STAT 513/413: Lecture 22
EM Algorithm

(A big splash in the maximum likelihood world)

Not everything has to be done numerically

We may sometimes solve the problem in closed form

And sometimes, we may solve by numerics only some part of it

EM Algorithm: a general scheme for reducing (some) more difficult problems into a (sequence of) easier ones

Not a scheme for those just following recipes needs invention

EM-algorithm, as formulated below, produces a sequence of iterates, with likelihood increasing along the sequence (-loglikehood decreasing)

(It does not produce the global maximum of the likelihood, *the textbook has it wrong*)

General structure

Suppose that data, y are generated by some random mechanism Y (random variable or vector: possibly many quantities are collapsed here into one letter Y or y)

The probability density of Y is $f(y;\vartheta)$; thus, when y is the observed value for Y, we may estimate ϑ via maximum likelihood

$$-\log f(y;\vartheta) \looparrowright \min_{\vartheta}!$$

Suppose we can see a way in which there were some additional data, z, generated by Z - so that we would have a density $g(y,z;\vartheta)$

Maximum likelihood estimation of ϑ would result in a problem

$$-\log g(y,z;\vartheta) \looparrowright \min_{\vartheta}! \qquad \text{(M-step)}$$

If this is an easier problem, then we can use it for solving an original one - if we know how to calculate

$$\mathrm{E}(-\log g(Y,Z;\vartheta)|Y=y;\vartheta) \qquad \text{(E-step)}$$

We start by some initial iterate ϑ_1; we do the E-step; we obtain something like in the M-step, we solve it; we obtain ϑ_2, and then we repeat E-step and M-step to obtain ϑ_3; etc.

Example: genetics

We have three letters, A, B, O - they are called *alleles*

Each individual is getting two of these, by random: their configuration, regardless of the order, determines the *genotype*

Thus, the genotypes are: AA, AB, AO, BB, BO, OO

Hardy-Weinberg law: the probability of genotype is derived from *independent* sampling of two alleles with probabilities p_A, p_B, p_O - as A, B, O are the only letters possible, we have $p_A + p_B + p_O = 1$

Thus, the genotypes have respectively probabilities

$$p_A^2,\ 2p_Ap_B,\ 2p_Ap_O,\ p_B^2,\ 2p_Bp_O,\ p_O^2$$

However, we do not observe genotypes but only phenotypes

A for AA and AO, B for BB and BO, AB for AB, and O for OO

with probabilities, respectively

$$p_A^2 + 2p_Ap_O,\ p_B^2 + 2p_Bp_O,\ 2p_Ap_B,\ p_O^2$$

And we would like to estimate p_A, p_B and p_O (and then we have also the estimates of probabilities of genotypes and phenotypes)

The easy problem

It is a "missing data problem" - the original motivation for the EM algorithm, where Z can be interpreted as "missing data" (although in some implementations purely imaginary)

If we observed genotypes, we would observe

$$n_{AA}, n_{AB}, n_{AO}, n_{BB}, n_{BO}, n_{OO} \qquad \text{of each genotype}$$

out of total n observations; the likelihood is

$$p_{AA}^{n_{AA}} p_{AB}^{n_{AB}} p_{AO}^{n_{AO}} p_{BB}^{n_{BB}} p_{BO}^{n_{BO}} p_{OO}^{n_{OO}} =$$

$$= (p_A^2)^{n_{AA}} (2p_A p_B)^{n_{AB}} (2p_A p_O)^{n_{AO}} (p_B^2)^{n_{BB}} (2p_B p_O)^{n_{BO}} (p_O^2)^{n_{OO}}$$

$$= p_A^{2n_{AA}+n_{AB}+n_{AO}} p_B^{2n_{BB}+n_{AB}+n_{BO}} p_O^{n_{AO}+n_{BO}+2n_{OO}} 2^{n_{AB}+n_{AO}+2n_{BO}}$$

so we have to minimize, under the condition $p_A + p_B + p_O = 1$

$$-(2n_{AA} + n_{AB} + n_{AO}) \log p_A$$

$$-(2n_{BB} + n_{AB} + n_{BO}) \log p_B - (n_{AO} + n_{BO} + 2n_{OO}) \log p_O$$

This yields $\hat{p}_A = \dfrac{2n_{AA} + n_{AB} + n_{AO}}{n}$, $\hat{p}_B = \dfrac{2n_{BB} + n_{AB} + n_{BO}}{n}$,

$$\hat{p}_O = \frac{n_{AO} + n_{BO} + 2n_{OO}}{n} \qquad \text{- M-step is easy}$$

