data is observed, there is only prior information about θ, and, together with the linking function and scaling parameter, one can form the loss function

$$L(\nu, \theta, \pi) = \omega \ell(\mathcal{X}, \theta) + \log \frac{\pi(\hat{\theta})}{\pi(\theta)},$$

where $\pi(\theta)$ is the density of Π and $\pi(\hat{\theta})$ is the value of the density where $\hat{\theta}$ maximizes $\pi(\theta)$. Then, (6) argue that since this is a sum of two loss functions with only one piece of information (the prior), the expected losses should be balanced between the two. In other words, ω should be chosen such that

$$\omega = \frac{\int \log \frac{\pi(\hat{\theta})}{\pi(\theta)} \pi(d\theta)}{\int \int \ell(\mathcal{X}, \theta)} m(dx, d\theta),$$

where $m(x, \theta)$ is a joint distribution for $\mathcal{X}$ and θ, where $\pi(\theta)$ is the marginal density of θ. The obvious drawback of this approach is that $m(x, \theta)$ must be specified, and has an impact on the determination of ω. Additionally, Π must be a proper prior for the above expectation to be well-defined. However, (6) show an empirical version produces a reasonable solution in a simple normal distribution example. Let $\ell(\mathcal{X}, \theta) = (\mathcal{X} - \theta)^2$ and let $\pi(\theta)$ be a normal density with mean zero and variance $1/\tau$. Then,

$$\omega = \frac{1}{2\sum_{i=1}^{n}(X_i - \bar{X}_{-i})^2}$$

where $\bar{X}_{-i}$ is the sample mean of the observations with X_i removed. Clearly, this choice of ω is asymptotically equal to $\frac{1}{2}\sigma^{-2}$, the correct scaling of the posterior, for instance, so that posterior credible sets are calibrated.

3.4 Variational approximations to Gibbs posteriors

In (2), the authors extend the work of (114) to variational approximations of Gibbs posteriors. Like Bayesian posteriors, Gibbs posteriors may be intractable in a given problem and require a great deal of computational effort to sample. The solution studied by (2) is to approximate the Gibbs posterior with a simpler distribution which is easier to sample. Often times, one can sample from this variational posterior in closed form without the need for Markov chain Monte Carlo (MCMC) or other algorithmic techniques. The challenge is to find a family of distributions large enough to approximate the Gibbs posterior with fidelity while also having computational advantages. The authors implement their work in the R package PACVB; see (78).

To choose a variational posterior it is necessary to define a family of distributions, say $\mathcal{F}$, defined on the parameter space Θ for consideration. Then, choose the posterior $\Pi^{\star}$ which satisfies

$$\Pi^{\star} = \arg\min_{\Pi \in \mathcal{F}} D_{KL}(\Pi, \Pi_n).$$

That is, choose the candidate distribution from $\mathcal{F}$ that comes closest to the Gibbs posterior in terms of Kullback-Leibler divergence.

Three common families of variational posteriors are

$$\mathcal{F}_1 = \{\Phi_{m,\sigma^2},\ m \in \mathbb{R}^d,\ \sigma^2 \in \mathbb{R}^+\}$$
$$,\mathcal{F}_2 = \{\Phi_{m,\sigma^2},\ m \in \mathbb{R}^d,\ \sigma^2 \in (\mathbb{R}^+)^d\}$$
$$,\mathcal{F}_3 = \{\Phi_{m,\Sigma},\ m \in \mathbb{R}^d,\ \Sigma \in S^{d+}\},$$

where S^{d+} denotes the set of $d \times d$ symmetric and positive definite matrices, and Φ_{m,σ^2} denotes the multivariate normal distribution function with mean vector m and covariance $\sigma^2 I_d$. The set $\mathcal{F}_2$ is called the mean-field approximation and is one of the more common families used in variational approximations.

One example considered in (2) is a classification problem. The data consists of response predictor pairs $\mathcal{X}^n = ((Y_1, X_1), (Y_2, X_2), ..., (Y_n, X_n))$ with $Y_i \in \{-1, 1\}$ and $X_i \in \mathbb{R}^d$. The empirical risk function used is the hinge loss function

$$R_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \max(0, 1 - Y_i X_i^{\top}\theta)$$

for a parameter vector $\theta \in \mathbb{R}^d$. For the family $\mathcal{F}_1$, (2) show that the optimal variational posterior has parameters m and σ maximizing

$$\frac{1}{n}\left\{\sum_{i=1}^{n}(1-X_iY_im)\Phi\left(\frac{1-X_iY_i}{\sigma\|X_iY_i\|_2}\right)+\sum_{i=1}^{n}\sigma\|X_iY_i\|_2\phi\left(\frac{1-X_iY_i}{\sigma\|X_iY_i\|_2}\right)\right\}-\frac{\sqrt{d}\|m\|_2^2}{2}+\frac{d}{2}(\log\sigma^2-\frac{1}{\sigma^2\sqrt{d}}). \tag{3.4.1}$$

The rationale behind the expression in Equation 3.4.1 is that the normal distribution with parameters m and σ maximizing this expression has the lowest posterior averaged risk (see Section 2.3) among distributions in $\mathcal{F}_1$. As shown in (2), Equation 3.4.1 is a convex function in the parameters, so can be maximized using standard methods. See Section 7.4.3 for another application of variational posteriors.

CHAPTER 4

GIBBS POSTERIOR ASYMPTOTICS

As discussed in Section 1.4.3, asymptotic theory is useful for determining if a method can be relied upon to give reasonable answers to research questions given substantial data. Such theory serves to complement subjective analyses and expert opinions and may help data analysts decide between competing methods, giving preference to one with a better convergence rate, for instance. This chapter discusses results, both new in this dissertation and by other authors, for consistency, convergence rates, and convergence in distribution of Gibbs posteriors.

4.1 Consistency

4.1.1 Consistency of Gibbs posteriors

Recall that a sequence of Bayesian posteriors Π_n is said to be consistent for the true parameter $\theta^\star$ if for every $\Pi-$measureable set A containing $\theta^\star$, $\Pi_n(A) \overset{\text{i.p.}}{\to} 1$ with respect to the distribution $P_{\theta^\star}$. This same definition is applicable to Gibbs posteriors. One way to show consistency of the Gibbs posterior is to start with a result on consistency of M-estimators, such as the one given in Section 1.2, and try to extend it to the Gibbs posterior, since the Gibbs posterior is built from the empirical risk function used in M-estimation. Suppose the conditions from Proposition 1.2.1 hold (with one inequality flipped due to minimization versus maximization),

$$\sup_{\theta\in\Theta}|R_n(\theta) - R(\theta)| \overset{\text{i.p.}}{\to} 0, \tag{4.1.1}$$

$$\sup_{\theta:d(\theta,\theta^\star)\geq\epsilon} R(\theta) > R(\theta^\star) \tag{4.1.2}$$

for some distance measure $d(\cdot,\cdot) : \Theta \times \Theta \mapsto \mathbb{R}^+$. Write the Gibbs posterior probability of the complement of the set $A = \{\theta \in \Theta : d(\theta, \theta^\star) < \epsilon\}$,

$$\Pi_n(A^c) = \frac{\int_{A^c} \exp(-\omega R_n(\theta))d\Pi(\theta)}{\int_\Theta \exp(-\omega R_n(\theta))d\Pi(\theta)}$$

for some $\epsilon > 0$. If it can be shown that $\Pi_n(A^c) \overset{\text{i.p.}}{\to} 0$ in $P-$probability, then the Gibbs posterior is consistent with respect to $d(\cdot,\cdot)$. One strategy is to treat the numerator and denominator of $\Pi_n(A^c)$ separately and bound the numerator from above and the denominator from below.

First, consider bounding the denominator from below. Multiply the denominator, denoted D_n, by $e^{n(\omega R(\theta^\star)+\alpha)}$ for a positive constant α,

$$e^{n(\omega R(\theta^\star)+\alpha)}D_n = e^{n(\omega R(\theta^\star)+\alpha)}\int_\Theta \exp(-n\omega R_n(\theta))d\Pi(\theta).$$

Bound this product from below by restricting the domain of integration,

$$e^{n(\omega R(\theta^\star)+\alpha)}D_n \geq \int_{\{\theta:\omega R_n\theta-\omega R(\theta^\star)\leq\alpha/2\}} \exp[-n\omega(R_n(\theta) - \omega R(\theta^\star) - \alpha)]d\Pi(\theta)$$

Add and subtract $\mathsf{R}(\theta)$ in the exponent of the integrand, and apply the inequality in the domain of integration to get

$$\begin{aligned}
&\int_{\{\theta:\omega R(\theta)-\omega R(\theta^\star)\leq\alpha/2\}} * \exp(-n[\omega R_n(\theta) - \omega R(\theta) + \omega R(\theta) - \omega R(\theta^\star) - \alpha])d\Pi(\theta) \\
&\geq e^{n\alpha/2}\int_{\{\theta:\omega R(\theta)-\omega R(\theta^\star)\leq\alpha/2\}} \exp(-n[\omega R_n(\theta) - \omega R(\theta)])d\Pi(\theta).
\end{aligned}$$

Since the above integrand is non-negative, use Fatou's Lemma to evaluate the limit

$$\begin{aligned}
&\liminf_{n\to\infty}\int_{\{\theta:\omega R(\theta)-\omega R(\theta^\star)|\leq\alpha/2\}} \exp(-n[\omega R_n(\theta) - \omega R(\theta)])d\Pi(\theta) \\
&\geq \int_{\{\theta:\omega R(\theta)-\omega R(\theta^\star)|\leq\alpha/2\}} \liminf_{n\to\infty} \exp(-n[\omega R_n(\theta) - \omega R(\theta)])d\Pi(\theta) \\
&\geq \int_{\{\theta:\omega R(\theta)-\omega R(\theta^\star)|\leq\alpha/2\}} \exp(-\limsup_{n\to\infty} n\omega|R_n(\theta) - R(\theta)|)d\Pi(\theta)
\end{aligned}$$

With this limit the denominator may be bounded from below by

$$e^{-n\omega\delta}\Pi(\{\theta : \omega R(\theta) - \omega R(\theta^\star)| \leq \alpha/2\})$$

in $P-$probability. Since $\alpha > 0$ is arbitrary and $\delta > 0$ vanishes as $n \to \infty$, $e^{n(\omega R(\theta^\star)+\alpha)}D_n$ diverges in $P-$probability as $n \to \infty$ as long as the prior distribution Π places positive mass on the set $\{\theta : \omega R(\theta) - \omega R(\theta^\star)| \leq \alpha/2\}$. Hence, D_n is bounded below by $Ce^{-n(\omega R(\theta^\star)+\alpha)}$ for some $C > 0$ in $P-$probability.

Next, bound the numerator from above. Write the numerator of $\Pi_n(A^c)$ as

$$N_n(A^c) = \int_{\{\theta:d(\theta,\theta^\star)>\epsilon\}} \exp(-n[\omega R_n(\theta)])d\Pi(\theta).$$

Add and subtract $R(\theta)$ from the exponent in the numerator to obtain

$$\int_{\{\theta:d(\theta,\theta^\star)>\epsilon\}} \exp(-n\omega[R_n(\theta) - R(\theta) + R(\theta)])d\Pi(\theta).$$

By the conditions in Equation 4.1.1, $R_n(\theta) - R(\theta)$ can be bounded uniformly over the set of integration by $\delta > 0$ in $P-$probability and $R(\theta) > R(\theta^\star) + \eta$ for some $\eta(\epsilon) > 0$. Then,

$$N_n(A^c) \leq e^{-n\omega[-\delta+R(\theta^\star)+\eta]}.$$

Combining the bounds on numerator and denominator,

$$\begin{aligned} \Pi_n(A^c) = \frac{N_n(A^c)}{D_n} &\lesssim \frac{e^{-n\omega[-\delta+R(\theta^\star)+\eta]}}{e^{-n(\omega R(\theta^\star)+\alpha)}} \\ &= e^{n\omega\delta+n\alpha-n\omega\eta} \end{aligned}$$

where $x \lesssim y$ means $x \leq cy$ for some constant $c > 0$. As $n \to \infty$, δ vanishes, but $\eta > 0$ is a fixed value dependent on ϵ. So, if α is chosen as, for instance, $\alpha < \omega\eta/2$, the bound vanishes in $P-$probability as $n \to \infty$.

The preceding calculations prove the following proposition.

Proposition 4.1.1 *Suppose that Equation 4.1.1 hold and that* Π *places positive probability on sets* $\{\theta : \omega R(\theta) - \omega R(\theta^\star)| \leq \alpha/2\}$ *for any* $\alpha > 0$*, then* $\Pi_n(\{\theta : d(\theta, \theta^\star) \leq \epsilon\}) \overset{i.p.}{\to} 1$ *in* $P-$*probability and* Π_n *is called consistent with respect to* $d(\cdot, \cdot)$*.*

4.1.2 Application of consistency to inference on a median

Section 1.3 described a statistical inference problem for a population median and mentioned that a consistent M-estimator for the median could be found. In particular, for a random sample $\mathcal{X}^n = (\mathcal{X}_1, \mathcal{X}_2, ..., \mathcal{X}_n)$ drawn from a distribution P with median $\theta^\star$, the estimator

$$\hat{\theta}_n := \arg\min_{\theta \in \mathbb{R}} \frac{1}{n} \left| \sum_{i=1}^{n} sign(\mathcal{X}_i - \theta) \right|$$

is consistent for $\theta^\star$. Note that $\hat{\theta}_n$ is not a unique minimum but may be taken to be the sample median for convenience. As noted in Section 1.3, the risk function is $R(\theta) = |P(\mathcal{X} \geq \theta) - P(\mathcal{X} < \theta)|$. Consistency may be shown by verifying the conditions of Equation 4.1.1; see Proposition 1.2.1. The uniform convergence of $R_n(\theta)$ to $R(\theta)$ follows from the Glivenko-Cantelli Theorem. Let $1(\cdot)$ denote the indicator function and $F_n(x) = \frac{1}{n} \sum_{i=1}^{n} 1(\mathcal{X}_i \leq x)$ denote the empirical distribution function for i.i.d. random variables $\mathcal{X}_1, \mathcal{X}_2, ..., \mathcal{X}_n$. The Glivenko-Cantelli Theorem states for i.i.d. random variables $\mathcal{X}_1, \mathcal{X}_2, ..., \mathcal{X}_n$ with distribution function F, $\sup_{x \in \mathbb{X}} |F_n(x) - F(x)| \to 0$ almost surely with respect to P. The empirical risk function may be written $R_n(\theta) = |\sum_{i=1}^n \frac{1}{n}1(\mathcal{X}_i \geq \theta) - \frac{1}{n}1(\mathcal{X}_i < \theta)|$. Write the absolute difference between the empirical risk and the risk function as

$$|R_n(\theta) - R(\theta)| = \left| |\sum_{i=1}^{n} \frac{1}{n}1(\mathcal{X}_i \geq \theta) - \frac{1}{n}1(\mathcal{X}_i < \theta)| - |P(\mathcal{X} \geq \theta) - P(\mathcal{X} < \theta)| \right|$$

and use the reverse triangle inequality to bound the difference as

$$|R_n(\theta) - R(\theta)| \leq \left| \sum_{i=1}^{n} \frac{1}{n}1(\mathcal{X}_i \geq \theta) - \frac{1}{n}1(\mathcal{X}_i < \theta) - (P(\mathcal{X} \geq \theta) - P(\mathcal{X} < \theta)) \right|.$$

Then, use the triangle inequality to obtain

$$|R_n(\theta) - R(\theta)| \leq \left| \sum_{i=1}^{n} \frac{1}{n}1(\mathcal{X}_i \geq \theta) - P(\mathcal{X} \geq \theta) \right| + \left| P(\mathcal{X} < \theta) - \frac{1}{n}\sum_{i=1}^{n} 1(\mathcal{X}_i < \theta) \right|.$$

Both of the terms on the right hand side of the above display converge uniformly to 0 almost surely with respect to P by the Glivenko-Cantelli Theorem. Hence, the difference $|R_n(\theta) - R(\theta)|$ converges uniformly to 0. Since $R(\theta) = |P(\mathcal{X} \geq \theta) - P(\mathcal{X} < \theta)|$ is clearly minimized uniquely at the population median, both conditions in Equation 4.1.1 hold. Any prior with positive support on the entire real line satisfies Proposition 4.1.1. For example, a Gibbs posterior using a normal prior with mean 0 and standard deviation 1 is consistent for $\theta^\star$ by Proposition 4.1.1.

The main challenge to applying Proposition 4.1.1 is verifying the uniform convergence in Equation 4.1.1. Since the linking function for the median boils down to a sum of indicator functions, the argument is essentially provided by the Glivenko-Cantelli Theorem. In Chapter 5, the uniform convergence condition is verified in a classification problem, and again, it is simplified by the fact that the empirical risk function can be expressed as a sum of indicators. When the functions involved are more complicated than indicators, the consistency result is still valid, but verifying the uniform convergence becomes more challenging.

4.2 Convergence rates of Gibbs posteriors

4.2.1 A result based on M-estimator convergence rates

In Section 1.3, M-estimation was introduced and the consistency of M-estimators was discussed. In Section 4.1, it was shown that sufficient conditions for consistency of M-estimators are also sufficient for consistency of the Gibbs posterior with only the addition of a mild condition on the prior distribution. In this section, this idea of basing Gibbs posterior asymptotics on M-estimator asymptotics will be pushed further, and it will be shown that Gibbs posteriors not only inherit the consistency of M-estimators, but also their precise convergence rate. Chapter 5 of (100) is a good resource on M-estimation, providing several results and examples for calculating the convergence rates of M-estimators.

Let $d(\cdot,\cdot)$ be a distance measure defined on the parameter space $d(\cdot,\cdot) : \Theta \times \Theta \mapsto \mathbb{R}^+$. Define the empirical distribution $\mathbb{P}_n = n^{-1} \sum_{i=1}^n I(\mathcal{X}_i)$. Next, for P and $\mathbb{P}_n$ defined before, let $\mathbb{G}_n f = n^{1/2}(\mathbb{P}_n f - Pf)$ be the empirical process. A good review of empirical processes is given in (100), Chapter 19. The M-estimator convergence rate relies on two assumptions: a well-separated property of the risk $R(\theta)$, and a bound on the expectation of the empirical process $\mathbb{G}_n$; see Equation 4.2.1 and Equation 4.2.2. The well-separated property states that the minimizer $\theta^\star$ of $R(\theta)$ is sufficiently unique; for instance, there is no parameter sequence

θ_n not converging to $\theta^\star$ with $R(\theta_n) \to R(\theta^\star)$. The bound on the empirical process sets the speed at which $R_n(\theta)$ converges to $R(\theta^\star)$, in probability, for θ in a neighborhood of $\theta^\star$. Given these conditions, in order for the Gibbs posterior based on $R_n(\theta)$ to inherit the convergence rate of the M-estimator, it is sufficient that the prior places sufficient mass on "risk function neighborhoods"; see Equation 4.2.3. The following result is applied to quantile regression in Section 4.2.3 and to a medical statistics application in Chapter 5. In the statement of Theorem 4.2.1 below, note $x \lesssim y$, $x \gtrsim y$ means $\exists C > 0$ s.t. $x \le Cy$, $x \ge Cy$.

Theorem 4.2.1 *Assume that for fixed constants $C_0, C_1, C_2 > 0$ and $\alpha \ge 2\beta$, for every n, and for every sufficiently small $\delta > 0$,*

$$\sup_{d(\theta,\theta^\star)>\delta} \{R(\theta^\star) - R(\theta)\} \le -C_1\delta^\alpha, \tag{4.2.1}$$

$$E \sup_{d(\theta,\theta^\star)<\delta} |\mathbb{G}_n(\ell(\mathcal{X},\theta) - \ell(\mathcal{X} - \theta^\star))| \le C_2\delta^\beta. \tag{4.2.2}$$

For the prior distribution Π on Θ, assume that

$$\Pi(\{\theta : R(\theta) - R(\theta^\star) < t_n\}) \gtrsim \exp(-C_0 n t_n). \tag{4.2.3}$$

where $t_n = n^{-\frac{\xi}{2\alpha+2\beta}}$ for some $0 < \xi < \alpha$. Let $r = r(\alpha, \beta) = (2\alpha - 2\beta)^{-1}$. Then the Gibbs posterior Π_n in (Equation 2.2.1) *satisfies $\Pi_n(A_n) = o_P(1)$ as $n \to \infty$, where $A_n = \{\theta : d(\theta, \theta^\star) > a_n n^{-r}\}$ for any diverging sequence a_n.*

For the proof of Theorem 4.2.1 it will be convenient to rewrite the posterior distribution in Equation 2.2.1 as

$$\Pi_n(A_n) = \frac{N_n(A_n)}{D_n} = \frac{\int_{A_n} e^{-n\{R_n(\theta)-R_n(\theta^\star)\}}\,\Pi(d\theta)}{\int_\Theta e^{-n\{R_n(\theta)-R_n(\theta^\star)\}}\,\Pi(d\theta)}.$$

For simplicity, it is assumed $\omega = 1$, but any constant $\omega > 0$ or suitably vanishing $\omega = \omega_n$ will do; see Proposition 4.2.1. Then the goal is to obtain appropriate bounds on the numerator and denominator.

Equation 4.2.1 and Equation 4.2.2 provide control over the numerator of the Gibbs posterior, which is described in the following Lemma.

Lemma 4.2.1 *Let* $s_n = a_n n^{-r}$ *where* $r = r(\alpha, \beta) = (2\alpha - 2\beta)^{-1}$*, and* a_n *is any diverging sequence. Then, assuming Equation 4.2.1 and Equation 4.2.2 hold, there exists* $K > 0$ *such that*

$$P\Big(\sup_{d(\theta,\theta^\star)>s_n}\{R_n(\theta^\star) - R_n(\theta)\} > -Ks_n^\alpha\Big) \to 0, \quad \textit{as } n \to \infty.$$

Proof of Lemma 4.2.1

Start with the identity

$$R_n(\theta^\star) - R_n(\theta) = \{R(\theta^\star) - R(\theta)\} - n^{-1/2}\mathbb{G}_n(l(\mathcal{X}, \theta) - l(\mathcal{X}, \theta^\star)),$$

where $\mathbb{G}_n f = n^{1/2}(\mathbb{P}_n f - Pf)$ is the empirical process. Next, since the supremum of a sum is no more than the sum of the suprema, see that

$$\sup_{d(\theta,\theta^\star)>\epsilon} \{R_n(\theta^\star) - R_n(\theta)\} \leq \sup_{d(\theta,\theta^\star)>\epsilon} \{R(\theta^\star) - R(\theta)\} + n^{-1/2} \sup_{d(\theta,\theta^\star)>\epsilon} |\mathbb{G}_n(\ell(\mathcal{X},\theta) - \ell(\mathcal{X},\theta^\star))|,$$

also taking the absolute value of the empirical process. From Equation 4.2.1, see that

$$\sup_{d(\theta,\theta^\star)>\epsilon} \{R_n(\theta^\star) - R_n(\theta)\} \leq -C_1\epsilon^\alpha + n^{-1/2} \sup_{d(\theta,\theta^\star)>\epsilon} |\mathbb{G}_n(\ell(\mathcal{X},\theta) - \ell(\mathcal{X},\theta^\star))|.$$

Now, following the proof of Theorem 5.52 from (100) or of Theorem 1 in (107), introduce "shells" $\{\theta : 2^m\epsilon < d(\theta,\theta^\star) \leq 2^{m+1}\epsilon\}$ for integers m. On these shells, use both Equation 4.2.1 and Equation 4.2.2. That is,

$$\begin{aligned}
&\sup_{d(\theta,\theta^\star)>s_n} \{R_n(\theta^\star) - R_n(\theta)\} > -Ks_n^\alpha \\
&\qquad \Longrightarrow \sup_{2^m s_n<d(\theta,\theta^\star)\leq 2^{m+1}s_n} \{R_n(\theta^\star) - R_n(\theta)\} > -Ks_n^\alpha \quad \exists\, m \geq 0 \\
&\qquad \Longrightarrow n^{-1/2} \sup_{2^m s_n<d(\theta,\theta^\star)<2^{m+1}s_n} |\mathbb{G}_n(\ell(\mathcal{X},\theta) - \ell(\mathcal{X},\theta^\star))| \geq C_1(2^m s_n)^\alpha - Ks_n^\alpha \\
&\qquad \Longrightarrow n^{-1/2} \sup_{d(\theta,\theta^\star)\leq 2^{m+1}s_n} |\mathbb{G}_n(\ell(\mathcal{X},\theta) - \ell(\mathcal{X},\theta^\star))| \geq C_1(2^m s_n)^\alpha - Ks_n^\alpha,
\end{aligned}$$

If $K \le C_1/2$, then $C_1(2^m s)^\alpha - Ks^\alpha \ge C_1(2^m s)^\alpha/2$ for all $m \ge 0$.

$$P\Big(\sup_{d(\theta,\theta^\star)>s_n} \{R_n(\theta^\star) - R_n(\theta)\} > -Ks_n^\alpha\Big)$$
$$\le \sum_{m\ge 0} P\Big(n^{-1/2} \sup_{d(\theta,\theta^\star)<2^{m+1}s_n} |\mathbb{G}_n(\ell(\mathcal{X},\theta) - \ell(\mathcal{X},\theta^\star))| \ge C_1(2^m s_n)^\alpha/2\Big)$$

To the summands, apply Markov's inequality and Equation 4.2.2 to get

$$P\Big(n^{-1/2} \sup_{d(\theta,\theta^\star)<2^{m+1}s_n} |\mathbb{G}_n(\ell(\mathcal{X},\theta) - \ell(\mathcal{X},\theta^\star))| \ge C_1(2^m s_n)^\alpha/2\Big) \le C_3 2^{m(\beta-\alpha)} a_n^{(\beta-\alpha)}$$

for some $C_3 > 0$ depending on C_1 and C_2. Since $\beta - \alpha < 0$, the sum over m converges and the upper bound vanishes since $a_n^{\beta-\alpha} \downarrow 0$, completing the proof.

The following lemma yields a necessary lower bound on the denominator of the Gibbs posterior distribution. The result is exactly Lemma 1 of (86). A simpler version, which is often applicable, is presented in Corollary 4.2.1; if the variance of the difference in the linking functions is bounded by a constant times the expectation of their difference, then the S_n neighborhoods may be simplified.

Lemma 4.2.2 *Let* t_n *be a sequence of positive numbers such that* $nt_n \to \infty$ *and set* $S_n = \{\theta : \max\{R(\theta) - R(\theta^\star), V(\ell(\mathcal{X},\theta) - \ell(\mathcal{X},\theta^\star))\} \le Mt_n\}$ *for some* $M > 0$*. Then* $\int e^{-[R_n(\theta)-R_n(\theta^\star)]}\, d\Pi \gtrsim \Pi(S_n)\exp(-2nt_n)$ *with* $P_{\theta^\star}$*−probability converging to* 1 *as* $n \to \infty$*.*

Corollary 4.2.1 *Let* t_n *be a sequence of positive numbers such that* $nt_n \to \infty$*. Suppose* $V(\ell(\mathcal{X},\theta) - \ell(\mathcal{X},\theta^\star)) \le M_1[R(\theta) - R(\theta^\star)]$ *for some constant* $M_1 > 1$*. A sufficient condi-*

tion is that $\ell(\mathcal{X},\theta)$ *is bounded almost surely. Then, there exists a constant* $M_2 > 0$ *such that* $\int e^{-[R_n(\theta)-R_n(\theta^\star)]}\, d\Pi \gtrsim \Pi(G_n)\exp(-2nt_n)$ *with* $P_{\theta^\star}$*−probability converging to* 1 *as* $n \to \infty$ *where* $G_n = \{\theta : \{R(\theta) - R(\theta^\star) \le M_2 t_n\}$.

Proof of Corollary 4.2.1

Choose $M > 0$ in Lemma 4.2.2. Set $M_2 = M/M_1$. Then, $G_n \subset S_n$ and the bound holds by Lemma 4.2.2.

Also, see that for a bounded linking function $\ell(\mathcal{X},\theta)$, there exists a constant $M_3 > 0$ such that

$$
\begin{aligned}
V(\ell(\mathcal{X},\theta) - \ell(\mathcal{X},\theta^\star)) &\le E_P\{[\ell(\mathcal{X},\theta) - \ell(\mathcal{X},\theta^\star)]^2\} \\
&= \int [\ell(x,\theta) - \ell(x,\theta^\star)]^2 P(dx) \\
&\le M_3 \int \ell(x,\theta) - \ell(x,\theta^\star) P(dx) \\
&= M_3[R(\theta) - R(\theta^\star)],
\end{aligned}
$$

thereby providing a simple criterion for applying Corollary 4.2.1.

