smfd19.tex
Day 19. 19.12.2014.

Exponential families (continued).

Now efficiency is not a Bayesian concept (it looks at the distribution of the statistic, so at values we could have seen but didn't, not just at the likelihood), nor is unbiasedness (for the same reason). However, sufficiency is important in Bayesian statistics also (above), as are exponential families.

First, we generalise the exponential family approach to cover several parameters and several sufficient statistics: call $f(x|\theta)$ a member of the *k-parameter exponential family* if

$$f(x|\theta) = \exp\{\Sigma_1^k A_j(\theta)B_j(x) + C(x) + D(\theta)\}.$$

Then by the Fisher-Neyman Factorisation Criterion, $B_1(x), \cdots, B_k(x)$ are sufficient statistics for the k parameters $A_1(\theta), \cdots, A_k(\theta)$. Suppose the prior is of the form

$$f(\theta) = f(\theta; a_1, \cdots, a_k, d) = \exp\{\Sigma_1^k a_j A_j(\theta) + dD(\theta) + c(a_1, \cdots, a_k, d)\}.$$

Then the posterior $f(\theta|x) \propto f(\theta)f(x|\theta)$, i.e. to

$$\exp\{\Sigma_1^k A_j(\theta)(a_j + B_j(x)) + (d+1)D(\theta)\},$$

i.e. to

$$f(\theta; a_1 + B_1(x), \cdots, a_k + B_k(x); d+1).$$

This is a $(k+1)$-dimensional exponential family. Its importance is that if the prior belongs to this family, so too does the posterior: the family is *closed under sampling*. This property is of crucial importance, partly because it is so mathematically convenient, partly because it shows up the structure of the problem. For instance, we shall return below to two of the examples we met in VII.2, where the relationship between prior and likelihood can now be seen in this light to be natural. The prior above is called the *natural conjugate family* to the exponential family above.

Example 1. *Bernoulli distribution.* For $x = 0, 1$,

$$f(x|\theta) = \theta^x(1-\theta)^{1-x} = \Big(\frac{\theta}{1-\theta}\Big)^x(1-\theta) = \exp\{x\log\Big(\frac{\theta}{1-\theta}\Big) + \log(1-\theta)\} :$$

here $k=1, A_1(\theta)=\log\Big(\frac{\theta}{1-\theta}\Big), B_1(x)=x, C(x)=0, D(\theta)=\log(1-\theta)$. The natural conjugate family is

$$\begin{aligned} f(\theta;a_1,d) &= \exp\{a_1A_1(\theta)+dD(\theta)+c(a_1,d)\} \\ &= \exp\{a_1\log\Big(\frac{\theta}{1-\theta}\Big)+d\log(1-\theta)+c(a_1,d)\} \\ &= \theta^{a_1}(1-\theta)^{d-a_1}\exp\{c(a_1,d)\}, \end{aligned}$$

which is *Beta* $B(a_1, d-a_1)$ as in VII.2.

2. *Normal distribution*, $N(\mu,\sigma^2)$: $\theta=(\mu,\sigma^2)$,

$$f(x|\theta)=\exp\{-\frac{1}{2}\frac{x^2}{\sigma^2}+\frac{x\mu}{\sigma^2}-\frac{1}{2}\frac{\mu^2}{\sigma^2}-\log\sigma-\frac{1}{2}\log 2\pi\},$$

$k=2, A_1(\theta)=1/\sigma^2, B_1(x)=-\frac{1}{2}x^2, A_2(\theta)=\mu/\sigma^2, B_2(x)=x, C(x)=0, D(\theta)=-\frac{1}{2}[\log(2\pi\sigma^2)+\mu^2/\sigma^2]$. The natural conjugate family is

$$f(\theta;a_1,a_2,d)=\exp\{a_1A_1(\theta)+a_2A_2(\theta)+dD(\theta)+c(a_1,a_2,d)\}$$

$$\propto(\sigma^2)^{-\frac{1}{2}d}\exp\{\frac{a_1}{\sigma^2}+\frac{a_2\mu}{\sigma^2}-\frac{1}{2}d\mu^2\sigma^2\}.$$

