RESEARCH ARTICLE

On the normalized power prior

Luiz Max Carvalho[1] | **Joseph G. Ibrahim**[2]

[1]School of Applied Mathematics, Getúlio Vargas Foundation (FGV), Rio de Janeiro, Brazil

[2]Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina

Correspondence
Luiz Max Carvalho, School of Applied Mathematics, Getúlio Vargas Foundation (FGV), Rio de Janeiro, Brazil.
Email: lmax.fgv@gmail.com

The power prior is a popular tool for constructing informative prior distributions based on historical data. The method consists of raising the likelihood to a discounting factor in order to control the amount of information borrowed from the historical data. However, one often wishes to assign this discounting factor a prior distribution and estimate it jointly with the parameters, which in turn necessitates the computation of a normalizing constant. In this article, we are concerned with how to approximately sample from joint posterior of the parameters and the discounting factor. We first show a few important properties of the normalizing constant and then use these results to motivate a bisection-type algorithm for computing it on a fixed budget of evaluations. We give a large array of illustrations and discuss cases where the normalizing constant is known in closed-form and where it is not. We show that the proposed method produces approximate posteriors that are very close to the exact distributions and also produces posteriors that cover the data-generating parameters with higher probability in the intractable case. Our results suggest that the proposed method is an accurate and easy to implement technique to include this normalization, being applicable to a large class of models. They also reinforce the notion that proper inclusion of the normalizing constant is crucial to the drawing of correct inferences and appropriate quantification of uncertainty.

KEYWORDS

doubly intractable, elicitation, historical data, normalization, power prior, sensitivity analysis

1 | BACKGROUND

Power priors[1,2] are a popular method of constructing informative priors and are widely used in fields such as medical research, where elicitation of informative priors is crucial. When historical data are available, power priors allow the elicitation of informative priors by borrowing information from the historical data. This is accomplished by raising the likelihood of the historical data to a scalar discounting factor a_0, usually taken to be $0 \leq a_0 \leq 1$. When $a_0 = 0$, the historical data receive no weight and thus no information is borrowed, whereas $a_0 = 1$ represents full borrowing of the information contained in the historical data to inform the prior. In many settings, this construction of an informative prior can be shown to be optimal in an information-processing sense.[3,4]

To make the presentation more precise, let the observed data be $D_0 = \{d_{01}, d_{02}, \ldots, d_{0N_0}\}$, $d_{0i} \in \mathcal{X} \subseteq \mathbb{R}^d$ and let $L(D_0|\theta)$ be a likelihood function assumed to be finite for all arguments $\theta \in \Theta \subseteq \mathbb{R}^q$. The simplest formulation of the power prior reads

$$\pi(\theta|D_0, a_0) \propto L(D_0|\theta)^{a_0} \pi(\theta), \tag{1}$$

Abbreviation: BCI, Bayesian (equal-tailed) credibility interval.

where π is called the *initial* prior and a_0 is a scalar, usually taken to be in $[0, 1]$. The scalar a_0 controls the amount of information from historical data that is included in the analysis of the current data.[1] A commonly adopted practice is to fix a_0 and assess the sensitivity of results to different values, including $a_0 = 0$ (no borrowing) and $a_0 = 1$ (full borrowing, see Ibrahim et al,[2] Section 5). One might also be interested in accommodating uncertainty about the relative weighting of the historical data by placing a prior π_A on a_0, indexed by a (set of) parameter(s) δ. This leads to what we will call the *unnormalized* power prior

$$\tilde{\pi}(\theta, a_0|D_0, \delta) = \pi(\theta|D_0, a_0)\pi_A(a_0|\delta) \propto L(D_0|\theta)^{a_0}\pi(\theta)\pi_A(a_0|\delta). \tag{2}$$

As observed by Neuenschwander et al,[5] this formulation does not lead to a correct joint posterior distribution for (θ, a_0) because the normalizing constant of (1),

$$c(a_0) := \int_\Theta L(D_0|\theta)^{a_0}\pi(\theta) \, d\theta, \tag{3}$$

is missing. The *normalized* power prior is defined as[6,7]:

$$\pi(\theta, a_0|D_0, \delta) = \frac{L(D_0|\theta)^{a_0}\pi(\theta)\pi_A(a_0|\delta)}{c(a_0)}. \tag{4}$$

In light of new (current) data $D = \{d_1, d_2, \ldots, d_N\}$, we thus have the joint posterior

$$p(\theta, a_0|D_0, D, \delta) \propto \frac{1}{c(a_0)}L(D|\theta)L(D_0|\theta)^{a_0}\pi(\theta)\pi_A(a_0|\delta). \tag{5}$$

In this setting, a_0 becomes a parameter we are interested in learning about in light of the data and thus we arrive at the marginal posterior

$$\begin{aligned} p(a_0|D_0, D, \delta) &= \int_\Theta p(\theta, a_0|D_0, D, \delta) \, d\theta, \\ &\propto \frac{\pi_A(a_0|\delta)}{c(a_0)} \int_\Theta L(D_0|\theta)^{a_0}\pi(\theta)L(D|\theta) \, d\theta, \end{aligned} \tag{6}$$

which involves the computation of not one, but two potentially high-dimensional integrals. Unfortunately, the posterior distribution in (5) is in the class of so-called doubly intractable distributions (see Section 5.2) and its exact computation depends on somewhat advanced Markov chain Monte Carlo (MCMC) techniques.[8]

In this article, we study the theoretical properties of the normalizing constant and use our findings to guide informed designs for sensitivity analysis. Furthermore, we explore a simple way to recycle computations from a sensitivity analysis in order to sample from an approximate joint posterior of a_0 and θ. This article is organized as follows: in Section 2, we present a few results on the propriety of the power prior and the properties of the normalizing constant, $c(a_0)$. We give general results as well as specific formulas for the exponential family of probability distributions and its conjugate prior. Section 3 discusses the computational aspects of approximating $c(a_0)$ when it is not known in closed-form and Section 4 brings a large array of illustrations of the normalized power prior in examples where the normalizing constant is known in closed-form and situations where it is not. We conclude with a discussion of the findings and future directions in Section 5.

2 | THEORY

We begin by describing a few results on the properties of the power prior and its normalized version. First, we show that the normalized power prior is always well-defined when the initial prior is proper in Theorem 1, for which we give an elementary proof in Appendix A.

Theorem 1. *Assume* $\int_{\mathcal{X}} L(x|\theta) \, dx < \infty$. *In addition, assume* π *is proper, that is,* $\int_\Theta \pi(\theta) \, d\theta = 1$. *Then,* $c(a_0) = \int_\Theta L(D_0|\theta)^{a_0}\pi(\theta) \, d\theta < \infty$ *for* $a_0 \geq 0$.

Proof. See Appendix A. ∎

Theorem 1 thus shows that the expression in (4) leads to a valid joint prior on (θ, a_0). While scientific interest usually lies with $a_0 \in [0, 1]$, showing the result holds also for $a_0 > 1$ might find use in other fields, such as the analysis of complex surveys, where the likelihood is raised to a power that is inversely proportional to a selection probability.[9] In many applications, one usually has a collection of historical data sets, with different sample sizes and particular (relative) reliabilities that one would like to include in a power prior analysis. Remark 1 extends Theorem 1 for the situation where multiple historical data sets are available and the analyst desires to include them simultaneously, each with a weight a_{0k}.

Remark 1. The power prior on multiple historical data sets is also a proper density.

Proof. See Appendix A. ∎

These two results give solid footing to the normalized power prior as well as tempered likelihood techniques, for which propriety is usually assumed without proof or proved only for specific cases.[6,9,10]

2.1 | Properties of the normalizing constant $c(a_0)$

In order to aid computation, it is useful to study some of the properties of the normalizing constant, $c(a_0)$, seen as a function of the scalar a_0. First, it is convenient to state the following proposition about the derivatives of the normalizing constant.

Proposition 1. *All of the derivatives of $c(a_0)$ exist, that is, $c \in C^\infty$.*

Proof. See Appendix A. ∎

We can now show that $c(a_0)$ is strictly convex (Lemma 1), which motivates specific algorithms for its approximation.

Lemma 1. *Assume $L(D_0|\theta)$ is continuous with respect to θ. Then the normalizing constant is a strictly convex function of a_0.*

Proof. See Appendix A. ∎

For the goals of this article, it would be useful to know more about the shape of $c(a_0)$, more specifically if and when its derivatives change signs. For computational stability reasons, one is usually interested in $l(a_0) := \log(c(a_0))$ instead of $c(a_0)$ and hence it is also useful to study the derivative of the log-normalizing constant, $l'(a_0) = c'(a_0)/c(a_0)$. A key observation is that $l'(a_0)$ changes signs at the same point as $c'(a_0)$ does, a feature that can be exploited when designing algorithms for approximating $l(a_0)$ (see Section 3.1).

Next, we state Remark 2, that shows that for the large class of discrete likelihoods (Bernoulli, Poisson, etc.), $c(a_0)$ is monotonic.

Remark 2. When $L(D|\theta)$ is a discrete likelihood, $c(a_0)$ is monotonically decreasing in a_0.

Proof. The proof is immediate from Lemma 1 and the fact that for a nondegenerate discrete likelihood the function $\log(L(D|\theta))$ is strictly negative and hence so is its expectation under the power prior (see Proposition 1). ∎

This will find application in the adaptive grid building described in Section 3.1.1.

2.1.1 | Exponential family

The methodology presented in this article is general and is applicable to all types of Bayesian models. Nevertheless, it might be of interest to provide theoretical results where they can be derived in closed-form. A large class of models routinely employed in applications is the exponential family of distributions, which includes the Gaussian and Gamma families, as well as the class of generalized linear models.[11,12] Here we give expressions for the normalizing constant when

the likelihood is in the exponential family. Furthermore, we also derive the marginal posterior of a_0 when the initial prior $\pi(\theta)$ is in the conjugate class.[13]

Suppose $L(D_0|\theta)$ is in the exponential family:

$$L(D_0|\theta) = \boldsymbol{h}(D_0) \exp \left[\eta(\theta)^T \boldsymbol{S}(D_0) - N_0 A(\theta) \right],$$

where $\boldsymbol{h}(D_0) := \prod_{i=1}^{N_0} h(d_{0i})$ and $\boldsymbol{S}(D_0) := \sum_{i=1}^{N_0} T(d_{0i})$. Thus we have

$$
\begin{aligned}
c(a_0) &= \int_\Theta \left\{ \boldsymbol{h}(D_0) \exp \left[\eta(\theta)^T \boldsymbol{S}(D_0) - N_0 A(\theta) \right] \right\}^{a_0} \pi(\theta) \, d\theta, \\
&= \boldsymbol{h}(D_0)^{a_0} \int_\Theta \exp \left(\eta(\theta)^T a_0 \boldsymbol{S}(D_0) \right) \exp \left(-a_0 N_0 A(\theta) \right) \pi(\theta) \, d\theta.
\end{aligned}
\tag{7}
$$

The derivative (see Proposition 1) evaluates to

$$c'(a_0) = \log(\boldsymbol{h}(D_0)) + \int_\Theta \left[\eta(\theta)^T a_0 \boldsymbol{S}(D_0) \right] f_{a_0}(D_0;\theta) \, d\theta - a_0 N_0 \int_\Theta f_{a_0}(D_0;\theta) A(\theta) \, d\theta, \tag{8}$$

where $f_{a_0}(D_0;\theta) := L(D_0|\theta)^{a_0} \pi(\theta)$.

These results can be refined further if we restrict the class of initial priors. If we choose $\pi(\theta)$ to be conjugate to $L(D_0|\theta)$,[13] that is,

$$\pi(\theta|\tau, n_0) = H(\tau, n_0) \exp\{ \tau^T \eta(\theta) - n_0 A(\theta) \},$$

we have

$$
\begin{aligned}
c(a_0) &= \boldsymbol{h}(D_0)^{a_0} H(\tau, n_0) \int_\Theta \exp \left[\eta(\theta)^T (\tau + a_0 \boldsymbol{S}(D_0)) - (n_0 + a_0 N_0) A(\theta) \right] \, d\theta, \\
&= \frac{\boldsymbol{h}(D_0)^{a_0} H(\tau, n_0)}{H \left([\tau + a_0 \boldsymbol{S}(D_0)]^T, n_0 + a_0 N_0 \right)}.
\end{aligned}
\tag{9}
$$

Following (6), the marginal posterior for a_0 is

$$p(a_0|D_0, D, \delta) \propto \frac{H \left([\tau + a_0 \boldsymbol{S}(D_0)]^T, n_0 + a_0 N_0 \right)}{H \left([\tau + a_0 \boldsymbol{S}(D_0) + \boldsymbol{S}(D)]^T, n_0 + a_0 N_0 + N \right)} \frac{\boldsymbol{h}(D)}{\boldsymbol{h}(D_0)^{a_0}} \pi_A(a_0|\delta). \tag{10}$$

3 | COMPUTATION

In this section, we propose a way to approximate $c(a_0)$ at a grid of values, while simultaneously picking the grid values themselves.

