Weighting the likelihood function

Bayesian and frequentist inference

Nancy Reid

February 23, 2008

1st Canada Mexico Statistics Meeting

Don Fraser, Ana-Maria Staicu, Grace Yun Yi, Elisabetta Maras

Bayesian posterior distribution

$$\mathrm{Pr}_m(\Psi \leq \psi \mid y) \doteq \Phi(r_B^*) = \Phi(r + \frac{1}{r} \log \frac{q_B}{r})$$

$Y \sim f(y;\theta) \quad Y = (Y_1, \dots, Y_n)$ — model

$\ell(\theta) = \ell(\psi,\lambda) = \log f(y;\psi,\lambda), \quad y \in R^n$ — log-likelihood

$r = \pm[2\{\ell(\hat\theta) - \ell(\hat\theta_\psi)\}]^{1/2}, \quad \psi \in R$ — likelihood root

$\hat\theta_\psi = (\psi, \hat\lambda_\psi)$ — constrained m.l.e.

$\pi_m(\psi \mid y) = \int \pi(\theta \mid y) d\lambda$

$\propto \int \exp \ell(\theta)\pi(\theta) d\lambda$ — marginal posterior

... Bayesian posterior

$$\mathrm{Pr}_m(\Psi \leq \psi \mid y) \doteq \Phi(r_B^*) = \Phi(r + \frac{1}{r}\log\frac{q_B}{r})$$

$$q_B = -\ell_{\mathrm{p}}'(\psi) j_{\mathrm{p}}(\hat{\psi})^{-1/2} \frac{|j_{\lambda\lambda}(\hat{\theta}_\psi)|^{1/2}}{|j_{\lambda\lambda}(\hat{\theta})|^{1/2}} \frac{\pi(\hat{\theta})}{\pi(\hat{\theta}_\psi)}$$

$\ell_{\mathrm{p}}(\psi) = \ell(\hat{\theta}_\psi)$ profile log-likelihood

$j(\theta) = -\ell''(\theta; y)$ observed information

$j(\theta) = \begin{bmatrix} j_{\psi\psi}(\theta) & j_{\psi\lambda}(\theta) \\ j_{\lambda\psi})\theta) & j_{\lambda\lambda}(\theta) \end{bmatrix}$ partitioned matrix

... Bayesian posterior

$$r_B^* \dot{\sim} N(0,1)$$

$$r_B^* = r + \frac{1}{r} \log \frac{q_B}{r}$$

r is the likelihood root

q_B is an adjusted score statistic

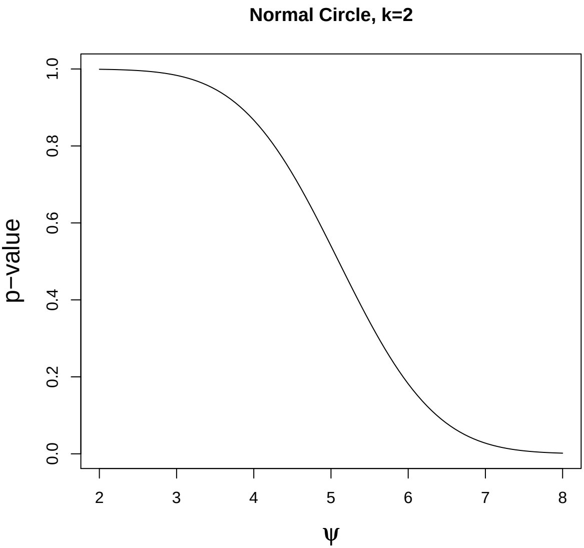
The approximation is very good!