Proof of Theorem 4.2.1

For the denominator, for a suitable sequence t_n specified below, it follows that $D_n \gtrsim e^{-3nt_n}$ by Lemma 4.2.2.

For the numerator, Lemma 4.2.1 provides a uniform bound, namely,

$$
P\Big(\sup_{|\theta-\theta^\star|>s_n} \{R_n(\theta^\star) - R_n(\theta)\} > -Ks_n^\alpha\Big) \to 0, \quad \text{as } n \to \infty
$$

where $s_n = a_n n^{-\frac{1}{2\alpha-2\beta}}$ as in Theorem 4.2.1. Therefore, $\log N_n(A_n) \leq -K a_n^\alpha n^{\frac{\alpha-2\beta}{2\alpha-2\beta}}$ with $P-$probability converging to 1.

Finally, put together the in-probability bounds on the numerator and denominator:

$$\Pi_n(A_n) = \frac{N_n(A_n)}{D_n} \lesssim \exp(-L(a_n^\alpha n^{\frac{\alpha-2\beta}{2\alpha-2\beta}} - nt_n))$$

for some $L > 0$. Choose $t_n = \tilde{a}_n n^{-\alpha/(2\alpha-2\beta)}$ such that $\tilde{a}_n$ diverges and $a_n^\alpha - \tilde{a}_n$ also diverges. This choice satisfies the requirement in Lemma 4.2.2. The numerator simplifies to $-L(a_n^\alpha - \tilde{a}_n) \to -\infty$, by construction. Hence, the bound vanishes, completing the proof.

When the conditions of Theorem 4.2.1 hold, the Gibbs posterior mean inherits the rate of convergence of the posterior.

Corollary 4.2.2 *Under the conditions of Theorem 4.2.1, if the prior mean for θ exists, then the posterior mean $\bar{\theta}_n$ satisfies $\bar{\theta}_n - \theta^\star = O_P(a_n n^{-r})$ as $n \to \infty$, for any sequence $a_n \to \infty$.*

Proof of Corollary 4.2.2

Set $s_n = a_n n^{-r}$. Next, define $\acute{s}_n = \acute{a}_n n^{-r(\gamma)}$, where $\acute{a}_n$ is such that $\acute{a}_n \to \infty$ but $\acute{a}_n / a_n \to 0$. Now partition $\mathbb{R}$ as $\{\theta : |\theta - \theta^\star| \leq \acute{s}_n\} \cup \{\theta : |\theta - \theta^\star| > \acute{s}_n\}$, and write

$$|\bar{\theta}_n - \theta^\star| \leq \int |\theta - \theta^\star| \Pi_n(d\theta) \leq \acute{s}_n + \int_{|\theta-\theta^\star|>\acute{s}_n} |\theta - \theta^\star| \Pi_n(d\theta), \tag{4.2.4}$$

where the first inequality is by Jensen. From the proof of Theorem 5.3.1, the posterior away from $\theta^\star$ is bounded by the prior times some $Z_n = o_P(1)$, uniformly in θ. That is,

$$\int_{|\theta-\theta^\star|>\acute{s}_n} |\theta-\theta^\star|\,\Pi_n(d\theta) \leq Z_n \int |\theta-\theta^\star|\,\Pi(d\theta).$$

In fact, one can bound Z_n more precisely:

$$Z_n \lesssim \exp\Big\{-M\Big(\acute{a}_n^{\alpha} n^{\frac{\alpha-2\beta}{2\alpha-2\beta}} - \tilde{a}_n n^{\frac{\alpha-2\beta}{2\alpha-2\beta}}\Big)\Big\},$$

for any divergent sequence $\tilde{a}_n$ such that $\acute{a}_n^{\alpha} - \tilde{a}_n \to \infty$. Dividing through Equation 4.2.4 by s_n see that $s_n^{-1}|\bar{\theta}_n - \theta^\star|$ is bounded by a constant times

$$\acute{a}_n/a_n + e^{-\zeta_n}\int |\theta-\theta^\star|\,\Pi(d\theta),$$

where $\zeta_n = Mn^{\frac{\alpha-2\beta}{2\alpha-2\beta}}[\acute{a}_n^{\alpha} - \tilde{a}_n] + \log a_n - \frac{1}{2\alpha-2\beta}\log n$. The first term in the upper bound goes to zero by the choice of $\acute{a}_n$. The second term goes to zero provided that the prior mean exists and $\zeta_n \to \infty$ as $n \to \infty$. The former condition was assumed, and the latter holds due to choosing $\acute{a}_n^{\alpha} - \tilde{a}_n \to \infty$. Then, $\tilde{\theta}_n - \theta^\star = o_P(s_n)$, as was to be proved.

4.2.2 Convergence rates and scaling of the Gibbs posterior

Proposition 4.2.1 *For (α, β) satisfying the assumptions of Theorem 4.2.1, write $r(\alpha,\beta) = (\alpha-2\beta)/(2\alpha-2\beta)$. Then the conclusion of Theorem 4.2.1 holds if the learning rate, $\omega = \omega_n$, appearing in Equation 2.2.1, vanishes no faster than $n^{-r(\alpha,\beta)}$.*

The result follows from the proof of Theorem 4.2.1, so a detailed proof is omitted.

4.2.3 Application to quantile regression

In Section 2.1, quantile regression was introduced as an example motivating the need for Gibbs posteriors. Recall the model for data $\mathcal{X} = (Y, X) \in \mathbb{R} \times \mathbb{R}^n$,

$$Q_\tau(Y \mid X) = X^\top \theta, \tag{4.2.5}$$

from Equation 2.1.1, where the τ^{th} conditional quantile of Y given X, denoted $Q_\tau(Y \mid X)$, is model as a linear combination of X.

In this section, quantile regression is used to provide an example of a model in which the Gibbs posterior convergence rate can be computed using Theorem 4.2.1. It is well-known that the M-estimator minimizing the quantile regression empirical risk in Equation 2.1.2 converges to the true parameter at rate $n^{-1/2}$. In order to show the $n^{-1/2}$ convergence rate applies also to the Gibbs posterior with linking function the empirical risk in Equation 2.1.2, it is necessary to verify Equation 4.2.1 and Equation 4.2.2 from Theorem 4.2.1 with $\alpha = 2$ and $\beta = 1$. These conditions typically require significant effort to verify. Corollary 5.53 in (100) Chapter 5 helps to identify α and β in Equation 4.2.1 and Equation 4.2.2 for linking functions $\ell(\mathcal{X}, \theta)$ that are Lipschitz and have corresponding risk functions $R(\theta)$ admitting a second-order Taylor expansion at $\theta^\star$. The following assumptions, 4.2.1 and 4.2.2 are sufficient for establishing the Lipschitz condition and the Taylor expansion in quantile regression.

Assumption 4.2.1 *The marginal distribution* G *of* X*, which is free of unknown parameters, is such that* $\mathsf{E}(\mathsf{X}\mathsf{X}^\top)$ *exists and is positive definite.*

Assumption 4.2.2 *The conditional distribution* Y*, given* $\mathsf{X} = x$*, has at least one finite moment and admits a continuous density* $f_x(y)$ *such that* $f_x(x^\top\theta^\star)$ *is bounded away from zero for* G*—almost all* x*.*

Proposition 4.2.2 *Consider i.i.d. data* $\mathcal{X}_i = (Y_i, X_i)$*,* $i = 1, \ldots, n$*, under the model given in Equation 2.1.1, with fixed* $\tau \in (0, 1)$*, and suppose that Assumptions 4.2.1 and 4.2.2 hold. If* $\theta^\star = \theta^\star(\tau)$ *is the true value, then the Gibbs posterior probability* $\Pi_n(A_n) = o_P(1)$ *as* $n \to \infty$*, where* $A_n = \{\theta : \|\theta - \theta^\star\| > a_n n^{-1/2}\}$ *and* a_n *is any diverging sequence, for any prior* Π *with continuous density bounded away from zero on a neighborhood of* $\theta^\star$*.*

Proof of Proposition 4.2.2 Proposition 4.2.2 confirms that the Gibbs posterior shares the same $n^{-1/2}$ convergence rate of the M-estimator presented in (53), Theorem 4.1. This proof is short and straightforward for the case of i.i.d. observations. A similar result for the case of independent but not i.i.d. observations is given in (91).

We can use Corollary 5.53 in (100) to show that $\alpha = 2$ and $\beta = 1$, yielding the $n^{-1/2}$ rate of convergence. Suppress the dependence on τ and refer to a generic parameter as θ and the true parameter minimizing the risk $R(\theta)$ as $\theta^\star$.

First, note that the linking function

$$\ell(\mathcal{X}, \theta) = |(y - x^\top\theta)(\tau - I_{\{y < x^\top\theta\}})|$$

for the quantile regression problem satisfies a Lipschitz property, i.e., that

$$|\ell(\mathcal{X},\theta) - \ell(\mathcal{X},\theta')| \leq (1+\tau)\|x\|\,\|\theta - \theta'\|.$$

This follows from looking at each of the three cases—$y > x^\top\theta_1 > x^\top\theta_2$, $x^\top\theta > y > x^\top\theta'$, and $x^\top\theta > x^\top\theta' > y$—and an application of the Cauchy–Schwartz inequality. The Lipschitz constant, $\mathcal{L}(x) = (1+\tau)\|x\|$, which depends only on x in this case, satisfies $\mathsf{E}_P(\mathcal{L}(x)^2) < \infty$ by Assumption 4.2.1. According to (100), Corollary 19.35, this implies that condition Equation 4.2.2 holds with $\beta = 1$.

Next, check that the risk function $R(\theta)$ admits a suitable second-order Taylor approximation at $\theta^\star$. Towards this, write

$$R(\theta) = \int_{\mathbb{X}} \Big[(\tau-1)\int_{-\infty}^{x^\top\theta}(y - x^\top\theta)f_x(y)\,dy + \tau\int_{x^\top\theta}^{\infty}(y - x^\top\theta)f_x(y)\,dy\Big]\,G(dx),$$

where G is the marginal distribution of X, defined on the space $\mathbb{X}$, and F_x is the conditional distribution function of Y, given $X = x$, and f_x is the corresponding density function. Differentiation with respect to θ gives

$$\dot{R}(\theta) = \int x\{F_x(x^\top\theta) - \tau\}\,G(dx) \quad \text{and} \quad \ddot{R}(\theta) = \int xx^\top f_x(x^\top\theta)\,G(dx).$$

Differentiating under the integral is permissible by the continuity and moment conditions imposed. Therefore, since $\dot{R}(\theta^\star) = 0$, the following second-order Taylor approximation holds:

$$R(\theta) = R(\theta^\star) + \tfrac{1}{2}(\theta - \theta^\star)^\top \ddot{R}(\theta^\star)(\theta - \theta^\star) + o(\|\theta - \theta^\star\|^2).$$

Based on the given assumptions, see that $\ddot{R}(\theta^\star)$ exists and is positive definite which, according to (100), page 76, implies Equation 4.2.1 with $\alpha = 2$.

Finally, by Theorem 4.2.1, conclude that the rate of convergence for the Gibbs posterior is n^{-r}, where $r = (2\alpha - 2\beta)^{-1} = \frac{1}{2}$. Therefore, the claimed $n^{-1/2}$ rate holds as long as the prior distribution for θ has a density function which is continuous and bounded away from 0 in a neighborhood of $\theta^\star$.

As a numerical illustration of Proposition 4.2.2, consider a small simulation experiment. For $\tau = 0.5$, consider the model

$$Y_i = \theta_0 + \theta_1 X_i + e_i, \quad i = 1, \ldots, n,$$

where $\theta_0 = 2$, $\theta_1 = 1$, $e_i \overset{\text{i.i.d.}}{\sim} \mathsf{N}(0, 4)$, and $X_i \overset{\text{i.i.d.}}{\sim} \mathsf{ChiSq}(2) - 2$. A total of 2000 sets of data were simulated from this model with sample sizes ranging from $n = 101$ to $n = 2100$. A Gibbs posterior was sampled on each data set using an improper prior, i.e. $\pi(x) \propto 1$, and the posterior mean of each parameter, denoted $\tilde{\theta}_0$ and $\tilde{\theta}_1$, was recorded. Since Proposition 4.2.2 determines the convergence rate of the Gibbs posterior to be $n^{-1/2}$, the posterior means should converge to θ_0 and θ_1 roughly at this rate. In other words, $\log|\tilde{\theta}_i - \theta_i| \approx b - \frac{1}{2}\log n$ for each $i = 1, 2$, since the Gibbs posterior convergence rate is proportional to $n^{-1/2}$. Figure 4.2.1 shows precisely this relationship for the simulated data.

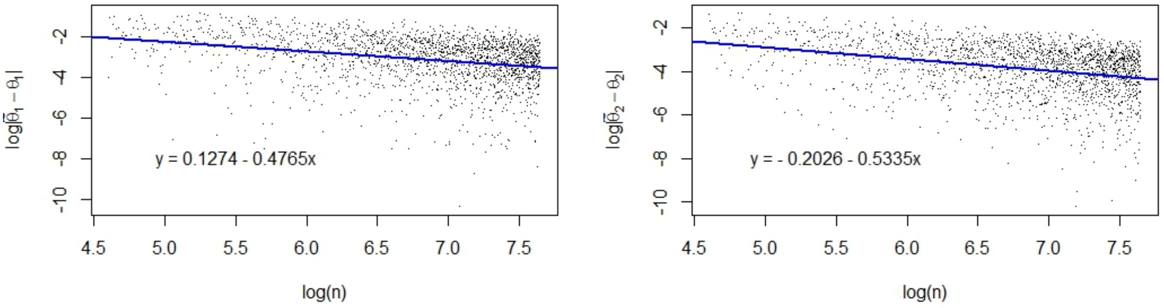


Figure 4.2.1. Left: the logarithm of the absolute difference between the Gibbs posterior mean values of the intercept for each data set and the true parameter value are regressed against the logarithm of the sample size. Right: the same regression is repeated for the slope parameter. The fitted least-squares line is shown in blue along with its equation.

4.3 Adaptive convergence rates for sets with smooth boundaries

4.3.1 Statistical estimation of infinite dimensional parameters

In previous sections the statistical inference problem has always concerned an interest parameter with a finite dimension, say a vector $\theta \in \mathbb{R}^p$ for an integer $p \geq 1$. However, there are important problems in which the parameter of interest may actually have *infinite* dimension. Some of these problems include probability density estimation, hazard function estimation in survival analysis, and regression. For concreteness, consider a regression problem with infinite dimensional regression function

$$Y = \theta(X) + \epsilon$$

where data $\mathcal{X} = (Y, X)$ is a response and predictor pair, ϵ is a mean-zero random variable, and $\theta(x)$ is an unknown function. It is typical to assume the predictor domain can be bounded, i.e. $X \in [c, d]$, and that the function $\theta(x)$ is bounded on this domain, i.e. $Y \in [a, b]$ for real numbers a, b, c, and d. The function $\theta(x)$ is said to be of infinite dimension because $\theta(x)$ can only be fully characterized by knowing the infinite parameter values $\theta(x)$ for $x \in [c, d]$. While some particular classes of functions, such as linear functions, can be characterized by a finite number of parameters, like a slope and intercept, this requires making more assumptions about $\theta(x)$ beyond boundedness on $[a, b]$.

Two successful methods for estimating general regression functions are linear smoothers and basis function expansions. An excellent introduction to these methods is presented in (104). Linear smoothers estimate $\theta(x)$ by a linear combination $\hat{\theta}(x) = \sum_{i=1}^{n} g_i(x)Y_i$ for functions $g_1, g_2, ..., g_n$ such that $\sum_{i=1}^{n} g_i(x) = 1$. An important special case of a linear smoother is the Nadaraya-Watson kernel estimator

$$\hat{\theta}_h(x) = \frac{\sum_{i=1}^{n} K(\frac{x-X_i}{h})Y_i}{\sum_{i=1}^{n} K(\frac{x-X_i}{h})};$$

see (74) and (105). The parameter $h > 0$ is called the bandwidth and influences the smoothness of $\hat{\theta}_h(x)$. When h is large, points far from x are given more weight than when h is small, which tends to smooth the estimate. For example, one type of kernel function is the Epanechnikov kernel $K(t) = \frac{3}{4}(1 - t^2)$ for $|t| < 1$ and 0 otherwise. For the Epanechnikov kernel, only points (Y_i, X_i) with $|x - X_i| \leq h$ will be included in the estimate of $\hat{\theta}_h(x)$.

The second common approach to function estimation uses basis function expansions. The Weierstrass Approximation Theorem from real analysis states than continuous, bounded functions on closed intervals can be approximated arbitrarily well by polynomials. In other words, there is real number sequence $(\beta_0, \beta_1, ...)$ such that $\theta(x) = \sum_{i=0}^{\infty} \beta_i x^i$ for all $x \in [c, d]$. This result suggests the approximation $\hat{\theta}(x) = \sum_{i=0}^{k} \beta_i x^i$ for some positive integer k, which is just a truncation of the series at k. Extensions of the Weierstrass theorem apply to other bases of functions rather than $(1, x, x^2, ...)$. Commonly used bases include trigonometric functions and splines. An analogous result for a trigonometric basis says that there exist constants $(a_0, a_1, ...)$ and $(b_1, b_2, ...)$ such that $\theta(x) = a_0 + \sum_{j=1}^{\infty} a_j \sin(jx) + \sum_{j=1}^{\infty} b_j \cos(jx)$. One type of spline function basis used in regression is the cubic b-spline basis, defined recursively as

$$
\begin{aligned}
B_{i,1}(x) &= 1(x \in [t_i, t_{i+1}]) \\
B_{i,k}(x) &= \frac{x - t_i}{t_{i+k-1} - t_i} B_{i,k-1}(x) + \frac{t_{i+k} - x}{t_{i+k} - t_{i+1}} B_{i+1,k-1}(x),
\end{aligned}
$$

where $t_{-2}, t_{-1}, t_0 < c$, and $t_{D+1}, t_{D+2}, t_{D+3} > d$ are called outer knots, while $t_1, ..., t_D \in [c, d]$ are called inner knots. Then, $\theta(x)$ is approximated by $\hat{\theta}_{D,\beta}(x) = \sum_{j=1}^{D} \beta_j B_{j,D}(x)$ where $\beta = (\beta_1, ..., \beta_D) \in (\mathbb{R}^+)^D$ is a vector of coefficients. Figure 4.3.1 shows an example of a linear smoother using an Epanechnikov kernel and a b-spline fit to data on the times between eruptions and duration of eruptions from the Old Faithful geyser.

4.3.2 Gibbs posterior convergence rate for a set with a smooth boundary

Sets with smooth boundaries are studied in (55), (68), (61), and (92) and are widely applied in image analysis. For example, the set of pixels depicting a tumor in a medical scan may be assumed to have a smooth boundary; see Figure 4.3.2 reproduced from (13). Recovery of the set from the background (or, equivalently, recovery of the boundary) could be helpful in diagnosis and treatment. Although the context can deviate from that of images made up of pixels, this analogy is fitting and simple to understand so will be carried on throughout this section.

Let $\Omega \subset \mathbb{R}^2$ be a bounded region that represents the frame of the image; typically, Ω will be a square, say, $[-\frac{1}{2}, \frac{1}{2}]^2$, but, generally, assume only that Ω is scaled to have unit Lebesgue measure. Data consists of pairs $\mathcal{X}_i = (X_i, Y_i)$, $i = 1, \ldots, n$, where X_i is a pixel location in Ω and Y_i is an intensity measurement at that pixel. The range of Y_i is context-dependent, and applications with both binary and real-valued cases are considered in Chapter 6. The model asserts that there is a region $\Gamma \subset \Omega$ such that the intensity distribution is different depending on whether the pixel is inside or outside Γ. Consider the following model for the joint distribution P_Γ of pixel location and intensity, $\mathcal{X} = (X, Y)$:

$$\begin{aligned} X &\sim g(x), \\ Y \mid (X = x) &\sim f_\Gamma(y)\, 1(x \in \Gamma) + f_{\Gamma^c}(y)\, I(x \in \Gamma^c), \end{aligned} \tag{4.3.1}$$

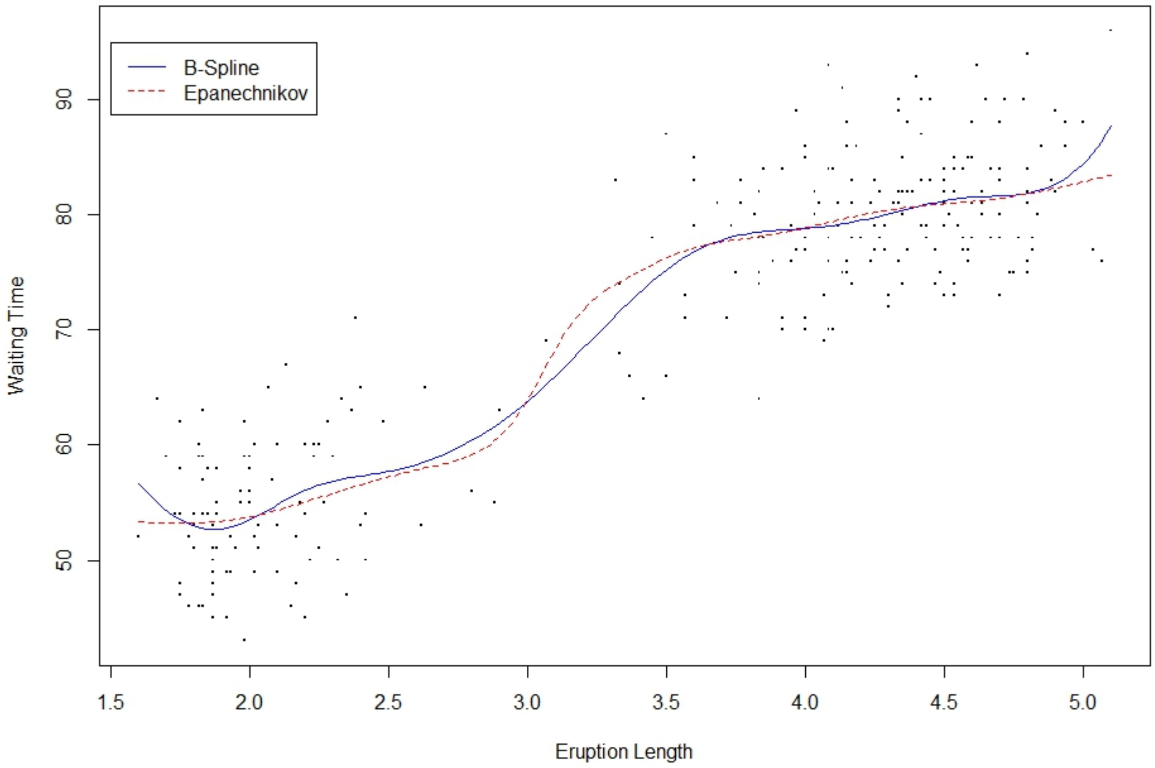


Figure 4.3.1. Waiting time and duration data of Old Faithful eruptions with curves fitted using a cubic b-spline (blue) and a Nadaraya-Watson kernel estimator using the Epanechnikov kernel (red).

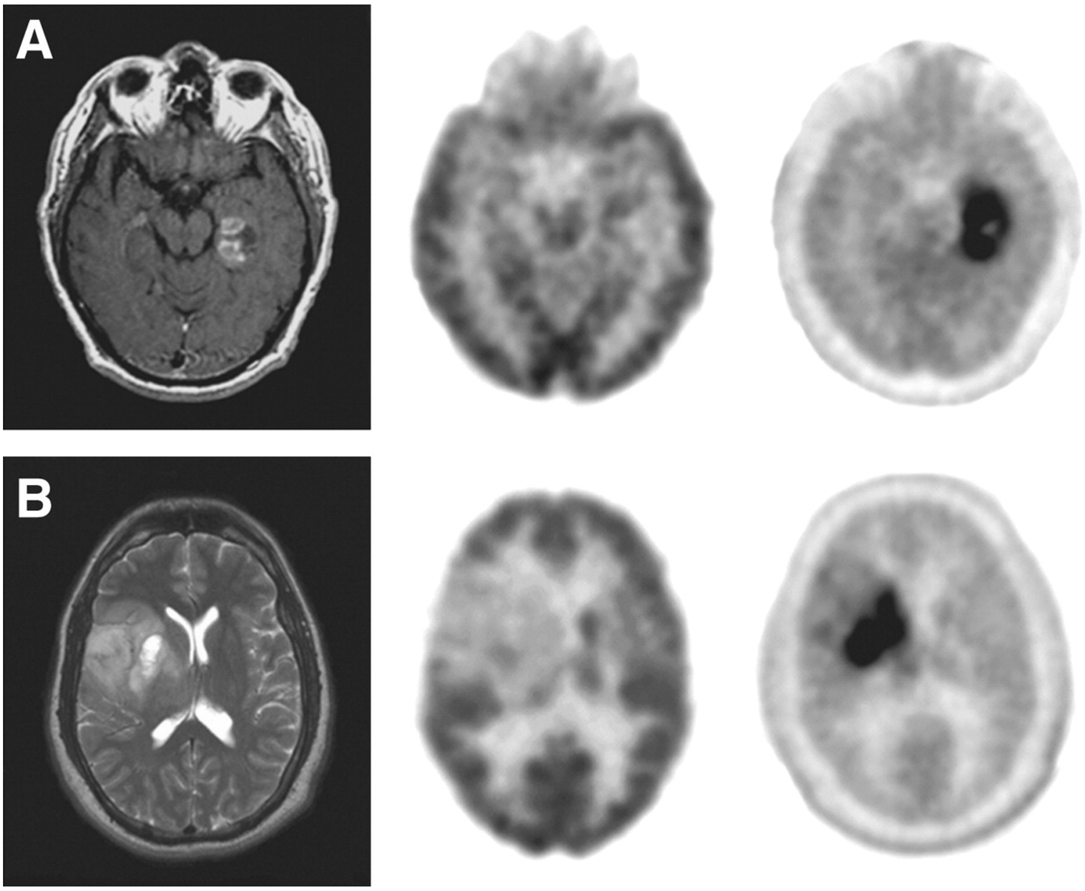


Figure 4.3.2. MRI (left), F-FDG PET (middle), and F-FDOPA PET (right) of glioblastoma (A) and grade II oligodendroglioma (B).

where g is a density on Ω, f_Γ and f_{Γ^c} are densities on the intensity space, and $I(\cdot)$ denotes an indicator function. That is, given the pixel location $X = x$, the distribution of the pixel intensity Y depends only on whether x is in Γ or Γ^c. Assume that there is a true, star-shaped region, denoted by $\Gamma^\star$, with a known reference point in its interior. Any point in $\Gamma^\star$ can be connected to the reference point by a line segment fully contained in $\Gamma^\star$. The observations $\{(X_i, Y_i) : i = 1, \ldots, n\}$ are i.i.d. samples from $P_{\Gamma^\star}$, and the goal is to make inference on $\Gamma^\star$ or, equivalently, its boundary $\gamma^\star := \partial\Gamma^\star$. The set $\Gamma^\star$ is assumed to have a smooth boundary, as described by Assumption 4.3.1. The ability of a statistical method to make inference on the image boundary will depend on how smooth the true boundary is. In (61), $\gamma^\star$ is interpreted as a function from the unit circle to the positive real line, and the authors formulate a Hölder smoothness condition for this function. Here, the boundary $\gamma^\star$ is treated as a function from the interval $[0, 2\pi]$ to the positive reals, and the smoothness condition is formulated on this arguably simpler version of the function. Since the reparametrization of the unit circle in terms of polar coordinates is smooth, it is easy to check that the Hölder smoothness condition in Assumption 4.3.1 is equivalent to that in (61).

Assumption 4.3.1 *The true boundary function* $\gamma^\star : [0, 2\pi] \to \mathbb{R}^+$ *is* α*-Hölder smooth, i.e., there exists a constant* $L = L_{\gamma^\star} > 0$ *such that*

$$|(\gamma^\star)^{([\alpha])}(\theta) - (\gamma^\star)^{([\alpha])}(\theta')| \leq L|\theta - \theta'|^{\alpha - [\alpha]}, \quad \forall\, \theta, \theta' \in [0, 2\pi], \tag{4.3.2}$$

where $(\gamma^\star)^{(k)}$ *denotes the* k^{th} *derivative of* $\gamma^\star$ *and* $[\alpha]$ *denotes the largest integer less than or equal to* α. *Denote this set of* α*-Hölder functions by* $\mathcal{H}(\alpha)$. *Following the description of* $\Gamma^\star$ *above, it is also assumed that the reference point is strictly interior to* $\Gamma^\star$ *meaning that it is contained in an open set itself wholly contained in* $\Gamma^\star$ *so that* $\gamma^\star$ *is uniformly bounded away from zero. Moreover, the density* g *for* X*, as in Equation 4.3.1, is uniformly bounded above by* $\overline{g} := \sup_{x\in\Omega} g(x)$ *and below by* $\underline{g} := \inf_{x\in\Omega} g(x) \in (0,1)$ *on* Ω.

