Maximum likelihood and EM algorithm (after the Chapter 8)

Pasha Zusmanovich, deCODE

Statistics Colloquium
March 30, 2007

What is likelihood and what it is good for?

Likelihood is just a conditional probability.

Formal definition

Given random events A and B, the **likelihood function** of A relative to B is:

$$\{\text{set of states of } B\} \to [0,1]$$
$$x \mapsto Pr(A \mid B = x).$$

Nothing fancy so far. Consider an ...

What is likelihood and what it is good for?

Example: alleles and genotypes

frequencies of alleles:
a: θ
A: $1 - \theta$

What is likelihood and what it is good for?

Example: alleles and genotypes

frequencies of alleles:
a: θ
A: $1-\theta$

$\Longrightarrow$

frequencies of genotypes:
aa: θ^2
aA: $2\theta(1-\theta)$
AA: $(1-\theta)^2$

What is likelihood and what it is good for?

Example: alleles and genotypes

frequencies of alleles:		frequencies of genotypes:	numbers:
a: θ	$\Longrightarrow$	aa: θ^2	n_{aa}
A: $1-\theta$		aA: $2\theta(1-\theta)$	n_{aA}
		AA: $(1-\theta)^2$	n_{AA}

The probability that numbers of genotypes would be exactly (n_{aa}, n_{aA}, n_{AA}):

$$f(\theta) = \frac{(n_{aa}+n_{aA}+n_{AA})!}{n_{aa}!n_{aA}!n_{AA}!}\theta^{2n_{aa}}(2\theta(1-\theta))^{n_{aA}}(1-\theta)^{2n_{AA}}$$

f is a likelihood function:
{ probability of alleles } $\longrightarrow$ { conditional probability of genotypes assuming given probability of alleles }.

What is likelihood and what it is good for?

Example: alleles and genotypes

frequencies of alleles:		frequencies of genotypes:	numbers:
a: θ	$\Longrightarrow$	aa: θ^2	n_{aa}
A: $1-\theta$		aA: $2\theta(1-\theta)$	n_{aA}
		AA: $(1-\theta)^2$	n_{AA}

The probability that numbers of genotypes would be exactly (n_{aa}, n_{aA}, n_{AA}):

$$f(\theta) = \frac{(n_{aa}+n_{aA}+n_{AA})!}{n_{aa}!n_{aA}!n_{AA}!}\theta^{2n_{aa}}(2\theta(1-\theta))^{n_{aA}}(1-\theta)^{2n_{AA}}$$

f is a likelihood function:
{ probability of alleles } $\rightarrow$ { conditional probability of genotypes assuming given probability of alleles }.

This is a model with parameter θ.
Question: Which parameter makes model the "best"?
Answer ...

What is likelihood and what it is good for?

Example: alleles and genotypes (continued)

Question: Which parameter makes model the "best"?
Answer: Those which makes the observed data more likely, i.e. which maximizes

$$f(\theta) = \frac{(n_{aa} + n_{aA} + n_{AA})!}{n_{aa}! n_{aA}! n_{AA}!} \theta^{2n_{aa}} (2\theta(1-\theta))^{n_{aA}} (1-\theta)^{2n_{AA}}$$

on $[0, 1]$.
Solution:

$$\hat{\theta} = \frac{2n_{aa} + n_{aA}}{2(n_{aa} + n_{aA} + n_{AA})}.$$

What is likelihood and what it is good for?

Example: alleles and genotypes (continued)

Question: Which parameter makes model the "best"?
Answer: Those which makes the observed data more likely, i.e. which maximizes

$$f(\theta) = \frac{(n_{aa} + n_{aA} + n_{AA})!}{n_{aa}! n_{aA}! n_{AA}!} \theta^{2n_{aa}} (2\theta(1-\theta))^{n_{aA}} (1-\theta)^{2n_{AA}}$$

on $[0, 1]$.
Solution:

$$\hat{\theta} = \frac{2n_{aa} + n_{aA}}{2(n_{aa} + n_{aA} + n_{AA})}.$$

But this is exactly the Hardy-Weinberg equilibrium!

What is likelihood and what it is good for?

Another example: linear regression

Fitting a line to the set of points on the plane $\{(x_1, y_1), \ldots, (x_n, y_n)\}$, assuming observations are independent, and errors are normally distributed. The model is:

$$Y = \beta_1 X + \beta_0 + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2).$$

What is the "probability" to have the observed data under the given model?

What is likelihood and what it is good for?

