smfd18(13a).tex

Day 18. 12.12.2013.

6. Hierarchical models; Markov Chain Monte Carlo (MCMC).

In the Bayesian paradigm, everything is random, including the parameters; also, the parameters are drawn from a prior, and we may have difficulty in choosing the prior. Such difficulties may be lessened if we draw the parameters of the prior from some 'prior prior', which will itself have parameters, called *hyperparameters*. Such a model is called a *hierarchical model*. Our main sources here are Robert [R] Ch. 8,9, Gelman et al. [GCSR] Ch. 5, 11.

Definition. A *hierarchical Bayes model* is a Bayesian model $(f(x|\theta), \pi(\theta))$ in which the prior $\pi(\theta)$ is decomposed into conditional distributions

$$\pi_1(\theta|\theta_1), \pi_2(\theta_1|\theta_2), \ldots, \pi_n(\theta_{n-1}|\theta_n)$$

and a marginal $\pi_{n+1}(\theta_n|\theta_n)$ such that

$$\pi(\theta) = \int \ldots \int \pi_1(\theta|\theta_1)\pi_2(\theta_1|\theta_2)\pi_n(\theta_{n-1}|\theta_n)\pi_{n+1}(\theta_n)d\theta_1 \ldots d\theta_{n+1}. \qquad (H)$$

The parameters θ_i are called *hyperparameters of level i*.

A hierarchical Bayes model is itself a Bayesian model, but the decomposition (H) is often useful – e.g., in MCMC (below), and in revealing structural information.

One rarely needs to go beyond $n = 2$, and we shall not do so. So we shall always have

$$\theta|\theta_1 \sim \pi_1(\theta|\theta_1), \qquad \theta_1 \sim \pi_2(\theta_1). \qquad (H)$$

Here the distribution of θ is a *mixture* of the θ_1, with *mixing distribution* π_2.

Example: Random effects in the linear model.

We may have a *mixed model*, with some *fixed effects*, as in IV, and some *random effects*. The classical instance of this is Henderson's work on the breeding of dairy cows (1950). The fixed effects are the objects of study – typically, diet, of interest for its effect on milk yield. The random effects are the animals – animals differ, just as people do. It is conventional to write the model equation here as

$$y = X\beta + Zu + \epsilon,$$

where

$$W = (X, Z)$$

is the $n \times (p+q)$ design matrix, X $(n \times p)$ and Z $(n \times q)$ are the design submatrices for the fixed and random effects. We take the random effects u and the error ϵ uncorrelated (independent when both are Gaussian, as we may as well assume here). The best linear unbiased estimator (BLUE) of IV.1 is conventionally called a *best linear unbiased predictor (BLUP)* here. These are the solutions of *Henderson's mixed model equations (MMEs).* Two different forms of the BLUP are given in [BF] 9.1. The use of Bayes' theorem is mentioned there. This is a hierarchical model with

$$y|\theta \sim N(\theta, \Sigma_1), \qquad \theta|\beta \sim N(X\beta, \Sigma_2).$$

Here the mean θ of y is decomposed into the fixed effects $X\beta$ and the random effects $Z\eta$, where $\eta \sim N(0, \Sigma_2)$.

Education.

Mixed models are widely used in educational studies (and more widely in Social Statistics). Here the fixed effects are the ones being studied – concerning, e.g., influence on performance of changes in syllabus, examination mode etc. The random effects are the pupils.

Finance.

Here the fixed effects are state of the economy, industrial sector etc. The random effects are the specific characteristics of the individual firms involved in the study.

Bayesian v. classical.

Strictly speaking, whether this procedure is classical or Bayesian depends on what our inference is about. The procedure is classical if the inference is about the fixed effects (β), but Bayesian if it is about the overall effects (θ).

Normal mean-variance mixtures (NMVM); normal variance mixtures (NVM).

