Default priors and model parametrization

Nancy Reid

O-Bayes09, June 6, 2009

Don Fraser, Elisabeta Marras, Grace Yun-Yi

Well-calibrated priors

- model $f(y;\theta), F(y;\theta)$;   log-likelihood $\ell(\theta) = \log f(y;\theta)$
- can we find priors that are guaranteed to be well-calibrated, at least approximately
- what structure do such priors need to have
- can we do this for all component parameters simultaneously

- Welch & Peers, 1963: $\pi(\theta)d\theta \propto i^{1/2}(\theta)d\theta$
  expected Fisher information (matrix) $i(\theta) = E\{-\ell'(\theta)\}$
- Peers, 1965; Tibshirani, 1989 parameter of interest $\theta = (\psi, \lambda)$: $\pi(\theta)d\theta \propto i_{\psi\psi}^{1/2}(\theta)g(\lambda)d\theta$
- matching priors using Edgeworth expansion

Well-calibrated priors

- model $f(y;\theta), F(y;\theta)$; log-likelihood $\ell(\theta) = \log f(y;\theta)$
- can we find priors that are guaranteed to be well-calibrated, at least approximately
- what structure do such priors need to have
- can we do this for all component parameters simultaneously

- Welch & Peers, 1963: $\pi(\theta)d\theta \propto i^{1/2}(\theta)d\theta$
  expected Fisher information (matrix) $i(\theta) = E\{-\ell''(\theta)\}$
- Peers, 1965; Tibshirani, 1989 parameter of interest $\theta = (\psi, \lambda)$: $\pi(\theta)d\theta \propto i_{\psi\psi}^{1/2}(\theta)g(\lambda)d\theta$
- matching priors using Edgeworth expansion

Approximate location models

- Location model: $Y \sim f(y-\theta) \Longrightarrow \pi(\theta)d\theta \propto d\theta$ scalar Y, θ
- Location model: $\theta \to \theta + d\theta, \quad y \to y + d\theta$
- $F(y;\theta)$ unchanged, i.e. $dF(y;\theta) = 0$

- General model: require $dF(y^0;\theta) = 0$ y^0 sample point
- $F_y(y^0;\theta)dy + F_{;\theta'}(y^0;\theta)d\theta = 0$ scalar Y, vector θ
- $$dy = -\frac{F_{;\theta}(y^0;\theta)}{F_y(y^0;\theta)}d\theta = V(\theta)d\theta$$
- sample $y_1, \ldots, y_n$: i.i.d Y_i, vector θ

$$dy_i = V_i(\theta)d\theta$$

Default prior

- $$dy = V(\theta)d\theta; \quad V(\theta) = \begin{bmatrix} V_1(\theta) \\ \vdots \\ V_n(\theta) \end{bmatrix}$$

  $n \times p$ matrix
- possible default prior $\pi(\theta)d\theta \propto |V'(\theta)V(\theta)|^{1/2}d\theta$
- convert to maximum likelihood coordinates
- $$\ell_\theta(\hat{\theta}; y) = 0 \Longrightarrow \ell_{\theta\theta}(\hat{\theta}; y)d\hat{\theta} + \ell_{\theta;y}(\hat{\theta}; y)dy = 0$$
- $$d\hat{\theta} = \hat{\jmath}^{-1}Hdy = \hat{\jmath}^{-1}HV(\theta)d\theta$$

  $\hat{\jmath} = j(\hat{\theta}^0)$: observed Fisher information
- proposed default prior

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

... default prior

- $$\pi(\theta)d\theta \propto |\hat{\jmath}^{-1} H V(\theta)| d\theta$$
- $V(\theta)$ links dy to $d\theta$
- $\hat{j}^{-1} H$ links dy to $d\hat{\theta}$
- gives right invariant prior for transformation parameter in transformation models
- provides extension of right invariant prior to general (continuous) models
- $V(\theta) = \left.\frac{dy}{d\theta}\right|_{y=y^0} = -\left.\frac{F_\theta(y;\theta)}{f(y;\theta)}\right|_{y=y^0}$
- $H = H(y^0; \hat{\theta}^0) = \left.\frac{\partial^2 \ell(\theta; y)}{\partial\theta\partial y}\right|_{\hat{\theta}^0, y^0}$
- $\hat{j} = j(\hat{\theta}^0) = -\ell_{\theta\theta'}(\hat{\theta}^0)$

