Missing Data and the EM algorithm

MSc Further Statistical Methods, Lecture 4
Hilary Term 2006

Steffen Lauritzen, University of Oxford; January 24, 2006

Missing data problems

case	A	B	C	D	E	F
1	a_1	b_1	$*$	d_1	e_1	$*$
2	a_2	$*$	c_2	d_2	e_2	$*$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	a_n	b_n	c_n	$*$	$*$	$*$

$*$ or *NA* denotes values that are *missing*, i.e. non-observed.

Examples of missingness

- non-reply in surveys, "missing" don't know, essentially an additional state for the variable in question
- recording error
- variable out of range
- just not recorded (e.g. too expensive)

Different types of missingness demand different treatment.

Notation for missingness

Data matrix Y, *missing data matrix* $M = \{M_{ij}\}$:

$$M_{ij} = \begin{cases} 1 & \text{if } Y_{ij} \text{ is missing} \\ 0 & \text{if } Y_{ij} \text{ is observed.} \end{cases}$$

Convenient to introduce the notation $Y = (Y_{\text{obs}}, Y_{\text{mis}})$, where Y_{mis} are conceptual and denote the data that were not observed.

Patterns of missingness

Univariate: $M_{ij} = 0$ unless $j = j^*$, e.g. an unmeasured response

Multivariate: $M_{ij} = 0$ unless $j \in J \subset V$, as above, just with multivariate response, e.g. in surveys

Monotone: There is an ordering of V so $M_{ik} = 0$ implies $M_{ij} = 0$ for $j < k$, e.g. drop-out in longitudinal studies.

Disjoint: Two subsets of variables never observed together. Controversial. Appears in Rubin's causal model.

General: none of the above. Haphazardly scattered missing values.

Latent: A certain variable is never observed. Maybe it is even unobservable.

Methods for analysis tend to get increasingly complex as we go down the list.

Methods for dealing with missing data

Complete case analysis: analyse only cases where all variables are observed. Can be adequate if most cases are present, but will generally give serious biases in the analysis. In survey's, for example, this corresponds to making inference about the population of responders, not the full population;

Weighting methods. For example, if a population total $\mu = \mathbf{E}(Y)$ should be estimated and unit i has been selected with probability π_i a standard method is the *Horwitz–Thompson estimator*

$$\hat{\mu} = \frac{\sum \frac{Y_i}{\pi_i}}{\sum \frac{1}{\pi_i}}.$$

To correct for non-response, one could let ρ_i be the response-probability, estimate this in some way as $\hat{\rho}_i$ and then let

$$\tilde{\mu} = \frac{\sum \frac{Y_i}{\pi_i \hat{\rho}_i}}{\sum \frac{1}{\pi_i \hat{\rho}_i}}.$$

Imputation methods: Find ways of estimating the values of the unobserved values as $\hat{Y}_{\text{mis}}$, then proceed as if there were complete data. Without care, this can give misleading results, in particular because the "sample size" can be grossly overestimated.

Model-based likelihood methods: Model the missing data mechanism and then proceed to make a proper likelihood-based analysis, either via the method of maximum-likelihood or using Bayesian methods. This

appears to be the most sensible way.

Typically this approach was not computationally feasible in the past, but modern algorithms and computers have changed things completely. Ironically, the efficient algorithms are indeed based upon imputation of missing values, but with proper corrections resulting.

Mechanisms of missingness

The data are *missing completely at random*, MCAR, if

$$f(M \mid Y, \theta) = f(M \mid \theta), \text{ i.e. } M \perp\!\!\!\perp Y \mid \theta.$$

Heuristically, the values of Y have themselves no influence on the missingness. Example is recording error, latent variables, and variables that are missing *by design* (e.g. measuring certain values only for the first m out of n cases). Beware: it may be counterintuitive that *missing by design is MCAR*.

