A multidimensional objective prior based on scoring rules

Isadora Antoniano-Villalobos

Ca' Foscari University of Venice, Italy

Joint work with:

Cristiano Villa, Newcastle University, UK
Stephen G. Walker, University of Texas at Austin, USA

The starting point is a **parameter of interest**, say $\theta \in \Theta \subset \mathbb{R}^d$, indexing a family of probability distributions $f(x|\theta)$.

The **Bayesian framework** requires the specification of a prior $q(\theta)$ supported on Θ.

In general, there are two options:

- Elicit the prior on the basis of prior information
- Use an **objective prior**, in the absence of information

Common objective approaches for the definition of priors are:

- Jeffreys prior
- Reference prior

Both of these depend on $f(x|\theta)$

Common objective prior approaches have known **drawbacks and limitations**:

- While they tend to be proper for bounded parameter spaces, e.g. $\Theta = (0, 1)$, they are often **improper** for $\Theta = (0, \infty)$ and $\Theta = (-\infty, \infty)$.
- For large or complex models, it is difficult to check posterior properness.
- Even for not-so-large models, prior independence is often assumed, to avoid issues when defining **multivariate objective priors**

OUR AIM: Finding objective priors for multiple parameters which are proper, heavy tailed and do not require an independence assumption

Objectivity and scoring rules

Fabrizio Leisen, Cristiano Villa, Stephen G. Walker (2020). On a class of objective priors from scoring rules (with discussion). *Bayesian Analysis* **15**, 1345–1523

Idea: For $\Theta \subset \mathbb{R}$, define the prior as the solution to the differential equation

$$S(q, q', q'') = 0,$$

where

- S is a **scoring rule** defined as a a weighted sum of the **log-score** and the **Hyvärinen score**
- q is the density of a possible prior for θ with q' and q'' the first two derivatives

The resulting prior has some interesting properties:

- It depends on Θ but not on $f(x|\theta)$
- By design, it is convex, proper, decreasing (and other desirable features)
- It **minimizes** a particular information criterion

Divergences, **Informations** and **Scores** are connected:

$$D(p, q) = I(p) + \int p\, S(q)$$

For example:

- Kullback-Leibler divergence, Shannon entropy and log-score

$$\int p \log(p/q) = p \log p + \int p\,(-\log q)$$

- Fisher Information divergence, Fisher information and Hyvärinen score

$$\int p(p'/p - q'/q)^2 = \int (p'/p)^2 + \int p\,[2q''/q - (q'/q)^2]$$

Proper scoring rules from Bregman divergences

Matthew Parry, M., A. Philip Dawid, Steffen Lauritzen (2012). Proper local scoring rules. *Annals of Statistics* 40, 561–592

Idea: Exploit the relation between D, I and S to define new scoring rules by choosing different divergences. In particular, considering the family of **Bregman divergences**:

$$D(p,q) = \int B_\phi(p,q); \quad B_\phi(p,q) = \phi(p) - \phi(q) - \phi_q(q)(p-q)$$

for a convex function $\phi : \mathbb{R}_+ \to \mathbb{R}$, where $\phi_q(q)$ denotes the derivative $\frac{d\phi(q)}{dq}$

For example:

- If $\phi(u) = u \log u$, B_ϕ is the Kullback-Leibler divergence
- If $\phi(u) = u^2$, B_ϕ is the L_2 norm

Objective priors from scoring rules derived from 2-dimensional Bregman divergences

Stephen G. Walker, Cristiano Villa (2021). An Objective Prior from a Scoring Rule. *Entropy* 23, 833.

Idea: Consider a **2-dimensional Bregman divergence:**

$$D(\mathbf{p},\mathbf{q}) = \int B_\phi(\mathbf{p},\mathbf{q}); \quad B_\phi(\mathbf{p},\mathbf{q}) = \phi(\mathbf{p}) - \phi(\mathbf{q}) - \phi_q(\mathbf{q})(p-q) - \phi_{q_\theta}(\mathbf{q})(p_\theta - q_\theta)$$

for a convex function $\phi : \mathbb{R}_+^2 \to \mathbb{R}$, where $\mathbf{p} = (p, p_\theta)$, $\mathbf{p} = (q, q_\theta)$,

$$q_\theta(\theta) = \frac{dq(\theta)}{d\theta}; \quad \phi_q(\mathbf{q}) = \frac{\partial \phi(\mathbf{q})}{\partial q}; \quad \phi_{q_\theta}(\mathbf{q}) = \frac{\partial \phi(\mathbf{q})}{\partial q_\theta}$$

