The Canadian Journal of Statistics
Vol. 50, No. 2, 2022, Pages 491–510
La revue canadienne de statistique

Avoiding prior–data conflict in regression models via mixture priors

Leonardo EGIDI*, Francesco PAULI, and Nicola TORELLI

Department of Economics, Business, Mathematics and Statistics, University of Trieste, Trieste, Italy

Key words and phrases: Bayesian model; generative model; mixture prior; prior–data conflict; regression.

MSC 2020: Primary 62F15; secondary 62J12.

Abstract: The Bayesian model consists of the prior–likelihood pair. A prior–data conflict arises whenever the prior allocates most of its mass to regions of the parameter space where the likelihood is relatively low. Once a prior–data conflict is diagnosed, what to do next is a hard question to answer. We propose an automatic prior elicitation that involves a two-component mixture of a diffuse and an informative prior distribution that favours the first component if a conflict emerges. Using various examples, we show that these mixture priors can be useful in regression models as a device for regularizing the estimates and retrieving useful inferential conclusions. *The Canadian Journal of Statistics* 50: 491–510; 2022 © 2021 The Authors. The Canadian Journal of Statistics/La revue canadienne de statistique published by Wiley Periodicals LLC on behalf of Statistical Society of Canada.

Résumé: Un modèle bayésien consiste à combiner une vraisemblance et une distribution a priori dans l'objectif d'estimer une distribution a posteriori pour un (ou des) paramètre(s) du modèle. On dit qu'il y a une inconsistance (ou conflit) entre les observations et la loi a priori lorsque cette dernière alloue la majeure partie de sa masse aux régions de l'espace des paramètres où la vraisemblance est relativement faible. Lorsqu'une telle situation se produit, il n'est pas toujours aisé d'y remédier. A cet effet, les auteurs proposent une solution qui offre une élicitation automatique de l'a priori, ilicitation basée sur un mélange de deux distributions a priori, l'une diffuse et l'autre informative. L'approche proposée favorisera la première composante du mélange en cas de conflit. Les auteurs illustrent l'utilité de cette approche à travers plusieurs exemples et modèles de régression tout en mettant en évidence ses capacités à servir comme dispositif et outil de régularisation des estimations et d'inférence statistique. *La revue canadienne de statistique* 50: 491–510; 2022 © 2021 Les auteurs. La revue canadienne de statistique/The Canadian Journal of Statistics, publiée par Wiley Periodicals LLC au nom de la Société statistique du Canada.

1. INTRODUCTION

The Bayesian model consists of the prior–likelihood pair (Gelman & Shalizi, 2013), and checking the model components is an essential statistical task. Recently, starting from the approaches of Box (1980), Rubin (1984) and Gelman, Meng & Stern (1996), simultaneous and joint checks of both components have been developed. However, along with these posterior predictive checks, the need for a statistical model to be *generative* has lately emerged. In particular, the ability to generate realistic hypothetical data is validated through prior-predictive checking (Gabry et al., 2019), and the plausibility of the prior is then somehow checked in a preliminary and separate way using a simulation perspective.

Additional Supporting Information may be found in the online version of this article at the publisher's website.

* *Corresponding author: legidi@units.it*

Correction added on 16 May 2022, after first online publication: CRUI-CARE funding statement has been added.

In this approach, a prior could be checked against the data and subsequently changed depending on the results of this check. Rather than merely representing the belief of the statistician before observing the data (Garthwaite, Kadane & O'Hagan, 2005), the analyst considers the prior distribution as a *device* that can convey information, regularize and suitably restrict the space of the unknown parameters (Gelman, 2017). As suggested by Gelman, Simpson & Betancourt (2017), to ensure a robust analysis we need to go beyond the standard Bayesian workflow where the prior is meant to be chosen without any reference to the data.

However, if we admit that the prior can be (judged to be) grossly inappropriate, these two model components (the prior and the sampling distribution) could contradict each other: when the prior concentrates most of its mass in low-density areas of the likelihood, we incur a *prior–data conflict* (Evans & Moshonov, 2006; Bousquet, 2008; Evans & Jang, 2011a; Al Labadi & Evans, 2017). In spite of the fact that this issue is recognized among Bayesian practitioners, remarkably few tools have been identified for resolving such a conflict once it has been identified. The Kullback–Leibler divergence criterion (Bousquet, 2008) and Bayesian *P*-values (Evans & Moshonov, 2006; Nott et al., 2016) have been developed for assessing the extent of prior–data conflict, but neither approach is used explicitly to elicit an improved prior distribution. In this article, we outline a procedure for including Bayesian *P*-values directly in the prior formulation to prevent a prior–data conflict. Following the logic outlined in Gelman et al. (2008), the usefulness of our procedure will be revealed in a regression context, where prior predictive checks highlight the importance of a careful elicitation. The main thread concerning use of a prior–data conflict to derive an improved prior beginning from an informative prior can be traced to Evans & Jang (2011a).

For the rest of this article, our working assumption will be that the sampling distribution is correct. Given a pair of priors p, q, where the first one is informative (and so could conflict with the data) and the second one is noninformative (so that it should not conflict with the data), our strategy is to combine them in a new mixture prior $\pi = \psi q + (1 - \psi)p$, and to develop an automatic procedure for estimating the mixture weight ψ in such a way that any conflict between p and the data no longer occurs. The resulting prior is then a robust alternative that lies between the informative prior p and the noninformative prior q. It would then be reasonable to choose the weight ψ such that as ψ approaches 1—meaning that a substantial prior–data conflict occurs—q is favoured; conversely, as ψ approaches 0—meaning that no conflict occurs—p is then a suitable prior.

The choice of the priors p and q is of primary importance in our approach. However, assigning a mixture prior weighted with an estimated ψ to a regression parameter does not guarantee robustness. Moreover, eliciting a rather informative prior, say a standard normal $\mathcal{N}(0, 1)$, may dramatically change its impact depending on the sampling model. In such extreme cases, we suggest and implement using a *predictive* informative prior. This is a data-driven prior distribution that depends on the sufficient statistic for the regression model, and is likely able to regularize the inferences, even when the informative prior and our proposed mixture fail. In practice, we do not believe that the dependence of our prior distribution on data is a major concern. We feel our proposed approach may be naturally located within the so-called *falsificationist Bayesian philosophy* (Gelman & Hennig, 2017), which openly deviates from subjective and objective Bayesian practices and in which the prior is open to falsification.

In a sense, it is immediate to see how the family of mixture priors $\{\psi q(\theta) + (1 - \psi)p(\theta); \psi \geq 0\}$ represents a natural priors' hierarchy *before viewing the data*; distinct priors can be in fact identified as ψ varies. Thus, our mixture prior is a device whose aim is to incorporate any possibility of prior–data conflict, allowing the absence of any such prior–data conflict as a particular case. So, it works as a sort of *built-in* prior with no need to routinely check many weakly informative priors and possibly to change them.

The rest of the article is organized as follows. Section 2 provides a quick overview of the prior–data conflict measures proposed by Evans & Moshonov (2006). In Section 3, we present our methodology, together with some theoretical results that quantify the extent of any prior–data conflict we can expect a priori when using π rather than p. We also justify the use of predictive informative priors in extreme cases. Section 4 explores several regression applications. Some concluding remarks and observations are provided in Section 5.

2. PRIOR–DATA CONFLICT: SOME BACKGROUND

We incur a prior–data conflict when we elicit a prior whose density mass concentrates on values of the parameter θ that are not supported by the data. In other words, such a conflict happens when the prior places its mass primarily on distributions in the sampling model for which the observed data are surprising (typically when only a few data points are observed). As mentioned by Evans & Moshonov (2006), Evans & Jang (2011a) and Nott et al. (2016), checking for prior–data conflicts takes a distinct perspective from verifying the appropriateness of the likelihood components.

We denote the sampling model for the data y in the sample space $\mathcal{Y}$ by $\{p(y|\theta) : \theta \in \Theta \subseteq R^d, d \geq 1\}$, where each $p(y|\theta)$ is a probability density on $\mathcal{Y}$ with respect to some support measure μ. The prior distribution $p(\theta)$ then leads to a prior predictive probability measure

$$M(A) = \int_\Theta \int_A p(y|\theta)p(\theta)\mu(dy)\nu(d\theta) = \int_A m(y)\mu(dy) \tag{1}$$

on $\mathcal{Y}$, where $A \subseteq \mathcal{Y}$, and $m(y) = \int_\Theta p(y|\theta)p(\theta)\nu(d\theta)$ is the density of M with respect to the measure μ, known as the prior predictive distribution for the sample y. For a function $T : \mathcal{Y} \to \mathcal{T}$, we may define the marginal prior predictive density

$$m_T(t) = \int_\Theta p(t|\theta)p(\theta)\nu(d\theta) \tag{2}$$

for T, where $p(t|\theta)$ is the marginal density for T. If T is a minimal sufficient statistic for the sampling model $p(y|\theta)$, it is well known that the posterior is the same whether we observe y or $T(y)$. In Evans & Moshonov (2006) and Evans & Jang (2011a), a prior–data conflict arises when the observed value $T(y_0) = t_0$ turns out to be *surprising* when compared with the distribution M_T:

$$P(t_0) = M_T(m_T(t) \leq m_T(t_0)). \tag{3}$$

The measure of surprise in Equation (3) is a prior predictive P-value according to the probability distribution M_T, and its purpose is to locate t_0 in the distribution M_T. Evans & Jang (2011b) stated a consistency result for Equation (3), proving that the limiting value for this tail probability as the data grow measures the extent to which the true value of the parameter is a surprising value with respect to the choice of the prior. If m_T is unimodal, Equation (3) represents the probability $P(t_0)$ such that the value t_0 falls in a low-density distribution area. It is really only when very small values of (3) are obtained that problems arise since then the data contradicts the prior. If $P(t_0)$ approaches zero, this means that t_0 lies in a region where M_T assigns very little probability, and then a prior–data conflict is likely to occur.