The data are *missing at random*, MAR, if

$$f(M \mid Y, \theta) = f(M \mid Y_{\text{obs}}, \theta), \text{ i.e. } M \perp\!\!\!\perp Y_{\text{mis}} \mid (Y_{\text{obs}}, \theta).$$

Heuristically, only the observed values of Y have influence on the missingness. By design, e.g. if individuals with certain characteristics of Y_{obs} are not included in part of study (where Y_{mis} is measured).

The data are *not missing at random*, NMAR, in all other cases.

For example, if certain values of Y cannot be recorded when they are out of range, e.g. in survival analysis.

The classifications above of the mechanism of missingness lead again to increasingly complex analyses.

It is not clear than the notion MCAR is helpful, but MAR is. Note that *if data are MCAR, they are also MAR.*

Likelihood-based methods

The most convincing treatment of missing data problems seems to be via modelling the missing data mechanism, i.e. *by considering the missing data matrix M as an explicit part of the data.*

The likelihood function then takes the form

$$L(\theta \,|\, M, y_{\text{obs}}) \propto \int f(M, y_{\text{obs}}, y_{\text{mis}} \,|\, \theta)\, dy_{\text{mis}} \qquad (1)$$

with

$$f(M, y_{\text{obs}}, y_{\text{mis}} \,|\, \theta) \propto L_{\text{mis}}(\theta) f(y_{\text{obs}}, y_{\text{mis}} \,|\, \theta), \qquad (2)$$

where the term $L_{\text{mis}}(\theta) \propto f(M \,|\, y_{\text{obs}}, y_{\text{mis}}, \theta)$ is based on an explicit model for the missing data mechanism.

Ignoring the missing data mechanism

The likelihood function *ignoring the missing data mechanism* is

$$L_{\text{ign}}(\theta \,|\, y_{\text{obs}}) \propto f(y_{\text{obs}} \,|\, \theta) = \int f(y_{\text{obs}}, y_{\text{mis}} \,|\, \theta)\, dy_{\text{mis}}. \quad (3)$$

When is $L \propto L_{\text{ign}}$ so the missing data mechanism can be ignored for further analysis? We will show this is true under the *following conditions:*

1. The data are *MAR*;
2. The parameters η governing the missingness are *separate* from parameters of interest ψ, i.e. the parameters vary in a product region, so that

information about the value of one does not restrict the other.

Ignorable missingness

If data are MAR and the missingness parameter is separate from the parameter of interest, we have $\theta = (\eta, \psi)$ and

$$L_{\text{mis}}(\theta) = L_{\text{mis}}(\eta) \propto f(M \,|\, y_{\text{obs}}, y_{\text{mis}}, \eta) = f(M \,|\, y_{\text{obs}}, \eta)$$

Hence, the factor L_{mis} is constant in (2) and can be taken outside in the integral in (1) so that, combining with (3) and

$$f(y_{\text{obs}}, y_{\text{mis}} \,|\, \theta) = f(y_{\text{obs}}, y_{\text{mis}} \,|\, \psi)$$

we get

$$L(\theta \,|\, M, y_{\text{obs}}) \propto L_{\text{mis}}(\eta) L_{\text{ign}}(\psi \,|\, y_{\text{obs}})$$

which shows that the missingness mechanism can be ignored when concerned with likelihood inference about ψ.

For a Bayesian analysis the parameters must in addition be *independent w.r.t. the prior:*

$$f(\eta, \psi) = f(\eta) f(\psi).$$

If the data are *NMAR* or the parameters are not separate, then the missing data mechanism cannot be ignored, care must be taken to model the mechanism $f(M \mid y_{\text{obs}}, y_{\text{mis}}, \theta)$ and the corresponding likelihood term must be properly included in the analysis.

The EM algorithm

The EM algorithm is an alternative to Newton–Raphson or the method of scoring for computing MLE in cases where the complications in calculating the MLE are due to *incomplete observation* and data are *MAR*, missing at random, with *separate parameters* for observation and the missing data mechanism, so the missing data mechanism can be ignored.

Data (X, Y) are the *complete data* whereas only *incomplete data* $Y = y$ are observed. (Rubin uses $Y = Y_{\text{obs}}$ and $X = Y_{\text{mis}}$).

