Expectation–maximisation algorithm

Christiana Kartsonaki

christiana.kartsonaki@dph.ox.ac.uk

CGAT

March 2, 2017

Introduction

Fitting a parametric model

Likelihood

function of the (unknown) parameters and the data

Maximum Likelihood Estimates (MLE) $\rightarrow$ parameter estimates which make the observed data most likely

Introduction

Fitting a parametric model

Likelihood

function of the (unknown) parameters and the data

Maximum Likelihood Estimates (MLE) $\rightarrow$ parameter estimates which make the observed data most likely

General approach, as long as tractable likelihood function exists

Calculate derivative of log likelihood, set to zero and solve for parameter, check that it is a maximum

Likelihood

Data $\mathbf{x}$

If $\mathbf{x}$ known we take $\ell(\theta; \mathbf{x}) = \log\{f(\mathbf{x}; \theta)\}$ and maximise it w.r.t. θ

In one or two dimensions the likelihood function can be tabulated.

May be very difficult in many dimensions and complex models.

Likelihood

Data **x**

If **x** known we take $\ell(\theta; \mathbf{x}) = \log\{f(\mathbf{x}; \theta)\}$ and maximise it w.r.t. θ

In one or two dimensions the likelihood function can be tabulated.

May be very difficult in many dimensions and complex models.

x not fully observed $\rightarrow$ not possible

Expectation–Maximisation (EM) algorithm $\rightarrow$ iterative procedure for maximising the likelihood

Expectation–Maximisation (EM) algorithm

EM algorithm $\rightarrow$ maximise a conditional likelihood for the unobserved data

Generalisation of maximum likelihood estimation to 'incomplete' data

Any problem that is simple to solve for complete data, but the available data are 'incomplete' in some way

The likelihood for the incomplete data may have multiple local maxima and no closed form solution, even if for the complete version it has a global maximum and closed form solution

Expectation–Maximisation (EM) algorithm

- Starting parameter values
- Use them to 'estimate' complete data
- Use estimates of complete data to update parameters

Repeat until converge (estimates close to each other).

EM algorithm

Incomplete data.

- Starting parameter values
- Supplement by synthetic data
- Analyse

EM algorithm

Incomplete data.

- Starting parameter values
- Supplement by synthetic data **E step**
- Analyse **M step**

EM algorithm

Incomplete data.

- Starting parameter values
- Supplement by synthetic data **E step**
- Analyse **M step**
- Use estimate to improve the synthetic data.
- Analyse.

EM algorithm

Incomplete data.

- Starting parameter values
- Supplement by synthetic data **E step**
- Analyse **M step**
- Use estimate to improve the synthetic data.
- Analyse.

Repeat until stable answer.

Incomplete data

Latent variables (present participle of *lateo* (lie hidden))

variables that are not directly observed but are rather inferred from other variables that are observed

Expectation–Maximisation (EM) algorithm

The fitting of certain models is simplified by treating the observed data as an incomplete version of an ideal dataset whose analysis would have been easy.

Key idea $\rightarrow$ estimate the log likelihood contribution from the missing data by its conditional value given the observed data

Expectation–Maximisation (EM) algorithm

Many versions of 'the' EM algorithm.

Expectation–Maximisation (EM) algorithm

Many versions of 'the' EM algorithm.

Rather than picking a single most likely completion of the missing assignments on each iteration, the EM algorithm computes probabilities for each possible completion of the missing data, using the current parameters $\hat{\theta}^{(t)}$

$\rightarrow$ create a weighted set of data consisting of all possible completions of the data

A modified version of maximum likelihood estimation that deals with weighted data provides new parameter estimates $\hat{\theta}^{(t+1)}$

'Estimates' completions of missing data given the current model (E step) and then estimates the model parameters using these completions (M step)

Expectation–Maximisation (EM) algorithm

complete data $y_C = (y_O, y_U)$

parameter θ

Expectation–Maximisation (EM) algorithm

complete data $y_C = (y_O, y_U)$

parameter θ

Expectation–Maximisation algorithm

initialize parameter: $\theta \leftarrow \theta^{(0)}$

Expectation–Maximisation (EM) algorithm

complete data $y_C = (y_O, y_U)$

parameter θ

Expectation–Maximisation algorithm

initialize parameter: $\theta \leftarrow \theta^{(0)}$

for $t = 1$ to ...

