

Bayesian Model Choice

Hoff Chapter 9, Liang et al 2007, Hoeting et al (1999), Clyde & George (2004) Statistical Science

October 26, 2022

Posterior Distribution

With a Normal-Gamma Prior $\text{NG}(\mathbf{b}_0, \Phi_0, \nu_0, \text{SS}_0)$, the posterior is $\text{NG}(\mathbf{b}_n, \Phi_n, \nu_n, \text{SS}_n)$: with

$$\begin{aligned}\boldsymbol{\beta} \mid \phi, \mathbf{Y} &\sim \text{N}(\mathbf{b}_n, (\phi \Phi_n)^{-1}) \\ \phi \mid \mathbf{Y} &\sim \text{G}\left(\frac{\nu_n}{2}, \frac{\text{SS}_n}{2}\right)\end{aligned}$$

and hyper-parameters

$$\begin{aligned}\Phi_n &= \mathbf{X}^T \mathbf{X} + \Phi_0 \\ \mathbf{b}_n &= \Phi_n^{-1} (\mathbf{X}^T \mathbf{X} \hat{\boldsymbol{\beta}} + \Phi_0 \mathbf{b}_0) \\ \nu_n &= n + \nu_0 \\ \text{SS}_n &= \text{SSE} + \text{SS}_0 + \hat{\boldsymbol{\beta}}^T \mathbf{X}^T \mathbf{X} \hat{\boldsymbol{\beta}} + \mathbf{b}_0^T \Phi_0 \mathbf{b}_0 - \mathbf{b}_n^T \Phi_n \mathbf{b}_n\end{aligned}$$

Marginal Distribution from Normal–Gamma

Theorem

Let $\boldsymbol{\theta} \mid \phi \sim N(m, \frac{1}{\phi}\Sigma)$ and $\phi \sim G(\nu/2, \nu\hat{\sigma}^2/2)$. Then $\boldsymbol{\theta}$ ($p \times 1$) has a p dimensional multivariate t distribution

$$\boldsymbol{\theta} \sim t_{\nu}(m, \hat{\sigma}^2\Sigma)$$

with density

$$p(\boldsymbol{\theta}) \propto \left[1 + \frac{1}{\nu} \frac{(\boldsymbol{\theta} - m)^T \Sigma^{-1} (\boldsymbol{\theta} - m)}{\hat{\sigma}^2} \right]^{-\frac{p+\nu}{2}}$$

Derivation

Marginal density $p(\boldsymbol{\theta}) = \int p(\boldsymbol{\theta} \mid \phi) p(\phi) d\phi$

$$\begin{aligned} p(\boldsymbol{\theta}) &\propto \int |\Sigma/\phi|^{-1/2} e^{-\frac{\phi}{2}(\boldsymbol{\theta}-m)^T \Sigma^{-1}(\boldsymbol{\theta}-m)} \phi^{\nu/2-1} e^{-\phi \frac{\nu \hat{\sigma}^2}{2}} d\phi \\ &\propto \int \phi^{p/2} \phi^{\nu/2-1} e^{-\phi \frac{(\boldsymbol{\theta}-m)^T \Sigma^{-1}(\boldsymbol{\theta}-m) + \nu \hat{\sigma}^2}{2}} d\phi \\ &\propto \int \phi^{\frac{p+\nu}{2}-1} e^{-\phi \frac{(\boldsymbol{\theta}-m)^T \Sigma^{-1}(\boldsymbol{\theta}-m) + \nu \hat{\sigma}^2}{2}} d\phi \\ &= \Gamma((p+\nu)/2) \left(\frac{(\boldsymbol{\theta}-m)^T \Sigma^{-1}(\boldsymbol{\theta}-m) + \nu \hat{\sigma}^2}{2} \right)^{-\frac{p+\nu}{2}} \\ &\propto \left((\boldsymbol{\theta}-m)^T \Sigma^{-1}(\boldsymbol{\theta}-m) + \nu \hat{\sigma}^2 \right)^{-\frac{p+\nu}{2}} \\ &\propto \left(1 + \frac{1}{\nu} \frac{(\boldsymbol{\theta}-m)^T \Sigma^{-1}(\boldsymbol{\theta}-m)}{\hat{\sigma}^2} \right)^{-\frac{p+\nu}{2}} \end{aligned}$$

Marginal Posterior Distribution of β

$$\begin{aligned}\beta \mid \phi, Y &\sim N(\mathbf{b}_n, \phi^{-1} \Phi_n^{-1}) \\ \phi \mid Y &\sim G\left(\frac{\nu_n}{2}, \frac{SS_n}{2}\right)\end{aligned}$$

Let $\hat{\sigma}^2 = SS_n/\nu_n$ (Bayesian MSE)

