

Some methodological remarks on conjugate priors, approximation of densities, linear models and hierarchical models

Abram Kagan

Dept. of Mathematics, University of Maryland, College Park

May 1 , 2008

How small can a family of conjugate priors be?

Let (x_1, \dots, x_n) be a sample from Poisson population $\text{Po}(\lambda)$ with λ as a parameter and let \mathcal{P} be a family of conjugate prior densities.

Suppose $\pi(\lambda) \in \mathcal{P}$. The posterior density depends on the data only through the sum $S = x_1 + \dots + x_n$ due to the sufficiency of the latter and is proportional to

$$\pi(\lambda)e^{-n\lambda}\lambda^S.$$

Thus, if $\pi(\lambda) \in \mathcal{P}$, then $C(S)\lambda^S e^{-n\lambda}\pi(\lambda) \in \mathcal{P}$ where $C(S)$ is the normalizing constant.

The sum S takes values in $\{0, 1, 2, \dots\}$. Suppose now that for a nonnegative integer M , the prior is $C(M)\lambda^M e^{-n\lambda}\pi(\lambda)$. The posterior is then proportional to $\lambda^{M+S} e^{-2n\lambda}\pi(\lambda)$ showing that if \mathcal{P} is a closed family of conjugate priors, then $\pi \in \mathcal{P}$ implies that

$$C(M)\lambda^M e^{-kn\lambda}\pi(\lambda) \in \mathcal{P}, \quad M = 0, 1, 2, \dots; \quad k = 1, 2, \dots$$

If one chooses

$$\pi(\lambda) = C(\alpha)\lambda^{\alpha-1} e^{-\beta\lambda}$$

with positive **integers** α and β , then the parameters of the posterior gamma distributions also will be positive integers. But since people prefer to allow arbitrary positive values of α , β , so are the parameters of the posterior.

Approximation of parametric families

As Morris notes, “the ADM can be used to fit any chosen Pearson family of distributions.” Let me remind that the Pearson families of densities $f(x)$ were defined by the relation

$$\frac{f'(x)}{f(x)} = \frac{a_0 + a_1x}{b_0 + b_1x + b_2x^2}.$$

One gets different families of densities depending on the discriminant of the quadratic polynomial in the denominator and some other characteristics of the coefficients. It's worthwhile to emphasize that the Pearson families are defined by logderivatives of the density, a characteristic of parametric families of densities almost as popular as the likelihood and, if the derivative is taken with respect to the parameter, known as the *Fisher score*.

Let $p(x; \theta)$ be a family of densities parameterized by a scalar or vector-values parameter θ . The nature of x is arbitrary, i. e., an observation takes value in an arbitrary (measurable) space $(\mathcal{X}, \mathcal{A})$. Let

$$J(x; \theta) = \partial \log p(x; \theta) / \partial \theta = p'(x; \theta) / p(x; \theta)$$

be the Fisher score (prime stands for differentiation with respect to θ ; for vector-values θ the Fisher score becomes a vector and the right hand side the gradient) and let

$$\phi_0(x) \equiv 1, \phi_1(x), \dots, \phi_m(x)$$

be functions of x (statistics) having finite second moments. Let us approximate $J(x; \theta)$ with linear combinations of $\phi_0, \phi_1, \dots, \phi_m$,

$$J(x; \theta) \rightarrow \tilde{J}(x; \theta) = \sum_0^m \lambda_i(\theta) \phi_i(x)$$

where $\lambda_0(\theta), \dots, \lambda_m(\theta)$ are minimizers with respect to $\lambda_1, \dots, \lambda_m$ of

$$\text{var}_\theta \left\{ J(x; \theta) - \sum_0^m \lambda_i \phi_i(x) \right\}.$$

The minimizing coefficients $\lambda_0(\theta), \dots, \lambda_m(\theta)$ depend only on the first and second moments of ϕ_0, \dots, ϕ_m ,

$$\int \phi_i(x)p(x; \theta)dx, \int \phi_i(x)\phi_k(x)p(x; \theta)dx, i, k = 0, 1, \dots, m.$$

In statistical terms, the above approximation means replacing the family $\{p(x; \theta)\}$ with an exponential family, though not directly, by approximating $p(x; \theta)$ with

$$\exp \sum_0^{m+1} \Lambda_i(\theta)\phi_i(x), \Lambda_{m+1}(\theta) \equiv 1,$$

but through the logderivative.

