The Formal Definition of Reference Priors

Dongchu Sun

University of Missouri and SAMSI

Jointly with *James O. Berger, Duke University*
and *José M. Bernardo, University of Valencia*

April 27, 2010, University of North Carolina–Chapel Hill

Outline

1. Introduction/Early Objective Priors
2. Jeffreys prior: Regular Cases
3. Reference Priors: Some Non-regular Cases
4. Reference Priors: A Formal Definition
5. Examples
6. Comments

Introduction–Use Noninformative Priors?

- A Bayesian analysis, which combines available prior information and data information, can be useful in practice.
- Ideally, one could use a subjective Bayesian analysis if there were enough information.
- In practice, however, the difficulties of subjective elicitation and time restrictions frequently limit us to use noninformative priors.
- As default priors depending only on model, used at the first stage; subjective priors can be used later.
- The most commonly used noninformative prior is the Jeffreys prior.

Earlier Objective Priors

1. Simon Laplace (1812) used a constant prior.
   (a) Strengths. It is simple and convenient; posterior normality.
   (b) Weakness. If $X \sim Bin(n, p)$, an constant prior for p is different from a constant prior for $log[p/(1-p)]$.
2. From 1838–1950, it was called inverse probability, by de Morgan.
3. For example, Egon Pearson (1925) recommended to use 'Jeffreys prior" for p: $p^{-1/2}(1-p)^{-1/2}$.

The Jeffreys Prior

- Consider a statistical model $\mathcal{M} \equiv \{p(\boldsymbol{x} \mid \theta),\, \boldsymbol{x} \in \boldsymbol{\mathcal{X}},\, \theta \in \Theta\}$.
- Under regularity conditions, the Fisher information

$$I(\theta) \;=\; -E^{\boldsymbol{x}|\theta}\left[\frac{\partial^2}{\partial\theta^2}\log p(\boldsymbol{x} \mid \theta)\right]$$

- Jeffreys prior: $\pi(\theta) \propto \sqrt{I(\theta)}$.
- $\boldsymbol{x}$ could be independent or dependent observations.
- Invariance. If ψ is 1-1 of θ: $\pi(\psi) = \pi(\theta)|d\theta/d\psi|$.

Simple Examples

- $Binomial\ (n, p)$. The Jeffreys prior for p is Beta $(.5, .5)$.
- $Normal\ (\mu, \sigma^2)$. The Jeffreys prior for μ is a constant.
- $Normal\ (\mu, \sigma^2)$. The Jeffreys prior for σ is $1/\sigma$.

Maximizing Missing Information

For any prior $\pi(\theta)$, consider the KL (Kullback-Leibler) divergence:

$$K[\pi(\cdot \mid \boldsymbol{x}^{(k)}),\ \pi] = \int \pi(\theta \mid \boldsymbol{x}^{(k)}) \log\left\{\frac{\pi(\theta \mid \boldsymbol{x}^{(k)})}{\pi(\theta)}\right\} d\theta,$$

where $\boldsymbol{x}^{(k)} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k\}$ is k iid replications of $\boldsymbol{x}$.

- The expected KL divergence: $E\,K[\pi(\cdot \mid \boldsymbol{x}^{(k)}),\ \pi]$.
- Maximizing asymptotically over all $\pi(\cdot)$, $\pi^R(\theta) \propto \sqrt{I(\theta)}$.
- The argument was heuristics: Bernardo's (1979) reference prior.
- A proof was given by Clarke & Barron (1994) under regularity conditions, need common support, compact set arguments, etc.

Reference Prior Without a Common Support

$x \sim uniform(0, \theta)$, $\theta > 0$.

- No Fisher information
- It is a scale family bacause $x/\theta \sim$ uniform $(0, 1)$. It should be $\pi(\theta) \propto 1/\theta$.

Other cases?

Reference Prior: Nonregular Cases

Ghosal and Samanta (1997) considered a class of nonregular cases.

- Let $p(x \mid \theta)$ be a pdf, $x \in S(\theta) = [a_1(\theta), a_2(\theta)]$, at least one of them is not a constant.
- Nonregular case: $S(\theta)$ for $\theta \in \Theta \subset \mathbb{R}$ is either increasing or decreasing. (So $a_1(\theta)$ and $a_2(\theta)$ are both monotonic.)

