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 \( \theta = (\psi, \lambda) \): \( \pi(\theta)d\theta \propto i^{1/2}_{\psi\psi}(\theta)g(\lambda)d\theta \)
- 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 $\theta = (\psi, \lambda)$: $\pi(\theta)d\theta \propto i^{1/2}(\theta)g(\lambda)d\theta$
- matching priors using Edgeworth expansion
Approximate location models

- Location model: \( Y \sim f(y - \theta) \implies \pi(\theta) d\theta \propto d\theta \)  
  scalar \( Y, \theta \)
- Location model: \( \theta \to \theta + d\theta, \ y \to y + d\theta \)
- \( F(y; \theta) \) 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 \( Y, \theta \)

\[
dy = -\frac{F_{\theta'}(y^0; \theta)}{F_y(y^0; \theta)} d\theta = V(\theta) d\theta
\]

- sample \( y_1, \ldots, y_n \):  
  i.i.d \( Y_i, \theta \)
\[
dy_i = V_i(\theta) d\theta
\]
Default prior

\[ dy = V(\theta) d\theta; \quad V(\theta) = \begin{bmatrix} V_1(\theta) \\ \vdots \\ V_n(\theta) \end{bmatrix} \text{ n × p matrix} \]

- possible default prior \( \pi(\theta) d\theta \propto |V'(\theta)V(\theta)|^{1/2} d\theta \)
- convert to maximum likelihood coordinates

\[ \ell_\theta(\hat{\theta}; y) = 0 \implies \ell_{\theta\theta}(\hat{\theta}; y) d\hat{\theta} + \ell_{\theta; y}(\hat{\theta}; y) dy = 0 \]

\[ d\hat{\theta} = \hat{j}^{-1} H dy = \hat{j}^{-1} H V(\theta) d\theta \]

\( \hat{j} = j(\hat{\theta}^0) \): observed Fisher information

- proposed default prior

\[ \pi(\theta) d\theta \propto |\hat{j}^{-1} H V(\theta)| d\theta \]
... default prior

\[ \pi(\theta) d\theta \propto |\hat{j}^{-1} HV(\theta)| d\theta \]

- \( V(\theta) \) 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

\[ V(\theta) = \frac{dy}{d\theta} \bigg|_{y=y^0} = -\frac{F_\theta(y; \theta)}{f(y; \theta)} \bigg|_{y=y^0} \]

\[ H = H(y^0; \hat{\theta}^0) = \frac{\partial^2 \ell(\theta; y)}{\partial \theta \partial y} \bigg|_{\hat{\theta}^0, y^0} \]

\[ \hat{j} = j(\hat{\theta}^0) = -\ell_{\theta \theta'}(\hat{\theta}^0) \]
Example: regression model

\[ y_1 = X_1 \beta + \sigma \epsilon_1, \ldots, y_n = X_n \beta + \sigma \epsilon_n \]

\[ V(\theta) = \frac{dy}{d\theta}\bigg|_{y_0} = \{X, (y^0 - X \beta)/(2\sigma)\} \]

\[ H = \begin{pmatrix} X'/\hat{\sigma}^2 \\ \hat{z}^0'/\hat{\sigma}^3 \end{pmatrix} \]

\[ \hat{j} = \text{diag}\{X'X/\hat{\sigma}^2, n/(2\hat{\sigma}^4)\} \]

\[
W(\theta) = \begin{cases} 
I & (X'X)^{-1}X'z^0(\theta)/(2\sigma) \\
2\hat{z}^0'\hat{\sigma}X/n & \hat{z}^0'z^0(\theta)\hat{\sigma}/(n\sigma) \\
\end{cases}
\]

\[
= \begin{cases} 
I & (\hat{\beta}^0 - \beta)/2\sigma^2 \\
0 & \hat{\sigma}^2/\sigma^2 \\
\end{cases}.
\]
... regression

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

\( V(\theta) = \begin{pmatrix} X & y^0 - X\beta \end{pmatrix} \begin{pmatrix} \sigma \end{pmatrix} \) \quad y = X\beta + \sigma \epsilon

\( \text{design 'residuals'} \)

\[ d\hat{\beta} = d\beta + (\hat{\beta}^0 - \beta) d\sigma^2 / 2\sigma^2 \]
\[ d\hat{\sigma}^2 = \hat{\sigma}^2 d\sigma^2 / \sigma^2. \]

\[ \pi(\theta)d\theta \propto d\beta d\sigma^2 / \sigma^2 \quad d\beta d\sigma / \sigma \]

nonlinear regression, \( y \sim N(x(\beta), \sigma^2) \), leads to \( d\tilde{\beta}d\sigma / \sigma \)

\( \tilde{\beta} \) coordinates for tangent plane \( \dot{x}(\hat{\beta}^0)(\beta - \hat{\beta}^0) \)

8 / 16
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
  \[ y_i \sim N(\mu_i, 1/n), \ i = 1, \ldots, k; \quad \psi = \sum (\mu_i^2)^{1/2} \]
- Default prior \( \pi(\mu) d\mu \propto d\mu \)
- Posterior \( \chi_k^2(n||y||^2) \)  
  Exact \( \chi_k^2(n||\psi||^2) \)
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’
- \( s(\psi) \triangleq \Phi(r + \frac{1}{r} \log \frac{q_B}{r}) \): Bayesian survivor value
- \( p(\psi) \triangleq \Phi(r + \frac{1}{r} \log \frac{q_F}{r}) \): 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
- \( q_B = q_F \iff \frac{\pi(\hat{\theta})}{\pi(\hat{\theta}_\psi)} = \ldots \)
- default prior along the curve \( \theta = \hat{\theta}_\psi \)  
  F&R, 2002
- need to extend to full parameter space
... details

\[ q_B = \ell_\psi(\hat{\theta}_\psi) \frac{|j_{\lambda\lambda}(\hat{\theta}_\psi)|^{1/2}}{|j(\hat{\theta})|^{1/2}} \frac{\pi(\hat{\theta})}{\pi(\hat{\theta}_\psi)} \]

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

\[ \frac{\pi(\hat{\theta}_\psi)}{\pi(\hat{\theta})} \propto \frac{\ell_\psi(\hat{\theta}_\psi)|\ell_\theta; V(\hat{\theta})||j_{\lambda\lambda}(\hat{\theta}_\psi)|}{|\ell; V(\hat{\theta}) - \ell; V(\hat{\theta}_\psi) \ell_\lambda; V(\hat{\theta}_\psi)||j(\hat{\theta})|} \]

along the profile curve \( C_\psi = \{\theta : \theta = (\psi, \hat{\lambda}_\psi)\} \)

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

\[ \pi(\theta) \propto |j_{(\theta\theta)}(\hat{\theta}_\psi)|^{1/2}|j_{(\lambda\lambda)}(\theta)|^{1/2} \]
any continuous model can be approximated to $O(n^{-1})$ by an exponential family model (at $y^0$)

canonical parameter

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

$V(\theta)$ the same matrix as in the default prior

cannection through location model approximation

→ ancillarity

→ flat prior
Conclusions

- calibrated priors are data dependent
- focus motivated by asymptotic theory for likelihood inference
- reference priors also targeted 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 \( \mu, \alpha, \sigma_\beta, \sigma_\gamma \)”
- 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.


