Model misspecification, Bayesian versus credibility estimation, and Gibbs posteriors

Liang Hong[*] and Ryan Martin[†]

July 18, 2019

Abstract

In the context of predicting future claims, a fully Bayesian analysis—one that specifies a statistical model, prior distribution, and updates using Bayes's formula—is often viewed as the gold-standard, while Bühlmann's credibility estimator serves as a simple approximation. But those desirable properties that give the Bayesian solution its elevated status depend critically on the posited model being correctly specified. Here we investigate the asymptotic behavior of Bayesian posterior distributions under a misspecified model, and our conclusion is that misspecification bias generally has damaging effects that can lead to inaccurate inference and prediction. The credibility estimator, on the other hand, is not sensitive at all to model misspecification, giving it an advantage over the Bayesian solution in those practically relevant cases where the model is uncertain. This begs the question: does robustness to model misspecification require that we abandon uncertainty quantification based on a posterior distribution? Our answer to this question is *No*, and we offer an alternative *Gibbs posterior* construction. Furthermore, we argue that this Gibbs perspective provides a new characterization of Bühlmann's credibility estimator.

Keywords and phrases: asymptotics; Bernstein–von Mises phenomenon; exponential family; robustness; uncertainty quantification.

1 Introduction

The classical credibility theory of Bühlmann (1967) is a cornerstone of the insurance industry. Designed mainly for premium estimation, the credibility estimator is simple and intuitive, bypassing the many possible challenges of a full Bayesian analysis. But the state-of-the-art in Bayesian analysis has changed a lot since the 1970s, computational and methodological

[*] Liang Hong is an Associate Professor in the Department of Mathematics, Robert Morris University, 6001 University Boulevard, Moon Township, PA 15108, USA. Tel.: 412-397-4024. Email address: hong@rmu.edu.

[†] Ryan Martin is an Associate Professor in the Department of Statistics, North Carolina State University, 2311 Stinson Drive, Raleigh, NC 27695, USA. Tel.: 919-515-1920. Email address: rgmarti3@ncsu.edu.

tools are now readily available to carry out very sophisticated analyses. And a major selling point of a Bayesian approach, compared to other existing frameworks, is that it provides uncertainty quantification, in the form of a posterior distribution, about any relevant feature of the claims distribution. Does this make the credibility theory irrelevant? From the perspective of robustness to model misspecification bias, we will argue here that the answer to this question is *no*. The perspective also eventually leads to a new characterization of the credibility estimator in terms of a so-called *Gibbs posterior*, thereby bridging the gap between the Bayesian and credibility frameworks.

After setting up the problem of predicting future claims and introducing the Bayesian and credibility approaches to this problem in Section 2, we proceed in Section 3 to discuss model misspecification and its effects on the Bayesian and credibility estimators. In particular, we demonstrate that the Bayesian posterior is sensitive to model misspecification, to the extent that inferences drawn can be misleading, even asymptotically. Specifically, under mild regularity conditions, the Bayesian posterior satisfies a Bernstein–von Mises property, i.e., is asymptotically normal, with vanishing variance. This is a desirable property when the model is correctly specified, since the mean of that normal distribution will equal the true value of the parameter asymptotically. When the model is incorrectly specified, however, a "true value of the parameter" does not exist, so the mean of the normal distribution will equal a "best approximation" relative to the posited model. That "best approximation" can lead to estimates which are arbitrarily far from the quantity being estimated, depending on the model quality, which is difficult to assess. The credibility estimator, on the other hand, by virtue of its simplicity and lack of commitment to any model, is perfectly robust to model misspecification, also converging to the true mean of the claims distribution. This stark contrast in how the two frameworks respond to model misspecification has important practical consequences. In particular, can one ever really be sure that a posited model is "correct"? If not, then how meaningful are those aforementioned selling points of the Bayesian framework? And those cases of exact credibility (e.g., Jewell 1975; Diaconis and Ylvisaker 1979), where the Bayes premium is equal to the credibility estimator provide little comfort since, even though the marginal posterior distribution for the mean might not be affected by misspecification, it is likely that the marginal posterior for every other feature is severely biased. These and other more subtle issues are addressed in Section 3.3.

The ubiquity of model misspecification and the havoc it can wreak on the Bayesian solution suggests that practitioners ought to avoid the risk altogether, sticking with the classical credibility theory. But to have a posterior distribution that quantifies uncertainty about the relevant unknowns, one might still be tempted towards a Bayesian analysis. The question is: *is it necessary that we give up on having probabilistic uncertainty quantification if we wish to avoid the risks of model misspecification?* Here, again, we argue that the answer is *no*, but since the Bayesian framework is tied directly to a statistical model through its dependence on the likelihood function, we require a new kind of posterior distribution construction, one that does not depend on a likelihood. For this, we recommend, in Section 4, a so-called *Gibbs posterior*, which has origins in machine learning (e.g., Zhang 2006ab) and has received considerable attention in statistics (e.g., Syring and Martin 2017, 2019ab; Wang and Martin

2019; Alquier 2008; Alquier and Ridgway 2017), even in an insurance application (Syring et al 2019). This framework proceeds to link data and quantities of interest through a discrepancy or risk function, rather than a likelihood. This discrepancy function is then combined with relevant prior information, very much like in Bayes's formula, leading to a posterior distribution that does not depend on any posited model. A novel feature of the Gibbs posterior is its *learning rate* parameter (e.g. Grünwald and van Ommen 2017, Bissiri et al. 2016, Syring and Martin 2019a) that controls the spread, and by properly tuning that spread, the Gibbs posterior uncertainty quantification can be made valid. Interestingly, the Bühlmann's classical credibility estimator can be characterized as the mean of a suitable Gibbs posterior, so that our proposed framework based on the former provides the aforementioned bridge between Bayes and credibility.

2 Background

2.1 Problem setup and model misspecification

Suppose the actuary has observable claims $X^n = (X_1, \ldots, X_n)$, assumed to be independent and identically distributed (iid) with common marginal distribution $P^\star$, having density function $p^\star$ with respect to a σ-finite measure ν on $\mathbb{X} = \mathbb{R}$ or $\mathbb{X} = [0, \infty)$, typically Lebesgue measure. One relevant feature of $P^\star$ is the mean $\mu^\star = \int x \, p^\star(x) \, \nu(dx)$, since that would be a best prediction of the next claim X_{n+1}. The actuary might also be interested in other features of $P^\star$, such as value-at-risk, conditional tail expectation, etc.

Since $P^\star$ is unknown, these features must be estimated based on the claims data. To this end, it is common to introduce a statistical model

$$\mathcal{P} = \{P_\theta : \theta \in \Theta\},$$

which is just a collection of probability distributions on $\mathbb{X}$, having densities p_θ with respect to ν, indexed by a parameter θ taking values in a parameter space Θ. For example, $\mathcal{P}$ might be the class of gamma distributions indexed by the parameter $\theta = (\alpha, \beta)$ that determines the shape and scale, respectively. Under a posited statistical model, those features of interest to the actuary are now described as functions of the model parameter θ. For example, the mean is $\mu_\theta = \int x \, p_\theta(x) \, \nu(dx)$, depending explicitly on θ, and similarly the variance is $\sigma_\theta^2 = \int (x - \mu_\theta)^2 \, p_\theta(x) \, \nu(dx)$; in our insurance context, these are referred to as the *individual premium* and *process variance*, respectively.

