T-61.5040

Learning Models and Methods, (5 cr) L

Spring 2007

Lecturer: **Petteri Pajunen**

Course Assistant: **Ville Viitaniemi**

LECTURE 11: 4.4.2007

MISSING DATA

Due to Easter Holiday, there are no exercises on 6.4 and no lecture on 11.4

Missing Data

- consider multivariate data $y_i \in \mathbb{R}^d$
- observe a few such vectors

$$y = (y_1, y_2, y_3, y_4, y_5) = \begin{bmatrix} 1 & 2 & 3 & 4 & 5 \\ 2 & 3 & 1 & 0 & 7 \\ 3 & 7 & 2 & 5 & 8 \end{bmatrix}$$

- likelihood: $p(y|\theta) = \prod_i p(y_i|\theta)$

- but what if your data is

$$\begin{aligned} y &= (y_1, y_2, y_3, y_4, y_5) \\ &= \begin{bmatrix} 1 & ? & 3 & 4 & ? \\ 2 & ? & 1 & 0 & 7 \\ 1 \text{ or } 7 & 7 & \text{in } [2,5] & \text{less than } 9 & 8 \end{bmatrix} \end{aligned}$$

- some components of y_i are *missing* (value is not a single number)
- this kind of data is not uncommon in real problems

- with missing data, the posterior $p(\theta|y)$ depends partly on *unobserved data*
- a heuristic solution: discard all vectors y_i that are not completely observed and proceed as usual
- but in the example, no vector is completely observed
- here, and often in practice, simply throwing away data is not optimal

- another heuristic: keep all the observed parts of y (call them y_{obs}), even if this means keeping only parts of vectors
- not generally optimal for two reasons:
  - the discarded components may carry some information directly, such as "less than 9"
  - the fact that a value is missing may give information about the nonobserved value, and therefore information about θ

- *missing values can contain lots of information*
- in predicting the absorbtion of an orally administered drug, bloodsamples are taken and the drug concentration is analyzed
- i.e. data consists of nonnegative values measured at certain times $t_1, t_2, \ldots, t_m$

- for technical reasons, drug concentration below some small value $a > 0$ cannot be measured
- these values become missing data
- but we do know that with high probability, these values are in the very small interval $[0, a]$
- discarding these values obviously loses information

- coin toss, probability of heads is θ
- observe some coin toss results and estimate θ
- for some reason, "heads" will not be observed with probability 0.5, and "tails" with probability 0.25
- just using observed data underestimates θ

- the probabilities define a *missing data mechanism*
- in this example, it is completely defined by data (observed and missing)
- y_{obs} alone does not define it here
- it would if the missingness probabilites were equal
- then using observed data would give correct results

- *general missing data mechanism*
- extend the coin tossing example so that "heads" goes missing with unknown probability ϕ
- now the missing data mechanism depends on data as well as ϕ
- ϕ belongs to the posterior, and data gives information about it: observing lots of "heads" suggests that ϕ is small

- in principle, one could solve the problem by constructing a model for data, θ, ϕ, and "missingness indicators"
- unknown quantities are missing values, θ, and ϕ
- this is complicated in practice: heuristics are often used, or one can make sure that the missing values can be ignored

Missing Data Heuristics

- there exist many heuristics for handling models with missing data
- some of them may be optimal in some cases, but generally they are not
- exercises illustrate some cases where these heuristics fail

- *list deletion*: observed vector y_i is discarded if one or more of the components are missing
- for scalar data, this corresponds to ignoring all missing data
- if the missing data mechanism is *ignorable* (defined later), then list deletion does not bias the results
- but information is lost when the data is multivariate since some observed components are discarded

- *imputation*: guess missing values
- *mean imputation* replaces missing values with the mean of the observed values
- *regression imputation* uses regression on observed data to guess the values
- even with reasonable imputed values, these methods underestimate posterior uncertainty

- above heuristics are generally non-Bayesian (imputation can be Bayesian if several values are simulated e.g. by MCMC)
- however, in some special cases they may correspond to the Bayesian procedure
- starting from the proper Bayesian method, we see some cases where missing data can be ignored
- at the same time we see the general way of handling missing data

Bayesian Treatment of Missing Data

- denote nonrigorously the complete data by $y = (y_{obs}, y_{mis})$ (**obs**erved, **mis**sing)
- the meaning of y_{obs} is that it represents the values of all observed data (y_{mis} correspondingly)
- define the complete data y as a matrix
- any of the elements can be missing (hence the difficulty of proper notation for y_{obs} and y_{mis})

