On the propriety of power priors

July 27, 2019

1 Background

- Power priors are cool. [bunch of citations for propriety of GLM, Cox model, exponential family, etc];
- In this paper we explore the situation where the initial prior π is proper;
- We also explore ways of approximating the normalising constant $c(a_0)$ for models where it cannot be written in closed-form.

Definition The data $D_0 = \{d_{01}, \ldots, d_{0N_0}\} \in \mathcal{X} \subseteq \mathbb{R}^p$ are assumed independent and identically distributed and the parameter $\theta \in \Theta \subseteq \mathbb{R}^q$ and $a_0 \geq 0$.

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

2 Results

We want to show under which conditions

$$\int_\Theta L(D_0 \mid \theta)^{a_0} \pi(\theta) d\theta < \infty.$$

Theorem 1. *Assume* $\int_\mathcal{X} L(x \mid \theta) dx < \infty$. *Denote* $f_{a_0}(D_0; \theta) := L(D_0 \mid \theta)^{a_0} \pi(\theta)$. *Then for* $a_0 > 0$, $\int_\Theta f_{a_0}(D_0; \theta) d\theta < \infty$.

Proof. First, note that for $0 < a_0 \leq 1$, the function $g(x) = x^{a_0}$ is concave. Then, by Jensen's inequality and the finiteness of $L(D_0 \mid \theta)$ for all of its arguments we have

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

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

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

with $1 \leq p \leq 2$, again falling into the same case, since we know that $L(D_0 \mid \theta)^p \pi(\theta)$ is normalisable. We can then show that for any $n \geq 1 \in \mathbb{N}$, $\int_\Theta f_{a_0}(D_0; \theta) d\theta < \infty$ 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 $n + 1$ we have for $n \leq a_0 \leq n + 1$ and $n - 1 \leq p_n \leq n$:

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

because $0 \leq a_0 - p_n \leq 1$ and $L(D_0 \mid \theta)^{p_n} \pi(\theta)$ is proper by hypothesis. □

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

Proof. Recall that the power prior on multiple historical data sets is of the form [Eq. 2.9](Ibrahim et al., 2015):

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

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

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

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

2.1 Properties of the normalising constant $c(a_0)$

Remark 2. *The normalising constant is a convex function of a_0.*

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

$$c(a_0) := \int_\Theta L(D_0 \mid \theta)^{a_0} \pi(\theta) d\theta, \tag{2}$$

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

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

$$c''(a_0) = \int_\Theta L(D_0 \mid \theta)^{a_0} \pi(\theta) [\log L(D_0 \mid \theta)]^2 d\theta, \tag{4}$$

and the integrals always exist (see Proposition 1 in Appendix A). From this we conclude that c is (strictly) convex and c' is monotonic, because c'' is always positive. □

For the goals of this paper, it would be useful to know more about the shape of $c(a_0)$, more specifically if and when its derivatives change signs. Let us first study an upper bound for $c'(a_0)$. By a straight-forward application of the Cauchy-Schwarz inequality, we also know that $c''(a_0) \geq c'(a_0)$. Now, to derive a lower bound on $c'(a_0)$, consider the inequality $1 - 1/x \leq \log(x)$, $x > 0$, which suggests

$$c'(a_0) \geq c(a_0) - \int_\Theta L(D_0 \mid \theta)^{a_0 - 1} \pi(\theta) d\theta.$$

2.1.1 Exponential family

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

$$L(D_0 \mid \theta) = \boldsymbol{h}(D_0) \exp\left(\eta(\theta)^T \left(\sum_{i=1}^{N_0} T(d_{0i})\right) - N_0 A(\theta)\right), \tag{5}$$

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

$$c(a_0) = \int_{\boldsymbol{\Theta}} \left[\boldsymbol{h}(D_0) \exp\left(\eta(\theta)^T \left(\sum_{i=1}^{N_0} T(d_{0i})\right) - N_0 A(\theta)\right)\right]^{a_0} \pi(\theta) d\theta, \tag{6}$$

$$= \boldsymbol{h}(D_0)^{a_0} \int_{\boldsymbol{\Theta}} \exp\left(\eta(\theta)^T \left(a_0 \sum_{i=1}^{N_0} T(d_{0i})\right)\right) \exp\left(-a_0 N_0 A(\theta)\right) \pi(\theta) d\theta. \tag{7}$$

The derivative evaluates to

$$c'(a_0) = \log(\boldsymbol{h}(D_0)) + \int_{\boldsymbol{\Theta}} \eta(\theta)^T \left(a_0 \sum_{i=1}^{N_0} T(d_{0i}) \right) f_{a_0}(D_0;\theta) d\theta - a_0 N_0 \int_{\boldsymbol{\Theta}} f_{a_0}(D_0;\theta) A(\theta) d\theta, \quad (8)$$