E step

$$Q(\theta \mid \theta^{(t)}) = \mathbb{E}\left\{\log f(y_C \mid \theta) \mid y_O, \theta^{(t)}\right\}$$

$Q(\theta \mid \theta^{(t)})$: conditional expectation
$f(y_C \mid \theta)$: complete-data likelihood

Expectation–Maximisation (EM) algorithm

complete data $y_C = (y_O, y_U)$

parameter θ

Expectation–Maximisation algorithm

initialize parameter: $\theta \leftarrow \theta^{(0)}$

for $t = 1$ to ...

E step

$$Q(\theta \mid \theta^{(t)}) = \mathbb{E}\left\{\log f(y_C \mid \theta) \mid y_O, \theta^{(t)}\right\}$$

$Q(\theta \mid \theta^{(t)})$: conditional expectation
$f(y_C \mid \theta)$: complete-data likelihood

M step

$$\theta^{(t+1)} = \underset{\theta}{\operatorname{argmax}} Q(\theta \mid \theta^{(t)})$$

Expectation–Maximisation (EM) algorithm

Expectation step: given the current estimate $\theta^{(t)}$ of θ, calculate $Q(\theta \mid \theta^{(t)})$ as a function of θ

Maximisation step: determine a new estimate $\theta^{(t+1)}$ as the value of θ which maximises $Q(\theta \mid \theta^{(t)})$

Some considerations

- Ascent property M step increases $Q \Rightarrow$ increases $L(\theta) = f(y_O; \theta)$
  The log likelihood $\ell(\theta, y_C)$ is never decreased at any iteration of the algorithm.
- Convergence to local maxima choose multiple initial values
- Convergence rate $\rightarrow$ depends on the amount of missing information
- E and/or M steps may be too complicated can be broken into smaller components, sometimes involving Monte Carlo simulation to compute the conditional expectations required for the E step

Precision

How flat or peaked the likelihood is around the maximum determines the precision of the estimate.

The EM algorithm on its own does not give this and much further development is involved in getting an idea of precision via EM.

Oakes (1999) *JRSS* B

Expectation–Maximisation (EM) algorithm

Y: observed data

U: unobserved variables

Our goal is to use the observed value y of Y for inference on a parameter θ, in models where we cannot easily calculate the density

$$f(y;\theta) = \int f(y \mid u;\theta) f(u;\theta) du$$

and hence cannot readily compute the likelihood for θ based only on y.

Expectation–Maximisation (EM) algorithm

We write the complete-data log likelihood based on both y and the value u of U as

$$\log f(y, u; \theta) = \log f(y; \theta) + \log f(u \mid y; \theta), \tag{1}$$

first term on the right $\rightarrow$ observed-data log likelihood $\ell(\theta)$

As the value of U is unobserved, the best we can do is to remove it by taking expectation of (1) with respect to the conditional density $f(u \mid y; \theta)$ of U given that $Y = y$.

Expectation–Maximisation (EM) algorithm

This yields

$$\mathbb{E}\{\log f(Y, U; \theta) \mid Y = y; \theta'\} = \ell(\theta) + \mathbb{E}\{\log f(U \mid Y; \theta) \mid Y = y; \theta'\},$$

which can be expressed as

$$Q(\theta; \theta') = \ell(\theta) + C(\theta; \theta').$$

Expectation–Maximisation (EM) algorithm

EM

Starting from an initial value θ' of θ,

1. Compute $Q(\theta;\theta') = \mathbb{E}\{\log f(Y,U;\theta) \mid Y = y;\theta'\}$ E step
2. With θ' fixed, maximise $Q(\theta;\theta')$ over θ, giving θ^* M step
3. Check if the algorithm has converged, using $\ell(\theta^*) - \ell(\theta')$ if available, or $|\theta^* - \theta'|$, or both. If not, set $\theta' = \theta^*$ and repeat.