3.1 | Efficiently approximating $c(a_0)$

While the exponential family is a broad class of models, for many models of practical interest $c(a_0)$ is not known in closed form, and hence must be computed approximately. As discussed above—and by Neuenschwander et al,[5] it is important to include $c(a_0)$ in the calculation of the posterior when a_0 is allowed to vary and assigned its own prior. An example where this would be important is when we need to normalize the power prior for use within an MCMC procedure. If one wants to avoid developing a customized MCMC sampler for this situation (see Section 5), one needs a simple yet accurate way of approximating $c(a_0)$ and its logarithm, $l(a_0)$.

Here we take the following approach to approximating $c(a_0)$: first, define a grid of values $\boldsymbol{a}^{\text{est}} = \{a_1^{\text{est}}, \dots, a_J^{\text{est}}\}$ for a typically modest grid size J. Using a marginal likelihood approximation method (see below), compute an estimate of $c(a_0)$

for each point in $\boldsymbol{a}^{\text{est}}$, obtaining a set of estimates. Consider an approximating function $g_{\boldsymbol{\xi}} : [0, \infty) \to (0, \infty)$, indexed by a set of parameters $\boldsymbol{\xi}$. For instance, $g_{\boldsymbol{\xi}}$ could be a linear model, a generalized additive model (GAM) or a Gaussian process. We can then use $\boldsymbol{a}^{\text{est}}$ and $\hat{c}(\boldsymbol{a}^{\text{est}})$ as data to learn about $\boldsymbol{\xi}$. Once we obtain an estimate $\hat{\boldsymbol{\xi}}$, $c(\cdot)$ can be approximated at any point z by the prediction $g_{\hat{\boldsymbol{\xi}}}(z)$.

In order to simplify implementation, in our applications we found it useful to create a grid of size $K \gg J$, $\boldsymbol{a}^{\text{pred}} = \{a_1^{\text{pred}}, \dots, a_K^{\text{pred}}\}$, and then compute the predictions $\boldsymbol{g}^{\text{pred}} := g_{\hat{\boldsymbol{\xi}}}(\boldsymbol{a}^{\text{pred}})$. We can then use this dictionary of values to obtain an approximate value of $c(a_0)$ by simple interpolation. This approach allows one to evaluate several approximating functions without having to implement each one separately.

A caveat of this grid approach is that the maximum end point needs to be chosen in advance, effectively bounding the space of a_0 considered. While for many applications, interest usually lies in $a_0 \in [0, 1]$, even when one is interested in $a_0 > 1$, one usually has a good idea of the range of reasonable of values, since this information is also useful in specifying the prior $\pi_A(a_0|\delta)$. In fact, prior information can be used to set the maximum grid value: let p be a fixed probability and then set the maximum grid value M such that

$$\int_0^M \pi(a) da = p.$$

One can pick $p = 0.9999$, for instance, so as to have high probability of not sampling any values of a_0 outside the grid. This path is not explored here, however.

3.1.1 | Adaptively building the estimation grid

Another approach is to build the estimation grid $\boldsymbol{a}^{\text{est}}$ adaptively. Since $c(a_0)$ is convex, we need to make sure our grid covers the region where its derivative changes signs (if it does) when designing both $\boldsymbol{a}^{\text{est}}$ and $\boldsymbol{a}^{\text{pred}}$. As discussed above, $l'(a_0)$ changes signs at the same point as $c'(a_0)$ does, and we shall exploit this to design our grids. One can get an estimate of $c'(a_0)$ directly from MCMC (see Equation A2 in Appendix A), since this is just the expected value of the log-likelihood under the power prior.[10] Letting $\theta = \{\theta^{(1)}, \theta^{(2)}, \dots, \theta^{(M)}\}$ be a collection of MCMC samples and $L(D|\theta)$ be the (fully normalized) likelihood function, a strongly consistent estimator for $c'(a_0)$ is:

$$\widehat{c'(a_0)} = \frac{1}{M} \sum_{i=1}^{M} \log L(D|\theta^{(i)}).$$

Noticing that $l'(a_0) = c'(a_0)/c(a_0)$ also gives a way to obtain $\widehat{l'(a_0)} = \widehat{c'(a_0)}/\widehat{c(a_0)}$, where the denominator is estimated using a marginal likelihood estimation method, for example, bridge sampling (see Section 3.2.2). In practice, this means that evaluating $l'(a_0)$ comes essentially "for free" once one does the computations necessary to estimate $l(a_0)$. The second derivative, $l''(a_0)$, can be estimated following a similar procedure.

In order to adaptively build the estimation grid, $\boldsymbol{a}^{\text{est}}$, we propose doing a bisection-type search. First, let m and M be the grid the endpoints and J be the budget on the total number of evaluations of $l(a_0)$. Further, fix two real constants $v_1, v_2 > 0$. In our computations, we have used $v_1 = v_2 = 10$.

1. Initialize the variables $Z = \{0\}$, which will store the visited values of a_0, $F = \{0\}$ which will store the values of $l(a_0)$ and $F' = \{\emptyset\}$ which will store the values of $l'(a_0)$;
2. Compute $l(m)$, $l'(m)$, $l(M)$, and $l'(M)$ and store these values in their respective variables. **If** $\text{sgn}(l'(m)) = \text{sgn}(l'(M))$, construct Z to be a regular grid of $J - 2$ values between m and M and compute/estimate $l(\cdot)$ at those values, building F and F' accordingly. **Else**, with $\text{sgn}(c'(m)) \neq \text{sgn}(c'(M))$, set $L^{(1)} = m$ and $U^{(1)} = M$ and make $J = J - 1$. Then, for the kth iteration $(k > 1)$:
   (a) Make $z^{(k)} = (L^{(k)} + U^{(k)})/2$, compute $l(z^{(k)})$, $l'(z^{(k)})$ and store $Z \leftarrow z^{(k)}$, $F \leftarrow l(z^{(k)})$ and $F' \leftarrow l'(z^{(k)})$.
   (b) Compare derivative signs: **if** $\text{sgn}(l'(z^{(k)})) = \text{sgn}(l'(m))$ set $L^{(k+1)} = z^{(k)}$ and $U^{(k+1)} = U^{(k)}$. Otherwise, set $L^{(k+1)} = L^{(k)}$ and $U^{(k+1)} = z^{(k)}$. Compute $\delta^{(k)} = |z^{(k)} - z^{(k-1)}|$ and set $J = J - 1$. **If** $J = 0$, stop.
   (c) **If** $J > 0$ but $\delta^{(k)} < v_1 m$, stop.

     i. Compute $A^{(k)} = \max(0, z^{(k)} - v_2 m)$ and $B^{(k)} = \min(z^{(k)} + v_2 m, M)$.

ii. Considering only the elements z_i of Z such that $A^{(k)} \leq z_i \leq B^{(k)}$, find the pair (z_i, z_{i+1}) such that $|z_i - z_{i+1}|$ is largest, make $z^{(k)} = |z_i - z_{i+1}|/2$, compute $l(z^{(k)})$, $l'(z^{(k)})$ and store $Z \leftarrow z^{(k)}$, $F \leftarrow l(z^{(k)})$ and $F' \leftarrow l'(z^{(k)})$. Set $J = J - 1$ and, if $J = 0$, stop.

Informally, the algorithm starts by approaching the point $\hat{a}$ at which $l'(\hat{a}) = 0$ and storing the values encountered on that path. Because we do not want to waste all of the computational budget in exploring too small a neighborhood around $\hat{a}$, we use v_1 to control the size of this neighborhood. Then, if the computational budget of J evaluations has not yet been exhausted, we "plug the gaps" in our collection of values of a_0, Z. Because these gaps matter more closer to the region around $\hat{a}$, we use v_2 to control the size of the neighborhood where we fill in the gaps.

The algorithm discussed in this section shares many similarities with power (tempered) posterior methods for computing marginal likelihoods.[10,14] A key difference is that while the aforementioned methods are concerned with estimating the normalizing constant for a single value of a_0 ($a_0 = 1$), here we are interested in approximating the whole $c(a_0)$ curve.

3.1.2 | When estimating marginal likelihoods directly is impractical

For very complex, parameter-rich models, it might be the case that estimating $l(a_0)$—or $c(a_0)$—at a few values of a_0 may still be very computationally costly. It may also be the case that the posterior density $L(D_0|\theta)\pi(\theta)$ is costly to compute. An alternative approach is to only evaluate $l'(a_0)$ instead, which should be cheaper, and then obtain an approximation of $l(a_0)$ *via* quadrature.

Van Rosmalen et al[15] propose a modification of the method of Friel and Pettitt[10] using the identity

$$l(a_0) = \int_0^{a_0} E_{f_a} \left[\log L(D_0|\theta) \right] \, da.$$

The method of Van Rosmalen et al exploits this identity by incrementing a_0 by a factor Δ_a (0.01, say) and running a batch of MCMC simulations to obtain a reliable estimate of the log-likelihood expectation until $a_0 \geq 1$. The desired integral can then be estimated from the "curve" of estimates of $l'(a_0)$ by the cumulative sum. This approach amounts to approximating the curve of $l'(a_0)$ and then integrating this approximate curve using midpoint quadrature.

The algorithm presented in Section 3.1.1 can be used in the context of expensive marginal likelihoods with very little change, namely, by simply not evaluating (estimating) $l(\cdot)$ at points Z. One can then approximate $l'(a_0)$—instead of $l(a_0)$—in the same way by estimating an approximate curve g_ξ, evaluating predictions on a fine grid (say, $K = 20\,000$) and then using midpoint integration to obtain an approximation of $l(a_0)$. We analyzed this approach on a limited set of examples and found that it yielded less accurate approximations when compared to the method approximating $c(a_0)$ directly (see Appendix D) and thus did not pursue the matter further. We note that these results are not unexpected, since the derivative-only approach of Van Rosmalen et al uses less information than the method proposed here. On the other hand, their approach should be much cheaper computationally and remains a competitive alternative when marginal likelihoods cannot be computed directly.

3.2 | Computational details

Easily extendable computer code and tutorials for implementing new models is available from https://github.com/maxbiostat/npowerPrioR and code to reproduce the analysis in this article https://github.com/maxbiostat/propriety_power_priors.

3.2.1 | Markov Chain Monte Carlo

The vast majority of the models discussed in this article lead to posterior distributions which cannot be written in closed-form and hence we must resort to numerical methods. Here we employ Hamiltonian—or hybrid—Monte Carlo (HMC), implemented in the Stan programming language[16] in order to estimate the expectations of interest. An excellent review of HMC can be found in Neal.[17] Unless stated otherwise, all of the analyses reported here are the result of

running four independent chains of 2000 iterations each, with the first 1000 removed as burn-in/warmup. Convergence was checked for by making sure all runs achieved a potential scale reduction factor (PSRF, $\hat{R}$) smaller than 1.01. For all expectations, we ensured Monte Carlo error (MCSE) was smaller than 5% of the posterior standard deviation.

3.2.2 | Bridge sampling

Our approach relies heavily on estimates of $l(a_0) = \log(c(a_0))$, which are (log) marginal likelihoods, at selected values of a_0. We employed bridge sampling[18,19] to compute marginal likelihoods using the methods implemented in the R package **bridgesampling**.[20]

Let $\boldsymbol{\Theta} \subseteq \mathbb{R}^d$ and let $(\boldsymbol{\Theta}, \mathcal{F}, P)$ be a probability space. Suppose P admits a density p with respect to a measure μ and consider computing

$$Z = \int_{\boldsymbol{\Theta}} q(t) \, d\mu(t),$$

where $p(\theta) = q(\theta)/Z$. The quantity Z is usually called the normalizing constant of p, and finds use in many applications in Statistics, particularly in Bayesian Statistics. In a Bayesian context, it is usual to compute the marginal likelihood $m(\boldsymbol{X}|\mathcal{M}) := \int_{\boldsymbol{\Theta}} L(\boldsymbol{X}|t, \mathcal{M}) d\pi(t)$, where π is the prior measure, and this quantity is the **evidence** in favor of model $\mathcal{M}$. Computing Z for most models of interest involves computing a high-dimensional integral which can seldom be solved in closed-form and is thus a difficult numerical task that requires specialized techniques. As before, denote $f_{a_0}(D_0; \theta) = L(D_0|\theta)^{a_0} \pi(\theta)$. Here we are interested in estimating

$$c(a_0, D_0) = \int_{\boldsymbol{\Theta}} f_{a_0}(D_0; \theta) d\theta.$$

While previously we have omitted the dependence of $c(a_0)$ on the data D_0 for clarity, here we shall write the full expression for completeness.

The method proposed initially by Meng and Wong[18] and extended by Meng and Schilling[19] gives the estimator

$$\hat{c}(a_0, D_0) = \frac{N^{-1} \sum_{j=1}^{N} h(\tilde{\theta}_j) f_{a_0}(D_0; \tilde{\theta}_j)}{M^{-1} \sum_{i=1}^{M} h(\theta_i^*) g(\theta_i^*)}, \tag{11}$$

where $h(\cdot)$ is the bridge distribution and $g(\cdot)$ is a proposal density. We then let $\tilde{\theta} = \{\tilde{\theta}_1, \tilde{\theta}_2, \ldots, \tilde{\theta}_N\}$ and $\theta^* = \{\theta_1^*, \theta_2^*, \ldots, \theta_M^*\}$ are sets of N and M samples from $f_{a_0}(D_0; \theta)$ and h, respectively.