The *complete data log-likelihood* is:

$$l(\theta) = \log L(\theta; x, y) = \log f(x, y; \theta).$$

The *marginal log-likelihood* or *incomplete data log-likelihood* is based on y alone and is equal to

$$l_y(\theta) = \log L(\theta; y) = \log f(y; \theta).$$

We wish to maximize l_y in θ but l_y is typically quite unpleasant:

$$l_y(\theta) = \log \int f(x, y; \theta)\, dx.$$

The EM algorithm is a method of maximizing the latter iteratively and alternates between two steps, one known as the *E-step* and one as the *M-step*, to be detailed below.

We let θ^* be and arbitrary but fixed value, typically the value of θ at the current iteration.

The E-step calculates the expected complete data log-likelihood ratio $q(\theta \,|\, \theta^*)$:

$$
\begin{aligned}
q(\theta \,|\, \theta^*) &= \mathbf{E}_{\theta^*}\left[\log \frac{f(X, y; \theta)}{f(X, y; \theta^*)} \,|\, Y = y\right] \\
&= \int \log \frac{f(x, y; \theta)}{f(x, y; \theta^*)} f(x \,|\, y; \theta^*)\, dx.
\end{aligned}
$$

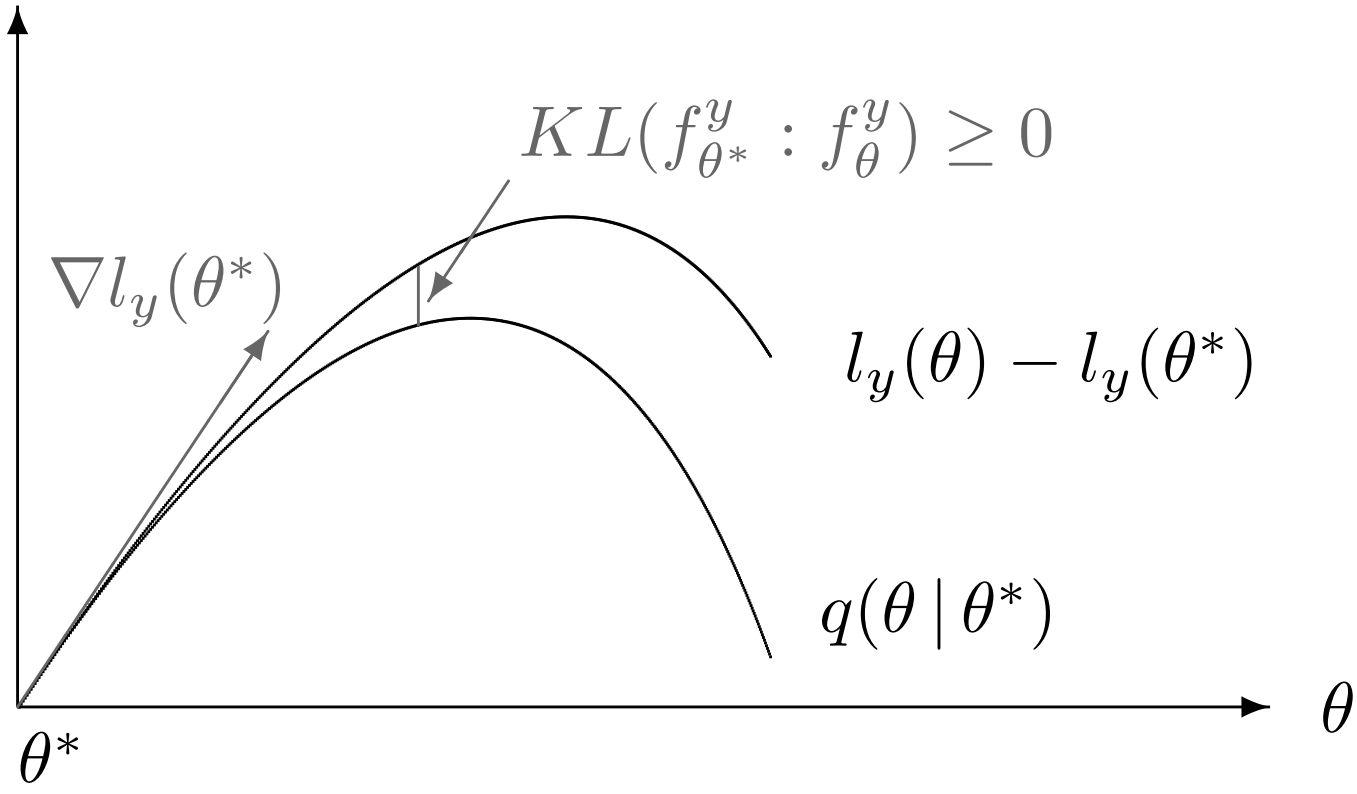
The M-step maximizes $q(\theta \,|\, \theta^*)$ in θ for for fixed θ^*, i.e. calculates

