

Handout on Sampling with Conjugate Posteriors

The general source of conjugate priors is the (natural, canonical) exponential family structure often assumed for data. Suppose that a data-vector \mathbf{X} (possibly but not necessarily an iid sample of constituent data-vectors) follows a density of the form

$$f_{\mathbf{X}}(\mathbf{x}|\vartheta) = h(\mathbf{x}) \exp\left(\sum_{j=1}^k T_j(\mathbf{x}) \vartheta_j - A(\vartheta)\right)$$

where the data may fall anywhere that the density factor is positive (a data-value region not depending on the parameter ϑ), and where the parameter space Θ is an open subset of the natural parameter space $\{\vartheta : \int h(\mathbf{x}) \exp(\sum_{j=1}^k T_j(\mathbf{x}) \vartheta_j) d\mathbf{x} < \infty\}$. Then a family of *conjugate* priors for this family of data densities is given by

$$\pi(\vartheta; \tau, \alpha) = k(\vartheta) \exp(\vartheta' \tau - \alpha A(\vartheta))$$

where τ is a parameter vector of the same dimension as ϑ , and the word ‘conjugate’ means that for all observed-data vectors \mathbf{X} , the posterior $f_{\vartheta|\mathbf{X}}(\vartheta|\mathbf{X})$ is a member of the same density family as π . The conjugate property is verified immediately by observing that, as functions of ϑ ,

$$f_{\vartheta|\mathbf{X}}(\vartheta|\mathbf{x}) \propto f_{\vartheta,\mathbf{X}}(\vartheta, \mathbf{x}) \propto k(\vartheta) \exp\left(\vartheta' \{\tau + T(\mathbf{x})\} - (\alpha + 1) A(\vartheta)\right)$$

so that the posterior density (normalized to integrate to 1 with respect to ϑ for each fixed \mathbf{x} vector) belongs to the same family $\pi(\cdot)$, with the parameters (τ, α) replaced by $(\tau + T(\mathbf{X}), \alpha + 1)$.

Important examples of conjugate-prior families created in this way are:

(1). Normal(μ, σ). Here the natural parameters are $\vartheta = (\nu, \rho) \equiv (\mu/\sigma^2, 1/\sigma^2)$, and the conjugate-prior family is

$$\pi(\vartheta) = \rho^{a/2} e^{-b\rho/2} \text{dnorm}(\nu, \mu_0, \sigma_0^2)$$

(2). Binomial(n, p). Here the natural parameter is $\vartheta = \log(\frac{p}{1-p})$, and the conjugate prior family is Beta(a, b).

(3). Poisson(λ). The natural parameter is $\vartheta = \log(\lambda)$, and the conjugate prior family is Gamma(a, b).

(4). Gamma(α, λ). The natural parameter is $\vartheta = (\alpha, \lambda)$, and the conjugate prior has a Gamma(a,b) form if α is fixed and known, but more generally is of the 3-parameter form

$$\pi(\vartheta) = k(\vartheta) e^{a\alpha - b\lambda - c(\log \Gamma(\alpha) - \alpha \ln(\lambda))}$$

These facts are proved and elaborated by computing the parameters of posterior densities from data, as follows:

(1.) If $Y_i \sim \mathcal{N}(\mu, \sigma^2)$ for $i = 1, \dots, n$ are iid given $\vartheta = (\nu, \rho) = (\mu/\sigma^2, 1/\sigma^2)$, then the prior density $\pi(\cdot)$ treating $\rho \sim \Gamma(\alpha/2, \lambda/2)$ and $\nu \sim \mathcal{N}(\mu_0, \sigma_0^2)$ as independent, leads after some algebra to the posterior density $f_{\vartheta|\mathbf{Y}}(\nu, \rho | \mathbf{Y})$

$$\begin{aligned} &= C \rho^{(n+\alpha-2)/2} \exp\left(-\frac{\rho}{2}\left(\lambda + \sum_{i=1}^n Y_i^2\right) + \frac{1}{2}\left(\sum_{i=1}^n Y_i + \frac{\mu_0}{\sigma_0^2}\right)^2 / (n/\rho + 1/\sigma_0^2)\right. \\ &\quad \left. - \frac{1}{2}(n/\rho + \mu_0/\sigma_0^2)\left(\nu - \frac{n\bar{Y} + \mu_0/\sigma_0^2}{n/\rho + 1/\sigma_0^2}\right)^2\right) \end{aligned}$$

There is an important limiting case of this prior-posterior conjugate family, corresponding to the fully *noninformative* (improper) *prior* with $\sigma_0^2 \rightarrow \infty$. In this case, it is easy to check that the (limiting) posterior makes $\rho \sim \Gamma(\frac{1}{2}(n + \alpha + 1), \frac{1}{2}(\lambda + (n-1)S_Y^2))$ and then, given \mathbf{Y} , ρ , conditionally $\nu \sim \mathcal{N}(\rho\bar{Y}, \rho/n)$.