The *Bessel function of the third kind*, K_λ (λ real) is defined (for our purposes) by the integral representation

$$K_\lambda(x) = \frac{1}{2}\int_0^\infty u^\lambda \exp\{-\frac{1}{2}(u + 1/u)\} du/u \qquad (x \geq 0).$$

Then for $\psi, \chi > 0$,

$$f(x) := \frac{(\psi/\chi)^{\frac{1}{2}\lambda}}{2K_\lambda(\sqrt{\psi\chi})} x^{\lambda-1} \exp\{-\frac{1}{2}(\psi x + \chi/x)\} \qquad (x > 0)$$

is a probability density, the *generalised inverse Gaussian (GIG).*

The distribution of $x \sim N(\mu + \beta\sigma^2, \sigma^2)$, where σ^2 is sampled randomly from GIG, forms a *normal mean-variance mixture (NMVM)*, with *mixing distribution* GIG. It is called the *generalised hyperbolic* distribution, GH. The case $\beta = 0$ is simpler; we then get a *normal variance mixture (NVM).*

The GH distributions have been much used in mathematical finance, specially for return distributions with intermediate return interval – say, daily returns (Bingham & Kiesel 2001; Barndorff-Nielsen 1970s-90s; Eberlein 1990s). The log-density is a (branch of a) hyperbola (hence the name). As a hyperbola has linear asymptotes, the log-density decays linearly at $\pm\infty$. By contrast, the Gaussian log-density (monthly returns) decays quadratically, while the Student t log-density (tick data) decays logarithmically.

The GH distributions can be defined in any number of dimensions. They have two important general properties:
1. They are *elliptical.* They are an important parametric special case within this semi-parametric setting; see I.6.2 D2, V.6 D6, VI.3 D10.
2. They are *self-decomposable*: they belong to the class SD of distributions of stationary $AR(1)$ time-series models,

$$X_t = \rho X_{t-1} + \epsilon_t.$$

Bayesian sampling; HM.

We return to (H), in the form

$$\pi(\theta|x) = \int \pi_1(\theta|x, \lambda)\pi_2(\lambda|x)d\lambda. \tag{H}$$

If we can sample efficiently from π_1 and π_2, we can use MCMC (in the form of a Bayesian sampling technique, *data augmentation* (Tanner & Wong, 1987)) to sample from π, by the following iterative algorithm.
Initialisation: Start with an arbitrary value λ_0.
Iteration: For $i = 1, \ldots, k$, generate
a. $\theta_i \sim \pi_1(\theta|x, \lambda_{i-1})$;
b. $\lambda_i \sim \pi_2(\lambda|x, \theta_i)$.
The generation of θ_i only depends on θ_{i-1}, not on previous values, so (θ_i) has the Markov property. Under suitable regularity conditions, this Markov chain will be ergodic, with limiting distribution π; furthermore, the approach to stationarity will often be geometrically fast.

The Hastings-Metropolis algorithm HM in this setting runs as follows.

To sample from a distribution π known up to a normalising factor, and given a transition kernel $q(\theta|\theta')$, HM proceeds as follows.
(i) Start with θ_0 arbitrary.
(ii) Update from θ_m to θ_{m+1} by:
1. Generate $\xi \sim q(.|\theta_m)$;
2. Define

$$\rho := \Big(\frac{\pi(\xi)q(\theta_m|\xi)}{\pi(\theta_m)q(\xi|\theta_m)}\Big) \wedge 1.$$

3. Take

$$\theta_{m+1} := \xi \text{ with probability } \rho, \quad \theta_m \text{ otherwise.}$$

Again under suitable regularity conditions, the Markov chain (θ_m) converges to the equilibrium distribution π as m increases. The convergence is often geometrically fast, again under suitable conditions.

Graphical models

It is possible to model complex statistical situations, with many variables, some of which are *conditionally independent given others*. Such conditional independence can be conveniently encoded, and represented visually, using *graphs* (in the sense of Graph Theory, an important branch of Combinatorial Theory). We must be brief here; we refer for a monograph treatment to Steffen L. LAURITZEN, *Graphical models*, OUP, 1996.

Graphical models originate in three different areas:
(i) Statistical Physics, in the work of Gibbs[1]. Here the idea is that particles can only interact with their immediate neighbours.
(ii) Genetics. This, incidentally, is one of the major application areas of heirarchical models, MCMC etc. (Human Genome Project, etc.).
(iii) Contingency tables. The analysis of complicated multi-dimensional contingency tables, where the data is counts cross-classified by characteristics, is important in the Social Sciences.