This section concerns computation of the convergence rate for a Gibbs posterior for $\Gamma^\star$. Models for the image boundary including the specific linking functions used along with several numerical examples will be presented in Chapter 6. The optimal rate of convergence for estimators of $\gamma^\star$ is $n^{-\alpha/(\alpha+1)}$ where $\alpha \geq 1$ summarizes the smoothness of $\gamma^\star$ as in Assumption 4.3.1; see (68). Denote an estimator of $\Gamma^\star$ by $\hat{\Gamma}$ and let $d(\hat{\Gamma}, \Gamma^\star)$ be a distance function. An estimator is said to be minimax with respect to $d(\cdot,\cdot)$ if the supremum of the expected distance over all $\Gamma^\star$ with $\partial\Gamma^\star \in \mathcal{H}(\alpha)$ between the estimated and true boundary functions is minimal over all estimators. In (68) the asymptotic minimax risk is shown to decay at rate $n^{-\alpha/(\alpha+1)}$,

$$\liminf_{n} \inf_{\hat{\Gamma}} \sup_{\partial\Gamma^\star\in\mathcal{H}(\alpha)} n^{\alpha/(\alpha+1)} E_P[d(\hat{\Gamma}, \Gamma^\star)] > 0.$$

Moreover, (68) shows existence of an estimator $\hat{\Gamma}$ obtaining this rate, i.e.

$$\sup_{\partial\Gamma^\star\in\mathcal{H}(\alpha)} E_P[d(\hat{\Gamma}, \Gamma^\star)] \lesssim n^{-\alpha/(\alpha+1)}.$$

The distance most often used to compute distance between sets in a subset of $\mathbb{R}^2$ is the Lebesgue measure of the symmetric difference $d(\Gamma, \Gamma^\star) = \lambda(\Gamma \Delta \Gamma^\star)$ where

$$\lambda(\Gamma \triangle \Gamma^\star) := \int_{\{t \in \Omega : (t \in \Gamma \cap t \notin \Gamma^\star) \bigcup (t \notin \Gamma \cap t \in \Gamma^\star)\}} dt.$$

The rest of this section presents sufficient conditions for the Gibbs posterior to concentrate on neighborhoods of $\Gamma^\star$ at nearly the minimax rate. The linking function $\ell(\mathcal{X}, \gamma)$, and, in particular, the difference $\ell(\mathcal{X}, \gamma) - \ell(\mathcal{X}, \gamma^\star)$ must satisfy certain properties laid out in 4.3.2. The first property says the exponential of this linking function difference can be bounded in terms of the L_1 difference $||\gamma - \gamma^\star||_1 = \int_0^{2\pi} |\gamma(x) - \gamma^\star(x)| g(x) dx$. The second property says that a sup-norm ball around $\gamma^\star$, denoted $B_\infty(\gamma^\star; r) := \{\gamma : \sup_{x \in [0,2\pi]} |\gamma(x) - \gamma^\star(x)| \leq r\}$, is contained in a neighborhood of $\gamma^\star$ characterized by the linking function differences $\ell(\mathcal{X}, \gamma) - \ell(\mathcal{X}, \gamma^\star)$. As in the proof of Theorem 4.2.1, these linking function differences are important in bounding posterior probabilities.

Assumption 4.3.2 *The linking function* $\ell(\mathcal{X}, \gamma)$ *satisfies*

$$0 < E(\exp(\ell(\mathcal{X}, \gamma) - \ell(\mathcal{X}, \gamma^\star))) < 1 - \rho ||\gamma - \gamma^\star||_1 < 1, \text{ and} \tag{4.3.3}$$

$$\{\theta : \max [R(\gamma) - R(\gamma^\star), V(\ell(\mathcal{X}, \gamma) - \ell(\mathcal{X}, \gamma^\star))] \geq C\delta\} \supseteq B_\infty(\gamma^\star; C_0\delta) \tag{4.3.4}$$

for some $\rho \in (0, 1)$*, some constants* $C, C_0, \delta > 0$.

Further, γ is parameterized as a linear combination of basis functions $\hat{\gamma}_{D,\beta}(x) = \sum_{j=1}^{D} \beta_j \phi_{j,D}(x)$ for a set of basis functions $\phi_{1,D}, ..., \phi_{D,D}$ which meet a certain approximation condition; see Assumption 4.3.3. General results are available on the error in approximating an α-Hölder smooth function by basis functions. For example, Theorem 6.10 in (82) implies that if $\gamma^\star$ satisfies Assumption 4.3.1, then the following approximation holds for b-splines.

Assumption 4.3.3 $\forall\ d > 0$, $\exists\ \beta_d^\star \in (\mathbb{R}^+)^d$ *such that* $\|\theta^\star - \hat{\theta}_{d,\beta_d^\star}\|_\infty \lesssim d^{-\alpha}$.

Similar approximations hold for other function bases, and all provide bounds on the uniform convergence of truncated basis function expansions to the true function in terms of the smoothness of $\gamma^\star$. Note that since $\gamma^\star(x) > 0$, one can consider all coefficients to be positive; i.e. $\beta_d^\star \in (\mathbb{R}^+)^d$, and see Lemma 1(b) in (85).

Using the basis function expansion as a parametrization of $\gamma^\star$, the prior distribution can be specified on the coefficients, $\beta = (\beta_1, ..., \beta_D)$, and number of basis functions, D. In problems with a finite parameter dimension, it is typically straightforward to ensure the prior puts non-negligible mass on neighborhoods of the true parameter, but this issue is more complicated for infinite dimensional parameters. The prior must meet several conditions, outlined in Assumption 4.3.4 below, in order to guarantee the prior is both sufficiently spread out yet puts enough mass near values of β and D yielding good approximations to $\gamma^\star$.

Assumption 4.3.4 *Let $\beta^\star_d$, for $d > 0$, be as in Assumption 4.3.3. Then there exists $C, m > 0$ such that the prior Π for (D, β) satisfies, for all $d > 0$,*

$$
\begin{aligned}
\log \Pi(D > d) &\lesssim -d \log d, \\
\log \Pi(D = d) &\gtrsim -d \log d, \\
\log \Pi(\|\beta - \beta^\star_d\|_1 \le kd^{-\alpha} \mid D = d) &\gtrsim -d \log\{1/(kd^{-\alpha})\}, \\
\log \Pi(\beta \notin [-m, m]^d \mid D = d) &\lesssim \log d - Cm.
\end{aligned}
$$

There are simple prior distributions meeting the conditions of Assumption 4.3.4, such as Poisson and exponential prior distributions; see Section 6.3.2.

The following lemma is a summary of various results derived in (61) towards proving their Theorem 3.3, which yields a convergence rate for a Bayesian posterior in a similar statistical problem. This result shows that priors meeting the conditions specified in 4.3.4 have two important properties: first, they place significant prior mass on sets of boundary functions that very closely approximate $\gamma^\star$, and second, they place almost all of their probability on a nice set Σ_n called a sieve. This sieve is a subset of the parameter space and its complexity grows with sample size n. As defined in Lemma 4.3.1, this sieve contains "nice" boundary functions, whose number of basis functions and maximum coefficient do not grow too quickly.

Lemma 4.3.1 *Let ϵ_n be as in Theorem 4.3.1 and let $D_n = (\frac{n}{\log n})^{\frac{1}{\alpha+1}}$. Then, $\|\theta^\star - \hat{\theta}_{D_n,\beta^\star}\|_\infty \le C\epsilon_n$ for some $C > 0$, $\beta^\star = \beta^\star_{D_n} \in (\mathbb{R}^+)^{D_n}$ from Assumption 4.3.3.*

1. *Define the neighborhood* $B_n^\star = \{(\beta, d) : \beta \in \mathbb{R}^d, d = D_n, \|\theta^\star - \hat{\theta}_{d,\beta}\|_\infty \le C\epsilon_n\}$. *Then* $\Pi(B_n^\star) \gtrsim \exp(-an\epsilon_n)$ *for some* $a > 0$ *depending on* C.

2. *Define the sieve* $\Sigma_n = \{\theta : \theta = \hat{\theta}_{d,\beta}, \beta \in \mathbb{R}^d, d \le D_n, \|\beta\|_\infty \le \sqrt{n/K_0}\}$. *Then* $\Pi(\Sigma_n^c) \lesssim \exp(-Kn\epsilon_n)$ *for some* K, $K_0 > 0$.

3. *The bracketing number of* Σ_n *satisfies* $\log N(\epsilon_n, \Sigma_n, \|\cdot\|_\infty) \lesssim n\epsilon_n$.

Theorem 4.3.1 states that a Gibbs posterior with linking function and prior satisfying Assumptions 4.3.2 and 4.3.4 concentrates on neighborhoods of $\Gamma^\star$ defined by the Lebesgue measure of the symmetric difference at nearly the minimax rate (up to a logarithmic factor). The same rate is given for a Bayesian posterior in (61).

Theorem 4.3.1 *Under Assumptions 4.3.1–4.3.4, for any positive sequence* $M_n \to \infty$

$$E_P\left[\Pi_n(\{\Gamma : \lambda(\Gamma\Delta\Gamma^\star) > M_n\epsilon_n\})\right] \to 0 \quad \textit{as } n \to \infty,$$

where $\epsilon_n = \{(\log n)/n\}^{\alpha/(\alpha+1)}$, *and* α *is the smoothness coefficient in Assumption 4.3.1.*

Proof of Theorem 4.3.1

Define the set

$$A_n = \{\gamma : \|\gamma^\star - \gamma\|_1 > M_0 M_n \epsilon_n\} \tag{4.3.5}$$

for some $M_0 > 0$ to be specified. For the sieve Σ_n in Lemma 4.3.1, part 2, see that $\Pi_n(A_n) \le \Pi_n(\Sigma_n^c) + \Pi_n(A_n \cap \Sigma_n)$. The goal is to show to show that both terms in the upper bound vanish, in $L_1(P)$, as $n \to \infty$.

It is helpful to start with a lower bound on $I_n = \int e^{-n\{R_n(\gamma)-R_n(\gamma^\star)\}}\,\Pi(d\gamma)$, the denominator in both of the terms discussed above. First, write

$$I_n \geq \int_{S_n} e^{-n\{R_n(\gamma)-R_n(\gamma^\star)\}}\,\Pi(d\gamma)$$

where S_n is defined in Lemma 4.2.2 with $t_n = \epsilon_n$ and $C > 0$ as in Assumption 4.3.2. From Lemma 4.2.2, see that $I_n \gtrsim \Pi(S_n)e^{-2C\epsilon_n}$, with P probability converging to 1 and by Assumption 4.3.2 $S_n \supseteq B_\infty(\gamma^\star; C_0\epsilon_n)$, so it follows from Lemma 4.3.1, part 1,

$$I_n \gtrsim \Pi\{B_\infty(\gamma^\star; C_0\epsilon_n)\}e^{-2C\epsilon_n} \gtrsim e^{-C_1 n\epsilon_n},$$

with P probability converging to one, and where $C_1 > 0$ is a constant depending on C_0 and C.

The next step is to bound $\Pi_n(\Sigma_n^c)$. Write this quantity as

$$\Pi_n(\Sigma_n^c) = \frac{N_n(\Sigma_n^c)}{I_n} = \frac{1}{I_n}\int_{\Sigma_n^c} e^{-n\{R_n(\gamma)-R_n(\gamma^\star\}}\,\Pi(d\gamma).$$

It will suffice to bound the expectation of $N_n(\Sigma_n^c)$. By Tonelli's theorem, independence, and Assumption 4.3.2, see that

$$\begin{aligned}
E_P[N_n(\Sigma_n^c)] &= \int_{\Sigma_n^c} E_P\left[e^{-n\{R_n(\gamma)-R_n(\gamma^\star)\}}\right] \Pi(d\gamma) \\
&= \int_{\Sigma_n^c} \left\{E_P\left[e^{-(\ell(\mathcal{X},\gamma)-\ell(\mathcal{X},\gamma^\star))}\right]\right\}^n \Pi(d\gamma) \\
&\leq \int_{\Sigma_n^c} \{1-\rho\|\gamma-\gamma^\star)\|_1\}^n \Pi(d\gamma) \\
&\leq \Pi(\Sigma_n^c).
\end{aligned}$$

By Lemma 4.3.1, part 2, see that $\Pi(\Sigma_n^c) \leq e^{-Kn\epsilon_n}$.

Next, bound $\Pi_n(A_n \cap \Sigma_n)$. Again, it will suffice to bound the expectation of $N_n(A_n \cap \Sigma_n)$. Choose a covering $A_n \cap \Sigma_n$ by sup-norm balls $B_j = B_\infty(\gamma_j; \omega M_0 M_n \epsilon_n)$, $j = 1, \ldots, J_n$, with centers γ_j in A_n and radii $\omega M_0 M_n \epsilon_n$, where $\omega \in (0, \frac{1}{gB})$. Also, from Lemma 4.3.1, part 3, see that J_n is bounded by $e^{K_1 n \epsilon_n}$ for some constant $K_1 > 0$. For this covering, immediately get

$$E_P[N_n(A_n \cap \Sigma_n)] \leq \sum_{j=1}^{J_n} E_P[N_n(B_j)].$$

For each j, using Tonelli, independence, and Assumption 4.3.2 again, write

$$E_P[N_n(B_j)] = \int_{B_j} \left\{E_P\left[e^{-(\ell(\mathcal{X},\gamma)-\ell(\mathcal{X},\gamma^\star))}\right]\right\}^n \Pi(d\gamma) \leq \int_{B_j} \{1-\rho\|\gamma-\gamma^\star\|_1\}^n \Pi(d\gamma).$$

For γ in B_j, since the center γ_j is in A_n, it follows that $\|\gamma - \gamma^\star\|_1$ is lower bounded by $\eta = M_0 M_n \epsilon_n(1 - \omega\bar{g}B)$. Therefore, from the bound on J_n,

$$E_P[N_n(A_n \cap \Sigma_n)] \leq \sum_{j=1}^{J_n} E_P[N_n(B_j)] \leq e^{-nhM_0M_n\epsilon_n} J_n \leq e^{-(\eta M_0 M_n - K_1)n\epsilon_n}.$$

Finally, it follows that

$$\begin{aligned}
\Pi_n(A_n) &\leq \Pi_n(A_n \cap \Sigma_n) + \Pi_n(\Sigma_n^c) \\
&= \frac{N_n(A_n \cap \Sigma_n)}{I_n} + \frac{N_n(\Sigma_n^c)}{I_n} \\
&\leq \frac{N_n(A_n \cap \Sigma_n)}{e^{-C_1 n\epsilon_n}} 1(I_n > e^{-C_1 n\epsilon_n}) + \frac{N_n(\Sigma_n^c)}{I_n} 1(I_n \leq e^{-C_1 n\epsilon_n}) \\
&\leq \frac{N_n(A_n \cap \Sigma_n)}{e^{-C_1 n\epsilon_n}} + 1(I_n \leq e^{-C_1 n\epsilon_n}).
\end{aligned}$$

Taking P-expectation and plugging in the bounds derived above, see that

$$E_P[\Pi_n(A_n)] \leq e^{-(\eta M_0 M_n - K_1 - C_1)n\epsilon_n}$$

and since $M_n \to \infty$, for large enough n, $M_0 M_n > (K_1 + C_1)/\eta$, so the upper bound vanishes as $n \to \infty$.

The last step is to show that the set A_n contains the set $\{\Gamma : \lambda(\Gamma \triangle \Gamma^\star) > M_n \epsilon_n\}$ for some choice of M_0. A simple conversion to polar coordinates gives

$$\begin{aligned}
\lambda(\Gamma \triangle \Gamma^\star) &= \int_{\Gamma \triangle \Gamma^\star} d\lambda \\
&= \int_0^{2\pi} \int_{\gamma(\theta) \wedge \gamma^\star(\theta)}^{\gamma(\theta) \vee \gamma^\star(\theta)} r \, dr \, d\theta \\
&= \frac{1}{2} \int_0^{2\pi} \{\gamma(\theta) \wedge \gamma(\theta^\star)\}^2 - \{\gamma(\theta) \vee \gamma(\theta^\star)\}^2 \, d\theta \\
&= \frac{1}{2} \int_0^{2\pi} |\gamma(\theta) - \gamma^\star(\theta)| \, |\gamma(\theta) + \gamma^\star(\theta)| \, d\theta.
\end{aligned}$$

Let $\underline{\gamma}^\star = \inf_\theta \gamma^\star(\theta)$, then it is easy to verify that

$$\underline{\gamma}^\star \leq |\gamma(\theta) + \gamma^\star(\theta)| \leq \operatorname{diam}(\Omega), \quad \forall \, \theta \in [0, 2\pi].$$

Therefore,

$$\tfrac{1}{2}\underline{\gamma}^\star \|\gamma - \gamma^\star\|_1 \leq \lambda(\Gamma \triangle \Gamma^\star) \leq \tfrac{1}{2} \operatorname{diam}(\Omega) \|\gamma - \gamma^\star\|_1. \tag{4.3.6}$$

Hence, if $M_0 > (\frac{1}{2}\underline{\gamma}^\star)^{-1}$, then $A_n \supset \{\Gamma : \lambda(\Gamma \Delta \Gamma^\star) > M_n \epsilon_n\}$, which implies

$$E_P \left[\Pi_n(\{\Gamma : \lambda(\Gamma \Delta \Gamma^\star) > M_n \epsilon_n\}) \right] \to 0$$

as $n \to \infty$, as was to be shown.

4.4 Asymptotic normality of the Gibbs posterior

In (14), the authors present a series of asymptotic results on Gibbs posteriors for a vector parameter. They provide conditions for the Gibbs posterior to converge at rate $n^{-1/2}$, determine when credible intervals are asymptotically calibrated, and show that the Gibbs posterior distribution converges to a normal distribution under certain conditions. They make use of the following assumptions.

Assumption 4.4.1 *The true parameter $\theta^\star$ belongs to the interior of a compact convex subset of d-dimensional Euclidean space.*

Assumption 4.4.2 *For any $\delta > 0$, there exists $\epsilon > 0$, such that*

$$P\{\sup_{|\theta-\theta^\star|\geq\delta} \frac{1}{n}(R_n(\theta^\star) - R_n(\theta)) > -\epsilon\} \overset{\text{i.p.}}{\to} 0 \; in \; P - probability.$$

Assumption 4.4.2 establishes uniform control on the numerator of the posterior probability away from $\theta^\star$. To see this, write the posterior probability of the set $A^c := \{\theta : |\theta - \theta^\star| \geq \delta\}$ as

$$\Pi_n(A^c) = \frac{N_n(A^c)}{D_n} = \frac{\int_{A^c} \exp(-n\omega R_n(\theta))d\Pi(\theta)}{\int_\Theta \exp(-n\omega R_n(\theta))d\Pi(\theta)},$$

and multiply and divide by $\exp(n\omega R_n(\theta^\star))$ to get

$$\frac{N_n(A^c)}{D_n} = \frac{\int_{A^c} \exp(-n\omega[R_n(\theta) - R_n(\theta^\star)])d\Pi(\theta)}{\int_\Theta \exp(-n\omega[R_n(\theta) - R_n(\theta^\star)])d\Pi(\theta)}.$$

Then, using Assumption 4.4.2, the numerator can be bounded above by

$$N_n(A^c) \leq e^{-\omega n^2 \epsilon}.$$

This uniform convergence is a strong assumption and is not always simple to verify.

Assumption 4.4.3 *For* θ *in an open neighborhood of* $\theta^\star$,

(i) $R_n(\theta) - R_n(\theta^\star) = (\theta - \theta^\star)\Delta_n(\theta^\star) + \frac{1}{2}(\theta - \theta^\star)' J_n(\theta^\star)(\theta - \theta^\star) + M_n(\theta)$,

(ii) $\Omega_n^{-1/2}(\theta^\star)\Delta_n(\theta^\star)/\sqrt{n} \xrightarrow{d} \mathcal{N}(0, I)$,

(iii) $J_n(\theta^\star) = O(1)$ *and* $\Omega_n(\theta^\star) = O(1)$ *are uniformly in* n *positive-definite constant matrices,*

(iv) for each $\epsilon > 0$ *there is a sufficiently small* $\delta > 0$ *and large* $C > 0$ *such that*

(a) $\limsup_{n\to\infty} P\{\sup_{C/\sqrt{n} \leq |\theta-\theta^\star| \leq \delta} \frac{|M_n(\theta)|}{n(\theta-\theta^\star)^2} > \epsilon\} < \epsilon$,

(b) $\limsup_{n\to\infty} P\{\sup_{|\theta-\theta^\star| \leq C/\sqrt{n}} |M_n(\theta)| > \epsilon\} = 0$.

Assumption 4.4.3 imposes rather strict requirements on the empirical risk function. Part (i) essentially requires a second-order Taylor expansion where Δ_n acts like a gradient, J_n acts like a second derivative matrix, and M_n is a remainder term. Part (iv) says that the remainder term $M_n(\theta)$ vanishes in $n^{-1/2}$-sized neighborhoods of $\theta^\star$ and is bounded above by approximately $n|\theta - \theta^\star|^2$ outside of these neighborhoods. This very simple example of inference on a population mean illustrates Assumption 4.4.3.

Suppose a data analyst can sample i.i.d. data $\mathcal{X}^n$ from a distribution P with mean $\theta^\star \in \mathbb{R}$ fulfilling Assumption 4.4.1 above. One possible choice of empirical risk function for estimating the mean is squared error loss, i.e. $R_n(\theta) = \frac{1}{n}\sum_{i=1}^n (\mathcal{X}_i - \theta)^2$. The empirical risk difference can be written

$$R_n(\theta) - R_n(\theta^\star) = \frac{1}{n}\sum_{i=1}^n 2\mathcal{X}_i(\theta^\star - \theta) + \theta^2 - (\theta^\star)^2.$$

Take $\Delta_n(\theta) = \frac{1}{n}\sum_{i=1}^n 2(\theta - \mathcal{X}_i)$ and $J_n(\theta) = 2$. Then,

$$R_n(\theta) - R_n(\theta^\star) = (\theta - \theta^\star)\Delta_n(\theta^\star) + \frac{1}{2}(\theta - \theta^\star)' J_n(\theta^\star)(\theta - \theta^\star)$$

exactly with remainder 0. Further, $\Delta_n(\theta^\star) = 2(\theta^\star - \frac{1}{n}\sum_{i=1}^n \mathcal{X}_i)$ so $\Omega_n^{-1/2}(\theta^\star)\Delta_n(\theta^\star)/\sqrt{n}$ is approximately standard normal for $\Omega_n(\theta^\star) = 4$.

Given the above assumptions, define the parameter $h := \sqrt{n}(\theta - \theta^\star) - J_n(\theta^\star)^{-1}\Delta_n(\theta^\star)/\sqrt{n}$ and the corresponding Gibbs posterior on the h space, $\Pi_{n,h}$. Proposition 4.4.1 below says that under Assumptions 4.4.1-4.4.3, the Gibbs posterior concentrates on a $n^{-1/2}$ neighborhood of $\theta^\star$ with size measured by the total variation norm. That is, the convergence rate of the Gibbs posterior is $n^{-1/2}$ in this setting. Further, when n is large, the Gibbs posterior is approximately normal with mean $\theta^\star + J_n(\theta^\star)^{-1}\Delta_n(\theta^\star)/n$ and covariance matrix $J_n(\theta^\star)^{-1}/n$.

Proposition 4.4.1 *Under Assumptions 4.4.1-4.4.3, and for any* $0 \le \alpha < \infty$,

$$||\Pi_{n,h}(h) - \Pi_{\infty,h}(h)||_{TVM(\alpha)} := \int_{H_n} |\Pi_{n,h}(h) - \Pi_{\infty,h}(h)| dh \overset{i.p.}{\to} 0,$$

where $H_n = \{\sqrt{n}(\theta - \theta^\star) - J_n(\theta^\star)^{-1}\Delta_n(\theta^\star)/\sqrt{n} : \theta \in \Theta\}$ *and*

$$\Pi_{\infty,h}(h) := \sqrt{\frac{\det J_n(\theta^\star)}{(2\pi)^d}}\exp(-\frac{1}{2}h' J_n(\theta^\star)h).$$

The authors of (14) highlight several problems meeting their assumptions including regression models, generalized method of moments models, quasi-likelihood models, and some time-series models. However, as demonstrated in Chapters 5 and 6, empirical risk functions failing Assumption 3 are common, such as those based upon misclassification error loss functions. In those cases the results of Sections 4.2.1 4.3.2 and provided by this dissertation are applicable.

CHAPTER 5

APPLICATION OF GIBBS MODELS TO THE MINIMUM CLINICALLY IMPORTANT DIFFERENCE

Portions of this chapter are reprinted from Journal of Statistical Planning and Inference, Vol 187, Syring, N. A. and Martin, R., "Gibbs posterior inference on the minimum clinically important difference", 67-77, Copyright (2017). See the statement of permission by Elsevier in the appendix.

5.1 Introduction

In clinical trials, often the main objective is assessing the efficacy of a treatment. However, experts have observed that statistical significance alone does not necessarily imply efficacy (42). For instance, a study with high power can detect statistically significant differences, but these may not translate to practical differences noticeable by the patients. As a result, a cutoff value different than a statistical critical value is desired that can separate patients with and without clinically significant responses. This cutoff is called the *minimum clinically important difference*, or *MCID* for short (43). Accurate inference on the MCID is crucial for clinicians and health policy-makers to make educated judgments about the effectiveness of certain treatments. Indeed, the U. S. Food and Drug Administration held a special workshop in 2012 on methodological developments towards improved inference on the MCID; see $\texttt{https://federalregister.gov/a/2012-27147}$.

The basic setup is that, in addition to a scalar diagnostic measure for each patient, which would be used to assess the statistical significance of a treatment, one also has access to a "patient-reported outcome," a binary indicator of whether or not the patient felt that the treatment was beneficial. Then, roughly, the MCID is defined as the cutoff value such that, if the diagnostic measure exceeds this cutoff, then the patient is likely to observe a benefit from the treatment. A more precise description of the problem setup is given in Section 4.2. The challenge in making inference on the MCID is in modeling the joint distribution for the diagnostic measure and patient-reported outcome. Given a model, standard likelihood-based methods—Bayesian or non-Bayesian—could be used, but specifying a sound model is difficult because the MCID is a rather complicated functional thereof. To avoid the potential bias caused by a misspecified parametric model and the inefficiencies that result from an overly-complex nonparametric model, a model-free approach is an attractive alternative. Recently, (37) propose a M-estimation framework for estimating the MCID, that does not require a model, but the distribution theory needed to provide valid tests or confidence intervals for the MCID based on their approach is apparently out of reach.

This chapter discusses how a Gibbs posterior can provide inference on the MCID without requiring a likelihood, thus avoiding the modeling step and the risk of misspecification while providing easy access to credible intervals, and shows that this new method compares favorably to the existing M-estimation method in terms of both large-sample theory and finite-sample performance. Construction of the Gibbs posterior takes advantage of the representation in (37) of the MCID as the minimizer of an expected loss. The given Gibbs posterior distribution is easy to compute and, with a suitable scaling, is shown to provide valid and efficient credible intervals for the MCID.

The focus in this paper is the MCID application, but some general comments about Gibbs posteriors are worth mentioning. First, the problem is related to that of model misspecification, and it is known ((9), (58), (101), (51), (17), and (77)) that, asymptotically, the posterior distribution behaves reasonably under misspecification provided that it is Gibbs-like in the sense that the negative log-likelihood used resembles a suitable loss function; a nice example of this type is (91). Second, although misspecification is usually viewed as a bad thing, there might be reasons to "misspecify on purpose." For example, one may not wish to spend the resources needed to flesh out a full model, including priors, and to compute the full posterior when, ultimately, it will be marginalized to the parameter of interest. The Gibbs posterior described here has the advantage of being defined directly on the parameter of interest, simplifying both prior specifications and posterior computations.

The remainder of this chapter is organized as follows. Section 5.2 introduces notation for the MCID problem and formulates its definition as a minimizer of an expected loss. This leads naturally to the M-estimator proposed in (37) and this section highlights two important improvements on their convergence rate result: first, the rate is improved and, second, the relationship between the rate and the local properties of the function in Equation 5.2.3 have been substantially clarified. In Section 5.3, after a motivating illustration, the Gibbs posterior distribution for the MCID is defined, and it is shown that the posterior, and the corresponding posterior mean, converge at the same rate as the M-estimator of (37). Simulation results are presented in Section 5.4, and the take away message is that the Gibbs posterior here presented, or a suitably scaled version thereof, provides quality inference on MCID, in terms of estimation accuracy and interval estimate coverage and length. Some concluding remarks are given in Section 5.5, and technical details are given in Section 5.6.

