Statistics and Probability Letters 183 (2022) 109335

Contents lists available at ScienceDirect

Statistics and Probability Letters

journal homepage: www.elsevier.com/locate/stapro

Effective sample size for a mixture prior

Leonardo Egidi

Department of Economics, Business, Mathematics and Statistics 'Bruno de Finetti', University of Trieste, Via Valerio 4/1, 34127, Trieste, Italy

ARTICLE INFO

Article history:
Received 19 July 2021
Received in revised form 5 November 2021
Accepted 14 December 2021
Available online 21 December 2021

MSC:
62F15

Keywords:
Statistical application
Clinical trial
Prior-data conflict

ABSTRACT

Mixture prior distributions are much used in statistical applications, such as clinical trials, especially to avoid prior-data conflicts. We explicitly prove that the effective sample size (ESS) of a mixture prior rarely exceeds the ESS of any individual mixture component density of the prior.

1. Introduction

In a Bayesian model the need of measuring and quantifying the amount of information contained in a prior distribution is of great theoretical and practical appeal. However, the task of assessing the impact of a prior distribution on the final inferential conclusions presents some technical difficulties, including the impossibility of building a unique philosophical and mathematical framework designed to satisfactorily achieve this aim. Among others, Morita et al. (2008) defined the effective sample size (ESS) for a parametric prior distribution as that integer value minimizing the distance between a candidate prior and a posterior distribution based on a noninformative prior—referred hereafter to as a noninformative posterior—defined in terms of the curvatures of their log-densities. The resulting ESS may be considered as the sample size due to the prior component, that is then added to the size of the experiment, conveyed by the likelihood: the larger this value, the higher the chance the chosen prior will dominate the inference.

An extra amount of prior size, possibly derived from historical information, could have dramatic consequences such as yielding the so-called prior-data conflict (Evans and Moshonov, 2006; Evans and Jang, 2011; Egidi et al., 2021). This is particularly true when dealing with clinical trials, for which the use of robust mixture priors has proven to alleviate prior-data conflicts (Schmidli et al., 2014; Egidi et al., 2021). Moreover, it is customary—e.g., in Bayesian variable selection (O'Hara and Sillanpää, 2009)— to choose a mixture prior in which one of the components is vague or noninformative and the other one is rather informative, according to the "spike and slab" prior philosophy (George and McCulloch, 1993). Given a couple of priors q, p, we can combine them in a new mixture prior for the parameter θ:

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

with $\psi \in [0, 1]$. Egidi et al. (2021) proposed for p to use a standard Gaussian in many applied problems, or a wildly informative data-dependent prior in small-sample scenarios, whereas Schmidli et al. (2014) develop a meta-analytic-predictive (MAP) prior derived from historical data by preliminarily fixing the weight ψ according to the experimenter's

E-mail address: legidi@units.it.

https://doi.org/10.1016/j.spl.2021.109335
0167-7152/

judgement. q is usually chosen as belonging to the same parametric family of p but with an inflated variance. Regarding the choice of q, almost any choice is possible: however, Egidi et al. (2021) propose to use weakly-informative priors in the spirit of Gelman et al. (2008).

Moreover, to capture and express a wide range of prior beliefs, Diaconis and Ylvisaker (1985) and others proposed to combine more conjugate prior distributions:

$$\pi(\theta) = \sum_{i=1}^{k} \psi_i p_i(\theta), \tag{2}$$

along with a hyperprior on ψ, so that the resulting mixture prior could incorporate distinct experts' opinions.

Although mixture priors turn out to be helpful in many statistical settings and may act as a valid approximation for any parametric prior (Dalal and Hall, 1983; Diaconis and Ylvisaker, 1985), it is not immediate to establish their amount of information and then compute their effective sample size. In this paper we adopt the ESS measure proposed by Morita et al. (2008) and we try to fill this gap by providing some guidelines along with some theoretical results to ease this computation and understand the final result. Some implementation details and simulation procedures follow the logic outlined by Egidi (2018).

