CONJUGAL BAYESIAN (UP)DATING

ROGER KOENKER

"We will all be Bayesians in 2020,and then we can be a united profession." D.V. Lindley's 1995 interview with A.F.M Smith, *Statistical Science*.

"I have lamented that Bayesian statisticians do not stick closely enough to the pattern laid down by Bayes himself: if they would only do as he did and publish posthumously we should all be saved a lot of trouble." [M. Kendall, On the Future of Statistics, JRSS(A), (1968), 131, 182-204].

1. Introduction

The seed from which the great forest of Bayesian statistics grew and was cut down to produce journal articles is the Gaussian mean model with conjugate priors. We briefly review some simple instances of this type before venturing into the "new growth" forest.

Article 18 of the UN Universal Declaration of Human Rights asserts that we are all entitled to our own opinions. However, Bayesians are encouraged to formulate opinions that are conjugal to their models for how observations of the world arise. (Religious dogma usually dictates how we see the world and what we believe about it, so this should hardly be surprising.)

Conjugal prior opinions have the property that they are capable of procreation when suitably coupled with associated probabilistic expressions of the likelihood of observed data. Offspring so generated must take the same form as the parental prior opinion, otherwise they are regarded as sterile, like a mule. We now illustrate this notion of mathematico-sexual compatibility with a serious of conjugal examples.

Example 1. Gaussian mean, known scale. Suppose we have observations $X_1, \cdots, X_n$ drawn iid from a $\mathcal{N}(\mu, \sigma^2)$ model with unknown μ and known σ^2. If our prior opinion about μ happens to be expressed as $\mathcal{N}(\mu_0, \sigma_0^2)$ then a simple computation reveals that our updated opinion about μ should be $\mathcal{N}(\mu_1, \sigma_1^2)$ where $\sigma_1^2 = (n\sigma^{-2} + \sigma_0^{-2})^{-1}$ and $\mu_1 = \sigma_1^2(n\bar{X}\sigma^{-2} + \sigma_0^{-2}\mu_0)$. If, to make things even more convenient, $\sigma_0^2 = \sigma^2$, then the prior acts just like we have one more observation, $X_{n+1} = \mu_0$. But more generally $\bar{X}$ and μ_0 get weighted according to their respective precisions, n/σ^2 for $\bar{X}$ and σ_0^{-2} for μ_0 and combined accordingly. A curious feature of this schema is that if $\bar{X}$ and μ_0 differ considerably relative to their respective precisions, then the updated posterior opinion may express the puzzling

Version: March 23, 2011. The material surveyed here is available from many sources, Appendix C of Jackman (2009) provides an especially nice collection of detailed proofs. These notes are part of an ongoing project I'm calling "Bayes for Those Who Hate Bayes," or more compactly: "Bayes in the Baño."

view that we now believe that μ lies, with high probability, in a region that neither the observations, nor the prior, suggest is very plausible.

The proof of this result relies on the following elementary algebraic identity:

$$\gamma_0(x-\mu_0)^2+\gamma_1(x-\mu_1)^2=(\gamma_0+\gamma_1)\left(x-\frac{\gamma_0\mu_0+\gamma_1\mu_1}{\gamma_0+\gamma_1}\right)^2+\frac{\gamma_0\gamma_1}{\gamma_0+\gamma_1}(\mu_1-\mu_0)^2.$$

Example 2. Gaussian mean, unknown scale I. Suppose that we don't presume to know σ^2 in the previous example, what then? There are many ways to express our prior ignorance about the joint distribution of μ and σ^2, but of these only a few are conjugal. A curious aspect of all this is that we are about to replace our admitted ignorance about σ^2, with a very precise expression of its probability distribution. The simplest of these conjugal priors is probably the joint density, $f(\mu,\sigma^2)=g(\mu|\sigma^2)h(\sigma^2)$ with g the conditional Gaussian density

$$g(\mu|\sigma^2)=\sqrt{n_0}\phi(n_0(\mu-\mu_0)^2/\sigma^2)/\sigma$$

and σ^2 as inverse gamma, invΓ, with density,

$$h(\sigma^2)=K(\sigma^2)^{-(\nu_0+2)/2}\exp\{-\nu\sigma_0^2/(2\sigma^2)\}.$$

This formulation of the prior essentially asserts that our prior belief about μ is based on having already seen n_0 observations from exactly the same process generating the X_i with μ_0 taken to be the sample mean from this "precognition." Cf. Spielberg (2002) Under these circumstances Bayesian mating of prior and likelihood yield the normal/inverse gamma posterior progeny:

$$\begin{aligned}\mu|\sigma^2,X &\sim \mathcal{N}(\mu_1,\sigma^2/n_1)\\ \sigma^2|X &\sim \text{invΓ}(\nu_1/2,\nu_1\sigma_1^2/2),\end{aligned}$$

where $\mu_1=(n_0\mu_0+n\bar{X})/(n_0+n)$, $n_1=n_0+n$, $\nu_1=\nu_0+n$, $\nu_1\sigma_1^2=\nu_0\sigma_0^2+S+n_0n(\mu_0-\bar{X})/n_1$, and $S=\sum(X_i-\bar{X})^2$. The proof of this involves some fairly tedious algebra.