Then the marginal posterior distribution of β is

$$\beta \mid Y \sim t_{\nu_n}(\mathbf{b}_n, \hat{\sigma}^2 \Phi_n^{-1})$$

Any linear combination $\mathbf{x}^T \beta$

$$\mathbf{x}^T \beta \mid Y \sim t_{\nu_n}(\mathbf{x}^T \mathbf{b}_n, \hat{\sigma}^2 \mathbf{x}^T \Phi_n^{-1} \mathbf{x})$$

has a univariate t distribution with ν_n degrees of freedom

Predictive Distribution

Suppose $Y^* | \beta, \phi \sim N(X^*\beta, I/\phi)$ and is conditionally independent of Y given β and ϕ

What is the predictive distribution of $Y^* | Y$?

$Y^* = X^*\beta + \epsilon^*$ and ϵ^* is independent of Y given ϕ

$$X^*\beta + \epsilon^* | \phi, Y \sim N(X^*b_n, (X^*\Phi_n^{-1}X^{*T} + I)/\phi)$$

$$Y^* | \phi, Y \sim N(X^*b_n, (X^*\Phi_n^{-1}X^{*T} + I)/\phi)$$

$$\phi | Y \sim G\left(\frac{\nu_n}{2}, \frac{\hat{\sigma}^2\nu_n}{2}\right)$$

$$Y^* | Y \sim t_{\nu_n}(X^*b_n, \hat{\sigma}^2(I + X^*\Phi_n^{-1}X^T))$$

Alternative Derivation

Conditional Distribution:

$$\begin{aligned}f(Y^* | Y) &= \frac{f(Y^*, Y)}{f(Y)} \\&= \frac{\iint f(Y^*, Y | \beta, \phi) p(\beta, \phi) d\beta d\phi}{f(Y)} \\&= \frac{\iint f(Y^* | \beta, \phi) f(Y | \beta, \phi) p(\beta, \phi) d\beta d\phi}{f(Y)} \\&= \iint f(Y^* | \beta, \phi) p(\beta, \phi | Y) d\beta d\phi\end{aligned}$$

Requires completing the square/quadratic!

Conjugate Priors

Definition

A class of prior distributions \mathcal{P} for θ is conjugate for a sampling model $p(y \mid \theta)$ if for every $p(\theta) \in \mathcal{P}$, $p(\theta \mid Y) \in \mathcal{P}$.

Advantages:

- ▶ Closed form distributions for most quantities; bypass MCMC for calculations
- ▶ Simple updating in terms of sufficient statistics “weighted average”
- ▶ Interpretation as prior samples - prior sample size
- ▶ Elicitation of prior through imaginary or historical data
- ▶ limiting “non-proper” form recovers MLEs

Choice of conjugate prior?

Jeffreys Prior

Jeffreys proposed a default procedure so that resulting prior would be invariant to model parameterization

$$p(\boldsymbol{\theta}) \propto |\mathcal{I}(\boldsymbol{\theta})|^{1/2}$$

where $\mathcal{I}(\boldsymbol{\theta})$ is the Expected Fisher Information matrix

$$\mathcal{I}(\boldsymbol{\theta}) = -\mathbb{E}\left[\frac{\partial^2 \log(\mathcal{L}(\boldsymbol{\theta}))}{\partial \theta_i \partial \theta_j}\right]$$

Fisher Information Matrix

Log Likelihood

$$\log(\mathcal{L}(\boldsymbol{\beta}, \phi)) = \frac{n}{2} \log(\phi) - \frac{\phi}{2} \|(I - P_X)Y\|^2 - \frac{\phi}{2} (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^T (X^T X) (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})$$

$$\frac{\partial^2 \log \mathcal{L}}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} = \begin{bmatrix} -\phi(X^T X) & -(X^T X)(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) \\ -(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^T (X^T X) & -\frac{n}{2} \frac{1}{\phi^2} \end{bmatrix}$$

$$E\left[\frac{\partial^2 \log \mathcal{L}}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}\right] = \begin{bmatrix} -\phi(X^T X) & 0_p \\ 0_p^T & -\frac{n}{2} \frac{1}{\phi^2} \end{bmatrix}$$

$$\mathcal{I}((\boldsymbol{\beta}, \phi)^T) = \begin{bmatrix} \phi(X^T X) & 0_p \\ 0_p^T & \frac{n}{2} \frac{1}{\phi^2} \end{bmatrix}$$

