STA 376

Spring Semester 2006
Mike West

January 19, 2006

1 EM and Mode Hunting

EM (Expectation-Maximisation) algorithms for mode-hunting: marginal posterior modes in a Bayesian model analysis, MLEs in problems of missing data or latent variables.

1.1 Relevant Theory: Entropy and Küllback-Leibler Divergence

- Two density functions $f(x)$ and $g(x)$ with common support
- x is general - discrete, real, multivariate, etc
- Entropy: $H_f = -\int \log(f(x))f(x)dx$
  - $H_f = E(-\log(f(x)))$
  - *more generally, formally:* $-\int \log(f(x))dF(x)$ with distribution F
- KL divergence of g from f:
$$K_{g|f} = \int \log(f(x)/g(x))f(x)dx$$
- $K_{g|f} = E(-\log(g(x))) - H_f$ where the expectation is with respect to $f(x)$
- *Key property:* $K_{g|f} \geq 0$ with equality if and only if $f(x) \equiv g(x)$ everywhere
  - *Proof:* (Lange, 10.4).
    Simply an application of Jensen's inequality, based on strict convexity of $-\log(w)$ for $w > 0$. The expected value of a convex function exceeds the function value at the expectation: $E(q(w)) \geq q(E(w))$ for any convex function of a random variable w. So, with $w = g(x)/f(x)$ and $q(w) = -\log(w)$ and taking the expectation with respect to $f(x)$,
$$K_{g|f} = E(-\log(g(x)/f(x))) \geq -\log(E(g(x)/f(x))) = 0.$$
- *One result:* $-\int \log(g(x))f(x)dx \geq H_f$

1.2 EM for Marginal Mode Hunting

EM traditionally derived for MLE evaluation in missing data problems. This is a special case of a Bayesian marginal posterior mode computation, and the general Bayesian setting is easier to understand and derive (see also Gelman et al, chapter 12).

- Statistical model defines a posterior $p(\theta, \tau|y)$ for parameters or latent variables θ, τ of arbitrary nature and dimension, and observed data of any kind y
- *Goal:* Compute marginal posterior mode(s) for θ : θ value that maximises $\log(p(\theta|y))$ (always numerically safer to work on log scale)
- *Problem:* Must marginalise over τ - problems of complexity such that the integration is hard.

1.3 Starting Point for EM Mode Hunting:

For any value of τ,

$$\log(p(\theta|y)) = \log(p(\theta, \tau|y)) - \log(p(\tau|\theta, y))$$

- Take expectation with respect to $p(\tau|\theta^0, y)$ for any specified value θ^0 (imagine this is an initial "guess" at the marginal posterior mode for θ)

$$\begin{aligned}\log(p(\theta|y)) &= \int \log(p(\theta, \tau|y))p(\tau|\theta^0, y)d\tau - \int \log(p(\tau|\theta, y))p(\tau|\theta^0, y)d\tau \\ &= Q(\theta|\theta^0) + R(\theta|\theta^0)\end{aligned}$$

  (both depend on y but notation drops that for clarity)
- *Second term:*
  - Match $f(x) \leftarrow p(\tau|\theta^0, y)$ and $g(x) \leftarrow p(\tau|\theta, y)$. Then

$$R(\theta|\theta^0) \geq -\int \log(p(\tau|\theta^0, y))p(\tau|\theta^0, y)d\tau$$

    with equality if and only if $\theta = \theta^0$. So $R(\theta|\theta^0)$ is *minimized* at $\theta = \theta^0$
- Consider any value $\theta = \theta^1$ such that

$$Q(\theta^1|\theta^0) > Q(\theta^0|\theta^0).$$

  Then:

$$\log(p(\theta^1|y)) = Q(\theta^1|\theta^0) + R(\theta^1|\theta^0) > Q(\theta^0|\theta^0) + R(\theta^0|\theta^0) = \log(p(\theta^0|y))$$

  - Generalised EM (GEM): *any* θ^1 such that

$$Q(\theta^1|\theta^0) > Q(\theta^0|\theta^0)$$

    increases the marginal posterior density
  - EM: Find $\theta = \theta^1$ to *maximise* $Q(\theta|\theta^0)$
    * "E"-step: Take the expectation to define Q
    * "M"-step: Maximise Q

- Algorithm:
  - Start anywhere: θ^i with $i = 0$.
  - Iterate: θ^{i+1} increases (GEM) or maximises (EM) the (objective) function $Q(\theta|\theta^i)$ over θ.
  - Above theory shows that this surely moves to higher marginal posterior density values, and so converges to a posterior mode
  - Local modes - will not escape. Multiple restarts generally needed. Can be very slow.
  - Problems in which computing Q is very hard are not good candidates for EM.
  - Often very easy to implement and compute in "standard" statistical model classes, ar least generating information as a starting point for further analysis.
  - Note that the iterations will also generate information about τ, often in terms of posterior expected values of elements of τ directly, conditional on the iterated values of θ. For example, $E(\tau|\theta^i, y)$ at the (approximate) posterior mode of θ.