See in particular Lauritzen, Ch. 4 (Contingency tables), Ch. 5 (Multivariate normal models), 7.3.1 (MCMC); also *EM algorithm* (two steps – expectation, maximisation), 7.4.1.

[1] J. W. Gibbs (1839-1903), American; one of the three founding fathers of Statistical Physics, with James Clerk Maxwell (1831-1879), Scottish, and Ludwig Boltzmann (1844-1906), German.

7. Further Bayesian aspects.

1. *Posterior means* [O'H] 1.25, p.15].

If t is an estimate of θ given data x, the mean squared error is

$$E[(t-\theta)^2|x] = E[t^2|x] - 2E[t\theta|x] + E[\theta^2|x] = t^2 - 2tE[\theta|x] + E[\theta^2|x]$$

(t is a statistic, that is, a function of the data x, so is known when x is known, and can be taken out of the expectation signs). Add and subtract $(E[\theta|x])^2$:

$$E[(t-\theta)^2|x] = (t - E[\theta|x])^2 + var(\theta|x).$$

Thus the value of t which minimises the posterior expected squared error is $t = E[\theta|x]$, the *posterior mean.* This now has two roles:
(i) minimising mean square error,
(ii) location summary of the posterior distribution.

2. *Repeated use of Bayes' Theorem* [O'H] 3.5, p. 66].

Suppose now our data x is partitioned into (x_1, x_2), where we observe x_1 first and x_2 second. With prior $f(\theta)$, we have two stages:
Stage 1. Posterior

$$f(\theta|x_1) = f(\theta)f(x_1|\theta)/f(x_1), \qquad f(x_1) = \int f(\theta)f(x_1|\theta)d\theta. \qquad (i)$$

Stage 2. The *prior* density for stage 2 is the *posterior* density above after stage 1. The *likelihood* is $f(x_2|\theta, x_1)$. So the posterior is

$$f(\theta|x_1,x_2) = f(\theta|x_1)f(x_2|\theta,x_1)/f(x_2|x_1), \qquad f(x_2|x_1) := \int f(\theta|x_1)f(x_2|\theta,x_1)d\theta. \qquad (ii)$$

Substitute $f(\theta|x_1)$ from (i) into (ii):

$$f(\theta|x_1,x_2) = \frac{f(\theta)f(x_1|\theta)f(x_2|\theta,x_1)}{f(x_1)f(x_2|x_1)}.$$

Now $f(x_2|x_1) := f(x_1,x_2)/f(x_1)$, so the denominator is $f(x_1,x_2)$. Similarly, the numerator is

$$f(\theta).\frac{f(\theta,x_1)}{f(\theta)}.\frac{f(\theta,x_1,x_2)}{f(\theta,x_1)} = f(\theta,x_1,x_2) = f(\theta)f(x_1,x_2|\theta).$$

So

$$f(\theta|x_1,x_2) = f(\theta).f(x_1,x_2|\theta)/f(x_1,x_2),$$

the usual result of Bayes' Theorem for updating by the whole data $x = (x_1, x_2)$ in one step. So:

Proposition. If data $x = (x_1, x_2)$ arrives in two stages, two applications of Bayes' Theorem, updating by x_1 first, then by x_2 given x_1, is equivalent to one application of Bayes' Theorem updating by $x = (x_1, x_2)$.

Corollary. If data $x = (x_1, \cdots, x_n)$ arrives successively in n stages, n applications of Bayes' Theorem – updating by x_i given $x_1, \cdots, x_{i-1}$ $(i = 1, \cdots, n)$ – are equivalent to one application of Bayes' theorem.

The systematic repeated use of Bayes' theorem is important in the subjects of Time Series (Ch. V) and Forecasting. In particular, the repeated *recursive* use of Bayes' theorem occurs in the *Kalman filter* (V.11), which is widely used – for instance, in engineering applications [on-line, or real-time, control of spacecraft, etc.] and in econometric time-series.

3. *Sufficiency* [O'H] 3.9, 69].