Although there may be many concerns about the use of these P-values to detect prior–data conflicts, in the rest of the article we will use $P(t_0)$ to detect a possible prior–data conflict. We will return to this point, which plays a key role in our approach, in Section 3.

3. A CLASS OF PRIORS FOR AVOIDING PRIOR–DATA CONFLICTS

3.1. The Mixture Prior

The goal is to elicit an informative prior $p(\theta)$, which we will henceforth call the *reference* informative prior for a problem of interest. Moreover, we want to avoid the possibility that this prior is in conflict with the observed data according to the P-value described in Equation (3), or that it dominates the inference when data are not fully informative. To dilute the effect of this choice, we may combine $p(\theta)$ with a noninformative prior $q(\theta)$, $\theta \in \mathbb{R}^1$ for simplicity, in a mixture prior $\pi(\theta)$ using the weight ψ, as follows:

$$\pi(\theta) = \psi q(\theta) + (1 - \psi)p(\theta). \tag{4}$$

The idea of using mixture priors to overcome a prior–data conflict was previously proposed, in the context of clinical trials, by Schmidli et al. (2014) and Mutsvari, Tytgat & Walley (2016). However, the authors were vague about the choice of the mixture weights, suggesting that this specification can be based on the degree of confidence of the clinical trial team in the relevance of the historical data, or more simply, that the larger weight, say $1 - \psi = 0.7$ or $1 - \psi = 0.9$, should be assigned to the informative prior. In our opinion, this is a *subjective* choice designed to correct a subjective source as an informative prior. To wisely use mixture priors in applied statistics, we believe the choice of the weight ψ should be automatic rather than subjective.

3.2. Choice of the Mixture Weight ψ

We propose a strategy for choosing ψ, which depends on the P-value in Equation (3) evaluated for π; in particular, ψ is such that the mixture prior does not imply a conflict. According to Equation (2), the marginal prior predictive density for the minimal sufficient statistic T under the prior π, $m_T^\pi = \int_\Theta p(t|\theta)\pi(\theta)\nu(d\theta)$, yields the prior predictive probability measure $M_T^\pi = \int_A m_T^\pi(t)\mu(dt)$. Using the mixture prior identified in Equation (4), it follows at once that

$$m_T^\pi = \psi m_T^q + (1 - \psi)m_T^p. \tag{5}$$

Given the observed statistic value $T(y_0) = t_0$, we propose to choose the smallest value of ψ such that the P-value $P^\pi(t_0) = M_T^\pi(m_T^\pi(t) \leq m_T^\pi(t_0))$ exceeds the threshold α and, consequently, a prior–data conflict no longer exists, i.e.,

$$\psi = \inf\{\psi | P^\pi(t_0) \geq \alpha\}. \tag{6}$$

In this context, α acts as a tuning parameter; its choice is connected to the degree of flexibility assumed by the experimenter. Evans & Moshonov (2006) do not address this issue; they only recognize a conflict in examples where the P-value is at most 0.05. Evans & Jang (2011a) allude to the fact that α is usually some cut-off value that depends on the application. Gelman et al. (2013) are even more vague, suggesting that a discrepancy in a statistic is found when its observed value falls in one of the tails of its replicated distribution.

The goal is not a decision about the existence of a conflict (the method is applied when there are clues that a conflict might occur), but it is vital for us to define a prior that incorporates the absence of a conflict as a particular case, and according to which the degree of this conflict (choice of α) may vary based on the individual cases and the investigator's judgement.

3.3. Specifying the Priors

The final degree of freedom provided to users of our approach is the choice of q, the noninformative prior. Although many definitions of noninformative priors have been proposed

in the past (Kass & Wasserman, 1996; Consonni et al., 2018), it is sufficient for our purposes to consider the absence of the possibility of any prior–data conflict as a necessary characteristic of any noninformative prior, as suggested by Evans & Moshonov (2006). In the examples in Section 4, we use the weakly informative priors proposed by Gelman et al. (2008), which, as shown in Evans & Jang (2011a), are unlikely to cause any prior–data conflict.

However, we need to characterize when the mixture prior identified in Equation (4) is weakly informative with respect to p; we do so by following the procedure proposed by Evans & Jang (2011a). To assess whether a base prior q is weakly informative with respect to an elicited prior p, these authors suggest evaluating

$$M_T^p(P^q(t_0) \leq x_\gamma), \tag{7}$$

where $P^q(t_0) = M_T^q(m_T^q(t) \leq m_T^q(t_0))$ is the P-value used to check whether or not there is prior–data conflict with respect to q, and $x_\gamma \in [0, 1]$ is a quantile of the distribution of $P^p(t_0)$. As Evans and Jang remark, if $m_T^p(t_0)$ has a continuous distribution when $t_0 \sim M_T^p$, then $x_\gamma = \gamma$. We say that q is *weakly informative relative to p* at level γ if the value identified in Equation (7) is at most x_γ. Moreover, the degree of weak informativity of a prior q relative to a prior p can be quantified via the ratio

$$\begin{aligned} r_{pq} &\equiv 1 - M_T^p(P^q(t_0) \leq x_\gamma)/x_\gamma \\ &= x_\gamma - M_T^p(P^q(t_0) \leq x_\gamma)/x_\gamma \\ &= M_T^p(P^p(t_0) \leq x_\gamma) - M_T^p(P^q(t_0) \leq x_\gamma)/x_\gamma, \end{aligned} \tag{8}$$

where the final equality holds under the assumption that the P-value $P^p(t_0)$ is uniformly distributed when $t_0 \sim M^p(t_0)$, as suggested by Evans & Jang (2011a). The ratio specified in Equation (8) represents, as a proportion, the reduction in any prior–data conflict that we can expect, a priori, when using q rather than p. In the mixture prior context, we want to check when π is weakly informative relative to p, and how much less informative than p it is via a measure analogous to the ratio identified in Equation (8) for the two priors p, π. The following result characterizes the notion of weak informativity for the mixture prior identified in Equation (4). The proof of the theorem may be found in the Appendix.

Theorem 1. *Suppose p and q are proper prior distributions for θ, the parameter of a statistical model $p(y|\theta)$, and $\pi(\theta) = \psi q(\theta) + (1 - \psi)p(\theta)$ is the mixture prior, for any $\psi \geq 0$. Then the ratio $r_{p\pi} \equiv [M^p(P^p(t_0) \leq x_\gamma) - M^p(P^\pi(t_0) \leq x_\gamma)]/x_\gamma$, which represents, as a proportion, the reduction in any prior–data conflict that we can expect, a priori, when using π rather than p, equals $\psi(P^q(t_0) - P^p(t_0))/x_\gamma$, with the numerator $\delta_{pq} \equiv \psi(P^q(t_0) - P^p(t_0))$ bounded between 0 and 1. Moreover, π is weakly informative relative to p at level γ if and only if $\delta_{pq} > 0$, i.e., whenever $\psi > 0$ and $P^q(t_0) > P^p(t_0)$.*

This result underlines the role played by ψ in the prior–data conflict context and quantifies the extent of the reduction in any prior–data conflict we can expect when using the mixture prior π rather than the informative prior p: $\delta_{pq} \approx 0$ (more or less the same quantity of prior–data conflict raised by p) when the informative prior p is entirely weighted towards the mixture prior ($\psi \approx 0, P^p(t_0) \approx P^q(t_0)$); conversely, $\delta_{pq} \approx 1$ when the noninformative prior q is entirely weighted towards the mixture prior ($\psi \approx 1, P^p(t_0) \approx 0$). In the following example we retrieve this result and the typical meaning of noninformative priors by comparing the mixture prior with a normal prior in the case of a normal likelihood for the observed data.