The performance of the estimator depends on the optimal choice of h, which in turn depends on $c(a_0, D_0)$, the target quantity. To overcome this difficulty, we use an iterative procedure to obtain the estimate from an initial guess $\hat{c}(a_0, D_0)^{(0)}$:

$$\hat{c}(a_0, D_0)^{(t+1)} = \frac{N}{M} \frac{\sum_{j=1}^{N} \frac{w(\tilde{\theta}_j)}{a w(\tilde{\theta}_j) + (1-a)\hat{c}(a_0, D_0)^{(t)}}}{\sum_{i=1}^{M} \frac{1}{a w(\theta_i^*) + (1-a)\hat{c}(a_0, D_0)^{(t)}}}, \tag{12}$$

where $w(\theta) = f_{a_0}(D_0; \theta)/g(\theta)$ and $a = M/(M + N)$. Numerically stable routines for computing (12) are implemented in the package **bridgesampling**.[20] As a note, the estimator in (11) assumes that the samples are independent and identically distributed, which is not the case when samples are obtained *via* MCMC, and hence M and N are replaced with estimates of the effective sample size (ESS). Since Stan achieves high efficiency (ESS/# samples) for most models considered here, this poses no problem. Here we use the default settings of the algorithm available in **bridgesampling**, meaning we take $g(\cdot)$ to be a multivariate proposal distribution. As explained by Gronau et al,[20] the bridge sampling algorithm is robust to the tail behavior of the target and proposal distributions as long as h is optimal, which is the case here.

3.2.3 | Generalized additive models

The approach in Section 3.1 (see also Section 3.2.2) allows us to obtain a set of J pairs $(a_{0i}, \hat{l}_i)$ from which the approximating function $g_\xi(a_0) \approx l(a_0)$ can be estimated. We now detail our tool of choice to construct g_ξ, the GAM.[21]

A GAM is a model for the conditional expectation of the dependent variable, $\mu_i := E[Y_i]$, of the form

$$g(\mu_i) = \boldsymbol{X}_i^* \boldsymbol{\eta} + \sum_{k=1}^{q} f_i(\boldsymbol{X}_i^*), \qquad (13)$$

where g is a link function, $\boldsymbol{\eta}$ is a vector of coefficients for the parametric components of the model, and the f_i are smooth functions of the covariates. In particular, here we are interested in the model

$$\hat{l}_i = \Delta + \sum_{k=1}^{q} b_k(a_{0i})\beta_k + \epsilon_i, \qquad (14)$$

and where b_k is the kth basis function, Δ is an intercept and we assume $\epsilon_i \sim \text{Normal}(0, \tau)$. We employ the routines in the **mgcv** package[22] in R to fit GAMs and make predictions.

In our applications, we employed $q = J$. This choice leads to overfitting, which in other settings would be undesirable. In our situation, however, overfitting is not a problem because the end goal is to predict the value of $l(a_0)$ within the measured range of the covariate a_0, $[m, M]$. Once we have fitted the model in (14), we have our approximating function $g_{\hat{\xi}}$, where $\hat{\xi} = \{\hat{\Delta}, \hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_q, \hat{\tau}\}$, which we can in turn use to predict $l(a_0)$ over a grid of K points covering $[m, M]$.

3.3 | Algorithm overview

In this section, we briefly describe how all the steps fit together.

1. Starting with a computational budget of J evaluations of the marginal likelihood, use the algorithm described in Section 3.1.1 to adaptively build a table of J values of $(a_0, l(a_0))$, $\boldsymbol{D}^{\text{est}}$ employing bridge sampling to estimate $l(a_0)$;
2. Using the "data" in $\boldsymbol{D}^{\text{est}}$, fit a GAM or other curve-fitting method to learn/emulate $l(a_0)$ (see Section 3.2.3);
3. Use the function fitted in the previous step to predict the value of $l(a_0)$ on a fine grid of values of a_0;
4. Using the dictionary of the previous step, sample from the approximate joint posterior, replacing $l(a_0)$ with its approximation $\hat{l}(a_0)$.

An schematic representation of the algorithm described in Section 3.1.1 can be found in Figure 1A, whereas a flowchart of the algorithm overview above is shown in Figure 1B.

We note that this structure is highly *modular*; one can estimate the marginal likelihoods in step 1 using a myriad of methods, not just bridge sampling. One can also replace the curve fitting step (step 2) with any continuous emulation function, such as Gaussian processes—see Park and Haran[23] for theoretical justification.

4 | APPLICATIONS

In this section, we discuss applications of the normalized power prior. We first discuss four examples where $c(a_0)$ is known in closed-form and use these as a benchmark for the approximations discussed in this article. Then we move on to explore two regression examples where the normalizing constant is not known in closed-form and thus only the approximation is available.

In all examples, we employed a Beta prior on a_0 with parameters $\eta = \nu = 1$ and thus restricted attention to $a_0 \in [0, 1]$ when approximating the normalizing constant—that is, we used $M = 1$. For all examples, we used $m = 0.05$ and employed budget of $J = 20$ evaluations of $l(a_0)$ via bridge sampling.

4.1 | Bernoulli likelihood

In this section, we revisit the Bernoulli example of Neuenschwander et al[5] and show how the approximation scheme proposed in Section 3.1 can be used, taking advantage of the fact that $c(a_0)$ is known exactly for this example. The historical
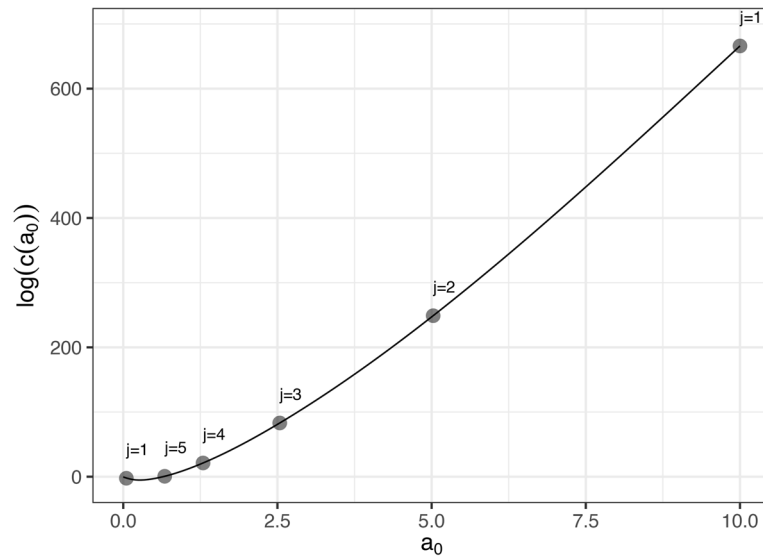
(A) Grid building.
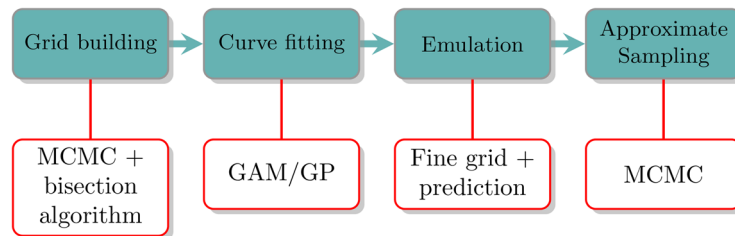
(B) Overview.

FIGURE 1 Schematic representation of the algorithm and workflow. In Panel A, we show the first five steps in the grid building algorithm: $l(m)$ and $l(M)$—that is, l at the endpoints—are computed simultaneously ($j = 1$) and then a bisection-type routine produces the next steps ($j = 2, 3, 4,$ and 5). Here we show the "zero-finding" phase of the algorithm, where consecutive steps go toward the point where the derivative changes signs. Panel B shows the whole workflow proposed here. The MCMC step can be done with Stan,[16] NIMBLE,[24] JAGS,[25] or any other suitable MCMC routine. Likewise, the curve fitting module can be carried out with a GAM, Gaussian process, or any other continuous curve-fitting method[23] [Colour figure can be viewed at wileyonlinelibrary.com]

data consist of N_0 Bernoulli trials $x_{0i} \in \{0, 1\}$. Suppose there were $y_0 = \sum_{i=1}^{N_0} x_{0i}$ successes. The model reads

$$\theta \sim \text{Beta}(c, d),$$
$$x_{0i}|\theta \sim \text{Bernoulli}(\theta).$$

This leads to a Beta posterior distribution for θ,

$$p(\theta|N_0, y_0, a_0) \propto \theta^{a_0 y_0 + c - 1}(1 - \theta)^{a_0(N_0 - y_0) + d - 1}, \tag{15}$$

and hence[5]:

$$c(a_0) = \frac{\mathcal{B}(a_0 y_0 + c, a_0(N_0 - y_0) + d)}{\mathcal{B}(c, d)}, \tag{16}$$

where $\mathcal{B}(w, z) = \frac{\Gamma(w)\Gamma(z)}{\Gamma(w+z)}$. The derivative, $c'(a_0)$, evaluates to

$$c'(a_0) = \frac{\mathcal{B}(z_0, w_0)(y_0[\psi_0(w_0) - \psi_0(z_0)] + N_0[\psi_0(z_0) - \psi_0(w_0 + z_0)])}{\mathcal{B}(c, d)}, \tag{17}$$

TABLE 1 Bernoulli example

Scenario	Data	Parameter	Unnormalized	Normalized	App. normalized
Scenario 1	$\frac{y_0}{N_0} = \frac{20}{100}$, $\frac{y}{N} = \frac{20}{100}$	θ	0.205 (0.135, 0.286)	0.204 (0.143, 0.270)	0.204 (0.144, 0.273)
		a_0	0.021 (0.001, 0.076)	0.576 (0.073, 0.979)	0.570 (0.074, 0.979)
Scenario 2	$\frac{y_0}{N_0} = \frac{10}{100}$, $\frac{y}{N} = \frac{200}{1000}$	θ	0.200 (0.176, 0.209)	0.197 (0.174, 0.223)	0.197 (0.173, 0.222)
		a_0	0.029 (0.001, 0.105)	0.361 (0.026, 0.916)	0.363 (0.027, 0.921)
Scenario 3	$\frac{y_0}{N_0} = \frac{200}{1000}$, $\frac{y}{N} = \frac{200}{1000}$	θ	0.201 (0.175, 0.227)	0.200 (0.181, 0.221)	0.200 (0.181, 0.221)
		a_0	0.002 (0.000, 0.007)	0.580 (0.077, 0.985)	0.572 (0.078, 0.979)
Scenario 4	$\frac{y_0}{N_0} = \frac{100}{1000}$, $\frac{y}{N} = \frac{200}{1000}$	θ	0.200 (0.176, 0.225)	0.195 (0.171, 0.221)	0.196 (0.172, 0.222)
		a_0	0.003 (0.000, 0.010)	0.056 (0.003, 0.161)	0.045 (0.003, 0.146)

Note: We compare estimates of both the response proportion θ and the power prior scalar a_0 using the unnormalized power prior, the exactly normalized prior (Equation 16) and an approximation obtained according to the method in Section 3.1. We approximated $l(a_0)$ using $J = 20$ grid points for estimation and $K = 20\,000$ points for prediction.

where $z_0 = a_0 y_0 + c$ and $w_0 = a_0(N_0 - y_0) + d$ and ψ_0 is the digamma function. If one observes new data $D = (N, y)$, one can then compute a posterior $p(\theta|a_0, D_0, D)$. In the situation where one lets a_0 vary by assigning it a prior $\pi_A(\cdot|\delta)$, one can write the marginal posterior for a_0 explicitly[5(eq.8)]:

$$p(a_0|D_0, D) \propto c(a_0)\pi_A(a_0|\delta)\mathcal{B}(a_0 y_0 + y + c - 1, a_0(N_0 - y_0) + (N - y) + d - 1). \tag{18}$$

Neuenschwander et al[5] thus consider the problem of estimating the probability of response θ in a survey where y of the N individuals are responders and $N - y$ are nonresponders. They consider four scenarios, that vary the historical ($D_0 = \{N_0, y_0\}$) and current data ($D = \{N, y\}$), detailed in Table 1. They employ flat Beta priors $\pi_A(a_0|\delta)$ with parameters $\eta = \nu = 1$ and $\pi(\theta)$ with parameters $c = d = 1$, which we also adopt here.

First, we show a sensitivity analysis where we computed the prior and posterior distributions for the quantity of interest θ for various ($J = 20$) values of a_0 in order to gauge how the discount factor affects the inferences reached. In Figure 2, we show the distribution of θ for various values of a_0 in each scenario. When the historical and current data are compatible ($y_0/N_0 = y/N$) as in scenarios 1 and 3, we see that the prior uncertainty encompasses the posterior for all values of a_0. In contrast, when there is incompatibility between the historical and current data sets, we see that the prior and posterior intervals stop overlapping for moderate values of a_0, an effect more prominent the larger N is. For scenario 2, we see overlap up until $a_0 \approx 0.30$, while for scenario 4, with more data, incompatibility starts to arise much earlier, around $a_0 \approx 0.05$.

We show that for all scenarios considered, posterior estimates of the response proportion θ and the power prior scalar a_0 are extremely consistent between the approximate normalization and the exact normalization given in (16). For all the scenarios considered, $l(a_0)$ looks approximately linear in a_0 as shown in Supplementary Figure E3. While we used a relatively fine grid ($K = 20\,000$) to create the $l(a_0)$ dictionary, we found that smaller values also gave good performance (data not shown, see below).