2. Review of prior effective sample size (ESS)

Given the parameter-vector $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^d$, we start eliciting two prior distributions $p(\boldsymbol{\theta}), q(\boldsymbol{\theta})$ by posing the following working assumptions on their moments:

$$\begin{aligned} &E_q(\boldsymbol{\theta}) = E_p(\boldsymbol{\theta}) \\ &\text{Corr}_p(\theta_i, \theta_j) = \text{Corr}_q(\theta_i, \theta_j), \ i \neq j \\ &\text{Var}_q(\theta_j) \gg \text{Var}_p(\theta_j), \ j = 1, \dots, d, \end{aligned} \tag{3}$$

such that p is referred to be a rather *informative* prior, whereas q is intended to be a *noninformative* prior distribution. The equality of the prior means is customary in many applications and frameworks, such as clinical trials—where we could suspect that a covariate has no effect in terms of regression purposes but we could be more or less confident about this finding— Bayesian Variable Selection (George and McCulloch, 1993) and comparison of priors (Evans and Jang, 2011). Here, the degree of ignorance intrinsic in our noninformative prior is completely translated in terms of a higher variability rather than in a different location.

According to Morita et al. (2008), the prior ESS of $p(\boldsymbol{\theta})$ with respect to the likelihood $f(\boldsymbol{y}|\boldsymbol{\theta})$ is defined as that integer which minimizes the distance between $p(\boldsymbol{\theta})$ and the noninformative posterior $q_n(\boldsymbol{\theta}|\boldsymbol{y})$. To define this distance, the negative second partial derivatives of the log-densities (the observed informations) are used for $j = 1, \dots, d$:

$$D_{p,j}(\boldsymbol{\theta}) = -\frac{\partial^2 \log(p(\boldsymbol{\theta}))}{\partial \theta_j^2}; \ D_{q,j}(n, \boldsymbol{\theta}, \boldsymbol{y}) = -\frac{\partial^2 \log(q(\boldsymbol{\theta}|\boldsymbol{y}))}{\partial \theta_j^2}. \tag{4}$$

In what follows, we will sometimes use the simplified notations p, q_n in place of $p(\boldsymbol{\theta}), q_n(\boldsymbol{\theta}|\boldsymbol{y})$ and $D_{p,j}, D_{q_n,j}$ in place of $D_{p,j}(\boldsymbol{\theta}), D_{q,j}(n, \boldsymbol{\theta}, \boldsymbol{y})$, respectively, where n is the data sample size. Let $D_{p,+} = \sum_{j=1}^{d} D_{p,j}$ and $D_{q_n,+} = \sum_{j=1}^{d} \int D_{q_n,j} f(\boldsymbol{y}) d\boldsymbol{y}$ denote the global information for the prior p and the posterior q_n, respectively. When $d = 1$, we will simply write D_p, D_{q_n}, suppressing the subscript '+'.

The distance between the prior p and the posterior q_n is then defined as the difference between the traces of the two information matrices:

$$\delta(n, \bar{\boldsymbol{\theta}}, p, q_n) = |D_{p,+}(\bar{\boldsymbol{\theta}}) - D_{q_n,+}(\bar{\boldsymbol{\theta}})|, \tag{5}$$

evaluated in $\bar{\boldsymbol{\theta}} = E_p(\boldsymbol{\theta})$, the prior informative mean—alternatively, one could also evaluate the curvature at the informative prior mode. Some discrepancy measures alternative to (5) could be adopted here, such as a member of the Rényi's class of divergence measures (Rényi, 1961), the Kullback–Leibler divergence. Even though the latter is an asymmetric divergence measure, as suggested by Nott et al. (2020) we could compute $\text{D}_{KL}(q_n(\boldsymbol{\theta}\|\boldsymbol{y})\|p(\boldsymbol{\theta})) = \int q_n(\boldsymbol{\theta}|\boldsymbol{y}) \log\left(\frac{q_n(\boldsymbol{\theta}|\boldsymbol{y})}{p(\boldsymbol{\theta})}\right) d\boldsymbol{\theta}$, that represents the amount of useful information, or information gain, about $\boldsymbol{\theta}$, that has been learned by discovering $\boldsymbol{y}$. In other words, the Kullback–Leibler divergence above measures the amount of information lost when the prior p is used in place of the posterior q_n.

The ESS for p is defined by Morita et al. (2008) as the integer minimizing the distance in (5):

$$ESS_p = \underset{n \in \mathbb{N}}{argmin}\{\delta(n, \bar{\boldsymbol{\theta}}, p, q_n)\}, \tag{6}$$

where the negative second log-prior derivative does not depend on n, whereas the posterior distribution always depends on the sample size n (see Table A.1 in the Support Information material for a quick overview about well-known Bayesian models). Thus, when D_{q_n} is linear in n, the distance defined in Eq. (5) reduces to a form such as $|a - n|$, where a is a constant: this is a continuous but not differentiable function in the minimum value $n = a$. When D_{q_n} is a polynomial in

the variable n of order greater than or equal to 2, then the distance $\delta(\cdot)$ is of the form $|a - n^b|$, $b \geq 2$, and represents a continuous and differentiable function.