The exponent is σ^2 times

$$-\frac{1}{2}d(\mu^2-\frac{2a_2\mu}{d}+a_1)=-\frac{1}{2}d[(\mu-\frac{a_2}{d})^2-a_1-\frac{{a_2}^2}{d^2}].$$

Writing $m:=a_2/d, b:=-a_1-{a_2}^2/2d$,

$$f(\theta;a_1,a_2,d)\propto(\sigma^2)^{-\frac{1}{2}d}\exp\{-\frac{1}{2}d(\mu-m)^2/\sigma^2-b/\sigma^2\}.$$

For σ known, this is a normal prior for μ, as in VII.2. With neither σ nor μ known (both parameters), this is the natural conjugate prior to the normal $N(\mu,\sigma^2)$. More generally, one can work with $(\sigma^2)^{-t}$ in place of $(\sigma^2)^{-\frac{1}{2}d}$. Here m, d, b (and t if present) are *hyperparameters* for the parameters μ, σ.

6. *Shrinkage* [O'H] 6.42, p. 159].

In the Bayesian paradigm the posterior gives a compromise between prior and likelihood. This 'pulls' the likelihood towards the prior, so 'pulls' a classical estimate towards a prior estimate. Similarly with several parameters. It

is thus typical of the Bayesian paradigm that estimators are less spread out than in the classical paradigm, a phenomenon known as *shrinkage*. Similar shrinkage effects occur in higher dimensions – the *James-Stein phenomenon*.

7. *Invariance and Jeffreys priors.*

Suppose we work with a parameter θ, with information per reading $I(\theta) = E[(\ell'(\theta)^2] = \int ((\log f)_\theta)^2 f(\theta)$. If we reparametrise to $\phi := g(\theta)$, then as $\partial/\partial\phi = (d\theta/d\phi)(\partial/\partial\phi)$,

$$I(\phi) = (d\theta/d\phi)^2 I(\theta).$$

The idea of choosing a prior which is large where the information is large is very attractive (and reminiscent of maximum-likelihood estimation!). Jeffreys suggested choosing a prior of the form

$$\pi(\theta) \propto \sqrt{I(\theta)}$$

– the square root to make the prior *invariant under reparametrisation*:

$$\pi(\phi)d\phi \propto \sqrt{I(\phi)}d\phi = \sqrt{I(\theta)}d\theta \propto \pi(\theta)d\theta : \qquad \pi(\phi)d\phi = \pi(\theta)d\theta$$

(both sides integrate to 1, so we can take equality here). There is an extension to higher dimensions, using the Fisher information matrix and the square root of the modulus of its determinant.

Bayesian procedures are in general not invariant under reparametrisation! This can be seen as a drawback, but Bayesians argue that one needs to incorporate a loss function (or utility function), and one should seek a parametrisation that suits this loss function.

Note. Sir Harold JEFFREYS (1891-1989) was primarily a geophysicist, and wrote an influential book *The Earth: Its Origin, History and Physical Constitution*, 1924[1]. He was also a pioneer of Bayesian statistics, and wrote an early book on it, *Theory of probability* (1st ed. 1939, 2nd ed. 1960, 3rd ed. 1983). He also wrote (with his wife) 'Jeffreys and Jeffreys', *Methods of mathematical physics*, CUP, 1946.

8. *The Bayes linear estimator.*

If $d(x)$ is a *linear* function, $a + b'z$, where $z = z(x)$ and b are vectors, the

[1]Jeffreys was the first to suggest that the earth's core is liquid – but he was a strong opponent of continental drift!

quadratic loss is

$$\begin{aligned} D &= E[(a + b'z - \theta)^2] \\ &= E[a^2 + 2ab'z + b'zz'b - 2a\theta - 2b'z\theta + \theta^2] \\ &= a^2 + 2ab'Ez + b'E(zz')b - 2aE\theta - 2b'E(z\theta) + E(\theta^2). \end{aligned}$$

Add and subtract $[E(\theta)]^2$, $(b'Ez)^2 = b'EzEz'b$ and $2b'EzE\theta$. Write $V := var\ z = E(zz') - EzEz'$ for the covariance matrix of z, $c := cov(\theta, z) = E(z\theta) - EzE\theta$ for the covariance vector between θ and the vector z.