When the actuary introduces a model $\mathcal{P}$, he/she is effectively assuming that $P^\star \in \mathcal{P}$. Of course, efforts can be made to justify such an assumption but, at the end of the day, it is still just an assumption that may or may not be true. When the model is correct, i.e., $P^\star \in \mathcal{P}$, we say that the model is *well-specified*, and it implies existence of a $\theta^\star \in \Theta$ such that $P_{\theta^\star} = P^\star$. An important consequence of the model being well-specified is that accurate estimation of $\theta^\star$ implies accurate estimation of any (smooth) feature of $P^\star$. On the other hand, when the model is incorrect, i.e., $P^\star \notin \mathcal{P}$, we say that the model is *misspecified*, which implies that there is no "true" value of θ. The actuary is, of course, unaware of this misspecification, so he/she will proceed to estimate the model parameter and relevant features of the claims distribution, but it is not immediately clear what will happen. Surely, there will be some features of the claims distribution that cannot be accurately estimated using methods based on a misspecified model, and we will generically refer to this deficiency as *model misspecification bias*. Here we explore the effect of model misspecification, e.g., what relevant features are unaffected by misspecification and under what conditions?

2.2 Bayesian estimation

The Bayesian approach is a normative framework for statistical inference. It starts with a statistical model $\mathcal{P}$, indexed by a parameter $\theta \in \Theta$, along with a prior distribution Π_0 on Θ, and applies Bayes's formula to construct a posterior distribution Π_n, depending on the claims data X^n, to be used for estimation, inference, prediction, etc. That is, the posterior distribution is defined as

$$\Pi_n(A) = \frac{\int_A L_n(\theta)\, \Pi_0(d\theta)}{\int_\Theta L_n(\theta)\, \Pi_0(d\theta)}, \quad A \subseteq \Theta, \tag{1}$$

where $L_n(\theta) = \prod_{i=1}^n p_\theta(X_i)$ is the likelihood function for θ based on data X^n. Only in certain special cases can the posterior distribution be written in closed-form, but Monte Carlo methods can be used to produce accurate numerical approximations.

Before proceeding with the discussion of how the posterior distribution will be used, it is important to first clarify what it represents. Rarely does the prior distribution Π_0 encode genuine prior beliefs about the model parameter θ. Indeed, the model itself is uncertain before and sometimes—as in our present case—even after data is observed, so it is not possible to have real prior information about a parameter that did not exist prior to specification of the model. There might be prior information available about certain features of the underlying $P^\star$, such as the mean $\mu^\star$, as discussed in Section 4, that can be used to help motivate a particular prior distribution for θ, but it generally does not *determine* the prior. Therefore, the prior is usually of a non- or partially-informative variety, and taken to have a relatively simple mathematical form, e.g., conjugate. Our point is that the prior is viewed simply as a device to get from data to a posterior, via Bayes's formula, not as a believable part of the model. So even though, mathematically, the Bayesian framework operates under a full joint distribution for data and parameter, i.e.,

$$\theta \sim \Pi_0 \quad \text{and} \quad (X_1, X_2, \ldots, X_n, \ldots) \mid \theta \overset{\text{iid}}{\sim} P_\theta,$$

the modern Bayesian does not take this as a genuine model for the claims data; he/she assumes, as above, that claims are iid from some $P^\star$, so questions about the behavior and performance of the posterior distribution under the iid setup are relevant, in both the well- and misspecified cases.

Various features of the posterior distribution might be of interest in a given application, but we are specifically interested in prediction of the next claim and, for this, the *posterior* *predictive distribution* is important. If Π_n is the posterior distribution as defined above, then the predictive density is

$$p_n(x) = \int_\Theta p_\theta(x)\, \Pi_n(d\theta), \quad n \geq 0.$$

This can be viewed as an estimate of the true density $p^\star$, the best guess of the distribution of X_{n+1}, based on data X^n. The case $n = 0$ corresponds to a "no-data" scenario and p_0 is referred to as the prior predictive density. The mean of the predictive distribution is called the *Bayes premium* and, by Fubini's theorem, has two different looking expressions:

$$\hat{\mu}_n^B := \int_{\mathbb{X}} x\, p_n(x)\, \nu(dx) = \int_\Theta \mu_\theta\, \Pi_n(d\theta).$$

As the notation suggests, $\hat{\mu}_n^B$ is the Bayes estimate of the mean $\mu^\star$ of $P^\star$ under squared-error loss based on the claims data X^n. In the the "no-data" scenario with $n = 0$, the prior predictive mean μ_0 is called the *collective premium*.

2.3 Credibility estimation

While the full Bayesian analysis described above is conceptually straightforward, the technology needed to actually carry out these computations for realistic models was unavailable until the early 1990s. Since important real-life problems existed long before the 1990s, Bühlmann (1967), building on ideas of Whitney (1918) and Bailey (1950), suggested a work-around that could achieve some of the benefits of a Bayesian analysis but without the computational burden. Specifically, Bühlmann's classical credibility theory seeks an estimator $\hat{\mu}_n^C$, linear in X^n, that minimizes the overall mean square prediction error. That is, define the function $B : \mathbb{X}^n \to \mathbb{R}$ to be solution to the optimization problem

$$\arg\min_{\hat{\mu}(\cdot)} \int_\Theta \int_{\mathbb{X}^n} \{\mu_\theta - \hat{\mu}(x^n)\}^2\, \nu^n(dx^n)\, \Pi_0(d\theta),$$

where the minimum is over all estimators $\hat{\mu}(x^n)$ that are linear in x^n. Then Bühlmann's recommendation is to set $\hat{\mu}_n^C = B(X^n)$. It is not too difficult to show that

$$\hat{\mu}_n^C = \frac{n}{n+\kappa}\, \bar{X}_n + \frac{\kappa}{n+\kappa}\, \mu_0, \tag{2}$$

where $\bar{X}_n = n^{-1} \sum_{i=1}^n X_i$ is the sample mean, and

$$\kappa = \frac{\int \sigma_\theta^2\, \Pi_0(d\theta)}{\int (\mu_\theta - \mu_0)^2\, \Pi_0(d\theta)}.$$

This is just a convex combination of the prior mean, μ_0, and the sample mean, $\bar{X}_n$, with the weight attached to the latter approaching unity as the data becomes more informative, i.e., as $n \to \infty$. Therefore, $\hat{\mu}_n^C$ is asymptotically equally as good as the optimal estimator, the sample mean, while accounting for some available prior information in finite-samples. Roughly speaking, κ indicates homogeneity of the policyholders with respect to the risk parameter θ. A relatively homogeneous pool of policyholders means a relatively small denominator of κ and vice versa. In certain cases, namely, when the model $\mathcal{P}$ is an exponential family (see Section 3.3), there is so-called *exact credibility* (Jewell 1975, Diaconis and Ylvisaker 1979) in the sense the Bayes premium actually equals the credibility estimator; this provides some additional—albeit unneeded—theoretical support for the credibility estimator. For a detailed account of the credibility theory, we refer readers to Bühlmann and Gisler (2005) and Klugman et al. (2008).

3 Effects of model misspecification

3.1 Bayesian case

In the well-specified model case, under certain relatively mild conditions, the Bayesian posterior satisfies a number of desirable asymptotic properties, e.g, posterior consistency, optimal or at least near-optimal concentration rates, a Bernstein–von Mises style posterior normality property, etc. However, when the model is misspecified, things are more complicated; in fact, it is not immediately clear what the desired asymptotic properties would be, given that there is no "true θ" around which we would hope the posterior to concentrate. While there is no "true θ," there is a "best θ" in the sense that it minimizes the Kullback–Leibler divergence of the model P_θ from the true distribution $P^\star$. More specifically, if

$$K(p^\star, p_\theta) := \int \log(p^\star/p_\theta)\, p^\star \, d\nu$$

denotes the Kullback–Leibler divergence of P_θ from $P^\star$, then point around which we hope the posterior will concentrate is the minimizer

$$\theta^\dagger = \arg\min_\theta K(p^\star, p_\theta).$$

General sufficient conditions for existence and uniqueness of the Kullback–Leibler minimizer are discussed in, e.g., Kleijn and van der Vaart (2006). For the smooth, finite-dimensional problems we have in mind here, finding a minimizer and showing that it is unique involves only basic calculus techniques; see Example 1 below. Of course, if the model is well-specified, so that there exists a "true value" $\theta^\star$, with $P^\star = P_{\theta^\star}$, then $\theta^\dagger = \theta^\star$, so the discussion here generalizes those like in Ghosh and Ramamoorthi (2003) and elsewhere; see Hong and Martin (2017b) for such a discussion in the context of insurance applications. This notion of minimizing the Kullback–Leibler divergence shows up in both the Bayesian (e.g., Berk 1966; Bunke & Milhaud 1998; Kleijn and van der Vaart; De Blasi and Walker 2013; Ramamoorthi et al. 2015) and non-Bayesian (e.g., Dahalhaus and Wefelmeyer 1996; Patilea 2001) literature on misspecification for finite- and infinite-dimensional models.