The harder problem

If we observe only phenotypes, then we observe

$$n_A, n_B, n_{AB}, n_O \qquad \text{of each phenotype}$$

out of total n observations; the likelihood is

$$(p_A^2 + 2p_A p_O)^{n_A}(p_B^2 + 2p_B p_O)^{n_B}(2p_A p_B)^{n_{AB}}(p_O^2)^{n_O}$$

so we have to minimize under the condition $p_A + p_B + p_O = 1$

$$-n_A \log(p_A^2 + 2p_A p_O) - n_B \log(p_B^2 + 2p_B)$$
$$-n_{AB} \log p_A - n_{AB} \log p_B - 2n_O \log p_O$$

We could perhaps do Newton method, but for that we would have to evaluate 3 first and 6 second partial derivatives (or better just 2 and 3, but expressing p_O as $1 - p_A - p_B$)

The E-step

The E-step is determined through probabilistic calculations: evaluating conditional expectations; to this end, we need a bit more notation

Random variables corresponding to numbers of phenotypes in the sample

$$\mathcal{N}_A, , \mathcal{N}_B, \mathcal{N}_{AB}, \mathcal{N}_O \quad \text{are observed as} \quad \mathrm{n}_A, \mathrm{n}_B, \mathrm{n}_{AB}, \mathrm{n}_O$$

The random variables corresponding to numbers of genotypes

$$\mathrm{N}_{AA}, \mathrm{N}_{AB}, \mathrm{N}_{AO}, \mathrm{N}_{BB}, \mathrm{N}_{BO}, \mathrm{N}_{OO}$$

are not observed, but they appear in the minimized -loglikelihood

$$-(\mathrm{N}_{AA} + \mathrm{N}_{AB} + \mathrm{N}_{AO}) \log \mathrm{p}_A$$

$$-(\mathrm{N}_{BB} + \mathrm{N}_{AB} + \mathrm{N}_{BO}) \log \mathrm{p}_B - (\mathrm{N}_{AO} + \mathrm{N}_{BO} + 2\mathrm{N}_{OO}) \log \mathrm{p}_O$$

and we have to calculate its *conditional* expectation for given (fixed, nonrandom, unknown) $\mathrm{p}_A, \mathrm{p}_B, \mathrm{p}_O$, conditional on

$$\mathcal{N}_A = \mathrm{n}_A, \quad \mathcal{N}_B = \mathrm{n}_B, \quad \mathcal{N}_{AB} = \mathrm{n}_{AB}, \quad \mathcal{N}_O = \mathrm{n}_O$$

The conditional expectation is that of a linear combination of the conditional expectations of genotype random variables...

... so we can treat them one by one

The last two phenotypes are easy: as $\mathsf{N}_{AB} = \mathcal{N}_{AB}$, we have

$$\mathsf{E}(\mathsf{N}_{AB}|\text{all conditions}) = \mathsf{E}(\mathsf{N}_{AB}|\mathcal{N}_{AB} = \mathfrak{n}_{AB}) = \mathfrak{n}_{AB}$$

Analogously

$$\mathsf{E}(\mathsf{N}_{O}|\text{all conditions}) = \mathsf{E}(\mathsf{N}_{O}|\mathcal{N}_{O} = \mathfrak{n}_{O}) = \mathfrak{n}_{O} \quad \text{as } \mathsf{N}_{O} = \mathcal{N}_{O}$$

The other two take a bit of calculations:

$$\mathsf{E}(\mathsf{N}_{AA}|\text{all conditions}) = \mathsf{E}(\mathsf{N}_{AA}|\mathcal{N}_{A} = \mathfrak{n}_{A})$$

$$= \mathfrak{n}_{A}\,\mathbb{P}[\text{Geno} = \text{AA}|\text{Pheno} = \text{A}]$$

why? because every observed phenotype A is genotype with AA with probability equal to that seen above, and for the expected value you sum $\mathfrak{n}_A$ of those; now use the definition of conditional probability

$$= \mathfrak{n}_{A}\,\frac{\mathbb{P}[\text{Geno} = \text{AA } \& \text{ Pheno} = \text{A}]}{\mathbb{P}[\text{Pheno} = \text{A}]}$$

$$= \mathfrak{n}_{A}\,\frac{\mathbb{P}[\text{Geno} = \text{AA}]}{\mathbb{P}[\text{Geno} = \text{AA}] + \mathbb{P}[\text{Geno} = \text{AO}]}$$

$$= \mathfrak{n}_{A}\,\frac{\mathfrak{p}_{A}^2}{\mathfrak{p}_{A}^2 + 2\mathfrak{p}_{A}\mathfrak{p}_{O}}$$