Example 1. *Comparing the mixture with a normal prior.*

Suppose we collect a sample $y = (y_1, \ldots, y_n)$ from a $\mathcal{N}(\theta, 1)$ distribution. The minimal sufficient statistic is $T(y) = \bar{y} \sim \mathcal{N}(\theta, 1/n)$. Suppose that the prior p on θ is $\mathcal{N}(\theta_0, \sigma_1^2)$, whereas the prior q on θ is a $\mathcal{N}(\theta_0, \sigma_2^2)$, with θ_0, σ_1^2, σ_2^2 known. We can combine p and q in a mixture prior $\psi q(\theta) + (1-\psi)p(\theta)$, with ψ estimated and known. Then, we can compute

$$\begin{aligned} P^{\pi}(t_0) &= M_T^{\pi}(m_T^{\pi}(t) \le m_T^{\pi}(t_0)) \\ &= \psi M_T^q(m^q(t) \le m^q(t_0)) + (1-\psi) M_T^p(m^p(t) \le m^p(t_0)) \\ &= \psi \left(1 - G_1\left(\frac{(t_0-\theta_0)^2}{1/n+\sigma_2^2}\right)\right) + (1-\psi)\left(1 - G_1\left(\frac{(t_0-\theta_0)^2}{1/n+\sigma_1^2}\right)\right), \end{aligned}$$

where G_k denotes the cumulative distribution function of the chi-squared random variable with k degrees of freedom. It follows that

$$\begin{aligned} &M_T^p(P^{\pi}(t_0) \le \gamma) \\ &= M_T^p\left(\psi\left(1 - G_1\left(\frac{(t_0-\theta_0)^2}{1/n+\sigma_2^2}\right)\right) + (1-\psi)\left(1 - G_1\left(\frac{(t_0-\theta_0)^2}{1/n+\sigma_1^2}\right)\right) \le \gamma\right) \\ &= M_T^p\left(1 - G_1\left(\frac{(t_0-\theta_0)^2}{1/n+\sigma_1^2}\right) + \psi\left(1 - G_1\left(\frac{(t_0-\theta_0)^2}{1/n+\sigma_2^2}\right)\right.\right. \\ &\quad \left.\left. - \left(1 - G_1\left(\frac{(t_0-\theta_0)^2}{1/n+\sigma_1^2}\right)\right)\right) \le \gamma\right) \\ &= M_T^p\left(1 - G_1\left(\frac{(t_0-\theta_0)^2}{1/n+\sigma_1^2}\right) \le \gamma - \psi\delta_{pq}\right) \\ &= M_T^p\left(\frac{(t_0-\theta_0)^2}{1/n+\sigma_1^2} \ge G_1^{-1}\left(1-\gamma+\psi\delta_{pq}\right)\right) \\ &= 1 - G_1\left(G_1^{-1}\left(1-\gamma+\psi\delta_{pq}\right)\right) \\ &= \gamma - \delta_{pq}, \end{aligned}$$

where $\delta_{pq} = \psi\left(P^q(t_0) - P^p(t_0)\right) = \psi\left(G_1\left(\frac{(t_0-\theta_0)^2}{1/n+\sigma_1^2}\right) - G_1\left(\frac{(t_0-\theta_0)^2}{1/n+\sigma_2^2}\right)\right)$. This quantity is at most γ iff $\psi > 0$ and $P^q(t_0) - P^p(t_0) > 0$, where the latter condition applies only when $\sigma_2^2 > \sigma_1^2$, as is customary for the case of two prior distributions p and q; see Evans & Jang (2011a).

3.4. Predictive Informative Priors

The choice of the statistic $T(y)$ and the two priors p and q plays a key role in our method. Unfortunately, when a serious prior–data conflict exists, we will need to replace the prior, and that would seem to suggest some dependence on the data; after all, we are replacing the prior because of the observed data. In many applied problems, it may then be useful to elicit a prior that does not cause any prior–data conflict. From this perspective, the prior should act as a device and its choice should guarantee robust inferential conclusions. If the mixture prior that we proposed

in Section 3.1 is revealed to be in conflict with the data after a first check (thus, ψ approaches one), the analyst could be tempted to reinforce his prior assumptions by defining a *predictive informative prior* $p_T(\theta)$, whose characteristic is to be centred at the sufficient statistic. Consider the location normal model (Evans & Moshonov, 2006), where $y_i \sim \mathcal{N}(\theta, \sigma^2)$, and $\theta \sim \mathcal{N}(0, 1)$; then, the predictive informative prior is $p_T(\theta) = \mathcal{N}(t_0, 1)$, where $t_0 \equiv T(y_0) = n^{-1} \sum_i y_i$ is the observed value of the minimal sufficient statistic. In general, we propose using the formulation

$$p_T(\theta_j) = \mathcal{N}(t_{0j}, 1) \tag{9}$$

for the predictive informative prior, where t_{0j} might be the maximum likelihood estimate or any other reasonable estimate for the parameter θ_j. The rationale here is that by choosing this prior which is centred on the maximum likelihood estimate we may have a readily available correction in the direction of the data and thus obtain a more robust prior. Some possible benefits associated with this particular choice will be revealed by the examples that we consider in Section 4.

Nevertheless, we claim that this is a very strong data-dependent prior, and its use should be restricted to extreme cases. For such a reason, in terms of a natural priors' hierarchy, the users are strongly encouraged to elicit a preliminary *reference* informative prior p, and possibly replace it with a predictive informative prior in the mixture prior identified in Equation (4) after checking for the possibility of prior–data conflict using $P(t_0)$ specified in Equation (3).

3.5. The Multi-Parameter Case

When $\theta \in \mathbb{R}^p$, we have a parameter-vector $\theta = (\theta_1, \theta_2, \ldots, \theta_p)$. To implement our procedure, we can assign a mixture prior to each of the p components of θ or, alternatively, we can define an approximation of $m(y)$ for only the component of interest and obtain a *pseudo*-prior predictive distribution. For example, if only θ_1, the initial element of θ, was of interest, we could use

$$m(y|\hat{\theta}_2, \hat{\theta}_3, \ldots, \hat{\theta}_p) = \int_{\theta_1} p(y|\theta_1, \hat{\theta}_2, \hat{\theta}_3, \ldots, \hat{\theta}_p) p(\theta_1) v(d\theta_1), \tag{10}$$

where $\theta_2, \theta_3, \ldots, \theta_p$ are replaced by consistent estimates $\hat{\theta}_2, \hat{\theta}_3, \ldots, \hat{\theta}_p$. We may then define the analogous pseudo-distribution

$$m(t|\hat{\theta}_2, \hat{\theta}_3, \ldots, \hat{\theta}_p) = \int_{\theta_1} p(t|\theta_1, \hat{\theta}_2, \hat{\theta}_3, \ldots, \hat{\theta}_p) p(\theta_1) v(d\theta_1) \tag{11}$$

for the marginal prior predictive density for T. We will rely on these pseudo quantities in Section 4 for regression models that involve more than one parameter.

The prior predictive distribution specified in Equation (11) uses consistent estimates for $\theta_2, \theta_3, \ldots, \theta_p$, and this may result in using the observed data twice. However, the alternative, i.e., evaluating the multi-dimensional integral of dimension p, would be computationally costly and would require an approximation. Evans & Moshonov (2007) proposed checking individual prior components in hierarchical priors, but their approach is based upon the existence of a set of ancillary statistics in cases where, moreover, the decomposition of the prior conforms to a certain structure. In fact, they focus on exponential models and group statistical models, and, as they report, even in those contexts only certain decompositions are seen to be amenable to their methodology. As an alternative, in their Example 2 they chose to use hierarchical checking without sufficient ancillaries. First they checked the marginal prior on the variance for a location-scale normal model. The conditional prior for the mean was checked subsequently.

The advantage that our proposed approach affords is that Equation (11) may be adopted for almost any prior structure, with no distinction between hierarchical, independent and other

forms of dependent priors. Furthermore, there is no need to require the existence of any ancillary statistics at this stage. The main drawbacks of our procedure are the possible double use of the data and the sensitivity of the prior predictive distribution to different estimators.

3.6. Computational Issues and Simulations

To compute the values $P(t_0)$ specified in Equation (3), we often need to use numerical or simulation methods, since the distributions M_T, M_T^{π} are often not available in an analytical form. For this reason, we usually approximate Equation (2) by drawing hypothetical replications $y_1^{rep}, \dots, y_n^{rep}$ from $m(y)$ and obtain the simulated distributions for $m_T(t), m_T^{\pi}(t)$. We are then able to compute the P-values and the mixture weight ψ as outlined in Equation (6).

In the Supplementary Materials that accompany this article, we provide the `R` source code required to simulate hypothetical data replications and compute the mixture weights ψ for the examples discussed in Section 4.

3.7. Summary

The procedure we propose may be summarized as follows:

(a) Choose a reference informative prior $p(\theta)$ and a noninformative prior $q(\theta)$.
(b) Choose a (possibly) sufficient statistic $T(y)$.
(c) If $m_T(t)$ is not analytically tractable, draw prior predictive values $y_1^{rep}, \dots, y_n^{rep}$ from $m(y)$ and obtain the simulated distribution of $m_T(t)$.
(d) Compute the P-value identified in Equation (3) and determine ψ as outlined in Equation (6).
(e) Carry out an analysis of the observed data assuming the reference mixture $\pi(\theta) = \psi q(\theta) + (1 - \psi)p(\theta)$.
(f) If a prior–data conflict seems possible, consider using an alternative robust mixture prior; adjust $p(\theta)$ to the predictive informative prior $p_T(\theta)$, and use the predictive mixture $\pi(\theta) = \psi q(\theta) + (1 - \psi)p_T(\theta)$, obtaining a new estimate for ψ.

3.8. A Simple Example

We refer to the location normal model example which Evans & Moshonov (2006) discussed as Example 1. Suppose $y = (y_1, \dots, y_n)$ is sampled from a $\mathcal{N}(\theta, 1)$ distribution, $\theta \in \mathbb{R}^1$, and the two possible priors p and q are $\theta \sim \mathcal{N}(\theta_0, \tau^2)$, and $\theta \sim \mathcal{N}(\theta_0, c\tau^2)$, $c >> 0$, respectively. The sample mean $T(y) = \bar{y} \sim \mathcal{N}(\theta, 1/n)$ is the minimal sufficient statistic. As Evans & Jang (2011a) show, the $\mathcal{N}(\theta_0, c\tau^2)$ prior is weakly informative with respect to the $\mathcal{N}(\theta_0, \tau^2)$ prior when $c > 1$.