  $\star$ $S(\theta) = [0, \theta]$, $\theta > 0$ is a nonregular case.
  $\star$ $S(\theta) = [\theta, \theta^{-1}]$, $\theta \in (0, 1]$ is a nonregular case.
  $\star$ $S(\theta) = [1 - \theta^{-1}, 1 + 2\theta^{-1}]$, $\theta \in [1, \infty)$ is a nonregular case.
  $\star$ $S(\theta) = [1 + \theta^{-1}, 3 + 2\theta]$, $\theta \in [1, \infty)$ is not a nonregular case.
  $\star$ $S(\theta) = [\theta, \theta^2]$, $\theta \in [1, \infty)$ is not a nonregular case.

Reference Prior: Nonregular Cases, Continue

Let $p(x \mid \theta)$ be a pdf, $x \in S(\theta) = [a_1(\theta), a_2(\theta)]$, $S(\theta)$ is monotonic.

- Assume $f(x \mid \theta) \to g_1(\theta)$ if $x \to a_1(\theta)$ and $\to g_2(\theta)$ if $x \to a_2(\theta)$.
- Let $c(\theta) = a_1'(\theta)g_1(\theta) - a_2'(\theta)g_2(\theta)$.
- Under regular conditions, the reference prior is $\pi^R(\theta) \propto |c(\theta)|$.
- **Example 1: Uniform on** $[0, \theta],\ \theta > 0$**.** Here $g_1(\theta) = g_2(\theta) = 1/\theta$, $a_1'(\theta) = 0$, and $a_2'(\theta) = 1$. Then $c(\theta) = -1/\theta$, and $\pi^R(\theta) \propto 1/\theta$.
- **Example 2: Uniform on** $[\theta, \theta^{-1}],\ \theta \in (0, 1]$**.**

  Here $g_1(\theta) = g_2(\theta) = \theta/(1 - \theta^2)$, $a_1'(\theta) = 1$, and $a_2'(\theta) = -1/\theta^2$. Then $c(\theta) = \theta(1 + \theta^2)/(1 - \theta^2) = \pi^R(\theta)$.

Goals

- Applying Bayes theorem to improper priors was not justifiable. Formalizing when this is legitimate is desirable.
- Previous attempts at a general definition of reference priors have had heuristic features, especially in situations where the reference prior is improper. Replacing the heuristics with a formal definition is desirable.

A Formal Definition of Reference Priors

Definition 1. *A strictly positive continuous function* $\pi(\theta)$ *is a permissible prior for model* $\mathcal{M} = \{p(\boldsymbol{x}\,|\,\theta),\, \boldsymbol{x} \in \boldsymbol{\mathcal{X}},\, \theta \in \Theta\}$ *if*

1. *for all* $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$, $\pi(\theta\,|\,\boldsymbol{x})$ *is proper, i.e.* $\int_\Theta p(\boldsymbol{x}\,|\,\theta)\,\pi(\theta)\,d\theta < \infty$;

2. *for some approximating compact sequence* Θ_i, *the corresponding posterior sequence is expected logarithmically convergent to* $\pi(\theta\,|\,\boldsymbol{x}) \propto p(\boldsymbol{x}\,|\,\theta)\,\pi(\theta)$, *i.e.,*

$$\lim_{i\to\infty} \int_{\boldsymbol{\mathcal{X}}} K[\pi(\cdot\,|\,\boldsymbol{x}),\, \pi_i(\cdot\,|\,\boldsymbol{x})]\, p_i(\boldsymbol{x})\, d\boldsymbol{x} = 0. \qquad (1)$$

where $p_i(\boldsymbol{x}) = \int_{\Theta_i} p(\boldsymbol{x}\,|\,\theta)\,\pi_i(\theta)\,d\theta$, $\pi_i(\theta) = \pi(\theta)1_{\Theta_i}(\theta)/\int_{\Theta_i}\pi(\theta)d\theta$.

Remark 1. *(a) If the posterior of θ based on π is proper,*

$$\lim_{i\to\infty} K[\pi(\cdot\,|\,\boldsymbol{x}),\ \pi_i(\cdot\,|\,\boldsymbol{x})] = 0,\ a.s.$$

(b) The 2nd condition, given in Berger & Bernardo (1992) and motivated by Fraser, Monette, and Ng (1985)'s example.

(c) Consider the location model $\mathcal{M} = \{f(x-\theta),\, \theta \in \mathbb{R},\, x \in \mathbb{R}\}$, where $f(t)$ is a pdf on $\mathbb{R}$.