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

If we choose $\pi(\theta)$ to be conjugate to $L(D_0 \mid \theta)$ (Diaconis and Ylvisaker, 1979), i.e.

$$\pi(\theta \mid \tau, n_0) = H(\tau, n_0) \exp\{\tau^T \eta(\theta) - n_0 A(\theta)\}, \quad (9)$$

we have

$$c(a_0) = \boldsymbol{h}(D_0)^{a_0} H(\tau, n_0) \int_{\boldsymbol{\Theta}} \exp\left[\eta(\theta)^T \left(\tau + a_0 \sum_{i=1}^{N_0} T(d_{0i}) \right) - (n_0 + a_0 N_0) A(\theta) \right] d\theta \quad (10)$$

2.2 Examples

2.2.1 Gaussian

Suppose one has N_0 historical observations $y_{i0} \in \mathbb{R}, i = 1, \ldots, N_0$. Here we will choose a Normal-gamma conjugate model:

$$\begin{aligned} \tau &\sim \text{Gamma}(\alpha_0, \beta_0), \\ \mu &\sim \text{Normal}(\mu_0, \kappa_0 \tau), \\ y_{0i} \mid \mu, \tau &\sim \text{Normal}(\mu, \tau), \end{aligned}$$

where the normal distribution is parametrised in terms of mean and precision (see below for a different parametrisation). The posterior distribution is again a normal-gamma distribution and the normalising constant is

$$\begin{aligned} c(a_0) &= \frac{\Gamma(\alpha_n)}{\Gamma(\alpha_0)} \frac{\beta_0^{\alpha_0}}{\beta_n^{\alpha_n}} \left(\frac{\kappa_0}{\kappa_n} \right)^2 (2\pi)^{-N_0 a_0/2}, \quad (11) \\ \alpha_n &= \alpha_0 + \frac{1}{2} a_0 N_0 \\ \kappa_n &= \kappa_0 + a_0 N_0 \\ \beta_n &= \beta_0 + \frac{1}{2} \left(a_0 \sum_{i=1}^{N_0} (y_{0i} - \bar{y})^2 + \left(\kappa_0 a_0 N_0 (\bar{y} - \mu_0)^2 \right) / \kappa_n \right), \end{aligned}$$

with $\bar{y} = \sum_{i=1}^{N_0} y_{0i}/N_0$.

2.2.2 Linear regression with a normal inverse-Gamma prior

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 parametrisation, which naturally leads to a normal inverse-gamma conjugate prior. The model is

$$\begin{aligned} \sigma^2 &\sim \text{Inverse-gamma}(\alpha_0, \beta_0), \\ \epsilon_i \mid \sigma^2 &\sim \text{Normal}(0, \sigma^2), \\ \beta \mid \sigma^2 &\sim \text{Normal}(\boldsymbol{\mu}_0, \sigma^2 \boldsymbol{\Lambda}_0^{-1}), \\ y_{0i} &= \boldsymbol{X}_{0i}^\top \boldsymbol{\beta} + \epsilon_i, \end{aligned}$$

where $\boldsymbol{\beta}$ is a $1 \times P$ vector of coefficients. The posterior is again a normal inverse gamma and thus

$$
\begin{aligned}
c(a_0) &= \sqrt{\frac{|\boldsymbol{\Lambda}_n|}{|\boldsymbol{\Lambda}_0^{-1}|}} \frac{\beta_0^{\alpha_0}}{\beta_n^{\alpha_n}} \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_n)} (2\pi)^{-N_0 a_0/2}, \\
\boldsymbol{\Lambda}_n &= \boldsymbol{X}_\star^\top \boldsymbol{X}_\star + \boldsymbol{\Lambda}_0^{-1}, \\
\boldsymbol{\mu}_n &= \boldsymbol{\Lambda}_n^{-1} \left(\boldsymbol{\Lambda}_0^{-1} \boldsymbol{\mu}_0 + \boldsymbol{X}_\star^\top \boldsymbol{y}_\star \right) \\
\alpha_n &= \alpha_0 + \frac{1}{2} a_0 N_0, \\
\beta_n &= \beta_0 + \frac{1}{2} \left(\boldsymbol{y}_\star^\top \boldsymbol{y}_\star + \boldsymbol{\mu}_0^\top \boldsymbol{\Lambda}_0^{-1} \boldsymbol{\mu}_0 - \boldsymbol{\mu}_n^\top \boldsymbol{\Lambda}_n \boldsymbol{\mu}_n \right),
\end{aligned}
\tag{12}
$$

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.