Suppose now that $x = (x_1, x_2)$, where x_1 is informative about θ, x_2 is uninformative. This is the idea of *sufficiency*, already encountered in classical statistics. We give a Bayesian treatment. To say that x_2 is uninformative means that x_2 cannot affect our views on θ, that is,
(i) $f(\theta|x) = f(\theta|x_1, x_2)$ does not depend on x_2, i.e.

$$f(\theta|x_1,x_2) = f(\theta|x_1), \qquad \text{or} \qquad \frac{f(\theta,x_1,x_2)}{f(x_1,x_2)} = \frac{f(\theta,x_1)}{f(x_1)}:$$

$$\frac{f(\theta,x_1,x_2)}{f(\theta,x_1)} = \frac{f(x_1,x_2)}{f(x_1)}, \qquad \text{i.e.} \qquad f(x_2|x_1,\theta) = f(x_2|x_1):$$

(ii) $f(x_2|x_1, \theta)$ does not depend on θ.
Either of (i), (ii), which are equivalent, can be used as the definition of sufficiency in a Bayesian treatment. Notice that (i) is essentially a Bayesian statement: it is meaningless in classical statistics, as there θ cannot have a density.

Now recall the classical Fisher-Neyman Factorisation Criterion for sufficiency: the likelihood $f(x|\theta)$ factorises as

(iii) $f(x|\theta)$, or $f(x_1, x_2|\theta)$, $= g(x_1, \theta)h(x_1, x_2)$,
for some functions g, h. As before:

Proposition. x_1 is sufficient for θ iff the Factorisation Criterion (iii) holds.

Proof. (ii) $\Rightarrow$ (iii):

$$\begin{aligned} f(x|\theta) = f(x_1, x_2|\theta) &= f(x_1|\theta)f(x_2|x_1, \theta) \qquad \text{(as in 2 above)} \\ &= f(x_1|\theta)f(x_2|x_1) \qquad \text{(by (ii))}, \end{aligned}$$

giving (iii).
(iii) $\Rightarrow$ (i): By Bayes' Theorem in the form 'posterior proportional to prior times likelihood', the factor $h(x_1, x_2)$ in (iii) can be absorbed into the constant of proportionality [which is unimportant: it can be recovered from the remaining terms, its role being merely to make these integrate to one]. Then x_2 drops out, so does not appear in the posterior, giving (i). //

Note. This proof is easier than the classical one! To a Bayesian, it is also more intuitive and revealing.

4. *Asymptotic normality* [O'H] 3.18, p. 74].

We recall (I.3) that in classical statistics, the maximum-likelihood estimator $\hat{\theta}$ of θ based on n i.i.d. readings $x_1, \cdots, x_n$ is asymptotically normal, with mean θ and variance $1/(nI(\theta))$, where $I(\theta)$ is the Fisher information per reading:

$$I(\theta) := E[(\ell'(\theta))^2] = -E[\ell''(\theta)], \qquad \ell(\theta) := \log f(x|\theta)$$

the log-likelihood (the likelihood itself is usually written $L(\theta)$ in classical statistics). This result needs some regularity conditions:
(i) enough smoothness to justify differentiating under the integral sign twice with respect to θ (as in the derivation of the above equation for the information, and in the proof of the Cramér-Rao inequality),
(ii) that the support of the likelihood (the region where it is positive) should not depend on θ.
Now the above is a large-sample result, in which the sample size n increases. So we expect that the data information will swamp the prior information. It does, and in the Bayesian case: [O'H], 3.18-26.

5. *Exponential families.*

A likelihood $f(x|\theta)$ belongs to the *exponential family* if it is of the form

$$f(x|\theta) = \exp\{a(\theta)u(x) + b(\theta) + k(x)\}$$

(as usual, we use vector notation: x, θ may be several-dimensional; see below). Exponential families (introduced in 1935-36 by Darmois, Pitman and Koopman) arise naturally in classical statistics. We quote: if a statistic $u(x)$ is minimum-variance ('efficient') and unbiased for θ, then the likelihood can be written in the above form (this follows from the conditions for equality in the Cramér-Rao inequality giving the minimum-variance bound, or 'information bound'). By the Fisher-Neyman Factorisation Criterion, $u(x)$ is sufficient for θ. So *efficiency implies sufficiency and membership of an exponential family.*

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

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