5.2 Minimum clinically important difference

5.2.1 Notation and definitions

In clinical trials for drugs or medical devices, it is standard to judge the effectiveness of the treatment based on statistical significance. However, it is possible that the treatment effect may be significantly different from zero in a statistical context, but the effect size is so small that the patients do not experience an improvement. To avoid the costs associated with bringing to market a treatment that is not clinically effective, it is advantageous to bring the patients' assessment of the treatment effect into the analysis. While the need for a measure of clinical significance is well-documented (48), it seems there is no universal definition of MCID and, consequently, there is no standard methodology to make inference on it. Recent efforts in this direction were made by (87) and (98). (37) provide a mathematically convenient formulation, described next, in which the MCID is expressed as a minimizer of a suitable loss function.

Let $\mathsf{Y} \in \{-1, 1\}$ denote the patient reported outcome with "$\mathsf{Y} = 1$" meaning that the treatment was effective and "$\mathsf{Y} = -1$" meaning that the treatment was not effective. Let X be a continuous diagnostic measure taken on each patient. Let P denote the joint distribution of

$\mathcal{X} = (X, Y)$, and p the marginal density of X with respect to Lebesgue measure. Given $\theta \in \mathbb{R}$, define the function $l(\mathcal{X}, \theta)$ by

$$\ell(\mathcal{X}, \theta) = \frac{1}{2}\{1 - Y \operatorname{sign}(X - \theta)\}, \quad (X, Y) \in \mathbb{R} \times \{-1, 1\}, \tag{5.2.1}$$

where $\operatorname{sign}(x) = 1$ is $x \geq 0$ and -1 otherwise, and write $R(\theta) = E_P[\ell(\mathcal{X}, \theta)]$ for the risk function, the expectation of $\ell(\mathcal{X}, \theta)$ with respect to the joint distribution P. Then the MCID, denoted by $\theta^\star$, is defined as

$$\theta^\star = \arg\min_\theta R(\theta). \tag{5.2.2}$$

That is, the MCID is the minimizer of the risk function R, and depends on the distribution P in a rather complicated way.

The intuition behind this definition is the alternative expression for $R(\theta)$:

$$R(\theta) = P\{Y \neq \operatorname{sign}(X - \theta)\},$$

i.e., $\theta^\star$ minimizes, over θ, the probability that $\operatorname{sign}(X - \theta)$ disagrees with Y. In other words, $\operatorname{sign}(X - \theta^\star)$ is the best predictor of Y in terms of minimum misclassification probability. Another representation of the MCID, as demonstrated by (37), that will be convenient below is as a solution to the equation $\eta(\theta) = \frac{1}{2}$, where

$$\eta(x) = P(Y = 1 \mid X = x) \tag{5.2.3}$$

is the conditional probability function. If η is continuous and strictly increasing, then $\theta^\star$ will be the unique solution to the equation $\eta(\theta) = \frac{1}{2}$. If η is only upper semi-continuous, then define $\theta^\star$ as $\inf\{x : \eta(x) \geq \frac{1}{2}\}$, and an argument similar to that in Lemma 1 of (37) shows that this $\theta^\star$ solves the optimization problem Equation 5.2.2.

5.2.2 M-estimator and its large-sample properties

(37) propose to estimate the MCID by minimizing an empirical risk. Let $\mathbb{P}_n = n^{-1} \sum_{i=1}^{n} \delta_{(X_i,Y_i)}$ be the empirical measure, based on the observations $\mathcal{X}_i = \{(X_i, Y_i) : i = 1, \ldots, n\}$, where $\delta_{(x,y)}$ is the point-mass measure at (x, y). Then the empirical risk is $R_n(\theta) = \mathbb{P}_n l(\mathcal{X}, \theta)$, and an M-estimator of MCID is obtained by minimizing $R_n(\theta)$, i.e.,

$$\hat{\theta}_n = \arg\min_{\theta} R_n(\theta). \tag{5.2.4}$$

Computation of the estimator is straightforward since it takes only finitely many values depending on the order statistics for the $\mathcal{X}$-sample. Therefore, a simple grid search is guaranteed to quickly identify the minimizer $\hat{\theta}_n$.

A shortcoming of this approach is that, due to the discontinuity of the loss function, an asymptotic normality result for the M-estimator does not seem possible; see Section 5.2.3. Therefore, valid confidence intervals for the MCID based on the M-estimator are not currently available. This provides motivation for a Bayesian approach, where credible intervals, etc, can be easily obtained, but some non-standard ideas are needed to deal with the fact that θ is defined by a loss function, not a likelihood; see Section 5.3. Bootstrap methods are available (see Section 5.4) but there is a general concern about their validity because the rate is not the usual $n^{-1/2}$.

Consistency and convergence rates for the M-estimator $\hat{\theta}_n$ have been studied by (37). The rates rely on the local behavior of the function η and of the marginal distribution of X around $\theta^\star$. Theorem 5.2.1 clarifies and substantially improves upon the rate result given in (37). The assumptions here are more efficient than theirs, and the differences are discussed below.

Assumption 5.2.1 *The marginal density* p *of* X *is continuous and bounded away from 0 and* ∞ *on an interval containing* $\theta^\star$*.*

Assumption 5.2.2 *The function* η *in Equation 5.2.3 is non-decreasing, upper semi-continuous, and satisfies* $\eta(\theta) > \eta(\theta^\star)$*) for all* $\theta > \theta^\star$*. Furthermore, there exists constants* $c > 0$*, and* $\gamma \geq 0$ *such that*

$$\min|\eta(\theta^\star \pm \epsilon) - \eta(\theta^\star)| > c\epsilon^\gamma, \quad \textit{for all small } \epsilon > 0, \tag{5.2.5}$$

where "min" is with respect to the two choices in "±."

The γ parameter may be interpreted as an"ease of identification" index, where smaller γ means that the η function is, in a certain sense, changing more rapidly near $\theta^\star$, making the MCID easier to identify. In particular, if η has a jump discontinuity at $\theta^\star$, then $\gamma = 0$, and this corresponds to the easiest case; if η is differentiable at $\theta^\star$, then $\gamma = 1$, the most difficult case; and if η is continuous but not differentiable at $\theta^\star$, then $\gamma \in (0, 1)$, an intermediate case. For a quick example of the latter case, intermediate ease of identification, fix $\alpha, \beta \in (0, 1)$, $\alpha \geq \beta$, and define $\eta(x)$, $x \in [-1, 1]$ as

$$\eta(x) = \begin{cases} \frac{1}{2}(1 - |x|^{\alpha}), & \text{if } x \in [-1, 0), \\ \\ \frac{1}{2}(1 + x^{\beta}), & \text{if } x \in [0, 1]. \end{cases} .$$

Clearly, the MCID is $\theta^{\star} = 0$, η is continuous but not differentiable there, and Equation 5.2.5 holds with $\gamma = \alpha$. Although this "ease of identification" index is non-standard, it appears to be the key determinant of the convergence rate. Indeed, the convergence rate of the M-estimator in Theorem 5.2.1 below improves as γ decreases to 0, explaining why $\gamma = 0$ and $\gamma = 1$ are the "easiest" and the "most difficult" cases, respectively.

Theorem 5.2.1 *Under Assumptions 5.2.1–5.2.2, the M-estimator $\hat{\theta}_n$ in Equation 5.2.4 satisfies $\hat{\theta}_n - \theta^{\star} = O_P(n^{-r})$ as $n \to \infty$, where $r = (1 + 2\gamma)^{-1}$, and γ is defined in Equation 5.2.5.*

Proof 5.2.1 *See Section 5.6.2.*

The assumptions presented here are different than those in the M-estimator convergence rate theorem of (37), so some comments are in order. In particular, they impose a Hölder continuity condition on η, as well as a "low noise assumption," in Equation (4) in their paper, which upper-bounds the P-probability assigned to events of the form $\{|\eta(X) - \frac{1}{2}| \leq \xi\}$. This implicitly requires that η not be too flat near $\theta^{\star}$, just like the condition (Equation 5.2.5), and together with their Hölder condition, they derive a locally uniform lower bound on the risk difference $R(\theta) - R(\theta^{\star})$, similar to the one derived in Lemma 5.6.1. However, the approach used here combined with more-direct assumptions appear to be more efficient, because the lower bound on $R(\theta) - R(\theta^{\star})$ has been improved and, consequently, a better convergence rate obtained. Indeed, in the case where η is differentiable at $\theta^\star$, the rate is $n^{-1/3}$ whereas (37) obtains $n^{-1/5}$ (up to logarithmic terms). Similarly, for the above example with powers $\alpha \geq \beta$, here the rate is seen to be n^{-r}, with $r = (1+2\alpha)^{-1}$ whereas (37) obtains $n^{-r'}$, with $r' = \{2(1+2\alpha) - \beta/\alpha\}^{-1}$. So, besides showing how the rate depends critically on the "ease of identification" index γ, these examples also highlight the significant improvements in the convergence rate calculations.

5.2.3 On smoothed versions of the problem

It was mentioned above that $R_n(\theta)$ not being smooth causes some problems in terms of limit distribution theory, etc. It would, therefore, be tempting to replace that non-smooth loss function by something smooth, and hope that the approximation error is negligible. One idea would be to introduce a nice parametric model for this problem. For example, consider a binary regression model, where $\eta(x) = F(\beta_0 + \beta_1 x)$ and F is some specified distribution function, such as logistic or normal. Then the MCID corresponds to the median lethal dose (for instance, (50) and (1)). Such a model is smooth so asymptotic normality holds. However, unless the true P has the specified form, there will be non-zero bias that cannot be overcome, even asymptotically; see Section 5.3.1. Since the bias is unknown, sampling distribution concentration around the wrong point cannot be corrected, so is of little practical value.

A slightly less extreme smoothing of the problem is to make a minor adjustment to the original loss function $\ell(\mathcal{X}, \theta)$. As in (37), introduce a smoothing parameter $\tau > 0$ and consider

$$\ell_\theta^\tau(x,y) = \min\bigl\{1, \bigl[1 - \tau^{-1}\, y \operatorname{sign}(x-\theta)\bigr]^+\bigr\},$$

where $u^+ = \max(u, 0)$ denotes the positive part. Write $R^\tau(\theta) = E_P[\ell(\mathcal{X}, \theta)^\tau]$. Based on arguments in (37), it can be shown that $R^\tau(\theta)$ converges uniformly to $R(\theta)$ as $\tau \to 0$, so, for small τ, the minimizer of R^τ would be close to $\theta^\star$. For fixed τ, one can define $R_n^\tau(\theta) = \mathbb{P}_n \ell(\mathcal{X}, \theta)^\tau$ just as before and consider an M-estimator $\hat{\theta}_n^\tau = \arg\min_\theta R_n^\tau(\theta)$. An asymptotic normality result for $\hat{\theta}_n^\tau$ is available, but the proper centering is not at $\theta^\star$ and the asymptotic variance is inversely proportional to τ. So, one could take $\tau = \tau_n$ vanishing with n in an effort to remove the bias, but a price must be paid in terms of the variance. Again, having an asymptotic normality result with either an unknown non-zero bias or a very large variance is of little practical value.

Based on these remarks, apparently there is no hope in trying to smooth out the problem to make it a standard one with the usual asymptotic distribution theory. So, in order to construct useful interval estimates, etc, one needs some different ideas.

5.3 A Gibbs posterior for MCID

5.3.1 Motivation

As discussed above, estimation of the MCID can be achieved without specifying a model, but the distribution theory needed to develop valid interval estimates is lacking. A Bayesian approach automatically provides uncertainty quantification, but it requires a model for the joint distribution P. To motivate this probability model-free Gibbs posterior that follows, the apparent sensitivity of some "standard" Bayesian posterior distributions—parametric and nonparametric—to the underlying P is demonstrated by example. To be clear, the claim is not that Bayesian methods, in general, are inappropriate for this MCID problem, only that the posterior can be particularly sensitive to the choice of model for P so a less-sensitive approach, if one were available, would be attractive.

Suppose the analysis begins with a model for P given by a joint density/mass function $f_\beta(x, y)$, depending on some parameter β, possibly infinite-dimensional, which would typically be different from θ. Given a prior for β, a posterior distribution for β can be readily obtained via Bayes theorem, which can be marginalized to get a posterior distribution for θ. In particular, logistic regression is a sort of black-box approach to study the relationship between a binary response and a quantitative predictor, so consider a Bernoulli model for Y, given $X = x$, where the success probability is $F(\beta_0 + \beta_1 x)$, where F is the standard logistic distribution function. In this case, the MCID is just the median lethal dose, i.e., $\theta = -\beta_0/\beta_1$. The choice of the logit link function F is quite rigid, but a more flexible nonparametric approach is available; see, e.g. (15).

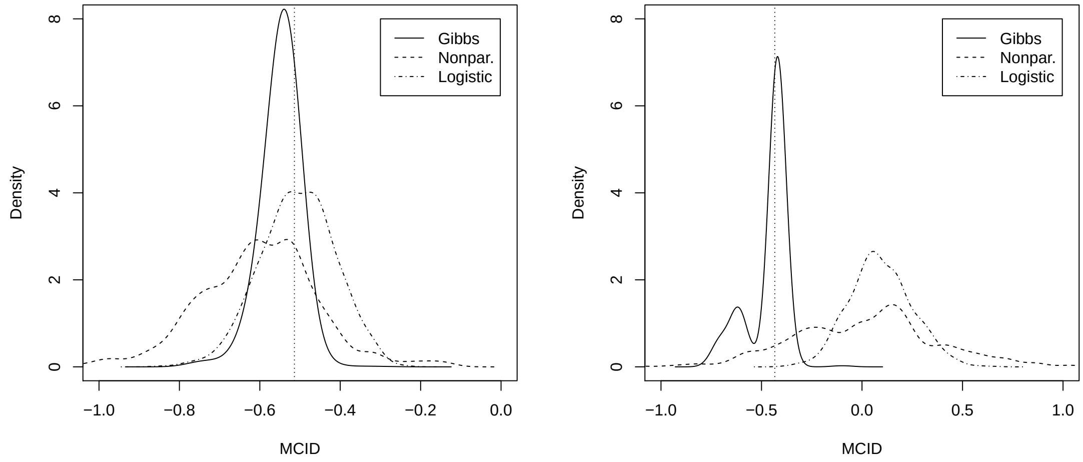
There are pros and cons to both of the approaches just described. Assuming that the logistic regression model is well-specified, inference on the MCID ought to be efficient. However, if the model is misspecified in some way, then there could be non-negligible bias that cannot be overcome, even asymptotically. The model that treats the link function nonparametrically is more flexible and, therefore, is less prone to bias, but at the cost of an increased computational burden and lower efficiency, i.e., the posterior for the MCID is more diffuse. Old-fashioned modeling would be a middle-ground between the extremes of a black-box logistic regression and an overly complex nonparametric regression, but this certainly requires some investment and, unfortunately, is not foolproof. The proposed Gibbs approach is an alternative middle-ground, one that avoids misspecification bias, computational and statistical inefficiency, and modeling investment.

For clarity, here is an illustration of the points just raised. In particular, compare the Gibbs posterior defined in Section 2.2 using the linking function in Equation 5.2.1 to both a standard Bayesian logistic regression and a nonparametric binary regression, (15). For the Bayesian logistic regression consider the vague priors for (β_0, β_1) given in (76), but note that the results do not appear to be sensitive to this choice. Suppose that the true model generating data (X, Y) has a distribution function F for X and, given $X = x$, Y is Bernoulli ± 1 with success probability $F(x)$. Consider two different forms of F, both two-component normal mixtures:

$$X \sim 0.7N(-1,1) + 0.3N(1,1) \quad \text{and} \quad X \sim 0.7N(-1,1) + 0.3N(3,1).$$

The true MCID may be calculated by solving Equation 5.2.3; it is equal to the median of the X distribution, specifically $\theta^\star = -0.514$ in the first example and $\theta^\star = -0.434$ in the second example. Of course, the logistic regression model is misspecified, but the nonparametric model should not be affected by this. But how will they perform in the two examples?

Plots of the marginal posterior density for θ are shown in Figure 5.3.1 for a simulated data set of size $n = 500$ obtained from each of the three methods—Gibbs, logistic regression, and nonparametric—one for each marginal distribution for X. In Panel (a) the posterior distributions for all three models put their mass near the true MCID. However, in Panel (b) the posterior distribution for the Bayesian logistic regression is clearly biased away from the true



(a) $X \sim 0.7\mathsf{N}(-1,1) + 0.3\mathsf{N}(1,1)$

(b) $X \sim 0.7\mathsf{N}(-1,1) + 0.3\mathsf{N}(3,1)$

Figure 5.3.1. Plots of (kernel estimates of) the posterior density for MCID. “Nonpar.” corresponds to the nonparametric binary regression model; “Logistic” corresponds to the posterior based on the genuine Bayes logistic model; “Gibbs” is the proposed likelihood-free Bayesian posterior; true MCID $\theta^\star$ marked with a dotted vertical line.

MCID. The nonparametric Bayesian posterior is very spread out, making it less informative for inference on the MCID. The Gibbs approach, however, is right on the mark in both cases, suggesting that it is neither sensitive to model misspecification nor does it suffer from the inefficiency of the nonparametric approach.

5.3.2 Posterior convergence rates

The Gibbs posterior convergence rate describes, roughly, the size of the neighborhood around $\theta^\star$ that it assigns nearly all its mass as $n \to \infty$. An important consequence of a posterior convergence rate result is that typical posterior summaries also have nice convergence rate properties; see Corollary 5.3.1.

It turns out that the posterior convergence rate result holds under virtually the same conditions as Theorem 5.2.1 for the M-estimator. The only additional condition needed concerns the prior, and it is very mild.

Assumption 5.3.1 *The prior distribution* Π *for* θ *has a density* π *which is continuous and bounded away from zero in a neighborhood of* $\theta^\star$*.*

Theorem 5.3.1 *Under Assumptions 5.2.1–5.3.1, the Gibbs posterior distribution satisfies* $\Pi_n(A_n) = o_P(1)$ *as* $n \to \infty$*, where* $A_n = \{\theta : |\theta - \theta^\star| > a_n n^{-r}\}$*,* $r = (1 + 2\gamma)^{-1}$ *for* γ *in* (Equation 5.2.5)*, and* a_n *is any diverging sequence.*

Proof 5.3.1 *See Section 5.6.*

Corollary 5.3.1 *Under the conditions of Theorem 5.3.1, if the prior mean for* θ *exists, then the posterior mean* $\tilde{\theta}_n$ *satisfies* $\tilde{\theta}_n - \theta^\star = O_P(a_n n^{-r})$ *as* $n \to \infty$*.*

Proof 5.3.2 *See Section 5.6.*

5.3.3 On scaling the loss function

In Lemma 5.3.1 below, it is shown that the Gibbs posterior may be scaled by a vanishing sequence without sacrificing the convergence rate in Theorem 5.3.1.

Lemma 5.3.1 *Under Assumptions 5.2.1–5.3.1, with* $r = (1+2\gamma)^{-1}$ *and* γ *defined in* (Equation 5.2.5)*, the conclusion of Theorem 5.3.1 holds if the loss function* $\ell(\mathcal{X}, \theta)$ *is scaled by a sequence* ω_n *that vanishes strictly more slowly than* $n^{-\gamma r}$*.*

Proof 5.3.3 *Similar to the proof of Theorem 5.3.1 in Section 5.6.*

By experimentation, with continuous η, it was found that a learning rate of approximately $\omega_n = cn^{-1/4}$, for $c \in (1, 2)$, worked well in terms of credible interval calibration. To avoid making an ad hoc choice of constant, the algorithm in (93) is used; see Chapter 7. This scaling algorithm is applied to each simulated data set, producing a different, data-dependent value of the scale parameter each time. Briefly, ω_n is determined by solving the equation that sets the Gibbs posterior credible interval coverage probability equal to the desired confidence level. The algorithm utilizes standard techniques including stochastic approximation, MCMC, and bootstrapping. In the following simulations, the algorithm succeeds in producing approximately calibrated credible intervals. In the numerical examples that follow, the ω_n selected by the algorithm is, on average, roughly $1.5n^{-1/4}$, which is consistent with the result in Lemma 5.3.1.

5.4 Numerical examples

Consider four examples to illustrate the performance of the Gibbs posterior for the MCID. Each example has a different marginal distribution for X:

Example 1. $X \sim 0.7\mathsf{N}(-1, 1) + 0.3\mathsf{N}(1, 1)$;

Example 2. $X \sim \mathsf{N}(1, 1)$;

Example 3. $X \sim \mathsf{unif}(-2, 4)$.

Example 4. $X \sim \mathsf{gam}(2, 0.5)$.

These examples cover a variety of distributions: bimodal, normal, flat, and skewed. In each example, n independent samples are taken from the respective marginal distributions, and then, given $X_i = x_i$, take Y_i as a ± 1 Bernoulli with probability $F(x_i)$, $i = 1, \dots, n$, where F is the distribution function of X and $\mathsf{ber}(p)$ denotes a Bernoulli distribution with success probability p. The relevant summaries are the bias and standard deviation of the estimators, and the coverage probability and length of the 90% interval estimates. Three sample sizes were considered, namely, $n = 250, 500, 1000$, and the results in Table 5.4.1 and Table 5.4.2 are based on 1000 Monte Carlo samples. The performance of the Gibbs posterior, using the scaling algorithm in (93) and a flat prior for θ, was compared to a baseline method, namely, the M-estimator and the corresponding percentile bootstrap confidence intervals.

Table 5.4.1 shows the empirical bias and standard deviation for both the M-estimator and the posterior mean while Table 5.4.2 shows the empirical coverage probability and length for the 90% interval estimates based on bootstrapping the M-estimator and on the Gibbs posterior sample. Here it can be seen that the additional flexibility of being able to choose the scaling parameter/sequence provides approximately calibrated posterior credible intervals for each n. Overall, the performance of the Gibbs posterior is comparable to the bootstrap, the take-away message being that a Bayesian-like approach need not sacrifice desirable frequentist properties. If reliable prior information is available, which is possible in medical applications where studies are replicated, then this can be readily incorporated into the analysis, naturally providing some improvements. For example, if an accurate, informative $N(-0.5, 1)$ prior is used in Example 1 for $n = 250$, the bias is reduced to 0.01 and the credible interval length is reduced to 0.83 with 90% coverage, an improvement over bootstrap confidence intervals. Additionally, the simulation examples demonstrate the robustness of the Gibbs model as it performs well in a

Example	Method	$n = 250$	$n = 500$	$n = 1000$
1	M-estimator	0.03 (0.21)	0.01 (0.16)	0.01 (0.12)
	Posterior mean	0.03 (0.22)	0.01 (0.17)	0.01 (0.13)
2	M-estimator	0.02 (0.16)	0.02 (0.12)	0.01 (0.10)
	Posterior mean	0.00 (0.15)	0.01 (0.12)	0.00 (0.10)
3	M-estimator	0.01 (0.12)	0.01 (0.10)	0.01 (0.08)
	Posterior mean	0.01 (0.12)	0.00 (0.10)	0.00 (0.07)
4	M-estimator	0.01 (0.12)	0.01 (0.09)	0.01 (0.07)
	Posterior mean	0.03 (0.10)	0.02 (0.08)	0.01 (0.06)

TABLE 5.4.1

ABSOLUTE EMPIRICAL BIAS (AND STANDARD DEVIATION) FOR POINT ESTIMATES.

variety of settings. On the other hand, standard models may not be robust. Bayesian logistic regression, for instance, sometimes performs worse than the Gibbs model, with an average MSE over 1000 simulations of 0.07 for Example 3 and 0.20 for Example 4.

5.5 Conclusion

In this chapter, motivated by a real application in medical statistics, a Gibbs model was developed and compared to existing techniques. In certain applications, like this MCID problem, the statistician may be reluctant to use a likelihood-based model due to fear of misspecification, computational difficulty, or for some other reason. The Gibbs model offers an alternative Bayesian-like approach that does not require a probability model, thus avoiding some of these potential challenges, and this advantage may make Gibbs models widely applicable. As demonstrated, the proposed Gibbs posterior is theoretically justified and provides quality point and

Example	Method	$n = 250$	$n = 500$	$n = 1000$
1	Bootstrap	0.91 (0.86)	0.91 (0.69)	0.93 (0.53)
	Gibbs	0.89 (0.89)	0.89 (0.69)	0.91 (0.55)
2	Bootstrap	0.91 (0.60)	0.91 (0.48)	0.92 (0.38)
	Gibbs	0.89 (0.61)	0.91 (0.50)	0.90 (0.38)
3	Bootstrap	0.90 (0.47)	0.90 (0.38)	0.91 (0.30)
	Gibbs	0.91 (0.48)	0.90 (0.37)	0.90 (0.30)
4	Bootstrap	0.92 (0.39)	0.92 (0.31)	0.92 (0.25)
	Gibbs	0.91 (0.41)	0.90 (0.31)	0.90 (0.24)

TABLE 5.4.2

EMPIRICAL COVERAGE PROBABILITY (AND MEAN LENGTH) OF 90% INTERVAL ESTIMATES.

interval estimates in practice. So, in a certain sense, the Gibbs posterior provides the best of both worlds: that is, one obtains a theoretically justifiable posterior distribution without unnecessary modeling and computations and without worry of model misspecification.

The technical details in this chapter are kept relatively simple due to the fact that θ is a scalar and $\ell(\mathcal{X};\theta)$ is bounded, but the methods can be applied more generally; see Chapter 3. For example, (37) proposed a generalization of the MCID problem in which θ is actually a function of some other covariates, thereby making the MCID "personalized" in a certain sense. Further work on extending both the theory and the computational methods presented here to this more general case is ongoing. A recent paper (see (96) and Chapters 3 and 5) develops a nonparametric Gibbs posterior for inference on a function and prove a smoothness-adaptive convergence rate theorem. The techniques developed therein may be able to be applied in the personalized MCID problem.

5.6 Technical details and proofs

5.6.1 Preliminary results

Here, for the sake of completeness, some basic facts about the empirical risk $R_n(\theta)$ and the risk $R(\theta)$ are summarized. Details will be given only for those results not taken directly from (37).

First, consider properties of the expected loss difference, $R(\theta) - R(\theta^\star)$. By definition of $\theta^\star$, and Assumption 4.2.2 about η, it is known the difference is strictly positive except at $\theta = \theta^\star$. To see this, (37) show that

$$R(\theta) - R(\theta^\star) = 2\int_{\theta^\star}^{\theta} \{\eta(x) - \frac{1}{2}\} p(x)\, dx. \tag{5.6.1}$$

Moreover, by continuity of p in Assumption 5.2.1 and almost everywhere continuity of η derived from Assumption 5.2.2, see that the derivative of $R(\theta){-}R(\theta^\star)$ could be zero only at $\theta = \theta^\star$, which implies that the function is uniformly bounded away from zero outside an interval containing $\theta^\star$. This latter point is important because asymptotic results of, say, the M-estimator require that the minimizer $\theta^\star$ be "well-separated" (e.g. (100) Theorem 5.7). The next goal is to find a lower bound on the expected loss difference in Equation 5.6.1 for parameter values far from the MCID. That is, calculate

$$\inf_{|\theta-\theta^\star|>\delta} R(\theta) - R(\theta^\star). \tag{5.6.2}$$

The following result specifies the bound in Equation 4.2.1 by showing $\alpha = 1 + \gamma$ for the MCID. This result is new and improves the bound given in (37).

Lemma 5.6.1 *Under Assumptions 5.2.1–5.2.2, there exists a constant* $c > 0$ *such that Equation 5.6.2 is lower-bounded by* $c\delta^{1+\gamma}$ *for all sufficiently small* $\delta > 0$.

Proof of Lemma 5.6.1

Since $\eta(x)$ is non-decreasing in x (Assumption 4.2.2) the infimum in Equation 5.6.2 occurs at the boundary, either at $\theta^\star + \delta$ or at $\theta^\star - \delta$. The two cases can be handled similarly, so it is sufficient to consider the case that the infimum is attained at $\theta^\star + \delta$. Monotonicity of η implies that $\eta(\theta^\star + \delta) > \eta(\theta^\star + \delta/2) > \eta(\theta^\star)$. Also, according to Assumption 5.2.1, the marginal density p is bounded away from zero on an interval containing $\theta^\star$, so let b be the infimum over the interval $[\theta^\star - \delta, \theta^\star + \delta]$. Using Equation 5.6.1, one can lower-bound $R(\theta^\star + \delta) - R(\theta^\star)$ as follows:

$$\begin{aligned}
R(\theta^\star + \delta) - R(\theta^\star) &= \int_{\theta^\star}^{\theta^\star + \delta} \{2\eta(x) - 1\} p(x)\, dx \\
&= \Big(\int_{\theta^\star}^{\theta^\star + \delta/2} + \int_{\theta^\star + \delta/2}^{\theta^\star + \delta}\Big)\{2\eta(x) - 1\} p(x)\, dx \\
&> b\,\delta\,\{\eta(\theta^\star + \delta/2) - \tfrac{1}{2}\} \\
&\geq b\,\delta\,\{\eta(\theta^\star + \delta/2) - \eta(\theta^\star)\}.
\end{aligned}$$

By Assumption 5.2.2, in particular, Equation 5.2.5, see that the difference in the last display is bounded below by $c_1(\delta/2)^\gamma$. Plugging this in at the end of the above display gives the advertised lower bound, $c\delta^{1+\gamma}$, where $c = bc_1/2^\gamma$.