Example 1. Suppose an actuary is entertaining the following gamma model:

$$\mathcal{P} = \left\{(\theta^3/2)x^2 e^{-\theta x}, x > 0 : \theta \in (0, \infty)\right\},$$

where $\Gamma(\alpha) = \int_0^\infty t^{\alpha-1} e^{-t} dt$ is the gamma function. But the observable claims $X_1, \ldots, X_n$ are in fact iid from the following lognormal model:

$$p^\star(x) = (x\sqrt{2\pi})^{-1} e^{-(\log x - 1)^2/2}, \quad x > 0.$$

Note that the true mean is $\mu^\star = \exp(1 + \frac{1}{2}) \approx 4.48$. In this case, it is easy to check that

$$K(p^\star, p_\theta) = \mu^\star \theta - 3 \log \theta + c,$$

where c is a constant that does not depend on θ and $\mu^\star$ denotes the true mean. It is now just a simple calculus exercise to show that the Kullback–Leibler minimizer $\theta^\dagger$ exists, is unique, and equals $3/\mu^\star = 0.669$.

Since our focus is on finite-dimensional cases and, in particular, nice exponential family models, we describe here some relatively recent results (Kleijn and van der Vaart 2012) on the so-called *Bernstein–von Mises phenomenon* under misspecification, where, as $n \to \infty$, the posterior Π_n resembles a normal distribution centered near $\theta^\dagger$ with a variance that is $O(n^{-1})$. In order to not disrupt the flow of our presentation, we give an incomplete statement of the result here, postponing discussion of the technical details until Appendix A. It will help to keep in mind that the posterior distribution itself is random because it depends on data X^n, so the forthcoming distributional convergence results also have stochastic qualifications. Here and throughout, we will use the notation $\dot{g}$ and $\ddot{g}$ to denote the gradient vector and Hessian matrix of a real-valued function g.

Theorem 1 (Kleijn and van der Vaart 2012)**.** *Let $P^\star$ be the true marginal distribution for the iid sequence $X_1, X_2, \ldots$ and consider the statistical model $\{P_\theta : \theta \in \Theta\}$, where $\Theta \subseteq \mathbb{R}^d$ for finite $d \geq 1$. Let $\theta^\dagger$ denote the unique Kullback–Leibler minimizer and suppose that*

- *Conditions 1–2 in Appendix A hold for the pair $(P^\star, \theta^\dagger)$, and*
- *the prior Π has a density that is positive and continuous in a neighborhood of $\theta^\dagger$.*

Then the posterior Π_n in (1) satisfies

$$\rho_{\text{TV}}\big[\Pi_n, \, \mathsf{N}_d\{\hat{\theta}_n, (nV_{\theta^\dagger})^{-1}\}\big] \to 0 \quad \text{in } L_1(P^\star) \text{ as } n \to \infty,$$

where ρ_{TV} is the total variation distance, $\hat{\theta}_n$ is a maximum likelihood estimator, and $V_{\theta^\dagger} = \ddot{k}^\star(\theta^\dagger)$, for $k^\star(\theta) = K(p^\star, p_\theta)$, is a positive definite $d \times d$ covariance matrix.

The conditions eluded to in Theorem 1 are rather mild, so this strong conclusion applies to a wide range of statistical models, including exponential families as discussed below. To rephrase those conclusions in more colloquial terms, under certain conditions, features of the posterior distribution can be accurately approximated, asymptotically, by those same features of a normal distribution with mean $\hat{\theta}_n$ and covariance $(nV_{\theta^\dagger})^{-1}$. For example, when n is large, the mean of the posterior is approximately $\hat{\theta}_n$ (which is approximately $\theta^\dagger$, see Appendix A) and a $100(1-\alpha)\%$ highest posterior density credible region is approximately

$$\{\vartheta \in \Theta : n(\vartheta - \hat{\theta}_n)^\top V_{\theta^\dagger}(\vartheta - \hat{\theta}_n) \leq c_\alpha\},$$

where c_α is the $(1-\alpha)$-quantile of the chi-square distribution with d degrees of freedom. Moreover, if θ is approximately normal under the posterior Π_n, with $O(n^{-1})$ variance, then we can immediately get a corresponding normal distribution approximation for any smooth function $g(\theta)$ of θ using the delta theorem. For example, if g is scalar-valued, then

$$\theta \sim \Pi_n \implies g(\theta) \sim \mathsf{N}\big(g(\hat{\theta}_n),\, \dot{g}(\theta^\dagger)^\top (nV_{\theta^\dagger})^{-1}\dot{g}(\theta^\dagger)\big), \quad \text{for large } n,$$

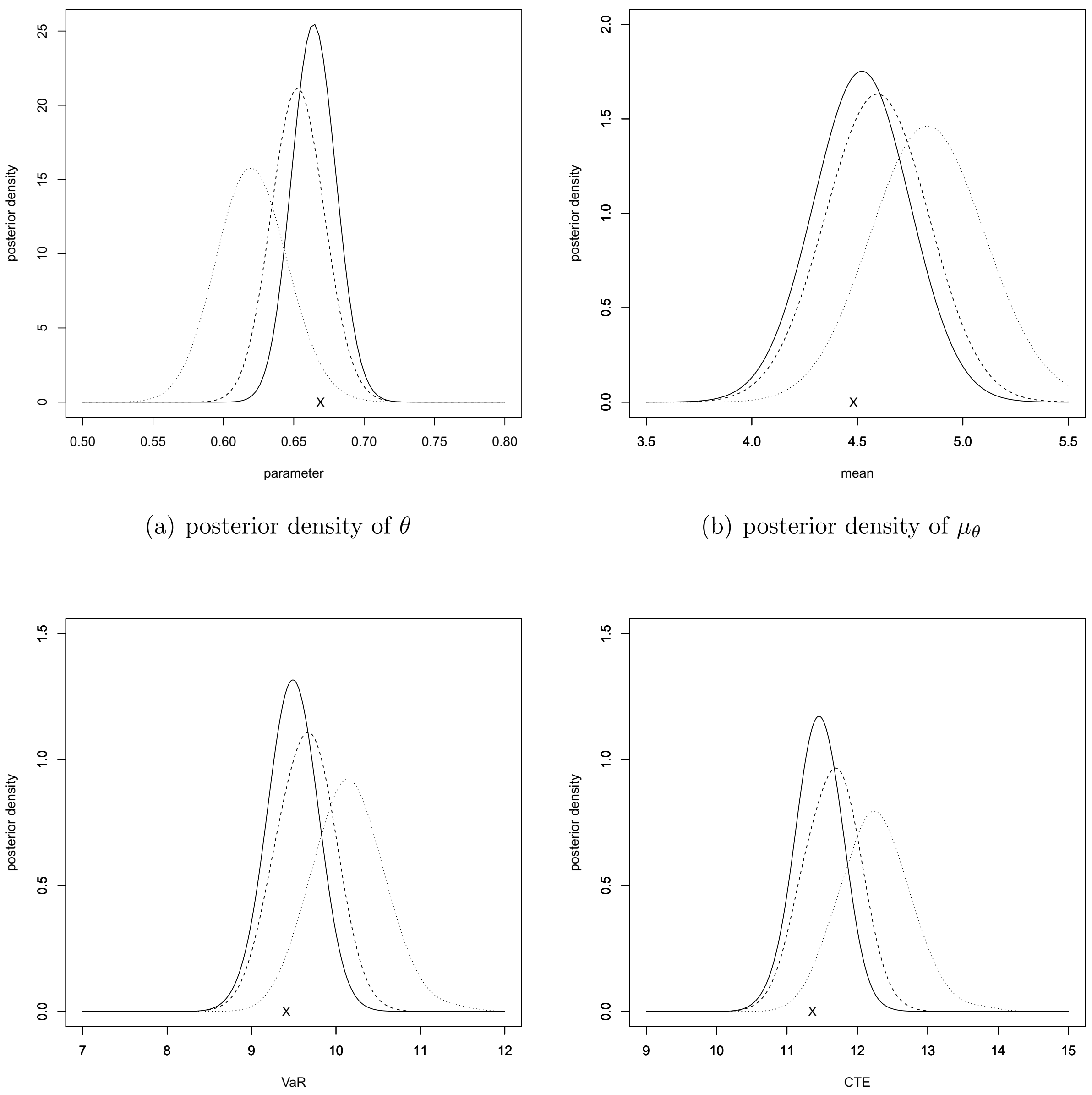
where, of course, $\dot{g}$ is assumed to be non-vanishing at $\theta^\dagger$. A particularly relevant choice of g is the mean function, $\theta \mapsto \mu_\theta$, and a particular consequence of the above discussion is that the Bayes premium satisfies $\hat{\mu}_n^B \to \mu_{\theta^\dagger}$. Other features can be handled similarly.