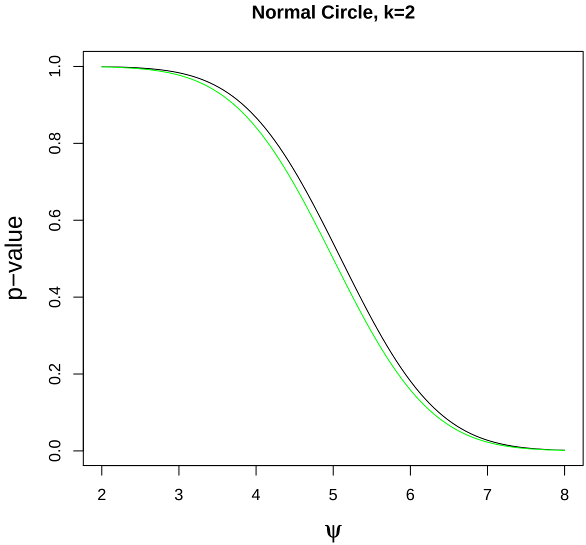
Example: Normal circle

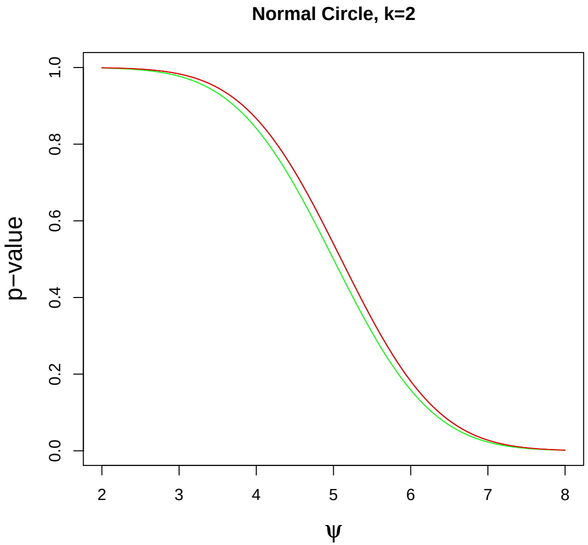
- $y_1 \sim N(\mu_1, 1/n), \ldots, y_k \sim N(\mu_k, 1/n)$
- parameter of interest $\psi = (\mu_1^2 + \cdots + \mu_k^2)^{1/2} = ||\mu||$
- prior $\pi(\mu) = 1$
- Exact marginal posterior $\quad \Pr\{\chi_k^2(n||y||^2) \geq n\psi^2\}$
- Third order

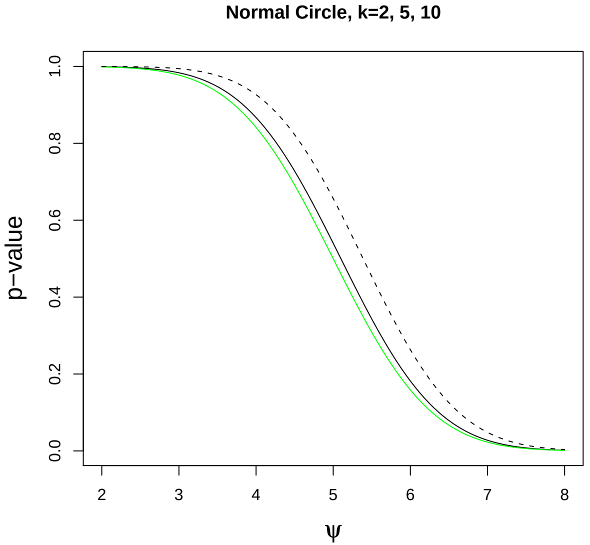
$$r_B^* = \sqrt{n}(\hat{\psi} - \psi) + \frac{1}{\sqrt{n}(\hat{\psi} - \psi)} \log\left\{\left(\frac{\hat{\psi}}{\psi}\right)^{(k-1)/2}\right\} \,\dot{\sim}\, N(0,1)$$