An then

The other expected values are calculated analogously; eventually (one can call it the use of the Bayes theorem)

$$E(N_{AA}|\text{all conditions}) = n_A \frac{p_A^2}{p_A^2 + 2p_A p_O}$$

$$E(N_{AO}|\text{all conditions}) = n_A \frac{2p_A p_O}{p_A^2 + 2p_A p_O}$$

$$E(N_{BB}|\text{all conditions}) = n_B \frac{p_B^2}{p_B^2 + 2p_B p_O}$$

$$E(N_{BO}|\text{all conditions}) = n_B \frac{2p_B p_O}{p_B^2 + 2p_B p_O}$$

$$E(N_{AB}|\text{all conditions}) = n_{AB}$$

$$E(N_{OO}|\text{all conditions}) = n_O$$

Note that, as expected

$$E(N_{AA}) + E(N_{AO}) = n_A \quad \text{and} \quad E(N_{BB}) + E(N_{BO}) = n_B$$

So finally

The data are n_A, n_B, n_{AB}, n_O, and also $n = n_A + n_B + n_{AB} + n_O$

The EM algorithm:

takes some p_A, p_B, p_O, and calculates the new iterates of them

using them first to calculating the conditional expectation of the -loglikelihood by plugging in the conditional expectations calculated above

obtaining thus an easy maximum likelihood formulation

which can be solved in closed form to give the new iterates

$$\widehat{p}_A = \frac{1}{n}\left(\frac{2n_A p_A^2}{p_A^2+2p_A p_O} + n_{AB} + \frac{n_A 2p_A p_O}{p_A^2+2p_A p_O}\right) = \frac{1}{n}\left(n_{AB} + 2n_A \frac{p_A^2+p_A p_O}{p_A^2+2p_A p_O}\right)$$

$$\widehat{p}_B = \frac{1}{n}\left(\frac{2n_B p_B^2}{p_B^2+2p_B p_O} + n_{AB} + \frac{n_B 2p_B p_O}{p_B^2+2p_B p_O}\right) = \frac{1}{n}\left(n_{AB} + 2n_B \frac{p_B^2+p_B p_O}{p_B^2+2p_B p_O}\right)$$

$$\widehat{p}_O = \frac{1}{n}\left(\frac{n_A 2p_A p_O}{p_A^2+2p_A p_O} + \frac{n_B 2p_B p_O}{p_B^2+2p_B p_O} + 2n_O\right)$$

And the process repeats by making those p_A, p_B, p_O...