The above is a mostly complete story of the effect of model misspecification on the Bayesian posterior distribution. While the story is relatively simple and elegant, it does not have a happy ending in general. That is, the villain—misspecification bias—delivers a major, sometimes fatal blow to the hero—the Bayesian posterior—casting doubt on any subsequent statistical analysis. See Section 3.3 for more details. Next is an example giving a particular instantiation of the general results discussed above.

Example 2. Consider same the setup in Example 1, with a gamma model, indexed by the rate parameter θ, and lognormal truth. Suppose the actuary takes a Bayesian approach and chooses the prior for θ to be the exponential distribution with hazard rate $\lambda = 0.2$. The argument presented in Appendix A shows that Theorem 1 applies. Therefore, $\widehat{\mu}_n^B \to \mu_{\theta^\dagger} = \mu^\star = 4.48$. The delta theorem also implies that the value-at-risk and conditional tail expectation at the $100q$-th level, with $q \in (0,1)$, of the predictive density, denoted by $\text{VaR}_n(q)$ and $\text{CTE}_n(q)$ respectively, satisfy $\text{VaR}_n(q) \to \text{VaR}_{\theta^\dagger}(q)$ and $\text{CTE}_n(q) \to \text{CTE}_{\theta^\dagger}(q)$. For example, if we take $q = 0.95$, then $\text{VaR}_{\theta^\dagger}(0.95) = 9.411$ and $\text{CTE}_{\theta^\dagger}(0.95) = 11.363$. But $\text{VaR}^\star(0.95) = 14.081$ and $\text{CTE}^\star(0.95) = 23.261$. This substantial gap between the true values and those the Bayesian solution points to clearly demonstrates the negative effect of model misspecification. To visualize the Bernstein–von Mises phenomenon in this case, we also perform a small simulation study. For three different sample sizes $200, 400$ and 600, Panel (a) of Figure 1 shows the posterior density $\pi_n(\theta)$. Panels (b), (c), and (d) demonstrate the corresponding posterior densities of $\hat{\mu}_\theta^B$, $\text{VaR}_\theta(0.95)$, and $\text{CTE}_\theta(0.95)$ along with the values of $\mu_{\theta^\dagger}$, $\text{VaR}_{\theta^\dagger}(0.95)$, and $\text{CTE}_{\theta^\dagger}(0.95)$.

3.2 Credibility case

As discussed in Section 2.3, a cornerstone of the classical credibility theory is simplicity, and that shines through here too in our assessment of the effect of model misspecification. Indeed, it is immediate from the credibility estimator's definition in (2) and the strong law of large numbers that $\hat{\mu}_n^C$ converges with $P^\star$-probability 1 to the true mean $\mu^\star$, regardless of whether $P^\star \in \mathcal{P}$ or $P^\star \not\in \mathcal{P}$. Therefore, the credibility estimator always identifies the true mean so, if premium estimation is the primary goal, the credibility estimator is superior to Bayes because it has no risk of model misspecification bias, no computational challenges, and no loss of efficiency. More details on this claim, along with connections to the so-called *exact credibility* case, are given in Section 3.3; see, also, Section 4.



(a) posterior density of θ

(b) posterior density of μ_θ

(c) posterior density of $\mathrm{VaR}_\theta(0.95)$

(d) posterior density of $\mathrm{CTE}_\theta(0.95)$

Figure 1: Illustration of the Bernstein–von Mises phenomenon under misspecification in Example 2. In all panels, three different sample sizes are taken: 200 (dotted), 400 (dashed) and 600 (solid). (a) Posterior densities of the parameter; (b) Posterior densities of the mean; (c) Posterior densities of VaR; (d) Posterior densities of CTE. The Kullback Leibler minimizer is marked by the symbol X.

3.3 Take-away messages

Take-away 1. In Section 3.1, we demonstrated that, under appropriate conditions, the Bayesian posterior distribution is going to concentrate its mass around the minimizer $\theta^\dagger$ of the Kullback–Leibler divergence of the model P_θ from the true distribution $P^\star$. This is the best possible outcome when the model is misspecified, but further investigation is still needed. First, there are some cases where this outcome is satisfactory. In particular, consider an exponential family model $\{P_\theta : \theta \in \Theta\}$ in its natural or canonical form, with density given by

$$p_\theta(x) = h(x) \exp\{\theta\, x - A(\theta)\},$$

where $h(x) > 0$ and A is determined by the constraint that p_θ must integrate to unity; see, e.g., Brown (1986) for details about the many nice properties possessed by this class of distributions. In addition to the well-known regularity of exponential families, which ensure that the conditions of Theorem 1 hold (see Appendix A), it can be shown that the Kullback–Leibler minimizer $\theta^\dagger$ satisfies the equation

$$\mu_{\theta^\dagger} = \mu^\star,$$

where the mean μ_θ is given by the expression $\dot{A}(\theta)$; see, e.g., Example 2.66 in Schervish (1995) or Example 2 in Bunke and Milhaud (1998). Consequently, if the Bayesian specifies an exponential family model that happens to be misspecified, then he/she will still be able to recover the individual premium asymptotically. In fact, if one takes a prior Π_0 conjugate to the exponential family model (Diaconis and Ylvisaker 1979), then the corresponding Bayes estimator $\hat{\mu}_n^B$ is exactly the credibility estimator $\hat{\mu}_n^C$, a case commonly referred to as *exact credibility* (Jewell 1975). This is the good news. The bad news is that most or all of the other relevant features of the posterior will be negatively affected by the model misspecification, since the feature $g(\theta^\dagger)$ of the limiting posterior distribution does not equal the corresponding feature of the true $P^\star$. The major selling point of a Bayesian approach is that it offers a normative framework for learning and making inference about any relevant feature of $P^\star$, but this argument only holds up when the model is well-specified. Under a misspecified model, the Bayesian approach will give incorrect or misleading answers, even asymptotically, to all or most relevant questions about $P^\star$.

Take-away 2. To follow up on the previous point, it would not make sense for a Bayesian to opt for an exponential family model *just because* the Bayes premium matches the credibility estimator. With the knowledge outlined above, suitable measures must be taken to keep our analyses safe from the effects of model misspecification bias (e.g., Grünwald 2018). One such measure would be to abandon the Bayesian framework for the simpler credibility estimator. This is a rather extreme measure because there is something desirable about having a "posterior" that provides uncertainty quantification. In Section 4 we present a strategy that balances between Bayesian and credibility estimation.