- *Assume that for some $\epsilon > 0$, $\lim_{|t|\to 0} |t|^{1+\epsilon} f(t) = 0$. Then $\pi(\theta) = 1$ is a permissible prior for the model $\mathcal{M}$.*
- *Assume that $f(t) = t^{-1}(\log t)^{-2}$, $t > e$ and $\pi(\theta) = 1$. Then for any compact set $\Theta_0 = [a, b]$ with $b - a \geq 1$,*

  $$\int_{\Theta_0} K\{\pi(\cdot\,|\,x), \pi_0(\cdot\,|\,x)\}\, p_0(x)\, dx = \infty.$$

  Thus $\pi(\theta) = 1$ is not a permissible prior for $\mathcal{M}$.

Maximizing Missing Information (MMI) Property

Definition 2. *Let* $\mathcal{M} \equiv \{p(\boldsymbol{x} \mid \theta),\, \boldsymbol{x} \in \boldsymbol{\mathcal{X}},\, \theta \in \Theta \in \mathbb{R}\}$*, be a model with one continuous parameter. Let* $\mathcal{P} = \{p(\theta) > 0 : \int_{\Theta} p(\boldsymbol{x} \mid \theta) p(\theta) d\theta < \infty\}$*. The function* $\pi(\theta)$ *is said to have the MMI property for model* $\mathcal{M}$ *given* $\mathcal{P}$ *if, for any compact set* $\Theta_0 \in \Theta$ *and any* $p \in \mathcal{P}$*,*

$$\lim_{k \to \infty} \{I\{\pi_0 \mid \mathcal{M}^k\} - I\{p_0 \mid \mathcal{M}^k\}\} \geq 0, \tag{2}$$

where π_0 *and* p_0 *are, respectively, the renormalized restrictions of* $\pi(\theta)$ *and* $p(\theta)$ *to* Θ_0*, and*

$$I\{p_0 \mid \mathcal{M}^k\} = \int_{\boldsymbol{\mathcal{X}}^k} K[p_0(\cdot \mid \boldsymbol{x}^{(k)}), p_0]\, p_0(\boldsymbol{x}^{(k)})\, d\boldsymbol{x}^{(k)}.$$

Definition of A Reference Prior

Definition 3. *A function* $\pi(\theta)$ *is a reference prior for model* $\mathcal{M} \equiv \{p(\boldsymbol{x} \mid \theta), \boldsymbol{x} \in \boldsymbol{\mathcal{X}}, \theta \in \Theta \in \mathbb{R}\}$ *given* $\mathcal{P} = \{p(\theta) > 0 : \int_\Theta p(\boldsymbol{x} \mid \theta)p(\theta)d\theta < \infty\}$ *if it is permissible and has the MMI property.*

- Implicitly, the reference prior on Θ is also the reference prior on any compact subset Θ_0. It is attractive and often stated as the practical way to proceed, when dealing with a restricted parameter space, but here it is simply a consequence of the definition.

- We feel that a reference prior needs to be both permissible and have the MMI property, the later is considerably more important. Others have defined reference priors only in relation to this property. For example, Clarke and Barron (1994).

- Is there a reference prior?
- Is it unique?

A Simple Constructive Formula for a Reference Prior

Theorem 1. For a model with pdf $p(\boldsymbol{x} \mid \theta)$, ($\boldsymbol{x}$ is the data & θ is a continuous unknown parameter), among the class of permissible priors with posterior consistency, the reference prior $\pi(\theta)$ is

$$\pi(\theta) = \lim_{k\to\infty} \frac{f_k(\theta)}{f_k(\theta_0)}, \tag{3}$$

$$f_k(\theta) = \exp\left\{ \int p(\boldsymbol{x}^{(k)} \mid \theta) \log\left[\pi^*(\theta \mid \boldsymbol{x}^{(k)})\right] d\boldsymbol{x}^{(k)} \right\}, \tag{4}$$

where θ_0 is an interior point of the parameter space Θ,

$\boldsymbol{x}^{(k)} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k\}$: k iid replications of $\boldsymbol{x}$,

and $\pi^*(\theta \mid \boldsymbol{x}^{(k)})$ is the posterior based on any fixed prior $\pi^*(\theta)$.