2.2.3 Bernoulli

The historical 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

$$
\begin{aligned}
\theta &\sim \text{Beta}(c, d), \\
x_{0i} \mid \theta &\sim \text{Bernoulli}(\theta).
\end{aligned}
$$

This leads to a Beta posterior distribution for θ,

$$
p(\theta | N_0, y_0, a_0) \propto \theta^{a_0 y_0 + c - 1} (1 - \theta)^{a_0 (N_0 - y_0) + d - 1}, \tag{13}
$$

and hence (Neuenschwander et al., 2009):

$$
c(a_0) = \frac{\mathcal{B}(a_0 y_0 + c, a_0 (N_0 - y_0) + d)}{\mathcal{B}(c, d)}, \tag{14}
$$

where $\mathcal{B}(w, z) = \frac{\Gamma(w)\Gamma(z)}{\Gamma(w+z)}$.

2.2.4 Gamma

The historical data consist of N_0 observations $\boldsymbol{y}_0 = \{y_{01}, \ldots, y_{0N_0}\}$, $y_{0i} > 0$. And the model is

$$
\begin{aligned}
\alpha &\sim \text{Gamma}(\eta_\alpha, \nu_\alpha), \\
\beta &\sim \text{Gamma}(\eta_\beta, \nu_\beta), \\
y_{0i} \mid \alpha, \beta &\sim \text{Gamma}(\alpha, \beta).
\end{aligned}
$$

The posterior distribution is

$$
p(\alpha, \beta | \boldsymbol{y}_0, a_0) \propto \frac{\beta^{a_0 N_0 \alpha}}{\Gamma(\alpha)^{a_0 N_0}} \boldsymbol{p}^{a_0(\alpha - 1)} \exp(-a_0 \boldsymbol{s} \beta) \times \alpha^{\eta_\alpha - 1} \exp(-\nu_\alpha \alpha) \times \beta^{\eta_\beta - 1} \exp(-\nu_\beta \beta), \tag{15}
$$

where $\boldsymbol{p} := \prod_{i=1}^{N_0} y_{0i}$ and $\boldsymbol{s} := \sum_{i=1}^{N_0} y_{0i}$. Noticing that the full conditional distribution of β is (the kernel of) a Gamma distribution leads to

$$
c(a_0) = \frac{\nu_\alpha^{\eta_\alpha} \nu_\beta^{\eta_\beta}}{\Gamma(\eta_\alpha)\Gamma(\eta_\beta)} \frac{1}{\boldsymbol{p}^{a_0}} \times \int_0^\infty \frac{\Gamma(a_0 N_0 \alpha + \eta_\beta)}{(a_0 \boldsymbol{s} + \nu_\beta)^{a_0 N_0 \alpha + \eta_\beta}} \frac{\alpha^{\eta_\alpha - 1} \exp(-(\nu_\alpha - a_0 \log \boldsymbol{p})\alpha)}{\Gamma(\alpha)^{a_0 N_0}} d\alpha. \tag{16}
$$

Whilst the expression in (16) cannot be obtained in closed-form, it can be approximated *via* quadrature.

2.2.5 Inverse Gaussian

The historical data consist of N_0 observations $\boldsymbol{y}_0 = \{y_{01}, \ldots, y_{0N_0}\}$, $y_{0i} > 0$. The model is

$$\begin{aligned} \mu \sim&\ \text{Gamma}(\alpha_\mu, \beta_\mu), \\ \lambda \sim&\ \text{Gamma}(\alpha_\lambda, \beta_\lambda), \\ y_{0i} \mid \mu, \lambda \sim&\ \text{Inverse-Gaussian}(\mu, \lambda). \end{aligned}$$

The posterior distribution is

$$\begin{aligned} p(\mu, \lambda | \boldsymbol{y}_0, a_0) \propto \left(\frac{\lambda}{2\pi}\right)^{\frac{1}{2}a_0 N_0} \boldsymbol{p}^{-\frac{3}{2}a_0} \exp\left(-a_0 \left[\frac{\boldsymbol{s}}{2\mu^2} - \frac{N_0}{\mu} + \frac{\boldsymbol{s}'}{2}\right]\lambda\right) \times \\ \mu^{\alpha_\mu - 1} \exp(-\beta_\mu \mu) \lambda^{\alpha_\lambda - 1} \exp(-\beta_\lambda \lambda), \end{aligned} \tag{17}$$