For example:

- If $\phi(u,v) = v^2/u$, B_ϕ is the Fisher information divergence

➔ In general, consider, $\phi(u,v) = u\,\alpha(v/u)$, which is convex whenever $\alpha : \mathbb{R} \to \mathbb{R}$ is convex

After some manipulation and assuming boundary conditions at the integration limits, the relation between D, I and S can be recovered:

$$\int B_\phi(\mathbf{p}, \mathbf{q}) = \int p\,\alpha(p_\theta/p) + \int p\left[\frac{d}{dx}\alpha_u(q_\theta/q) - \alpha(q_\theta/q) + (q_\theta/q)\alpha_u(q_\theta/q)\right],$$

where α_u denotes the derivative $\frac{d\alpha(u)}{du}$

The score $S(\mathbf{q}) = S(q, q_\theta, q_{\theta\theta})$

$$\begin{aligned} S(\mathbf{q}) &= \frac{d}{d\theta}\alpha_u(q_\theta/q) - \alpha(q_\theta/q) + (q_\theta/q)\alpha_u(q_\theta/q) \\ &= \alpha_u u(q_\theta/q)\frac{q q_{\theta\theta} - q_\theta^2}{q^2} - \alpha(q_\theta/q) + (q_\theta/q)\alpha_u(q_\theta/q) \end{aligned}$$

is called an **order-2 local score** or **2-local score**, since it depends on the distribution only through the density q and its first two derivatives q_θ and $q_{\theta\theta}$, evaluated at the local point θ

➔ An **objective prior** for θ is defined as the solution to the differential equation

$$S(q, q_\theta, q_{\theta\theta}) = 0$$

Example: Let $\alpha(u) = u^{-2}$, thus $\phi(\mathbf{p}) = p\alpha(p_\theta/p) = p^3/p_\theta^2$ and

$$S(\mathbf{q}) = 3\left(\frac{q}{q_\theta}\right)^2\left[\frac{2q\,q_{\theta\theta}}{q_\theta^2} - 3\right]$$

Solving $S(\mathbf{q}) = 0$ results in a prior

$$q(\theta) = \frac{a}{(a+\theta)^2}, \quad \theta \in [0, \infty)$$

➔ This is a **Lomax distribution** with **scale** $a > 0$ and **shape** $k = 1$

- Heavy-tailed distribution related to the generalized Pareto
- $\mathbb{E}_q[\theta] = \infty$
- $q(\theta)$ is decreasing and convex
- Invariance to the transformation $t(\theta) = 1/\theta$ holds iif $a = 1$

➔ A prior with similar properties for $\theta \in \Theta = (-\infty, \infty)$ can be obtained through symmetrization:

$$q(\theta) = \frac{a}{2(a+|\theta|)^2}$$

We begin with the prior for $\theta \in [0, \infty)$:

$$q(\theta) = \frac{a}{(a+\theta)^2} \quad \text{i.e. } \theta \sim L(a, 1)$$

and consider a second parameter $\tau \in [0, \infty)$

➔ Definition of the **joint prior** for $(\theta, \tau) \in [0, \infty)^2$ requires the definition of $q(\tau|\theta)$:

- If the support of τ is $[0, \infty)$ for all θ, $q(\tau|\theta)$ should also be a **Lomax distribution**
- If a priori **independence is not assumed**, the parameters of $q(\tau|\theta)$ may depend on θ, thus

$$q(\tau|\theta) = \frac{\tilde{a}(\theta)^{\tilde{k}(\theta)}\tilde{k}(\theta)}{(\tilde{a}(\theta) + \tau)^{\tilde{k}(\theta)+1}} \quad \text{i.e. } \tau|\theta \sim L\left(\tilde{a}(\theta), \tilde{k}(\theta)\right)$$

- The **joint prior** should not depend on the order in which the two parameters are considered. In other words,

$$q(\theta)q(\tau|\theta) = q(\tau)q(\theta|\tau)$$

Our proposal: A prior for a 2-dimensional parameter

By symmetry, τ and θ should have the same marginal distribution and the following equality should hold

$$\frac{a}{(a+\theta)^2}\,\frac{\tilde{a}(\theta)^{\tilde{k}(\theta)}\tilde{k}(\theta)}{(\tilde{a}(\theta)+\tau)^{\tilde{k}(\theta)+1}} = \frac{\tilde{a}}{(\tilde{a}+\tau)^2}\,\frac{a(\tau)^{k(\tau)}k(\tau)}{(a(\tau)+\theta)^{k(\tau)+1}}$$

This is achieved iif $k(\tau) = \tilde{k}(\theta) = 2$, $a = \tilde{a}$, $\tilde{a}(\theta) = a + \theta$ and $a(\tau) = a + \tau$.