- define a matrix I with the same dimensions as y to indicate missing components
- i.e. $I_{ij} = 0$ if y_{ij} is missing and $I_{ij} = 1$ if y_{ij} is observed
- I is known, since we know which parts of y are missing

- in general, we would like to compute $p(\theta|y)$
- since $p(\theta|y) = p(\theta|y_{obs}, y_{mis})$, this posterior is useless in practice since we don't know y_{mis}
- lets review what we have:
  - knowns: y_{obs} and missingness indicators I
  - unknowns: y_{mis}, θ, and ϕ (parameters for missingness)
  - we want to know $p(\theta|y_{obs}, I)$

- the full posterior is obtained using the product rule:

$$\begin{aligned} p(\theta, \phi, y_{mis} | y_{obs}, I) &= \frac{p(\theta, \phi, y_{mis}, y_{obs}, I)}{p(y_{obs}, I)} \\ &\propto p(\theta, \phi, y, I) \\ &= p(I, y | \theta, \phi) p(\theta, \phi) \end{aligned}$$

- in the following, we assume that the complete-data likelihood can be written as

$$p(I, y | \theta, \phi) = p(I | y, \phi) p(y | \theta)$$

- meaning:
  - $p(I | y, \theta, \phi) = p(I | y, \phi)$: missingness does not depend on θ if y, ϕ are known
  - $p(y | \theta, \phi) = p(y | \theta)$: given θ, data does not depend on ϕ
- marginalizing the full posterior over y_{mis} and ϕ gives

$$p(\theta | y_{obs}, I) \propto \int \int p(I | y, \phi) p(y | \theta) p(\theta, \phi) dy_{mis} d\phi$$

Missing Data Mechanisms

- sometimes the difficult marginalization can be simplified
- this depends on properties of the *missing data mechanism*

$$p(I | y, \phi)$$

- this is in general a function of all data (observed and missing), the indicators and possibly ϕ

Missing Completely at Random

- *missing completely at random (MCAR)* holds when

$$p(I | y, \phi) = p(I | \phi)$$

- the probability does not depend on the values of y_{obs} and y_{mis}
- example: coin toss where each toss is missing with a fixed probability

Missing at Random

- *missing at random (MAR)* holds when

$$p(I|y,\phi) = p(I|y_{obs},\phi)$$

- missingness does not depend on values y_{mis}
- generic example: $y_i = (a_i, b_i)$, always observe a_i but missingness of b_i depends on a_i only
- MAR is very useful, since with another assumption it simplifies handling of missing data

Observed at Random

- *observed at random (OAR)* holds when

$$p(I|y,\phi) = p(I|y_{mis},\phi)$$

- missingness does not depend on observed values y_{obs}
- this case is not very useful on its own

Neither MAR nor OAR

- missing data is *neither MAR nor OAR* when

$$p(I|y,\phi)$$

  cannot be simplified by removing values of y
- this is the general case, and thus requires the full marginalization derived above
- sometimes called non-ignorable but this can be confusing, as seen later
- in the drug absorption example, missingness depends on the value y compared with the threshold:

model is neither MAR nor OAR

- note that in the above definitions, the given condition must hold for *all pairs* y, I
- for example in MAR, $p(I|y, \phi)$ must be a function of only those values y_{ij} for which $I_{ij} = 1$
- this must hold for any possible pair y, I

- if $p(\theta|y_{obs}) = p(\theta|y_{obs}, I)$, then the missing data mechanism can be ignored
- this missing data property is called *ignorability*
- warning! even though *neither MAR nor OAR* is called non-ignorable, it does not mean that all other cases are ignorable!

- what can we do when missing data is ignorable?
- for ignorable missing data mechanism we get

$$\begin{aligned} p(\theta|y_{obs}) &\propto p(y_{obs}|\theta)p(\theta) \\ &= \int p(y|\theta)p(\theta)dy_{mis} \end{aligned}$$

- so it is enough to integrate the missing data out from the complete data model
- without ignorability, one ends up integrating over $p(y|\theta, I)$ which is usually complicated

- ignorability can be guaranteed in certain special cases
- MAR (missing at random) means that

$$p(I|y, \phi) = p(I|y_{obs}, \phi)$$

- i.e. the missing values do not affect the indicators
- MAR alone does not guarantee ignorability!