Jeffreys Prior

Jeffreys Prior

$$\begin{aligned}p_J(\boldsymbol{\beta}, \phi) &\propto |\mathcal{I}((\boldsymbol{\beta}, \phi)^T)|^{1/2} \\&= |\phi(\mathbf{X}^T \mathbf{X})|^{1/2} \left(\frac{n}{2} \frac{1}{\phi^2} \right)^{1/2} \\&\propto \phi^{p/2-1} |\mathbf{X}^T \mathbf{X}|^{1/2} \\&\propto \phi^{p/2-1}\end{aligned}$$

Improper prior $\iint p_J(\boldsymbol{\beta}, \phi) d\boldsymbol{\beta} d\phi$ not finite

Formal Bayes Posterior

$$p(\beta, \phi | Y) \propto p(Y | \beta, \phi) \phi^{p/2-1}$$

if this is integrable, then renormalize to obtain formal posterior distribution

$$\begin{aligned}\beta | \phi, Y &\sim N(\hat{\beta}, (X^T X)^{-1} \phi^{-1}) \\ \phi | Y &\sim G(n/2, \|Y - X\hat{\beta}\|^2/2)\end{aligned}$$

Limiting case of Conjugate prior with $b_0 = 0$, $\Phi = 0$, $\nu_0 = 0$ and $SS_0 = 0$

Jeffreys did not recommend using this Posterior does not depend on dimension p

Independent Jeffreys Prior

- ▶ Treat β and ϕ separately (“orthogonal parameterization”)
- ▶ $p_{IJ}(\beta) \propto |\mathcal{I}(\beta)|^{1/2}$
- ▶ $p_{IJ}(\phi) \propto |\mathcal{I}(\phi)|^{1/2}$

$$\mathcal{I}((\beta, \phi)^T) = \begin{bmatrix} \phi(X^T X) & 0_p \\ 0_p^T & \frac{n}{2} \frac{1}{\phi^2} \end{bmatrix}$$

$$p_{IJ}(\beta) \propto |\phi X^T X|^{1/2} \propto 1$$

$$p_{IJ}(\phi) \propto \phi^{-1}$$

Independent Jeffreys Prior is

$$p_{IJ}(\beta, \phi) \propto p_{IJ}(\beta)p_{IJ}(\phi) = \phi^{-1}$$

Formal Posterior Distribution

With Independent Jeffreys Prior

$$p_{IJ}(\beta, \phi) \propto p_{IJ}(\beta)p_{IJ}(\phi) = \phi^{-1}$$

Formal Posterior Distribution

$$\begin{aligned}\beta \mid \phi, Y &\sim N(\hat{\beta}, (X^T X)^{-1} \phi^{-1}) \\ \phi \mid Y &\sim G((n-p)/2, \|Y - X\hat{\beta}\|^2/2) \\ \beta \mid Y &\sim t_{n-p}(\hat{\beta}, \hat{\sigma}^2 (X^T X)^{-1})\end{aligned}$$

Bayesian Credible Sets $p(\beta \in C_\alpha) = 1 - \alpha$ correspond to frequentist Confidence Regions

$$\frac{\lambda^T \beta - \lambda^T \hat{\beta}}{\sqrt{\hat{\sigma}^2 \lambda^T (X^T X)^{-1} \lambda}} \sim t_{n-p}$$

Zellner's g -prior

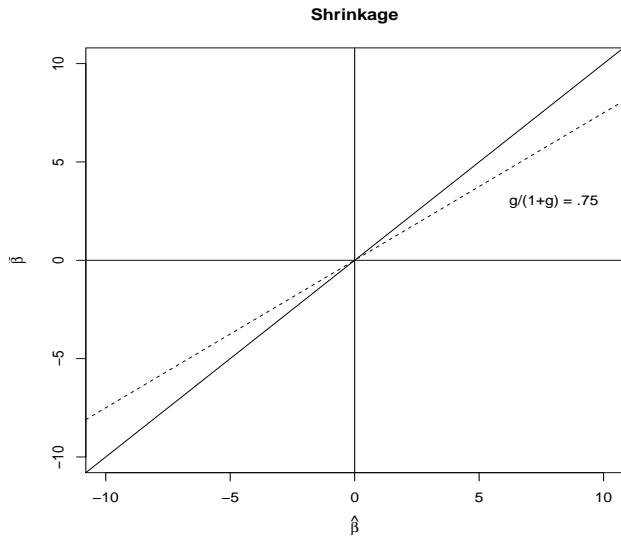
Zellner's g -prior(s) $\beta \mid \phi \sim N(b_0, g(X^T X)^{-1}/\phi)$

$$\beta \mid Y, \phi \sim N\left(\frac{g}{1+g}\hat{\beta} + \frac{1}{1+g}b_0, \frac{g}{1+g}(X^T X)^{-1}\phi^{-1}\right)$$