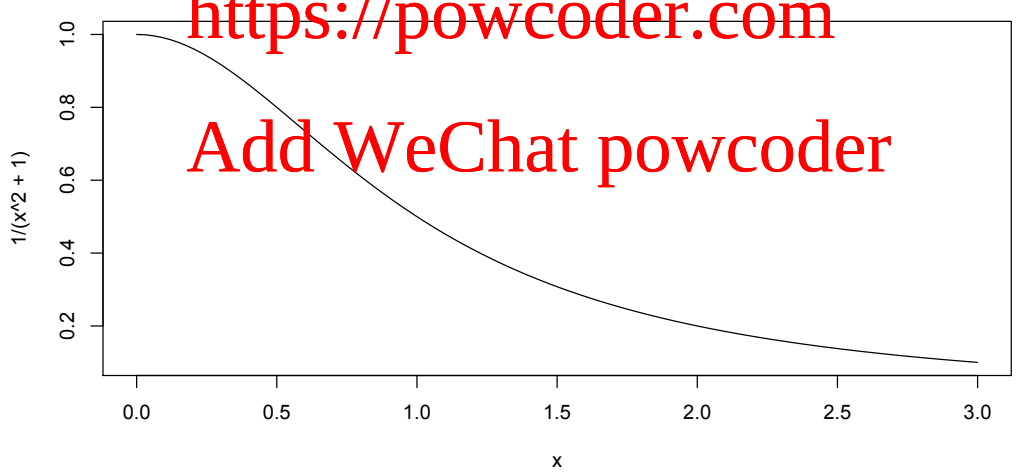
...until convergence

One more example: fitting the Cauchy distribution

The system of equations to solve is

$$\sum_{i=1}^{n} \frac{y_i - \mu}{\sigma^2 + (y_i - \mu)^2} = 0 \quad \text{and} \quad \sum_{i=1}^{n} \frac{(y_i - \mu)^2}{\sigma^2 + (y_i - \mu)^2} = \frac{n}{2}$$

We concentrate on the first one - as the second one may be not that difficult: if we interepret it as an equation for $\sigma > 0$, then the left-hand side is monotonous in σ.



For the purpose of finding μ out of the first equation, we may consider in what follows σ known and fixed (we might even consider it equal to 1 for notational simplicity, but that could mislead some)

The easy problem?

The similar problem we know would be easy is the same one when the distribution of the y_i's is not Cauchy but normal. In such a case, the first equation becomes

$$\sum_{i=1}^{n} \frac{y_i - \mu}{\sigma} = 0 \qquad \text{with an easy solution } \hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

How can the actual problem correspond to this one? Well, every statisticians know that Cauchy is in fact t distribution with 1 degree of freedom. And that t distribution is obtained as

$$\frac{U}{\sqrt{Z/k}} \text{ where } U \text{ is } N(0,1),\ Z \text{ is } \chi^2(k) \text{ and they are independent}$$

so that Cauchy is in particular

$$\frac{U}{\sqrt{Z}} \text{ where } U \text{ is } N(0,1),\ Z \text{ is } \chi^2(1) \text{ and they are independent}$$

So, let us try to make some use of it

Hard vs. easy

The actual problem assumes that $Y_1, Y_2, \ldots, Y_n$ are independent

and $\dfrac{Y_i - \mu}{\sigma}$ is Cauchy $\equiv$ t with 1 df

That is $\dfrac{Y_i - \mu}{\sigma}$ has the same distribution as $\dfrac{U_i}{\sqrt{Z_i}}$

whenever U_i is $N(0,1)$, Z_i is $\chi^2(1)$ and they are independent

and thus $\sqrt{Z_i}\dfrac{Y_i - \mu}{\sigma}$ has the same distribution as U_i

so then in the actual problem, if Z_i is given (known: write z_i)

and $\sqrt{z_i}\dfrac{Y_i - \mu}{\sigma}$ is $N(0,1)$

This is not exactly the easy problem as before - but is still easy

(... = fill in the details, if necessary)

the maximum likelihood estimate of μ (for known σ) is

$$\hat{\mu} = \frac{\sum_{i=1}^{n} z_i y_i}{\sum_{i=1}^{n} z_i}$$ the weighted mean (by z_i)

So, M-step will be again easy...

... and now the remainder of it

We need now to write the full joint likelihood of Y_i and Z_i - which is the product of *conditional* densities of all Y_i (each for given Z_i) and then the product of all densities of Z_i. As the parameters μ and σ will be only in the first part, the conditional densities, we end up after taking logs (and minus) with (relevant) negative loglikehood

$$\sum_{i=1}^{n} z_i \frac{(y_i - \mu)^2}{2\sigma^2} + \log\left(\frac{\sigma}{\sqrt{z_i}}\right)$$

We can drop the second term if σ is known, so we end up with

$$\sum_{i=1}^{n} z_i \frac{(y_i - \mu)^2}{2\sigma^2}$$

(... = fill in the details, if necessary)

Now, we need to take the conditional expectation of it, given Y_i (and μ and σ). Luckily again, it is linear in z_i, so it boils down to taking the conditional expectation of z_i itself

Densities, densities

For taking the conditional expectation of Z_i, we eventually need its density: some more educated statisticians know - or can find out - that

the general $\chi^2(k)$ density is, for $x > 0$ $\quad \frac{(1/2)^{k/2}}{\Gamma(k/2)} x^{k/2-1} e^{-x/2}$

and it is a special case of the density of the Gamma distribution

$\Gamma(\alpha, \beta)$, which is for $x > 0$ $\quad \frac{\beta^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$ (isn't it?)