$$
\theta^{**} = \arg\max_{\theta} q(\theta \,|\, \theta^*).
$$

We will show that *After an E-step and subsequent M-step, the likelihood function has never decreased.*

The picture on the next overhead should show it all.

Expected and complete data likelihood



$$l_y(\theta) - l_y(\theta^*) = q(\theta \,|\, \theta^*) + KL(f^y_{\theta^*} : f^y_\theta)$$

$$\nabla l_y(\theta^*) = \left.\frac{\partial}{\partial\theta} l_y(\theta)\right|_{\theta=\theta^*} = \left.\frac{\partial}{\partial\theta} q(\theta \,|\, \theta^*)\right|_{\theta=\theta^*}.$$

Kullback-Leibler divergence

The *KL divergence* between f and g is

$$KL(f:g) = \int f(x) \log \frac{f(x)}{g(x)}\, dx.$$

Also known as *relative entropy* of g with respect to f.

Since $-\log x$ is a convex function, Jensen's inequality gives

$KL(f:g) \geq 0$ *and* $KL(f:g) = 0$ *if and only if* $f = g$, since

$$KL(f:g) = \int f(x) \log \frac{f(x)}{g(x)}\, dx \geq -\log \int f(x) \frac{g(x)}{f(x)}\, dx = 0,$$

so KL divergence defines an (asymmetric) distance measure between probability distributions.

Expected and marginal log-likelihood

Since $f(x \,|\, y; \theta) = f\{(x, y); \theta\}/f(y; \theta)$ we have

$$\begin{aligned} q(\theta \,|\, \theta^*) &= \int \log \frac{f(y;\theta) f(x \,|\, y;\theta)}{f(y;\theta^*) f(x \,|\, y;\theta^*)} f(x \,|\, y;\theta^*)\, dx \\ &= \log f(y;\theta) - \log f(y;\theta^*) \\ &\quad + \int \log \frac{f(x \,|\, y;\theta)}{f(x \,|\, y;\theta^*)} f(x \,|\, y;\theta^*)\, dx \\ &= l_y(\theta) - l_y(\theta^*) - KL(f^y_{\theta^*} : f^y_\theta). \end{aligned}$$

Since the KL-divergence is minimized for $\theta = \theta^*$, differentiation of the above expression yields

$$\left.\frac{\partial}{\partial\theta} q(\theta \,|\, \theta^*)\right|_{\theta=\theta^*} = \left.\frac{\partial}{\partial\theta} l_y(\theta)\right|_{\theta=\theta^*}.$$

Let now $\theta_0 = \theta^*$ and define the iteration

$$\theta_{n+1} = \arg\max_{\theta} q(\theta \,|\, \theta_n).$$

Then

$$\begin{aligned} l_y(\theta_{n+1}) &= l_y(\theta_n) + q(\theta_{n+1} \,|\, \theta_n) + KL(f^y_{\theta_{n+1}} : f^y_{\theta_n}) \\ &\geq l_y(\theta_n) + 0 + 0. \end{aligned}$$

So the log-likelihood never decreases after a combined E-step and M-step.

It follows that any limit point must be a saddle point or a local maximum of the likelihood function.