Second, some approximation properties are needed for the class of functions

$$\mathcal{L}_\delta := \{\ell(\mathcal{X}, \theta) - \ell(\mathcal{X}, \theta^\star) : |\theta - \theta^\star| < \delta\}, \quad \delta > 0.$$

(37) shows, using the standard partition in the classical Glivenko–Cantelli theorem (e.g. (100), Example 19.6), that the $L_1(P)$ ϵ-bracketing number $N_{[]}(\epsilon, \mathcal{L}_\infty, L_1(P))$ is proportional to ϵ^{-1}. This is enough to show that the class $\mathcal{L}_\infty$ is Glivenko–Cantelli, from which a uniform law of large numbers follows. However, better rates can be obtained by using a local bracketing, i.e., of $\mathcal{L}_\delta$ for finite δ, and Assumptions 5.2.1–5.2.2. Such considerations allow us to remove the unnecessary logarithmic term on the rate presented in Theorem 1 of (37).

Lemma 5.6.2 $N_{[]}(\epsilon, \mathcal{L}_\delta, L_1(P)) \lesssim \delta/\epsilon$.

Proof of Lemma 5.6.2

For the standard Glivenko–Cantelli theorem partition, which is used in (37), one needs to partition the interval $[0, 1]$ into k intervals of length less than ϵ, so k must be greater than $1/\epsilon$, but can be taken less than $2/\epsilon$. By Assumption 5.2.1, see that $\delta \lesssim P(|X - \theta^\star| < \delta) \lesssim \delta$. For the local bracketing, this means one only needs partition an interval of length proportional to δ into intervals of length less than ϵ. Therefore, the total number of intervals is $\lesssim \delta/\epsilon$, as was to be shown.

From this and the fact that the brackets are pairs of indicator functions, one can get a bound on the $\mathsf{L}_2(\mathsf{P})$ bracket number, i.e., $\mathsf{N}_{[]}(\epsilon, \mathcal{L}_\delta, \mathsf{L}_2(\mathsf{P})) \lesssim (\delta/\epsilon)^2$; see Example 19.6 in (100). Then the bracketing integral is

$$J_{[]}(\delta, \mathcal{L}_\delta, \mathsf{L}_2(\mathsf{P})) := \int_0^\delta \{\log \mathsf{N}_{[]}(\epsilon, \mathcal{L}_\delta, \mathsf{L}_2(\mathsf{P}))\}^{1/2}\, d\epsilon \lesssim \delta. \tag{5.6.3}$$

Finally, a maximal inequality is needed for the empirical process $\mathbb{G}_n(\ell(\mathcal{X},\theta) - \ell(\mathcal{X},\theta^\star))$ for θ near $\theta^\star$. This result provides the value of the bound in Equation 4.2.2. Together, Lemma 5.6.1 and Lemma 5.6.3 below provide the Gibbs posterior convergence rate. Towards proving Lemma 5.6.3, (37) show that $g(\theta) = I_{\{|\theta-\theta^\star| \leq \delta\}}$ is an envelop function for $\mathcal{L}_\delta$, with $\|g\|_{\mathsf{L}_2(\mathsf{P})} \lesssim \delta^{1/2}$. Then, given the bound Equation 5.6.3 on the bracketing integral, the maximal inequality in Corollary 19.35 of (100) provides the following.

Lemma 5.6.3 $E_P\{\sup_{|\theta-\theta^\star|<\delta} |\mathbb{G}_n(l(\mathcal{X},\theta) - l(\mathcal{X},\theta^\star))|\} \lesssim \delta^{1/2}$.

5.6.2 Proofs from Section 5.2.2

Proof of Theorem 5.2.1

Similar to the proof of Theorem 2 in (107) and of Theorem 5.52 in (100). The M-estimator $\hat\theta_n$, the global minimizer of R_n, satisfies $R_n(\hat\theta_n) \leq R_n(\theta^\star) + \zeta_n$ for *any* ζ_n. Let $K > 0$ be as in Lemma 4.2.1 in Section 3.1, and take $\zeta_n = K s_n^{1+\gamma}$, where $s_n = a_n n^{-r}$, $r = (1+2\gamma)^{-1}$, and a_n is any divergent sequence. Then see that

$$|\hat\theta_n - \theta^\star| > s_n \implies \sup_{|\theta-\theta^\star|>s_n} \{R_n(\theta^\star) - R_n(\theta)\} \geq -K s_n^{1+\gamma}.$$

By Lemma 4.2.1, the latter event has vanishing P-probability, which implies that $\mathsf{P}(|\hat{\theta} - \theta^\star| > s_n) \to 0$. Therefore, $\hat{\theta}_n - \theta^\star = o_{\mathsf{P}}(s_n)$ or, since a_n is arbitrary, $\hat{\theta}_n - \theta^\star = O_{\mathsf{P}}(n^{-r})$.

5.6.3 Proofs from Section 5.3.2

Proof of Theorem 5.3.1

From Lemma 4.2.1 and Lemma 5.6.1 one may get an exponential bound on the numerator $N_n(A_n)$, i.e., $N_n(A_n) \leq \exp\{-Kns_n^{1+\gamma}\}$ with P-probability approaching 1. For the denominator D_n, by Corollary 4.2.1, for $t_n = \tilde{a}_n n^{-\frac{1+\gamma}{1+2\gamma}}$, where $\tilde{a}_n$ is as in the proof of Theorem 4.2.1, see that $D_n \gtrsim \Pi(\Theta_n)e^{-2nt_n}$, where $\Theta_n = \{\theta : R(\theta) - R(\theta^\star) \leq t_n\}$. The claim is that

$$\Theta_n \supseteq \{\theta : |\theta - \theta^\star| \leq Ht_n\},$$

for some $H > 0$. To see this, first note that if θ is close to $\theta^\star$, then by an argument similar to that in the proof of Lemma 5.6.1 and using the boundedness of η,

$$R(\theta) - R(\theta^\star) \lesssim \int_{\theta^\star}^{\theta} p(x)\, dx.$$

The remaining term in the upper bound is the marginal P-probability assigned to the small interval around $\theta^\star$ which, by Assumption 4.2.1, can be bounded by a constant times $|\theta - \theta^\star|$. Therefore, $R(\theta) - R(\theta^\star) \lesssim |\theta - \theta^\star|$ so, if $|\theta - \theta^\star| \lesssim t_n$, then $R(\theta) - R(\theta^\star) \lesssim t_n$. Under Assumption 5.3.1, one can bound $\Pi(\Theta_n) \gtrsim t_n$. Thus,

$$D_n \gtrsim \exp(-2n^{\frac{\gamma}{1+2\gamma}} - \log n^{\frac{1+\gamma}{1+2\gamma}}) \gtrsim \exp(-C_0 n^{\frac{\gamma}{1+2\gamma}}), \quad \text{with } \mathsf{P}\text{-probability approaching } 1,$$

for some $C_0 > 0$. Putting together the bounds on the numerator and denominator see that, for some constant $M > 0$,

$$\frac{N_n(A_n)}{D_n} \lesssim \exp\Big\{-M\Big(a_n^{1+\gamma} n^{\frac{\gamma}{1+2\gamma}} - \tilde{a}_n n^{\frac{\gamma}{1+2\gamma}}\Big)\Big\},$$

and since $a_n^{1+\gamma} - \tilde{a}_n \uparrow \infty$ by construction, the upper bound vanishes, completing the proof.

Proof of Corollary 5.3.1

Set $s_n = a_n n^{-r}$ for a_n an arbitrary divergent sequence. Next, define $\tilde{s}_n = \tilde{a}_n n^{-r(\gamma)}$, where $\tilde{a}_n$ is such that $\tilde{a}_n / a_n \to 0$, e.g., $\tilde{a}_n = \log a_n$. Now partition $\mathbb{R}$ as $\{\theta : |\theta - \theta^\star| \leq \tilde{s}_n\} \cup \{\theta : |\theta - \theta^\star| > \tilde{s}_n\}$, and write

$$|\tilde{\theta}_n - \theta^\star| \leq \int |\theta - \theta^\star| \Pi_n(d\theta) \leq \tilde{s}_n + \int_{|\theta - \theta^\star| > \tilde{s}_n} |\theta - \theta^\star| \Pi_n(d\theta), \tag{5.6.4}$$

where the first inequality is by Jensen. From the proof of Theorem 5.3.1, the posterior away from $\theta^\star$ is bounded by the prior times some $Z_n = o_P(1)$, uniformly in θ. That is,

$$\int_{|\theta - \theta^\star| > \tilde{s}_n} |\theta - \theta^\star| \Pi_n(d\theta) \leq Z_n \int |\theta - \theta^\star| \Pi(d\theta).$$

In fact, one can bound Z_n more precisely:

$$Z_n \lesssim \exp\Big\{-M\Big(\tilde{a}_n^{1+\gamma} n^{\frac{\gamma}{1+2\gamma}} - n^\beta\Big)\Big\}, \quad \text{sufficiently small } \beta > 0.$$

Dividing through Equation 4.2.4 by s_n see that $s_n^{-1}|\tilde{\theta}_n - \theta^\star|$ is bounded by a constant times

$$\tilde{a}_n/a_n + e^{-\zeta_n} \int |\theta - \theta^\star| \, \Pi(d\theta),$$

where $\zeta_n = M\tilde{a}_n^{1+\gamma} n^{\frac{\gamma}{1+2\gamma}} - Mn^\beta + \log a_n - r \log n$. The first term in the upper bound goes to zero by the choice of $\tilde{a}_n$. The second term goes to zero provided that the prior mean exists and $\zeta_n \to \infty$ as $n \to \infty$. The former condition was assumed, and the latter can be easily arranged by choosing β sufficiently small, so $\tilde{\theta}_n - \theta^\star = o_P(s_n)$.

CHAPTER 6

APPLICATION OF GIBBS MODELS TO INFERENCE ON THE BOUNDARY OF A NOISY IMAGE

6.1 Introduction

In image analysis, the boundary or edge of the image is one of the most important features of the image, and extraction of this boundary is a critical step. An image consists of pixel locations and intensity values at each pixel, and the boundary can be thought of as a curve separating pixels of higher intensity from those of lower intensity. Applications of boundary detection are wide-ranging, e.g., (69) use boundary detection to identify important features in pictures of natural settings, (60) identifies boundaries in medical images, and in (112) boundary detection helps classify the type and severity of wear on machines. For images with noiseless intensity, boundary detection has received considerable attention in the applied mathematics and computer science literature; see, e.g., (115), (66), (60), and (3). However, these approaches suffer from a number of difficulties. First, they can produce an estimate of the image boundary, but do not quantify estimation uncertainty. Second, these methods use a two-stage approach where the image is first smoothed to filter out noise and then a boundary is estimated based on a calculated intensity gradient. This two-stage approach makes theoretical analysis challenging, and no convergence results are available for these methods. Third, in the following examples, these methods perform poorly on noisy data, and one reason for this is that the intensity gradient is less informative for the boundary when data are observed with noise. In the statistics literature, (35) take a Bayesian approach to boundary detection and emphasize borrowing information to recover boundaries of multiple, similar objects in an image. Boundary detection using wombling is also a popular approach; see (62), with applications to geography (64), public health (65), and ecology (22). However, these techniques are used with areal or spatially aggregated data and are not suitable for the pixel data encountered in image analysis.

Section 6.2, describes the image boundary problem, following the setup in Section 4.3.2 and (61). In (61), the authors take a fully Bayesian approach, modeling the probability distributions of the pixel intensities both inside and outside the image. This approach is challenging because it often introduces nuisance parameters in addition to the image boundary. Section 6.3 presents a Gibbs model that avoids this issue.

The asymptotic convergence properties of the Gibbs posterior were investigated in Section 4.3.2 where sufficient conditions were provided for the Gibbs posterior to converge at the minimax optimal rate relative to neighborhoods of the true boundary measured by the Lebesgue measure of a symmetric difference and adaptively to the unknown smoothness of the boundary. In Section 6.4, specific linking functions and priors are chosen to meet the conditions of Theorem 4.3.1. Further, since the Gibbs posterior concentrates at the optimal rate without requiring a model for the pixel intensities, it can be claimed that the inference on the image boundary is robust.

Computation of the Gibbs posterior is relatively straightforward and Section 6.5 presents a reversible jump MCMC method; `R` code to implement to the proposed Gibbs posterior inference is available at `https://github.com/nasyring/GibbsImage` . A comparison of inference based on the proposed Gibbs posterior to that based on the fully Bayes approach in (61) is shown in Section 6.6. For smooth boundaries, the two methods perform similarly, providing very accurate estimation. However, the Gibbs posterior is easier to compute, thanks to there being no nuisance parameters, and is notably more accurate than the Bayes approach when the model is misspecified or when the boundary is not everywhere smooth. The technical details needed to verify the conditions of Theorem 4.3.1 are found in Section 6.7.

6.2 Problem formulation

The problem setup is given in Section 4.3.2 and is briefly reviewed here. The frame of an image is denoted $\Omega \subset \mathbb{R}^2$. Pixels $\mathcal{X}_i = (X_i, Y_i)$, $i = 1, \ldots, n$, where X_i is a pixel location in Ω and Y_i is an intensity measurement are sampled according to

$$
\begin{aligned}
X &\sim g(x), \\
Y \mid (X = x) &\sim f_\Gamma(y)\, 1(x \in \Gamma) + f_{\Gamma^c}(y)\, I(x \in \Gamma^c),
\end{aligned}
$$

where g is a density on Ω, f_Γ and f_{Γ^c} are densities on the intensity space, and $I(\cdot)$ denotes an indicator function.

The density g for the pixel locations is of no interest and, as is common, it may be considered known. So the question is how to handle the two conditional distributions, f_Γ and f_{Γ^c}. (61) take a fully Bayesian approach, modeling both f_Γ and f_{Γ^c}. By specifying these parametric models, they are obligated to introduce priors and carry out posterior computations for the corresponding parameters. Besides the efforts needed to specify models and priors and to carry out posterior computations, there is also a concern that the models for the pixel intensities might be incorrect, potentially biasing the inference on $\Gamma^\star$. Since the forms of f_Γ and f_{Γ^c}, as well as any associated parameters, are of no inferential interest in the boundary detection problem, it is natural to ask if inference can be carried out robustly, without modeling the pixel intensities.

This question is answered affirmatively, and a Gibbs model for Γ is presented in Section 6.3. In the present context, suppose there is a linking function $\ell(\mathcal{X}, \Gamma)$ that measures how well an observed pixel location–intensity pair (x, y) agrees with a particular region Γ. The defining characteristic of $\ell(\mathcal{X}, \Gamma)$ is that $\Gamma^\star$ should be the unique minimizer of $\Gamma \mapsto R(\Gamma)$, where $R(\Gamma) = E_{P_{\Gamma^\star}}(\ell(\mathcal{X}, \Gamma))$ is the risk. A main contribution here, in Section 6.3.1, is specification of a loss function that meets this criterion. A necessary condition to construct such a loss function is that the distribution functions F_Γ and F_{Γ^c} are stochastically ordered. Imagine a gray-scale image; then, stochastic ordering means this image is lighter, on average, inside the boundary than outside the boundary, or vice-versa. In the specific context of binary images, this means that it is assumed, without loss of generality, $f_{\Gamma^\star} > f_{\Gamma^{\star c}}$, while for continuous images, again without loss of generality, $F_{\Gamma^\star}(y) < F_{\Gamma^{\star c}}(y)$ for all $y \in \mathbb{R}$. The Gibbs posterior is defined as usual, in Equation 2.2.1.

Proper scaling of the loss in the Gibbs model by a learning rate ω is important, and some context-specific scaling is provided; see Sections 6.4 and 6.5.2. The choice of prior Π for Γ is discussed in Section 6.3.2. Together, the linking function $\ell(\mathcal{X}, \Gamma)$ and the prior for Γ define the Gibbs model, no further modeling is required.

6.3 Gibbs model for the image boundary

6.3.1 Linking function

To start, consider inference on the image boundary when the pixel intensity is binary, i.e., $Y_i \in \{-1, +1\}$. In this case, the conditional distributions, f_Γ and f_{Γ^c}, in Equation 4.3.1 must be Bernoulli, so the likelihood is known. Even though the parametric form of the conditional distributions is known, the Gibbs approach only requires prior specification and posterior computation related to Γ, whereas the Bayes approach must also deal with the nuisance parameters in these Bernoulli conditional distributions. More generally, this binary case is relatively simple and will provide insights into how to formulate a Gibbs model in the more challenging continuous intensity problem.

For the binary case, a reasonable choice for the linking function $\ell(\mathcal{X}, \Gamma)$ is the following weighted misclassification error loss, depending on a parameter h:

$$\ell(\mathcal{X}, \Gamma) = \ell(\mathcal{X}, \Gamma, \mid h) = h\, I(y = +1, x \in \Gamma^c) + I(y = -1, x \in \Gamma). \tag{6.3.1}$$

Note that putting $h = 1$ in Equation 6.3.1 gives the usual misclassification error loss. In order for the Gibbs model to work the risk, or expected loss, must be minimized at the true $\Gamma^\star$ for some h. Picking h to ensure this property holds necessitates making a connection between the probability model in Equation 4.3.1 and the loss in Equation 6.3.1. The condition in

Equation 6.3.2 below is just the connection needed. With a slight abuse of notation, let $f_{\Gamma^\star}$ and $f_{\Gamma^{\star c}}$ denote the conditional probabilities for the event $Y = +1$, given $X \in \Gamma^\star$ and $X \in \Gamma^{\star c}$, respectively. Recall the stochastic ordering assumption implies $f_{\Gamma^\star} > f_{\Gamma^{\star c}}$.

Proposition 6.3.1 *Using the notation in the previous paragraph, if* h *is such that*

$$f_{\Gamma^\star} > \frac{1}{1+h} > f_{\Gamma^{\star c}}, \tag{6.3.2}$$

then the risk function $R(\Gamma) = E_{P_{\Gamma^\star}}(\ell(\mathcal{X}, \Gamma))$ *is minimized at* $\Gamma^\star$.

Proof 6.3.1 *See Section 6.7.1.*

Either one—but not both—of the above inequalities can be made inclusive and the result still holds. The condition in Equation 6.3.2 deserves additional explanation. For example, if it is known $f_{\Gamma^\star} \geq \frac{1}{2} > f_{\Gamma^{\star c}}$, then take $h = 1$, which means that in Equation 6.3.1 there is a penalty on both intensities of 1 outside Γ and intensities of -1 inside Γ by a loss of 1. If, however, it is known the overall image brightness is higher so that $f_{\Gamma^\star} \geq \frac{4}{5} > f_{\Gamma^{\star c}}$ then take $h = 1/4$ in Equation 6.3.1 and penalize bright pixels outside Γ by less than dull pixels inside Γ. To see why this loss balancing is so crucial, suppose the second case above holds so that $f_{\Gamma^\star} = 4/5$ and $f_{\Gamma^{\star c}} = 3/4$, but put $h = 1$ anyway. Then, in Equation 6.3.1, $1(y = +1, x \in \Gamma^c)$ is very often equal to 1 while $1(y = -1, x \in \Gamma)$ is very often 0. This will likely minimize the expected loss then by incorrectly taking Γ to be all of Ω so that the first term in the loss vanishes. Knowing a good choice of h corresponds to having some prior information about $f_{\Gamma^\star}$ and $f_{\Gamma^{\star c}}$, but it is also possible to use the data to estimate a good value of h and this data-driven strategy is described in Section 6.5.2.

In the continuous case it is assumed that the pixel intensity takes its value in $\mathbb{R}$. The proposed strategy is to modify the misclassification error loss Equation 6.3.1 by working with a suitably discretized pixel intensity measurement. In particular, consider the following version of the misclassification error, depending on parameters (c, k, z):

$$\ell(\mathcal{X}, \Gamma) = \ell(\mathcal{X}, \Gamma, | \, c, k, z) = k \, I(y > z, x \in \Gamma^c) + c \, I(y \le z, x \in \Gamma). \tag{6.3.3}$$

Again, it can be claimed that, for suitable (c, k, z), the risk function is minimized at $\Gamma^\star$. Let F_Γ and F_{Γ^c} denote the distribution functions corresponding to the densities f_Γ and f_{Γ^c} in Equation 4.3.1, respectively. Recall the stochastic ordering assumption implies $F_{\Gamma^\star}(z) < F_{\Gamma^{\star c}}(z)$.

Proposition 6.3.2 *If* (c, k, z) *in Equation 6.3.3 satisfies*

$$F_{\Gamma^\star}(z) < \frac{k}{k+c} < F_{\Gamma^{\star c}}(z), \tag{6.3.4}$$

then the risk function $R(\Gamma) = E_{P_{\Gamma^\star}}(\ell(\mathcal{X}, \Gamma))$ *is minimized at* $\Gamma^\star$.

Proof 6.3.2 *See Section 6.7.2.*

Again, either one—but not both—of the above inequalities in Equation 6.3.4 can be made inclusive and the result still holds. The parameters k and c in Equation 6.3.3 determine the scale of the loss as mentioned in Section 6.1. This implies that the true image $\Gamma^\star$ can be identified by working with a suitable version of the loss Equation 6.3.3. A similar condition to Equation 6.3.4, see Assumption 6.4.1 in Section 6.4, says what scaling is needed in order for the Gibbs posterior to concentrate at the optimal rate. Although the conditions on the scaling all involve the unknown distribution $P_{\Gamma^{\star}}$, a good choice of (c, k, z) can be made based on the data alone, without prior information, and this strategy is discussed in Section 6.5.2.

6.3.2 Prior specification

A prior distribution for the boundary of the region Γ is specified by first expressing the pixel locations x in terms of polar coordinates (θ, r), an angle and radius, where $\theta \in [0, 2\pi]$ and $r > 0$. The specific reference point and angle in Ω used to define polar coordinates are essentially arbitrary, subject to the requirement that any point in $\Gamma^{\star}$ can be connected to the reference point by a line segment contained in $\Gamma^{\star}$. In (61), the authors tested the influence of the reference point in simulations and found it to have little influence on the results. Using polar coordinates the boundary of Γ can be determined by the parametric curve $(\theta, \gamma(\theta))$.

Whether one is taking a Bayes or Gibbs approach, a natural strategy to model the image boundary is to express γ as a linear combination of suitable basis functions, i.e., $\gamma(\theta) = \hat{\gamma}_{D,\beta}(\theta) = \sum_{j=1}^{D} \beta_j B_{j,D}(\theta)$. In (61), the authors use the eigenfunctions of the squared exponential periodic kernel as their basis functions. Here a model based on free knot b-splines is considered, where the basis functions are defined recursively as in Section 4.3.1. Note that the coefficient vector β is restricted to be positive because the function values $\gamma(\theta)$ measure the radius of a curve from the origin. In the simulations in Section 6.6, the coefficients $\beta_2, ..., \beta_D$ are free parameters, while β_1 is calculated deterministically to force the boundary to be closed, i.e. $\gamma(0) = \gamma(2\pi)$, and it is required that $t_1 = 0$ and $t_D = 2\pi$; all other inner knots are free. The model based on the b-spline representation seems to perform as well as the eigenfunctions used in (61) for smooth boundaries, but a bit better for boundaries with corners; see the examples in Section 6.6.

Therefore, the boundary curve is γ is parametrized by an integer D and a D-vector β. A prior Π on (D, β) may be introduced hierarchically as follows: D has a Poisson distribution with rate μ_D and, given D, the coordinates $\beta_1, \ldots, \beta_D$ of β are iid exponential with rate μ_β. These choices satisfy the technical conditions on Π detailed in Section 4.3.2 and Assumption 4.3.4. In the numerical experiments in Section 6.6, the values $\mu_D = 12$ and $\mu_\beta = 10$ are used.

6.4 Gibbs posterior convergence

The Gibbs model depends on two inputs, namely, the prior and the linking function. In order to ensure that the Gibbs posterior enjoys desirable asymptotic properties, some conditions on both of these inputs are required. The first assumption listed below concerns the linking function, which functions like a misclassification loss function; the second concerns the true image boundary $\gamma^\star = \partial\Gamma^\star$; and the third concerns the prior. Here the focus is on the continuous intensity case, since the only difference between this and the binary case is that the latter provides the discretization itself.

Assumption 6.4.1 *Loss function parameters* (c, k, z) *in Equation 6.3.3 satisfy*

$$F_{\Gamma^\star}(z) < \frac{e^k - 1}{e^{c+k} - 1} \quad \textit{and} \quad F_{\Gamma^{\star c}}(z) > \frac{e^k - 1}{e^k - e^{-c}}. \tag{6.4.1}$$

Compared to the Equation 6.3.4 that was enough to allow the linking function to identify the true $\Gamma^\star$, Equation 6.4.1 in Assumption 6.4.1 is only slightly stronger. This can be seen from the following inequality:

$$\frac{e^k - 1}{e^k - e^{-c}} > \frac{k}{k + c} > \frac{e^k - 1}{e^{c+k} - 1}.$$

However, if (c, k) are small, then the three quantities in the above display are all approximately equal, so Assumption 6.4.1 is not much stronger than what is needed to identify $\Gamma^\star$. Again, these conditions on (c, k, z) can be understood as providing a meaningful scale to the linking function. Intuitively, the scale of the linking function between observations receiving no loss versus some loss, expressed by parameters k and c, should be related to the level of information in the data. When $F_{\Gamma^\star}(z)$ and $F_{\Gamma^{\star c}}(z)$ are far apart, the data can more easily distinguish between $F_{\Gamma^\star}$ and $F_{\Gamma^{\star c}}$, so one may assign larger losses than when $F_{\Gamma^\star}(z)$ and $F_{\Gamma^{\star c}}(z)$ are close and the data are relatively less informative.

The Gibbs posterior based on the particular linking functions given in Section 6.3.1 along with Assumption 6.4.1 and the prior specified in Section 6.3.2 converges at the minimax rate as described in Section 4.3.2. The main effort needed is to verify the conditions of Assumption 4.3.2 sufficient for Theorem 4.3.1.

Theorem 6.4.1 *With a slight abuse of notation, let* Π *denote the prior for* Γ*, induced by that on* (D, β)*, and* Π_n *the corresponding Gibbs posterior from Equation 2.2.1 using the linking* *function in Equation 6.3.3. Under Assumptions 6.4.1, 4.3.1, 4.3.3, and 4.3.4, for any positive sequence* $M_n \to \infty$

$$P_{\Gamma^\star}\Pi_n(\{\Gamma : \lambda(\Gamma^\star \triangle \Gamma) > M_n \epsilon_n\}) \to 0 \quad \textit{as } n \to \infty,$$

where $\epsilon_n = \{(\log n)/n\}^{\alpha/(\alpha+1)}$ *and* α *is the smoothness coefficient in Assumption 4.3.1.*

See Section 6.7 for a proof of Theorem 6.4.1.

6.5 Computation

6.5.1 Sampling algorithm

A reversible jump MCMC scheme is used, as in (32), to sample from the Gibbs posterior. These methods have been used successfully in Bayesian free-knot spline regression problems; see, e.g., (19) and (20). Although the sampling procedure is more complicated when allowing the number and locations of knots to be random versus using fixed knots, the resulting spline functions can do a better job fitting curves with low smoothness.