The **joint prior** for $(\theta, \tau) \in [0, \infty)^2$ is therefore

$$q(\theta, \tau) = \frac{2a}{(a+\theta+\tau)^3}$$

➔ This is a **bivariate Lomax distribution** with **scale** $a > 0$ and **shape** $k = 1$

- Heavy-tailed distribution related to the bivariate generalized Pareto
- $\mathbb{E}_q[\theta] = \mathbb{E}_q[\tau] = \infty$ but $\mathbb{E}_q[\theta] = a + \theta$, $\mathbb{E}_q[\tau] = a + \tau$
- $q(\theta, \tau)$ is decreasing and convex
- Invariance to the transformation $t(\theta, \tau) = (1/\theta, 1/\tau)$ holds iif $a = 1$

➔ A prior with similar properties for $(\theta, \tau) \in \Theta = (-\infty, \infty) \times [0, \infty)$ can be obtained through symmetrization:

$$q(\theta, \tau) = \frac{a}{(a+|\theta|+\tau)^2}$$

Prior derivation from Bregman 2-local scores

Consider a 3-dimensional **Bregman divergence:**

$$D(\mathbf{p}, \mathbf{q}) = \int B_\phi(\mathbf{p}, \mathbf{q});$$

$$B_\phi(\mathbf{p}, \mathbf{q}) = \phi(\mathbf{p}) - \phi(\mathbf{q}) - \phi_q(\mathbf{q})(p - q) - \phi_{q_\theta}(\mathbf{q})(p_\theta - q_\theta) - \phi_{q_\tau}(\mathbf{q})(p_\tau - q_\tau)$$

for a convex function $\phi : \mathbb{R}^3_+ \to \mathbb{R}$, where $\mathbf{p} = (p, p_\theta, p_\tau)$, $\mathbf{p} = (q, q_\theta, q_\tau)$,

$$q_\theta(\theta, \tau) = \frac{\partial q(\theta, \tau)}{\partial \theta}; \quad q_\tau(\theta, \tau) = \frac{\partial q(\theta, \tau)}{d\tau};$$

$$\phi_q(\mathbf{q}) = \frac{\partial \phi(\mathbf{q})}{\partial q}; \quad \phi_{q_\theta}(\mathbf{q}) = \frac{\partial \phi(\mathbf{q})}{\partial q_\theta}; \quad \phi_{q_\tau}(\mathbf{q}) = \frac{\partial \phi(\mathbf{q})}{\partial q_\tau}$$

The resulting **bivariate 2-local score** is

$$S(\mathbf{q}) = -\phi_q(\mathbf{q}) + \frac{\partial \phi_{q_\theta}(\mathbf{q})}{\partial \theta} + \frac{\partial \phi_{q_\tau}(\mathbf{q})}{\partial \tau} = 4\left(\frac{q}{q_\theta}\right)^3 \left[\frac{3q\, q_{\theta\theta}}{q_\theta^2} - 4\right] + 4\left(\frac{q}{q_\tau}\right)^3 \left[\frac{3q\, q_{\tau\tau}}{q_\tau^2} - 4\right]$$

Solving $S(\mathbf{q}) = 0$, under symmetry conditions, results in a prior

$$q(\theta, \tau) = \frac{2a}{(a + \theta + \tau)^3}$$

➔ Once again, this is a **bivariate Lomax distribution** with **scale** $a > 0$ and **shape** $k = 1$:

$$(\theta, \tau) \sim L_2(a, k)$$

In dimension $d \geq 2$

In general, for $\boldsymbol{\theta} = (\theta_1, \dots, \theta_d) \in [0, \infty)^d$ we consider the **Bregman divergence** of dimension $d + 1$ induced by

$$B_\phi(\mathbf{p}, \mathbf{q}) = \phi(\mathbf{p}) - \phi(\mathbf{q}) - \phi_q(\mathbf{q})(p - q) - \sum_{i=1}^{d} \phi_{q_i}(\mathbf{q})(p_i - q_i)$$

for a convex function $\phi : \mathbb{R}_+^{d+1} \to \mathbb{R}$, where $\mathbf{p} = (p, p_1, \dots, p_d)$, $\mathbf{p} = (q, q_1, \dots, q_d)$,

$$q_i(\boldsymbol{\theta}) = \frac{\partial q(\boldsymbol{\theta})}{\partial \theta_i}; \quad \phi_q(\mathbf{q}) = \frac{\partial \phi(\mathbf{q})}{\partial q}; \quad \phi_{q_i}(\mathbf{q}) = \frac{\partial \phi(\mathbf{q})}{\partial q_i}$$