Mixtures

Consider a sample $Y = (Y_1, \ldots, Y_n)$ from individual densities

$$f(y; \alpha, \mu) = \{\alpha\phi(y - \mu) + (1 - \alpha)\phi(y)\}$$

where ϕ is the normal density

$$\phi(y) = \frac{1}{\sqrt{2\pi}} e^{-y^2/2}$$

and α and μ are both unknown, $0 < \alpha < 1$.

This corresponds to a fraction α of the observations being contaminated, or originating from a different population.

Incomplete observation

The likelihood function becomes

$$L_y(\alpha, \mu) = \prod_i \{\alpha\phi(y_i - \mu) + (1 - \alpha)\phi(y_i)\}$$

is quite unpleasant, although both Newton–Raphson and the method of scoring can be used.

But suppose we knew which observations came from which population?

In other words, let $X = (X_1, \ldots, X_n)$ be i.i.d. with $P(X_i = 1) = \alpha$ and suppose that the conditional distribution of Y_i given $X_i = 1$ was $\mathcal{N}(\mu, 1)$ whereas given $X_i = 0$ it was $\mathcal{N}(0, 1)$, i.e. that X_i was indicating whether Y_i was contaminated or not.

Then the marginal distribution of Y is precisely the mixture distribution and the 'complete data likelihood' is

$$\begin{aligned} L_{x,y}(\alpha,\mu) &= \prod_i \alpha^{x_i}\phi(y_i-\mu)^{x_i}(1-\alpha)^{1-x_i}\phi(y_i)^{1-x_i} \\ &\propto \alpha^{\sum x_i}(1-\alpha)^{n-\sum x_i}\prod_i \phi(y_i-\mu)^{x_i} \end{aligned}$$

so taking logarithms we get (ignoring a constant) that

$$\begin{aligned} l_{x,y}(\alpha,\mu) &= \sum x_i \log\alpha + \left(n-\sum x_i\right)\log(1-\alpha) \\ &\quad -\sum_i x_i(y_i-\mu)^2/2. \end{aligned}$$

If we did not know how to maximize this explicitly,

differentiation easily leads to:

$$\hat{\alpha} = \sum x_i/n, \quad \hat{\mu} = \sum x_i y_i / \sum x_i.$$

Thus, when complete data are available the frequency of contaminated observations is estimated by the observed frequency and the mean μ of these is estimated by the average among the contaminated observations.

E-step and M-step

By taking expectations, we get the E-step as

$$\begin{aligned} q(\alpha, \mu \,|\, \alpha^*, \mu^*) &= \mathbf{E}_{\alpha^*,\mu^*}\{l_{X,y}(\alpha, \mu) \,|\, Y = y\} \\ &= \sum x_i^* \log \alpha + \left(n - \sum x_i^*\right) \log(1 - \alpha) \\ &\quad - \sum_i x_i^* (y_i - \mu)^2/2 \end{aligned}$$

where

$$x_i^* = \mathbf{E}_{\alpha^*,\mu^*}(X_i \,|\, Y_i = y_i) = P_{\alpha^*,\mu^*}(X_i = 1 \,|\, Y_i = y_i).$$

Since this has the same form as the complete data likelihood, just with x_i^* replacing x_i, the M-step simply

becomes

$$\alpha^{**} = \sum x_i^*/n, \quad \mu^{**} = \sum x_i^* y_i / \sum x_i^*,$$

i.e. here the mean of the contaminated observations is estimated by a weighted average of all the observations, the weight being proportional to the probability that this observation is contaminated. In effect, x_i^* act as *imputed values* of x_i.

The imputed values x_i^* needed in the E-step are calculated as follows:

$$\begin{aligned} x_i^* &= \mathbf{E}(X_i \,|\, Y_i = y_i) = P(X_i = 1 \,|\, Y_i = y_i) \\ &= \frac{\alpha^* \phi(y_i - \mu^*)}{\alpha^* \phi(y_i - \mu^*) + (1 - \alpha^*)\phi(y_i)}. \end{aligned}$$