Another example: linear regression

Fitting a line to the set of points on the plane $\{(x_1, y_1), \ldots, (x_n, y_n)\}$, assuming observations are independent, and errors are normally distributed. The model is:

$$Y = \beta_1 X + \beta_0 + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2).$$

What is the "probability" to have the observed data under the given model?

$$P(Y \text{ lies in } \delta\text{-neighbourhood of } y_i | X = x_i) \approx \text{density}(Y)|_{X=x_i, Y=y_i} \cdot 2\delta,$$

so "probability" is replaced by density. If X is fixed,

$$Y - \beta_1 X - \beta_0 \sim N(0, \sigma^2) \Rightarrow Y \sim N(\beta_1 X + \beta_0, \sigma^2).$$

What is likelihood and what it is good for?

Another example: linear regression (continued)

Maximizing

$$\text{density}(Y)|_{X=x_i, Y=y_i} = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} exp\Big(- \frac{(\beta_1 x_i + \beta_0 - y_i)^2}{2\sigma^2}\Big)$$

$$= \Big(\frac{1}{\sqrt{2\pi}\sigma}\Big)^n exp\Big(- \frac{1}{2\sigma^2} \sum_{i=1}^{n} (\beta_1 x_i + \beta_0 - y_i)^2\Big)$$

is equivalent to minimizing

$$\sum_{i=1}^{n} (\beta_1 x_i + \beta_0 - y_i)^2.$$

What is likelihood and what it is good for?

Another example: linear regression (continued)

Maximizing

$$\text{density}(Y)|_{X=x_i, Y=y_i} = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} exp\Big(- \frac{(\beta_1 x_i + \beta_0 - y_i)^2}{2\sigma^2}\Big)$$

$$= \Big(\frac{1}{\sqrt{2\pi}\sigma}\Big)^n exp\Big(- \frac{1}{2\sigma^2} \sum_{i=1}^{n} (\beta_1 x_i + \beta_0 - y_i)^2\Big)$$

is equivalent to minimizing

$$\sum_{i=1}^{n} (\beta_1 x_i + \beta_0 - y_i)^2.$$

But this is exactly the least squares!

What is likelihood and what it is good for?

Refined formal definition

Assuming a random variable X has a density function $f(x, \theta)$ parametrized by θ, the likelihood function is:

$$\theta \mapsto f(x, \theta).$$

"Conceptual" definition

Likelihood is the probability of observed data under the given model.

Thus, the maximum likelihood correspond to the model (in the given parametrized class of models) which makes the observerd data "most likely".
One usually maximize $log\, f(x, \theta)$ instead of $f(x, \theta)$ (**log-likelihood function**). Ok, since *log* is monotonic. But ...

Why logarithm?

- Turns multiplicative things to additive.

- Diminishes the “long tail”.

Why logarithm?

- Turns multiplicative things to additive. In most cases on practice, the likelihood function is the product of several functions. E.g., if $X_1, \ldots, X_n$ are independent random variables, then their likelihood function:

$$f(x_1, \ldots, x_n, \theta) = f(x_1, \theta) \ldots f(x_n, \theta),$$

  so logarithm turns multiplicative things to additive and easier to deal with. (And logarithm is the **only** "good" function taking multiplication to addition).
- Diminishes the "long tail".

Why logarithm?

- Turns multiplicative things to additive. In most cases on practice, the likelihood function is the product of several functions. E.g., if $X_1, \ldots, X_n$ are independent random variables, then their likelihood function:

$$f(x_1, \ldots, x_n, \theta) = f(x_1, \theta) \ldots f(x_n, \theta),$$

  so logarithm turns multiplicative things to additive and easier to deal with. (And logarithm is the **only** "good" function taking multiplication to addition).
- Diminishes the "long tail". A random variable with values in $\mathbb{R}^+$ (say, results of a measurement) tends to have a skewed distribution to the right because there is lower limit but not upper limit. Passing to log diminishes this skewness.

What is likelihood and what it is good for?

Maximum likelihood behaves nicely asymtotically

Taylor series:

$$\ell(\theta) = \ell(\hat{\theta}) + \frac{1}{2}(\theta - \hat{\theta})^2 \ell''(\hat{\theta}) + \dots$$

$i(\theta) = E(-\ell''(\theta))$ – **Fisher information**.

$\hat{\theta} \sim N(\theta_0, i(\theta_0)^{-1})$ as number of samples $\to \infty$.

Could be used to assess the precision of $\hat{\theta}$.

What is likelihood and what it is good for?