Recall

$$\pi(\theta) = \lim_{k\to\infty} \frac{f_k(\theta)}{f_k(\theta_0)},$$
$$f_k(\theta) = \exp\left\{ \int p(\boldsymbol{x}^{(k)} \mid \theta) \log\left[\pi^*(\theta \mid \boldsymbol{x}^{(k)})\right] d\boldsymbol{x}^{(k)} \right\}.$$

- It is interesting that this expression holds (under mild conditions) for any type of continuous parameter, regardless of the asymptotic nature of the posterior.
- A second use of the expression is that it allows straightforward computation of the reference prior numerically. This is illustrated for a difficult nonregular problem, and for a problem for which analytical determination of the reference prior seems very difficult.

Example 3. Uniform on $(a_1(\theta), a_2(\theta))$

Here $0 < a_1(\theta) < a_2(\theta)$ are both strictly monotonic increasing functions on $\Theta = (\theta_0, \infty)$ with derivatives satisfying $0 < a_1'(\theta) < a_2'(\theta)$.

— is not a regular model;

— has no group invariance structure;

— doesn't belong to Ghosal & Samanta (1997) nonregular models.

Example 3. Uniform on $(a_1(\theta), a_2(\theta))$, Continue

Here $0 < a_1(\theta) < a_2(\theta)$ are both strictly monotonic increasing functions on $\Theta = (\theta_0, \infty)$ with derivatives satisfying $0 < a_1'(\theta) < a_2'(\theta)$.

Theorem 2. Define

$$b_k \equiv b_k(\theta) = \frac{a_2'(\theta) - a_1'(\theta)}{a_k'(\theta)}, \quad k = 1, 2.$$

The reference prior of θ is

$$\pi(\theta) = \frac{a_2'(\theta) - a_1'(\theta)}{a_2(\theta) - a_1(\theta)} \exp\left\{ b_1 + \frac{1}{b_1 - b_2}\left[b_1 \psi\left(\frac{1}{b_1}\right) - b_2 \psi\left(\frac{1}{b_2}\right)\right]\right\},$$

where $\psi(z)$ is the digamma function, $\psi(z) = \frac{d}{dz}\log(\Gamma(z))$, $z > 0$.

Example 4. Uniform on (θ, θ^2), $\theta \geq 1$

This is a special case of Theorem 2 with $\theta_0 = 1$, $a_1(\theta) = \theta$, and $a_2(\theta) = \theta^2$. Then $b_1 = 2\theta - 1$ and $b_2 = (2\theta - 1)/(2\theta)$. It is easy to show that $b_2^{-1} = b_1^{-1} + 1$. For the digamma function,

$$\psi(z+1) \;=\; \psi(z) + \frac{1}{z}, \text{ for } z > 0, \tag{5}$$

so that $\psi(1/b_1) = \psi(1/b_2) - b_1$. The reference prior for θ is then

$$\pi(\theta) = \frac{2\theta - 1}{\theta(\theta - 1)} \exp\left\{\psi\left(\frac{2\theta}{2\theta - 1}\right)\right\}, \quad \theta \geq 1. \tag{6}$$

Numerical Computation of the Reference Prior

Analytical derivation of reference priors may be technically demanding in complex models. However, Formula (3) may also be used to obtain an approximation to the reference prior through its numerical evaluation. Moderate values of k (to simulate the asymptotic posterior) are typically sufficient to obtain a good approximation to the reference prior.

A Pseudo Code

1. Starting values:
   - choose a moderate value for k;
   - choose an arbitrary positive function $\pi^*(\theta)$, say $\pi^*(\theta) = 1$;
   - choose the number m of samples to be simulated.

2. For any given θ value, **repeat**, for $j = 1, \ldots, m$:
   - simulate a RS $\{\boldsymbol{x}_{1j}, \ldots, \boldsymbol{x}_{kj}\}$ of size k from $p(\boldsymbol{x} \mid \theta)$;
   - compute numerically $c_j = \int_\Theta \prod_{i=1}^k p(\boldsymbol{x}_{ij} \mid \theta)\, \pi^*(\theta) d\theta$;
   - evaluate $r_j(\theta) = \log[\, \prod_{i=1}^k p(\boldsymbol{x}_{ij} \mid \theta)\, \pi^*(\theta)/c_j\,]$.
   - compute $\pi(\theta) = \exp[\, m^{-1} \sum_{j=1}^m r_j(\theta)\,]$ and **store** $\{\theta, \pi(\theta)\}$.

3. **Repeat** routines (2) for all θ values where the pair $\{\theta, \pi(\theta)\}$ is required.