To summarize the algorithm, start with the prior distribution Π for (D, β) as discussed in Section 6.3.2. Next, initialize values of D, the knot locations $\{t_{-2}, ..., t_{D+3}\}$, and the values of $\beta_2, ..., \beta_D$. The value of β_1 is then calculated numerically to force closure. In the examples below, $D = 12$ with $t_{-2} = -2$, $t_{-1} = -1$, $t_0 = -0.5$, $t_{13} = 2\pi + 0.5$, $t_{14} = 2\pi + 1$, $t_{15} = 2\pi + 2$ and $t_1, ..., t_{12}$ evenly spaced in $[0, 2\pi]$. Set inner knots $t_0 = 0$ and $t_D = 2\pi$ while the other inner knot locations remain free to change in birth, death, and relocation moves; also set $\beta_2 = \beta_3 = ... = \beta_{12} = 0.1$. Then the following three steps constitute a single iteration of the reversible jump MCMC algorithm to be repeated until the desired number of samples are obtained:

1. Use Metropolis-within-Gibbs steps to update the elements of the β vector, again solving for β_1 to force closure at the end. In the examples, a normal proposal distribution centered at the current value of the element of the β vector, and with standard deviation 0.10 is used.
2. Randomly choose to attempt either a birth, death, or relocation move to add a new inner knot, delete an existing inner knot, or move an inner knot.
3. Attempt the jump move proposed in Step 2. The β vector must be appropriately modified when adding or deleting a knot, and again β_1 must be chosen to force closure. Details on the calculation of acceptance probabilities for each move can be found in (19) and (20).

R code to implement this Gibbs posterior sampling scheme, along with the empirical loss scaling method described in Section 6.5.2, is available at `https://github.com/nasyring/GibbsImage`

.

6.5.2 Loss scaling

It is not clear how to select (c, k, z) to satisfy Assumption 6.4.1 without knowledge of $F_{\Gamma^\star}$ and $F_{\Gamma^{\star c}}$. However, it is fairly straightforward to select values of (c, k, z) based on the data which are likely to meet the required condition. First, some notion of optimal (c, k, z) values must be defined. If $F_{\Gamma^\star}$ and $F_{\Gamma^{\star c}}$ were known, then z could be selected to maximize $F_{\Gamma^{\star c}}(z) - F_{\Gamma^\star}(z)$ because this choice of z gives the point at which $F_{\Gamma^\star}$ and $F_{\Gamma^{\star c}}$ are most easily distinguished. Then,

(c, k) would be be chosen as the largest values such that Equation 6.4.1 holds. Intuitively, (c, k) should be large so that the linking function in Equation 6.3.3 is more sensitive to departures from $\gamma^\star$.

Since $F_{\Gamma^\star}$ and $F_{\Gamma^{\star c}}$ are not known, $F_{\Gamma^\star}(z)$ and $F_{\Gamma^{\star c}}(z)$ should be estimated from the data. In order to do this, an estimate of $\gamma^\star$ is needed in order to define the regions $\Gamma^\star$ and $\Gamma^{\star c}$. For a grid of z values $z_1, z_2, ..., z_g$, minimize Equation 6.3.3 for each z_j where $(c, k) = (c_j, k_j)$ are taken to satisfy $\frac{k_j}{k_j+c_j} = n^{-1}|\{i : y_i \leq z_j\}|$. The resulting classifiers obtained from these minimizations are used to estimate the regions $\Gamma^\star$ and $\Gamma^{\star c}$ and $F_{\Gamma^\star}$ and $F_{\Gamma^{\star c}}$ are estimated by their corresponding empirical versions $\hat{F}_{\Gamma^\star}$ and $\hat{F}_{\Gamma^{\star c}}$, respectively. Then, z is chosen from among the z_j such that $z = \arg\max_{z_j} \hat{F}_{\Gamma^{\star c}}(z_j) - \hat{F}_{\Gamma^\star}(z_j)$. Based on this choice of z, choose the final values of (c, k) to satisfy Equation 6.4.1 replacing $F_{\Gamma^\star}(z)$ and $F_{\Gamma^{\star c}}(z)$ by their estimates $\hat{F}_{\Gamma^\star}(z)$ and $\hat{F}_{\Gamma^{\star c}}(z)$.

Based on the simulations in Section 6.6, this method produces values of (c, k, z) very close to their optimal values. Importantly, the estimated (c, k) are more likely to be smaller than their optimal values than larger, which makes the estimates more likely to satisfy Equation 6.4.1. This is a consequence of the stochastic ordering of $F_{\Gamma^\star}$ and $F_{\Gamma^{\star c}}$. Unless the classifier obtained by minimizing Equation 6.3.3 is perfectly accurate, it will tend to mix together samples from $F_{\Gamma^\star}$ and $F_{\Gamma^{\star c}}$ in the estimates. If the estimate of $F_{\Gamma^\star}(z)$ is contaminated with some observations from $F_{\Gamma^{\star c}}$, $F_{\Gamma^\star}(z)$ will tend to be overestimated, and vice versa $F_{\Gamma^{\star c}}(z)$ will tend to be underestimated. These errors will cause (c, k) to be underestimated, and therefore more likely to satisfy Equation 6.4.1.

6.6 Numerical examples

The Gibbs model was tested on data from both binary and continuous images following much the same setup as in (61). The pixel locations in $\Omega = [-\frac{1}{2}, \frac{1}{2}]^2$ are sampled by starting with a fixed $m \times m$ grid in Ω and making a small random uniform perturbation at each grid point. Several different pixel intensity distributions are considered. Two types of shapes for $\Gamma^\star$ are considered: an ellipse with center $(0.1, 0.1)$, rotated at an angle of 60 degrees, with major axis length 0.35 and minor axis length 0.25; and a centered equilateral triangle of height 0.5. The ellipse boundary will test the sensitivity of the model to boundaries which are off-center while the triangle tests the model's ability to identify non-smooth boundaries.

Four binary intensity images and four continuous intensity images are used as examples and compared with the Bayesian method in (61). Codes for implementing their fully Bayesian approach are available via CRAN in the *BayesBD* package; see (95).

B1. Ellipse image, $m = 100$, and $F_{\Gamma^\star}$ and $F_{\Gamma^{\star c}}$ are Bernoulli with parameters 0.5 and 0.2, respectively.

B2. Same as B1 but with triangle image.

B3. Ellipse image, $m = 500$, and $F_{\Gamma^\star}$ and $F_{\Gamma^{\star c}}$ are Bernoulli with parameters 0.25 and 0.2, respectively.

B4. Same as B3 but with triangle image.

C1. Ellipse image, $m = 100$, and $F_{\Gamma^\star}$ and $F_{\Gamma^{\star c}}$ are $N(4, 1.5^2)$ and $N(1, 1)$, respectively.

C2. Same as C1 but with triangle image.

C3. Ellipse image, $m = 100$, and $F_{\Gamma^\star}$ and $F_{\Gamma^{\star c}}$ are $0.2\,N(2, 10) + 0.8\,N(0, 1)$, a normal mixture, and $N(0, 5)$, respectively.

C4. Ellipse image, $m = 100$, and $F_{\Gamma^\star}$ and $F_{\Gamma^{\star c}}$ are t distributions with 3 degrees of freedom and non-centrality parameters 1 and 0, respectively.

For binary images, the likelihood must be Bernoulli, so the Bayesian model is correctly specified in cases B1–B4. For the continuous examples in C1–C4, a Gaussian likelihood is assumed for the Bayesian model. Then, cases C1 and C2 will show whether or not the Gibbs model can compete with the Bayesian model when the model is correctly specified, while cases C3 and C4 will demonstrate the superiority of the Gibbs model over the Bayesian model when there is model misspecification. Again, the Gibbs model has the added advantage of not having to specify priors for or sample values of the mean and variance hyperparameters associated with the normal conditional distributions.

Each example scenario was replicated 100 times for both the Gibbs and Bayesian models, each time producing a posterior sample of size 4000 after a burn in of 1000 samples. The errors—Lebesgue measure of the symmetric difference—were recorded for each run along with the estimated linking function parameters for the Gibbs models for continuous images. The results are summarized in Table 6.6.1 and Table 6.6.2. The Gibbs model is competitive with the fully Bayesian model in Examples B1–B4, C1, and C2, when the likelihood is correctly specified. When the likelihood is misspecified, there is a chance that the Bayesian model will fail, as in Examples C3 and C4. However, the Gibbs model does not depend on a likelihood, only the stochastic ordering of $F_{\Gamma^\star}$ and $F_{\Gamma^{\star c}}$, and it continues to perform well in these non-

TABLE 6.6.1

AVERAGE ERRORS (AND STANDARD DEVIATIONS) FOR EACH EXAMPLE.

Model	B1	B2	B3	B4	C1	C2	C3	C4
Bayes	0.00 (0.00)	0.02 (0.00)	0.01 (0.00)	0.02 (0.01)	0.03 (0.03)	0.04 (0.03)	0.11 (0.06)	0.10 (0.05)
Gibbs	0.01 (0.00)	0.01 (0.00)	0.02 (0.01)	0.02 (0.01)	0.01 (0.00)	0.01 (0.00)	0.01 (0.01)	0.01 (0.01)

TABLE 6.6.2

AVERAGE (AND OPTIMAL) VALUES OF THE PARAMETERS (C, K, Z).

Parameter	C1	C2	C3	C4
c	1.45 (1.86)	1.47 (1.86)	0.80 (1.27)	0.71 (0.80)
k	2.30 (2.36)	2.29 (2.36)	0.26 (0.34)	0.71 (0.75)
z	2.43 (2.40)	2.39 (2.40)	-1.83 (-1.76)	0.46 (0.46)

Gaussian examples. From Table 6.6.2, it can be seen that the empirical method described in Section 6.5.2 is able to select parameters for the loss function in Equation 6.3.3 close to the optimal values and meeting Assumption 6.4.1.

Figure 6.6.1 shows the results of the Bayesian and Gibbs models for one simulation run in each of Examples B1–B2 and C1–C4. The 95% credible regions, as in (61), are highlighted in gray around the posterior means. That is, let $u_i = \sup_\theta\{|\gamma_i(\theta) - \hat{\gamma}(\theta)|/s(\theta)\}$, where $\gamma_i(\theta)$ is the i^{th} posterior boundary sample, $\hat{\gamma}(\theta)$ is the pointwise posterior mean and $s(\theta)$ the pointwise standard deviation of the $\gamma(\theta)$ samples. If τ is the 95th percentile of the u_i's, then a 95% uniform credible band is given by $\hat{\gamma}(\theta) \pm \tau\, s(\theta)$. The results of cases B2 and C2 suggest that free-knot b-splines may do a better job of approximating non-smooth boundaries than the kernel basis functions used by (61).

6.7 Proofs

6.7.1 Proof of Proposition 6.3.1

By the definition of the loss function in Equation 6.3.1, for a fixed h and for any $\Gamma \subset \Omega$, it follows that

$$\begin{aligned}
\ell(\mathcal{X}, \Gamma) - \ell(\mathcal{X}, \Gamma^\star) &= h\, I(Y = +1, X \in \Gamma^c) - h\, I(Y = +1, X \in \Gamma^{\star c}) \\
&\quad + I(Y = -1, X \in \Gamma) - I(Y = -1, X \in \Gamma^\star) \\
&= h\, I(Y = +1, X \in \Gamma^\star \setminus \Gamma) - I(Y = -1, X \in \Gamma^\star \setminus \Gamma) \\
&\quad + I(Y = -1, X \in \Gamma \setminus \Gamma^\star) - h I(Y = +1, X \in \Gamma \setminus \Gamma^\star).
\end{aligned}$$

Then the expectation of the loss difference above is

$$P_g(X \in \Gamma^\star \setminus \Gamma)\,(h f_{\Gamma^\star} + f_{\Gamma^\star} - 1) + P_g(X \in \Gamma \setminus \Gamma^\star)\,(1 - f_{\Gamma^{\star c}} - h f_{\Gamma^{\star c}}),$$

where P_g is the probability relative to the marginal distribution g of X. This quantity is zero if and only if $\Gamma = \Gamma^\star$. It can also be lower bounded by

$$P_g(X \in \Gamma \triangle \Gamma^\star)\, \min\{h f_{\Gamma^\star} + f_{\Gamma^\star} - 1, 1 - f_{\Gamma^{\star c}} - h f_{\Gamma^{\star c}}\}.$$

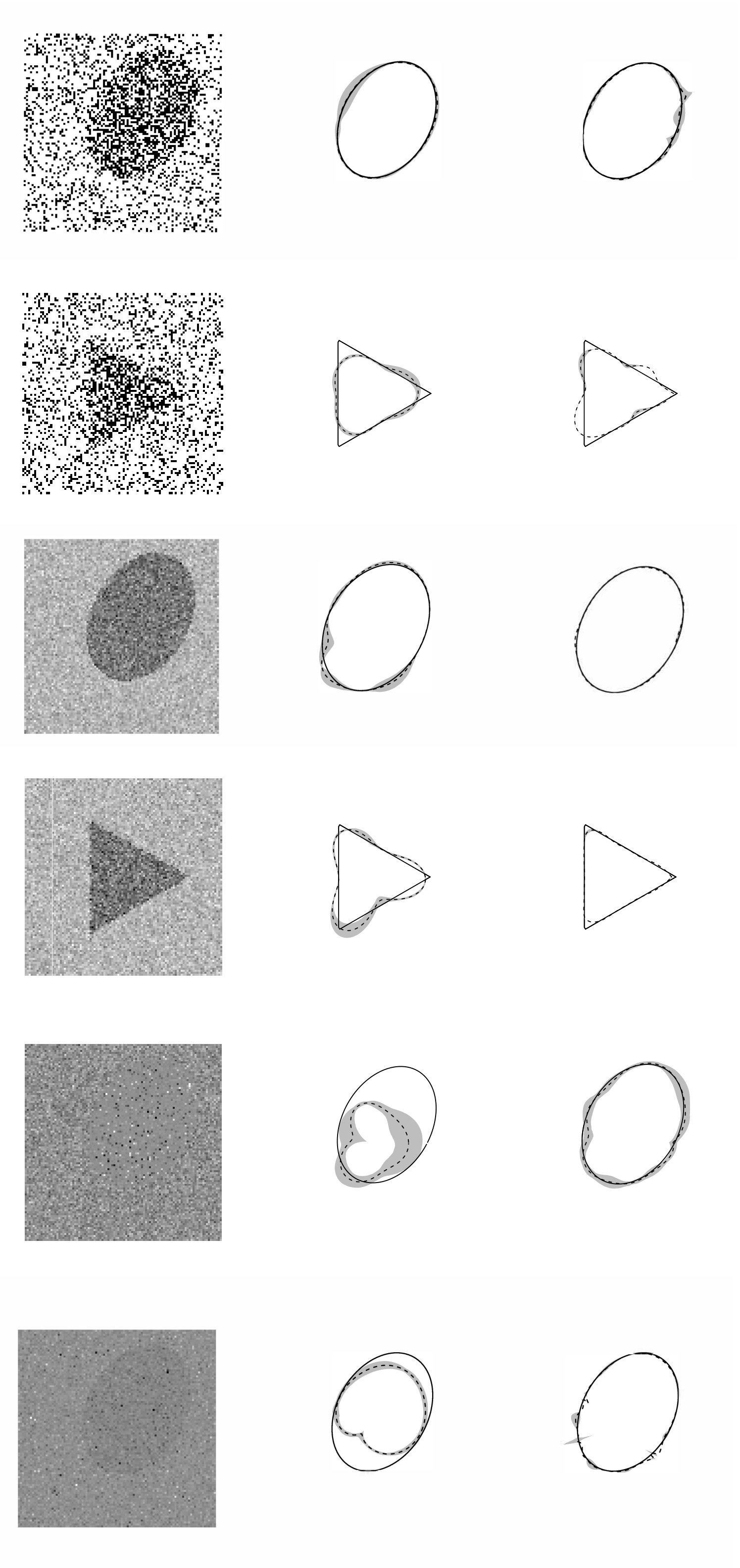


Figure 6.6.1. From top, Examples B1–B2, C1–C4. In each row, the observed image is on the left, the Bayesian posterior mean estimator is in the middle, and the Gibbs posterior mean estimator is on the right. Solid lines show the true image boundary, dashed lines are the estimates, and gray regions are 95% credible bands.

Given the condition (Equation 6.3.2) in Proposition 6.3.1, both terms in the minimum are positive. Therefore, $R(\Gamma) \geq R(\Gamma^\star)$ with equality if and only if $\Gamma = \Gamma^\star$, proving the claim.

6.7.2 Proof of Proposition 6.3.2

The proof here is very similar to that of Proposition 6.3.1. By the definition of the loss function in Equation 6.3.3, for any fixed (c, k, z) and for any $\Gamma \subset \Omega$, get

$$\begin{aligned}
\ell(\mathcal{X}, \Gamma) - \ell(\mathcal{X}, \Gamma^\star) &= k\, I(Y \geq z, X \in \Gamma^c) - k\, I(Y \geq z, X \in \Gamma^{\star c}) \\
&\qquad + c\, I(Y < z, X \in \Gamma) - c\, I(Y < z, X \in \Gamma^\star) \\
&= k\, I(Y \geq z, X \in \Gamma^\star \setminus \Gamma) - c\, I(Y < z, X \in \Gamma^\star \setminus \Gamma) \\
&\qquad + c\, I(Y < 1, X \in \Gamma \setminus \Gamma^\star) - k\, I(Y \geq z, X \in \Gamma \setminus \Gamma^\star).
\end{aligned}$$

Then, the expectation of the loss difference above is given by

$$P_g(X \in \Gamma^\star \setminus \Gamma)\{k - kF_{\Gamma^\star}(z) - cF_{\Gamma^\star}(z)\} + P_g(X \in \Gamma \setminus \Gamma^\star)\{cF_{\Gamma^{\star c}}(z) - k + kF_{\Gamma^{\star c}}(z)\},$$

where, again, P_g is the probability relative to the marginal distribution g of X. This quantity is zero if and only if $\Gamma = \Gamma^\star$. It can also be lower bounded by

$$P_g(X \in \Gamma \triangle \Gamma^\star)\ \min\big\{k - kF_{\Gamma^\star}(z) - cF_{\Gamma^\star}(z), cF_{\Gamma^{\star c}}(z) - k + kF_{\Gamma^{\star c}}(z)\big\}.$$

Given the condition (Equation 6.3.4) in Proposition 6.3.2, both terms in the minimum are positive. Therefore, $R(\Gamma) \geq R(\Gamma^\star)$ with equality if and only if $\Gamma = \Gamma^\star$, proving the claim.

6.7.3 Preliminary results

The following lemma draws a connection between the distance defined by the Lebesgue measure of the symmetric difference and the sup-norm between the boundary functions.

Lemma 6.7.1 *Suppose $\Gamma^\star$, with boundary $\gamma^\star = \partial\Gamma^\star$, satisfies Assumption 4.3.1, in particular, $\underline{\gamma}^\star := \inf_{\theta\in[0,2\pi]} \gamma^\star(\theta) > 0$. Take any $\Gamma \subset \Omega$, with $\gamma = \partial\Gamma$, such that $\lambda(\Gamma\triangle\Gamma^\star) > \delta$, and any $\widetilde{\Gamma} \subset \Omega$ such that $\tilde{\gamma} = \partial\widetilde{\Gamma}$ satisfies $\|\tilde{\gamma} - \gamma\|_\infty < \omega\delta$, where $\omega \in (0, 1)$. Then*

$$\lambda(\widetilde{\Gamma}\triangle\Gamma^\star) > \frac{4\delta}{\underline{\gamma}^\star}\Big(\frac{1}{\mathrm{diam}(\Omega)} - \pi\omega\Big),$$

where $\mathrm{diam}(\Omega) = \sup_{x,x'\in\Omega} \|x - x'\|$ is the diameter of Ω. So, if $\omega < \{\pi\,\mathrm{diam}(\Omega)\}^{-1}$, then the lower bound is a positive multiple of δ.

Proof 6.7.1 *Recall the connection between the symmetric difference-based distance and the L_1 distance between boundary functions from the proof of Theorem 4.3.1,*

$$\tfrac{1}{2}\underline{\gamma}^\star\|\gamma - \gamma^\star\|_1 \le \lambda(\Gamma\triangle\Gamma^\star) \le \tfrac{1}{2}\,\mathrm{diam}(\Omega)\|\gamma - \gamma^\star\|_1. \tag{6.7.1}$$

Next, if $\lambda(\Gamma\triangle\Gamma^\star) > \delta$, which is positive by Assumption 4.3.1, then it follows from the right-most inequality in Equation 4.3.6 that $\mathrm{diam}(\Omega)\|\gamma - \gamma^\star\|_1 > 2\delta$ and, by the triangle inequality,

$$\mathrm{diam}(\Omega)\{\|\gamma - \tilde{\gamma}\|_1 + \|\tilde{\gamma} - \gamma^\star\|_1\} > 2\delta.$$

See that $\|\gamma-\tilde{\gamma}\|_1 \leq 2\pi\|\gamma-\tilde{\gamma}\|_\infty$ *which, by assumption, is less than* $2\pi\omega\delta$*. Consequently,*

$$\mathrm{diam}(\Omega)\{2\pi\omega\delta + \|\tilde{\gamma}-\gamma^\star\|_1\} > 2\delta$$

and, hence,

$$\|\tilde{\gamma}-\gamma^\star\|_1 > \frac{2\delta}{\mathrm{diam}(\Omega)} - 2\pi\omega\delta.$$

By the left-most inequality in Equation 4.3.6, see that

$$\lambda(\widetilde{\Gamma}\triangle\Gamma^\star) > \frac{4\delta}{\underline{\gamma}^\star\,\mathrm{diam}(\Omega)} - \frac{4\pi\omega\delta}{\underline{\gamma}^\star} = \frac{4\delta}{\underline{\gamma}^\star}\Big(\frac{1}{\mathrm{diam}(\Omega)} - \pi\omega\Big),$$

which is the desired bound. It follows immediately that the lower bound is a positive multiple of δ *if* $\omega < \{\pi\,\mathrm{diam}(\Omega)\}^{-1}$.

6.7.4 Proof of Theorem 6.4.1

Theorem 6.4.1 can be proven by verifying the conditions in Assumption 4.3.2 necessary for Theorem 4.3.1.

First show that if Equation 6.4.1 holds, then $\mathsf{E}_{\mathsf{P}_{\Gamma^\star}} \exp\{-(\ell(\mathcal{X},\Gamma)-\ell(\mathcal{X},\Gamma^\star))\} < 1-\rho\lambda(\Gamma^\star\triangle\Gamma)$ for a constant $\rho \in (0,1)$.

From the proof of Proposition 6.3.2, see that

$$\begin{aligned}\ell(\mathcal{X},\Gamma)-\ell(\mathcal{X},\Gamma^\star) &= k\,I(y \geq z, x \in \Gamma^\star\setminus\Gamma) - c\,I(y < z, x \in \Gamma^\star\setminus\Gamma)\\ &\quad + c\,I(y < z, x \in \Gamma\setminus\Gamma^\star) - k\,I(y \geq z, x \in \Gamma\setminus\Gamma^\star).\end{aligned}$$

The key observation is that, if $x \not\in \Gamma \triangle \Gamma^\star$, then the linking function difference is 0 and, therefore, the exponential of the linking function difference is 1. Taking expectation with respect $P_{\Gamma^\star}$, get

$$\begin{aligned} E_{P_{\Gamma^\star}} \exp\{-(\ell(\mathcal{X}, \Gamma) - \ell(\mathcal{X}, \Gamma^\star))\} = P_g(X \not\in \Gamma^\star \triangle \Gamma) \\ + [\exp(-k)(1 - F_{\Gamma^\star}(z)) + \exp(c) F_{\Gamma^\star}(z)] P_g(X \in \Gamma^\star \setminus \Gamma) \\ + [\exp(-c) F_{\Gamma^{\star c}}(z) + \exp(k) - \exp(k) F_{\Gamma^{\star c}}(z)] P_g(X \in \Gamma \setminus \Gamma^\star). \end{aligned}$$

From Equation 6.4.1, see that

$$\kappa := \max\{\exp(-k)(1 - F_{\Gamma^\star}(z)) + \exp(c) F_{\Gamma^\star}(z), \exp(-c) F_{\Gamma^{\star c}}(z) + \exp(k) - \exp(k) F_{\Gamma^{\star c}}(z)\} < 1,$$

so that

$$\begin{aligned} P_{\Gamma^\star} \exp\{-(\ell(\mathcal{X}, \Gamma) - \ell(\mathcal{X}, \Gamma^\star))\} &\le 1 - P_g(X \in \Gamma \triangle \Gamma^\star) + \kappa P_g(X \in \Gamma \triangle \Gamma^\star) \\ &= 1 - (1 - \kappa) P_g(X \in \Gamma \triangle \Gamma^\star). \end{aligned}$$

Then the claim follows, with $\rho = (1 - \kappa)\underline{g} < 1$, since $P_g(X \in \Gamma \triangle \Gamma^\star) \ge \underline{g} \lambda(\Gamma \triangle \Gamma^\star)$.

Next, show that if Equation 6.4.1 holds, then

$$\{\theta : \max\left[R(\gamma) - R(\gamma^\star), V(\ell(\mathcal{X}, \gamma) - \ell(\mathcal{X}, \gamma^\star))\right] \ge C\delta\} \supseteq B_\infty(\gamma^\star; C_0\delta)$$

for some constants $C, C_0, \delta > 0$.

From Proposition 6.3.2 and Lemma 6.7.1

$$\begin{aligned} R(\Gamma) - R(\Gamma^\star) &\leq P_g(X \in \Gamma \triangle \Gamma^\star) \min\{k - kF_{\Gamma^\star}(z) - cF_{\Gamma^{\star c}}(z), cF_{\Gamma^{\star c}}(z) + kF_{\Gamma^{\star c}}(z)\} \\ &\leq \tfrac{1}{2} V_1 \, \overline{g} \, \mathrm{diam}(\Omega) \, \|\gamma - \gamma^\star\|_1 \end{aligned}$$

where $V_1 = V_{c,k,z}$ is the $\min\{\cdots\}$ term in the above display. Further

$$\begin{aligned} V(\ell(\mathcal{X}, \gamma) - \ell(\mathcal{X}, \gamma^\star)) \leq P_g(X \in \Gamma \triangle \Gamma^\star) \max\{&k^2(1 - F_{\Gamma^\star}(z)) + c^2 F_{\Gamma^\star}(z), \\ &k^2(1 - F_{\Gamma^{\star c}}(z)) + c^2 F_{\Gamma^{\star c}}(z)\} \\ \leq\; &V_2 \, \overline{g} \, \mathrm{diam}(\Omega) \, \|\gamma - \gamma^\star\|_1 \end{aligned}$$

where $V_2 = V_{c,k,z}$ is the $\max\{\cdots\}$ term in the above display.

Let $B_\infty(\gamma^\star; r)$ denote the set of regions Γ with boundary functions $\gamma = \partial\Gamma$ that satisfy $\|\gamma - \gamma^\star\|_\infty \leq r$. If $\Gamma \in B_\infty(\gamma^\star; C_0\epsilon_n)$, then we have

$$\|\gamma - \gamma^\star\|_1 \leq 2\pi \overline{g} C_0 \epsilon_n$$

and, therefore, $\max\{R(\Gamma) - R(\Gamma^\star), \, V(\ell(\mathcal{X}, \gamma) - \ell(\mathcal{X}, \gamma^\star))\} \leq C\epsilon_n$, where $C = C_0 \pi \max\{V_1, V_2\} \overline{g}^2 \, \mathrm{diam}(\Omega)$. With Assumption 4.3.2 verified, the claim follows.

CHAPTER 7

DETERMINING THE SCALE PARAMETER

7.1 Introduction

An advantage of Bayesian and other more general Bayesian-like methods that base their inference on a suitable posterior distribution is that uncertainty quantification, in the form of credible regions for the unknown parameters, is readily available. For this uncertainty quantification to be meaningful, it is common to require that the specified credibility level agrees, at least approximately, with the frequentist coverage probability, i.e., that the 95% credibility regions read off from the posterior are approximately 95% confidence regions. In this case, it is said that the posterior credible region is *calibrated.* For well-specified Bayesian models, one often has a Bernstein–von Mises theorem available to justify a calibration claim, but when the model is misspecified in at least one of several possible ways, calibration often fails. For example, (52) derived a Bernstein–von Mises theorem for Bayesian posteriors under model misspecification, and pointed out that, even if concentration target and rate are correct, misspecification can still cause a lack of calibration; see page 362 in their paper and Section 7.2 below. Similarly, the commonly used variational Bayes posteriors (e.g. (45; 41)) often lack the desired calibration property, and correcting this is listed as one of the important open problems in (7).

To address this problem, a scalar tuning parameter (or learning rate) is introduced to the posterior, intended to control the spread of the posterior distribution. Often times, the resulting scaled posterior may be interpreted as a Gibbs posterior as in Equation 2.2.1, but the setup in this chapter can be generalized to other situations not fitting this definition precisely; see Section 7.4 below. Having introduced an extra parameter into the posterior, it is then proposed to select this tuning parameter such that the corresponding posterior credible regions are calibrated in the sense described above, and an algorithm is presented, based on bootstrap and other Monte Carlo techniques, to implement this idea efficiently.