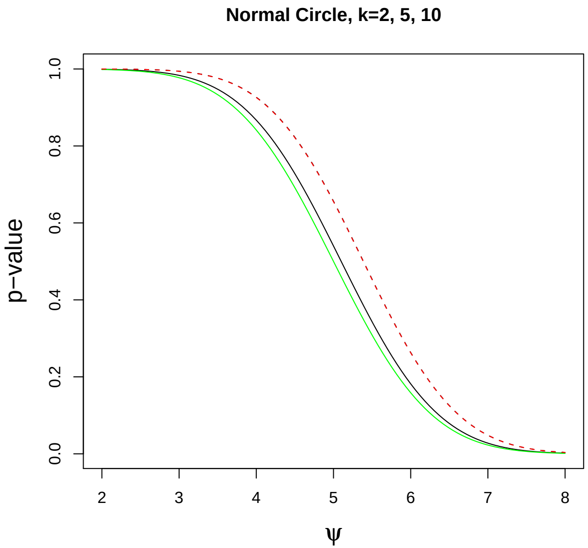
- Normal approximation to posterior $\sqrt{n}(\hat{\psi} - \psi) \,\dot{\sim}\, N(0,1)$

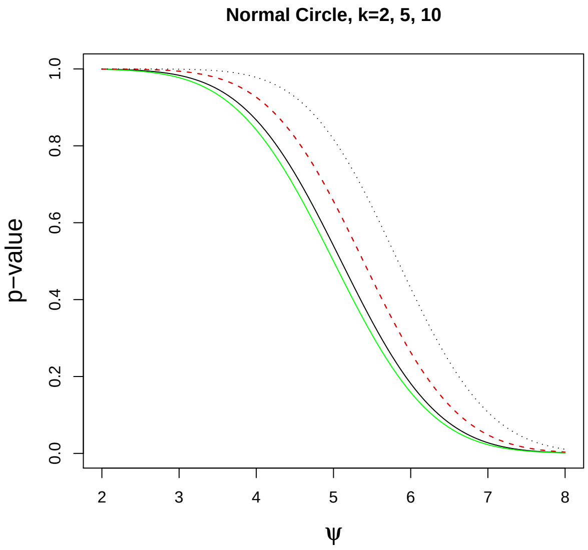

Normal Circle, k=2
p−value
1.0
0.8
0.6
0.4
0.2
0.0
2
3
4
5
6
7
8
ψ

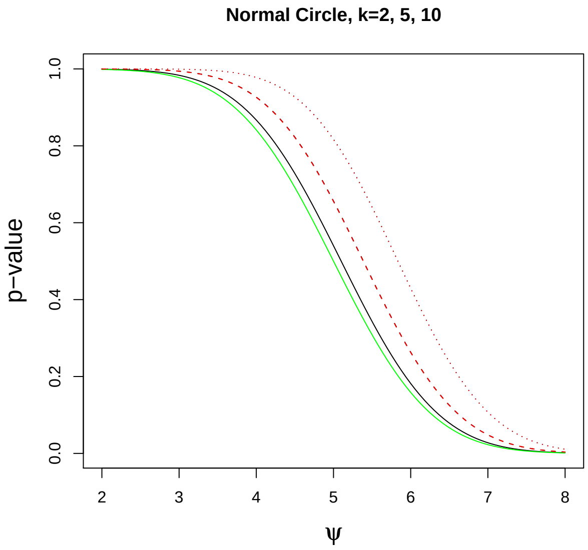

Normal Circle, k=2
p−value
0.0
0.2
0.4
0.6
0.8
1.0
2
3
4
5
6
7
8
ψ


Normal Circle, k=2
p−value
0.0
0.2
0.4
0.6
0.8
1.0
2
3
4
5
6
7
8
ψ


Normal Circle, k=2, 5, 10
p−value
1.0
0.8
0.6
0.4
0.2
0.0
2
3
4
5
6
7
8
ψ


Normal Circle, k=2, 5, 10
p−value
1.0
0.8
0.6
0.4
0.2
0.0
2
3
4
5
6
7
8
ψ


Normal Circle, k=2, 5, 10
p−value
0.0
0.2
0.4
0.6
0.8
1.0
2
3
4
5
6
7
8
ψ


Normal Circle, k=2, 5, 10
p−value
1.0
0.8
0.6
0.4
0.2
0.0
2
3
4
5
6
7
8
ψ

Frequentist P-value

$$\text{Pvalue}(\psi) \doteq \Phi(r_F^*) = \Phi(r + \frac{1}{r}\log\frac{q_F}{r})$$

