T-61.5040 Oppivat mallit ja menetelmät
T-61.5040 Learning Models and Methods
Pajunen, Viitaniemi

Solutions to exercise 11, 13.4.2007

Problem 1.

i) We use the same notation as in the lecture notes: the observed measurement values are denoted by by y_{obs}, the missing values by y_{mis}, and $y = (y_{\text{obs}}, y_{\text{mis}})$. We also have a set of indicators $I = \{I_i\}$ where $I_i = 0$ if the measurement y_i is missing, and 1 otherwise. Parameters ϕ dictate the missingness.

We are interested in the posterior distribution of the weight θ of our object. This can be written as $p(\theta|y_{\text{obs}}, I) = p(y_{\text{obs}}, I|\theta)p(\theta)/p(y_{\text{obs}}, I)$. For a flat prior $p(\theta)$ this is proportional to the likelihood

$$
\begin{aligned}
p(y_{\text{obs}}, I|\theta) &= \int p(y_{\text{obs}}, y_{\text{mis}}, I|\theta) dy_{\text{mis}} && (1)\\
&= \int p(y, I|\theta) dy_{\text{mis}} && (2)\\
&= \int p(I|y, \theta)p(y|\theta) dy_{\text{mis}} && (3)\\
&= p(y_{\text{obs}}|\theta) \int p(I|y, \theta)p(y_{\text{mis}}|\theta) dy_{\text{mis}} && (4)
\end{aligned}
$$

We want to apply to this the assumption $p(I|y, \phi) = p(I|y_{\text{obs}}, \phi)$. In order to do this we introduce an additional assumption $p(\phi|\theta) = p(\phi)$ which guarantees $p(\phi|y) = p(\phi)$ (since $p(y|\phi, \theta) = p(y|\theta)$). Combined with the model assumption $p(y|\theta, \phi) = p(y|\theta)$ we also have $p(\phi|y, \theta) = p(\phi)$. Then

$$p(I|y, \theta) = \int p(I, \phi|y, \theta) d\phi = \int p(I|\phi, y)p(\phi|y, \theta) d\phi = \int p(I|\phi, y_{\text{obs}})p(\phi) d\phi = p(I|y_{\text{obs}}).$$

Here we used $p(I|y, \phi, \theta) = p(I|y, \phi)$.

Plugging the result into (4) gives

$$p(y_{\text{obs}}, I|\theta) = p(y_{\text{obs}}|\theta)p(I|y_{\text{obs}}) \underbrace{\int p(y_{\text{mis}}|\theta) dy_{\text{mis}}}_{=1}.$$

The only part that depends on θ is $p(y_{\text{obs}}|\theta)$ and thus

$$p(\theta|y_{\text{obs}}, I) \propto p(y_{\text{obs}}|\theta) \propto p(\theta|y_{\text{obs}})$$

which leads to

$$p(\theta|y_{\text{obs}}, I) = p(\theta|y_{\text{obs}}) = N(\theta|\bar{y}, \frac{1}{91}),$$

where $\bar{y}$ denotes the mean of the observed values.

As we saw, in this MAR system (missing at random, revealed by the given assumption $p(I|y, \phi) = p(I|y_{\text{obs}}, \phi)$) the additional assumption $p(\theta, \phi) = p(\theta)p(\phi)$ leads to ignorability, i.e. $p(\theta|y_{\text{obs}}, I) = p(\theta|y_{\text{obs}})$. In the lectures this was seen to be the case in general. Had we not assumed $p(\theta, \phi) = p(\theta)p(\phi)$, we wouldn't have been able to obtain a simple expression for the likelihood.

ii) If the value $y_i > 100$, then the corresponding indicator $I_i = 0$. To simplify calculations, we make the missingness mechanism deterministic by also assuming that when $y_i \leq 100$, $I_i = 1$. This implies that the missingness distribution $p(I|y, \phi)$ is not MAR, OAR (observed at random) nor MCAR (missing completely at random). The deterministic dependence of I on the data y makes it straightforward to calculate the likelihood

$$
\begin{aligned}
p(y_{\text{obs}}, I|\theta) &= \int p(y|\theta)p(I|y) dy_{\text{mis}} && (5)\\
&= \int p(y_{\text{obs}}|\theta)p(y_{\text{mis}}|\theta) \prod_{i_{\text{all}}} p(I_i|y_i) dy_{\text{mis}} && (6)\\
&= p(y_{\text{obs}}|\theta) \prod_{i_{\text{mis}}} \int p(y_i|\theta)p(I_i|y_i) dy_i && (7)\\
&= p(y_{\text{obs}}|\theta) \prod_{i_{\text{mis}}} \int_{100}^{\infty} p(y_i|\theta) dy_i && (8)
\end{aligned}
$$

since all the likelihoods $p(I_i|y_i)$ are either one or zero. In (6) all the likelihoods corresponding to observed values i equal one as these values were observed in the first place, eliminating the terms from the product.

The missing value likelihoods are zero in some part of the integrated range and one elsewhere. In (7) we are evaluating the likelihood of the actual values of I if the missing data values were known. In the I that we have actually observed, all the components corresponding to the missing data equal zero, so we need to consider only $p(I_i = 0|y_i)$. If the missing data y_i was below 100, the likelihood of I_i being zero is zero. But if it is more than 100, the likelihood is unity. Thus the likelihood term cuts off from the integration range areas where any y_i is below 100. In the left over area the likelihood is one and disappears from the product.

In the product of integrals, each term is the probability mass contained in a Normal distribution $N(y|\theta, 1)$ in the interval $[100, \infty)$. The result is different from that in the part i) due to the missing data not being MAR.