At the beginning, it seems like we will need only $\chi^2(1)$, that is, the special case with $k = 1$, which is actually $\Gamma(1/2, 1/2)$...

... if we are good in calculus - otherwise, some trickery we have seen in the Bayesian analysis can help us

The magic of $\propto$ again

Conditional on $Z_i = z_i$, the density of Y_i is $N\left(\mu, \frac{\sigma^2}{z_i}\right)$

that is, the density of $Y_i|Z_i = z_i$ is $\propto e^{-z_i\frac{(y_i-\mu)^2}{2\sigma^2}}$

The joint density of Y_i and Z_i is the product of the latter and the density of $\Gamma(1/2, 1/2)$

that is $\propto e^{-z_i\frac{(y_i-\mu)^2}{2\sigma^2}} z_i^{-\frac{1}{2}} e^{-\frac{z_i}{2}} = e^{-\frac{1}{2}\left(1+\frac{(y_i-\mu)^2}{\sigma^2}\right)z_i} z_i^{-\frac{1}{2}}$

which says that the conditional distribution of $Z_i|Y_i = y_i$ is $\propto$ same

that is, it is $\Gamma(\alpha, \beta) = \Gamma\left(\frac{1}{2}, \frac{1}{2}\left(1+\frac{(y_i-\mu)^2}{\sigma^2}\right)\right)$

and then every statistician can find out (and some calculate, and some perhaps even remember) that its expected value is

$$\frac{\alpha}{\beta} = \frac{1}{1+\frac{(y_i-\mu)^2}{\sigma^2}} = E(Z_i|Y_i = y_i)$$

The finale: EM-algorithm recipe

So the combination of E- and M-step is the weighted mean...

More precisely (σ is still known!): select μ_1

Calculate weights $z_i = \dfrac{1}{1 + \dfrac{(y_i - \mu_1)^2}{\sigma^2}}$

Calculate the weighted mean $\mu_2 = \dfrac{\sum_{i=1}^{n} z_i y_i}{\sum_{i=1}^{n} z_i}$

and repeat...

Extension: regression

Why did people bother with it if it is so complicated? And moreover, one dimensional problem (in μ)??

Because it works exactly the same way in the regression setting:

not only when $y_i = \mu + \sigma\varepsilon_i$ with ε_i Cauchy errors

but also when $y_i = x_i^\top \beta + \sigma\varepsilon_i$ with ε_i still Cauchy errors (and σ still known)

The EM-algorithm goes in an analogous way: select β_1

Calculate weights $z_i = \dfrac{1}{1 + \dfrac{(y_i - x_i^\top \beta_1)^2}{\sigma^2}}$

Calculate β_2 as a weighted least squares estimate, solving

$$\sum_{i=1}^{n} z_i (y_i - x_i^\top \beta)^2 \rightsquigarrow \min_{\beta} !$$

and repeat...

So it works via EM... but after all, it is nothing but...

Finally: mixtures beware

Example 14.5 (EM algorithm for a mixture model). In this example, the EM algorithm is applied to estimate the parameters of the quadratic form introduced in Example 14.4. Recall that the problem can be formulated as estimation of the rate parameters of a mixture of gamma random variables. Although the EM algorithm is not the best approach for this problem, as an exercise we repeat the estimation for $k = 3$ components (two unknown parameters) as outlined in Example 14.4.

The EM algorithm first updates the posterior probability p_{ij} that the i^{th} sample observation y_i was generated from the j^{th} component. At the t^{th} step,

$$p_{ij}^{(t)} = \frac{\frac{1}{k} f_j(y_i|y, \lambda^{(t)})}{\sum_{j=1}^{k} \frac{1}{k} f_j(y_j|y, \lambda^{(t)})},$$

where $\lambda^{(t)}$ is the current estimate of the parameters $\{\lambda_j\}$, and $f_j(y_i|y, \lambda^{(t)})$ is the Gamma$(1/2, 1/(2\lambda_j^{(t)}))$ density evaluated at y_i. Note that the mean of the j^{th} component is λ_j so the updating equation is

$$\mu_j^{(t+1)} = \frac{\sum_{i=1}^{m} p_{ij}^{(t)} y_i}{\sum p_{ij}^{(t)}}.$$

In order to compare the estimates, we generate the data from the mixture Y using the same random number seed as in Example 14.4.

Good for their problem, but not (entirely) good for ours