Take-away 3. We have focused more on the effect of model misspecification on the *location* of the posterior distribution, i.e., that for a given feature $g(\theta)$ of P_θ, the posterior will center around $g(\theta^\dagger)$ which might be very different from the target feature of $P^\star$. A more subtle issue is the effect that model misspecification has on the *spread* of the posterior. Recall that, in Theorem 1, it happens that the posterior will asymptotically center around $\hat{\theta}_n$, the maximum likelihood estimator. When the model is misspecified, there is no longer a connection between the likelihood function and the claims distribution and that changes the status of $\hat{\theta}_n$ to a M-estimator (e.g., van der Vaart 1998, Chapter 5). This is important because, in the well-specified case, there is a direct connection between the variance of the maximum likelihood estimator and the second derivative of the log-likelihood function, consequence of the identity (e.g., Brown 1986, Section 4.3)

$$\int \dot{\ell}_\theta \, \dot{\ell}_\theta^\top \, p_\theta \, d\nu = - \int \ddot{\ell}_\theta \, p_\theta \, d\nu, \tag{3}$$

where $\ell_\theta = \log p_\theta$. However, as is well known, the variance of a M-estimator is given by the so-called *sandwich formula* (e.g., Müller 2013). For illustration, consider our setting where we are interested in the marginal posterior distribution of μ_θ. The delta theorem argument above says that the posterior variance is

$$n^{-1} \, \dot{\mu}_{\theta^\dagger}^\top V_{\theta^\dagger}^{-1} \dot{\mu}_{\theta^\dagger}.$$

However, the asymptotic variance of the M-estimator, $\hat{\theta}_n$, is

$$n^{-1} \dot{\mu}_{\theta^\dagger} V_{\theta^\dagger}^{-1} M_{\theta^\dagger} V_{\theta^\dagger}^{-1} \dot{\mu}_{\theta^\dagger},$$

where, compared to the left-hand side of (3),

$$M_\theta = \int \dot{\ell}_\theta \, \dot{\ell}_\theta^\top \, p^\star \, d\nu. \tag{4}$$

Note that $V_{\theta^\dagger}$ is like the matrix on the right-hand side of (3), but with expectation according to $p^\star$. In general, $M_{\theta^\dagger}$ and $V_{\theta^\dagger}$ will be different, so the two asymptotic variances disagree. Consequently, the spread of the posterior can be too narrow or too wide, casting doubt on the validity of the uncertainty quantification contained therein. The Bayesian framework offers no remedy to correct for the variance mismatch, but see Section 4.

Take-away 4. Our focus here has been on finite-dimensional cases where the specification of a statistical model $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ can be a serious restriction. For Bayesians who are concerned about misspecification and want to avoid this risk, one option is to expand the model to include a wider collection of distributions, a so-called *Bayesian nonparametric* approach where the unknown distribution is the "parameter," making it an infinite-dimensional problem. By expanding the scope of the model, one enjoys a number of benefits, in particular, the more flexible model automatically eliminates some but not all of the risk of misspecification bias; see, e.g., Kleijn and der Vaart (2006), DeBlasi and Walker (2013). and Ramamoorthi et al. (2015). But this added flexibility does not come for free—there are costs in terms of both computational and statistical efficiency when one expands to an infinite-dimensional model—so the actuary wanting to take a Bayesian approach has a difficult choice to make.

Hong and Martin (2016, 2017a, 2018) discuss the benefits of a nonparametric formulation, the relevant theoretical properties, and implementation details in an insurance context; see, also, Richardson and Hartman (2019). Fortunately, a middle-ground between the finite- and infinite-dimensional Bayes formulation—retaining the benefits of both, but without the shortcomings—is possible and we discuss this in Section 4.

4 A new perspective on credibility theory

As discussed above, the Bayesian approach is attractive because it results in a posterior probability distribution that incorporates available prior information and quantifies uncertainty about any unknown feature of $P^\star$, but it is sensitive to model misspecification. The credibility estimator, on the other hand, provides only a point estimate, but it is optimal in terms of estimation accuracy and is not sensitive to model misspecification. Is there a middle-ground that offers the benefits of each approach but without their shortcomings? Our desiderata are as follows:

- incorporates available prior information and returns a sort of "posterior;"
- the "posterior" is robust, i.e., not sensitive to model misspecification;
- estimates derived from the "posterior" are optimal;
- and uncertainty quantification derived from the "posterior" should be *valid* in the sense that a $100(1-\alpha)\%$ credible region, for $\alpha \in (0,1)$, should have approximately $1-\alpha$ coverage probability under $P^\star$.

For the first desideratum, a Bayesian-style prior-to-posterior updating would be nice, but it cannot involve a likelihood because that is what leads to sensitivity to model misspecification. To be robust in the sense of the second desideratum, the updating should be through something other than a likelihood. Moreover, the third and fourth desiderata require, roughly, that this "posterior" have right center and spread. So the crucial step is identifying a suitable substitute for the likelihood in Bayes's formula.

Before that, however, there is one point that deserves emphasis. The above desiderata cannot be achieved simultaneously for all relevant features of $P^\star$. What makes this uniformity possible in the Bayesian formulation[1] is that the assumed model is part of the posterior construction, and that is precisely what leads to its sensitivity to model misspecification. The price of achieving robustness as in the second desideratum is that we have to choose a relevant feature of $P^\star$ on which to focus. In what follows, we will work with the mean μ of the claims distribution, since that is most relevant to the prediction task, but other features can be investigated similarly; it is also possible to investigate finitely many features simultaneously, but we will not discuss this here.

The remainder of this section is devoted to the construction of a so-called *Gibbs posterior* for the mean of $P^\star$. The name is derived from its connections to the Gibbs distribution often

[1] Even in the well-specified case, the Bayesian cannot make reliable inference on *all features*, due to certain limitations on marginalization. This *false confidence* phenomenon is described in Martin (2019)

encountered in statistical mechanics, but that is not important for our purposes. Some relatively recent references on Gibbs posteriors include Zhang (2006ab) and Bissiri et al. (2016). Given a generic value, μ, of the mean of $P^\star$, possibly different from the true value, $\mu^\star$, one way to measure its feasibility relative to a set of claims X^n is via the discrepancy function

$$D_n(\mu) = \frac{1}{n} \sum_{i=1}^{n} (X_i - \mu)^2. \tag{5}$$

That is, if $D_n(\mu') > D_n(\mu)$, then we say that μ is more feasible than μ' relative to the data. (Of course, this is not the only way one can measure feasibility but, in light of the connections we find below between our Gibbs posterior and those in the Bernstein–von Mises phenomenon of Theorem 1, we think this is a very reasonable approach.) Suppose that prior information about the mean of $P^\star$ is available, and that it can be encoded into a prior distribution Π; note the immediate advantage of only having to specify prior information about a feature of interest compared to a model parameter, e.g., a shape parameter, that may not have any direct real-world interpretation. Then we can define the Gibbs posterior distribution as

$$\Pi_n(d\mu) \propto e^{-\omega\, n D_n(\mu)}\, \Pi(d\mu), \quad \mu \in \mathbb{R}, \tag{6}$$

where the scale parameter $\omega > 0$ is called the *learning rate* (e.g. Grünwald and van Ommen 2017, Bissiri et al. 2016, Syring and Martin 2019a), related to the fourth desideratum, and will be discussed below. For different features of interest or for an entirely different context, the Gibbs posterior (6) looks the same, just the discrepancy function might be different. Other examples are presented in Syring and Martin (2017, 2019ab), Syring et al (2019), and Wang and Martin (2019).