As Evans and Jang show, the prior predictive distribution of $\bar{y}$ with respect to the prior p is $\mathcal{N}(\theta_0, \tau^2 + 1/n)$. Given the observed value $T(y_0) = t_0$, we want to assess whether or not this value lies in one of the tails of the prior predictive distribution via the P-value:

$$M_T^p(m_T^p(t) \le m_T^p(t_0)) = 2(1 - \Phi(|t_0 - \theta_0|/(\tau^2 + 1/n)^{1/2})). \tag{12}$$

We combine p and q in the mixture prior identified in Equation (4), and obtain the P-value

$$\begin{aligned} M_T^{\pi}(m_T^{\pi}(t) \le m_T^{\pi}(t_0)) = 2[1 - (&\psi\Phi(|t_0 - \theta_0|/(c\tau^2 + 1/n)^{1/2}) \\ &+ (1 - \psi)\Phi(|t_0 - \theta_0|/(\tau^2 + 1/n)^{1/2}))]. \end{aligned} \tag{13}$$

We may adjust our reference informative prior and define the predictive informative prior $p_T(\theta) = \mathcal{N}(\bar{y}, \tau^2)$. To illustrate the benefits that derive from using our procedure, we simulate two distinct samples for y, from $\mathcal{N}(0, 1)$ and from $\mathcal{N}(10, 1)$, respectively, with the following parameter settings: $n = 100$, $\theta_0 = 0$, $\tau = 4$, $c = 100$; in addition, we fix $\alpha = 0.25$. The first

sample is fictitiously simulated under the assumption that there is no prior–data conflict between our informative prior p and the data y, whereas the second sample is generated with the explicit intention of a conflict with the informative prior centred at 0. For the first sample we obtain $M_T^p = 0.985$ and $\psi = 0$; thus, as we were expecting, a prior–data conflict does not occur. For the second sample we obtain $M_T^p = 0.01$ for the informative prior, which indicates a prior–data conflict. For the mixture prior $\theta \sim \psi\mathcal{N}(\theta_0, c\tau^2) + (1-\psi)\mathcal{N}(\theta_0, \tau^2)$, $M_T^\pi = 0.25$ with the weight ψ estimated at 0.25; $M_T^\pi = 1, \psi = 0$ for the mixture prior with the predictive prior in place of the reference p, $\theta \sim \psi\mathcal{N}(\theta_0, c\tau^2) + (1-\psi)\mathcal{N}(\bar{y}, \tau^2)$.

Even in the second sample, where a prior–data conflict occurs, adjusting the reference informative prior is not strictly required: the starting informative prior p, when weighted and combined with a diffuse prior, already prevents any prior–data conflict up to the fixed threshold α.

4. APPLICATIONS

In this section we consider three distinct examples of regression models for which the mixture priors introduced in the previous section prove to be beneficial. Although complete/quasi-complete separation in logistic regression is not strictly the cause of a prior–data conflict, nevertheless it may be viewed as the implicit cause in a broader context in which default priors are often not able to regularize the inferences and yield poor answers.

For simplicity and technical convenience, in each of the following examples we elicit the standard normal $\mathcal{N}(0, 1)$ as the reference informative prior: rather than eliciting an actual informative prior as in Al Labadi, Baskurt & Evans (2018), we are motivated to use one reference prior and checking when it does cause prior–data conflicts. Of course, the impact of such a prior varies in relation to the likelihood and the sufficient statistic (Gelman, Simpson & Betancourt, 2017).

To implement our procedure, we set $\alpha = 0.05$ and fixed the number of hypothetical data replications required to compute the P-values identified in Equation (3) to be 10^3 in each example. These values were chosen following some sensitivity tests. Computational steps and other details may be found in the Supplementary Material, i.e., `R` source code that accompanies this article.

4.1. Logistic Regression and Separation

Experiments such as clinical trials may contain much historical information and the analyst may rely on this source of past information, considering it as a sort of baseline for similar and future studies. This is the underlying mechanism in Bayesian inference: given sequential observation of the data points y_1 and y_2 collected at two distinct times t_1 and t_2, each with sampling distribution $p(y|\theta)$, the posterior $p(\theta|y_1)$ usually acts as the prior for the new data point y_2, and the update is then proportional to $p(\theta|y_1, y_2) \propto p(\theta|y_1)p(y_1, y_2|\theta)$.

Now consider the following imaginary—but realistic—situation typical of causal inference, where the parameter θ represents the probability of contracting a particular disease. Suppose we want to collect the binary response y_{ij}, our dependent variable, where $y_{ij} = 1$ if the ith subject in the jth sample has the disease, and $y_{ij} = 0$ otherwise. We suppose that for each selected subject we also collect some individual predictors, x_{ij} and z_{ij}, say the plasma level of a protein of interest and the sex of the subjects in the sample, respectively ($z_{ij} = 1$ if the ith subject in the jth sample is a male, 0 otherwise). We, the analysts, collect a total of five samples, each of length $n = 100$ during the year 2019 in the same hospital, assuming that there are no subjects' ties: each sample $y_1, \ldots, y_5$ is associated with the predictors $(x_1, z_1), \ldots, (x_5, z_5)$. However, we immediately realize a quasi-complete separation (Zorn, 2005; Gelman et al., 2008; Sauter & Held, 2016) scenario arising in the fifth sample: each male in the fifth sample is not affected by the disease, regardless

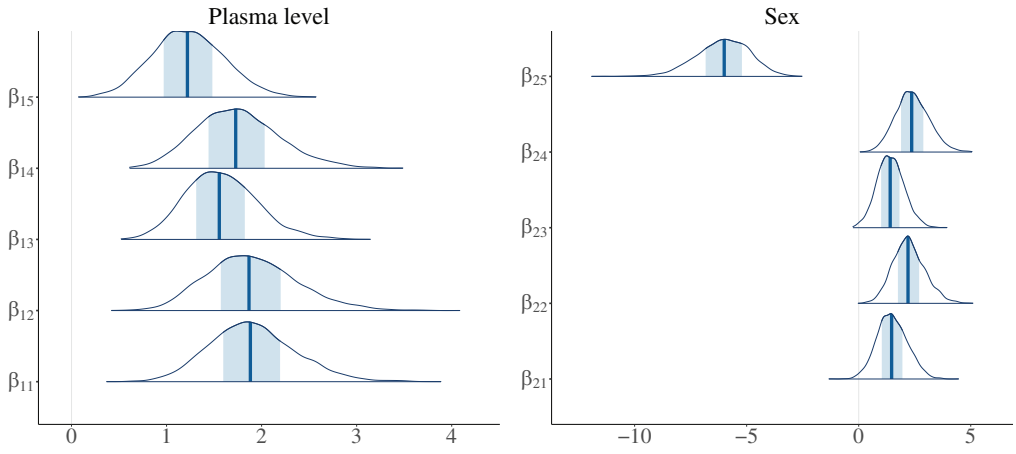


FIGURE 1: Logistic regression, five separate analyses with weakly informative priors: posterior marginal distributions along with 50% intervals (light blue areas) from separate posterior analyses of the coefficients β_1, β_2 across the five different experiments (`rstanarm` package, 2000 iterations).

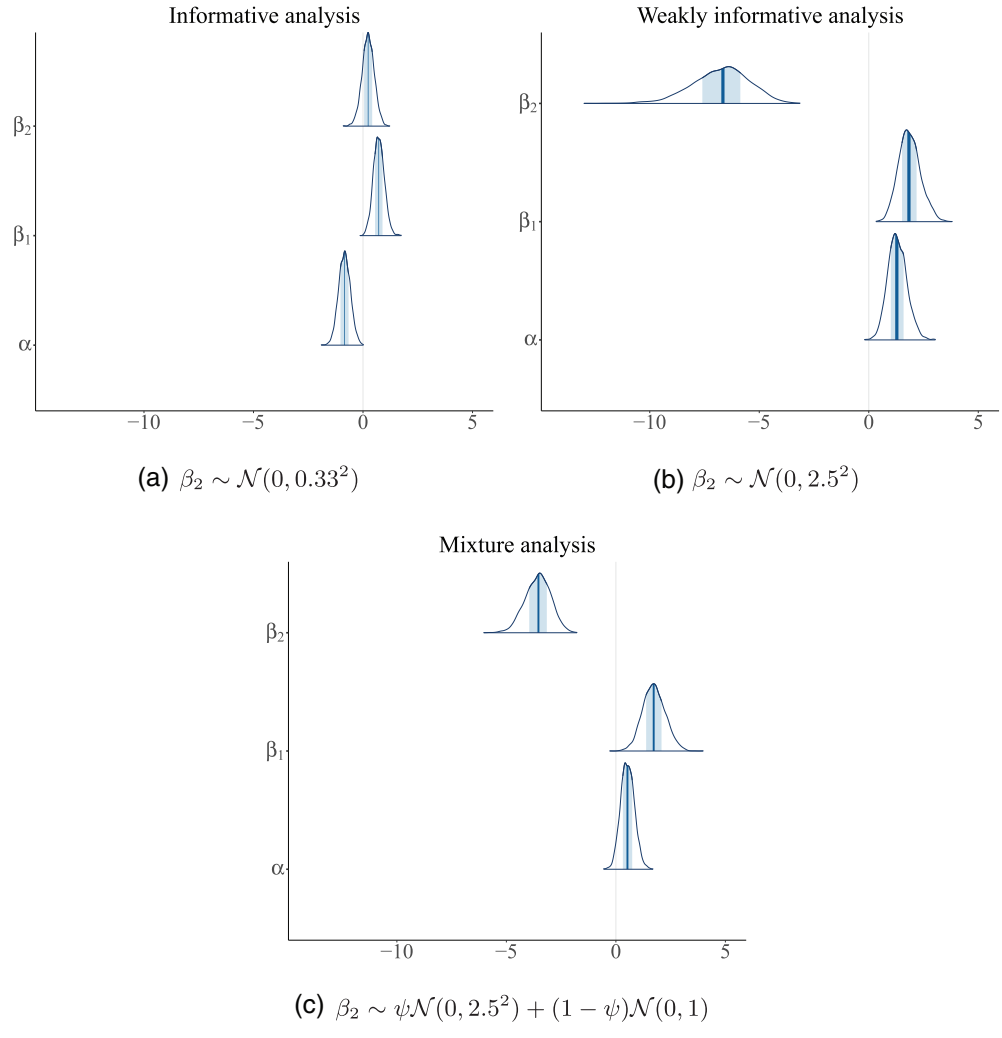
of the plasma level x_5. Thus, one of the model's covariates almost perfectly predicts the outcome variable y_5, i.e., $y_{i5} = 0$ if $x_{i5} = 1$, whereas y_{i5} can be 0 or 1 if $x_{i5} = 0$.