Connection with some fancy areas of Mathematics

Back to alleles and genotypes example: model with **inbreeding coefficient** λ:

frequencies of alleles:	frequencies of genotypes:	numbers:
a: θ	aa: $\theta^2 + \theta(1-\theta)\lambda$	38
A: $1-\theta$	aA: $2\theta(1-\theta)(1-\lambda)$	95
	AA: $(1-\theta)^2 + \theta(1-\theta)\lambda$	53

(some real blood groups data from UK, 1947)

Scoring equations are equivalent to:

What is likelihood and what it is good for?

Connection with some fancy areas of Mathematics

Back to alleles and genotypes example: model with **inbreeding coefficient** λ:

frequencies of alleles:	frequencies of genotypes:	numbers:
a: θ	aa: $\theta^2 + \theta(1-\theta)\lambda$	38
A: $1-\theta$	aA: $2\theta(1-\theta)(1-\lambda)$	95
	AA: $(1-\theta)^2 + \theta(1-\theta)\lambda$	53

(some real blood groups data from UK, 1947)

Scoring equations are equivalent to:

$$372\theta^3\lambda^2 - 744\theta^3\lambda - 558\theta^2\lambda^2 + 372\theta^3 + 1131\theta^2\lambda + 186\theta\lambda^2 - 573\theta^2 - 668\theta\lambda + 201\theta + 148\lambda = 0;$$

$$186\theta^2\lambda^2 - 372\theta^2\lambda - 186\theta\lambda^2 + 186\theta^2 + 387\theta\lambda - 201\theta - 148\lambda + 53 = 0.$$

What is likelihood and what it is good for?

Connection with some fancy areas of Mathematics

Back to alleles and genotypes example: model with **inbreeding coefficient** λ:

frequencies of alleles:	frequencies of genotypes:	numbers:
a: θ	aa: $\theta^2 + \theta(1-\theta)\lambda$	38
A: $1-\theta$	aA: $2\theta(1-\theta)(1-\lambda)$	95
	AA: $(1-\theta)^2 + \theta(1-\theta)\lambda$	53

(some real blood groups data from UK, 1947)

Scoring equations are equivalent to:

$$372\theta^3\lambda^2 - 744\theta^3\lambda - 558\theta^2\lambda^2 + 372\theta^3 + 1131\theta^2\lambda + 186\theta\lambda^2 - 573\theta^2 - 668\theta\lambda + 201\theta + 148\lambda = 0;$$

$$186\theta^2\lambda^2 - 372\theta^2\lambda - 186\theta\lambda^2 + 186\theta^2 + 387\theta\lambda - 201\theta - 148\lambda + 53 = 0.$$

Statistics + Algebraic Geometry = **Algebraic Statistics**.

What is likelihood and what it is good for?

Advantages (to summarize)

- Agrees with intuition.
- Confirmed by other methods.
- "Nice" asymptotic behavior.
- Very good practical results.
- Universal.
- Connection with other areas of Mathematics.

What is likelihood and what it is good for?

Advantages (to summarize)

- Agrees with intuition.
- Confirmed by other methods.
- "Nice" asymptotic behavior.
- Very good practical results.
- Universal.
- Connection with other areas of Mathematics.

Disadvantages

- No "theoretical" justification.
- Could be bad for small samples.
- No way to compare "disjoint" models.
- "Bayesian" issue ...

What is likelihood and what it is good for?

"Bayesian" issue:

$$Pr(data|model) = \frac{Pr(model|data)Pr(data)}{Pr(model)}.$$

What is likelihood and what it is good for?

"Bayesian" issue:

$$Pr(data|model) = \frac{Pr(model|data)Pr(data)}{Pr(model)}.$$

Philosophical mumbo-jumbo:

- M. Forster and E. Sober, *Why likelihood?*, The Nature of Scientific Evidence (ed. M. Taper and S. Lele), Univ. of Chicago Press, 2004, 153–165
  `http://philosophy.wisc.edu/forster/Likelihood/default.htm`
- B. Fitelson, *Likelihoodism, bayesianism, and relational confirmation*, Synthese, to appear
  `http://fitelson.org/research.htm`

EM algorithm

Finding the maximum of likelihood function could be difficult.

Example: alleles and phenotypes

Assume A is **dominant**, and we observe only **phenotypes**:

frequencies of alleles:
a: θ
A: $1-\theta$

frequencies of genotypes:
aa: θ^2
aA: $2\theta(1-\theta)$
AA: $(1-\theta)^2$

numbers of phenotypes:
a: 38
A: 148

EM algorithm

Finding the maximum of likelihood function could be difficult.