- ▶ Invariance: Require posterior of $X\beta$ equal the posterior of $XH\alpha$ ($a_0 = H^{-1}b_0$) ($b_0 = 0$)
- ▶ Choice of g ?
- ▶ $\frac{g}{1+g}$ weight given to the data
- ▶ Fixed g effect does not vanish as $n \rightarrow \infty$
- ▶ Use $g = n$ or place a prior distribution on g

Shrinkage

Posterior mean under g -prior with $b_0 = 0$ $\frac{g}{1+g}\hat{\beta}$



Ridge Regression

- ▶ If $X^T X$ is nearly singular, certain elements of β or (linear combinations of β) may have huge variances under the g -prior (or flat prior) as the MLEs are highly unstable!
- ▶ **Ridge regression** protects against the explosion of variances and ill-conditioning with the conjugate prior:

$$\beta \mid \phi \sim N\left(0, \frac{1}{\phi\lambda} I_p\right)$$

- ▶ Posterior for β (conjugate case)

$$\beta \mid \phi, \lambda, Y \sim N\left((\lambda I_p + X^T X)^{-1} X^T Y, \frac{1}{\phi} (\lambda I_p + X^T X)^{-1}\right)$$

- ▶ induces shrinkage as well!

Model Choice ?

- ▶ Redundant variables lead to unstable estimates
- ▶ Some variables may not be relevant ($\beta_j = 0$)
- ▶ Can we infer a "good" model from the data?
- ▶ Expand model hierarchically to introduce another latent variable γ that encodes models \mathcal{M}_γ $\gamma = (\gamma_1, \gamma_2, \dots, \gamma_p)^T$ where

$$\gamma_j = 0 \Leftrightarrow \beta_j = 0$$

$$\gamma_j = 1 \Leftrightarrow \beta_j \neq 0$$

- ▶ Find Bayes factors and posterior probabilities of models \mathcal{M}_γ
- ▶ 2^p models!

Zellner's g-prior

Centered model:

$$Y = 1_n \alpha + X^c \beta + \epsilon$$

where X^c is the centered design matrix where all variables have had their mean subtracted

- ▶ $p(\alpha, \phi) \propto 1/\phi$
- ▶ $\beta_\gamma \mid \alpha, \phi, \gamma \sim N(0, g\phi^{-1}(X_\gamma^c' X_\gamma^c)^{-1})$

which leads to marginal likelihood of γ that is proportional to

$$p(Y \mid \gamma) = C(1 + g)^{\frac{n-p-1}{2}} (1 + g(1 - R_\gamma^2))^{-\frac{(n-1)}{2}}$$

where R^2 is the usual coefficient of determination for model \mathcal{M}_γ .
Trade-off of model complexity versus goodness of fit

Lastly, assign distribution to space of models

Sketch

- ▶ Integrate out β_γ using sums of normals
- ▶ Find inverse of $I_n + gP_{X_\gamma}$ (properties of projections)
- ▶ Find determinant of $\phi(I_n + gP_{X_\gamma})$
- ▶ Integrate out intercept (normal)
- ▶ Integrate out ϕ (gamma)
- ▶ algebra to simplify quadratic forms to R_γ^2

Or integrate α , β_γ and ϕ (complete the square!)

Priors on Model Space

$$p(\mathcal{M}_\gamma) \Leftrightarrow p(\gamma)$$

- ▶ $p(\gamma_j = 1) = .5 \Rightarrow P(\mathcal{M}_\gamma) = .5^p$ Uniform on space of models
 $p_\gamma \sim \text{Bin}(p, .5)$
- ▶ $\gamma_j \mid \pi \stackrel{\text{iid}}{\sim} \text{Ber}(\pi)$ and $\pi \sim \text{Beta}(a, b)$ then $p_\gamma \sim \text{BB}_p(a, b)$

$$p(p_\gamma \mid p, a, b) = \frac{\Gamma(p+1)\Gamma(p_\gamma+a)\Gamma(p-p_\gamma+b)\Gamma(a+b)}{\Gamma(p_\gamma+1)\Gamma(p-p_\gamma+1)\Gamma(p+a+b)\Gamma(a)\Gamma(b)}$$

- ▶ $p_\gamma \sim \text{BB}_p(1, 1) \sim \text{Unif}(0, p)$

Posterior Probabilities of Models

- Calculate analytically under enumeration

$$p(\mathcal{M}_\gamma | Y) = \frac{p(Y | \gamma)p(\gamma)}{\sum_{\gamma' \in \Gamma} p(Y | \gamma')p(\gamma')}$$

Express as a function of Bayes factors and prior odds!

- Use MCMC over Γ - Gibbs, Metropolis Hastings
- Do we need to run MCMC over γ , β_γ , α , and ϕ ?

Inference/Decisions ?