

QTL mapping in mice, cont.

Lecture 11, Statistics 246

February 26, 2004

Maximum likelihood fitting of a two-component normal mixture model using the EM algorithm

This is the simplest case. After dealing with it, we'll turn to our QTL mixture models.

Suppose that y_1, \dots, y_n are i.i.d. random variables having normal mixture density $(1-p)\phi((y-\mu_0)/\sigma) + p\phi((y-\mu_1)/\sigma)$, where ϕ is the standard normal density. Our interest is usually in estimating p, μ_0, μ_1 and σ , though in our QTL problem, we know p .

One approach to the problem is to introduce i.i.d Bernoulli r.v. z_1, \dots, z_n , such that $pr(z_i = 1) = p$, and $pr(y_i | z_i) = \phi((y - \mu_{z_i})/\sigma)$. In this case, the marginal distribution of y_1, \dots, y_n is the above normal mixture. In the spirit of the **EM-algorithm**, we consider the **full data** to be $(y_1, z_1), \dots, (y_n, z_n)$ with z_1, \dots, z_n **missing**; equivalently, the z_1, \dots, z_n are viewed as random variables **augmenting** the y_1, \dots, y_n .

EM generalities

Write $y = (y_1, \dots, y_n)$, $z = (z_1, \dots, z_n)$, and $\theta = (\rho, \mu_0, \mu_1, \sigma)$. The EM proceeds by considering the expectation of **full** data log-likelihood $\log pr(y, z; \theta)$, given y , rather than the **observed** data log-likelihood $\log pr(y; \theta)$. It alternates between the **E-step**, which involves forming $Q(\theta, \theta^{old}) = E^{old} \{ \log pr(y, z; \theta) \mid y \}$, where E^{old} denotes the expectation (given y) taken with an initial value θ^{old} , and the **M-step**, which involves calculating $\theta^{new} = \arg \max_{\theta} Q(\theta, \theta^{old})$.

The key fact, which justifies the widespread use of the EM-algorithm, is that if we follow this 2-step process, $\log pr(y; \theta^{new}) \geq \log pr(y; \theta^{old})$.

Exercise. Prove this last fact. (Hint: Consider the ratio $pr(y; \theta^{new})/pr(y; \theta^{old})$. Multiply top and bottom by $pr(y, z; \theta^{old})$, sum over z , and use Jensen's inequality.)

The EM-algorithm could not be presented as having anything to do with maximum likelihood were it not for the fact that under certain conditions,

$$E^{\theta} \{ \log pr(y, z; \theta) \mid y \}, = \log pr(y; \theta).$$

Exercise. Describe circumstances under which this is true, and prove³ it.

EM generalities, cont.

Of course the whole point of using the EM algorithm is that the E and M steps should both be “easy”, e.g. given by simple closed form expressions, while direct maximization should be “hard”. As we will shortly see, this is the case in our normal mixture model problem, up to a point. But for the reasons which follow, not everyone will deal with normal mixture models via the EM.

As always, a price is paid for “ease”. If we do not have ready access to the observed data log-likelihood, $\log pr(\theta; y)$ - and usually when we do, we would not consider using the EM-algorithm - we have to work harder to get the asymptotic standard errors and confidence intervals for the MLE, θ^∞ . There are also a number of important details to consider with the EM: starting points, stopping rules, the rate of convergence, the issue of local vs global maxima, and so on. Some of these are issues whatever approach one adopts, while others are EM-specific. The EM is not a magic bullet, but its conceptual simplicity, and its ease coding have made it very popular in genetics and bioinformatics. Many problems do indeed benefit from the missing or augmented data perspective.

Normal mixture models via the EM

In the normal mixture problem, $-2\log pr(\theta; y, z)$ with known p takes the form

$$\begin{aligned} n \log \sigma^2 + \frac{1}{\sigma^2} \sum_i [y_i^2 - 2y_i \mu_{z_i} + \mu_{z_i}^2] \\ = n \log \sigma^2 + \frac{1}{\sigma^2} \sum_i [y_i^2 - 2y_i \{I(z_i = 0)\mu_0 + I(z_i = 1)\mu_1\} \\ + \{I(z_i = 0)\mu_0^2 + I(z_i = 1)\mu_1^2\}]. \end{aligned}$$

Because of the linearity of this expression, the E-step simply consists of replacing $I(z_i = 0)$ by $pr(z_i = 0 | y_i, \theta^{old})$, and similarly for $I(z_i = 1)$. This is a quite general phenomenon.