Example: regression model

- $y_1 = X_1\beta + \sigma\epsilon_1, \ldots, y_n = X_n\beta + \sigma\epsilon_n$
- $V(\theta) = \left.\dfrac{dy}{d\theta}\right|_{y^0} = \{X, (y^0 - X\beta)/(2\sigma)\}$
- $H = \begin{pmatrix} X'/\hat{\sigma}^2 \\ \hat{z}^{0\prime}/\hat{\sigma}^3 \end{pmatrix}$
- $\hat{j} = \mathrm{diag}\{X'X/\hat{\sigma}^2, n/(2\hat{\sigma}^4)\}$
- $$\begin{aligned} W(\theta) &= \left\{ \begin{matrix} I & (X'X)^{-1}X'z^0(\theta)/(2\sigma) \\ 2\hat{z}^{0\prime}\hat{\sigma}X/n & \hat{z}^{0\prime}z^0(\theta)\hat{\sigma}/(n\sigma) \end{matrix} \right\} \\ &= \left\{ \begin{matrix} I & (\hat{\beta}^0 - \beta)/2\sigma^2 \\ 0 & \hat{\sigma}^2/\sigma^2 \end{matrix} \right\}. \end{aligned}$$

... regression

- $y \sim N(X\beta, \sigma^2), \quad \theta = (\beta, \sigma^2)$

- $V(\theta) = \left(X \qquad \dfrac{y^0 - X\beta}{\sigma} \right) \qquad y = X\beta + \sigma\epsilon$

- *design*   *'residuals'*

-

$$\begin{aligned} d\hat{\beta} &= d\beta + (\hat{\beta}^0 - \beta)d\sigma^2/2\sigma^2 \\ d\hat{\sigma}^2 &= \hat{\sigma}^2 d\sigma^2/\sigma^2. \end{aligned}$$

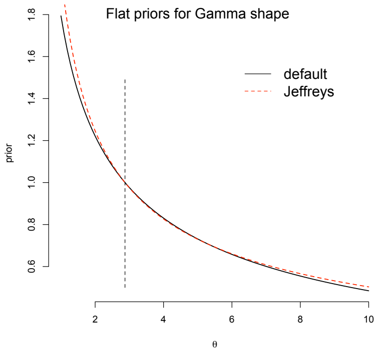
-

$$\pi(\theta)d\theta \propto d\beta d\sigma^2/\sigma^2 \qquad d\beta d\sigma/\sigma$$

- nonlinear regression, $y \sim N(x(\beta), \sigma^2)$, leads to $d\tilde{\beta}d\sigma/\sigma$
- $\tilde{\beta}$ coordinates for tangent plane $\dot{x}(\hat{\beta}^0)(\beta - \hat{\beta}^0)$

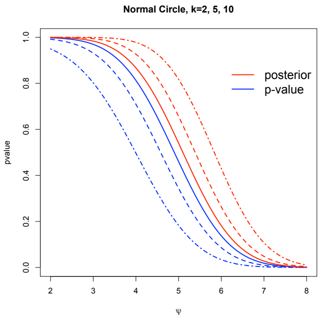
Scalar parameter

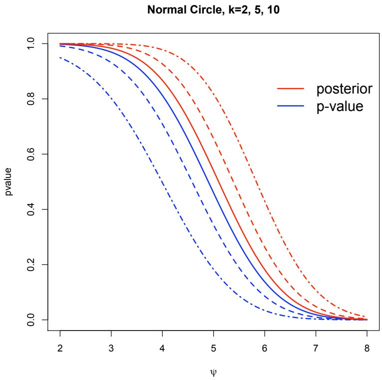
- default prior is data dependent, changes with y^0
- based on approximate local location model
- Jeffreys' prior $\pi_J(\theta) d\theta \propto i^{1/2}(\theta) d\theta$ gives frequentist matching to second order
- depends on model, not data   unconditional

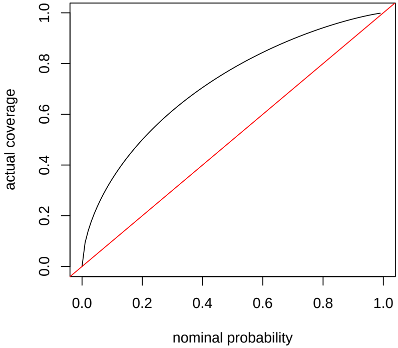


Targetted priors

- If parameter of interest is curved, then prior needs to be targetted on the parameter of interest
- marginalization paradox (Dawid, Stone and Zidek)
- Example: Normal circle
  $y_i \sim N(\mu_i, 1/n), i = 1, \ldots, k; \quad \psi = \sum(\mu_i^2)^{1/2}$ Stein
- Default prior $\pi(\mu)d\mu \propto d\mu$
- Posterior $\chi^2_k(n||y||^2)$ Exact $\chi^2_k(n||\psi||^2)$