$$r = \pm[2\{\ell(\hat{\theta}) - \ell(\hat{\theta}_\psi)\}]^{1/2}$$

$$q_F = \frac{|\ell_{;V}(\hat{\theta}) - \ell_{;V}(\theta) \quad \ell_{\lambda;V}(\hat{\theta}_\psi)|}{|\ell_{\theta;V}(\hat{\theta})|} \frac{|j(\hat{\theta})|^{1/2}}{|j_{\lambda\lambda}(\hat{\theta}_\psi)|^{1/2}}$$

Normal circle

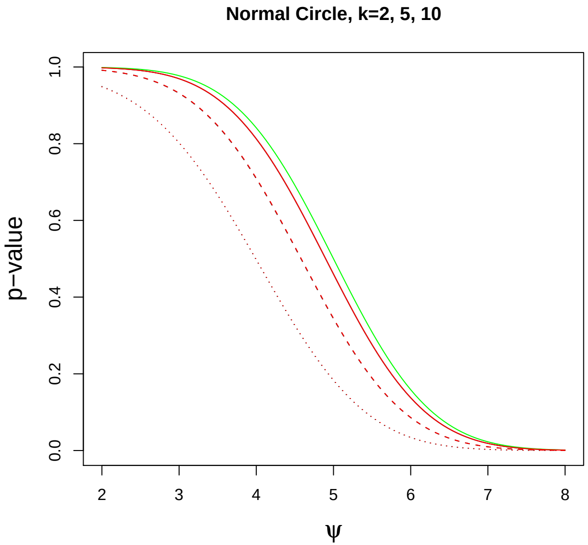
- exact P-value(ψ) $\qquad \Pr\{\chi^2_k(n\psi^2) \geq n||y||^2\}$
- approx:

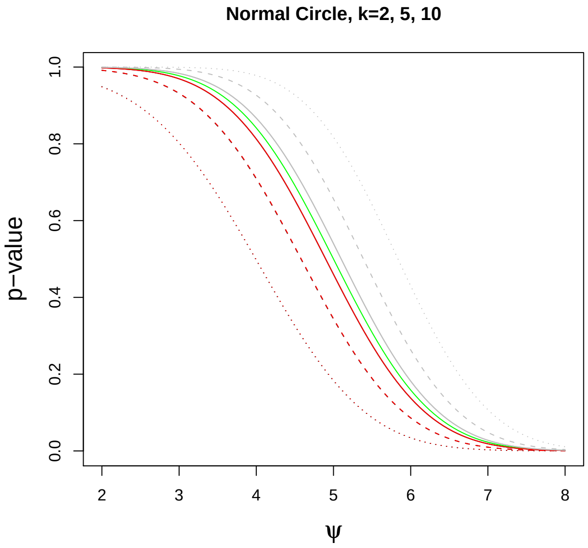
$$r_F^* = \sqrt{n}(\hat{\psi} - \psi) - \frac{1}{\sqrt{n}(\hat{\psi} - \psi)} \log\left\{\left(\frac{\hat{\psi}}{\psi}\right)^{(k-1)/2}\right\} \dot{\sim} N(0,1)$$

- Bayes:

$$r_B^* = \sqrt{n}(\hat{\psi} - \psi) + \frac{1}{\sqrt{n}(\hat{\psi} - \psi)} \log\left\{\left(\frac{\hat{\psi}}{\psi}\right)^{(k-1)/2}\right\} \dot{\sim} N(0,1)$$