Example: alleles and phenotypes

Assume A is **dominant**, and we observe only **phenotypes**:

frequencies of alleles:	frequencies of genotypes:	numbers of phenotypes:
a: θ	aa: θ^2	a: 38
A: $1-\theta$	aA: $2\theta(1-\theta)$	A: 148
	AA: $(1-\theta)^2$	

Scoring equation amounts to: $38/\theta^2 - 148/(1-\theta^2) = 0$, i.e. is biquadratic. Suppose we don't know how/don't want to solve it. What to do?

EM algorithm

Finding the maximum of likelihood function could be difficult.

Example: alleles and phenotypes

Assume A is **dominant**, and we observe only **phenotypes**:

frequencies of alleles:	frequencies of genotypes:	numbers of phenotypes:
a: θ	aa: θ^2	a: 38
A: $1-\theta$	aA: $2\theta(1-\theta)$	A: 148
	AA: $(1-\theta)^2$	

Scoring equation amounts to: $38/\theta^2 - 148/(1-\theta^2) = 0$, i.e. is biquadratic. Suppose we don't know how/don't want to solve it.
What to do?
Introduce back missing numbers n_{aA} and n_{AA} (**hidden parameters**) and iterate.

EM algorithm

Example: alleles and phenotypes (continued)

EM algorithm

Example: alleles and phenotypes (continued)

E Step 1: initial genotype numbers: $n_{aA} = n_{AA} = 148/2 = 74.00$

EM algorithm

Example: alleles and phenotypes (continued)

E Step 1: initial genotype numbers: $n_{aA} = n_{AA} = 148/2 = 74.00$

M Step 2: find MLE for those numbers: $\theta = (2 \cdot 38 + 74.00)/(2 \cdot 186) = 0.40$

EM algorithm

Example: alleles and phenotypes (continued)

E Step 1: initial genotype numbers: $n_{aA} = n_{AA} = 148/2 = 74.00$

M Step 2: find MLE for those numbers: $\theta = (2 \cdot 38 + 74.00)/(2 \cdot 186) = 0.40$

E Step 3: for $\theta = 0.40$, find genotype frequencies: for aA: $2{\cdot}0.40{\cdot}(1{-}0.40) = 0.48$ and for AA: $(1{-}0.40)^2 = 0.36$, and for them, genotype numbers: $n_{aA} = 186 \cdot 0.48 = 89.28$, $n_{AA} = 148 - 89.28 = 58.72$

EM algorithm

Example: alleles and phenotypes (continued)

E Step 1: initial genotype numbers: $n_{aA} = n_{AA} = 148/2 = 74.00$

M Step 2: find MLE for those numbers: $\theta = (2 \cdot 38 + 74.00)/(2 \cdot 186) = 0.40$

E Step 3: for $\theta = 0.40$, find genotype frequencies: for aA: $2{\cdot}0.40{\cdot}(1{-}0.40) = 0.48$ and for AA: $(1{-}0.40)^2 = 0.36$, and for them, genotype numbers: $n_{aA} = 186 \cdot 0.48 = 89.28$, $n_{AA} = 148 - 89.28 = 58.72$

M Step 4: find MLE for those numbers: $\theta = (2 \cdot 38 + 89.28)/(2 \cdot 186) = 0.44$

EM algorithm

Example: alleles and phenotypes (continued)

E Step 1: initial genotype numbers: $n_{aA} = n_{AA} = 148/2 = 74.00$

M Step 2: find MLE for those numbers: $\theta = (2 \cdot 38 + 74.00)/(2 \cdot 186) = 0.40$

E Step 3: for $\theta = 0.40$, find genotype frequencies: for aA: $2{\cdot}0.40{\cdot}(1{-}0.40) = 0.48$ and for AA: $(1{-}0.40)^2 = 0.36$, and for them, genotype numbers: $n_{aA} = 186 \cdot 0.48 = 89.28$, $n_{AA} = 148 - 89.28 = 58.72$

M Step 4: find MLE for those numbers: $\theta = (2 \cdot 38 + 89.28)/(2 \cdot 186) = 0.44$

E Step 5: for $\theta = 0.44$, find genotype frequencies: for aA: $2{\cdot}0.44{\cdot}(1{-}0.44) = 0.49$ and for AA: $(1{-}0.44)^2 = 0.31$ and genotype numbers: $n_{aA} = 186 \cdot 0.49 = 91.14$, $n_{AA} = 148 - 91.14 = 56.86$

EM algorithm

Example: alleles and phenotypes (continued)