Exercise: Check these assertions and carry out the calculations.

Normal mixture models via the EM, cont

Having completed this, the M-step leads to

$$\mu_0^{new} = \sum_i pr(z_i = 0 | y_i, \theta^{old}) y_i / \sum_i pr(z_i = 0 | y_i, \theta^{old}),$$

with a similar formula for μ_1^{new} .

Exercise: Derive these formulae, and also one for σ^{new} .

With many data sets, these quantities converge after a several hundred iterations.

Exercise: Repeat this exercise for the case of unknown mixing probability p , obtaining the MLE of p . (Guess the answer first.)

Putting it together: the EM-algorithm in mouse QTL mapping.

Now let's work out the contribution of one mouse to the full data log-likelihood function $\log pr(\mathbf{y}, \mathbf{m}, \mathbf{g} ; \theta)$, for an F_2 intercross. Here \mathbf{y} denotes *all* the phenotype (QT) data, \mathbf{m} all the marker data, and \mathbf{g} all the “missing” genotype data at a putative QTL located at a position z on a chromosome, and $\theta = (\mu_A, \mu_H, \mu_B, \sigma)$ are the parameters in the **single** QTL model.

Let non-bold symbols denote the same data for just one mouse.

First, note that we can factorize this mouse's contribution as follows,

$$pr(y, m, g ; \theta) = pr(m) \times pr(g | m) \times pr(y | g; \theta),$$

since $pr(y | m, g ; \theta) = pr(y | g ; \theta)$ (*why?*).

Now $pr(m)$ can be ignored, as it does not involve the parameters, while $pr(g | m)$ gives the probabilities of this mouse having A , H or B at z , given all *its* marker data, i.e. the mixture component probabilities for *this* mouse.

As previously mentioned, this is an HMM calculation, i.e. another EM.

Finally, $pr(y | g; \theta)$ is just the mixture distribution for that mouse, and as in the previous discussion, the current probabilities of the mouse being A , H or B on the basis of its y weight its contributions to $Q(\theta, \theta^{old})$.

Multiple QTL methods

Why consider multiple QTL at once?

- **To separate linked QTL.** If two QTL are close together on the same chromosome, our one-at-a-time strategy may have problems finding either (e.g. if they work in opposite directions, or interact). Our LOD scores won't make sense either.
- **To permit the investigation of interactions.** It may be that interactions greatly strengthen our ability to find QTL, though this is not clear.

The main reason for considering multiple QTL simultaneously is

- **To reduce residual variation.** If QTL exist at loci other than the one we are currently considering, they should be in our model. For if they are not, they will be in the error, and hence reduce our ability to detect the current one. See below.

Abstractions/Simplifications

First pass: For a BC, assume

- Complete marker data
- QTL are at marker loci
- QTL act additively

The problem

n backcross mice; M markers in all, with at most a handful expected to be near QTL

x_{ij} = genotype (0/1) of mouse i at marker j

y_i = phenotype (trait value) of mouse i

$$Y_i = \mu + \sum_{j=1}^M \Delta_j x_{ij} + \varepsilon_j \quad \text{Which } \Delta_j \neq 0 ?$$

⇒ Variable selection in linear models (regression)

Variable selection in linear models: a few generalities

There is a huge literature on the selection of variables in linear models. Why? First, if we leave out variables which should be in our model, we **risk biasing** the parameter estimates of those we include, and our “error” has systematic components, reducing our ability to identify those which should be there. Second, when we includes variables in a model which do not need to be there, we **reduce the precision** of the estimates of the ones which are there. We need to include **all** the relevant (perhaps important is a better word), and **no** irrelevant variables in our model. This is an instance of classic statistical trade-off of bias and variance: too few variables, higher bias; too many, higher variance. The literature contains a host of methods of addressing this problem, known as model or variable selection methods. Some are general, i.e. not just for linear models, e.g. the Akaike information criterion (AIC), the Bayes information criterion (BIC), or cross-validation (CV). Others are specifically designed for linear models, e.g. ridge regression, Mallows’ C_p , and the lasso. We don’t have the time to offer a thorough review of this topic, but we’ll comment on and compare a few approaches.

Special features of our problem

We have a known joint distribution of the covariates (marker genotypes), and these are Markovian

We may have lots of missing covariate information, but it can be dealt with.

Our aim is to identify the key players (markers), not to minimize prediction error or find the “true model”

Conclusion: Our problem is not a generic linear model selection problem, and this offers the hope that we can do better than we might in general.