(2.) Based on $Y \sim \text{Binom}(n, p)$ conditionally given p , with $p \sim \text{Beta}(a, b)$, it is easily checked that the posterior density is

$$f_{p|Y}(p|Y) \sim \text{Beta}(a + Y, b + n - Y)$$

(3.) Based on $Y \sim \text{Poisson}(n\lambda)$ conditionally given λ , with $\lambda \sim \Gamma(a, b)$, it is easily checked that the posterior density is

$$f_{\lambda|Y}(\lambda|Y) \sim \Gamma(a + Y, b + n)$$

(4.) Finally, based on a sample $Y_i, 1 \leq i \leq n$, of iid Gamma(α, λ) variates conditionally given (α, λ) , where $\pi(\alpha, \lambda) = C(a, b, c) \exp(a\alpha - b\lambda + c(\alpha \ln(\lambda) - \ln \Gamma(\alpha)))$, we find the posterior to have the same form with the parameters (a, b, c) replaced by $(a + \sum_{i=1}^n \ln Y_i, b + n\bar{Y}, c + n)$.

One might hope that these conjugate prior relationships would lead to tractable posterior densities for linear and generalized-linear regression models involving the distributional families just discussed, but this hope is only partially realized.

First, and most satisfactorily, consider the normal example, with k -vector predictors X_i fixed (with first components 1, in the usual case where the regression relationship includes an intercept term) for $i = 1, \dots, n$. Then assume, conditionally given a p -vector parameter β of regression coefficients and a variance parameter σ^2 , that independent scalar responses are observed, distributed as $Y_i \sim \mathcal{N}(\beta'X_i, \sigma^2)$. Again take $\rho = 1/\sigma^2$, and now define a p -vector transformed parameter by $\nu = \beta/\sigma^2$. Let the prior density of ρ, ν make these parameter components independent, with

$$\rho \sim \Gamma\left(\frac{a}{2}, \frac{\lambda}{2}\right) \quad , \quad \nu \sim \mathcal{N}_{MVN}(\underline{\mu}_0, \Sigma_0)$$

Then algebra of the same sort required to compute example (1) above, leads to a closed form posterior density. The general formula for the posterior is:

$$\begin{aligned} & \rho^{(n+a-2)/2} \exp\left(-\frac{\rho}{2}\left(\lambda + \sum_{i=1}^n Y_i^2\right)\right) \cdot |\det(\rho^{-1} \sum_i X_i^{\otimes 2} + \Sigma_0^{-1})|^{-1/2} \\ & \cdot \exp\left(\frac{1}{2}\left(\sum_{i=1}^n Y_i X_i + \Sigma_0^{-1} \underline{\mu}_0\right)' (\rho^{-1} \sum_i X_i^{\otimes 2} + \Sigma_0^{-1})^{-1} \left(\sum_{i=1}^n Y_i X_i + \Sigma_0^{-1} \underline{\mu}_0\right)\right) \\ & \cdot \mathcal{N}_{MVN}\left(\nu, (\rho^{-1} \sum_i X_i^{\otimes 2} + \Sigma_0^{-1})^{-1} \left(\sum_{i=1}^n Y_i X_i + \Sigma_0^{-1} \underline{\mu}_0\right), (\rho^{-1} \sum_i X_i^{\otimes 2} + \Sigma_0^{-1})^{-1}\right) \end{aligned}$$

In the noninformative-prior limit, $\Sigma_0 \rightarrow \infty$ (in the sense that the minimum eigenvalue converges to $+\infty$), it is readily checked that the posterior density makes ρ given (Y_1, \dots, Y_n) distributed as $\Gamma(\frac{1}{2}(n+a+p), \frac{1}{2}(\lambda + RSS))$, where p is the dimension of the predictors X_i and coefficient vector β , and where $RSS = \sum_{i=1}^n (Y_i - X_i' \hat{\beta})^2$, with $\hat{\beta}$ the least-squares coefficient estimator based on $(Y_i, X_i, i = 1, \dots, n)$. Finally, given the data $\{Y_i\}_{i=1}^n$ and parameter ρ , the posterior conditional density of $\nu = \beta\rho$ is found to be $\mathcal{N}_{MVN}(\rho \hat{\beta}, \rho(\mathbf{X}^{tr} \mathbf{X})^{-1})$, where \mathbf{X} is the $n \times p$ design matrix with i 'th row X_i , where we have use the fact that $\mathbf{X}^{tr} \mathbf{X} = \sum_i X_i^{\otimes 2}$.