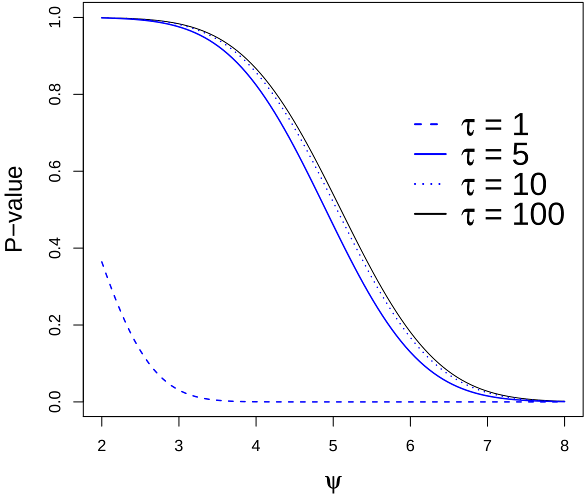
- $r_B^* - r_F^* \dot{\sim} \frac{k-1}{\psi\sqrt{n}}$

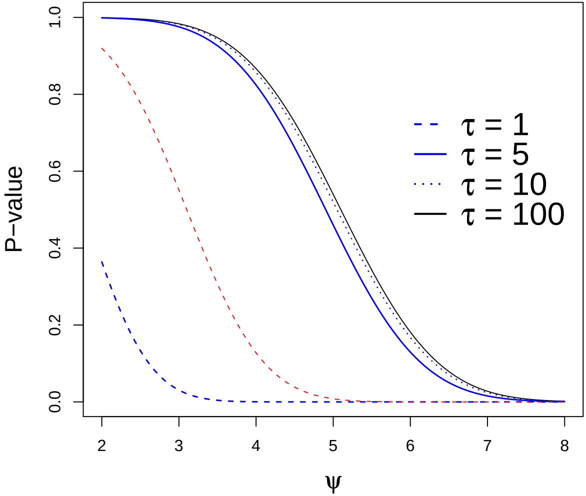

Normal Circle, k=2, 5, 10
p−value
ψ
0.0
0.2
0.4
0.6
0.8
1.0
2
3
4
5
6
7
8

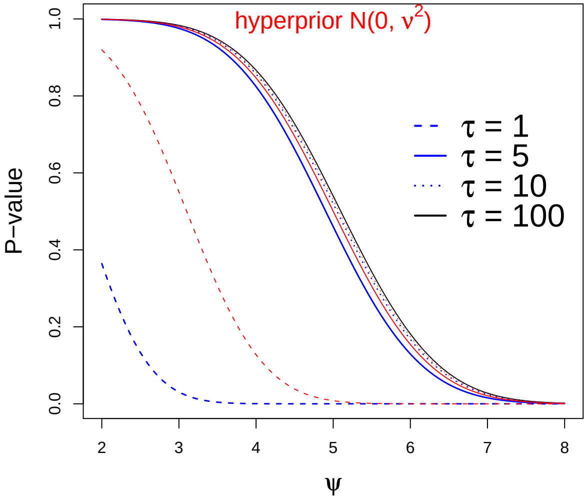

Normal Circle, k=2, 5, 10
p−value
0.0
0.2
0.4
0.6
0.8
1.0
2
3
4
5
6
7
8
ψ

Normal circle: $y_i \sim N(\mu_i, 1) : \psi = ||\mu||$

- $$q_B = -\ell_p'(\psi) j_p(\hat{\psi})^{-1/2} \frac{|j_{\lambda\lambda}(\hat{\theta}_\psi)|^{1/2}}{|j_{\lambda\lambda}(\hat{\theta})|^{1/2}} \frac{\pi(\hat{\theta})}{\pi(\hat{\theta}_\psi)}$$
- $$r_B^* = \sqrt{n}(\hat{\psi} - \psi) + \frac{1}{\sqrt{n}(\hat{\psi} - \psi)} \log \left\{ \left(\frac{\hat{\psi}}{\psi} \right)^{(k-1)/2} \frac{\pi(\hat{\theta})}{\pi(\hat{\theta}_\psi)} \right\}$$
- (i) : $\mu_i \sim N(0, \tau^2)$
  (ii): $\mu_i \sim N(a, \tau^2), \quad a \sim N(0, \nu^2)$