Note that this posterior no longer has a fixed scale; our prior belief revealed that we were uncertain about σ^2 and after seeing the data we are still uncertain. If we are interested in the posterior marginal distribution of μ it is easily approximated by simulation: draw σ^2's from the specified inverse gamma distribution and plug them into the expressions for μ_1 and σ_1^2 to draw Gaussians. Now mumble the magical Bayesian incantation: *Simulation is Revelation.*

A limiting case of the foregoing takes $\nu_0=0$, in which case $\sigma_1=S/n$ and we are back to something closely resembling classical likelihood based results. However, $\nu_0=0$, while yielding a proper inverse gamma posterior, is not itself proper, and therefore does not satisfy the conditions set forth above for conjugality.

Note also that in contrast to the previous example, the scale parameter of the posterior for σ^2 has a term involving $(\mu-\bar{X})^2$, so when the prior and the data disagree the uncertainty about both μ and σ^2 are increased.

Example 3. Gaussian mean, unknown scale II. Suppose instead of prior beliefs about μ depending on σ^2 as in the previous example, we were to assume that μ and σ^2

were stochastically independent with $\mu \sim \mathcal{N}(\mu_0, \sigma_0^2)$ and $\sigma^2 \sim \text{invΓ}(\nu_0/2, \nu\sigma_0^2/2)$. With this prior, Bayesian copulation yields the posterior,

$$\begin{aligned} \mu|\sigma^2, X &\sim \mathcal{N}(\mu_1, \sigma_2^2/n_1) \\ \sigma^2|X &\sim \text{invΓ}(\nu_1/2, \nu_1\sigma_1^2/2), \end{aligned}$$

where $\nu_1 = \nu_0 + n$, $\mu_1 = \sigma_1^2(n\sigma^{-2}\bar{X} + \sigma_0^{-2}\mu_0)$, $\sigma_2^2 = (n\sigma^{-2} + \sigma_0^{-2})^{-1}$, and $S = \sum(X_i - \bar{X})^2$. Note that when the prior takes this independent form, the gap between μ_0 and $\bar{X}$ is no longer contributing to the scale of the posterior for σ^2.

Example 4. Gaussian Regression, unknown scale I. Expanding slightly on the earlier Example 2, suppose that we now have observations $Y_1, \cdots, Y_n$ that are multivariate Gaussian, $\mathcal{N}(X\beta, \sigma^2 I_n)$ where X is a matrix of full column rank, p. We now need a multivariate Gaussian prior for the p-vector, β, which we will take to be $\mathcal{N}(\beta_0, \sigma^2\Omega_0)$ and retain the inverse gamma prior for σ^2. The posterior becomes,

$$\sigma^2|Y \sim \text{invΓ}(\nu_1/2, \nu_1\sigma_1^2/2),$$

but now,

$$\beta|\sigma^2 Y \sim \mathcal{N}(\beta_1, \sigma^2\Omega_1)$$

where $\beta_1 = \Omega_1(\Omega_0^{-1}\beta_0 + X^\top X\hat{\beta})$, $\Omega_1 = (\Omega_0^{-1} + X^\top X)^{-1}$, $\nu_1 = \nu_0 + n$, $\nu_1\sigma_1^2 = \nu_0\sigma_0^2 + S + R$, $\hat{\beta} = (X^\top X)^{-1}X^\top Y$, $S = \|Y - X\hat{\beta}\|^2$, and $R = \|\beta_0 - \hat{\beta}\|^2_{\Omega_1}$. This matrix weighted average is a natural generalization of the earlier Example 2.

Example 5. Gaussian Regression, unknown scale II. Similarly we can extend Example 3 to the regression setting. With independent priors, $\beta \sim \mathcal{N}(\beta_0, \Omega_0)$ and $\sigma^2 \sim \text{invΓ}(\nu_0/2, \nu_0\sigma_0^2/2)$, Bayesian updating yields the posterior,

$$\begin{aligned} \sigma^2|Y &\sim \text{invΓ}(\nu_1/2, \nu_1\sigma_1^2/2), \\ \beta|\sigma^2 Y &\sim \mathcal{N}(\beta_1, \sigma^2\Omega_1) \end{aligned}$$

with $\nu_1 = \nu_0 + n$, $\nu_1\sigma_1^2 = \nu_0\sigma_0^2 + S$, $\beta_1 = \Omega_1(\Omega_0^{-1}\beta_0 + \sigma^{-2}X^\top X\hat{\beta})$, and $\Omega_1 = (\Omega_0^{-1} + \sigma^{-2}X^\top X)^{-1}$.

Such examples can be expanded *ad nauseum* typically continuing with Student t priors for $\hat{\beta}$, but we will resist the temptation to move in this direction.

References

Jackman, S. (2009): *Bayesian Analysis for the Social Sciences*. Wiley, New York.
Spielberg, S. (2002): "Minority Report," 20th Century Fox.