We are asked to perform a Bayesian logistic regression at the end of 2019, assuming that the response probability $p_{ij} \equiv \Pr(Y_{ij} = 1)$ associated with the ith patient in the jth sample is

$$\text{logit}(p_{ij}) = \alpha + \beta_1 x_{ij} + \beta_2 z_{ij},$$

where $\text{logit}(x) = \log(x/(1-x)),\ x \in [0,1]$. As a preliminary attempt, we decide to perform five logistic regressions treating each experiment as if it was *independent* of the others. Let $\beta_{11}, \beta_{12}, \ldots, \beta_{15}$ and $\beta_{21}, \beta_{22}, \ldots, \beta_{25}$ denote the regression parameters β_1 and β_2 associated with plasma level and sex, respectively, which correspond to the five measurements. Figure 1 shows the resulting posterior intervals (50% credibility areas are coloured in light blue) for β_1 and β_2 obtained with the `R` package `rstanarm` (Goodrich et al., 2018) using the default weakly informative priors: β_1, which is displayed in the left panel, is rather similar across the five samples, whereas in the right panel of the same figure, β_{25} has the opposite sign to $\beta_{21}, \ldots, \beta_{24}$, due to separation.

Now consider fitting the Bayesian logistic regression for the fifth experiment conditional on what we observed in the previous four experiments. Initially we decide to carry out our analysis without worrying about separation in the data. If we are fully informative Bayesian analysts, we should use the posterior derived from the four experiments as the new prior for the fifth experiment, and then update the results. Otherwise, we could use a standard weakly informative prior as suggested by Gelman et al. (2008) for the logistic regression, say a $\mathcal{N}(0, 10^2)$ for the intercept and $\mathcal{N}(0, 2.5^2)$ for the parameters β_1 and β_2. Posterior 50% intervals and marginal posterior distributions from the two analyses are reported in the top row of Figure 2. The posterior distributions for the parameter β_2 are completely different in the two frameworks. It appears that the weakly informative prior favours the probability of no disease in males too strongly. For a male subject ($z_{i5} = 1$), the estimated odds ratio that we obtain using the posterior median is $p_{i5}/(1-p_{i5}) = \exp(-5.5) = 0.004$; as we will see, we feel that this posterior estimate could dramatically underestimate the probability of disease for a male subject, especially if this



(a) $\beta_2 \sim \mathcal{N}(0, 0.33^2)$

(b) $\beta_2 \sim \mathcal{N}(0, 2.5^2)$

(c) $\beta_2 \sim \psi\mathcal{N}(0, 2.5^2) + (1-\psi)\mathcal{N}(0, 1)$

FIGURE 2: Logistic regression, fifth experiment: posterior marginal distributions along with 50% intervals (light blue areas) from analysis of the coefficients β_1, β_2 in the fifth experiment, under (a) informative priors, (b) weakly informative prior and (c) mixture prior with reference informative prior, where $T(y, z) = \sum_{i=1}^{n} y_{i5} z_{i5} = 0$ and $\psi = 0$.

result is used as prior information for future studies. Conversely, the informative prior for β_2, $\beta_2 \sim \mathcal{N}(0, 0.33^2)$, estimated from the first four experiments is likely to conflict with the data, with an odds ratio of about 1.64 when $z_{i5} = 1$. We need something in between. To implement our mixture prior

(a) First we need to choose the sufficient statistic with respect to β_2: $T(y, z) = \sum_{i=1}^{n} y_{i5} z_{i5}$ has observed value $T(y_0, z_0) = 0$ for the fifth experiment.
(b) Run our source code to (i) draw hypothetical values from $m(y|\hat{\alpha}, \hat{\beta}_1, \beta_2)$, with $\hat{\alpha}$ and $\hat{\beta}_1$ consistent estimates for α and β, and (ii) compute the weights. We obtain $\psi = 0$, meaning no weight is associated with the prior q.
(c) Choose $p(\beta_2) = \mathcal{N}(0, 1)$, $q(\beta_2) = \mathcal{N}(0, 2.5^2)$.

(d) Run the Bayesian logistic regression with the following mixture prior for β_2: $\beta_2 \sim \psi\mathcal{N}(0, 2.5^2) + (1-\psi)\mathcal{N}(0,1)$.
(e) In such a case, since $T(y_0, z_0) = 0$, the predictive informative prior $p_T(\beta_2)$ coincides with the reference informative prior $p(\beta_2)$.

The 50% posterior intervals and marginal posterior distributions from our procedure are displayed in the lower panel of Figure 2. The posterior interval for β_2 is sensibly narrower than the same interval under the weakly informative prior; moreover, the posterior median, about -3.5, makes more sense since it represents a compromise between the median for β_2 under the strongly informative analysis (about 0.5) and the same under the weakly informative analysis (about -5.5). As a further confirmation, the odds ratio for a male is about 0.03 under the mixture prior, somehow lying between the unrealistic value 0.004 (weakly informative prior) and 1.64 (informative prior).

As a final comment, we feel that our procedure is even more robust than the weakly informative prior in case of separation arising in logistic regression. In this case, the standard normal prior absorbs the information required to obtain meaningful posterior estimates, and thus there is no need to use the adjusted predictive prior.

4.2. A Bioassay Experiment

We consider now a well-known small-sample experiment previously analyzed by Racine et al. (1986) and Gelman et al. (2008), in which the choice of a prior may strongly affect the final inference. Table 1 summarizes the data collected from 20 animals that were exposed to four different doses of a toxin, where x_i represents the ith of k dose levels, measured on a logarithmic scale, given to n_i animals, of which y_i died. We assume the typical binomial model

$$y_i | p_i \sim \text{Bin}(n_i, p_i),$$

where p_i represents the probability of death for animals given dose x_i, and

$$\text{logit}(p_i) = \alpha + \beta x_i.$$

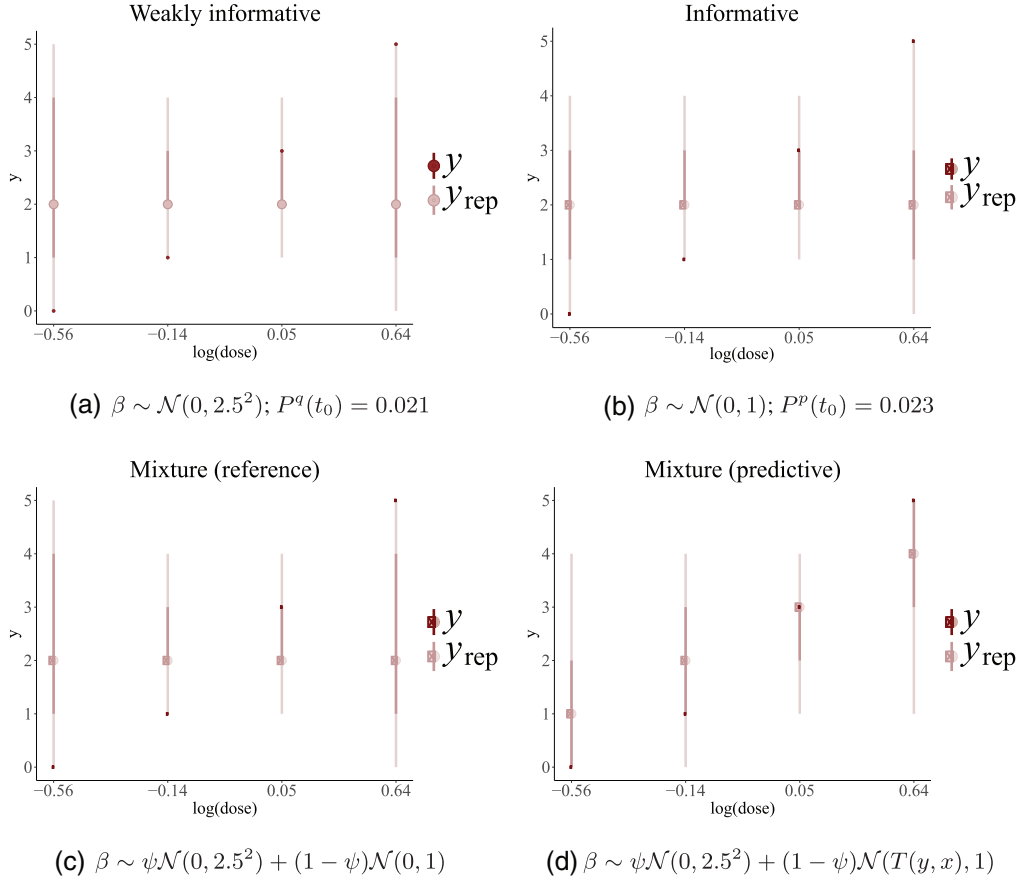
As suggested by Racine et al. (1986), prior information may be available either in the form of the results of a previous experiment using the same substance or in the form of assessments elicited from one or more expert toxicologists.