The marginal posteriors of a_0 obtained for each scenario are shown in Figure 3. As shown in Table 1, the approximately normalized power prior is in close agreement with the closed-form solution, for a variety of shapes the distribution takes across scenarios. In particular, scenarios 1 and 3 are designed such that posterior estimates of a_0 should be around 1/2 in order to reflect the fact that current data are compatible with historical data. On the other hand, scenarios 2 and 4 are designed such that there is mild incompatibility between historical and current data, and this is reflected in the properly normalized posteriors for a_0, whereas the unnormalized posteriors yield counter-intuitive results.

Looking closely at Figure 3D, however, we notice that while the mean and BCI of the approximately normalized posterior are not significantly different from the exactly normalized distribution, the shape of the marginal posterior density for a_0 does show some inconsistencies. This serves as a warning that the approximate normalization does not work equally well in all situations, and may be susceptible to nonlinearities in the sense that a small error in approximating $c(a_0)$ might have a big impact on the estimates.
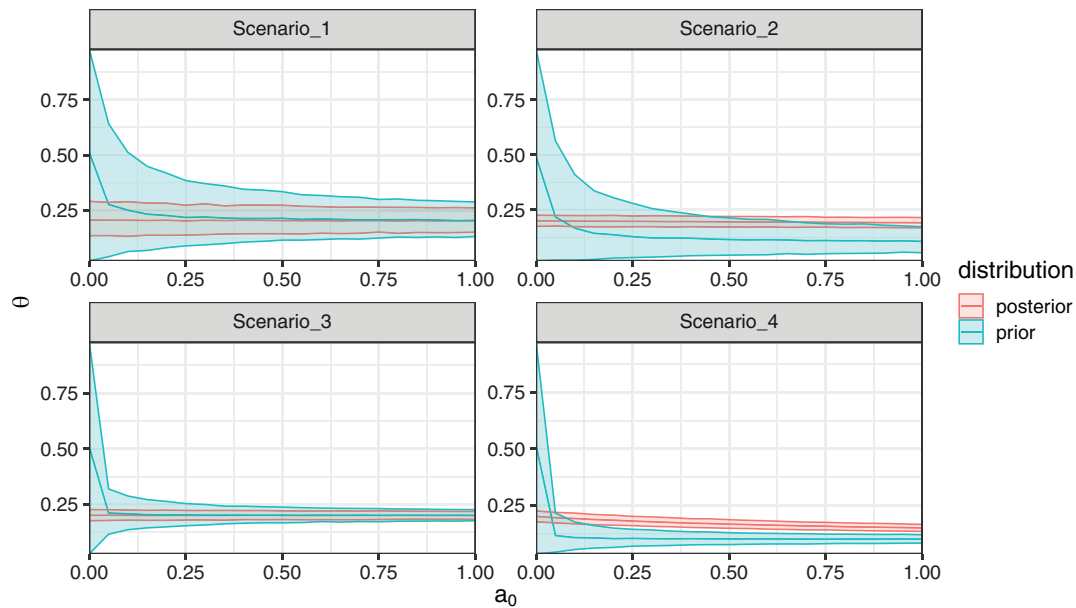
FIGURE 2 Sensitivity analysis for the Bernoulli example. We show the prior and posterior distribution for θ as the discounting factor a_0 varies. Solid lines show the posterior mean, while shaded areas depict the 95% Bayesian credibility interval (BCI). Colors show the distribution in question: the prior $\pi_{a_0}(\theta) = L(D_0|\theta)^{a_0} \pi(\theta)$ or the posterior $p_{a_0}(\theta) = L(D|\theta)\pi_{a_0}(\theta)$ [Colour figure can be viewed at wileyonlinelibrary.com]
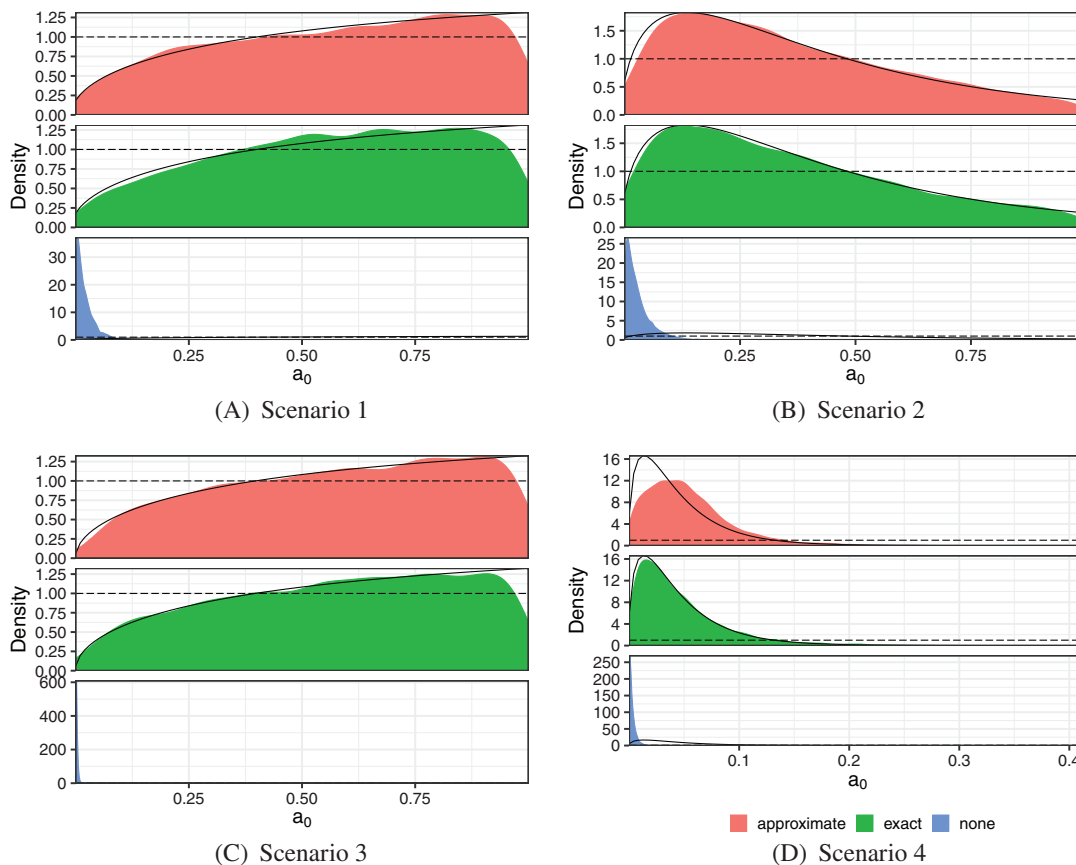

FIGURE 3 Marginal distributions of a_0 for the Bernoulli example. Solid lines show the marginal posterior of a_0 as given by (18), normalized via quadrature. Colors (and horizontal tiles) show the normalization method used: none (unnormalized), exact (normalized), or approximate. Horizontal dashed line marks the Beta prior with parameters $\eta = \nu = 1$. Please note that the x-axes differ between panels [Colour figure can be viewed at wileyonlinelibrary.com]

TABLE 2 Parameter estimates for the linear regression example

Parameter	True	None	Exact	App. $K = 50$	App. $K = 10\,000$
β_0	-1	$-0.56\ (-0.98, -0.16)$	$-0.91\ (-1.12, -0.59)$	$-0.92\ (-1.12, -0.63)$	$-0.92\ (-1.12, -0.68)$
β_1	1	$0.78\ (0.38, 1.18)$	$0.89\ (0.66, 1.09)$	$0.88\ (0.69, 1.08)$	$0.89\ (0.67, 1.08)$
β_2	0.5	$0.32\ (-0.05, 0.70)$	$0.50\ (0.24, 0.70)$	$0.50\ (0.28, 0.70)$	$0.51\ (0.29, 0.69)$
β_3	-0.5	$-0.68\ (-1.04, -0.34)$	$-0.57\ (-0.79, -0.39)$	$-0.58\ (-0.78, -0.39)$	$-0.57\ (-0.78, -0.38)$
σ^2	4	$3.7\ (2.8, 4.8)$	$4.4\ (3.6, 5.0)$	$4.4\ (3.8, 5.1)$	$4.4\ (3.8, 5.0)$
a_0	$-$	$0.00\ (0.00, 0.00)$	$0.48\ (0.05, 0.97)$	$0.43\ (0.11, 0.94)$	$0.49\ (0.09, 0.97)$

Note: We report the posterior mean and 95% BCI for the regression parameters $\boldsymbol{\beta}$, response variance σ^2, and the power prior scalar a_0. We employed a Beta($\eta = \nu = 1$) as prior for a_0.

4.1.1 | Linear regression with a normal inverse-Gamma prior

To conclude the examples for which we know $c(a_0)$ in closed-form, we present a popular model for Bayesian linear regression. Suppose $\boldsymbol{X}_0$ is a $N_0 \times P$ full-rank matrix of predictors and $\boldsymbol{y}_0 = \{y_{01}, \ldots, y_{0N_0}\}$ is a vector of observations. For illustrative purposes, we will employ a mean and variance parameterization, which naturally leads to a normal inverse-Gamma conjugate prior. The model is

$$\sigma^2 \sim \text{Inverse Gamma}(\alpha_0, \gamma_0),$$
$$\epsilon_i | \sigma^2 \sim \text{Normal}(0, \sigma^2),$$
$$\beta | \sigma^2 \sim \text{Normal}(\boldsymbol{\mu}_0, \sigma^2 \boldsymbol{\Lambda}_0^{-1}),$$
$$y_{0i} = \boldsymbol{X}_{0i}^\top \boldsymbol{\beta} + \epsilon_i,$$

where $\boldsymbol{\beta}$ is a $1 \times P$ vector of coefficients and $\boldsymbol{\Lambda}_0$ is a $P \times P$ variance-covariance matrix controlling the prior variance of the coefficients. The posterior is again a normal inverse Gamma and thus

$$c(a_0) = \sqrt{\frac{|\boldsymbol{\Lambda}_n|}{|\boldsymbol{\Lambda}_0^{-1}|}} \frac{\gamma_0^{\alpha_0}}{\gamma_n^{\alpha_n}} \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_n)} (2\pi)^{-N_0 a_0/2},$$
$$\boldsymbol{\Lambda}_n = \boldsymbol{X}_\star^\top \boldsymbol{X}_\star + \boldsymbol{\Lambda}_0^{-1},$$
$$\boldsymbol{\mu}_n = \boldsymbol{\Lambda}_n^{-1} \left(\boldsymbol{\Lambda}_0^{-1} \boldsymbol{\mu}_0 + \boldsymbol{X}_\star^\top \boldsymbol{y}_\star \right),$$
$$\alpha_n = \alpha_0 + \frac{1}{2} a_0 N_0,$$
$$\gamma_n = \gamma_0 + \frac{1}{2} \left(\boldsymbol{y}_\star^\top \boldsymbol{y}_\star + \boldsymbol{\mu}_0^\top \boldsymbol{\Lambda}_0^{-1} \boldsymbol{\mu}_0 - \boldsymbol{\mu}_n^\top \boldsymbol{\Lambda}_n \boldsymbol{\mu}_n \right), \tag{19}$$

where $\boldsymbol{X}_\star = \sqrt{a_0} \boldsymbol{X}_0$ and $\boldsymbol{y}_\star = \sqrt{a_0} \boldsymbol{y}_0$, and $|A|$ denotes the determinant of A.

As a first experiment, we generate $N_0 = 1000$ data points, drawing the columns of $\boldsymbol{X}_0$ from a standard normal distribution and using $\boldsymbol{\beta} = \{-1, 1, 0.5, -0.5\}$. The response variable Y_0 is generated using a normal distribution with variance $\sigma^2 = 4$, that is, $y_{0i} \sim \text{Normal}(\boldsymbol{\beta}^T \boldsymbol{X}_{0i}, 4)$. For the current data, we generate $N = 100$ points using the same data-generating process. To complete the model specification, we set $\alpha_0 = 1/2$, $\gamma_0 = 2$, and $\boldsymbol{\Lambda}_0 = \frac{3}{2} \boldsymbol{I}_P$, where $\boldsymbol{I}_P$ is the $P \times P$ identity matrix.

Results of the power prior analysis of this data are shown in Table 2 and indicate that while parameter recovery is similar for the exactly normalized and approximately normalized posteriors, the approximate method does not recover the lower tail of the marginal posterior of a_0 well.

Next, we explore the behavior of our approach when the dimension of the problem increases, with the goal of ascertaining if and how the performance of the method deteriorates with increasing dimension. We devised four scenarios where we keep constant the ratio $N_0/P = 10$ and make $P = 5, 10, 50, 100$ (see Table 3). For the current data, we fixed $N_0 = 100$, which leads to a near-identification configuration for Scenario D. For these experiments, we replaced the default configurations on Stan by increasing the number of iterations (from 2000 to 5000) and maximum tree size (`max_treedepth` from 10 to 15) and decreasing the step size (`adapt_delta` from 0.8 to 0.95).