where $\boldsymbol{p} := \prod_{i=1}^{N_0} y_{0i}$, $\boldsymbol{s} := \sum_{i=1}^{N_0} y_{0i}$ and $\boldsymbol{s}' := \sum_{i=1}^{N_0} \frac{1}{y_{0i}}$. Similar to the Gamma case, one can marginalise over λ by noticing its full conditional distribution is a Gamma distribution (see Appendix):

$$\begin{aligned} c(a_0) = \frac{\beta_\mu^{\alpha_\mu} \beta_\lambda^{\alpha_\lambda}}{\Gamma(\alpha_\mu)\Gamma(\alpha_\lambda)} \times (2\pi)^{-\frac{1}{2}a_0 N_0} \boldsymbol{p}^{-\frac{3}{2}a_0} \Gamma\left(\frac{a_0 N_0 + 2\alpha_\lambda}{2}\right) \times \\ \int_0^\infty \left(\frac{a_0 \boldsymbol{s}}{2\mu^2} - \frac{a_0 N_0}{\mu} + \frac{a_0 \boldsymbol{s}'}{2} + \beta_\lambda\right)^{-\frac{a_0 N_0 + 2\alpha_\lambda}{2}} \mu^{\alpha_\mu - 1} \exp(-\beta_\mu \mu) d\mu. \end{aligned} \tag{18}$$

Again, while the integral in (18) is not known in closed-form, it can be approximated *via* quadrature.

2.2.6 Poisson

Suppose the historical data consists of N_0 observations $y_{0i} \in \{0, 1, \ldots\}$. The model is

$$\begin{aligned} \lambda \sim&\ \text{Gamma}(\alpha_0, \beta_0), \\ y_{0i} \mid \lambda \sim&\ \text{Poisson}(\lambda). \end{aligned}$$

The posterior distribution is

$$p(\lambda \mid \boldsymbol{y}_0) \propto \frac{1}{\boldsymbol{p'}^{a_0}} \lambda^{a_0 \boldsymbol{s}} \exp(-a_0 N_0 \lambda) \times \lambda^{\alpha_0 - 1} \exp(-\beta_0 \lambda), \tag{19}$$

where $\boldsymbol{s} := \sum_{i=0}^{N_0} y_{0i}$ and $\boldsymbol{p}' := \prod_{i=0}^{N_0} y_{0i}!$, leading to

$$c(a_0) = \frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)} \frac{1}{\boldsymbol{p'}^{a_0}} \frac{\Gamma(a_0 \boldsymbol{s} + \alpha_0)}{(a_0 N_0 + \beta_0)^{a_0 \boldsymbol{s} + \alpha_0}}. \tag{20}$$

2.3 Efficiently computing $c(a_0)$

For many models, $c(a_0)$ is not known in closed form, and hence must be approximated if we wish to compute it. An example where this would be important is when we need to normalise the power prior for use within a Markov chain Monte Carlo (MCMC) procedure.

Here we take the following approach to approximating $c(a_0)$: first, define a grid of values $\boldsymbol{a}^{\text{est}} = \{a_1^{\text{est}}, \ldots, a_J^{\text{est}}\}$ for a typically modest number 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 $g_{\boldsymbol{\xi}} : \mathbb{R}^+ \to \mathbb{R}^+$, indexed by a set of parameters $\boldsymbol{\xi}$. For instance, $g_{\boldsymbol{\xi}}$ could be a linear model, a generalised additive model (GAM) or a Gaussian

process. We can then use $\boldsymbol{a}^{\text{est}}$ and $\hat{\boldsymbol{c}}$ 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 $K \gg J$, $\boldsymbol{a}^{\text{pred}} = \{a_1^{\text{pred}}, \ldots, a_K^{\text{pred}}\}$, 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 key insight is that $c(a_0)$ is convex, so we need only to make sure our grid covers the region where its derivative changes signs when designing both $\boldsymbol{a}^{\text{est}}$ and $\boldsymbol{a}^{\text{pred}}$.

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. For many applications, interest usually lies with $a_0 \in [0, 1]$, and in applications where 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 $\pi(a_0)$. 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 a high chance of not sampling any values outside the grid.

Effect of discretising the approximating function. In order to evaluate the effect of discretisation of $g_{\boldsymbol{\xi}}$ on inferences, we evaluated the effect of varying the prediction grid size K. For three examples where we know the truth, namely the Gaussian and Bernoulli distributions and a linear regression with a normal inverse-Gamma prior, we compared estimates of parameters – including a_0 – and marginal likelihoods obtained by using the true $c(a_0)$ and the discretised $g_{\boldsymbol{\xi}}$ with $K = 50, 100, 1000,$ and $10,000$. For prediction, we use a simple, linear grid, with end points $a_1^{\text{pred}} = 0$ and $a_K^{\text{pred}} = 10$. For all experiments we fixed the estimation grid size at $J = 10$ with the same endpoints as the prediction grid $a_1^{\text{est}} = 0$, $a_K^{\text{est}} = 10$, but placing most $(J - 3 = 7)$ points on the $[0, 1]$ interval.