The ESS defined in (6) is a very useful index of a prior's informativeness, and can be computed for parameters' subvectors; moreover ESS values may be also used to monitor the prior's reliability in the stage of elicitation process. When the ESS for a given prior is particularly high and close to the whole experiment's sample size, the experimenter could be tempted to think that inferential conclusions are dominated by the prior rather than the data. As previously remarked, alternative distances or divergence measures could be easily—through use of suited R libraries, e.g.—applied in (5) and (6) to find the correspondent value of ESS: different methods can yield slightly different results, perhaps sensitivity checks should always be implemented.

To gain generality, let us consider now the mixture prior $\pi(\boldsymbol{\theta}) = \sum_{i=1}^{k} \psi_i p_i(\boldsymbol{\theta})$ defined by considering k prior distributions, each contributing the mixture through a weight ψ_i, to possibly reflect multiple prior opinion beliefs about the parameter, as suggested by Diaconis and Ylvisaker (1985). The effective sample size ESS_π may be computed for the mixture prior analogously as in (6), and upon some mild conditions the following theorem holds.

Theorem 1. *Suppose we elicit a sequence of prior distributions $p_1(\boldsymbol{\theta}), p_2(\boldsymbol{\theta}), \ldots, p_k(\boldsymbol{\theta})$ such that the following assumptions hold:*

(i) $E_{p_1}(\boldsymbol{\theta}) = E_{p_2}(\boldsymbol{\theta}) = \cdots = E_{p_k}(\boldsymbol{\theta})$;
(ii) $Var_{p_1}(\theta_j) \gg Var_{p_i}(\theta_j)$ *for* $j = 1, 2, \ldots, d$ *and* $i \neq 1$, *so that the prior* p_1 *is the most informative;*
(iii) $\bar{\boldsymbol{\theta}} = mode(p_i(\boldsymbol{\theta}))\ \forall i = 1, 2, \ldots, k$;
(iv) ψ_i *are deterministic (fixed) coefficients such that* $\sum_{i=1}^{k} \psi_i = 1$.

Given the mixture prior $\pi(\boldsymbol{\theta}) = \sum_{i=1}^{k} \psi_i p_i(\boldsymbol{\theta})$ *and denoting* $H_{i,j}(\bar{\boldsymbol{\theta}}) = \frac{\partial^2 p_i(\bar{\boldsymbol{\theta}})}{\partial \theta_j^2}|_{\boldsymbol{\theta}=\bar{\boldsymbol{\theta}}}$ *and* $H_{i,+}(\bar{\boldsymbol{\theta}}) = \sum_{j=1}^{d} H_{i,j}(\bar{\boldsymbol{\theta}})$, *we then obtain the following relationship:*

$$ESS_\pi \leq ESS_{p_1} \Leftrightarrow H_{1,+}(\bar{\boldsymbol{\theta}}) \leq \sum_{j=1}^{d} p_1(\bar{\boldsymbol{\theta}}) \frac{\sum_{i=2}^{k} \psi_i H_{i,j}(\bar{\boldsymbol{\theta}})}{\sum_{i=2}^{k} \psi_i p_i(\bar{\boldsymbol{\theta}})}. \tag{7}$$

For a formal proof, see the Support Information material. Although an analytical solution of the ESS for the mixture priors is not available in closed-form, Formula (7) provides an upper bound and yields an intuitive result. The interpretation is that whatever the weights $\boldsymbol{\psi}$ and the priors $p_1, p_2, \ldots, p_k$ used in the mixture, its information is lower than or equal to the information contained in p_1 if and only if the degree of informativity given by the second derivative of p_1 does not exceed a threshold depending on the informativity expressed by all the other $k-1$ priors. The second derivative of a prior distribution evaluated in the maximum value yields an intuitive amount of informativity: as depicted in Figure D.1 from the Support Information material for some Gaussian priors, the higher the information provided by the prior, and the lower is the second derivative value.