TABLE 3 Scaling of the algorithm with dimension, linear regression example

		Scenario			
		A	**B**	**C**	**D**
	Normalization	$N_0 = 50, P = 5$	$N_0 = 100, P = 10$	$N_0 = 500, P = 50$	$N_0 = 1000, P = 100$
CI width (inclusion)	None	0.79 (0.8)	0.64 (1)	0.89 (0.94)	1.26 (0.87)
	Approximate	0.67 (0.6)	0.49 (0.9)	0.39 (0.98)	0.28 (1)
	Exact	0.67 (0.6)	0.49 (0.9)	0.38 (0.98)	0.28 (1)
MSE $\boldsymbol{\beta}$ ($\times 10^{-2}$)	None	6	3	5	1.6
	Approximate	5	1.62	0.8	0.31
	Exact	5	1.6	0.8	0.31
MRAE $l(a_0)$ ($\times 10^{-4}$)	–	7.7	0.87	0.73	0.52

Note: For each scenario, we show the mean relative absolute error (MRAE) of the estimated $l(a_0)$ for $J = 20$ points. We show the average width of the (95%) credibility intervals (CIs) as well as the CIs that included the true data-generating coefficients ("inclusion") for the unnormalized, approximately normalized, and exactly normalized posteriors. We also show the mean squared error (MSE) in the estimation of $\boldsymbol{\beta}$. For comparison, the MRAE for $N_0 = 1000$ and $P = 5$ was 0.05×10^{-4}.

The results in Table 3 suggest that even in the extreme case of scenario D, with $P = 100$ parameters and $N_0 = 1000$ data points our approach is able to accurately approximate the normalizing constant and the approximately normalized posteriors compare favorably to their exactly normalized counterparts. Perhaps counterintuitively, the MRAE in the estimation of the (log) normalizing constant decreases with dimension. We hypothesize this is the effect of the function $l(a_0)$ increasing in absolute value while the estimation method (bridge sampling) does not lose precision at quite the same rate, leading to relatively more precise estimates for values of a_0 closer to 1. In general, failing to account for the normalizing constant leads to broader credibility intervals and worse estimates of the coefficients (using the marginal posterior mean) in terms of MSE.

We present the estimated $l(a_0)$ in each scenario in Supplementary Figure E4 and show that the derivative-based method discussed briefly in Section 3.1.2 performs worse as the dimension of the problem increases, as expected. This is because it gets progressively harder to reliably estimate the derivative of $l(a_0)$ as the dimension of the parameter space increases. As a general takeaway we note that while the method remains accurate for this admittedly simple but high-dimensional problem, we needed to change the computational specifications to increase precision (eg, increase the number of iterations) and also use a finer approximation grid for $l(a_0)$, in particular, $K = 50\,000$ and $K = 100\,000$ points. For scenario D, even using $K = 50\,000$ did not lead to a good approximation (Supplementary Figure E5D).

4.2 | Logistic regression

Next, we approach a problem for which $c(a_0)$ cannot be written in closed-form. Logistic regression is very popular model for binary outcomes in the presence of explanatory variables (covariates). Taking $\boldsymbol{Y}_0 = \{y_{01}, y_{02}, \dots, y_{0N_0}\}$ with $y_{0i} \in \{0, 1\}$ and a (assumed full rank) $N_0 \times P$ matrix of covariates $\boldsymbol{X}_0$ as historical data, the model we consider here is

$$y_{0i} \sim \text{Bernoulli}(\theta_i),$$
$$\theta_i = \frac{\exp(\alpha + \boldsymbol{X}_{0i}^T \boldsymbol{\beta})}{1 + \exp(\alpha + \boldsymbol{X}_{0i}^T \boldsymbol{\beta})},$$
$$\alpha \sim \text{Normal}(0, 1),$$
$$\beta_i \sim \text{Normal}(0, 1),$$

where α is the intercept and $\boldsymbol{\beta}$ is a P-dimensional vector of coefficients. Since we do not have the benefit of a closed-form $c(a_0)$ in this example, we simulate data with known parameters and study how parameters are recovered as a function of
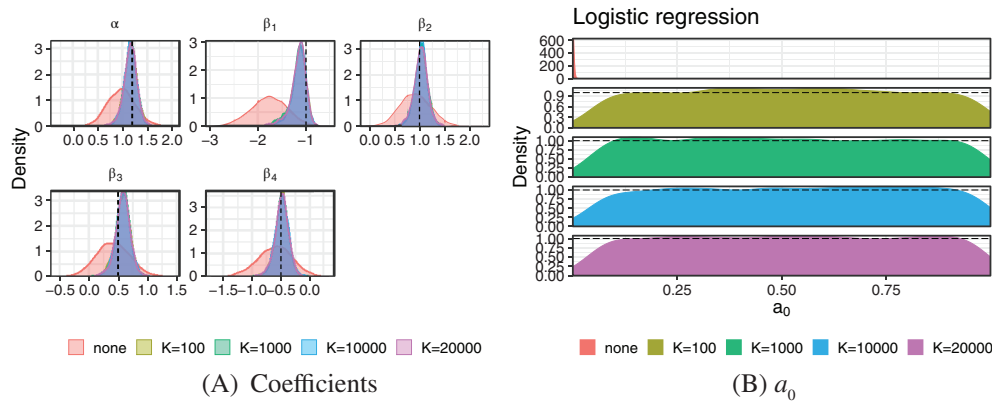
FIGURE 4   Results for the logistic regression example. Panel A shows the marginal posterior distributions for model parameters, with colors again pertaining to the approximation scheme. Vertical dashed lines show the "true" parameter values of the data-generating process. Horizontal dashed lines show the prior density of a Beta($\eta = 1, \nu = 1$) for a_0. In Panel B, the subpanels (and colors) correspond to the posterior distribution of the parameter a_0 when $c(a_0)$ is accounted for using various grid sizes K and when it is not included [Colour figure can be viewed at wileyonlinelibrary.com]

the grid size K. First, we generate $N_0 = 1000$ historical data points $(\boldsymbol{Y}_0, \boldsymbol{X}_0)$, where the matrix of covariates is constructed in the same manner as in the linear regression example. We set $\alpha = 1.2$ and $\boldsymbol{\beta} = \{-1, 1, 0.5, -0.5\}$. For the current data, we use the same data-generating process to create a set of $N = 100$ new data points $(\boldsymbol{Y}, \boldsymbol{X})$. A prior sensitivity analysis (PSA) is shown in Supplementary Figure E6. The chief idea is that a properly normalized power prior would allow one to capture the similarities between the historical and current data, while an analysis lacking the proper normalization would yield counter-intuitive and suboptimal results, as demonstrated in the previous examples. The results shown in Figure 4A seem to support this intuition, since the approximately normalized power prior leads to posterior estimates that better recover the generating parameters, while the unnormalized prior leads to more diffuse posteriors that do not capture the full information contained in the data.

In addition, we see that the approximate posteriors start to stabilize for $K > 1000$, showcasing the increased difficulty of this multi-dimensional problem (see also Section 4.1.1). The bimodal marginal posterior for a_0 (Figure 4B) suggests high uncertainty about the compatibility of the current data and historical data, which is unsurprising given the small number of current observations. In order to gauge the dependence of these results on our specific parameter choices, in Supplementary Figure E7A,B, we show results for a similar setup with $\alpha = 0.2$ and $\boldsymbol{\beta} = \{-10, 1, 5, -5\}$. Results indicate that in this scenario with larger (absolute value) coefficients, our approximation still works well, albeit with less posterior coverage of the data-generating values. Overall, the results of this section suggest that our approach works well in a setting where the normalizing constant is not known in closed-form and leads to posterior estimates that appropriately incorporate the information in the historical data.

4.3 | Survival model with cure fraction

As a final illustration, we show an application of the approximation scheme to an elaborate survival model, namely, the cure rate model proposed by Chen et al.[26] This model allows one to accommodate situations where a significant proportion of subjects is cured. The model can be described generatively as follows. Let N be the number of carcinogenic cells left after initial treatment, assumed to follow a Poisson distribution with rate θ. Now let $Z_j, j = 1, 2, \ldots, N$ be i.i.d. random variables with distribution function $F(t) = 1 - S(t)$. The variable of interest is then $T = \min(Z_j), 0 \leq j \leq N$, the time of relapse. Suppose we observe i.i.d. data $\boldsymbol{Y} = \{y_1, y_2, \ldots, y_n\}$ with $\boldsymbol{W} = \{w_1, w_2, \ldots, w_n\}$ being indicators of whether observations are (right) censored. If we also have a $n \times p$ matrix of covariates $\boldsymbol{X}$, we can then write the likelihood after marginalizing over the latent variables:

$$L(\boldsymbol{\beta}, \boldsymbol{\psi} | \boldsymbol{Y}, \boldsymbol{W}) = \prod_{i=1}^{n} (\theta_i f(y_i | \boldsymbol{\phi}))^{w_i} \exp\left[-\theta_i (1 - S(y_i | \boldsymbol{\phi}))\right], \qquad (20)$$

where $\theta_i = \exp(X_i^T \beta)$, with β a vector of coefficients and $\psi = (\alpha, \lambda)$ the parameters of a Weibull distribution, that is,

$$f(y_i|\psi) = \alpha y_i^{\alpha-1} \exp\left[\lambda - y_i^\alpha \exp(\lambda)\right].$$

For more details, see Chen et al.[26]

To illustrate the use of a normalized power prior, we will consider a situation where one wants to analyze data from a current clinical trial in light of historical information provided by an earlier study which includes many of the same covariates and measurements. In particular, we consider data from a two-arm clinical trial on phase III melanoma conducted by the Eastern Cooperative Oncology Group, denoted E1684. In this study, patients were assigned to either a interferon treatment (IFN) or observation, and survival was defined as time from randomization to death. We have $n = 284$ measurements for this data set. As historical data, we employ the data from an earlier essay, denoted E1673, for which we have $n_0 = 650$ data points. We consider three covariates: (standardized) age, sex, and performance status (PS), that is, whether the patient was fully active or other. Since in this article we are chiefly concerned with situations where the initial priors are proper, we modify the prior modeling of Reference 26 to include proper priors for all parameters. Namely, we employ the following prior structure:

$$\beta \sim \text{Normal}(\mathbf{0}, \sigma_\beta^2 I_P),$$

$$\alpha \sim \text{Gamma}(\delta_0, \tau_0),$$

$$\lambda \sim \text{Normal}(\mu_0, \sigma_0^2),$$

with $\sigma_\beta^2 = 10$, $\delta_0 = 1$, $\tau_0 = 0.01$, $\mu_0 = 0$, and $\sigma_0^2 = 10\,000$. Letting $D_0 = \{Y_0, W_0, X_0\}$, the normalized joint power prior and posterior are, respectively,

$$\pi(\beta, \psi, a_0|D_0) = \frac{L(\beta, \psi|D_0)^{a_0} \pi(\beta, \psi) \pi_A(a_0)}{c(a_0)}, \tag{21}$$

and

$$p(\beta, \psi, a_0|D_0, D) \propto \frac{L(\beta, \psi|D_0)^{a_0} \pi(\beta, \psi) L(\beta, \psi|D) \pi_A(a_0)}{c(a_0)}, \tag{22}$$

where $D = \{Y, W, X\}$ and we take $\pi_A(\cdot)$ to be a Beta prior with parameters η and ν, taking values as in Table 4. In their original analysis of the melanoma data, Chen et al[26] employed an unnormalized power prior. Here, we revisit their analysis (Table 4 therein) and employ a normalized power prior which can then be compared to the results with an unnormalized prior (Table 4). The sensitivity analysis presented in Supplementary Figure E6 suggests that all model parameters are very sensitive to small values of a_0, whereas posteriors stabilize for values of $a_0 > 0.1$.

The results in Table 4 show that including the normalization factor $c(a_0)$ leads to substantially different parameter estimates. In particular, the normalized prior leads to the posterior (equal-tailed, 95%) Bayesian credibility interval (BCI) for the coefficient of age and sex excluding zero, showing that when information is properly accounted for through correct normalization of the power prior, inferences might change. Looking at the posterior estimates for a_0, we note that these changes likely stem from the fact that when employing a (approximately) normalized power prior, we give the historical data more weight and thus effectively increase the amount of data entering the model.

5 | DISCUSSION

5.1 | Starting from a sensitivity analysis

The starting point for the methodology presented here is a PSA, in which one computes the distribution $L(D_0|\theta)^{a_0} \pi(\theta)$ for a range of values of a_0 in order to gauge how sensitive the resulting prior is to the discounting (tempering) parameter. The class of models amenable to such an analysis thus comprises models: (i) that are well-established/studied and thus there is little need to test different likelihood functions or even initial priors; and (ii) for which one is able to compute the estimates in reasonable time such that a sensitivity analysis of the sort discussed here is feasible.