Marginal likelihood estimation. We employed bridge sampling to compute marginal likelihoods implemented in the R package **bridgesampling** (Gronau et al., 2017). We used HMC as before, with four independent chains of 2000 iterations each, removing half of each chain as burn-in/warm-up, resulting in 4000 samples being kept.

TODO:

- check proofs;

3 Questions

1. **The shape of** $c(a_0)$: From both a theoretical and practical point of view, it'd be nice to know when $c(a_0)$ is non-monotonic like for the Gaussian case [and a very special choice of hyperparameters] and when it's just a straight line, as for most models I've tested. Appendix B contains some rudimentary analysis for the Gaussian case...

2. **Exponential family**: it could be potentially fruitful to explore the results in Eq. 7 a bit more and maybe help answer the question above; under which conditions will a likelihood in the exponential family lead to a monotonic (and hence linear since we know it's definitely convex) $c(a_0)$?

3. **Sensitivity to** a_0: (i) is this important? (ii) how to define sensitivity? one thing to look at is variance reduction as $a_0 \to \infty$... At any rate, we know empirically that for some models (e.g. logistic regression) including the normalising constant is of utmost importance, whereas for others (e.g. Bernoulli), it doesn't seem to matter much. **Important:** When I say "matters" (or not), I'm talking about the effect on estimates of actual parameters, like regression coefficients or success probability. Including $c(a_0)$ always makes a HUGE difference for "estimates" of a_0, obviously.

4. **Situations with** $a_0 > 1$ Should we pursue this? It seems to me we should at least consider it, since we show that the power prior is proper for $a_0 > 1$ also.

Acknowledgements

LMC would like to thank Leo Bastos (PROCC/Fiocruz) for helpful discussions and Dr. Beat Neuenschwander for clarifications regarding his paper. This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) Finance Code 001. LMC is supported by a postodoctoral fellowship from CAPES.

References

Diaconis, P. and Ylvisaker, D. (1979). Conjugate priors for exponential families. *The Annals of statistics*, pages 269–281.

Gronau, Q. F., Singmann, H., and Wagenmakers, E.-J. (2017). Bridgesampling: An r package for estimating normalizing constants. *arXiv preprint arXiv:1710.08162*.

Ibrahim, J. G., Chen, M.-H., Gwon, Y., and Chen, F. (2015). The power prior: theory and applications. *Statistics in medicine*, 34(28):3724–3749.

Neuenschwander, B., Branson, M., and Spiegelhalter, D. J. (2009). A note on the power prior. *Statistics in Medicine*, 28(28):3562–3566.

Piegorsch, W. W. and Casella, G. (1985). The existence of the first negative moment. *The American Statistician*, 39(1):60–62.

A Additional results

Proposition 1. *The first and second derivatives of* $c(a_0)$ *always exist, i.e.,* $c \in \mathcal{C}^2$.

Proof. Here we will assume that $L(D_0 \mid \theta) > 0 \, \forall \theta \in \Theta$. Let $f(\theta) = L(D_0 \mid \theta)^{a_0}\pi(\theta)$. To see that

$$c'(a_0) = \int_\Theta L(D_0 \mid \theta)^{a_0}\pi(\theta) \log L(D_0 \mid \theta) d\theta < \infty,$$

make $X = L(D_0 \mid \theta)$ and apply Jensen's inequality to get $E_f[\log(X)] \leq \log\left(E_f[X]\right)$, then apply Theorem 1. Now, to show that

$$c''(a_0) = \int_\Theta L(D_0 \mid \theta)^{a_0}\pi(\theta)[\log L(D_0 \mid \theta)]^2 d\theta < \infty,$$

we will first assume that $L(D_0 \mid \theta) < 1 \, \forall \theta \in \Theta$, because if $L(D_0 \mid \theta) > 1$, the integral is bounded above by $c(a_0 + 2)$ and hence is clearly finite. For $0 < x \leq 1$, the usual bounds on $\log(x)$ can be manipulated to get

$$(x-1)\log(x) \leq \log(x)\log(x) \leq -\frac{\log(x)}{x}$$

Hence all that remains is to show that $-\infty < E_f[\log(X)/X]$. Again, we can manipulate the bounds on $\log(x)$ to get $1/x - 1/x^2 \leq \log(x)/x$, so we need to show that $E[1/X^2] < \infty$. This holds for $a_0 \geq 2$, but we shall try to show the result in general. Define the random variable $Y = [L(D_0 \mid \theta)]^2$, let G be the push-forward measure on Y induced by Π_θ, and let $g(y)$ be its density. Then we can write