So far we have achieved the first desideratum, going from a prior to a Gibbs posterior. Turning to robustness, note that there is no model explicitly being assumed in the definition of the Gibbs posterior. The reader might notice that we did implicitly assume a sort of Gaussian likelihood (see below), but there is a substantial practical difference between "explicit" and "implicit" in this case, and the learning rate parameter allows us to correct for the shortcomings of the implicit model that is sure to be wrong. That discrepancy function in (5) was not chosen for its Gaussian-like form, but rather for where it is minimized. In particular, it is easy to check that $D_n(\mu)$ is minimize at $\mu = \bar{X}_n$, the sample mean; similarly, the expected discrepancy, $D(\mu) = \int (x-\mu)^2\, p^\star(x)\, \nu(dx)$, the pointwise limit of $D_n(\mu)$, is minimized at $\mu = \mu^\star$, the true mean. This is relevant to the third desideratum because, just like the Bayesian posterior whose mode is near the maximum likelihood estimator, which tends to be close to the Kullback–Leibler minimizer, this property implies that the Gibbs posterior will concentrate around $\bar{X}_n$, the optimal estimator of $\mu^\star$.

Before proceeding to the fourth desideratum, we first do some simplification of the Gibbs posterior distribution in our present case focused on the mean of $P^\star$. That is,

$$\Pi_n(d\mu) \propto e^{-\omega\, n(\mu - \bar{X}_n)^2}\, \Pi(d\mu), \quad \mu \in \mathbb{R}.$$

Ignoring the prior, this resembles a normal distribution, centered at $\bar{X}_n$, with variance proportional to $(\omega n)^{-1}$. This makes clear that, through an appropriate choice of ω, the Gibbs posterior spread could be made such that the derived uncertainty quantification is valid in the sense described above. To be more precise, consider a $\mathsf{N}(\mu_0, \sigma_0^2)$ prior for μ. Then the Gibbs posterior is exactly

$$\mathsf{N}\big(\hat{\mu}_n^C, \tfrac{\kappa\sigma_0^2}{n+\kappa}\big) = \mathsf{N}\big(\hat{\mu}_n^C, (2n\omega + 1)^{-1}\big),$$

where $\kappa = (2\omega\sigma_0^2)^{-1}$ and $\hat{\mu}_n^C$ is the corresponding credibility estimator in (2). Compare this to the conclusion of Theorem 1 plus delta theorem presented in Section 3.1. Thus, *we have a new characterization of Bühlmann's credibility estimator as the mean of a Gibbs posterior.* Finally, to address the fourth desiderata, remember that the asymptotic variance of $\hat{\mu}_n^C$ is the same as that of $\bar{X}_n$ and is equal to $n^{-1}\sigma^{\star 2}$, where $\sigma^{\star 2}$ is the variance of $P^\star$. And since both the Gibbs posterior and the limiting sampling distribution of $\bar{X}_n$ are normal, to achieve the validity condition described in the fourth desideratum, it suffices to tune ω such that the two variances are asymptotically equal. That is, we need

$$\frac{\sigma^{\star 2}}{n} \approx \frac{1}{2n\omega + 1} \quad \text{or} \quad \omega \approx \frac{1}{2\sigma^{\star 2}}.$$

Of course, we do not know $\sigma^{\star 2}$ but it can be easily estimated and that estimate can be plugged in to tune ω. In other cases, achieving the appropriate tuning might be more challenging, but methods are available in Syring and Martin (2019a), Lyddon et al. (2019), etc.

The key point is that, if we are specifically interested in the mean of $P^\star$, then the Gibbs posterior achieves all four desiderata described above, thereby retaining the benefits of both a Bayesian posterior and credibility estimator, while avoiding the risk of model misspecification bias. Finally, the reader may have noticed that, in the Gibbs presentation above, we allowed the generic μ to range over the entire real line whereas, at least for claims data, we know that the mean is positive. In that case, the Gibbs posterior can be truncated to $[0, \infty)$, hence a truncated normal. For large n, this truncation has only a negligible effect, so all of what was discussed above would carry over unchanged to the truncated case.

5 Concluding remarks

In this paper we have investigated the asymptotic convergence properties of Bayesian posterior distributions when the model is misspecified. The general conclusion is that model misspecification has damaging effects on the Bayesian posterior distribution, at least for some features of the quantity of interest, that can lead to misleading inferences. One situation where the effect of model misspecification is seemingly mild is in the case where the model is an exponential family, so that the Bayesian marginal posterior distribution for μ_θ, the mean of the loss distribution, will be correctly centered around the true mean asymptotically. Even in this case, the spread of the posterior distribution is affected by misspecification bias so that uncertainty quantification based on that distribution would not be meaningful. The credibility estimator, on the other hand, is not affected by model misspecification at all so, if one cares only about prediction and is concerned about model misspecification bias, then one should abandon the Bayesian approach for the credibility estimator. But it is still possible to work with a "posterior" distribution that is not sensitive to model misspecification, but one has to go beyond a Bayesian framework. Here we suggested a Gibbs posterior construction and provided a new characterization of the classical credibility estimator as the mean of our Gibbs posterior distribution.

A shortcoming of the Bayesian approach is that it offers no remedy for misspecification bias. The remedy we recommended here was to construct a posterior distribution in a different way, using a discrepancy function instead of a proper likelihood. But this is not the only way. Recently, Grünwald has written about a so-called *generalized Bayesian* approach wherein one takes a proper likelihood as usual, but introduces a learning rate parameter—a power on the likelihood—like we had in our Gibbs formulation. He argues that, by taking that learning rate to be suitably small, one can correct for model misspecification in the sense that predictions will still be accurate; see, e.g., Grünwald (2012) and Grünwald and van Ommen (2017). In our insurance context, prediction is the motivation so Grünwald's *SafeBayes* approach seems promising and deserves further investigation.

Acknowledgments

This work is partially supported by the National Science Foundation, DMS–1811802.

A Details from Section 3.1

Here we fill in the left-out details in our discussion of the Bernstein–von Mises result in Section 3.1, based on Kleijn and van der Vaart (2012). In particular, we precisely state and explain their two key conditions, central to the approximate normality of the Bayesian posterior distribution in Theorem 1.

We adopt the notation used in van der Vaart (1998) where, if f is a real-valued measurable function and μ is a measure, then μf denotes the integral $\int f(x)\,\mu(dx)$ of f with respect to μ. Also, recall that we have a model with densities p_θ indexed by the parameter $\theta \in \Theta$, and $L_n(\theta) = \prod_{i=1}^n p_\theta(X_i)$ denotes the likelihood function for that model based on data X^n.

The first condition pertains to the regularity of the posited model with respect to the true distribution $P^\star$. It is related to those familiar conditions presented in introductory math-stat courses for establishing asymptotic normality of the maximum likelihood estimators.

Condition 1: Regularity. The model $\{p_\theta : \theta \in \Theta\}$ satisfies a local asymptotic normality condition at $\theta^\dagger$ relative to $P^\star$. That is, there exists random vectors $\Delta_{n,\theta^\dagger}$, bounded in $P^\star$-probability, and a positive definite matrix $V_{\theta^\dagger}$ such that, for any compact H,

$$\sup_{h \in H} \Bigl| \log \frac{L_n(\theta^\dagger + n^{-1/2}h)}{L_n(\theta^\dagger)} - h^\top V_{\theta^\dagger} \Delta_{n,\theta^\dagger} - \tfrac{1}{2} h^\top V_{\theta^\dagger} h \Bigr| \to 0 \quad \text{in } P^\star\text{-probability}. \tag{7}$$

Intuitively, this condition says that, in a certain strong sense, the log-likelihood function is approximately quadratic around $\theta^\dagger$. This approximately quadratic shape is precisely what drives the asymptotic normality of the posterior distribution. While in some cases—including the one we focus on below—it is possible to verify Condition 1 directly, but, for our discussion of Condition 2 below, we list here the general sufficient conditions for establishing Condition 1 presented in Kleijn and van der Vaart (2012).