Example 4. Uniform on $(\theta, \theta^2),\ \theta \geq 1$: Revisit

The reference prior: $\pi(\theta) = \dfrac{2\theta - 1}{\theta(\theta - 1)} \exp\left\{\psi\left(\dfrac{2\theta}{2\theta - 1}\right)\right\}, \quad \theta \geq 1.$

- We present the reference prior numerically calculated with the algorithm for nine θ values, uniformly log-spaced and rescaled to have $\pi(2) = 1$; $m = 1000$ samples of $k = 500$ observations were used to compute each of the nine $\{\theta_i, \pi(\theta_i)\}$ points.
- These nine points are clearly almost perfectly fitted by the exact reference prior, shown by a continuous line; indeed the nine points were accurate to within four decimal points.

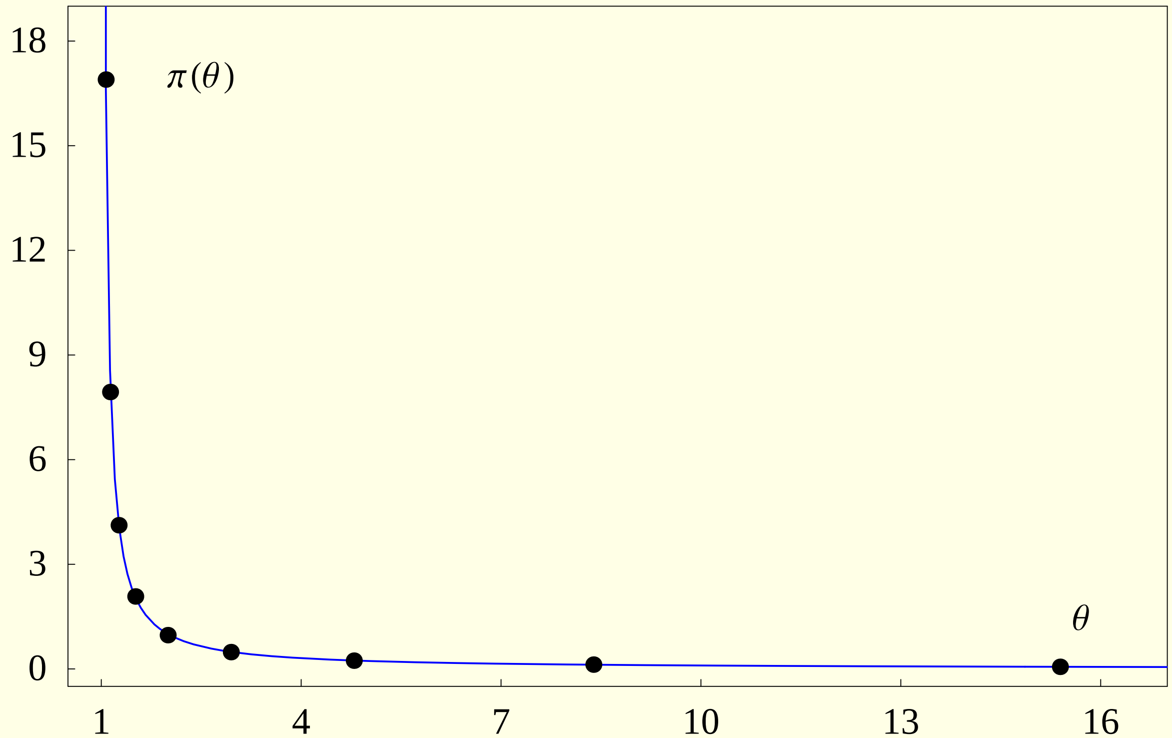

Figure 1. Numerical reference prior for the uniform model on (θ, θ^2)

The computation was done before obtaining the analytic reference prior. A nearly perfect fit was $\pi(\theta) = 1/(\theta - 1)$. It was guessed to be the actual reference prior. This guess was wrong, but over the computed range it is indeed nearly proportional to $1/(\theta - 1)$.

Example 5: A Triangular Distribution

- The use of a symmetric triangular density on $(0,1)$ can be traced back to the 18th century to Simpson (1855).
- Schmidt (1934) noticed that it has the same density of two iid uniform random variables on $(0,1)$.
- Ayyangar (1941) studied the asymmetric triangular distribution on $(0,1)$,

$$p(x \mid \theta) = \begin{cases} 2x/\theta, & \text{for } 0 < x \leq \theta, \\ 2(1-x)/(1-\theta), & \text{for } \theta < x < 1, \end{cases} \qquad 0 < \theta < 1.$$

- Johnson and Kotz (1999) revisited nonsymmetric triangular distributions in the context of modeling prices.