1.0
0.8
0.6
0.4
0.2
0.0
P−value
2
3
4
5
6
7
8
ψ
τ = 1
τ = 5
τ = 10
τ = 100


1.0
0.8
0.6
0.4
0.2
0.0
P−value
2
3
4
5
6
7
8
ψ
τ = 1
τ = 5
τ = 10
τ = 100


hyperprior N(0, ν²)
τ = 1
τ = 5
τ = 10
τ = 100
P−value
0.0
0.2
0.4
0.6
0.8
1.0
2
3
4
5
6
7
8
ψ

Matching priors

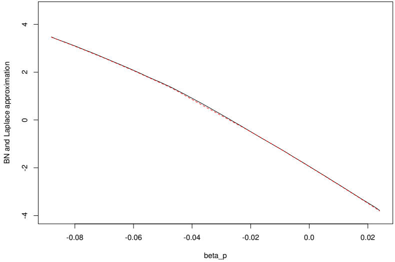
- posterior quantile: $\mathrm{Pr}_{\theta|y}(\theta \geq \theta_\pi^{(1-\alpha)}(y) \mid y) = \alpha$
- matching prior: $\mathrm{Pr}_{Y|\theta}(\theta_\pi^{(1-\alpha)}(Y) \leq \theta) = \alpha + O(n^{-1})$
- θ scalar: $\pi(\theta) \propto i^{1/2}(\theta)$ $\qquad i(\theta) = Ej(\theta) = E\{-\ell''(\theta)\}$
- $\theta = (\psi, \lambda)$: $\pi(\theta) \propto i_{\psi\psi}^{1/2}(\theta) g(\lambda)$ $\qquad i_{\psi\lambda}(\theta) = 0$
- using r_B^*:

$$\frac{\pi(\hat\theta)}{\pi(\hat\theta_\psi)} = \frac{i_{\psi\psi}^{1/2}(\hat\theta) g(\hat\lambda)}{i_{\psi\psi}^{1/2}(\hat\theta_\psi) g(\hat\lambda_\psi)}$$

- invariant to choice of g — A.M. Staicu, 2007

Example: logistic regression

77 observations, 6 covariates, inference for β_6



method	lower bound	upper bound	p-value for 0
$\Phi(r^*)$	-0.058	0.00029	0.052
$\Phi(r_B^*)$	-0.058	0.00028	0.052
1st order	-0.063	-0.00062	0.047

Strong matching

- $s(\psi) = \Phi(r + \frac{1}{r}\log\frac{q_B}{r})$: Bayesian survivor value
- $p(\psi) = \Phi(r + \frac{1}{r}\log\frac{q_F}{r})$: Frequentist *p*-value
- $q_B = -\ell_p'(\psi) j_p(\hat{\psi})^{-1/2} \frac{|j_{\lambda\lambda}(\hat{\theta}_\psi)|^{1/2}}{|j_{\lambda\lambda}(\hat{\theta})|^{1/2}} \frac{\pi(\hat{\theta})}{\pi(\hat{\theta}_\psi)}$
- $q_F = \frac{|\ell_{;V}(\hat{\theta}) - \ell_{;V}(\theta) \quad \ell_{\lambda;V}(\hat{\theta}_\psi)|}{|\ell_{\theta;V}(\hat{\theta})|} \frac{|j(\hat{\theta})|^{1/2}}{|j_{\lambda\lambda}(\hat{\theta}_\psi)|^{1/2}}$
- $q_B = q_F \Leftrightarrow \frac{\pi(\hat{\theta})}{\pi(\hat{\theta}_\psi)} = \ldots$
- default prior along the curve $\theta = \hat{\theta}_\psi$ F&R, 2002

Example: normal circle $y_i \sim N(\mu_i, 1)$

- $$r_F^* = \sqrt{n}(\hat{\psi} - \psi) - \frac{1}{\sqrt{n}(\hat{\psi} - \psi)} \log \left\{ \left(\frac{\hat{\psi}}{\psi} \right)^{(k-1)/2} \right\} \dot{\sim} N(0,1)$$

- $$r_B^* = \sqrt{n}(\hat{\psi} - \psi) + \frac{1}{\sqrt{n}(\hat{\psi} - \psi)} \log \left\{ \left(\frac{\hat{\psi}}{\psi} \right)^{(k-1)/2} \right\} \dot{\sim} N(0,1)$$