$$f_Y(y) = L(D_0 \mid \theta)^{a_0}\pi(\theta) = y^{a_0/2}g(y).$$

Notice that asking that $E[1/X^2] < \infty$ is equivalent to asking that $E[Y^{-1}] < \infty$. Corollary 2.1 of Piegorsch and Casella (1985) states that if Y is a positive random variable and $f'_Y(0)$ exists and is finite, then $E[Y^{-1}] < \infty$. Since

$$f'_Y(y) = y^{(a_0-2)/2}\left(\frac{a_0}{2}g(y) + yg'(y)\right),$$

we have $f'_Y(0) = 0$ and hence we conclude $E[1/X^2]$ is finite, completing the proof. □

B The derivative of $c(a_0)$ for the normal case

Define

$$\begin{aligned}
c(a_0) &= g(a_0)h(a_0)w(a_0)z(a_0),\\
g(a_0) &:= \frac{\Gamma\left(\alpha_0 + \frac{N_0}{2}a_0\right)}{\Gamma(\alpha_0)},\\
h(a_0) &:= \frac{\beta_0^{\alpha_0}}{(\beta_0 + \Delta a_0)^{\alpha_0 + \frac{N_0}{2}a_0}},\\
w(a_0) &:= \left(\frac{\kappa_0}{\kappa_0 + N_0 a_0}\right)^2,\\
z(a_0) &:= (2\pi)^{-N_0 a_0/2},
\end{aligned}$$

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,

$$c' = hwzg' + gwzh' + ghzw' + ghwz'. \tag{21}$$

Notice that only the first term of (21) 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:

$$c'(a_0) > 0 \implies \frac{N_0}{2}\psi_0\left(\alpha_0 + \frac{N_0}{2}a_0\right) > \frac{|h'(a_0)|}{h(a_0)} + \frac{|w'(a_0)|}{w(a_0)} + \frac{|z'(a_0)|}{z(a_0)}. \tag{22}$$

Since

$$\frac{|h'(a_0)|}{h(a_0)} = \frac{\Delta\left(\alpha_0 + \frac{N_0}{2}a_0\right)}{\Delta a_0 + \beta_0} + \frac{N_0}{2}\log\left(\Delta a_0 + \beta_0\right), \tag{23}$$

$$\frac{|w'(a_0)|}{w(a_0)} = \frac{2N_0}{a_0 N_0 + \kappa_0}, \tag{24}$$

$$\frac{|z'(a_0)|}{z(a_0)} = \log(2\pi)\frac{N_0}{2}, \tag{25}$$

we arrive at

$$\frac{N_0}{2}\psi_0\left(\alpha_0 + \frac{N_0}{2}a_0\right) > \frac{\Delta\left(\alpha_0 + \frac{N_0}{2}a_0\right)}{\Delta a_0 + \beta_0} + \frac{N_0}{2}\log\left(\Delta a_0 + \beta_0\right) + \frac{2N_0}{a_0 N_0 + \kappa_0} + \log(2\pi)\frac{N_0}{2}, \tag{26}$$

$$\psi_0\left(\alpha_0 + \frac{N_0}{2}a_0\right) > \frac{\Delta\left(2\alpha_0 + N_0 a_0\right)}{N_0\left(\Delta a_0 + \beta_0\right)} + \log\left(\Delta a_0 + \beta_0\right) + \frac{4}{a_0 N_0 + \kappa_0} + \log(2\pi). \tag{27}$$

C Full conditional of λ for an Inverse Gaussian likelihood

We need to show that the full conditional of λ will **always** be a Gamma distribution. Recalling that $\mu > 0$, we need to prove that

$$\left(-\frac{a_0 N_0}{\mu} + \frac{a_0 \boldsymbol{s}}{2\mu^2} + \frac{a_0 \boldsymbol{s}'}{2} + \beta_\lambda\right) = \frac{-2a_0 N_0 \mu + a_0 \boldsymbol{s} + (a_0 \boldsymbol{s}' + 2\beta_\lambda)\mu^2}{2\mu^2} \geq 0,$$