$$D = (a + b'Ez - E\theta)^2 + b'(varz)b - 2b'cov(z, \theta) + var\theta :$$

$$D = (a + b'Ez - E\theta)^2 + b'Vb - 2b'c + var\theta. \tag{1}$$

Write $b^* := V^{-1}c$, $D^* := var(\theta) - c'V^{-1}c$. Then this becomes

$$D = (a + b'Ez - E\theta)^2 + (b - b^*)'V(b - b^*) + D^* \tag{$*$}$$

(the quadratic terms check as $b^{*T}Vb* = c^TV^{-1}VV^{-1}c = c^TV^{-1}c$, the linear terms as $c = Vb^*$).

The third term on the right in $(*)$ does not involve a, b, while the first two are non-negative (the first is a square, the second a quadratic form with matrix V, non-negative definite as V is a covariance matrix). So the expected quadratic loss D is minimised by choosing $b = b^*$, $a = -b^{*\prime}Ez + E\theta$. Then

$$d(x) = E\theta + cV^{-1}(z - Ez), \qquad c := cov(z, \theta), \quad V := var(z).$$

This gives the *Bayes linear estimator* of θ based on data $z = z(x)$. This is the best approximation to the posterior mean (in the sense of mean-square error) among the class of linear estimators (in $z = z(x)$).

Distributional assumptions.

The Bayes linear estimator depends only on first and second moments: $E\theta$, Ez, $c = cov(z, \theta)$, $V = var(z)$. So we do not need to know the full likelihood, just the first and second moments of $(\theta, z(x))$, the parameter and the function z in which we want the estimator to be linear.

Application. We have met this in the *Kalman filter* (V.11).

9. Bayesian solution of the equity premium puzzle.

Following Markowitz (I.5), we should diversify our financial savings into a range of assets in our portfolio – including cash (invested risklessly – e.g.,

by buying Government bonds, or 'gilts', or putting it in the bank or building society – which we suppose riskless here, discounting such disasters as the Icelandic banking crisis, Northern Rock, RBS etc.) and risky stock. There is no point in taking risk unless we are paid for it, so there will be an excess return – equity premium – $\mu - r$ of the risky stock (return μ) over the riskless cash (return r), to be compared with the volatility σ of the risky stock via the *Sharpe ratio* (or *market price of risk*) $\lambda := (\mu - r)/\sigma$). Historical data show that the observed excess return seems difficult to explain.

A Bayesian solution to this 'equity premium puzzle' (the term is due to Mehra & Prescott (1985)) has been put forward by Jobert, Platania and Rogers: there is no equity premium puzzle, if one uses a Bayesian analysis to reflect fully one's uncertainty in modelling this situation. See

[JPR] A. JOBERT, A. PLATANIA & L. C. G. ROGERS, *A Bayesian solution to the equity premium puzzle*. Preprint, Cambridge (available from Chris Rogers' homepage: Cambridge University, Statistical Laboratory).

The Twenties Example [JPR]. One observes daily prices of a stock for T years, with an annual return rate of 20% and an annual volatility of 20%. How large must T be to give confidence intervals of $\pm 1\%$ for (i) the volatility, (ii) the mean? Answers: (i) about 11; (ii) about 1,550!!

This illustrates what is called *mean blur*; see e.g.

D. G. LUENBERGER, *Investment Science*, OUP, 1997.

Rough explanation: for the mean, only the starting and final values matter (so effective sample size is 2); for the volatility, everything matters.

For non-Bayesian approaches, see e.g. Maenhout, *Rev. Fin. Studies* (2004), Barillas, Hansen & Sargent, *J. Econ. Th.* (2009).

10. Bayesian Non-parametrics.

We have discussed Bayesian statistics at some length in this Ch. VII, and (more briefly) Non-parametric statistics in Ch. VI. It is natural to wonder whether the two can be combined. This is indeed happening. The process has been enormously helped by the growth of modern computer power. Those interested can investigate this for themselves: e.g., Googling "Bayesian non-parametrics" produced 7,990 hits and "Bayesian nonparametrics" 30,700. There are lots of connections with machine learning, for example, and lots of applications.

NHB 19.12.2014