- $$r_B^* = r_F^* \iff \pi(\mu) d\mu \propto ||\mu||^{-(k-1)} d\mu \qquad (\psi = ||\mu||)$$

Extending strong matching: scalar parameter

- Location model: $Y \sim f(y - \theta) \Longrightarrow \pi(\theta)d\theta \propto d\theta$
- $$Y \sim f\{y - \beta(\theta)\} \Longrightarrow \pi(\theta)d\theta \propto d\beta(\theta)$$
- there exists a transformation $\beta(\theta)$ such that any model $f(y;\theta)$ is well approximated by a location model
- well approximated:
  - has location form (to $O(n^{-3/2})$)
  - agrees with original model at the data point
  - agrees with the first derivative of the model at the data point
- $\beta(\theta) = \int \frac{\ell'(\theta)}{\ell_{;V}(\theta)} d\theta$

Approximate location models

- Location model: $Y \sim f(y - \theta) \quad \theta \to \theta + d\theta, \quad y \to y + dy$
- $F(y;\theta)$ unchanged, i.e. $dF(y;\theta) = 0$
- General model $Y \sim f(y;\theta)$ require $dF(y^0;\theta) = 0$
- $F_y(y^0;\theta)dy + F_{;\theta}(y^0;\theta)d\theta = 0$ (scalar or vector θ)
- $$dy = -\frac{F_{;\theta}(y^0;\theta)}{F_y(y^0;\theta)}d\theta = V(\theta)d\theta$$
- $y_1, \ldots, y_n : V(\theta) = \begin{bmatrix} V_1(\theta) \\ \vdots \\ V_n(\theta) \end{bmatrix}$ $n \times p$ matrix

Default prior

- $$dy = V(\theta)d\theta$$
- $$\ell_\theta(\hat{\theta}; y) = 0 \Longrightarrow \ell_{\theta\theta}(\hat{\theta}; y)d\hat{\theta} + \ell_{\theta;y}(\hat{\theta}; y)dy = 0$$
- $$d\hat{\theta} = \hat{\jmath}^{-1} H dy$$
- proposed default prior

$$\pi(\theta)d\theta \propto |\hat{\jmath}^{-1} H V(\theta)| d\theta$$

Example $y \sim N(X\beta, \sigma^2)$

- $\theta = (\beta, \sigma)$: length $p = q + 1$
- $V(\hat{\theta}) = \begin{pmatrix} X & \frac{y^0 - X\hat{\beta}}{\hat{\sigma}} \end{pmatrix}$
- *design* *residuals*
- $V(\theta) = \begin{pmatrix} X & \frac{y^0 - X\beta}{\sigma} \end{pmatrix}$
- $\pi(\theta) d\theta \propto d\beta d\sigma / \sigma$

... not calibrated

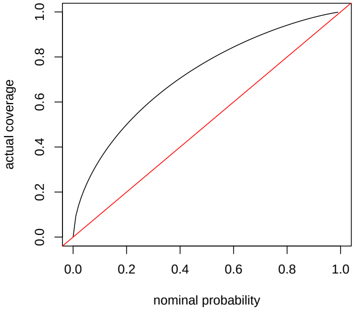
- If parameter of interest is curved, then prior needs to be targetted on the parameter of interest
- marginalization paradox (Dawid, Stone and Zidek)
- proposal: $\pi(\theta)d\theta \propto |\hat{\jmath}^{-1}HV(\theta)|d\theta = J_\theta(\theta)d\theta$, say
- in model with ψ fixed, construct prior for λ (vector) using the same construction
- leading to: $\pi(\lambda \mid \psi)d\lambda \propto J_{\lambda|\psi}(\lambda)d\lambda$
- recalibrate the ratio, so that posterior agrees with adjusted profile log-likelihood
- details ...