When the sample size is small, the role of the prior may be particularly relevant. In this application, we want to combine two aspects of the prior specification. On one hand, the prior should not be in conflict with the observed data. Nonetheless, a Bayesian model should always be generative, and simulations from the prior predictive distribution should be reasonable.

Consider first the scenario where substantial information may not be incorporated in the prior, leading to the elicitation of weakly informative priors, namely $\alpha \sim \mathcal{N}(0, 10^2)$ and $\beta \sim \mathcal{N}(0, 2.5^2)$. The log(dose) is rescaled to have mean 0 and standard deviation 0.5. As is evident from Figure 3a, where we have displayed four predictive intervals, one for each of the four values of x, fake data generated under these priors are meaningless in this small-sample application: the intervals range from 0 to 5, covering the entire support of the observed data, and posterior medians are constant with respect to x. Thus, even in the absence of substantial prior information, weakly informative priors are too vague in this context and do not yield useful replications under the assumed prior marginal distribution. The same result is obtained for the reference informative prior $\mathcal{N}(0, 1)$ (Figure 3b) for the same bioassay experiment.

TABLE 1: Bioassay experiment: x_i is the dose (log g/ml), n_i the number of animals and y_i the number of deaths.

Dose x_i	Animals n_i	Deaths y_i
−0.86	5	0
−0.30	5	1
−0.05	5	3
0.73	5	5



(a) $\beta \sim \mathcal{N}(0, 2.5^2)$; $P^q(t_0) = 0.021$

(b) $\beta \sim \mathcal{N}(0, 1)$; $P^p(t_0) = 0.023$

(c) $\beta \sim \psi\mathcal{N}(0, 2.5^2) + (1-\psi)\mathcal{N}(0, 1)$

(d) $\beta \sim \psi\mathcal{N}(0, 2.5^2) + (1-\psi)\mathcal{N}(T(y, x), 1)$

FIGURE 3: Bioassay experiment: medians (light red dots) and predictive intervals from the prior marginal distribution against the observed y_i (dark red dots), where $\alpha \sim \mathcal{N}(0, 10^2)$, assuming for β (a) weakly informative prior, (b) informative prior, (c) mixture with reference informative prior, where $\psi = 0.5$ and (d) mixture with predictive prior, where $T(y, x) = \sum x_i y_i = 3.24$ and $\psi = 0.3$.

TABLE 2: `Prostate` dataset: list of covariates.

Variable	Description
lpsa	Level of prostate-specific antigen
lcavol	Log(cancer volume)
lweight	Log(prostate weight)
age	Age
lbph	Log(benign prostatic hyperplasia amount)
svi	Seminal vesicle invasion
lcp	Log(capsular penetration)
gleason	Gleason score
pgg45	Percentage Gleason scores 4 or 5

Evans & Jang (2011a) (see Figure 4a in their paper) have suggested that this reference informative prior is not more informative than $\mathcal{N}(0, 2.5^2)$ and could instead be considered the standard noninformative prior for this problem when one focuses on the probabilities instead of the regression coefficients (Al Labadi, Baskurt & Evans, 2018). Moreover, even if $P^q(t_0) = 0.1073$ when $q(\alpha, \beta) = \mathcal{N}(0, 10^2) \times \mathcal{N}(0, 2.5^2)$ according to Evans & Jang (2011a), the degree of prior–data conflict raised by the weakly informative $q(\alpha, \beta)$ and the informative prior $q(\alpha)p(\beta) = \mathcal{N}(0, 10^2) \times \mathcal{N}(0, 1)$ amounts in our case to 0.021 and 0.023, respectively. This misalignment with their result is justified by the fact that they use m_T to compute the P-value $P^q(t_0)$ identified in Equation (3), whereas we use the pseudo-prior predictive distribution $m(t|\hat{\alpha}, \beta)$ specified in Equation (11), with $\hat{\alpha}$ estimated from the data. As Figure 3a,b shows, both these priors are far from the data, they are centred at zero, and this choice means the binomial model in this small-sample scenario cannot fulfill its generative function when these priors are chosen.

We need perhaps a prior able to regularize the inferences and to provide plausible replications from the prior predictive distribution in the context of small datasets. To implement the mixture prior that we advocated in Section 3, we need a sufficient statistic for the model, such as $T(y, x) = (\sum_{i=1}^4 x_i y_i, \sum_{i=1}^4 y_i)$. Thus, we need to estimate the weight ψ such that the mixture prior does not lead to a prior–data conflict. Using the reference mixture prior $\beta \sim \psi q(\beta) + (1 - \psi)p(\beta)$ (see Figure 3c), where $q(\beta) = \mathcal{N}(0, 2.5^2)$, $p(\beta) = \mathcal{N}(0, 1)$, the estimated weight is $\psi = 0.5$, which does not improve the situation. Thus, our reference informative prior $p(\beta)$ is not generative, and hence is not very informative with respect to the parameter β, and is therefore likely to lead to a prior–data conflict. In such a situation, we definitely need to use the predictive informative prior, $\beta \sim \psi\mathcal{N}(0, 2.5^2) + (1 - \psi)\mathcal{N}(\sum_{i=1}^4 x_i y_i, 1)$. The predictive intervals displayed in Figure 3d, where $\psi = 0.3$, are now narrower, and the posterior medians clearly vary with the different dose levels, replicating the pattern observed in the original data.

Marginal posterior distributions and posterior 50% intervals for β that were obtained using weakly informative, reference mixture, and predictive mixture priors are displayed in Figure 4; the distributions obtained under weakly informative and reference mixture coincide, whereas the marginal posterior under the predictive mixture prior yields narrower posterior intervals. In such a case, the reference prior $\mathcal{N}(0, 1)$ is not able to regularize the estimates, and using the predictive mixture prior is clearly preferable.

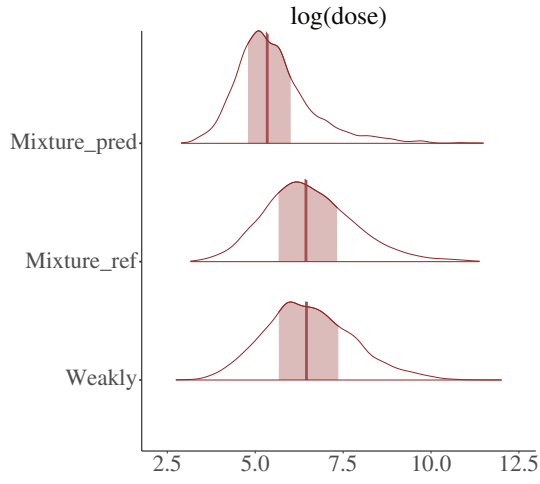


FIGURE 4: Bioassay experiment: marginal posterior distribution and posterior 50% intervals for β under weakly informative, reference mixture, and predictive mixture priors (see the text for the details).

4.3. Linear Regression with Multiple Predictors

The dataset `Prostate` in the `R` package `lasso2` was used by Stamey et al. (1989) and Tibshirani (1996) to investigate the correlation between the level of prostate-specific antigen and other covariates for men who were about to undergo a radical prostatectomy. See Table 2 for a full list of the covariates. We assume a simple linear model for the response measurement y_i representing the amount of prostate-specific antigen as the dependent variable:

$$y_i = \beta_1 + \sum_{j=1}^{p} \beta_{j+1} x_{ij} + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2), \quad i = 1, \ldots, n. \tag{14}$$

The variate x_{ij} denotes the value of the jth covariate for the ith unit, and each β_j is an unknown regression coefficient.

A natural first approach here is to use LASSO (least absolute shrinkage and selection operator) regression as developed in Tibshirani (1996) to shrink toward zero a subset of coefficients that are not associated with influential predictors. LASSO estimates $\pm$ standard errors are displayed in Figure 5; the coefficients β_4, β_5, β_7, β_8 and β_9, associated with age, lbph, lcp, gleason and pgg45, respectively, are shrunk toward zero. The plot reveals a possible identifiability problem with the intercept β_1, which has a standard error that is large when compared with the standard errors of the other regression coefficients. A rough solution could be to drop the intercept term from the model, but this could yield undesirable effects in the global model and, in general, a lack of interpretation for the remaining coefficients.