A1. $\theta \mapsto \log p_\theta(x)$ is differentiable at $\theta^\dagger$ for $P^\star$-almost all x;

A2. there exists an open neighborhood U of $\theta^\dagger$ and a $P^\star$-square integrable function $x \mapsto m_{\theta^\dagger}(x)$ such that, for all $\theta_1, \theta_2 \in U$,

$$|\log p_{\theta_1}(x) - \log p_{\theta_2}(x)| \leq m_{\theta^\dagger}(x) \|\theta_1 - \theta_2\|, \quad \text{for } P^\star\text{-almost all } x;$$

A3. $\theta \mapsto k^\star(\theta) = K(p^\star, p_\theta)$ has a second-order Taylor approximation at $\theta^\star$,

$$k^\star(\theta) - k^\star(\theta^\dagger) = \tfrac{1}{2}(\theta - \theta^\dagger)^\top V_{\theta^\dagger}(\theta - \theta^\dagger) + o(\|\theta - \theta^\dagger\|^2), \quad \theta \to \theta^\dagger,$$

where $V_{\theta^\dagger}$ is a positive definite matrix.

The second condition is concerned with the size of the neighborhood around $\theta^\dagger$ to which the posterior assigns the majority of its mass.

Condition 2: Posterior concentration rate. The posterior distribution has a root-n concentration rate around $\theta^\dagger$ with respect to $P^\star$, i.e., for any sequence $a_n \to \infty$,

$$\Pi_n(\{\theta \in \Theta : \|\theta - \theta^\dagger\| > a_n n^{-1/2}\}) \to 0 \quad \text{in } L_1(P^\star) \text{ as } n \to \infty. \tag{8}$$

This condition says that the posterior distribution will assign nearly all of its mass to a neighborhood of $\theta^\dagger$ of radius roughly $O(n^{-1/2})$. This is important to the asymptotic normality result because it guarantees that the region where the quadratic approximation of the log-likelihood is inaccurate can be effectively ignored. In certain cases, the posterior concentration rate property can be checked directly. For example, suppose that the posterior mean vector and covariance matrix can be evaluated in closed form, and write these as m_n and C_n, respectively. Then it follows from Markov's inequality and the usual bias–variance decomposition of mean square error, we get

$$\Pi_n(\{\theta \in \Theta : \|\theta - \theta^\dagger\| > a_n n^{-1/2}\}) \leq \frac{n}{a_n^2} \bigl\{ \|m_n - \theta^\dagger\|^2 + \operatorname{tr}(C_n) \bigr\}.$$

Therefore, if the expected $\|m_n - \theta^\dagger\|^2$ and $\operatorname{tr}(C_n)$ are $O(n^{-1})$ as $n \to \infty$, which is common in finite-dimensional problems, then the root-n posterior concentration rate results holds. But for cases where the posterior mean and variance are not available in closed-form, then indirect methods are needed to verify Condition 2 above, and Kleijn and van der Vaart (2012) offer the following sufficient conditions.

B1. For all θ a neighborhood of $\theta^\dagger$, $P^\star(p_\theta / p_{\theta^\dagger}) < \infty$;

B2. for $m_{\theta^\dagger}$ as in A2 above, $P^\star \exp(s m_{\theta^\dagger}) < \infty$ for some $s > 0$;

B3. the matrix $M_{\theta^\dagger}$ defined in (4) is invertible; and

B4. for every $\varepsilon > 0$, there exists a sequence of test functions, $\psi_n : \mathbb{X}^n \to [0,1]$ such that

$$P^{\star n}\psi_n \to 0 \quad \text{and} \quad \sup_{\theta : \|\theta - \theta^\dagger\| > \varepsilon} Q_\theta^n(1 - \psi_n) \to 0,$$

where Q_θ is the measure with density given by $q_\theta(x) = p^\star(x)p_\theta(x)/p_{\theta^\dagger}(x)$.

Next we illustrate the above for the one-parameter exponential family model with densities $p_\theta(x) = h(x)\exp\{\theta x - A(\theta)\}$; of course, other models can be handled in a similar way. Recall that the Kullback–Leibler minimizer, $\theta^\dagger$, satisfies the relation $\mu_{\theta^\dagger} = \mu^\star$, where $\mu_\theta = \dot{A}(\theta)$. That is, the best approximation of $P^\star$ in the exponential family model is the one that matches the mean. As an aside, the maximum likelihood estimator is the unique solution to the equation $\dot{A}(\theta) = \bar{X}_n$, so consistency follows from the law of large numbers and the continuous mapping theorem.

Starting with Condition 1, we could verify A1–A3, but it is no more difficult to check (7) directly. Since $\theta \mapsto A(\theta)$ is smooth, we have the following Taylor approximation around $\theta^\dagger$:

$$A(\theta) - A(\theta^\dagger) = \dot{A}(\theta^\dagger)(\theta - \theta^\dagger) + \tfrac{1}{2}\ddot{A}(\theta^\dagger)(\theta - \theta^\dagger)^2 + o(|\theta - \theta^\dagger|^2).$$

Plugging this in to the log-likelihood function gives

$$\log \frac{L_n(\theta^\dagger + hn^{-1/2})}{L_n(\theta^\dagger)} = hn^{-1/2}\bigl\{n\bar{X}_n - n\dot{A}(\theta^\dagger)\bigr\} - \tfrac{1}{2}\ddot{A}(\theta^\dagger)h^2 + o(1), \quad n \to \infty.$$

Since $\dot{A}(\theta^\dagger) = \mu^\star$, it follows from the central limit theorem that

$$\Delta_{n,\theta^\dagger} := n^{-1/2}\ddot{A}(\theta^\dagger)^{-1}\bigl\{n\bar{X}_n - n\dot{A}(\theta^\dagger)\bigr\}$$

is asymptotically normal and, therefore, bounded in $P^\star$-probability. And since the "$o(1)$" term above is independent of both h and data, Condition 1 holds, with $V_{\theta^\dagger} = \ddot{A}(\theta^\dagger)$.

It would be possible to check Condition 2 directly if we had a particular exponential family form. For instance, if the actuary chooses an exponential prior with hazard rate 1/5 for the model in Example 2. Then the posterior Π_n will have a gamma distribution with a shape parameter $3n + 1$ and rate parameter $1/5 + \sum_{i=1}^n X_i^{-1}$. Then we have closed-form expressions for the mean and variance and the strategy outlined above can be followed to confirm that (8) holds. But for a general exponential family, we need to check Condition 2 indirectly using B1–B4. For these exponential family models, B1 is a consequence of B2. For B2 in this case, it is easy to check that we can take

$$m_{\theta^\dagger}(x) = |x| + \text{constant},$$

so the integrability assumption in B2 holds if the tails of $P^\star$ are sufficiently thin. Next, it is easy to check that $M_{\theta^\dagger}$ is just a covariance matrix of $P^\star$ so, existence of an inverse is only a mild assumption. So it turns out that B4 is the only non-trivial sufficient condition to check. We do not give all the details here, but we do believe that it is worth describing the test construction, etc. (Note that the construction of such tests is straightforward when the model is well-specified; it is the model misspecification that complicates the matter.)