In the other examples (2)–(4), one might want respectively to take $\log(p/(1-p))$, $\log(\lambda)$, and $-\log(\lambda)$, for the i 'th observed data value, to have a regression form $\beta'X_i$ with fixed explanatory predictor variables X_i . However, there is no choice of prior density for β in these examples

that gives the resulting parameters p, λ their conjugate prior-posterior form described above. For example in (2), the logistic regression model $\log(p/(1-p)) = \beta' X_i$ with $\beta \sim \mathcal{N}(\underline{\mu}_0, \text{diag}(\sigma_0^2))$ leads to fairly intractable posterior distributions. This does not mean they cannot be used, only that closed form integrations to calculate posteriors are impossible to do with reasonable accuracy and speed.

Concluding Steps on Posterior Predictive Sampling

We provide now the computing formulas for posterior and posterior predictive sampling in our normal-regression example. Here the parameters are $\rho = 1/\sigma^2$, $\nu = \beta/\sigma^2 = \rho\beta$. The formulas given on the previous page imply first that (in terms of the $n \times p$ design matrix \mathbf{X} and $n \times 1$ data vector \mathbf{Y})

$$f_{\rho|\mathbf{Y}}(\rho|\mathbf{Y}) = C \rho^{(n+a-2)/2} \exp\left(-\frac{\rho}{2}(\lambda + \|\mathbf{Y}\|^2)\right) \cdot |\det(\rho^{-1} \mathbf{X}^{tr} \mathbf{X} + \Sigma_0^{-1})|^{-1/2} \\ \cdot \exp\left(\frac{1}{2}(\mathbf{X}^{tr} \mathbf{Y} + \Sigma_0^{-1} \underline{\mu}_0)'(\rho^{-1} \mathbf{X}^{tr} \mathbf{X} + \Sigma_0^{-1})^{-1}(\mathbf{X}^{tr} \mathbf{Y} + \Sigma_0^{-1} \underline{\mu}_0)\right)$$

where C is determined by the requirement that this density integrate to 1 over $\rho \in (0, \infty)$. Next, conditionally given ρ, \mathbf{Y} , the posterior distribution for $\beta = \nu/\rho$ is

$$\beta \sim \mathcal{N}_{MVN}\left((\mathbf{X}^{tr} \mathbf{X} + \rho \Sigma_0^{-1})^{-1}(\mathbf{X}^{tr} \mathbf{Y} + \Sigma_0^{-1} \underline{\mu}_0), \rho^{-2}(\rho^{-1} \mathbf{X}^{tr} \mathbf{X} + \Sigma_0^{-1})^{-1}\right)$$

For any fixed p -vector \mathbf{x}_0 , this allows us for example to conclude

$$\beta' \mathbf{x}_0 \sim \mathcal{N}_{MVN}\left(\mathbf{x}_0'(\mathbf{X}^{tr} \mathbf{X} + \rho \Sigma_0^{-1})^{-1}(\mathbf{X}^{tr} \mathbf{Y} + \Sigma_0^{-1} \underline{\mu}_0), \mathbf{x}_0'(\rho \mathbf{X}^{tr} \mathbf{X} + \rho^2 \Sigma_0^{-1})^{-1} \mathbf{x}_0\right)$$

In any case, conditionally given $\sigma^2 = 1/\rho$, $\beta = \rho\nu$, and \mathbf{Y} , a new posterior predictive dataset \mathbf{Y}^* of observations Y_i^* for $1 \leq i \leq n$ would consist of independent observations depending only on β, σ^2 through $Y_i^* \sim \mathcal{N}(\beta' X_i, \sigma^2)$.

To accomplish posterior simulations, the only slightly tricky step is therefore to calculate and invert the conditional distribution function of ρ given \mathbf{Y} by calculating and inverting numerically

$$F_{\rho|\mathbf{Y}}(\rho|\mathbf{Y}) = \int_{r=0}^{\rho} f_{\rho|\mathbf{Y}}(r|\mathbf{Y}) dr$$

All of these steps are implemented in the log `PredSamp.LR`.