- One simple, venerable example is random sampling from a T distribution under a standard reference prior: $(x_i|\theta) \sim T_k(\mu, \sigma^2)$ independently, with $\theta = (\mu, \sigma)$ with $p(\theta) \propto \sigma^{-2}$. Here τ stands for the set of n implicit random scales that mix normal distributions to generate the T. That is, the model is equivalent to $(x_i|\theta, \tau) \sim N(\mu, \sigma^2/\tau_i)$ where $\tau_i \sim Ga(k/2, k/2)$ independently.
- Another key practical example is multiple shrinkage prior modelling in regression.

  Regression setup: data $n-$vector $z = H\beta + \nu$ where H is fixed $n \times p$ design matrix, β is $p-$vector of regression parameters, and $\nu \sim N(0, \phi^{-1}I)$ for some precision ϕ. Hierarchical/multiple shrinkage prior $\beta|\tau \sim N(0, T)$ where $T = diag(\tau_1, \ldots, \tau_p)$ and τ is just the set of these values. Often use (conditionally conjugate) inverse gamma priors over these shrinkage parameters: $\tau_i^{-1} \perp\!\!\!\perp Ga(a/2, b/2)$ for specified (a, b). Interest focuses on β and the EM can be applied easily and usefully to compute posterior modes for $\theta = (\beta, \phi)$ in this setting. Evaluate under the traditional reference prior $p(\theta) \propto \phi^{-1}$.
- A second standard and useful example is the traditional normal hierarchical model (random effects) for 1-way Anova data, as developed in Gelman et al (section 12.5).

1.4 Missing Data & Traditional View of EM

Contexts in which τ represents missing data or latent variables: Usual alternative notation is $\tau = z$ and the *full* or *complete* data is $x = (y, z)$. Problems are often those in which the model and inference is tractable if z were in fact also observed.

- Recall the key definition:

$$Q(\theta|\theta^0) = \int \log(p(\theta, \tau|y))p(\tau|\theta^0, y)d\tau$$

- Use the identity

$$p(\theta, \tau|y) = p(y, \tau|\theta)p(\theta)/p(y)$$

  inside the integral defining the Q function to get

$$Q(\theta|\theta^0) = \int \log(p(y, \tau|\theta))p(\tau|\theta^0, y)d\tau + \log(p(\theta)) - \log(p(y))$$

  Denote the integral here by

$$Q^{MLE}(\theta|\theta^0) = \int \log(p(y, \tau|\theta))p(\tau|\theta^0, y)d\tau$$

  so that

$$Q^{MLE}(\theta|\theta^0) = Q(\theta|\theta^0) - \log(p(\theta)) + \text{constant}$$

- Maximising $Q(\theta|\theta^0) - \log(p(\theta))$ generates the EM for (local) MLEs.
- In cases of $p(\theta) \propto \text{constant}$, the marginal posterior is just $p(\theta|y) \propto p(y|\theta)$ so that posterior modes are exactly (local or global) MLEs. In such cases, maximising Q is the same as maximising Q^{MLE} since $\log(p(\theta))$ is constant.
- $Q^{MLE}(\theta|\theta^0)$ is the traditional (G)EM criterion function in missing data problems, when $\tau = z$ is missing data rather than "parameters". Technically the same thing of course. Using the z notation,

$$Q^{MLE}(\theta|\theta^0) = \int \log(p(y, z|\theta))p(z|\theta^0, y)dz$$

  - the expected value of the log of the complete data density (given θ) from the statistical model, with expectation with respect to the current "best guess" of the distribution of the missing data having seen the observed data.
- In many applications where this is feasible, the data are conditionally independent under the assumed model, so that $p(z|\theta, y) = p(z|\theta)$, not dependent on y. EM works well in many such problems; it can be very hard to derive and implement in problems where the dependence on y of this distribution is intricate.
- MCMC usually applies much more easily in many such problems.