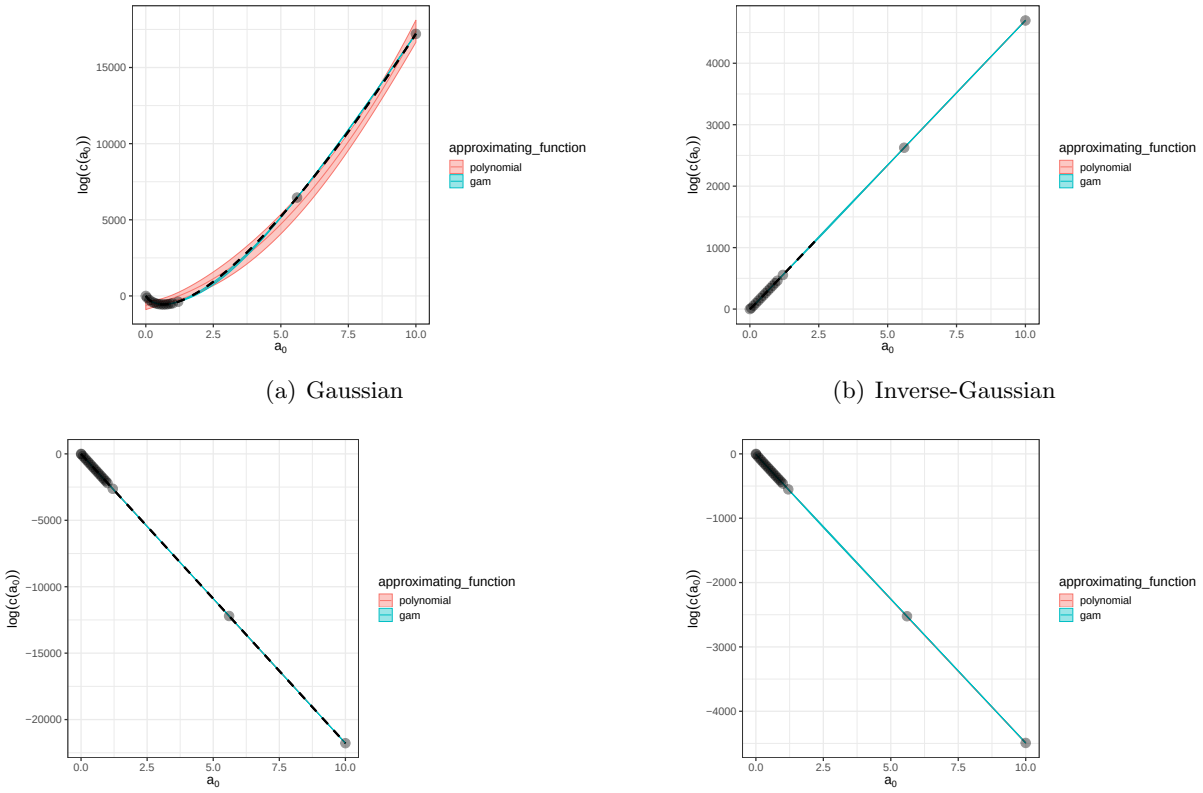
whence, equivalently,

$$\boldsymbol{s}'\mu^2 - 2N_0\mu + \boldsymbol{s} \geq 0.$$

Since $\boldsymbol{s}' > 0$, this happens whenever

$$4N_0^2 - 4\boldsymbol{s}'\boldsymbol{s} \leq 0,$$

which is always true, by direct application of Titu's lemma.