... details

- $\pi(\theta)d\theta \propto J_\theta(\theta)d\theta$
- $\pi(\lambda \mid \psi)d\lambda \propto J_{\lambda|\psi}(\lambda)d\lambda$
- adjustment for nuisance information to ensure that
- $\pi_m(\psi \mid y) \propto \exp \ell_A(\psi; y)\pi(\psi)$
- $\ell_A(\psi)$ is an adjusted profile log-likelihood for ψ

Location based priors not calibrated

- Example $y_i \sim N(\mu_i, 1)$, $\psi = ||\mu||$, default prior $\propto d\mu$
- Posterior probability limits for ψ do not have frequentist coverage



... not calibrated

- source: Heinrich, J. (2005) Proceedings of Phystat05
- n events from Poisson with rate $\epsilon s + b$, with s of interest and additional Poisson measurements of b and ϵ
- 'flat' priors for s, ϵ, b

Heinrich, 2005

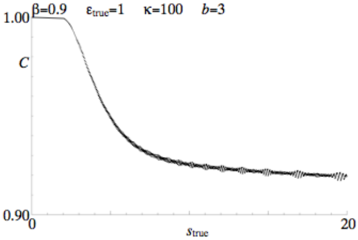


Fig. 2. Typical single channel case. Coverage for 90% credibility level upper limits, acceptance uncertainty = 10%, background uncertainty = zero.

Heinrich, 2005

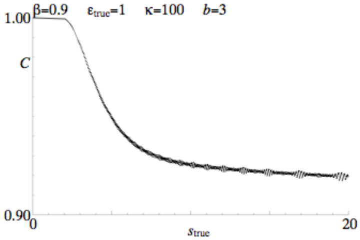


Fig. 2. Typical single channel case. Coverage for 90% credibility level upper limits, acceptance uncertainty = 10%, background uncertainty = zero.

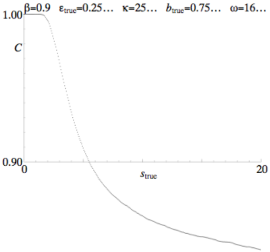


Fig. 7. 4 independent channels. Coverage for 90% credibility level upper limits, acceptance uncertainty = 40%/channel, Background uncertainty = 29%/channel.

... not calibrated

- Cox and Hinkley, 1974, after Mitchell, 1969:
  $E(y \mid x) = \alpha + \beta(1 - x)^\rho, \quad x \in (0, 1)$
- Bayesian probit regression (Jones, 2008; Siddhartha and Chib, 1984): $p(y = 1) = \Phi(\alpha + \beta x_i)$: flat priors on α, β
- hierarchical Poisson models (Gelman et al, 2007):
  $E(y_{ij}) = c_0 x_{ij} \exp(\mu + \alpha_i + \beta_j + \gamma_{ij})$
- "non-informative uniform priors on $\mu, \underline{\alpha}, \sigma_\beta, \sigma_\gamma$"

Conclusions

- calibrated priors are data dependent
- focus motivated by asymptotic theory for likelihood inference
- other asymptotic approaches consider different measures, such as K-L distance from prior to posterior (leading to reference priors)
- always need to target on parameter of interest
- marginalization to curved parameters using flat priors may lead to poorly calibrated inferences

Some references

- Fraser, D.A.S. and Reid, N. (2002). Strong matching of frequentist and Bayesian inference. *J. Statist. Plan. Infer.* **103**, 263–285.
- Staicu, A.M. and Reid, N. (2007). Uniqueness of matching priors. submitted
- Staicu, A.M. (2007). On some aspects of likelihood methods with applications in biostatistics. PhD Thesis, U Toronto.
- Datta, G.S. and Ghosh, M. (1995). Some remarks on noninformative priors. *J. Amer. Statist. Assoc.* **90**, 1357–1363.
- Dawid, A.P., Stone, M., and Zidek, J.V. (1973). Marginalization paradoxes in Bayesian and structural inference. *J. Roy. Statist. Soc.* B **35**, 189–233.
- Bernardo, J.M. (1979). Reference posterior distributions for Bayesian inference. *J. Roy. Statist. Soc.* B **41**, 113–147 (with discussion).