Example: *ABO* allele frequencies

Blood group: A, AB, B, O

$ABO \rightarrow$ 3 alleles

A and B codominant, both dominant over O

assume Hardy–Weinberg equilibrium and random sampling

$p_A = \frac{2N_{AA}+N_{Aa}}{2N} = n_{AA} + \frac{1}{2}n_{Aa}$

dominance $\rightarrow$ cannot distinguish the homozygotes and the heterozygotes

Example: *ABO* allele frequencies

Hardy–Weinberg equilibrium

Genotypic frequencies $\rightarrow$ always determine the allelic frequencies

The reverse is not necessarily true

Under some assumptions, for an autosomal gene with 2 alleles (A and a) with frequencies p and q respectively, we have 3 genotypes (AA, Aa and aa) whose frequencies are

p^2, $2pq$ and q^2

$= (p+q)^2$

G. H. Hardy, W. Weinberg (1908) (independently)

For more than 2 alleles:

$$\begin{aligned}(p_1 + p_2 + \ldots + p_n)^2 &= p_1^2 + 2p_1p_2 + \ldots + 2p_1p_n \\ &\quad + p_2^2 + \ldots \\ &\quad + p_n^2 + \ldots + 2p_{n-1}p_n\end{aligned}$$

Example: *ABO* allele frequencies

Phenotype (Blood group)	Genotype	Phenotype counts	Genotype counts	Expected frequency
A	$AA + AO$	n_A	$n_{AA} + n_{AO}$	$p^2 + 2pr$
B	$BB + BO$	n_B	$n_{BB} + n_{BO}$	$q^2 + 2pr$
AB	AB	n_{AB}	n_{AB}	$2pq$
O	OO	n_O	n_{OO}	r^2
Total		n	n	1

Example: *ABO* allele frequencies

History

Bernstein (1925)

$$r' = \sqrt{\frac{n_O}{n}}$$

$$\mathbb{E}\left(\frac{n_A + n_O}{n}\right) = (p + r)^2 = (1 - q)^2 \;\Rightarrow\; q' = 1 - \sqrt{\frac{n_A + n_O}{n}}$$

and similarly

$$p' = 1 - \sqrt{\frac{n_B + n_O}{n}}$$

- do not generally add up to 1

Example: *ABO* allele frequencies

History

Bernstein (1930)

$$d = 1 - (p' + q' + r')$$

therefore

$$p'' = p' \left(1 + \frac{d}{2}\right)$$

$$q'' = q' \left(1 + \frac{d}{2}\right)$$

$$r'' = \left(r + \frac{d}{2}\right) \left(1 + \frac{d}{2}\right)$$

- still don't add up to 1, but difference smaller

Example: *ABO* allele frequencies

History

Wiener (1929)

$$r''' = \sqrt{\frac{n_O}{n}}$$

$$\mathbb{E}\left(\frac{n_A + n_O}{n}\right) = (p + r)^2 \;\Rightarrow\; p''' = \sqrt{\frac{n_A + n_O}{n}} - \sqrt{\frac{n_O}{n}}$$

and similarly

$$q''' = \sqrt{\frac{n_B + n_O}{n}} - \sqrt{\frac{n_O}{n}}$$

- don't add up either

Example: *ABO* allele frequencies

Phenotype (Blood group)	Genotype	Phenotypic counts	Genotypic counts	Expected frequency
A	$AA + AO$	n_A	$n_{AA} + n_{AO}$	$p^2 + 2pr$
B	$BB + BO$	n_B	$n_{BB} + n_{BO}$	$q^2 + 2pr$
AB	AB	n_{AB}	n_{AB}	$2pq$
O	OO	n_O	n_{OO}	r^2
Total		n	n	1