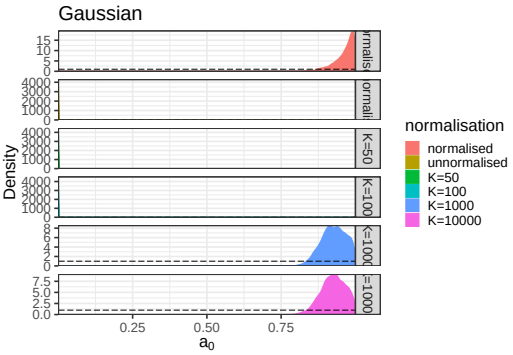
(a) Gaussian

(b) Inverse-Gaussian

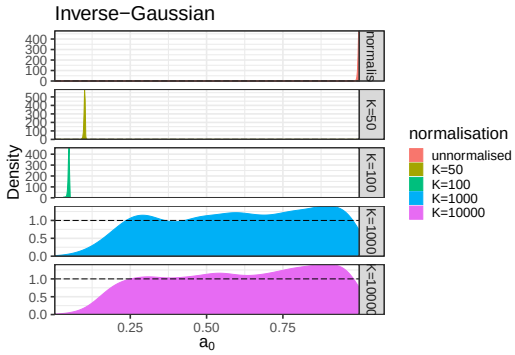
(c) Linear regression

(d) Logistic regression

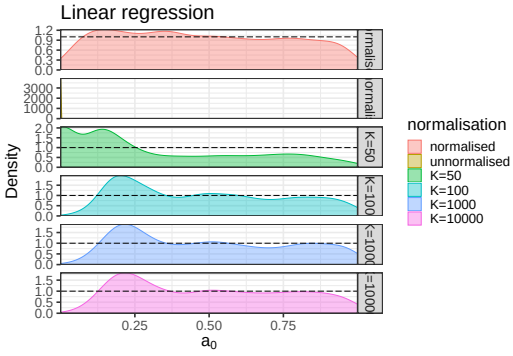
Figure 1: **The normalising constant $c(a_0)$ for several models**. The plots show $\log c(a_0)$ for a host of models. Notice how for the inverse-Gaussian model $\log c(a_0)$ is **increasing** in a_0. Dashed line shows the true $c(a_0)$, when it exists or can be computed.



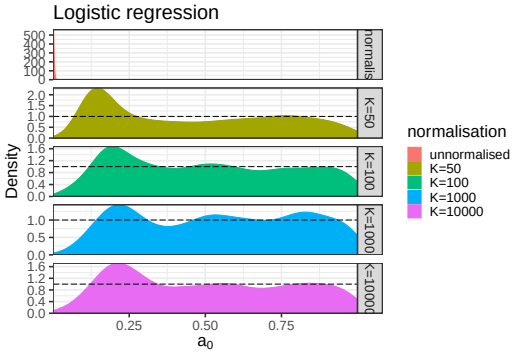
(a) Gaussian



(b) Inverse-Gaussian

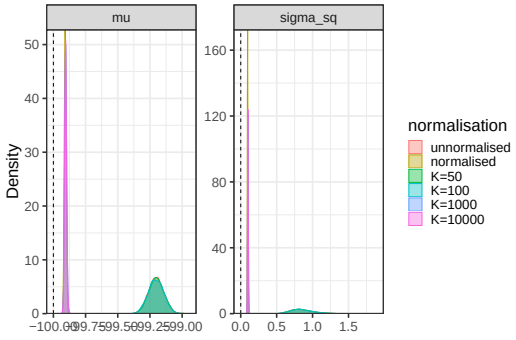


(c) Linear regression

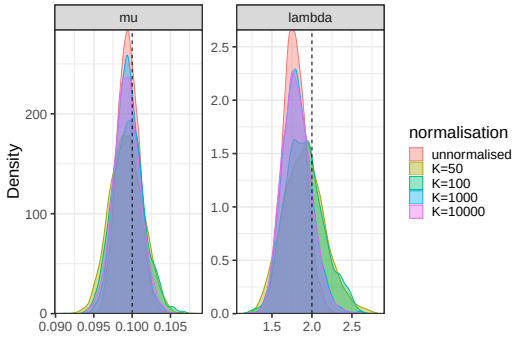


(d) Logistic regression

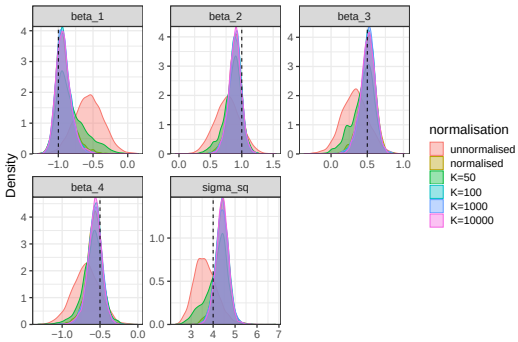
Figure 2: **The posterior distribution of a_0 under normalised and unnormalised models**. The panels (and colours) correspond to the posterior distribution of the parameter a_0 when $c(a_0)$ is accounted for and when it is not included. Horizontal dashed lines show the prior density of a Beta($\nu = 1, \eta = 1$).



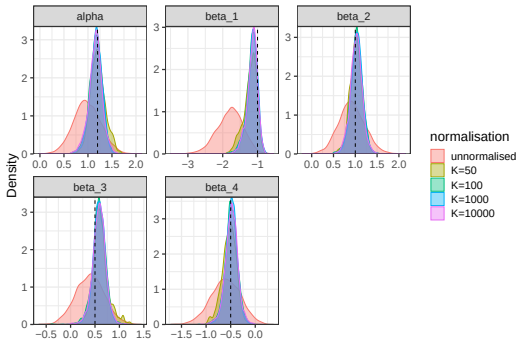
(a) Gaussian



(b) Inverse-Gaussian



(c) Linear regression



(d) Logistic regression

Figure 3: **The posterior distribution of parameters of interest under normalised and unnormalised models**. The panels (and colours) correspond to the posterior distribution of the parameter a_0 when $c(a_0)$ is accounted for and when it is not included. Vertical lines mark the data-generating parameter values.