TABLE 4 Results for the cure fraction rate model

Prior on a_0, Beta(η, ν)	Parameter	Unnormalized	App. normalized
$\eta = 1, \nu = 1$	Intercept	0.10 (−0.11, 0.31)	0.47 (0.24. 0.70)
	Age	0.09 (−0.05,0.23)	0.15 (0.03, 0.26)
	Sex	−0.13 (−0.44, 0.18)	−0.31 (−0.48, −0.13)
	PS	−0.23 (−0.76, 0.25)	−0.04 (−0.33, 0.36)
	α	1.30 (1.13, 1.48)	1.02 (0.93, 1.12)
	λ	−1.36 (−1.63, −1.12)	−1.80 (−2.03, −1.55)
	a_0	0.00 (0.00, 0.00)	0.41 (0.19, 0.94)
$\eta = 50, \nu = 50$	Intercept	0.20 (−0.02, 0.44)	0.50 (0.29, 0.71)
	Age	0.10 (−0.04, 0.24)	0.15 (0.05, 0.26)
	Sex	−0.17 (−0.44, 0.10)	−0.33 (−0.49, −0.19)
	PS	−0.19 (−0.72, 0.27)	0.07 (−0.25, 0.37)
	α	1.17 (1.02, 1.32)	1.01 (0.92, 1.10)
	λ	−1.47 (−1.75, −1.21)	−1.83 (−2.05, −1.62)
	a_0	0.03 (0.02, 0.04)	0.48 (0.37, 0.60)
$\eta = 100, \nu = 100$	Intercept	0.28 (0.05. 0.52)	0.51 (0.29, 0.72)
	Age	0.11 (−0.02, 0.16)	0.16 (0.05, 0.26)
	Sex	−0.20 (−0.43, 0.02)	−0.33 (−0.49, −0.18)
	PS	−0.15 (−0.61, 0.27)	0.07 (−0.27, 0.37)
	α	1.11 (0.98, 1.24)	1.01 (0.92, 1.09)
	λ	−1.56 (−1.84, −1.30)	−1.83 (−2.05, −1.63)
	a_0	0.07 (0.05, 0.08)	0.49 (0.42, 0.56)
$\eta = 200, \nu = 1$	Intercept	0.38 (0.14, 0.63)	0.54 (0.36, 0.72)
	Age	0.13 (0.01, 0.25)	0.17 (0.09, 0.26)
	Sex	−0.25 (−0.45, 0.06)	−0.36 (−0.49, −0.24)
	PS	−0.09 (−0.52, 0.31)	0.15 (−0.11, 0.39)
	α	1.05 (0.93, 1.17)	1.00 (0.93, 1.07)
	λ	−1.68 (−1.96, −1.48)	−1.89 (−2.06, −1.73)
	a_0	0.14 (0.12. 0.16)	1.00 (0.98, 1.00)

Note: For several choices of prior for a_0, we show posterior means and 95% BCIs for the model coefficients as well as $\psi = \{\alpha, \lambda\}$ and a_0 under no normalization and approximate normalization using the methods proposed in Section 3.1. We employed $J = 20$ evaluations to estimate $c(a_0)$ and $K = 2E4$ points for the approximation grid.

Taking these conditions as given, we then first propose a simple way of picking a fixed budget of, say, $J = 20$, values for a_0 at which to compute the power prior distribution using a bisection-type algorithm based on the theoretical results in Section 2.1. Since there are many instances in which one would wish to represent the uncertainty about a_0 as probability distribution, we propose a way to recycle computations in order to approximately sample from the joint posterior of $(a_0, \boldsymbol{\theta})$ where $\boldsymbol{\theta}$ are the parameters of interest. This requires computing the normalizing constant (Equation 3) as observed by Neuenschwander et al.[5]

Sensitivity to the effects of normalization varies between models and data configurations; the Bernoulli model in Section 4.1 shows little difference in parameter estimates between unnormalized and normalized posteriors, while for the regression and survival examples—Sections 4.1.1 and 4.3, respectively—parameter estimates, in particular their precisions are affected more strongly. In terms of shape, the normalizing constant $c(a_0)$ seen as function of the discounting scalar a_0 is usually monotonic, at least for majority of the examples we have considered. For "discrete" likelihoods, such

as the Poisson (see Appendix B) and negative binomial models, we know $c(a_0)$ is monotonically decreasing (Remark 2), which simplifies the grid-building step of our algorithm.

The notable exception is the somewhat artificial example of a Gaussian likelihood presented in Appendix C (Figure C1A) for which $c(a_0)$ resembles a convex parabola, illustrating the importance of the results in Section 2.1, which tell us that the normalizing constant is a strictly convex—but not necessarily monotonic—function of the discounting scalar, a_0. This problem was devised so as to test our ability to approximate $c(a_0)$ in a difficult setting, namely, when it is not monotonic and varies over a large range—we give more motivation for a theoretical analysis in Appendix C.1. Results indicate the method proposed here is able to correctly approximate the normalizing constant and thus provide a usable technique when $c(a_0)$ is not known in closed-form.

5.2 | The normalized power prior as a doubly intractable problem

The normalized power prior is closely related to the class of doubly intractable problems, which encompasses Markov random fields[27] and exponential random graph models[28] and many others. For a review, see Park and Haran.[8]

To see how our problem fits into the doubly intractable framework, we can rewrite Equation (5) as

$$p(a_0, \theta | D_0, D, \delta) \propto L(D | \theta, a_0) \pi(\theta, a_0 | D_0, \delta),$$
$$\propto L(D | \theta, a_0) h(\theta | a_0, D_0) \pi_A(a_0 | \delta),$$

with $h(\theta | a_0, D_0) := c(a_0)^{-1} L(\theta | D_0)^{a_0} \pi(\theta)$ playing the part of an intractable likelihood where θ is seen as data. With the exception of the double Metropolis-Hastings algorithm of Liang,[29] most computational tools available rely on the ability to simulate from $h(\theta | a_0, D_0)$ relatively easily, which is often not the case in our setting (see below).

On the other hand, the fact that $c(a_0)$ is univariate allows our approximation scheme to be feasible for many models. In contrast, extending our approach to multiple historical data sets (Remark 1) would thus be a nontrivial task, since one would need to "observe" $c(a_{01}, a_{02}, \dots, a_{0M})$ at many points (on a M-dimensional grid) in order to obtain a good approximation.

5.3 | Exact and inefficient or inexact and efficient?

In contrast to many existing algorithms such as auxiliary variable MCMC,[30] the methodology we put forth in this article does not lead to sampling from the exact joint posterior of a_0 and θ. Our method is what Park and Haran[8] call an "asymptotically inexact" algorithm because we replace the true (power) prior with an approximate density with normalizing constant $g_{\hat{\xi}}(a_0)$.

While it would obviously be preferable to have an exact algorithm, it is important to strike a balance between simplicity and exactitude. The noise in an inexact algorithm can be decomposed into approximation error and Monte Carlo error, whereas the noise an exact algorithm comes solely from the Monte Carlo approximation and thus can, in theory, be made arbitrarily small. In practice, however, it is entirely possible for the error from a suboptimally implemented exact algorithm to be larger than that of an efficient inexact method. Almost all available state-of-the-art exact samplers for doubly intractable problems require careful consideration of the proposal distributions, as in the case of the double Metropolis-Hastings sampler of Liang,[29] and/or the ability to easily simulate from the intractable likelihood,[8,31,32] which in our case is not feasible.

Here we have devised a simple framework that employs the very efficient dynamic Hamiltonian Monte Carlo implemented in Stan[16] and requires very little programming effort to include virtually any model. Our results show that the adaptive grid-building with GAM-based approximation works well for a range of problems and this gives us confidence that in this instance one should prefer an efficient inexact algorithm to a potentially inefficient exact one.

5.4 | Current limitations and future directions

The method presented here can be improved in many respects. First, it is possible that better approximations to $c(a_0)$ could be devised by using custom curve-fitting methods that incorporate the fact that $c'(a_0)$—and $l'(a_0)$—is monotonically

increasing (see Section 2.1), such as the Gaussian process methods discussed by Riihimäki and Vehtari[33] and Wang and Berger.[34] In particular, the Gaussian process-based function emulation approach of Park and Haran[23] has good theoretical properties and guarantees that the approximate target is close to the true target in the total variation sense.

Second, extending the methodology here to multiple historical data sets is straightforward only under the assumption of independence between the a_{0k}. The grid-based approach that we have shown to work well here is going to scale poorly with dimension in the sense that if one has K historical data sets and decides to use J points for the sensitivity analysis, one ends up computing KJ posteriors. Moreover, under nonindependence incorporating uncertainty about multiple weights at once would necessitate careful consideration of the prior distribution over $\boldsymbol{a_0}$.

Finally, while our inexact approach makes it possible for practitioners to perform sensitivity analyses and sample from the approximate joint posterior efficiently, this should not discourage the development of more efficient exact algorithms. The main challenge for the normalized power prior in particular is that it is not easy to sample from $h(\theta|a_0, D_0)$, making it difficult to implement auxiliary variable-type algorithms. This points to double-Metropolis-type algorithms[29] as the most promising class of exact algorithms to be developed for the analysis of the normalized power prior.

CONFLICT OF INTEREST
The authors declare no conflicts of interest.

AUTHOR CONTRIBUTIONS
LMC and JGI conceived the study, LMC wrote the computer code. Both authors wrote the article.

DATA AVAILABILITY STATEMENT
All of the data used in this article can be found at https://github.com/maxbiostat/propriety_power_priors.

ORCID
Luiz Max Carvalho https://orcid.org/0000-0001-5736-5578
Joseph G. Ibrahim https://orcid.org/0000-0003-2428-6552

REFERENCES
1. Ibrahim JG, Chen MH. Power prior distributions for regression models. *Stat Sci*. 2000;15(1):46-60.
2. Ibrahim JG, Chen MH, Gwon Y, Chen F. The power prior: theory and applications. *Stat Med*. 2015;34(28):3724-3749.
3. Zellner A. Information processing and Bayesian analysis. *J Econ*. 2002;107(1-2):41-50.
4. Ibrahim JG, Chen MH, Sinha D. On optimality properties of the power prior. *J Am Stat Assoc*. 2003;98(461):204-213.
5. Neuenschwander B, Branson M, Spiegelhalter DJ. A note on the power prior. *Stat Med*. 2009;28(28):3562-3566.
6. Duan Y, Smith EP, Ye K. Using power priors to improve the binomial test of water quality. *J Agric Biol Environ Stat*. 2006;11(2):151.
7. Duan Y, Ye K, Smith EP. Evaluating water quality using power priors to incorporate historical information. *Environmetrics Official J Int Environmetrics Soc*. 2006;17(1):95-106.
8. Park J, Haran M. Bayesian inference in the presence of intractable normalizing functions. *J Am Stat Assoc*. 2018;113(523):1372-1390.
9. Savitsky TD, Toth D. others. Bayesian estimation under informative sampling. *Electron J Stat*. 2016;10(1):1677-1708.
10. Friel N, Pettitt AN. Marginal likelihood estimation via power posteriors. *J Royal Stat Soc Ser B (Stat Methodol)*. 2008;70(3):589-607.
11. Nelder JA, Wedderburn RW. Generalized linear models. *J Royal Stat Soc Ser A (General)*. 1972;135(3):370-384.
12. McCullagh P, Nelder J. *Generalized Linear Models*. 2nd ed. Boca Raton, FL: Chapman & Hall; 1989.
13. Diaconis P, Ylvisaker D. Conjugate priors for exponential families. *Ann Stat*. 1979;7(2):269-281.
14. Gelman A, Meng XL. Simulating normalizing constants: from importance sampling to bridge sampling to path sampling. *Stat Sci*. 1998;13(2):163-185.
15. van Rosmalen J, Dejardin D, van Norden Y, Löwenberg B, Lesaffre E. Including historical data in the analysis of clinical trials: Is it worth the effort? *Stat Methods Med Res*. 2018;27(10):3167-3182.
16. Carpenter B, Gelman A, Hoffman M, et al. Stan: a probabilistic programming language. *J Stat Softw*. 2017;76(1):1-32. https://doi.org/10.18637/jss.v076.i01.
17. Neal RM. MCMC using Hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*. Vol 2. Boca Raton, FL: CRC Press; 2011:11.

18. Meng XL, Wong WH. Simulating ratios of normalizing constants via a simple identity: a theoretical exploration. *Stat Sin*. 1996;6(4):831-860.

19. Meng XL, Schilling S. Warp bridge sampling. *J Comput Graph Stat*. 2002;11(3):552-586.

20. Gronau QF, Singmann H, Wagenmakers EJ. Bridgesampling: an R package for estimating normalizing constants; 2017. arXiv preprint arXiv:1710.08162.

21. Wood SN. *Generalized Additive Models: An Introduction with R*. Boca Raton, FL: CRC Press; 2017.

22. Wood SN. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *J Royal Stat Soc Ser B (Stat Methodol)*. 2011;73(1):3-36.

23. Park J, Haran M. A function emulation approach for doubly intractable distributions. *J Comput Graph Stat*. 2020;29(1):66-77.

24. de Valpine P, Turek D, Paciorek CJ, Anderson-Bergman C, Lang DT, Bodik R. Programming with models: writing statistical algorithms for general model structures with NIMBLE. *J Comput Graph Stat*. 2017;26(2):403-413.

25. Plummer M. JAGS: a program for analysis of Bayesian graphical models using Gibbs sampling. Paper presented at: Proceedings of the 3rd International Workshop on Distributed Statistical Computing; Vol. 124, 2003:1-10; Vienna, Austria.

26. Chen MH, Ibrahim JG, Sinha D. A new Bayesian model for survival data with a surviving fraction. *J Am Stat Assoc*. 1999;94(447):909-919.