In Section 2.3, alternative methods to select this scale parameter are reviewed. These proposals are reasonable, but they do not provide any guarantees that the uncertainty quantification coming from the corresponding posterior distribution is meaningful. In contrast, the proposal here is designed specifically to make the corresponding posterior credible regions calibrated, at least approximately. The claimed calibration follows immediately from its construction, and the simulations presented in Section 7.4, covering several different models and types of posteriors, demonstrate the effectiveness of the proposed method.

The remainder of this chapter is organized as follows. Section 7.2 explains the intuition behind the proposed approach. The *general posterior calibration* algorithm is presented in Section 7.3 and its basic properties are discussed. Section 7.4 contains several examples, including a Gibbs posterior in quantile regression, a misspecified Bayes posterior in linear regression, and a variational Bayes posterior in a mixture model, and Section 7.5 makes some concluding remarks.

7.2 Problem formulation

Suppose there is data $\mathcal{X}^n = (X_1, \ldots, X_n)$ consisting of i.i.d. observations from a distribution P; here, each X_i could be a vector or even a response–predictor variable pair, i.e., $\mathcal{X}_i = (X_i, Y_i)$. The quantity of interest is a parameter $\theta = \theta(P)$, a feature of the underlying distribution P, taking values in Θ. Consider the following general construction of a posterior distribution for inference on θ.

- Connect data $\mathcal{X}$ to a full set of parameters $\eta \in N$ through either a statistical model for P, as in Bayes or other likelihood-based settings, or a suitable linking function, as in Gibbs or M-estimation settings.
- Introduce a prior Π for the full parameter η, and a scale $\omega > 0$ to weight the information about η in the data with that in the prior.
- Combine the prior, scale, and likelihood/linking function to get a posterior distribution for η.
- Integrate to get the corresponding marginal posterior for θ, denoted by $\Pi_{n,\omega}$.

This general recipe includes both the Bayesian and Gibbs posterior procedure, as well as variational Bayes, as demonstrated in Section 7.4.3. It also covers classical empirical Bayes or other posteriors based on data-dependent priors (e.g. (24; 71; 36)). The one technical requirement necessary is that the posterior $\Pi_{n,\omega}$ be consistent in the sense that it concentrates, asymptotically, on the actual value $\theta(P)$ for each fixed ω. Consistency must be verified case-by-case, but this is standard; see Section 7.4. Given that the posterior $\Pi_{n,\omega}$ is approximately centered around $\theta^\star$, the use of credible regions to quantify uncertainty is reasonable.

The proposed choice of scale is based on calibrating the posterior credible regions to be used for uncertainty quantification. Fix a level $\alpha \in (0, 1)$ and, for concreteness, consider the highest posterior density credible regions defined as

$$C_{\omega,\alpha}(\mathcal{X}^n) = \{\theta : \pi_{n,\omega}(\theta) \geq c_\alpha\}, \tag{7.2.1}$$

where $\pi_{n,\omega}$ is the density function corresponding to the posterior $\Pi_{n,\omega}$, and c_α is a constant chosen so that the $\Pi_{n,\omega}$-probability assigned to $C_{\omega,\alpha}(\mathcal{X}^n)$ is equal to $1-\alpha$. The scale parameter ω controls the spread of the posterior and, thereby, the size of these credible regions. The proposal here is to choose ω so that the credible regions are of the appropriate size to be calibrated, i.e., so that their coverage probability, $P\{C_{\omega,\alpha}(\mathcal{X}^n) \ni \theta(P)\}$, is approximately equal to $1 - \alpha$; see Section 7.3.

To better understand this proposal, recall that, in the classical setting of a well-specified Bayesian model with suitable regularity, the credible region will be calibrated, at least asymptotically, when $\omega = 1$. There has been recent interest in the misspecified case and, in particular, (52) showed that even if a Bernstein–von Mises theorem holds, the posterior credible regions might not be calibrated. Roughly speaking, misspecification affects the shape of the posterior contours, which may be the wrong shape compared to the sampling distribution of the corresponding M-estimator. Varying the scale parameter can provide a conservative solution to this problem: the posterior contours can be stretched enough that they contain a differently-shaped

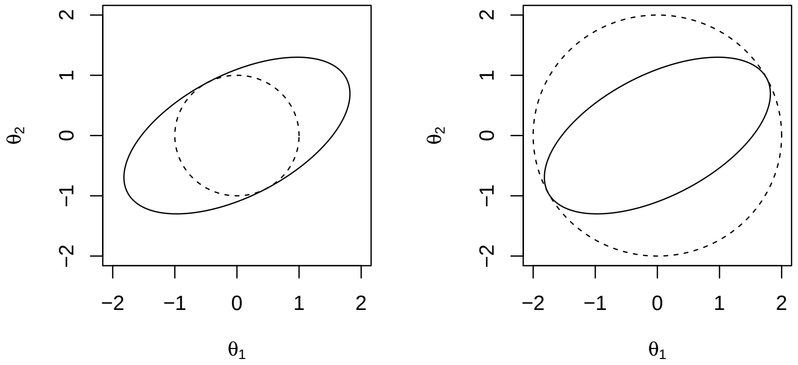


Figure 7.2.1. Contours of the asymptotic distribution of the M-estimator (solid) and those of the asymptotic Gibbs posterior (dashed). Left: $\omega = 1$; Right: $\omega = 1/4$.

but suitably calibrated confidence region; see Figure 7.2.1 for an illustration and Remark 2 below.

7.3 Posterior calibration algorithm

As discussed previously, the goal is to select the scale parameter ω such that the corresponding posterior credible region are calibrated in the sense that the credibility level agrees, at least approximately, with the nominal coverage probability. To this end, for a desired significance level $\alpha \in (0, 1)$, and the preferred credible region $C_{\omega,\alpha}(\mathcal{X}^n)$ as in Equation 7.2.1, define the coverage probability function

$$c_\alpha(\omega \mid P) = P\{C_{\omega,\alpha}(\mathcal{X}^n) \ni \theta(P)\},$$

i.e., the P-probability that the credible region $C_{\omega,\alpha}(\mathcal{X}^n)$ contains the target $\theta(\mathsf{P})$. Then calibration requires that ω be such that

$$c_\alpha(\omega \mid \mathsf{P}) = 1 - \alpha, \tag{7.3.1}$$

i.e., that the $100(1-\alpha)\%$ posterior credible region is also a $100(1-\alpha)\%$ confidence region. Of course, in practice, this equation cannot be solved because P is not known. The approach described below is designed to get around this practical roadblock. Before proceeding, note that solving Equation 7.3.1 is a fixed-n exercise, so the aim is to get exact calibration in finite samples. Asymptotic approximations come into play, however, because P is unknown in real applications, but the numerical illustrations in Section 7.4 demonstrate that it is, in fact, possible to achieve exact calibration, at least in some cases.

To build up the intuition, start by assuming that P is known; later the more realistic case of unknown P is considered. Even in this unrealistic case, it is generally not possible to solve for ω in Equation 7.3.1 explicitly, so numerical methods are required. It is possible to solve Equation 7.3.1 via stochastic approximation (80; 56; 7) by iterating according to the rule

$$\omega^{(t+1)} = \omega^{(t)} + \kappa_t\{\hat{c}_\alpha(\omega^{(t)} \mid \mathsf{P}) - (1-\alpha)\}, \qquad t \geq 0 \tag{7.3.2}$$

where $\hat{c}_\alpha(\omega \mid \mathsf{P})$ is a Monte Carlo approximation to the coverage probability, obtained by simulating new copies of the data $\mathcal{X}^n$ from P, and (κ_t) is a non-stochastic sequence such that $\sum_t \kappa_t = \infty$ and $\sum_t \kappa_t^2 < \infty$. For the numerical results in Section 7.4, $\kappa_t = (t+1)^{-3/4}$ is used.

In this case, if $c_\alpha(\omega \mid P)$ is continuous and monotone decreasing in ω, both very reasonable assumptions, then it follows from the almost supermartingale convergence theorem of (79), that $\omega^{(t)} \to \omega^\star$ P-almost surely, as $t \to \infty$, where $\omega^\star$ is the solution to Equation 7.3.1.

For the realistic case where P is unknown, the proposed approach changes in two ways. First, since it is not possible to sample new copies of $\mathcal{X}^n$ from P, replace simulation from P with simulation from $\mathbb{P}_n$, i.e., instead sample with replacement from the observed data $\mathcal{X}^n$. Second, since $\theta(P)$ is also not known, there is no way to check if a given credible region $C_{\omega,\alpha}(\mathcal{X}^n)$ covers it. Instead, use $\theta(\mathbb{P}_n)$ in place of $\theta(P)$. This results in an empirical version of the coverage probability $c_\alpha(\omega \mid P)$, namely,

$$c_\alpha(\omega \mid \mathbb{P}_n) = \mathbb{P}_n\{C_{\omega,\alpha}(\mathcal{X}^n) \ni \theta(\mathbb{P}_n)\}, \tag{7.3.3}$$

and the proposal is to find ω such that

$$c_\alpha(\omega \mid \mathbb{P}_n) = 1 - \alpha. \tag{7.3.4}$$

In practice, $c_\alpha(\omega \mid \mathbb{P}_n)$ cannot be evaluated either, but bootstrap will provide a Monte Carlo estimator, which is denoted by $\hat{c}_\alpha(\omega \mid \mathbb{P}_n)$. Then, proceed to solve Equation 7.3.1 by using the same stochastic approximation procedure described above for the known-P case. Collectively, these steps to solve this equation make up the *general posterior calibration* (GPC) algorithm. An R code implementation for each of the examples in Section 7.4 is available at `https://github.com/nasyring/GPC`.

Algorithm 1 – General Posterior Calibration.

Fix a convergence tolerance $\epsilon > 0$ and an initial guess $\omega^{(0)}$ of the calibration parameter. Take B bootstrap samples $\tilde{\mathcal{X}}_1^n, \ldots, \tilde{\mathcal{X}}_B^n$ of size n. Set $t = 0$ and do:

1. Construct credible regions $C_{\omega^{(t)},\alpha}(\tilde{\mathcal{X}}_b^n)$ for each $b = 1, \ldots, B$.
2. Evaluate the empirical coverage $\hat{c}_\alpha(\omega^{(t)} \mid \mathbb{P}_n)$ as in Equation 7.3.3.
3. If $\left|\hat{c}_\alpha(\omega^{(t)} \mid \mathbb{P}_n) - (1-\alpha)\right| < \epsilon$, then stop and return $\omega^{(t)}$ as the output; otherwise, update $\omega^{(t)}$ to $\omega^{(t+1)}$ according to Equation 7.3.2, set $t \leftarrow t + 1$, and go back to Step 1.

Remark 1 In most applications, the credible regions $C_{\omega,\alpha}(\mathcal{X}^n)$ will not be available in closed form, so posterior sampling will be needed in such cases. But despite having several moving parts—bootstrap, MCMC, and stochastic approximation—the proposed GPC algorithm is relatively computationally inexpensive. For example, in the quantile regression problem in Section 7.4.1, with a two-dimensional parameter, sample size $n = 100$, $B = 200$ bootstrap samples, and $M = 2000$ posterior samples, the algorithm took less than 10 seconds to converge on a Windows desktop computer with a 4.0 GHz Intel Core i7 processor. It seems reasonable that minimal extra computational investment is a fair trade for calibrated posterior credible regions.

Remark 2 The workhorses of the GPC algorithm, namely, bootstrap, stochastic approximation, and MCMC, are widely used and, on their own, theoretically sound. But having them all working together in tandem makes a general theoretical analysis of the GPC algorithm very challenging. Here, some technical insight is provided as to why the algorithm works, leaving a complete theoretical analysis for future research.

Let the interest parameter be defined via a risk function $R(\theta) = E_P(\ell(\mathcal{X}, \theta))$, so that $\theta(P) = \arg\min R(\theta)$. In this case, a Gibbs posterior $\Pi_{n,\omega}$ is defined as in Equation 2.2.1, where $R_n(\theta)$ is the empirical risk. Under suitable regularity conditions, the Gibbs posterior will resemble a normal distribution, centered at $\theta(\mathbb{P}_n) = \arg\min R_n(\theta)$, with asymptotic covariance matrix $\omega^{-1}\Sigma_n$, where $\Sigma_n = (nV_{\theta(\mathbb{P}_n)})^{-1}$ and V_θ is the second derivative matrix of $R(\theta)$. So, the credible region $C_{\omega,\alpha}(\mathcal{X}^n)$ will look roughly like

$$\{\theta : \omega(\theta - \theta(\mathbb{P}_n))^\top \Sigma_n^{-1}(\theta - \theta(\mathbb{P}_n)) \le \xi_\alpha\},$$

where ξ_α is the appropriate chi-square quantile. On the other hand, the asymptotic covariance matrix of $\theta(\mathbb{P}_n)$ is given by $\Psi_n = n^{-1}V_{\theta(P)}^{-1} M V_{\theta(P)}^{-1}$, where $M = E_P(\dot{\ell}(\mathcal{X}, \theta)\dot{\ell}(\mathcal{X}, \theta)^\top)$ and $\dot{\ell}(\mathcal{X}, \theta)$ is the derivative of $\theta \mapsto \ell(\mathcal{X}, \theta)$, so an asymptotic confidence region is

$$\{\theta : (\theta - \theta(\mathbb{P}_n))^\top \Psi_n^{-1}(\theta - \theta(\mathbb{P}_n)) \le \xi_\alpha\}.$$

In general, $\omega\Sigma_n^{-1}$ and Ψ_n^{-1} are different, so the credible region has a different shape than the confidence region and, therefore, may not be calibrated. But the GPC algorithm will take ω roughly equal to the smallest eigenvalue of $\Psi_n^{-1/2}\Sigma_n\Psi_n^{-1/2}$, so that the credible region contains the aforementioned confidence region, making the latter conservatively calibrated; exact calibration is possible only if $\Psi_n = c\Sigma_n$ for some scalar $c > 0$. Since $\Psi_n^{-1/2}\Sigma_n\Psi_n^{-1/2}$ has a limit as $n \to \infty$, then it can be expected that the GPC solution to Equation 7.3.3 will converge to the smallest eigenvalue of that limiting matrix, hence asymptotic calibration.

For a quick proof-of-concept, suppose the data $\mathcal{X}^n$ are iid P and the population mean $\theta = \int t \, d\mathsf{P}(t)$ is the quantity of interest. Here, take $\mathsf{P} = \mathsf{N}(0, 1)$, so that $\theta = 0$. Consider three posterior distributions: a Bayes model using the correct normal likelihood; a Gibbs posterior using $R_n(\theta) = \sum_{i=1}^{n} (\mathcal{X}_i - \theta)^2$, and a misspecified Bayesian posterior with a Laplace likelihood. For the well-specified Bayes and the Gibbs posteriors, it is expected that the GPC algorithm will select $\omega \approx 1$ and $\omega \approx 0.5$, respectively; for the Laplace model, based on the V_θ and M calculations in Remark 2, $\omega \approx 0.64$. Figure 7.3 plots the mean trajectories of the ω values obtained from the GPC algorithm, with error bars, as a function of n, and the results largely follow the above expectations.

7.4 Applications

7.4.1 Quantile regression

Recall the quantile regression model defined in Equation 2.1.1. In quantile regression, for fixed $\tau \in (0, 1)$ and data $\mathcal{X} = (Y, X)$, the interest is in the τ^{th} quantile of the response $Y \in \mathbb{R}$, given the covariates $X \in \mathbb{R}^{p+1}$, where dimension $p + 1$ represents an intercept and p covariates. In this formula, the vector θ depends on τ but, for notational simplicity, this dependence is omitted. This model specifies no parametric form for the conditional distribution of Y given X. Inference on the quantile regression coefficient θ may be carried out using asymptotic approximations ((54), Theorem 4.1) or by using the bootstrap (39). A Bayesian approach would also be attractive, but no distributional form for the conditional distribution is given in (Equation 2.1.1), hence no likelihood. A workaround that has been considered by several authors (e.g. (111; 90; 89)) is to use a (misspecified) asymmetric Laplace likelihood. This

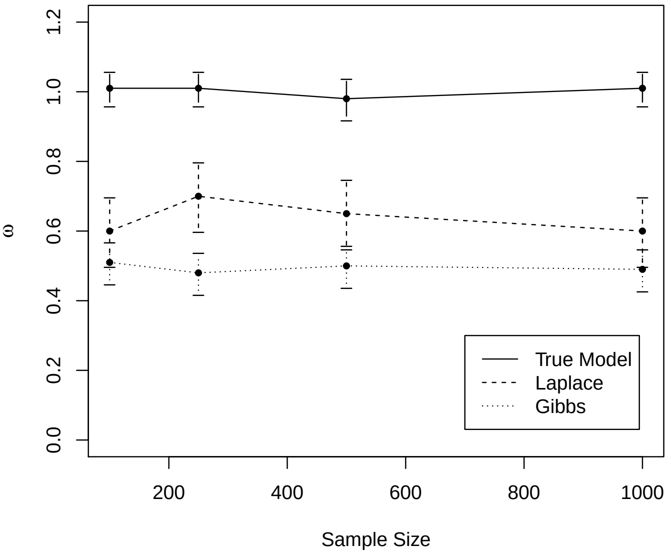


Figure 7.3.1. Mean choice of ω over 100 simulated standard normal data sets of sizes 100, 250, 500, and 1000 using the true likelihood, a Gibbs model, and a Laplace likelihood. Vertical bars represent two standard deviations from the mean.

corresponds to a Gibbs model Equation 2.2.1 using the empirical risk given in Equation 2.1.2 based on the usual check-loss function.

It follows from (52) that the Gibbs posterior based on Equation 2.1.2 satisfies a Bernstein–von Mises theorem. Despite the desirable convergence result, the variance mismatch discussed in Section 7.2 causes the credible regions to be too large and over-cover, a sign of inefficiency. On the other hand, the GPC algorithm calibrates the intervals exactly, for all n, without loss of efficiency in terms of interval lengths.

To demonstrate this, a simulation example presented in (110) is revisited for comparison. For $\tau = 0.5$, the model they consider is

$$Y_i = \theta_0 + \theta_1 X_i + e_i, \quad i = 1, \dots, n,$$

where $\theta_0 = 2$, $\theta_1 = 1$, $e_i \overset{\text{i.i.d.}}{\sim} \mathsf{N}(0, 4)$, and $X_i \overset{\text{i.i.d.}}{\sim} \mathsf{ChiSq}(2) - 2$. For this model, the authors showed numerically that their proposed Bayesian empirical likelihood approach ("BEL.s") produced credible intervals with approximate coverage near the nominal 95% level. Moreover, compared to the Bayesian method with misspecified asymmetric Laplace likelihood ("BDL") or, equivalently, the Gibbs posterior with ω chosen by averaging residuals, their method is shown to be more efficient in terms of interval length. The results for these methods are presented in Table 7.4.1, along with the results from the posterior intervals scaled by the algorithm.

There are two key observations to be made. First, the GPC method calibrates the credible intervals to have exact 95% coverage across the range of n, while the other methods tend to

		Coverage Probability					Average Length				
n		BEL.s	BDL	Normal	$\omega \equiv 0.8$	GPC	BEL.s	BDL	Normal	$\omega \equiv 0.8$	GPC
100	θ_0	0.97	0.98	0.95	0.96	0.95	1.06	1.11	1.00	1.00	0.91
	θ_1	0.98	0.98	0.98	0.98	0.95	0.58	0.58	0.55	0.52	0.47
400	θ_0	0.95	0.98	0.95	0.95	0.95	0.50	0.55	0.50	0.49	0.46
	θ_1	0.97	0.98	0.97	0.96	0.95	0.26	0.28	0.25	0.25	0.23
1600	θ_0	0.96	0.97	0.96	0.95	0.95	0.25	0.28	0.25	0.24	0.23
	θ_1	0.96	0.98	0.96	0.96	0.95	0.13	0.14	0.12	0.12	0.11

TABLE 7.4.1

COMPARISON OF 95% POSTERIOR CREDIBLE INTERVALS OF THE MEDIAN REGRESSION PARAMETERS FROM FIVE METHODS: BEL.S; BDL; NORMAL; THE CONFIDENCE INTERVAL COMPUTED USING THE ASYMPTOTIC NORMALITY OF THE M-ESTIMATOR; $\omega \equiv 0.8$, THE SCALED POSTERIOR WITH ω FIXED EQUAL TO 0.8; AND GPC. COVERAGE PROBABILITY AND AVERAGE INTERVAL LENGTHS ARE COMPUTED OVER 5000 SIMULATED DATA SETS FOR THE GPC METHOD, NORMAL INTERVALS, AND FIXED-ω INTERVALS. RESULTS FOR BEL.S AND BDL ARE TAKEN FROM (110) AND WERE CALCULATED FROM 1000 SIMULATED DATA SETS.

over-cover. Second, GPC credible intervals tend to be shorter than those of the other methods, especially for $n = 100$. All three methods have a $n^{-1/2}$ convergence rate so, for large n, one cannot expect to see substantial differences between the various methods. Therefore, the small-n case should be the most important and, at least in this case, the credible intervals calibrated using the GPC algorithm are clearly the best.

Finally, considering that in smooth models ω should account for the difference in asymptotic variance between the posterior and the M-estimator, it is reasonable to ask if a calibration algorithm is needed at all, i.e., can approximate calibration be obtained with a fixed value of ω based on these asymptotic variances? A comparison of the asymptotic variance of the posterior with that of the M-estimator shows that $0.80\Sigma_n^{-1} \approx \Psi_n^{-1}$; therefore, take $\omega \equiv 0.80$ in an attempt to calibrate posterior credible intervals with a fixed scaling. Table 7.4.1 shows that the GPC algorithm is still better than using a fixed scale based on asymptotic normality, especially at smaller sample sizes where the normal approximation is less justifiable.

7.4.2 Linear regression

Consider the usual multiple linear regression model for data $\mathcal{X}_i = (Y_i, X_i) \in \mathbb{R} \times \mathbb{R}^p$

$$Y_i = \beta_0 + X_i^\top \beta + \sigma e_i, \quad i = 1, \ldots, n, \tag{7.4.1}$$

where $\beta \in \mathbb{R}^p$ is the vector of slope coefficients, $\sigma > 0$ is an unknown scale parameter, and $e_1, \ldots, e_n$ are assumed to be i.i.d. $N(0, 1)$. Suppose, however, that the constant error variance assumption is violated, in particular, $e_i \sim N(0, \sigma^2 \|X_i\|)$, $i = 1, \ldots, n$, independent. This choice of predictor-dependent variance is a less-stylized version of that in (34). The proposed model is, therefore, misspecified, but the goal is still to obtain calibrated inference on $\theta = (\beta_0, \beta)$.

The Jeffreys prior is a reasonable default choice with density $\pi(\eta) \propto (\sigma^2)^{-3/2}$ (40) for the full parameter $\eta = (\theta, \sigma^2)$. Since this prior is probability-matching for the location-scale model (e.g. (16)), it can be expected that the posterior credible intervals are approximately calibrated for this linear regression. However, for a misspecified model, calibration might fail; in fact, as shown in Table 7.4.2, the credible intervals are too narrow and tend to undercover.

To investigate the performance of the proposed GPC method compared to several others, a simulation study was done. The simulated data sets have $n = 50$ observations. Each $X_i \in$

$\mathbb{R}^3$ is multivariate normal with zero mean and unit variance for each element, and pairwise correlation 0.5 for X_{i1} and X_{i2} and zero otherwise. To sample Y_i parameter values of $\beta_0 = 0$, $\beta = (1, 2, -1)^\top$, and $\sigma = 1$ were used. Although the error variance contains $\|X_i\|$, the regular tests for constant variance do not detect the heteroscedasticity. Table 7.4.2 shows the estimated coverage probability and mean lengths of several posterior credible intervals for the components of θ. Besides those scaled by the GPC algorithm, a misspecified Bayes approach that fixes $\omega \equiv 1$ was considered, and posteriors with scale ω chosen by the method in (38) and the *SafeBayes* method in (34), Algorithm 1 were compared. The results in Table 7.4.2 show that for this example SafeBayes performs similarly to GPC, while the method in (38) does not improve upon the misspecified Bayesian model in terms of calibration.

Figure 7.4.1 shows a boxplot comparison of the scale parameters chosen by the three posterior scaling methods for the misspecified Bayesian posterior. The GPC algorithm, along with the SafeBayes method, tends to produce smaller values of ω than the method of Holmes and Walker. Small values of ω mean higher posterior variance and wider credible intervals, which explains these method's improvement in calibration. While both GPC and SafeBayes pick $\omega \approx 0.8$ on average, the distribution of ω is much more concentrated using the GPC algorithm.

7.4.3 Variational inference for a normal mixture model

Variational inference offers a competing method to MCMC for approximating the posterior distribution. This approach specifies a family of distributions—often a normal family—as candidate posteriors and then chooses the parameters of that family to minimize the Kullback–Leibler divergence from the true posterior. The variational posterior is simple by construction

		β_0	β_1	β_2	β_3
Misspecified Bayes	coverage	0.94	0.89	0.88	0.87
	length	0.99(0.15)	1.16(0.20)	1.16(0.20)	1.01(0.17)
GPC	coverage	0.98	0.94	0.94	0.93
	length	1.17(0.18)	1.36(0.23)	1.36(0.24)	1.18(0.20)
SafeBayes	coverage	0.96	0.93	0.94	0.92
	length	1.19(0.26)	1.40(0.31)	1.39(0.33)	1.21(0.28)
Holmes and Walker	coverage	0.91	0.84	0.80	0.82
	length	0.87(0.18)	1.01(0.22)	1.01(0.22)	0.87(0.18)

TABLE 7.4.2

EMPIRICAL COVERAGE PROBABILITIES OF 95% CREDIBLE INTERVALS AND AVERAGE INTERVAL LENGTHS (AND STANDARD DEVIATIONS) CALCULATED USING 5000 SIMULATIONS FROM THE MODEL DESCRIBED IN SECTION 7.4.2.

and, if carefully chosen, will be consistent (e.g. (103)), but as noted in (7), misspecification causes the variational posterior variance to be too small.

As an example, consider the normal mixture model presented in (7), i.e., $Y_1, \ldots, Y_n$ are iid observations from the mixture model

$$\sum_{k=1}^{K} \pi_k N(\mu_k, \sigma_k^2). \tag{7.4.2}$$

The full parameter η consists of the mixture weights $(\pi_1, \ldots, \pi_K)$, means $(\mu_1, \ldots, \mu_K)$, and variances $(\sigma_1^2, \ldots, \sigma_K^2)$, but in what follows consider inference only on the means. A variational posterior can be constructed for η following Algorithm 2 in (7), which approximates the posterior by a multivariate normal. The additional scale factor ω in the modified variational posterior $\Pi_{n,\omega}$ only adjusts the overall scale of this multivariate normal. Therefore, if $m_1, \ldots, m_K$ and

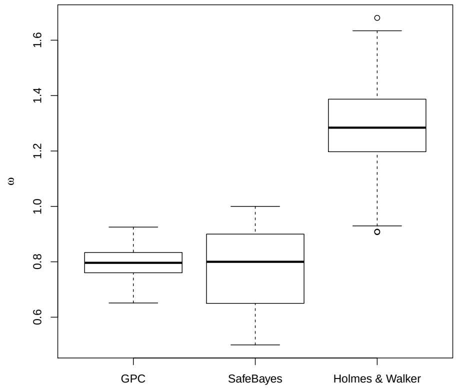


Figure 7.4.1. Boxplots of ω for the model described in Section 7.4.2 using GPC, SafeBayes (34), and the method in (38) over 5000 simulated data sets.

$\nu_1, \ldots, \nu_K$ are the means and variances, respectively, of this variational posterior for the mixture means $\mu_1, \ldots, \mu_K$, then the corresponding ω-scaled variational posterior $100(1-\alpha)\%$ credible intervals are of the form

$$\mu_k \pm z^\star_{\alpha/2}\, \omega\, \nu_k^{1/2}, \quad k = 1, \ldots, K.$$

It is straightforward to incorporate this variational posterior setup into the GPC algorithm; the computational investment is in carrying out the optimization needed for the variational approximation at each bootstrap step, but then the credible intervals are available in closed-form so no posterior sampling is needed.

The claim is that the GPC algorithm will properly scale the variational posterior, calibrating the corresponding credible intervals, correcting the under-estimation of variance noted in

		μ_1	μ_2
GPC	coverage	0.96	0.96
	length	0.67 (0.08)	0.67 (0.08)
VI	coverage	0.92	0.92
	length	0.55 (0.03)	0.55 (0.03)

TABLE 7.4.3

EMPIRICAL COVERAGE PROBABILITY AND AVERAGE LENGTH (AND STANDARD DEVIATION) OF THE CREDIBLE INTERVALS FOR (μ_1, μ_2) BASED ON THE GPC ALGORITHM AND THE VARIATIONAL POSTERIOR (VI) IN (7) OVER 5000 SIMULATED DATA SETS FROM THE MIXTURE MODEL IN Equation 7.4.2.