Example: *ABO* allele frequencies

Complete data: $X = (n_{AA}, n_{AO}, n_{BB}, n_{BO}, n_{AB}, n_{OO})$ (genotype counts)

Observed data: $Y = (n_A, n_B, n_{AB}, n_O)$ (phenotype counts)

n (total number of individuals in the sample), in terms of the complete data

n_A in terms of the complete data

n_B in terms of the complete data

n_{AB} in terms of the complete data

n_O in terms of the complete data

complete data likelihood – assume HWE

Example: *ABO* allele frequencies

If a person has type A (B), the underlying genotype could be either AA (BB) or AO (BO).

$n_A = n_{AA} + n_{AO}$

$n_B = n_{BB} + n_{BO}$

$n_{AB} = n_{AB}$

$n_O = n_{OO}$

The likelihood for the complete data is simple:

$(n_{AA}, n_{AO}, n_{BB}, n_{BO}, n_{AB}, n_{OO}) \sim \text{Multinomial}(n, p^2, 2pr, q^2, 2pr, 2pq, r^2)$

Example: *ABO* allele frequencies

Phenotype (Blood group)	Genotype	Phenotypic counts	Genotypic counts	Expected frequency
A	$AA + AO$	$n_A = 186$	$n_{AA} + n_{AO}$	$p^2 + 2pr$
B	$BB + BO$	$n_B = 38$	$n_{BB} + n_{BO}$	$q^2 + 2pr$
AB	AB	$n_{AB} = 13$	n_{AB}	$2pq$
O	OO	$n_O = 284$	n_{OO}	r^2
Total		n	n	1

Clarke et al. (1959) *BMJ*

$p =?$

$q =?$

$r =?$

Example: *ABO* allele frequencies

Available data are incomplete

$n_{AA}, n_{AO}, n_{BB}, n_{BO}$ are unknown

Observed data: $n_O = (n_A, n_B, n_{AB}, n_O)$

Unobserved data: $n_U = (n_{AA}, n_{AO}, n_{BB}, n_{BO})$

Complete data: $n_C = (n_{AA}, n_{AO}, n_{BB}, n_{BO}, n_{AB}, n_{OO})$

$n_{AA} + n_{AO} = n_A$

$n_{BB} + n_{BO} = n_B$

$n_O = n_{OO}$

Example: *ABO* allele frequencies

EM algorithm

- Start from estimates of the allele frequencies and use them to calculate the expected frequencies of all genotypes (step E), assuming Hardy–Weinberg equilibrium
- Use these 'hypothetical'/'augmented' complete genotypic frequencies to obtain new estimates of the allele frequencies, using maximum likelihood (step M)

Then use these new allele frequency estimates in a new E step, iterating until the values converge.

History

Ceppellini et al. (1955) – estimation of gene frequencies

Dempster et al. (1977)

Other applications

- Allele frequencies – recessive
- Haplotype frequencies
- Welch–Baum algorithm – Hidden Markov Models
- RNAseq
- Mixture densities

Haplotypes

Co-dominant markers:

- count number of chromosomes (e.g. $2N$)
- count number of alleles (e.g. n)
- allele frequency $\rightarrow$ simple proportion ($n/2N$)

Unphased genotypes of individuals from some populations, where each unphased genotype consists of unordered pairs of SNPs taken from homologous chromosomes of the individual

Haplotypes: contiguous blocks of SNPs inherited from a single chromosome

Unknown phase

Haplotypes can't always be counted directly – focusing on unambiguous genotypes introduces bias

Excoffier and Slatkin (1995) *Mol Biol Evol*

Haplotypes

Assuming that each individual's genotype is a combination of two haplotypes (one maternal and one paternal), the goal of haplotype inference is to determine a small set of haplotypes that best explain all of the unphased genotypes observed in the population