In the special case of a univariate two-component mixture with $k = 2$ where ψ (1-ψ) is the weight assigned to the noninformative (informative) prior p_2 (p_1), Eq. (7) simplifies to:

$$p_1''(\bar{\theta}) \leq \frac{p_1(\bar{\theta})}{p_2(\bar{\theta})} p_2''(\bar{\theta}), \tag{8}$$

which is always true, where the ratio $p_1(\bar{\theta})/p_2(\bar{\theta}) \geq 1$, and $p_2''(\bar{\theta}) \geq p_1''(\bar{\theta})$ due to the considerations above. Intuitively, the less informative is p_2, and the closer is the right term to zero, making $p_1''(\bar{\theta}) \leq 0$, which is true for hypothesis. Conversely, as p_2 is closer to p_1, the right term in (8) tends to $p_2''(\bar{\theta})$, and we get $p_1''(\bar{\theta}) \leq p_2''(\bar{\theta})$, that is also true. Thus, the ESS provided by a two-component mixture consisting of the couple of priors p_1 and p_2 does never exceed the ESS provided by the most informative prior p_1.

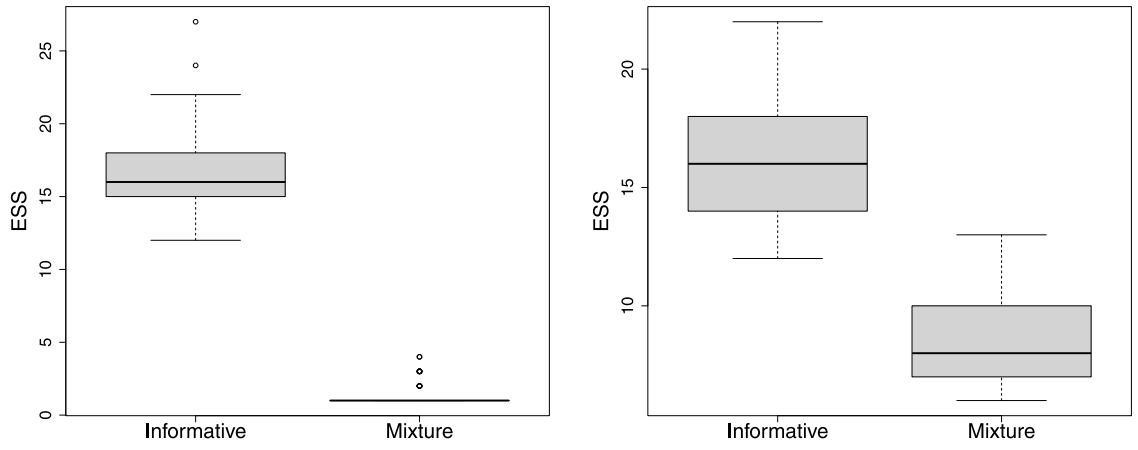
Suppose now to elicit a prior distribution for the mixture weights $\boldsymbol{\psi}$ of the mixture prior, rather than considering them as a fixed/deterministic quantity. Then we have the following result.

Theorem 2. *Assume the same conditions of Theorem 1 hold. If assumption (iv) is replaced by*

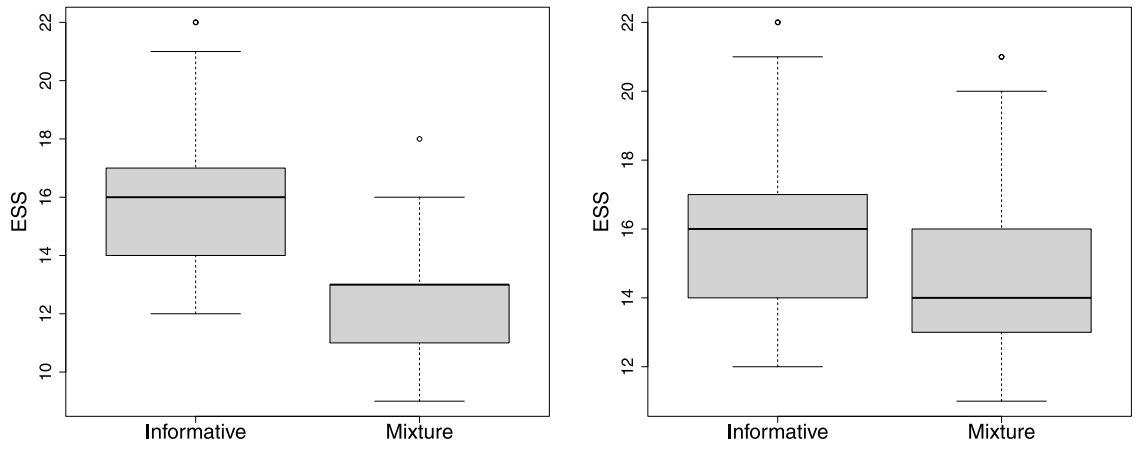
(ivb) $\boldsymbol{\psi} \sim f_{\boldsymbol{\psi}}$, *with* $f_{\boldsymbol{\psi}}$ *a suited prior for* $\boldsymbol{\psi}$, *then* $ESS_\pi \leq ESS_{p_1}$ *iff:*