The null hypothesis is $P^\star$ and the alternative is $\{Q_\theta : |\theta - \theta^\dagger| > \varepsilon\}$, for fixed $\varepsilon > 0$, and a classical Neyman–Pearson style likelihood ratio test is an obvious choice. Note that, in the ratio $p^\star/q_\theta$, the true density gets canceled out, so it is effectively just a comparison between two distributions in the model, namely, P_θ and $P_{\theta^\dagger}$. Therefore, the intuition is that we would reject the null hypothesis if the data are more consistent with $\{P_\theta : |\theta - \theta^\dagger| > \varepsilon\}$ than with $P_{\theta^\dagger}$. Now for the details. In our one-parameter setting, it is enough to test $\{Q_\theta : \theta = \theta^\dagger \pm \varepsilon\}$ so let us focus on testing $P^\star$ versus $Q_{\theta^\dagger+\varepsilon}$; the other case can be handled in an analogous way, and we will put the two together below. Define the test

$$\psi_n^+ = \begin{cases} 1 & \text{if } n^{-1}\sum_{i=1}^n \log\{p^\star(X_i)/q_{\theta^\dagger+\varepsilon}(X_i)\} < 0 \\ 0 & \text{otherwise.} \end{cases}$$

By the law of large numbers,

$$\frac{1}{n}\sum_{i=1}^n \log\frac{p^\star(X_i)}{q_{\theta^\dagger+\varepsilon}(X_i)} \to K(p^\star, p_{\theta^\dagger+\varepsilon}) - K(p^\star, p_{\theta^\dagger}), \quad P^\star\text{-almost surely,}$$

Since the limit above positive by definition of $\theta^\dagger$, the first conclusion in B4 holds, i.e., $P^{\star n}\psi_n^+ \to 0$. Establishing the second conclusion in B4, namely, $Q_\theta^n(1-\psi_n^+) \to 0$, requires a good amount of care so we refer the interested reader to the proof of Theorem 3.2 (middle of page 369) in Kleijn and van der Vaart (2012) for details. Now define a second test ψ_n^- that replaces $\theta^\dagger + \varepsilon$ with $\theta^\dagger - \varepsilon$; the same analysis above applies to this test also. Now combine these two tests as follows:

$$\psi_n = \max(\psi_n^+, \psi_n^-).$$

The two conditions in B4 hold for ψ_n because they hold for both components, ψ_n^+ and ψ_n^-. Since we have now checked all of the sufficient conditions, it follows that the Bernstein–von Mises phenomenon holds for this exponential family model.

References

Alquier, P. (2008). PAC-Bayesian bounds for randomized empirical risk minimizers. *Mathematical Methods of Statistics* 17(4), 279–304.

Alquier, P. and Ridgway, J. (2019). Concentration of tempered posteriors and of their variational approximations. *Annals of Statistics*, to appear; `https://arxiv.org/abs/1706.09293`

Bailey, A. (1950). "Credibility Procedures," *Proceedings of the Casualty Actuarial Society* XXXVII 7–23 and 94–115.

Berk, R.H. (1966). Limiting behavior of posterior distributions when the model is incorrect. *Annals of Statistics* 37, 51–58.

Bissiri, P., Holmes, C. and Walker, S. (2016). A general framework for updating belief distribution. *Journal of Royal Statistical Society, Series B–Statistical Methodology* 78(5), 1103–1130.

Brown, L.D. (1986). *Fundamentals of Statistical ExponentialFamilies with Applications in Statistical Decision Theory.* Lecture Notes-Monograph Series, Vol. 9. Hayward, CA: Institute of Mathematical Statistics.

Bühlmann, H. (1967). Experience rating and credibility. *ASTIN Bulletin* 4, 199–207.

Bühlmann, H. and Gisler, A. (2005). *A Course in Credibility Theory and its Applications*, New York: Springer.

Bunke O. and Milhaud, X. (1998). Asymptotic behavior of Bayes estimates under possibly incorrect models *Annals of Statistics* 26(2), 617–644.

Dahalhaus, R. and Wefelmeyer, W. (1996). Asymptotically optimal estimation in misspecified time series models. *Annals of Statistics* 24(3), 952–974.

De Blasi, P. and Walker, S. (2013). Bayesian asymptotics with misspecified models. *Statistica Sinica* 23, 169–187.

Diaconis, P. and Ylvisaker, D. (1979). Conjugate priors for exponential families. *Annals of Statistics* 7(2), 269–281.

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

Grünwald, P.D. (2012). The safe Bayesian: Learning the learning rate via the mixability gap. In *Proceedings of the 23rd International Conference on Algorithmic Learning Theory*, 169–183.

Grünwald, P.D. (2018). Safe Probability. *Journal of Statistical Planning and Inference* 195, 47–63.

Grünwald, P.D. and van Ommen T. (2017). Inconsistency of Bayesian inference for misspecified linear models, and a proposal for repairing it. *Bayesian Analysis* 12(4), 1069–1103.

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

Hong, L. and Martin, R. (2016). Discussion on "Credibility estimation of distribution functions with applications to experience rating in general insurance", *North American Actuarial Journal*, 20 (1), 95–98.

Hong, L. and Martin, R. (2017a). A flexible Bayesian nonparametrics model for predicting future insurance claims. *North American Actuarial Journal* 21, 228–241.

Hong, L. and Martin, R. (2017b). A review of Bayesian asymptotics in general insurance applications. *European Actuarial Journal* 7, 231-255.

Hong, L. and Martin, R. (2018). Dirichlet process mixture models for insurance loss data, *Scandinavian Actuarial Journal* 6, 545–554.

Jewell, W.S. (1975). Regularity conditions for exact credibility. *ASTIN Bulletin* 8, 336–341.

Kleijn, B.J.K. and van der Vaart, A.W. (2006). Misspecification in infinite-dimensional Bayesian statistics. *Annals of Statistics* 34(2), 837–877.

Kleijn, B.J.K. and van der Vaart, A.W. (2012). The Bernstein-Von-Mises theorem under misspecification. *Electronic Journal of Statistics* 6, 354–381.

Klugman, S. A., Panjer, H. H., and Willmot, G. E. (2008). *Loss Models: from Data to Decisions*, Third Edition. Hoboken: Wiley.

Lyddon, S., Holmes, C. and Walker, S. (2019). General Bayesian updating and the loss-likelihood bootstrap. *Biometrika* 106(2), 465–478.

Martin, R. (2019). False confidence, non-additive beliefs, and valid statistical inference. *International Journal of Approximate Reasoning* 113, 39–73.

Müller, U. K. (2013). Risk of Bayesian inference in misspecified models, and the sandwich covariance matrix. *Econometrica* 81(5): 1805–1849.

Patilea, V. (2001). Convex models, MLE and misspecification. *Annals of Statistics* 29(1), 94–123.

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

Richardson, R. and Hartman, B. (2018). Bayesian nonparametric regression models for modeling and predicting healthcare claims. *Insurance: Mathematics and Economics* 83, 1–8.

Schervish, M.J. (1995). *Theory of Statistics.* New York: Springer.

Syring, N. (2017). Gibbs posterior distributions: new theory and applications (doctoral dissertation). University of Illinois at Chicago. Retrieved from `http://hdl.handle.net/10027/22219`.

Syring, N. and Martin, R. (2017). Gibbs posterior inference on the minimum clinically important difference. *Journal of Statistical Planning and Inference* 187, 67–77.

Syring, N. and Martin, R. (2019a). Calibrating general posterior credible regions, *Biometrika* 106(2), 479–486.

Syring, N. and Martin, R. (2019b). Robust and rate-optimal Gibbs posterior inference on the boundary of a noisy image, *Annals of Statistics*, to appear, `https://arxiv.org/abs/1606.08400`.

Syring, N., Hong, L. and Martin, R. (2019). Gibbs posterior inference on value-at-risk. *Scandinavian Actuarial Journal*, to appear. `https://doi.org/10.1080/03461238.2019.1573754`.

van der Vaart, A. W. (1998). *Asymptotic Statistics.* New York: Cambridge University Press.

Wang, Z. and Martin, R. (2019). Model-free posterior inference on the area under the receiver operating characteristic curve, `https://arxiv.org/abs/1906.08296`.

Whitney, A. (1918). The theory of experience rating. *Proceedings of Casualty Actuarial Society* 4, 274–292.

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

Zhang, T. (2006b). Information theoretical upper and lower bounds for statistical estimation. *IEEE Transaction on Information Theory* 52, 1307–1321.