Bayes - frequentist $\approx \Phi\left\{\frac{(k-1)}{\psi\sqrt{n}}\right\}$

not fixed by hierarchy of priors

Targetted priors: strong matching

- use Laplace approximations to posterior and to frequentist *p*-value
- structure of approximations makes comparison ‘easy’
- $s(\psi) \doteq \Phi(r + \frac{1}{r}\log\frac{q_B}{r})$: Bayesian survivor value
- $p(\psi) \doteq \Phi(r + \frac{1}{r}\log\frac{q_F}{r})$: Frequentist *p*-value
- $r = r(\psi) = \pm\sqrt{2}\{\ell(\hat\theta) - \ell(\psi, \hat\lambda_\psi)\}$
- q_B contains the prior; q_F various information functions and sample space derivatives
- $q_B = q_F \Leftrightarrow \dfrac{\pi(\hat\theta)}{\pi(\hat\theta_\psi)} = \ldots$
- default prior along the curve $\theta = \hat\theta_\psi$ F&R, 2002
- need to extend to full parameter space

... details

- $q_B = \ell_\psi(\hat\theta_\psi) \dfrac{|j_{\lambda\lambda}(\hat\theta_\psi)|^{1/2}}{|j(\hat\theta)|^{1/2}} \dfrac{\pi(\hat\theta)}{\pi(\hat\theta_\psi)}$
- $q_F = \dfrac{|\ell_{;V}(\hat\theta) - \ell_{;V}(\hat\theta_\psi) \quad \ell_{\lambda;V}(\hat\theta_\psi)|}{|\ell_{\theta;V}(\hat\theta)|} \dfrac{|j(\hat\theta)|^{1/2}}{|j_{\lambda\lambda}(\hat\theta_\psi)|^{1/2}}$
- $\dfrac{\pi(\hat\theta_\psi)}{\pi(\hat\theta)} \propto \dfrac{\ell_\psi(\hat\theta_\psi)|\ell_{\theta;V}(\hat\theta)||j_{\lambda\lambda}(\hat\theta_\psi)|}{|\ell_{;V}(\hat\theta) - \ell_{;V}(\hat\theta_\psi) \quad \ell_{\lambda;V}(\hat\theta_\psi)||j(\hat\theta)|}$
- along the profile curve $\mathcal{C}_\psi = \{\theta : \theta = (\psi, \hat\lambda_\psi)\}$
- based on exponential family approximation at y^0
- use observed information off the curve to extend prior to parameter space
- to get a version that when integrated via Laplace, brings us back to r_f^* approximation
- $\pi(\theta) \propto |j_{(\theta\theta)}(\hat\theta_\psi)|^{1/2} |j_{(\lambda\lambda)}(\theta)|^{1/2}$

... details

- any continuous model can be approximated to $O(n^{-1})$ by an exponential family model (at y^0)
- canonical parameter

$$\varphi(\theta) = \ell_{;V}(\theta; y^0) = \sum \ell_{y_i}(\theta; y^0) V_i(\hat{\theta}^0)$$

- $V(\theta)$ the same matrix as in the default prior
- connection through location model approximation

  $\rightarrow$ ancillarity

  $\rightarrow$ flat prior

Conclusions

- calibrated priors are data dependent
- focus motivated by asymptotic theory for likelihood inference
- reference priors also targetted on parameter of interest
- marginalization to curved parameters using flat priors may lead to poorly calibrated inferences
- hierarchical Poisson models:
  $E(y_{ij}) = c_0 x_{ij} \exp(\mu + \alpha_i + \beta_j + \gamma_{ij})$
- "non-informative uniform priors on $\mu, \underline{\alpha}, \sigma_\beta, \sigma_\gamma$"
- difficulties and opportunities with new large data sets
- checking sensitivity to prior specification can be done simply using asymptotic approximation $\Phi(r_B^*)$
- connections to Empirical Bayes?

Some references

- Fraser, D.A.S., Marras, E., Reid, N. and Yi, G. (2009). Default priors for Bayesian and frequentist inference.
- Fraser, D.A.S. and Reid, N. (2002). Strong matching of frequentist and Bayesian inference. *J. Statist. Plan. Infer.* **103**, 263–285.
- Berger, J.O., Bernardo, J. and Sun, D. (2009).The formal definition of reference priors. *Ann. Statist.* **37**, 905–938.
- 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).