$$H_{1,+}(\bar{\boldsymbol{\theta}}) \leq \sum_{j=1}^{d} p_1(\bar{\boldsymbol{\theta}}) \frac{\sum_{i=2}^{k} \psi_i H_{i,j}(\bar{\boldsymbol{\theta}})}{\sum_{i=2}^{k} \psi_i p_i(\bar{\theta}_j)} + k \frac{\sum_{i=1}^{k} H_{i,i}(\bar{\boldsymbol{\theta}}) p_i(\bar{\boldsymbol{\theta}})}{\sum_{i=1}^{k} \psi_i p_i(\bar{\boldsymbol{\theta}})}. \tag{9}$$

The result above extends the upper bound stated in Theorem 1 by acknowledging a supplementary part of information contained in the hyperprior distribution for the mixture weights. The more information is provided by the hyperprior for $\boldsymbol{\psi}$, and the lower is the upper bound for the information provided by p_1 in Eq. (9): intuitively, as the mixture weights information increases—thus, as the information of the mixture prior increases—we need an even more informative prior



(a) $\boldsymbol{\psi} = (0.1, 0.4, 0.5)$ (b) $\boldsymbol{\psi} = (1/3, 1/3, 1/3)$



(c) $\boldsymbol{\psi} = (0.5, 0.2, 0.3)$ (d) $\boldsymbol{\psi} = (0.7, 0.1, 0.2)$

Fig. 1. Scenario A: ESS for 100 simulated datasets under the informative $p_1(\boldsymbol{\theta}) \sim \text{Dirichlet}(\boldsymbol{\theta}|\boldsymbol{\alpha})$ and the mixture prior $\sum_{i=1}^{3} \psi_i p_i(\boldsymbol{\theta})$ under different choices for the mixture weights $\boldsymbol{\psi}$, where $p_2(\boldsymbol{\theta}) \sim \text{Dirichlet}(\boldsymbol{\theta}|\boldsymbol{\alpha}/c)$, $p_3(\boldsymbol{\theta}) \sim \text{Dirichlet}(\boldsymbol{\theta}|\boldsymbol{\alpha}/3c)$, $\boldsymbol{\alpha} = (5, 5, 10)$ and $c = 10$.

p_1 to ensure that $ESS_\pi \leq ESS_{p_1}$. In case of a uniform hyperprior $\boldsymbol{\psi} \sim \mathcal{U}(0, 1)$—i.e., scarce information for $\boldsymbol{\psi}$—then $H_{i,i}(\bar{\boldsymbol{\theta}}) = 0 \; \forall i = 1, 2, \ldots, k$, thus the second addendum in (9) is zero, and Formula (7) arises as a special case. By resuming, when no (or very scarce information) is plugged into this hyperprior the $\boldsymbol{\psi}$'s do not influence the final mixture ESS and they behave as they were deterministic factors.

3. Simulation study

We consider a multinomial-Dirichlet model to perform a simulation study and assess the effective sample size provided by a rather informative prior and a mixture prior. For each $n = 1, 2, \ldots, 50$ we simulate M datasets replications $\boldsymbol{y}^{(m)} \sim \text{Multin}(n, \boldsymbol{\theta})$, where $\boldsymbol{\theta} \in [0, 1]^3$ and $m = 1, 2, \ldots, M$. We fixed $\boldsymbol{\theta}_0 = (1/3, 1/3, 1/3)$ as the true value parameter and we assume $p_1(\boldsymbol{\theta}) \sim \text{Dirichlet}(\boldsymbol{\theta}|\boldsymbol{\alpha})$, $p_2(\boldsymbol{\theta}) \sim \text{Dirichlet}(\boldsymbol{\theta}|\boldsymbol{\alpha}/c)$, $p_3(\boldsymbol{\theta}) \sim \text{Dirichlet}(\boldsymbol{\theta}|\boldsymbol{\alpha}/3c)$, with $\boldsymbol{\alpha} = (5, 5, 10)$. The mixture prior is $\pi(\boldsymbol{\theta}) = \sum_i \psi_i p_i(\boldsymbol{\theta})$, then to assess how the ESS of the two priors—p_1 and π—varies:

scenario A we fix $c = 10$ and let vary the mixture weights $\boldsymbol{\psi}$;

scenario B we fix the mixture weights $\boldsymbol{\psi} = (1/3, 1/3, 1/3)$ and let vary c.