E Step 1: initial genotype numbers: $n_{aA} = n_{AA} = 148/2 = 74.00$

M Step 2: find MLE for those numbers: $\theta = (2 \cdot 38 + 74.00)/(2 \cdot 186) = 0.40$

E Step 3: for $\theta = 0.40$, find genotype frequencies: for aA: $2{\cdot}0.40{\cdot}(1{-}0.40) = 0.48$ and for AA: $(1{-}0.40)^2 = 0.36$, and for them, genotype numbers: $n_{aA} = 186 \cdot 0.48 = 89.28$, $n_{AA} = 148 - 89.28 = 58.72$

M Step 4: find MLE for those numbers: $\theta = (2 \cdot 38 + 89.28)/(2 \cdot 186) = 0.44$

E Step 5: for $\theta = 0.44$, find genotype frequencies: for aA: $2{\cdot}0.44{\cdot}(1{-}0.44) = 0.49$ and for AA: $(1{-}0.44)^2 = 0.31$ and genotype numbers: $n_{aA} = 186 \cdot 0.49 = 91.14$, $n_{AA} = 148 - 91.14 = 56.86$

M Step 6: find MLE for those numbers: $\theta = (2 \cdot 38 + 91.14)/(2 \cdot 186) = 0.44$

EM algorithm

Example: alleles and phenotypes (continued)

E Step 1: initial genotype numbers: $n_{aA} = n_{AA} = 148/2 = 74.00$

M Step 2: find MLE for those numbers: $\theta = (2 \cdot 38 + 74.00)/(2 \cdot 186) = 0.40$

E Step 3: for $\theta = 0.40$, find genotype frequencies: for aA: $2 \cdot 0.40 \cdot (1-0.40) = 0.48$ and for AA: $(1-0.40)^2 = 0.36$, and for them, genotype numbers: $n_{aA} = 186 \cdot 0.48 = 89.28$, $n_{AA} = 148 - 89.28 = 58.72$

M Step 4: find MLE for those numbers: $\theta = (2 \cdot 38 + 89.28)/(2 \cdot 186) = 0.44$

E Step 5: for $\theta = 0.44$, find genotype frequencies: for aA: $2 \cdot 0.44 \cdot (1-0.44) = 0.49$ and for AA: $(1-0.44)^2 = 0.31$ and genotype numbers: $n_{aA} = 186 \cdot 0.49 = 91.14$, $n_{AA} = 148 - 91.14 = 56.86$

M Step 6: find MLE for those numbers: $\theta = (2 \cdot 38 + 91.14)/(2 \cdot 186) = 0.44$

Stop!

EM algorithm

Advantages

- Reduces MLE problem to another more manageable (MLE) problem.
- Agrees with results obtained by other means.
- Works on practice.

EM algorithm

Advantages

- Reduces MLE problem to another more manageable (MLE) problem.
- Agrees with results obtained by other means.
- Works on practice.

Disadvantages

- No theoretical justification.

Maximum likelihood and ME algorithm at deCODE

Associations studies

`nemo` by Daníel Gudbjartsson.

Typical input data: list of affected and unaffected individuals, list of markers (e.g. SNPs), list of genotypes (per marker and per individual).

Maximum likelihood and ME algorithm at deCODE

Associations studies

`nemo` by Daníel Gudbjartsson.

Typical input data: list of affected and unaffected individuals, list of markers (e.g. SNPs), list of genotypes (per marker and per individual).

Haplotypes inference from genotypes

Maximum parsimony vs. maximum likelihood.

Maximum likelihood and ME algorithm at deCODE

Associations studies

`nemo` by Daníel Gudbjartsson.

Typical input data: list of affected and unaffected individuals, list of markers (e.g. SNPs), list of genotypes (per marker and per individual).

Haplotypes inference from genotypes

Maximum parsimony vs. maximum likelihood.

Example (0,1 – homozygote, 2 – heterozygote):

genotypes:

2120

2102

1221

Maximum likelihood and ME algorithm at deCODE

Associations studies

`nemo` by Daníel Gudbjartsson.

Typical input data: list of affected and unaffected individuals, list of markers (e.g. SNPs), list of genotypes (per marker and per individual).

Haplotypes inference from genotypes

Maximum parsimony vs. maximum likelihood.

Example (0,1 – homozygote, 2 – heterozygote):

genotypes:		parsimonial solution:
2120	$\Longleftarrow$	0100 + 1110
2102		0100 + 1101
1221		1011 + 1101

That's all.

Slides at `http://justpasha.org/tmp/presentation.pdf` .