If now (x_1, \dots, x_n) is a sample from $p(x; \theta)$, the equation

$$\sum_1^n \tilde{J}(x_i; \theta) = 0$$

has all the properties of the maximum likelihood equation.

Exponential families play the role of the least favorable: they minimize the Fisher information on the parameter in the class of families with a given structure of the first and second moments of statistics ϕ_1, \dots, ϕ_m .

Linear models and Cramér-Rao inequality

Here I want to show a connection between the Cramér-Rao classical inequality and linear models. More precisely, the inequality turns out a direct corollary of (i) monotonicity of the Fisher information and (ii) a lower bound for the information in linear models.

We are in the setup of the previous section when an observation X has a density $p(x; \theta) = p(x; \theta_1, \dots, \theta_s)$ depending on an s -variate parameter θ . Assuming the Fisher vector score

$$\mathbf{J} = \begin{pmatrix} J_1 \\ \vdots \\ J_s \end{pmatrix}, \quad J_r = \frac{\partial \log p(x; \theta)}{\partial \theta_r}, \quad r = 1, \dots, s$$

well defined with the covariance matrix (assumed positive definite)

$$I_X(\theta) = E_\theta(\mathbf{J}\mathbf{J}')$$

called the matrix of Fisher information on θ in X .

We need the following two facts:

- *Monotonicity of $I(\theta)$.* If $S = S(X)$ is a statistic then

$$I_S(\theta) \leq I_X(\theta). \quad (1)$$

- *Lower bound for $I(\theta)$ in linear models.* For a linear model

$$Y = A\theta + e$$

with an $(n \times s)$ design matrix A , $E(e) = 0$ and the covariance matrix $V(e) = E(e'e) = V$ assumed positive definite,

$$I_Y(\theta) \geq A'V^{-1}A.$$

Now let us treat an unbiased estimator $\tilde{\theta}(X)$ of θ with positive definite covariance matrix $V_{\tilde{\theta}}$ as a new observation Y . Since $E_{\theta}(\tilde{\theta}) = \theta$, $Y = \tilde{\theta}(X)$ follows a general linear model with $A = I_s$, the $(s \times s)$ identity matrix. Thus,

$$I_{\tilde{\theta}}(\theta) \geq V_{\tilde{\theta}}^{-1}(\theta).$$

On the other side, since $\tilde{\theta}(X)$ is a statistic, monotonicity implies

$$I_X(\theta) \geq I_{\tilde{\theta}}(\theta).$$

Thus, the multivariate Cramér-Rao inequality is obtained from

$$I_X(\theta) \geq V_{\tilde{\theta}}(\theta)$$

by taking the inverse.

If Y has an n -variate normal distribution $N_n(A\theta, V)$ and the covariance matrix V does not depend on θ , the lower bound in (ii) is attained. Dependence of the covariance matrix on the regression parameter θ increases the information in Y and, at least in principle, can be used for reducing the covariance matrix of the LSE.

Two comments on hierarchical Bayes models

In hierarchical Bayes models the pdf of an (observable) random element X is determined by an (unobservable) parameter θ which is assumed a random variable: given θ , the conditional pdf of X is $f(x|\theta)$ while the pdf of θ is determined by another (also unobservable) parameter η . Namely, given η , the pdf of θ is $g(\theta|\eta)$. Assume the (prior) pdf $\pi(\eta)$ of η such that $E(|\eta|) < \infty$.

The intuitive relation

$$E(\eta|X = x) = E\{E(\eta|\theta)|X = x\}$$

does not hold true in general.

The setup discussed in Alan Zaslavsky's talk and in one of Carl Morris' papers:
 $\theta_1, \dots, \theta_n$ iid $N(\mu, \sigma^2)$ and observations

$$X_i = \theta_i + \epsilon_i$$

where ϵ 's are iid and independent of θ 's.

For frequentists, the problem here is of robust estimation of μ from

$$X_i = \mu + \epsilon_i + \xi_i$$

with ϵ 's representing the main source of noise with known distribution and ξ 's being a small perturbation.

The information on μ in an individual observation is

$$I_X(\mu)I_\epsilon(\mu) - \sigma^2(\xi) + E[(J'_\epsilon)^2] + \dots$$

showing that here the Gaussian distribution is the most favorable.