- The nonsymmetric triangular distribution does not have a sufficient statistic of finite dimension.

- Although $\log[p(x \mid \theta)]$ is differentiable for all θ values, the formal Fisher information is negative, so Jeffreys-rule prior does not exist.

- Figure 2 presents a numerical calculation of the reference prior at thirteen θ values, equally spaced on $(0, 1)$, and rescaled to have $\pi(1/2) = 2/\pi$; $m = 2500$ samples of $k = 2000$ observations were used to compute each of the thirteen $\{\theta_i, \pi(\theta_i)\}$ points. Interestingly, these points are nearly perfectly fitted by the (proper) prior $\pi(\theta) = Beta(\theta \mid 1/2, 1/2) \propto \theta^{-1/2}(1-\theta)^{-1/2}$, shown by a continuous line.

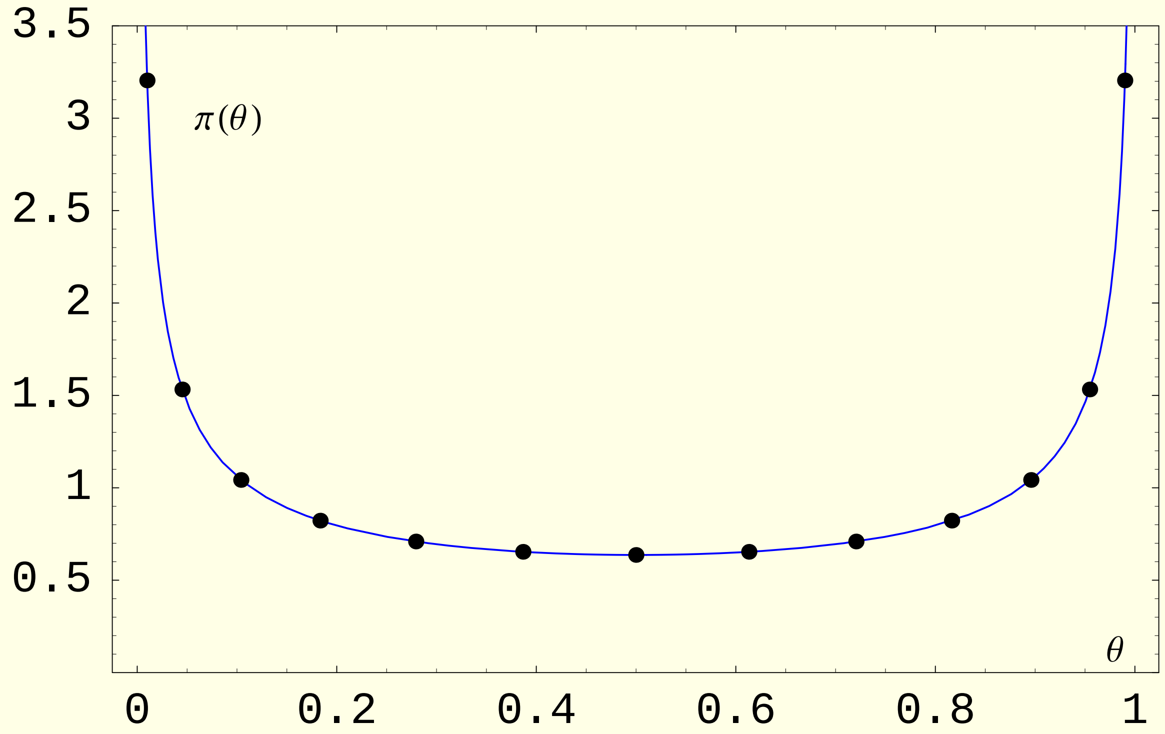

Figure 2. Numerical reference prior for the triangular distribution on $(0, 1)$

Analytical derivation of the reference prior does not seem to be feasible in this example, but there is an interesting heuristic argument based on asymptotically efficiency, suggesting that the $Beta(\theta \mid 1/2, 1/2)$ prior is indeed the reference prior.

Comments

- It is based on Berger, Bernardo, & Sun (2009).
- Give a formal rule for one continuous parameter cases.
- No regular conditions are needed.
- Multi-parameter cases: Lin, Berger and Sun (2010, in preparation)
  - ⋆ It depends on the order of importance
  - ⋆ It depends on the choice of compact sets
- Under semi-invariance structure.
- Spatial and temporal models.
- It depends on KL divergence. Other divergence?
- Frequentist matching?