We let

$$\phi(\mathbf{p}) = p\alpha\left(\frac{p_1}{p}, \dots \frac{p_d}{p}\right)$$

for

$$\alpha(\mathbf{u}) = \sum_{i=1}^{d} u_i^{-(d+1)}; \quad \mathbf{u} = (u_1, \dots, u_d)$$

The resulting **multivariate 2-local score** is

$$S(\mathbf{q}) = -\phi_q(\mathbf{q}) + \sum_{i=1}^{d} \frac{\partial \phi_{q_i}(\mathbf{q})}{\partial \theta_i} = (d+2) \sum_{i=1}^{d} \left(\frac{q}{q_i}\right)^{d+1} \left[\frac{(d+1) q \, q_{ii}}{q_i^2} - (d+2)\right],$$

where

$$q_{ii}(\boldsymbol{\theta}) = \frac{\partial^2 q(\boldsymbol{\theta})}{\partial \theta_i^2}$$

Solving $S(\mathbf{q}) = 0$ we obtain the **joint prior** for $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d) \in [0, \infty)^d$,

$$q(\boldsymbol{\theta}) = \frac{da}{\left(a + \sum_{i=1}^{d} \theta_i\right)^{d+1}}$$

➔ This is a **multivariate Lomax distribution** with **scale** $a > 0$ and **shape** $k = 1$:

$$\boldsymbol{\theta} \sim L_d(a, k)$$

The same prior can be obtained by the conditional construction, sequentially deriving $q(\theta_{i+1}|\theta_1, \ldots, \theta_i)$, and

$$\theta_{i+1}|\theta_1, \ldots, \theta_i \sim L\Big(a + \sum_{j=1}^{i} \theta_j, i+1\Big)$$

- For a Lomax distribution $L(a, k)$:
  - The larger the shape parameter, the "lighter" the tail:

    The expectation is finite whenever $k > 1$

    The variance is finite whenever $k > 2$

- Intuitively, while the joint prior is heavy tail, it does not assign "too much" mass on the tails of the multivariate distribution

The **joint prior** for $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d) \in (-\infty, \infty)^r \times [0, \infty)^{d-r}$ can be obtained by symmetrization:

$$q(\boldsymbol{\theta}) = \frac{da}{2^r \left(a + \sum_{i=1}^{r} |\theta_i| + \sum_{i=r+1}^{d} \theta_i\right)^{d+1}}$$

We consider $X \sim \text{Weibull}(\theta, \beta)$ and draw 250 independent samples of size n for $\theta = 1$ and $\beta = \{0.5, 1, 100\}$ (see Sun, 1997). We compare our Lomax prior with the reference prior via relative MSE with respect to the posterior mean and coverage for 95% credible intervals

n=30	MSE - β				MSE - θ		
		$\beta = 0.5$	$\beta = 1$	$\beta = 10$	$\beta = 0.5$	$\beta = 1$	$\beta = 10$
	Reference	3.92	3.90	3.91	3.13	3.13	3.12
	Lomax	3.33	3.27	3.21	2.59	2.61	2.69
	COV - β				COV - θ		
		$\beta = 0.5$	$\beta = 1$	$\beta = 10$	$\beta = 0.5$	$\beta = 1$	$\beta = 10$
	Reference	0.90	0.91	0.91	0.91	0.92	0.91
	Lomax	0.91	0.91	0.90	0.95	0.96	0.96
n=100	MSE - β				MSE - θ		
		$\beta = 0.5$	$\beta = 1$	$\beta = 10$	$\beta = 0.5$	$\beta = 1$	$\beta = 10$
	Reference	1.85	1.86	1.94	1.36	1.37	1.37
	Lomax	1.77	1.75	1.76	1.29	1.29	1.30
	COV - β				COV - θ		
		$\beta = 0.5$	$\beta = 1$	$\beta = 10$	$\beta = 0.5$	$\beta = 1$	$\beta = 10$
	Reference	0.94	0.95	0.94	0.94	0.93	0.94
	Lomax	0.93	0.93	0.92	0.95	0.94	0.94