27. Besag J. Spatial interaction and the statistical analysis of lattice systems. *J Royal Stat Soc Ser B (Methodol)*. 1974;36(2):192-225.

28. Robins G, Pattison P, Kalish Y, Lusher D. An introduction to exponential random graph (p*) models for social networks. *Soc Netw*. 2007;29(2):173-191.

29. Liang F. A double Metropolis–Hastings sampler for spatial models with intractable normalizing constants. *J Stat Comput Simul*. 2010;80(9):1007-1022.

30. Møller J, Pettitt AN, Reeves R, Berthelsen KK. An efficient Markov chain Monte Carlo method for distributions with intractable normalising constants. *Biometrika*. 2006;93(2):451-458.

31. Murray I, Ghahramani Z, MacKay D. MCMC for doubly-intractable distributions; 2012. arXiv preprint arXiv:1206.6848.

32. Stoehr J, Benson A, Friel N. Noisy Hamiltonian Monte Carlo for doubly intractable distributions. *J Comput Graph Stat*. 2019;28(1):220-232.

33. Riihimäki J, Vehtari A. Gaussian processes with monotonicity information. Paper presented at: Proceedings of the 13th International Conference on Artificial Intelligence and Statistics. Sardinia, Italy: Chia Laguna Resort; Vol 9, 13-15 May. 2010:645-652. https://proceedings.mlr.press/v9/.

34. Wang X, Berger JO. Estimating shape constrained functions using Gaussian processes. *SIAM/ASA J Uncertain Quant*. 2016;4(1):1-25.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

APPENDIX A. ADDITIONAL RESULTS AND PROOFS

Proof of Theorem 1. Denote $f_{a_0}(D_0; \theta) := L(D_0|\theta)^{a_0}\pi(\theta)$. First, note that $c(0) = 1$ because π is proper. For $0 < a_0 \leq 1$, the function $g(x) = x^{a_0}$ is concave and thus, by Jensen's inequality and the finiteness of $L(D_0|\theta)$ for all of its arguments we have

$$c(a_0) = \int_\Theta f_{a_0}(D_0; \theta) \ d\theta \leq \left[\int_\Theta L(D_0|\theta)\pi(\theta) \ d\theta\right]^{a_0} < \infty.$$

Rewrite $f_{a_0}(D_0; \theta) = L(D_0|\theta)^{a_0-1}L(D_0|\theta)\pi(\theta)$. If $1 \leq a_0 \leq 2$, we have the Jensen's inequality case above, since we know that $L(D_0|\theta)\pi(\theta)$ is normalizable (proper). Similarly, if $2 \leq a_0 \leq 3$, we can write

$$f_{a_0}(D_0; \theta) = L(D_0|\theta)^{a_0-p}L(D_0|\theta)^p\pi(\theta),$$

with $1 \leq p \leq 2$, again falling into the same case, since we know that $L(D_0|\theta)^p\pi(\theta)$ is normalizable. We can then show that for any $n \in \mathbb{N}$, $\int_\Theta f_{a_0}(D_0; \theta) \, d\theta < \infty$ for $n - 1 \leq a_0 \leq n$. The base case for $1 \leq n \leq 3$ is established. Now suppose the hypothesis holds for $n \geq 3$. For $n \leq a_0 \leq n + 1$ and $n - 1 \leq p_n \leq n$:

$$\int_\Theta L(D_0|\theta)^{a_0-p_n}L(D_0|\theta)^{p_n}\pi(\theta) \ d\theta < \infty,$$

because $0 \leq a_0 - p_n \leq 1$ and $L(D_0|\theta)^{p_n}\pi(\theta)$ is proper by hypothesis, establishing the case for $n + 1$. ∎

Remark 3 (Improper initial priors). If π is improper but $L(\theta|D_0)\pi(\theta)$ is integrable, that is, the posterior is proper, then Theorem 1 holds for $a_0 > 0$.

Proof. Analogous to the proof of Theorem 1, only excluding the boundary case $a_0 = 0$. ∎

Now let us prove Remark 1:

Proof. Recall that the power prior on multiple historical data sets is of the form[2(eq.2.9)]:

$$\pi(\theta|\boldsymbol{D}, \boldsymbol{a}_0) \propto \prod_{k=1}^{M} L(\theta|D_k)^{a_{0k}} \pi_0(\theta).$$

Assume, without loss of generality, that $L(\theta|D_k)^{a_{0k}} > 1$ for all θ and let $m := \max(\boldsymbol{a}_0)$ with $\boldsymbol{a}_0 := \{a_{01}, a_{02}, \ldots, a_{0M}\}$. Then $\pi(\theta|\boldsymbol{D}, a_0)$ is bounded above by

$$g(\theta) := \prod_{k=1}^{M} L(\theta|D_k)^m \pi_0(\theta) = \left[\prod_{k=1}^{M} L(\theta|D_k)\right]^m \pi_0(\theta) = L(\theta|\boldsymbol{D})^m \pi_0(\theta),$$

which is normalizable following Theorem 1. To relax the assumption made in the beginning, notice that this construction also bounds the case $0 \leq L(\theta|D_k)^{a_{0k}} \leq 1$ (for some k) above. ∎

We first prove Proposition 1.

Proof. First, we will assume that $L(D_0|\theta) > 0 \; \forall \theta \in \Theta$. Now, consider the change of variables $\theta \mapsto l$, with $l = \log(L(D|\theta))$. Then we write

$$h_{a_0}(l) = \frac{\exp(a_0 l)g(l)}{z(a_0)},$$

where $g(l)$ is a nonnegative function that accommodates the transform $\theta \mapsto l$ with respect to the prior π and $z(a_0)$ is the appropriate normalizing constant, guaranteed to exist by Theorem 1. The moment-generating function (MGF) of l is

$$M_t(l) = E_h[\exp(tl)] = \int_{-\infty}^{\infty} \frac{\exp((t + a_0)l)g(l)}{z(a_0)} \; dl.$$

Since $E_h[l^r] \equiv \frac{d^r c(a_0)}{da_0^r}$, all that remains is to show that $M_r(l)$ exists for all $r \geq 0$. Under the change of variables discussed above, Theorem 1 shows that

$$\int_{-\infty}^{\infty} \exp(wl)g(l) \; dl < \infty,$$

for $w > 0$. Making $w = t + a_0$ concludes the proof. ∎

Now we establish Lemma 1.

Proof. Define the normalizing constant as a function $c : [0, \infty) \to (0, \infty)$,

$$c(a_0) := \int_{\Theta} L(D_0|\theta)^{a_0} \pi(\theta) \; d\theta, \tag{A1}$$

which is positive and continuous on its domain. The first and second derivatives are

$$c'(a_0) = \int_{\Theta} L(D_0|\theta)^{a_0} \pi(\theta) \log L(D_0|\theta) \; d\theta, \tag{A2}$$

$$c''(a_0) = \int_{\Theta} L(D_0|\theta)^{a_0} \pi(\theta)[\log L(D_0|\theta)]^2 \; d\theta, \tag{A3}$$

and the integrals always exist (as per Proposition 1). Differentiation under the integral sign is justified because both $L(D_0|\theta)^{a_0}\pi(\theta)$ and $L(D_0|\theta)^{a_0}\pi(\theta)\log L(D_0|\theta)$ are continuous with respect to θ. From this we conclude that c is (strictly) convex and c' is monotonic, because c'' is always positive. ∎

APPENDIX B. POISSON LIKELIHOOD

Now consider another simple discrete example, the modeling of counts. Suppose the historical data consist of N_0 observations $y_{0i} \in \{0, 1, \dots\}$, assumed to come from a Poisson distribution. For simplicity, we will again consider the conjugate case:

$$\lambda \sim \text{Gamma}(\alpha_0, \beta_0),$$
$$y_{0i}|\lambda \sim \text{Poisson}(\lambda).$$

The posterior distribution is

$$p(\lambda|\boldsymbol{y}_0) \propto \frac{1}{\boldsymbol{p'}^{a_0}} \lambda^{a_0\boldsymbol{s}} \exp(-a_0 N_0 \lambda) \times \lambda^{\alpha_0-1} \exp(-\beta_0\lambda), \tag{B1}$$

where $\boldsymbol{s} := \sum_{i=0}^{N_0} y_{0i}$ and $\boldsymbol{p'} := \prod_{i=0}^{N_0} y_{0i}!$, leading the closed-form expression

$$c(a_0) = \frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)} \frac{1}{\boldsymbol{p'}^{a_0}} \frac{\Gamma(a_0\boldsymbol{s} + \alpha_0)}{(a_0 N_0 + \beta_0)^{a_0\boldsymbol{s}+\alpha_0}}. \tag{B2}$$

For this model, we have (see Remark 2):

$$c'(a_0) = \left[-\log(\boldsymbol{p'}) - \frac{N_0(\alpha_0 + \boldsymbol{s}a_0)}{\beta_0 + N_0 a_0} - \boldsymbol{s}\log(\beta_0 + N_0 a_0) + \boldsymbol{s}\psi_0(\alpha_0 + \boldsymbol{s}a_0) \right] c(a_0). \tag{B3}$$

We can use this example to study the quality of the approximation to $c(a_0)$ as the number of grid points K increases. For the experiment in this section, we simulated $N_0 = 200$ historical data points $\boldsymbol{y}_0$ with $\lambda = 2$ and $N = 100$ current data points $\boldsymbol{y}$ with the same rate parameter. The prior hyperparameters are $\alpha_0 = \beta_0 = 2$. Figure B1 shows the resulting marginal posteriors for a_0 and λ using several values of the grid size, K, in the approximation grid for $l(a_0)$. Even for relatively small
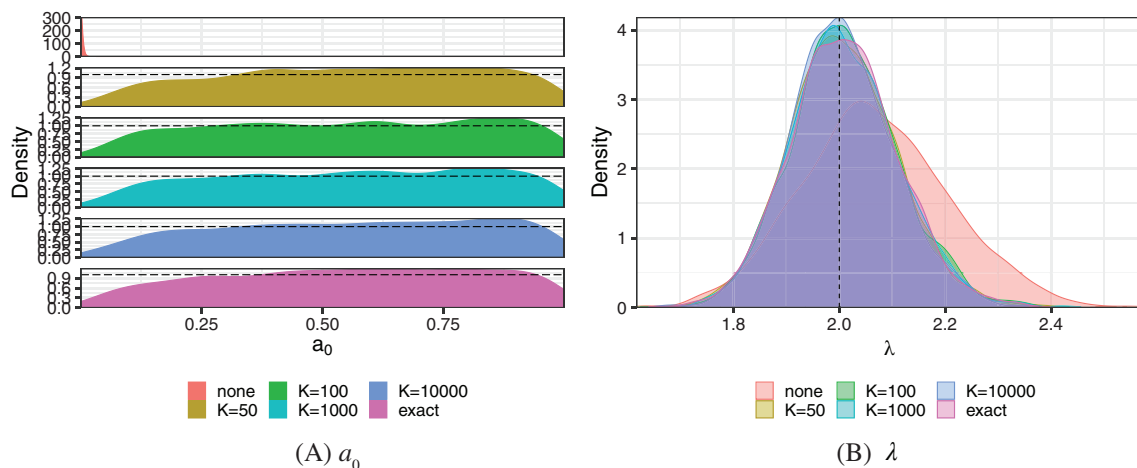

FIGURE B1 Results for the Poisson example. Panels (and colors) correspond to various values of the grid size, K, as well as the results with no normalization. In Panel A, we show the marginal posterior for a_0, using horizontal dashed lines to show the prior density of a Beta($\eta = 1, \nu = 1$). In Panel B, we show the resulting posterior distributions for the rate parameter, λ. The vertical dashed line marks the true value of λ [Colour figure can be viewed at wileyonlinelibrary.com]

values of K, such as $K = 50$, the approximately normalized posteriors are very similar to the posterior obtained with exact normalization, both for a_0 and λ.

APPENDIX C. GAUSSIAN LIKELIHOOD WITH UNKNOWN MEAN AND VARIANCE

Now we move on to study a case where $c(a_0)$ is nonmonotonic and thus presents a more challenging setting. Suppose one has N_0 historical observations $y_{i0} \in \mathbb{R}$, $i = 1, \ldots, N_0$, which come from a Gaussian distribution with parameters μ and τ. Here we will choose a normal-Gamma conjugate model:

$$\tau \sim \text{Gamma}(\alpha_0, \beta_0),$$
$$\mu \sim \text{Normal}(\mu_0, \kappa_0 \tau),$$
$$y_{0i} | \mu, \tau \sim \text{Normal}(\mu, \tau),$$

where the normal distribution is parameterized in terms of mean and precision (see below for a different parameterization). The posterior distribution is again a normal-Gamma distribution and the normalizing constant is

$$c(a_0) = \frac{\Gamma(\alpha_n)}{\Gamma(\alpha_0)} \frac{\beta_0^{\alpha_0}}{\beta_n^{\alpha_n}} \left(\frac{\kappa_0}{\kappa_n}\right)^2 (2\pi)^{-N_0 a_0/2},$$

$$\alpha_n = \alpha_0 + \frac{1}{2} a_0 N_0,$$

$$\kappa_n = \kappa_0 + a_0 N_0,$$

$$\beta_n = \beta_0 + \frac{1}{2} \left(a_0 \sum_{i=1}^{N_0} (y_{0i} - \bar{y})^2 + \left(\kappa_0 a_0 N_0 (\bar{y} - \mu_0)^2\right) / \kappa_n\right), \tag{C1}$$

with $\bar{y} = N_0^{-1} \sum_{i=1}^{N_0} y_{0i}$. In Section C.1, we give a closed-form expression for $c'(a_0)$ and characterize the point of inflection of $c(a_0)$ by giving the conditions for $c'(a_0) = 0$.