4.3.1. Weakly informative priors

Following Gelman et al. (2008), we assigned weakly informative priors to each of the coefficients in the regression: the intercept $\beta_1 \sim \mathcal{N}(0, 10^2)$, whereas $\beta_j \sim \mathcal{N}(0, 2.5^2)$ for $j = 2, \ldots, 9$, and $\sigma \sim$ Exponential(1). We fit the model with the `rstanarm` package, specifying 2000 Hamiltonian Monte Carlo simulations and checking the convergence of the Markov chains using the Gelman–Rubin statistic $\hat{R}$ ($\hat{R} \leq 1.1$ for all the parameters). Figure 6a displays the resulting

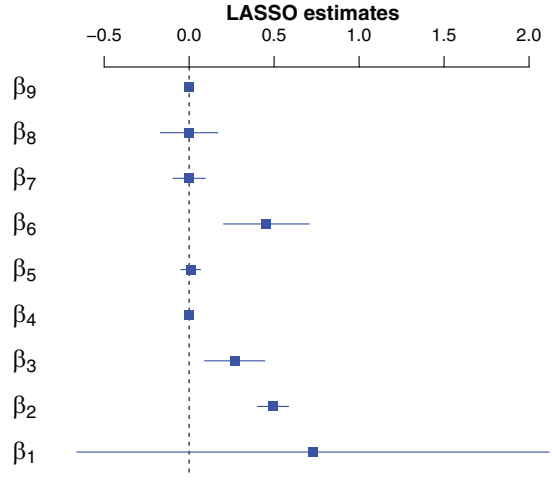


FIGURE 5: LASSO estimates for β (estimate $\pm$ standard error) parameters, `Prostate` dataset. Model fit obtained via the `R` package `lasso2`.

posterior intervals for the components of the β vector; the Bayesian model with underlying weakly informative priors for the regression coefficients results in posterior interval estimates that are rather similar to the corresponding LASSO-based intervals. The regression coefficients β_4, β_5, β_7, β_8 and β_9, associated with age, lbph, lcp, gleason and pgg45, respectively, are all shrunk toward zero, whereas β_2, β_3 and β_6, associated with lcavol, lweight and lbph, respectively, are greater than zero. The estimate of the intercept β_1 reveals a problem; the parameter is not identifiable. The suspicion here is that a prior–data conflict arose with respect to the parameter β_1, and the data are not fully informative. To fix the conflict and properly estimate the intercept, we need a suitable remedy.

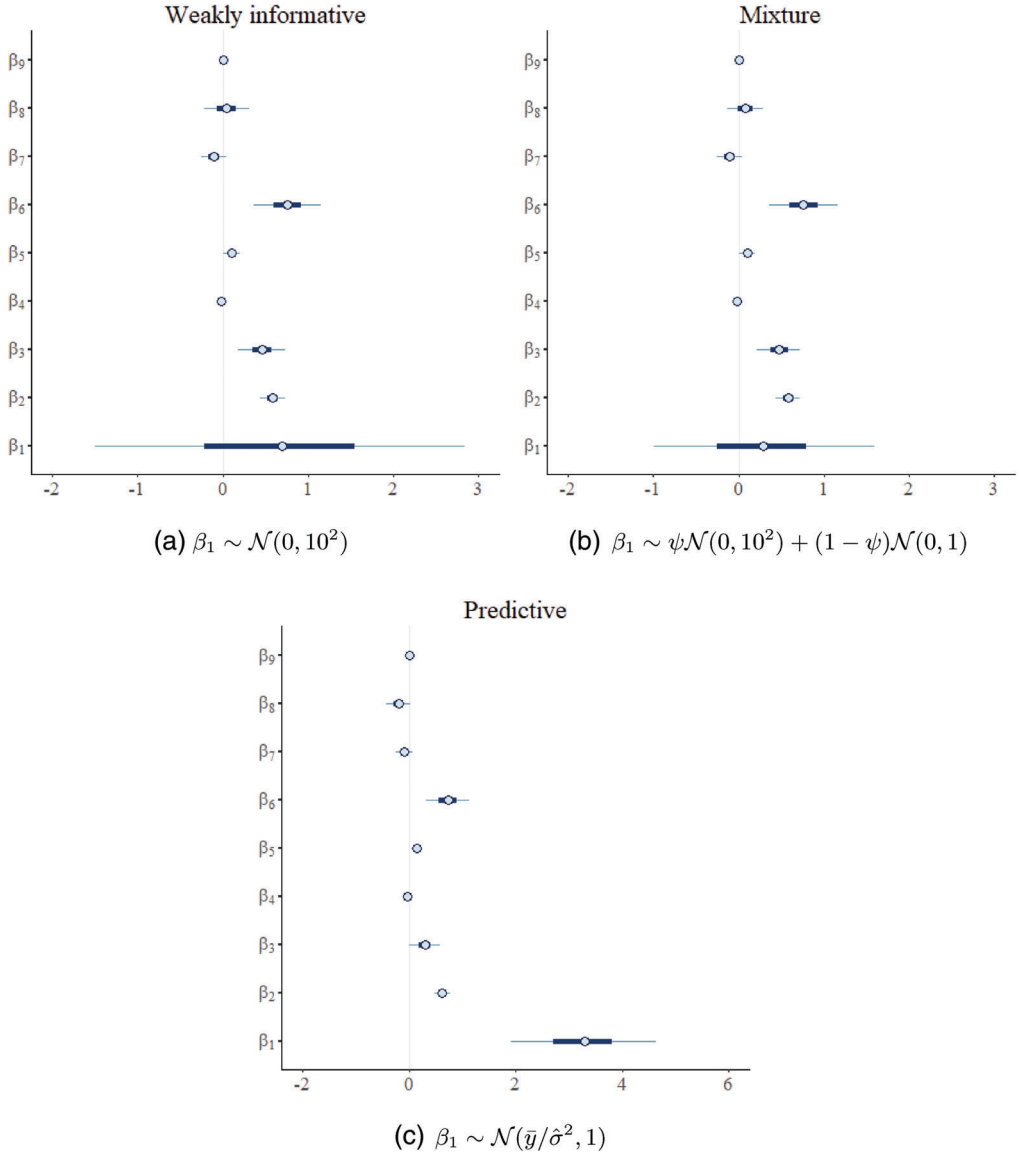
4.3.2. Mixture priors

We now must choose the informative and the diffuse prior. For the latter, we end up selecting the same weakly informative prior $\mathcal{N}(0, 10^2)$ used in the previous analysis; for the former, we start with a reference standard normal prior $\mathcal{N}(0, 1)$, and then eventually update it using the procedure we described in Section 3.7.

To implement the mixture prior, we refer to Equation (10) and consider the pseudo-prior predictive distribution $m(y|\hat{\beta}_2, \hat{\beta}_3, \dots, \hat{\beta}_9)$. The main steps are the following:

(a) Choose the sufficient statistic, $T(x, y) = X^T y$, where X is the $n \times (p + 1)$ predictor matrix.
(b) Run our source code to sample hypothetical replications from $m(y|\hat{\beta}_2, \hat{\beta}_3, \dots, \hat{\beta}_9)$ under the pseudo-prior predictive distribution and estimate ψ. We obtain $\psi = 0$.
(c) Run the linear regression with the reference mixture prior $\beta_1 \sim \psi\mathcal{N}(0, 10^2) + (1 - \psi)\mathcal{N}(0, 1)$.
(d) Consider the predictive prior distribution $p_T(\beta_1) = \mathcal{N}(\bar{y}/\hat{\sigma}^2, 1)$, where $\hat{\sigma}^2$ denotes an estimate for σ^2. From the LASSO model, we obtained $\hat{\sigma}^2 = 0.52$.
(e) Carry out the linear regression with $\beta_1 \sim \psi\mathcal{N}(0, 10^2) + (1 - \psi)\mathcal{N}(\bar{y}/\hat{\sigma}^2, 1)$; the estimated value of ψ was equal to 0.

Figure 6b displays the resulting posterior intervals for the β parameters that we obtained using the reference mixture prior $\beta_1 \sim \psi\mathcal{N}(0, 10^2) + (1 - \psi)\mathcal{N}(0, 1)$, with $\psi = 0$. The interval



(a) $\beta_1 \sim \mathcal{N}(0, 10^2)$ (b) $\beta_1 \sim \psi\mathcal{N}(0, 10^2) + (1-\psi)\mathcal{N}(0, 1)$

(c) $\beta_1 \sim \mathcal{N}(\bar{y}/\hat{\sigma}^2, 1)$

FIGURE 6: Posterior 50% intervals (dark blue segments) and 95% intervals (light blue segments) for β parameters of the multiple linear regression model for the `Prostate` dataset. $\beta_2, \ldots, \beta_9 \sim \mathcal{N}(0, 2.5^2)$. The intercept β_1 is assigned: (a) weakly informative prior, (b) mixture prior, $\psi = 0$ and (c) predictive mixture, $\psi = 0$. `rstan` package (Stan Development Team, 2018a), 2000 HMC iterations.

for the intercept is narrower than the corresponding interval estimate that we obtained using the weakly informative prior. Clearly, the estimate of β_1 is more stable. Figure 6c displays the posterior intervals for the β parameters under the predictive mixture prior $\beta_1 \sim \psi\mathcal{N}(0, 10^2) + (1-\psi)\mathcal{N}(\bar{y}/\hat{\sigma}^2, 1)$, with $\psi = 0$. Evidently, the posterior estimates for the regression coefficients $\beta_2, \ldots, \beta_9$ are almost unchanged with respect to LASSO, weakly informative, and reference mixture analyses. However, the intercept β_1 is now estimated with much more precision, and

the 95% posterior interval does not contain zero. Somehow, we added the essential information required to estimate β_1, and we checked that this information was actually relevant by simulating hypothetical data from $m(y|\hat{\beta}_2, \hat{\beta}_3, \dots, \hat{\beta}_9)$. As a final remark, we feel that choosing to assign weakly informative priors may not prevent poor estimation when there is a (partial) lack of information in the observed data.