Single sample results on real data: $n = 19$ times to breakdown (minutes) of an insulating fluid between electrodes at a voltage of 34 KV (see Ellah, 2012)

- Observations:

0.96	4.15	0.19	0.78	8.01	31.75	7.35	6.50	8.27	33.91
32.52	3.16	4.85	2.78	4.67	1.31	12.06	36.71	72.89	

- Posterior summaries:

	Reference			Lomax		
	Mean	Variance	95% C.I.	Mean	Variance	95% C.I.
θ	0.8	0.02	(0.55,1.10)	0.73	0.02	(0.48,1.02)
β	16.84	44.93	(8.51,31.89)	11.11	15.08	(5.07,20.36)

- Maximum likelihood estimates: $\hat{\theta} = 0.77$ and $\hat{\beta} = 12.22$

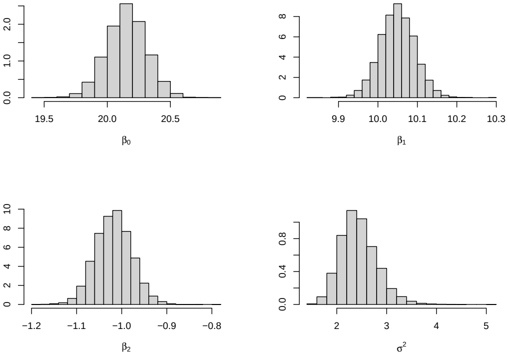
We consider 250 independent samples of size $n = 100$ from a linear regression model with two covariates, coefficients $\beta = (20, 10, -1)$ and variance $\sigma^2 = 2$. We compare our Lomax prior with a vague prior and Zelner's g prior via MSE with respect to the maximum a posteriori and coverage for 95% credible intervals

	MSE			COV		
Parameter	Lomax	Vague	Zellner's g	Lomax	Vague	Zellner's g
β_0	0.998	0.998	0.998	0.95	0.94	0.95
β_1	1.000	1.000	1.000	0.97	0.96	0.97
β_2	1.001	1.002	1.001	0.97	0.98	0.98
σ^2	1.098	1.095	1.094	0.93	0.92	0.92

EXAMPLE 2: LINEAR REGRESSION

Single sample results on simulated data: sample of size $n = 100$ from the linear regression model with intercept $\beta_0 = 20$, coefficients $\beta_1 = 10$ and $\beta_2 = -1$, and variance σ^2=2

Posterior histograms:



Posterior summary:

	Median	95% C.I.
β_0	20.15	(19.84, 20.47)
β_1	10.05	(9.96, 10.14)
β_2	-1.02	(-1.10, -0.94)
σ^2	2.41	(1.83, 3.26)

Thank you!

References

James O. Berger, José M. Bernardo, and Dongchu Sun. The formal definition of reference priors. *The Annals of Statistics*, 37:905–938, 2009.

Ahmed H.A. Ellah. Bayesian and non-bayesian estimation of the inverse weibull model based on generalized order statistics. *Intelligent Information Management*, 4:23–31, 2012.

Dennis V. Lindley and Nozer D. Singpurwalla. Multivariate distributions for the life lengths of components of a system sharing a common environment. *Journal of Applied Probability*, 23 (2):418–431, 1986.

Tapan Kumar Nayak. Multivariate lomax distribution: Properties and usefulness in reliability theory. *Journal of Applied Probability*, 24(1):170–177, 1987.

Matthew Parry, A. Philip Dawid, and Steffen Lauritzen. Proper local scoring rules. *The Annals of Statistics*, 40(1):561 – 592, 2012.

Dongchu Sun. A note on noninformative priors for weibull distributions. *Journal of Statistical Planning and Inference*, 61:319–338, 1997.

Stephen G. Walker and Cristiano Villa. An objective prior from a scoring rule. *Entropy*, 23(7), 2021.

A. Zellner. On Assessing Prior Distributions and Bayesian Regression Analysis with g Prior Distributions. *In Goel, P.; Zellner, A. (eds.). Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti. Studies in Bayesian Econometrics and Statistics. Vol. 6.*, pages 233–243, 1986.