To make the discussion concrete, we generate $N_0 = 50$ data points from a Gaussian distribution with parameters $\mu = -0.1$ and $\tau = 10^6$. We construct the Gamma prior on τ with $\alpha_0 = \beta_0 = 1$ and assign a Gaussian prior on μ, with parameters $\mu_0 = 0$ and $\kappa_0 = 5$. This choice of hyperparameters leads to a function $c(a_0)$—Equation (C1)—that resembles a concave up parabola (Figure C1A). We then generate $N = 200$ new points from the same distribution to be used as current data. The points show the values of $l(a_0)$ and $l'(a_0)$ estimated using the algorithm described in Section 3.1.1 of the main text, which exploits the derivatives of $c(a_0)$ to place more points closer to the region where $c'(a_0)$ (and $l'(a_0)$) changes signs.

We show the resulting marginal posteriors for μ and τ as well as a_0 under no normalization, exact and approximate normalization with various K in Figure C1B,C. The first observation is that approximations with $K > 100$ seem to produce marginal posteriors for a_0 that resembles the exactly normalized distribution quite closely, even in this setting, where $c(a_0)$ is nonlinear.

In terms of parameter posteriors, we find that the posterior is not very sensitive to the value of a_0, as shown by the overlap between marginal posteriors with no normalization as well as exact and approximate normalization. Even in this setting the approximately normalized marginal posteriors match their exact counterparts closely.

We also evaluate the performance of adaptively building the grid of a_0 by comparing the mean absolute error (MAD) and root mean squared error (RMSE) of the estimated function $g_{\hat{\xi}}$ to the true (exact) normalization when using either a uniform grid or the adaptive grid. Over the whole range of $a_0 \in [0, 10]$, we found that the uniform grid leads to an estimated function with lower MAD (0.59 vs 0.82) and lower RMSE (0.99 vs 1.18). When considering only the range $a_0 \in [0, 1]$, the support of the prior—$\pi_A(a_0|\delta)$, we find the opposite: the adaptive grid outperforms uniform with MAD 2.10 vs 0.10 and RMSE 2.73 vs 0.13. This suggests that the adaptive scheme would produce better results in situations where the region where the derivative changes lies within the support of the prior.

C.1 The derivative of $c(a_0)$ for the normal case

In this section, we give more detail on the analysis of the Gaussian example of the previous section. Define

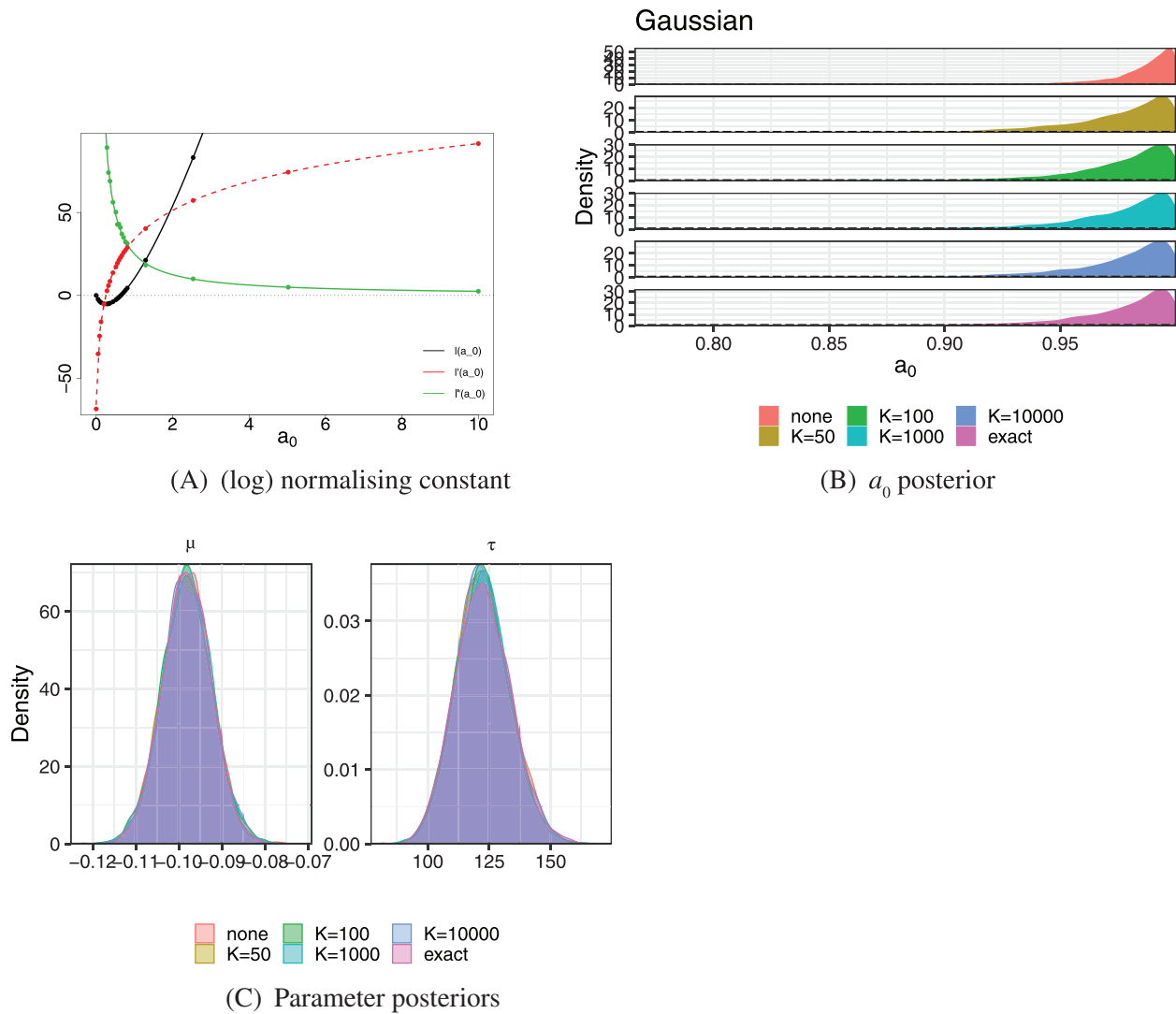$$c(a_0) = g(a_0)h(a_0)w(a_0)z(a_0),$$

(A) (log) normalising constant

(B) a_0 posterior

(C) Parameter posteriors

FIGURE C1   Results for the Gaussian example. In Panel A, we show $l(a_0) := \log(c(a_0))$ (black) and its first two derivatives (red and green, respectively). Points show the $J = 20$ estimates of $l(a_0)$, $l'(a_0)$, and $l''(a_0)$ obtained using the algorithm in Section 3.1.1. In Panel B, we show the marginal posterior of a_0 under no normalization, exact normalization, or approximate normalization with various grid sizes (K) in each subpanel. Horizontal dashed lines show the prior density of a Beta($\eta = 1, \nu = 1$). In Panel C, we show the marginal posteriors for μ and τ under no normalization, exact normalization, or approximate normalization with various grid sizes (K) in each subpanel [Colour figure can be viewed at wileyonlinelibrary.com]

$$g(a_0) := \frac{\Gamma\left(\alpha_0 + \frac{N_0}{2}a_0\right)}{\Gamma(\alpha_0)},$$

$$h(a_0) := \frac{\beta_0^{\alpha_0}}{(\beta_0 + \Delta a_0)^{\alpha_0 + \frac{N_0}{2}a_0}},$$

$$w(a_0) := \left(\frac{\kappa_0}{\kappa_0 + N_0 a_0}\right)^2,$$

$$z(a_0) := (2\pi)^{-N_0 a_0/2},$$

with $\Delta = \frac{1}{2}\left(\sum_{i=1}^{N_0}(y_{0i} - \bar{y})^2 + \frac{\kappa_0}{\kappa_n}N_0(\bar{y} - \mu_0)^2\right)$. Thus, dropping dependency on a_0 for notational compactness, we have

$$c' = hwzg' + gwzh' + ghzw' + ghwz'. \tag{C2}$$

Notice that only the first term of (C2) is positive. Since $g'(a_0) = \frac{N_0}{2}\psi_0\left(\alpha_0 + \frac{N_0}{2}a_0\right)g(a_0)$, we can write the following inequality:

$$c'(a_0) > 0 \Rightarrow \frac{N_0}{2}\psi_0\left(\alpha_0 + \frac{N_0}{2}a_0\right) > \frac{|h'(a_0)|}{h(a_0)} + \frac{|w'(a_0)|}{w(a_0)} + \frac{|z'(a_0)|}{z(a_0)}.$$

Since

$$\frac{|h'(a_0)|}{h(a_0)} = \frac{\Delta\left(\alpha_0 + \frac{N_0}{2}a_0\right)}{\Delta a_0 + \beta_0} + \frac{N_0}{2}\log(\Delta a_0 + \beta_0),$$

$$\frac{|w'(a_0)|}{w(a_0)} = \frac{2N_0}{a_0 N_0 + \kappa_0},$$

$$\frac{|z'(a_0)|}{z(a_0)} = \log(2\pi)\frac{N_0}{2},$$

we arrive at

$$\frac{N_0}{2}\psi_0\left(\alpha_0 + \frac{N_0}{2}a_0\right) > \frac{\Delta\left(\alpha_0 + \frac{N_0}{2}a_0\right)}{\Delta a_0 + \beta_0} + \frac{N_0}{2}\log(\Delta a_0 + \beta_0) + \frac{2N_0}{a_0 N_0 + \kappa_0} + \log(2\pi)\frac{N_0}{2},$$

$$\psi_0\left(\alpha_0 + \frac{N_0}{2}a_0\right) > \frac{\Delta(2\alpha_0 + N_0 a_0)}{N_0(\Delta a_0 + \beta_0)} + \log(\Delta a_0 + \beta_0) + \frac{4}{a_0 N_0 + \kappa_0} + \log(2\pi). \qquad (C3)$$

APPENDIX D. COMPARING APPROXIMATIONS OF THE LOG-NORMALIZING CONSTANT

In this section, we study two approaches to estimating $l(a_0)$, using four examples where it is known in closed-form. First, we consider the main approach discussed in this article, which consists of estimating $l(a_0)$ at a grid of $J = 15$ points of a_0 and using a GAM as the approximating function g_ξ to approximate $l(a_0)$ directly. We then evaluate the fitted function at a grid of $K = 20\,000$ values to form a vector $\boldsymbol{l}_{\text{direct}}$.

Another approach, adopted by Van Rosmalen et al[15]—VR2018 henceforth, is to use estimates of $l'(a_0)$ (see Equation A2) via the average log-likelihood. To achieve this, one starts at $a_0 = 0$ and computes an estimate of $l'(a_0)$, increments a_0 by Δ_a and then repeats this operation until $a_0 \geq 1$. An estimate of $l(a_0)$ at a point a_0 by using the cumulative sum up to that point. This approximate $l(a_0)$ function can then be interpolated at a fine grid of values for a_0 using linear interpolation.

We then compare the estimated values with the true values, $\boldsymbol{l}_{\text{true}}$, by computing the root mean squared error, $\hat{r} = \sqrt{\frac{1}{K}\sum_{i=1}^{K}\left(l_{\text{est}}^{(i)} - l_{\text{true}}^{(i)}\right)^2}$.

We show results in terms RMSE and running time for the Bernoulli, Poisson, Gaussian, and linear regression in Table D1. For these experiments, we used $J = 20$, $m = 0.05$ for our algorithm and $\Delta_a = 0.01$ for VR2018.

TABLE D1 Mean root squared error comparison of methods for approximating $l(a_0)$

Model	This article		VR2018	
	RMSE	Run time (seconds)	RSME	Run time (seconds)
Bernoulli	0.13	12.76	8.96	10.21
Poisson	0.07	21.22	4.61	40.56
Linear regression	0.03	23.14	1.69	52.77
Gaussian	0.021	18.10	0.31	18.04

Note: We used $J = 20$ points to construct $\boldsymbol{a}^{\text{est}}$ and use a GAM to approximate either $l(a_0)$ or $l'(a_0)$. In the latter case, we evaluate the fitted function on a fine grid ($K = 20\,000$ points) and obtain an approximation of $l(a_0)$ via midpoint integration (see text).

As expected, estimates (predictions) derived using direct estimation of $l(a_0)$ are substantially more accurate. The running times of the methods are also comparable, because while the computations for VR2018 are cheaper, one also needs to do more iterations—with $\Delta_a = 0.01$ there are 101 grid points to evaluate, in contrast to $J = 20$ for our method. All experiments were conducted on an Intel Core i7-8565U laptop with 16GB of RAM running Ubuntu 20.04, using R 4.0.4 and **rstan** version 2.21.2. Code to reproduce the analyses can be found at https://github.com/maxbiostat/propriety_power_priors.