- the indicators can still depend on ϕ
- if missing data mechanism is MAR and condition $p(\phi|\theta) = p(\phi)$ (denote as $\phi \perp\!\!\!\perp \theta$) holds, then we obtain ignorability
- antiexample: data are measurements of an object weighing θ kilos. If the scale has problems in weighting heavy objects, then ϕ (failure probability) and θ (the weight) are not independent
- model can still be MAR, since given ϕ, the missingness values I can be independent of y_{mis}

- MAR, $\phi \perp\!\!\!\perp \theta \implies$ ignorability:

$$
\begin{aligned}
p(\theta|y_{obs}, I) &\propto \int\int p(I|y,\phi)p(y|\theta)p(\theta,\phi)dy_{mis}d\phi \\
&\overset{MAR}{=} \int\int p(I|y_{obs},\phi)p(y|\theta)p(\phi|\theta)p(\theta)dy_{mis}d\phi \\
&\overset{\phi\perp\!\!\!\perp\theta}{=} \int\int p(I|y_{obs},\phi)p(y|\theta)dy_{mis}p(\theta)p(\phi)d\phi \\
&= p(\theta)\int p(y|\theta)dy_{mis}\int p(I|y_{obs},\phi)p(\phi)d\phi \\
&\propto p(\theta)p(y_{obs}|\theta) \\
&\overset{BT}{\propto} p(\theta|y_{obs})
\end{aligned}
$$

- as always, closed-form results are often difficult
- there are two main approaches (assume ignorability):
  - approximate $p(\theta|y_{obs})$ e.g. by the EM-algorithm
  - simulate $p(\theta, y_{mis}|y_{obs})$
- obtaining simulated values of y_{mis} is called *multiple imputation*
- these imputed values can be used later to apply methods that cannot handle missing data

EM for Missing Data

- missing data y_{mis} can be thought of as latent variables
- the EM-algorithm for an ignorable case is easy

E-step: $q(y_{mis}) = p(y_{mis}, \theta|y_{obs})$

M-step: maximize

$$\int \log p(y_{mis}, \theta|y_{obs})q(y_{mis})dy_{mis}$$

Data Augmentation

- assume that missingness is ignorable and $p(\theta|y_{obs})$ is difficult to approximate
- $p(y_{mis}|y_{obs}, \theta)$ is generally simpler
- also the posterior $p(\theta|y_{mis}, y_{obs}) = p(\theta|y)$ is often simpler than $p(\theta|y_{obs})$
- two unknowns θ, y_{mis}, and two "easy" full conditional posteriors

- *data augmentation*:
  1. set $i = 1$, choose initial approximation $p_1(\theta|y_{obs})$
  2. simulate $\theta_1, \ldots, \theta_m$ from $p_i(\theta|y_{obs})$ and then simulate y^j_{mis}, $j = 1, \ldots, m$ from
  $$p(y_{mis}|\theta^j, y_{obs})$$
  3. approximate $p_{i+1}(\theta|y_{obs})$ as a mixture
  $$p_{i+1}(\theta|y_{obs}) \propto \sum_{j=1}^{m} p(\theta|y^j_{mis}, y_{obs})$$
  4. set $i := i + 1$ and go to step 2

- DA uses the same full conditional posteriors as the Gibbs Sampler
- if you simulate just one y_{mis} ($m = 1$), then DA is Gibbs Sampler
- the DA algorithm can be shown to converge in a certain sense
- as in Gibbs Sampler, the important part is to be able to simulate the full conditional posteriors

- $p(y_{mis}|y_{obs}, \theta)$ is often easy to simulate
- but $p(\theta|y)$ is as difficult as without missing data
- heuristically, one may simulate the distributions using any simulation methods, for example MCMC
- one should be careful to try to obtain an independent set of simulated values

- assume you have a DA/Gibbs Sampler for nonmissing data
- in the ignorable case,just add y_{mis} to the model
- the steps to simulate θ are unchanged (ϕ and I cause complications in non-ignorable cases)
- reason: the unknown y_{mis} is assumed to be known in these steps

- y_{mis} is simulated from $p(y_{mis}|y_{obs}, \theta)$
- this is often easy due to known θ
- a possible complication is that y_{mis} is not always independent of y_{obs}, given θ
- example: $p(y|\theta) = N(y|\mu, \Sigma)$, $y \in \mathbb{R}^2$
- if $y = [y_{obs}, y_{mis}]'$ (observe only the first component), then $p(y_{mis}|y_{obs}, \theta)$ is not $p(y_{mis}|\theta)$

- in the exercises, the Gibbs sampler is used to simulate a Normal model with missing data
- consider Normal data in $\mathbb{R}^2$ where some of the second components are missing
- compare list deletion, mean imputation, and Gibbs sampling in estimating a Normal model
- demo: `gibbsmissing2.R`