(7). To demonstrate this, a simple simulation study is carried out. Take $K = 2$, $\pi_1 = \pi_2 = 1/2$, $(\mu_1, \mu_2) = (-2, 2)$, and $\sigma_1 = \sigma_2 = 1$. Table 7.4.3 shows the empirical coverage probabilities and mean lengths of the 95% credible intervals based on Algorithm 2 in (7) and the GPC algorithm. Apparently, the GPC algorithm corrects the underestimated variance of the variational posterior, producing credible intervals that are slightly conservative.

7.5 Discussion

The sensitivity of Bayesian credible sets to the posited probability model makes obtaining calibrated inference a challenging problem. The linear regression example demonstrates this sensitivity when the model is taken for granted. However, misspecification can happen in a variety of settings, and not always unintentionally. In quantile regression, the model is determined by a risk function rather than a likelihood, making traditional Bayesian inference using the true likelihood elusive. And, other times, computational considerations make variational posteriors an attractive alternative to a fully Bayesian analysis. The GPC algorithm may provide a solution in all of these settings by correcting model misspecification to produce, at least approximately, calibrated inferences.

Although the focus in this paper is on misspecified models, it may still be desirable to apply the GPC algorithm even when the true likelihood is used. The reason is that the GPC algorithm can aid in producing calibrated inferences for the given sample size, regardless of the prior distribution used. This facilitates the use of informative priors, if available, instead of default priors, while still gaining the desired calibration property.

Finally, while it is clear that the GPC algorithm produces approximately calibrated credible sets, a detailed theoretical study is needed. The techniques used herein—stochastic approximation, MCMC, and bootstrap—each are theoretically sound on their own, but very complicated when used in tandem. Further work in this direction may help provide guidance, but the lack of completely rigorous theory does not take away from the encouraging examples shown throughout the paper.

CHAPTER 8

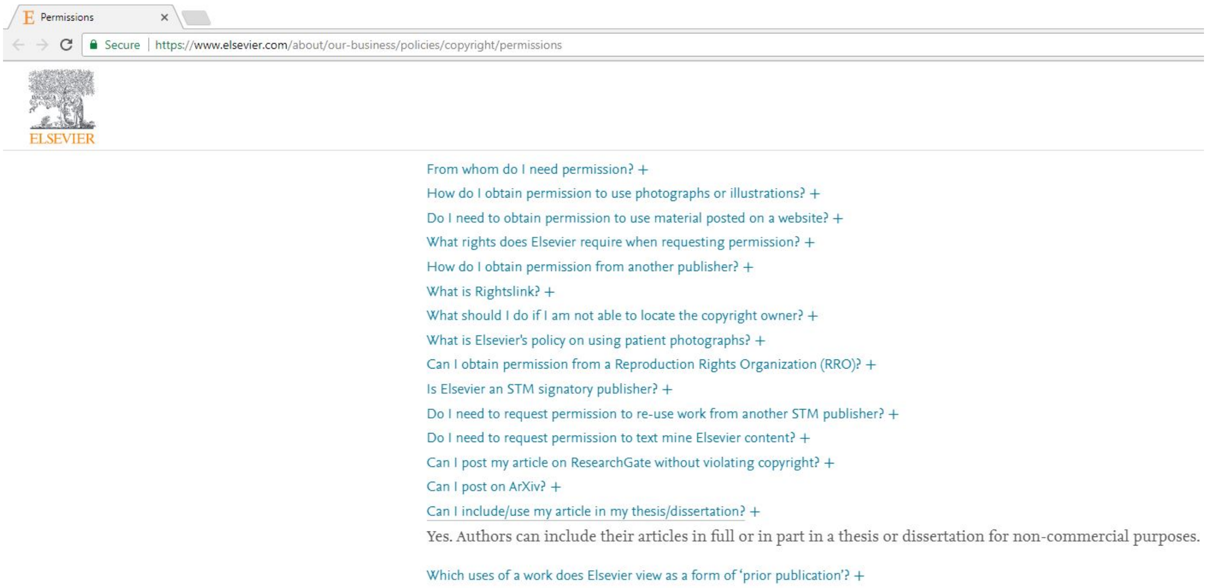
CONCLUSION

The literature on Gibbs models is still somewhat limited, but the motivation for alternative methods for posterior inference both for robustness to the underlying probability model and for good predictive properties is well-developed. This dissertation has sought to further the cause of these techniques through convincing applications and new theoretical foundations. Two general theorems, detailed in Chapter 4, provide techniques for calculating posterior convergence rates for Gibbs models in a wide variety of settings, including infinite-dimensional problems. Two in-depth applications, covered in Chapters 5 and 6, utilize these results and also provide examples of the utility of Gibbs models.

Challenges remain in this field. Although Chapter 7 provides an algorithm for choosing an advantageous scale for the Gibbs posterior, the technique is computationally demanding for parameters with more than a few dimensions. It remains to be seen if an existing technique or a new one can provide a method for selecting a meaningful scale for general models. A robustness argument is made in Chapter 1 to motivate the Gibbs model and it is again evident in both applications, but it is not clear how the data analyst should choose given multiple candidate linking functions, one or more of which may be likelihoods. In some inference problems, like the MCID problem, the parameter is defined via a loss function, which makes a Gibbs model a natural choice. But, in other cases, like quantile regression and the image boundary problem, even when the parameter can be understood to minimize a certain loss function, probability models are the default approach for most practitioners. One challenge to popularizing Gibbs models is to convince data analysts to think in terms of linking data and parameter more generally, not necessarily via a probability model.

There are several new and exciting applications for Gibbs posteriors to be considered, with potential for new theoretical advances as well. The MCID problem has already been extended in the M-estimation setting in (37). The more general approach consists of parametrizing the MCID not as a scalar, but as a function of covariates. Then, each patient can possess their own MCID depending on their own characteristics. The resulting MCID function is referred to as the personalized MCID, and estimation of the personalized MCID using M-estimation has already been studied. A Gibbs posterior approach could certainly be applied to the personalized MCID, and many of the techniques used in Section 4.3 and Chapter 6 could be helpful. Another potential new application of Gibbs posteriors is in modeling the Lévy density, which describes a jump-diffusion Lévy process, with applications in finance. An M-estimation method for estimating the Lévy density is described in (21).

APPENDICES



https://www.elsevier.com/about/our-business/policies/copyright/permissions

From whom do I need permission? +

How do I obtain permission to use photographs or illustrations? +

Do I need to obtain permission to use material posted on a website? +

What rights does Elsevier require when requesting permission? +

How do I obtain permission from another publisher? +

What is Rightslink? +

What should I do if I am not able to locate the copyright owner? +

What is Elsevier's policy on using patient photographs? +

Can I obtain permission from a Reproduction Rights Organization (RRO)? +

Is Elsevier an STM signatory publisher? +

Do I need to request permission to re-use work from another STM publisher? +

Do I need to request permission to text mine Elsevier content? +

Can I post my article on ResearchGate without violating copyright? +

Can I post on ArXiv? +

Can I include/use my article in my thesis/dissertation? +

Yes. Authors can include their articles in full or in part in a thesis or dissertation for non-commercial purposes.

Which uses of a work does Elsevier view as a form of 'prior publication'? +

Figure .0.1. Elsevier grants permission to the author to reproduce previously published material in this dissertation.

CITED LITERATURE

1. Agresti, A.: Categorical Data Analysis. New York, Wiley-Interscience [John Wiley & Sons], second edition, 2002.

2. Alquier, P., Ridgway, J., and Chopin, N.: On the properties of variational approximations of Gibbs posteriors. J. Mach. Learn. Res., 17:1–41, 2016.

3. Anam, S., Uchino, E., and Suetake, N.: Image boundary detection using the modified level set method and a diffusion filter. Procedia Comput. Sci., 22:192–200, 2013.

4. Barron, A. and Cover, T.: Minimum complexity density estimation. IEEE Trans. Inf. Theory, 37:1034–1054, 1991.

5. Berger, J. O., Bernardo, J. M., and Sun, D.: The formal definition of reference priors. Ann. Stat., 37(2):905–938, 2009.

6. Bissiri, P., Holmes, C., and Walker, S.: A general framework for updating belief distributions. J. R. Stat. Soc. Ser. B Stat. Methodol., 78(5):1103–1130, 2016.

7. Blei, D. M., Kucukelbir, A., and McAuliffe, J. D.: Variational Inference: A Review for Statisticians. To appear in: *J. Amer. Statist. Assoc*, 2017.

8. Buchinsky, M. and Hahn, J.: An alternative estimator for the censored regression model. Econometrica, 66:653–671, 1998.

9. Bunke, O. and Milhaud, X.: Asymptotic behavior of Bayes estimates under possibly incorrect models. Ann. Statist., 26(2):617–644, 1998.

10. Casella, G. and Berger, R. L.: Statistical Inference. New York, Duxbury, 2002.

11. Catoni, O.: Pac-Bayesian Supervised Classification: The Thermodynamics of Statistical Learning, volume 56. Institute of Mathematical Statistics, 1997.

12. Catoni, O.: A PAB Bayesian approach to adaptive classification. Preprint 840, 2003. Laboratoire de Probabilités et Modèles Aléatoires, Université Paris 6, Paris.

13. Chen, W., Silverman, D. H., Sibylle, D., Czernin, J., Kamdar, N., Pope, W., Satyamurthy, N., Schiepers, C., and Cloughesy, T.: F-FDOPA PET imaging of brain tumors: Comparison study with F-FDG PET evaluation of diagnostic accuracy. Journal of Nuclear Medicine, 47(6):904–911, 2006.

14. Chernozhukov, V. and Hong, H.: An MCMC approach to classical estimation. J. Econom., 115:293–346, 2003.

15. Choudhuri, N., Ghosal, S., and Roy, A.: Nonparametric binary regression using a Gaussian process prior. Stat. Methodol., 4:227–243, 2007.

16. Datta, G. and Mukerjee, R.: Probability Matching Priors: Higher Order Asymptotics. New York, Springer, 2004.

17. De Blasi, P. and Walker, S. G.: Bayesian asymptotics with misspecified models. Statist. Sinica, 23:169–187, 2013.

18. de Finetti, B.: La prévision: ses lois logiques, ses sources subjectives. Annales de l'Institut Henri Poincaré, pages 1–68, 1937.

19. Denison, D., Mallick, B., and Smith, A.: Automatic Bayesian curve fitting. J. R. Stat. Soc. Ser. B Stat. Methodol., 60(2):333–350, 1998.

20. DiMatteo, I., Genovese, C., and Kass, R.: Bayesian curve-fitting with free-knot splines. Biometrika, 88(4):1055–1071, 2001.

21. Figueroa-López, J. E.: Nonparametric estimation for lévy models based on discrete-sampling. In IMS Lecture Notes-Monograph Series (LNMS). Optimality: The Third Erich L. Lehmann Symposium, ed. J. Rojo, volume 57, pages 117–146. 2009.

22. Fitzpatrick, M., Preisser, E., Porter, A., Elkinton, J., Waller, L., Carlin, B., and Ellison, A.: Ecological boundary detection using Bayesian areal wombling. Ecology, 91(12):3448–3455, 2010.

23. Fraser, D. A. S.: Is bayes posterior just quick and dirty confidence? Statist. Sci., 26(3):299–316, 2011.

24. Fraser, D. A. S., Reid, N., Marras, E., and Yi, G. Y.: Default priors for Bayesian and frequentist inference. J. R. Stat. Soc. Ser. B Stat. Methodol., 72(5):631–654, 2010.

25. Geman, S. and Geman, D.: Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-6(6):721–741, 1984.

26. Ghosh, J. K., Delampady, M., and Samanta, T.: An Introduction to Bayesian Analysis. New York, Springer, 2006.

27. Ghosh, J. K. and Ramamoorthi, R. V.: Bayesian Nonparametrics. New York, Springer, 2003.

28. Ghosh, J. K. and Ramamoorthi, R. V.: Bayesian Nonparametrics. New York, Springer-Verlag, 2003.

29. Ghosh, M.: Objective priors: an introduction for frequentists. Statist. Sci., 26(2):187–202, 2011.

30. Gibbs, J.: Elementary Principles in Statistical Mechanics Developed with Especial Reference to the Rational Foundation of Thermodynamics.. New York, Charles Scribner's Sons, 1902.

31. Gleser, L. J. and Hwang, J. T.: The nonexistence of 100(1)diameter in errors-in-variables and related models. Ann. Statis., 15(4):1351–1362, 1987.

32. Green, P.: Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika, 82(4):711–732, 1995.

33. Grünwald, P. and Van Ommen, T.: Inconsistency of Bayesian inference for misspecified linear models, and a proposal for repairing it. Unpublished manuscript, `arXiv:1412.3730`, 2014.

34. Grünwald, P. and Van Ommen, T.: Inconsistency of Bayesian inference for misspecified linear models, and a proposal for repairing it. Unpublished manuscript, `arXiv:1412.3730`, 2016.

35. Gu, K., Pati, D., and Dunson, D.: Bayesian multiscale modeling of closed curves in point clouds. JASA, 109(508):1481–1494, 2015.

36. Hannig, J., Iyer, H., Lai, R., and Lee, T.: Generalized fiducial inference: A review and new results. J. Amer. Statist. Assoc., 111:1346–1361, 2016.

37. Hedayat, S., Wang, J., and Xu, T.: Minimum clinically important difference in medical studies. Biometrics, 71:33–41, 2015.

38. Holmes, C. and Walker, S. G.: Assigning a value to a power likelihood in a Bayesian model. Biometrika, 104:497–503, 2017.

39. Horowitz, J.: Bootstrap methods for median regression models. Econometrica, 66:1327–1351, 1998.

40. Ibrahim, J. G. and Laud, P. W.: On Bayesian analysis of generalized linear models using Jeffreys's prior. J. Amer. Statist. Assoc., 86(416):981–986, 1991.

41. Jaakola, T. S. and Jordan, M. I.: A variational approach to Bayesian logistic regression models and their extensions. Sixth International Workshop on Artificial Intelligence and Statistics, 82, 1997.

42. Jacobson, N. and Truax, P.: Clinical significance: a statistical approach to defining a meaningful change in psychotherapy research. J. Consult. Clin. Psych., 59:12–19, 1991.

43. Jaescheke, R., Signer, J., and Guyatt, G.: Measurement of health status: ascertaining the minimum clinically important difference. Control. Clin. Trials, 10:407–415, 1989.

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

45. Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K.: An introduction to variational methods for graphical models. Mach. Learn., 37:183–233, 1999.

46. Kass, R. E. and Raftery, A. E.: Bayes factors. JASA, 90(430), 1995.

47. Kato, K.: Quasi-Bayesian analysis of nonparametric instrumental variables models. Annals of Statistics, 41(5):2359–2390, 2013.

48. Kaul, S. and Diamond, G. A.: Trial and error: How to avoid commonly encountered limitations of published clinical trials. J. Am. Coll. Cardiol., 55(5):415–427, 2010.

49. Keener, R. W.: Theoretical Statistics: Topics for a Core Course. New York, Springer, 2010.

50. Kelly, G. E.: The median lethal dose—design and estimation. The Statistician, 50(1):41–50, 2001.

51. Kleijn, B. J. K. and van der Vaart, A. W.: Misspecification in infinite-dimensional Bayesian statistics. Ann. Statist., 34(2):837–877, 2006.

52. Kleijn, B. and van der Vaart, A.: The Bernstein-Von-Mises theorem under misspecification. Electron. J. Stat., 6:354–381, 2012.

53. Koenker, R.: Quantile Regression. Cambridge University Press, Cambridge, 2005.

54. Koenker, R.: Quantile Regression. Cambridge University Press, Cambridge, 2005.

55. Korostelev, A. P. and Tsybakov, A. B.: Lecture Notes in Statistics: Minimax Theory of Image Re-construction, volume 82. New York, Springer, 1993.

56. Kushner, H. J. and Yin, G. G.: Stochastic approximation and recursive algorithms and applications. New York, Springer-Verlag, second edition, 2003.

57. Langford, J.: Tutorial on practical prediction theory for classification. J. Mach. Learn. Res., 6:273–306, 2005.

58. Lee, J. and MacEachern, S. N.: Consistency of Bayes estimators without the assumption that the model is correct. J. Statist. Plann. Inference, 141(2):748–757, 2011.

59. Li, C., Jiang, W., and Tanner, M.: General oracle inequalities for Gibbs posterior with application to ranking. JMLR: Workshop and Conference Proceedings, 30:1–10, 2013.

60. Li, C., Xu, C., Gui, C., and Fox, M.: Distance regularized level set evolution and its application to image segmentation. IEEE Trans. Image Process., 19(12):3243–3254, 2010.

61. Li, M. and Ghosal, S.: Bayesian detection of image boundaries. 2015. To appear in *Ann. Statist.* , available at `arXiv:1508.05847`.

62. Liang, S., Banerjee, S., and Carlin, C.: Bayesian wombling for spatial point processes. Biometrics, 65:1243–1253, 2009.

63. Liu, C. and Martin, R.: Inferential Models: Reasoning with Uncertainty. Monographs in Statistics and Applied Probability Series. Chapman & Hall, 2015.

64. Lu, H. and Carlin, B.: Bayesian areal wombling for geographical boundary analysis. Geogr. Anal., 37:265–285, 2004.

65. Ma, H. and Carlin, B.: Bayesian multivariate areal wombling for multiple disease boundary analysis. Bayesian Anal., 2(2):281–302, 2007.

66. Maini, R. and Aggarwal, H.: Study and comparison of various image edge detection techniques. IJIP, 3(1):1–11, 2009.

67. Mammen, E. and Tsybakov, A.: Asymptotic minimax recovery of sets with smooth boundaries. Ann. Statist., 23:502–524, 1995.

68. Mammen, E. and Tsybakov, A.: Asymptotical minimax recovery of sets with smooth boundaries. Ann. Stat., 23:502–524, 1995.

69. Martin, D., Fowlkes, C., and Malik, J.: Learning to detect natural image boundaries using local brightness, color, and texture cues. IEEE Trans. Pattern Anal. Mach. Intell., 26(5):530–549, 2004.

70. Martin, R. G. and Liu, C.: Inferential models: resoning with uncertainty. New York, Chapman and Hall/CRC Press, 2015.

71. Martin, R. and Walker, S. G.: Asymptotically minimax empirical Bayes estimation of a sparse normal mean vector. Electron. J. Stat., 8(2):2188–2206, 2014.

72. McAllester, D. A.: Some PAC-Bayesian theorems. Mach. Learn., 37:355–363, 1999.

73. McAllester, D.: PAC-Bayesian model averaging. COLT '99, pages 164–170, 1999.

74. Nadaraya, E. A.: On estimating regression. Theory Probab. Appl, 9(1):141–142, 1964.

75. O'Hagan, A.: Fractional bayes factors for model comparison. J. R. Statis. Soc B, 57(1):99–138, 1995.

76. Polson, N., Scott, J., and Windle, J.: Bayesian inference for logistic models using Pólya-gamma latent variables. J. Amer. Statist. Assoc., 108(504):1339–1349, 2013.

77. Ramamoorthi, R. V., Sriram, K., and Martin, R.: On posterior concentration in misspecified models. Bayesian Analysis, 10(4):759–789, 2015.

78. Ridgway, J.: PACVB: Variational Bayes (VB) Approximation Of Gibbs Posteriors With Hinge Losses, 2013. R package version 1.1.1.

79. Robbins, H. and Siegmund, D.: A convergence theorem for non-negative almost supermartingales and some applications. In Optimizing Methods in Statistics, ed. J. S. Rustagi, pages 233–258. New York, Academic Press, 1971.

80. Robbins, H. and Monro, S.: A stochastic approximation method. Ann. Math. Stat., 22:400–407, 1951.

81. Savage, L. J.: The Foundations of Statistics, 2nd ed.. New York, Dover, 1972.

82. Schumaker, L.: Spline Functions Basic Theory. New York, Wiley, 2007.

83. Schumaker, L.: Spline Functions Basic Theory. New York, Wiley, 2007.

84. Shawe-Taylor, J. and Williamson, R.: A PAC analysis of a Bayesian estimator. New York, Association for Computing Machinery, 1997.

85. Shen, W. and Ghosal, S.: Adaptive Bayesian procedures using random series prior. Scand. J. Stat., 42:1194–1213, 2015.

86. Shen, X. and Wasserman, L.: Rates of convergence of posterior distributions. Ann. Statist., 29(3):687–714, 2001.

87. Shiu, S. Y. and Gatsonis, C.: The predictive receiver operating characteristic curve for the joint assessment of the positive and negative predictive values. Phil. Trans. Roy. Soc. A, 366(1874):2313–2333, 2008.

88. Sriram, K.: A sandwich likelihood correction for Bayesian quantile regression based on the misspecified assymetric Laplace density. Stat. Probab. Lett., 107:18–26, 2015.

89. Sriram, K.: A sandwich likelihood correction for Bayesian quantile regression based on the misspecified asymmetric Laplace density. Stat. Probab. Lett., 107:18–26, 2015.

90. Sriram, K., Ramamoorthi, R., and Ghosh, P.: Posterior consistency of Bayesian quantile regression based on the misspecified asymmetric Laplace density. Bayesian Analysis, 8(2):479–504, 2013.

91. Sriram, K., Ramamoorthi, R., and Ghosh, P.: Posterior consistency of Bayesian quantile regression based on the misspecified asymmetric Laplace density. Bayesian Analysis, 8(2):479–504, 2013.

92. Syring, N. and Li, M.: BayesBD: An R package for Bayesian inference on image boundaries. Accepted to R Journal, `arXiv:1612.04271`, 2017+.

93. Syring, N. and Martin, R.: Calibrating posterior credible regions. Unpublished manuscript, `arXiv:1509.00922`, 2017.

94. Syring, N. and Martin, R.: Gibbs posterior inference on the minimum clinically important difference. J. Stat. Plan. Inference, 187(0):67–77, 2017.

95. Syring, N. and Li, M.: BayesBD: Bayesian inference for image boundaries, 2017. R package version 1.2.

96. Syring, N. and Martin, R.: Robust Gibbs posterior inference on the boundary of a noisy image. Unpublished manuscript, `arXiv:1606.08400`, 2016.

97. Syring, N. A.: Gibbs Posterior Distributions: New Theory and Applications. Doctoral dissertation, University of Illinois at Chicago, 2017.

98. Turner, D., Schünemann, H. J., Griffith, L. E., Beaton, D. E., Griffiths, A. M., Critch, J. N., and Guyatt, G. H.: The minimal detectable change cannot reliably replace the minimal important difference. J. Clin. Epidemiol., 63(1):28–36, 2010.

99. van der Vaart, A. W.: Asymptotic Statistics. New York, Cambridge, 1998.

100. van der Vaart, A. W.: Asymptotic Statistics. Cambridge University Press, Cambridge, 1998.

101. Walker, S. G.: Bayesian inference with misspecified models. J. Statist. Plann. Inference, 143(10):1621–1633, 2013.

102. Wang, B. and Titterington, D.: Convergence properties of a general algorithm for calculating variational Bayesian estimates for a normal mixture model. Bayesian Anal., 1:625–650, 2006.

103. Wang, B. and Titterington, D. M.: Inadequacy of interval estimates corresponding to variational bayesian approximations. In AISTATS, 2005.

104. Wasserman, L. A.: All of Nonparametric Statistics. New York, Springer, 2005.

105. Watson, G. S.: Smooth regression analysis. Sankhya, 50(1), 1982.

106. White, H.: Maximum likelihood estimation of misspecified models. Econometrica, 50(1), 1982.

107. Wong, W. H. and Shen, X.: Probability inequalities for likelihood ratios and convergence rates of sieve MLEs. Ann. Statist., 23(2):339–362, 1995.

108. Yang, Y. and He, X.: Bayesian empirical likelihood for quantile regression. Ann. Stat., 40(2):1102–1131, 2012.

109. Yang, Y., Wang, H., and He, X.: Posterior inference in Bayesian quantile regression with assymetric Laplace likelihood. Int. Stat. Rev., 84(3):327–344, 2016.

110. Yang, Y. and He, X.: Bayesian empirical likelihood for quantile regression. Ann. Stat., 40(2):1102–1131, 2012.

111. Yu, K. and Moyeed, R. A.: Bayesian quantile regression. Stat. Probab. Lett., 54(4):437–447, 2001.

112. Yuan, W., Chin, K., Hua, M., Dong, G., and Wang, C.: Shape classification of wear particles by image boundary analysis using machine learning algorithms. Mech. Syst. Signal Pr., 72-73:346–358, 2016.

113. Zhang, T.: From ϵ-entropy to KL-entropy: analysis of minimum information complexity density estimation. Annals of Statistics, 34(5):2180–2210, 2006.

114. Zhang, T.: Information theoretical upper and lower bounds for statistical estimation. IEEE Trans. Inf. Theory, 52(4):1307–1321, 2006.

115. Ziou, D. and Tabbone, S.: Edge detection techniques – an overview. Pattern Recognition and Image Analysis C/C of Raspoznavaniye Obrazov I Analiz Izobrazhenii, 8:537–559, 1998.

VITA

Nicholas Aaron Syring

Contact Information

Chicago, IL

Email: `nasyrin@gmail.com`

Phone: 779-400-5642

URL: `http://homepages.math.uic.edu/~nsyring2/home.html`,

`https://github.com/nasyring`

Education

- Ph.D. Statistics, University of Illinois at Chicago, Advisors: Dr. Min Yang and Dr. Ryan Martin (NCSU), 12/2017 (anticipated)
- M.S. Statistics, Northern Illinois University, 2013
- B.S. Actuarial Science, Illinois State University, 2009

Employment

Allstate Insurance, *Data Scientist Intern*, Summer 2017

North Carolina State University, *Teaching Assistant*, 2016–2017

Google Summer of Code, Summer 2016

University of Illinois at Chicago, *Teaching Assistant*, 2013–2017

University of Illinois at Chicago, *Research Assistant*, 2015

NASA Langley Area Research Center, *Visiting Research Scientist*, Summer 2014

Capital One, *Statistician Intern*, Summer 2013

Northern Illinois University, *Teaching Assistant*, 2011–2013

State Farm Life Insurance Company, *Actuarial Analyst*, 2009–2011

Teaching

@ North Carolina State University

- ST311: Introduction to Statistics, Instructor, Spring 2017
- ST371: Introduction to Probability and Distribution Theory, TA, Fall 2016

@ University of Illinois at Chicago

- MATH165: Business Calculus, TA, Fall 2017
- STAT381: Applied Statistical Methods I, Instructor, Spring 2016
- MATH121: Precalculus, TA (2 times)
- MATH180: Calculus I, TA (2 times)

@ Northern Illinois University

- STAT350: Introduction to Probability & Statistics, grader (4 times)
- STAT382: Theory of Interest, Financial Derivatives, & Statistics, TA (2 times)
- STAT481: Probabilistic Foundations in Actuarial Science, TA (2 times)

Journal Articles

Submitted

- N. Syring and R. Martin. *Robust and Rate-Optimal Gibbs Posterior Inference on the Boundary of a Noisy Image.*
- N. Syring and R. Martin. *Calibrating General Posterior Credible Regions.*

Published or Accepted

- N. Syring and M. Li. *BayesBD: An R Package for Bayesian Inference on Image Boundaries.* Accepted to R Journal. 2017+.
- N. Syring and R. Martin. *Gibbs Posterior Inference on the Minimum Clinically Important Difference.* Journal of Statistical Planning and Inference. 187 (2017): 67-77. `http://dx.doi.org/10.1016/j.jspi.2017.03.001.`
- C. Liu, R. Martin, and N. Syring. *Efficient Simulation from a Gamma Distribution with Small Shape Parameter.* Comput. Stat. (2016). `https://doi.org/10.1007/s00180-016-0692-0.`

Invited Seminar Talks

- Joint Statistical Meetings Invited Poster Presentation

  Inferential Models for Instrumental Variables

  Baltimore–07/2017

- Summer Research Conference

  Image Boundary Detection via a Gibbs Model

  IIT–05/2016

- Undergraduate Mathematics Seminar

  Misspecified Statistical Models: What happens when the model is wrong?

  Wheaton College–11/2015

- Statistics Seminar

  Scaling the Gibbs posterior

  University of Illinois at Chicago–09/2015

- Statistics Seminar

  On Bayesian inference without a model

  University of Illinois at Chicago–11/2014

Professional Service

- Committee Member, Statistics Graduate Student Leadership Committee, University of Illinois at Chicago, 2014-2016.

- Organizer, Graduate Statistics Seminar, University of Illinois at Chicago, 2014-2016.

- Ad-hoc reviewer of journal papers for: *Bulletin of the Iranian Mathematical Society.*