Haplotypes

Assuming that each individual's genotype is a combination of two haplotypes (one maternal and one paternal), the goal of haplotype inference is to determine a small set of haplotypes that best explain all of the unphased genotypes observed in the population

Observed data: unphased genotypes

Latent variables: assignments of unphased genotypes to pairs of haplotypes

Parameters: frequencies of each haplotype in the population

E step: using the current haplotype frequencies to estimate probability distributions over phasing assignemnts for each unphased genotype

M step: using the expected phasing assignments to refine estimates of haplotype frequencies

RNA-Seq

Challenges in transcript assembly and abundance estimation, arising from the ambiguous assignment of reads to isoforms

Baum–Welch algorithm

$\rightarrow$ estimate parameters of a Hidden Markov Model (HMM)

EM algorithm

e.g. Copy Number Variants (CNVs)

- uses the forward-backward algorithm
- HMM: joint probability of a collection of 'hidden' (latent) and observed discrete random variables
  - ith hidden variable given the $(i-1)$th hidden variable is independent of previous hidden variables
  - current observed variables depend only on the current hidden state

Clustering

E step of EM algorithm $\rightarrow$ assigns clusters to each data point based in its relative density under each mixture component

M step $\rightarrow$ recomputes the component density parameters based on the current clusters

k mixture components, each with a Gaussian density

'soft' version of k-means clustering

probabilistic (rather than deterministic) assignments of points to cluster centers

Variational Bayes

Variational Bayes $\rightarrow$ family of techniques for approximating intractable integrals arising in Bayesian inference and machine learning

Typically used in complex statistical models consisting of observed and latent variables and unknown parameters, as might be described by a graphical model

- To provide an approximation to the posterior probability of the unobserved variables

  Alternative to Monte Carlo sampling methods (particularly, Markov chain Monte Carlo methods such as Gibbs sampling) for a Bayesian approach for complex distributions that are difficult to directly evaluate or sample from

- To derive a lower bound for the marginal likelihood of the observed data

  Used for model selection – higher marginal likelihood for a given model indicates a better fit (thus greater probability that the model in question was the one that generated the data).

Variational Bayes

Can be seen as an extension of the EM algorithm from maximum a posteriori (MAP) estimation of the single most probable value of each parameter to fully Bayesian estimation which computes (an approximation to) the entire posterior distribution of the parameters and latent variables.

It finds a set of optimal parameter values, and it has the same alternating structure as does EM, based on a set of equations that cannot be solved analytically.

Discussion

- The EM algorithm enables parameter estimation in probabilistic models with incomplete data.
- Can be used in cases where a complex problem can be formulated in terms of a simple problem if the data were complete.
  If log likelihood of the data that would have been observed if complete has an appreciably simpler functional form than that of the data actually observed, then EM useful.
- Many variations.
- Standard errors not produced directly.

References


Ceppellini, R., Siniscalco, M. and Smith, C. A. B. (1955). The estimation of gene frequencies in a random-mating population *Annals of Human Genetics* **20**, 97–115.


Clarke, C. A., Price Evans, D. A., McConnell, R. B. and Sheppard, P. M. (1959). Secretion of blood group antigens and peptic ulcer. *British Medical Journal* **1**, 603–607.


Cox, D. R. and Oakes, D. (1984). *Analysis of Survival Data.* Chapman and Hall / CRC.


Davison, A. C. (2003). *Statistical Models.* Cambridge University Press.


Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, Series B, **39** (1), 1–38.


Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The elements of statistical learning: Data Mining, Inference, and Prediction* (Vol. 2, No. 1). New York: Springer.


Excoffier, L. and Slatkin, M. (1995). Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Molecular Biology and Evolution* **12** (5), 921–927.


Oakes, D. (1999). Direct calculation of the information matrix via the EM algorithm. *ournal of the Royal Statistical Society*, Series B, **61** (2), 479–482.