Results for the ESS from Scenario A are depicted in Fig. 1, whereas the correspondent pattern for the distance $\delta(n, \bar{\boldsymbol{\theta}}, p, q_n)$ defined in Eq. (5) is provided in Figure E.1 in the Support Information material. As it may appear evident,

the larger (smaller) is the weight ψ_1 assigned to the informative prior p_1 and the closer (further) is the ESS_{p_1} to ESS_π. When each of the distributions in the mixture is assigned the same weight (panel (b)), the resulting ESS under the mixture is sensitively lower than the ESS implied by the informative prior. The results plotted in Fig. 1 are rather intuitive and confirm that the information raised by the mixture never exceeds the information contained in the informative prior if and only if theorem's condition in (7) applies, and this happens for all of the four cases in Scenario A.

The results for Scenario B in Figures E.2, E.3 in the Support Information material highlight an important but paradoxical feature. As c increases—then, as the noninformative Dirichlet becomes even more noninformative—the ESS for π tends to approximate the ESS under the informative prior p_1 (panel (c) and (d)). This is less intuitive and deserves a quick technical consideration: the curvature of $\log(\pi(\boldsymbol{\theta}))$ will approximate the curvature of $\log(p_1(\boldsymbol{\theta}))$ as c will be increased, and the information carried by the two priors will tend to coincide. This counterintuitive fact may be read twofold: we could need another notion of distance, possibly less sensitive to the values of the hyperparameter c, or we could use other noninformative priors whose functional form does not depend on c, such as Jeffreys priors (see Section H for a practical application on a phase I trial).

The relevant R code for the simulation study and the computation of the mixture ESS is provided in the Support Information material.

Data sharing

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.spl.2021.109335. The following supporting information containing R code and theorems' proofs is available as part of the online article: https://github.com/LeoEgidi/Support-material-ESS-paper

References

Dalal, S., Hall, W., 1983. Approximating priors by mixtures of natural conjugate priors. J. R. Stat. Soc. Ser. B Stat. Methodol. 45 (2), 278–286.
Diaconis, P., Ylvisaker, D., 1985. Quantifying prior opinion. Bayesian statistics 2. In: Proceedings of the Second Valencia International Meeting", September, vol. 6. pp. 133–156.
Egidi, L., 2018. Developments in Bayesian Hierarchical Models and Prior Specification with Application to Analysis of Soccer Data.
Egidi, L., Pauli, F., Torelli, N., 2021. Avoiding prior-data conflict in regression models via mixture priors. Canad. J. Statist. http://dx.doi.org/10.1002/cjs.11637 (in press).
Evans, M., Jang, G.H., 2011. Weak informativity and the information in one prior relative to another. Statist. Sci. 26 (3), 423–439.
Evans, M., Moshonov, H., 2006. Checking for prior-data conflict. Bayesian Anal. 1 (4), 893–914.
Gelman, A., Jakulin, A., Pittau, M.G., Su, Y.-S., et al., 2008. A weakly informative default prior distribution for logistic and other regression models. Ann. Appl. Stat. 2 (4), 1360–1383.
George, E.I., McCulloch, R.E., 1993. Variable selection via Gibbs sampling. J. Amer. Statist. Assoc. 88 (423), 881–889.
Morita, S., Thall, P.F., Müller, P., 2008. Determining the effective sample size of a parametric prior. Biometrics 64 (2), 595–602.
Nott, D.J., Wang, X., Evans, M., Englert, B.-G., 2020. Checking for prior-data conflict using prior-to-posterior divergences. Statist. Sci. 35 (2), 234–253.
O'Hara, R.B., Sillanpää, M.J., 2009. A review of Bayesian variable selection methods: what, how and which. Bayesian Anal. 4 (1), 85–117.
Rényi, A., 1961. On measures of entropy and information. In: Proceedings of the 4th Berkeley Symposium on Mathematical Statistics and Probability. University of California Press, pp. 547–561.
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 (4), 1023–1032.