5. DISCUSSION

How to proceed once a prior–data conflict is detected is a tricky question to answer. Moreover, there are no automatic procedures for dealing with an eventual lack of robustness of the posterior estimates and also with a lack of parameter identifiability that arises from the marginal posterior distributions. However, the prior is an essential tool for regularizing inferences. To achieve these goals, we proposed a two-component mixture model that combines an informative and a noninformative prior such that a prior–data conflict between the data and the informative prior is avoided. Our approach is based on the prior–data conflict measures developed by Evans & Moshonov (2006) and offers a new insight into a reasoned elicitation. If the mixture prior is not capable of fixing the issue, we are able to extend the method and consider a predictive prior in place of the reference informative prior chosen before the experiment. We justify our proposed priors by providing theoretical tools that measure the degree of informativity with respect to a reference informative prior. In terms of a broader interpretation, the family of mixture priors $\{\psi q(\theta) + (1 - \psi)p(\theta); \psi \geq 0\}$ represents a natural hierarchy of priors *before seeing the data*; distinct priors can be identified as ψ varies.

As motivated by the applications, this class of priors could be beneficial for regression models where prior–data conflicts may arise with respect to a subset of parameters, and the resulting inference may be misleading if the priors are even slightly misspecified. Generally speaking, use of our proposed mixture of priors seems to regularize the inference in a broad sense, not just in the case when a prior–data conflict arises.

One major concern in our method is that we advocate choosing data-dependent priors. However, data-dependent priors are widely used in applied statistics with convincing motivations (Wasserman, 2000; Gelman et al., 2008; Goodrich et al., 2018), and we feel our mixture prior $\psi\ q(\theta) + (1 - \psi)p(\theta)$ somehow depends on the observed data in a marginal sense; the mixture weight ψ is chosen using a prior-predictive check that is carried out *before* the model is fitted (Box, 1980; Gabry et al., 2019). In addition, this prior-predictive check is the same tool adopted by Evans & Moshonov (2006) and Evans & Jang (2011a) to assess whether or not a prior–data conflict has arisen and, if so, to replace the prior. Thus, as argued by Gelman et al. (2008), we do not believe that the dependence of our prior distribution on the observed data represents a major concern, because the inferential conclusions are derived from proper posteriors.

Further research is warranted to implement our proposed approach in more complex settings, such as hierarchical models, and to provide appropriate computational software that is easy to use with such a choice of prior. Some tools for checking the robustness of our proposed approach would also be desirable.

ACKNOWLEDGEMENTS

Open Access Funding provided by Universita degli Studi di Milano-Bicocca within the CRUI-CARE Agreement.

REFERENCES

Al Labadi, L. & Evans, M. (2017). Optimal robustness results for relative belief inferences and the relationship to prior–data conflict. *Bayesian Analysis*, 12, 705–728.

Al Labadi, L., Baskurt, Z., & Evans, M. (2018). Statistical reasoning: Choosing and checking the ingredients, inferences based on a measure of statistical evidence with some applications. *Entropy*, 20, 289.

Bousquet, N. (2008). Diagnostics of prior–data agreement in applied Bayesian analysis. *Journal of Applied Statistics*, 35, 1011–1029.

Box, G. E. P. (1980). Sampling and Bayes' inference in scientific modelling and robustness. *Journal of the Royal Statistical Society Series A*, 143, 383–430.

Consonni, G., Fouskakis, D., Liseo, B., & Ntzoufras, I. (2018). Prior distributions for objective Bayesian analysis. *Bayesian Analysis*, 13, 627–679.

Evans, M. & Jang, G. H. (2011a). Weak informativity and the information in one prior relative to another. *Statistical Science*, 23, 423–439.

Evans, M. & Jang, G. H. (2011b). A limit result for the prior predictive applied to checking for prior–data conflict. *Statistics and Probability Letters*, 81, 1034–1038.

Evans, M. & Moshonov, H. (2006). Checking for prior–data conflict. *Bayesian Analysis*, 1, 893–914.

Evans, M. & Moshonov, H. (2007). Checking for prior–data conflict with hierarchically specified priors. In Upadhyay, A. K., Singh, U., & Dey, D. (Eds.) *Bayesian Statistics and its Applications*, Anamaya Publishers, New Delhi, 145–159.

Gabry, J., Simpson, D., Vehtari, A., Betancourt, M., & Gelman, A. (2019). Visualization in Bayesian workflow. *Journal of the Royal Statistical Society Series A*, 182, 389–402.

Garthwaite, P. H., Kadane, J. B., & O'Hagan, A. (2005). Statistical methods for eliciting probability distributions. *Journal of the American Statistical Association*, 100, 680–701.

Gelman, A. (2017). Prior choice recommendations wiki!. *Statistical Modeling, Causal Inference, and Social Science Blog*. http://andrewgelman.com.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2013). *Bayesian Data Analysis*, 3rd ed., Chapman & Hall/CRC, Boca Raton.

Gelman, A. & Hennig, C. (2017). Beyond subjective and objective in statistics. *Journal of the Royal Statistical Society Series A*, 180, 967–1033.

Gelman, A., Jakulin, A., Pittau, M. G., & Su, Y. S. (2008). A weakly informative default prior distribution for logistic and other regression models. *Annals of Applied Statistics*, 2, 1360–1383.

Gelman, A., Meng, X. L., & Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, 6, 733–807.

Gelman, A. & Shalizi, C. R. (2013). Philosophy and the practice of Bayesian statistics. *British Journal of Mathematical and Statistical Psychology*, 66, 8–38.

Gelman, A., Simpson, D., & Betancourt, M. (2017). The prior can generally only be understood in the context of the likelihood. *Entropy*, 19, 555.

Goodrich, B., Gabry, J., Ali, I., & Brilleman, S. (2018). `rstanarm`: Bayesian applied regression modeling via Stan. R package version 2.17.4. http://mc-stan.org/.

Kass, R. & Wasserman, L. (1996). The selection of prior distributions by formal rules. *Journal of the American Statistical Association*, 91, 1343–1370.

Mutsvari, T., Tytgat, D., & Walley, R. (2016). Addressing potential prior–data conflict when using informative priors in proof-of-concept studies. *Pharmaceutical Statistics*, 15, 28–36.

Nott, D. J., Wang, X., Evans, M., & Englert, B. G. (2016). Checking for prior–data conflict using prior to posterior divergences. arXiv preprint arXiv:1611.00113.

Racine, A., Grieve, A. P., Fluhler, H., & Smith, A. F. M. (1986). Bayesian methods in practice: experiences in the pharmaceutical industry. *Journal of the Royal Statistical Society Series C*, 35, 93–150.

Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Annals of Statistics*, 12, 1151–1172.

Sauter, R. & Held, L. (2016). Quasi-complete separation in random effects of binary response mixed models. *Journal of Statistical Computation and Simulation*, 86, 2781–2796.

Schmidli, H., Gsteiger, S., Roychoudhury, S., O'Hagan, A., Spiegelhalter, D., & Neuenschwander, B. (2014). Robust meta-analytic-predictive priors in clinical trials with historical control information. *Biometrics*, 70, 1023–1032.

Stamey, T. A., Kabalin, J. N., McNeal, J. E., Johnstone, I. M., Freiha, F., Redwine, E. A., & Yang, N. (1989). Prostate specific antigen in the diagnosis and treatment of adenocarcinoma of the prostate. II. Radical prostatectomy treated patients. *Journal of Urology*, 141, 1076–1083.

Stan Development Team. (2018). *RStan: The R interface to Stan. R package version 2.18*. http://mc-stan.org.

Stan Development Team. (2018). *The Stan Math Library. R package version 2.18*. http://mc-stan.org.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society Series B*, 58, 267–288.

Wasserman, L. (2000). Asymptotic inference for mixture models by using data-dependent priors. *Journal of the Royal Statistical Society Series B*, 62, 159–180.

Zorn, C. (2005). A solution to separation in binary response models. *Political Analysis*, 13, 157–170.

APPENDIX

Proof of Theorem 1

Proof. The numerator of $r_{p\pi}$ is equal to

$$\begin{aligned} M_T^p(P^p(t_0) \leq x_\gamma) - M_T^p(P^p(t_0) \leq\ & x_\gamma - \psi(P^q(t_0) - P^p(t_0))) \\ =\ & x_\gamma - x_\gamma + \psi(P^q(t_0) - P^p(t_0)) \\ =\ & \psi(P^q(t_0) - P^p(t_0)) \\ \equiv\ & \delta_{pq}, \end{aligned}$$

where the first equality holds since $P^p(t_0)$ is uniformly distributed when $t_0 \sim M_T^p$. Then, $r_{p\pi} > 0 \Leftrightarrow \delta_{pq} > 0$, which happens when $\psi > 0$ and $P^q(t_0) - P^p(t_0) > 0$. ■

Received 12 August 2019
Accepted 26 January 2021