Again, the posterior is proportional to the likelihood:

$$p(\theta|y_{\text{obs}}, I) \propto p(y_{\text{obs}}, I|\theta) = \int p(y_{\text{obs}}, I|\theta, \phi)p(\phi|\theta) d\phi = p(y_{\text{obs}}|\theta) \prod_{i_{\text{mis}}} \int_{100}^{\infty} p(y_i|\theta) dy_i.$$

iii) The data can be thought to be generated i.i.d. as follows:
- take a weight y from $N(y|\theta, 1)$
- see if it is over 100 kg
- if it is, discard it
- otherwise report it
- repeat until we get 91 weights.

This time we don't have any information on how many weights were not reported. Therefore we have less information than in the part ii). We can think of each observed data point as having been generated from a Normal distribution which is truncated at $y = 100$. This must be normalized, since its integral is $\int_{-\infty}^{100} N(y|\theta, 1)dy \neq 1$. So we get

$$p(y_{\text{obs}}|\theta) = \prod_{i_{\text{obs}}} p(y_i|\theta) = \prod_{i_{\text{obs}}} \left[N(y_i|\theta, 1) / \int_{-\infty}^{100} N(y|\theta, 1)dy \right].$$

This likelihood grows without bound as θ grows. That is, if we use a flat prior we should deduce from our measurements that the object we're weighing is likely to be infinitely heavy.

Problem 2.

i) MCAR, since $p(I|y, \phi) = p(I|\phi)$. MCAR is ignorable if ϕ is independent of θ.

ii) We assume that the gender is observed in all the cases even in the alcohol usage remains unknown, i.e. no gender data is missing. We denote $y_i = (a_i, x_i)$ The mechanism is MAR, since $p(I|y, \phi) = p(I|y_{obs}, \phi)$: the missingness depends on the gender, which is always observed. It is not true that $p(I|y, \phi) = p(I|\phi)$ and thereby the mechanism is not MCAR.

The mechanism is ignorable if $p(\phi|\theta) = \int p(\phi, x|\theta)dx$ equals $p(\phi)$. Since ϕ is determined by the gender, it holds that $p(\phi|x, \theta) = p(\phi|x)$. Using this on the integrand results in

$$\int p(\phi|x, \theta)p(x|\theta)dx = \int p(\phi|x)p(x|\theta)dx$$

If x and θ are independent, this reduces to

$$\int p(\phi|x)p(x)dx = \int p(\phi, x)dx = p(\phi).$$

Thus the mechanism is ignorable if x and θ are independent.

iii) MCAR, since $p(I|y, \phi) = p(I|\phi)$. This time ignorability does not hold since $\phi = \theta$.

iv) Since $p(I|y, \phi)$ cannot be simplified, the mechanism is non-ignorable (also not MAR, MCAR, or OAR). The data y_i defines the probability for missingness and therefore no y_i's can be removed from $p(I|y, \phi)$.

Problem 3.

We need conditional marginal posteriors of the unknown quantities:

$$p(\mu|y, \Sigma, I, \phi), \quad p(\Sigma|\mu, y, I, \phi), \quad p(y_{mis}|\mu, \Sigma, I, y_{obs}, \phi)$$

The first distribution is

$$\begin{aligned} p(\mu|y, \Sigma, I, \phi) &\propto p(y, I|\mu, \Sigma, \phi)p(\mu|\Sigma, \phi) = p(y|\mu, \Sigma)p(I|y, \phi)p(\mu|\Sigma) \\ &\propto p(y|\mu, \Sigma)p(\mu|\Sigma) = p(\mu, y|\Sigma) = p(\mu|y, \Sigma)p(y|\Sigma) \\ &\propto p(\mu|y, \Sigma) \end{aligned}$$

Here we assumed the independence of $\theta = (\mu, \Sigma)$ and ϕ. Analogous derivation gives the second distribution as

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

Therefore the two first distributions are just like in the non-missing situation, so they are known by the assumption given in the problem.

It remains to simulate missing data, given parameters μ, Σ. The distribution

$$p(y_{mis}|\mu, \Sigma, I, y_{obs}, \phi) = \frac{p(y, I|\mu, \Sigma, \phi)}{p(I, y_{obs}|\mu, \Sigma, \phi)} = \frac{p(y|\mu, \Sigma, I)p(I|y, \phi)}{p(I, y_{obs}|\mu, \Sigma, \phi)}$$

By the given assumption, $p(I|y, \phi) = p(I|y_{obs}, \phi)$. (This does not generally follow from the ignorability.) Then the above distribution has only one term containing y_{mis} and it is $p(y|\mu, \Sigma, I)$. Therefore

$$p(y_{mis}|\mu, \Sigma, I, y_{obs}, \phi) \propto p(y|\mu, \Sigma, I) = p(y_{mis}|\mu, \Sigma, I, y_{obs})p(y_{obs}|\mu, \Sigma, I) \propto p(y_{mis}|\mu, \Sigma, I, y_{obs}).$$

The Gibbs sampling thus requires us to simulate the last distribution, which is quite straightforward.

Since the parameters are known, each $y(i)$ is conditionally independent from each other. For each i we need to simulate the missing components $y_{mis}(i)$ of the normally distributed vector y_i given the observed components $y_{obs}(i)$ and the parameters of the distribution. Let's figure out how to do this.

From now on we consider one observation i at a time. For brevity, we write $y_{mis}(i) = u$ and $y_{obs}(i) = v$ and drop the conditioning by μ, Σ, I from the notation. We also write

$$\text{Var}(u) = \Sigma_u, \ \text{Var}(v) = \Sigma_v, \ \text{Cov}(u, v) = \Sigma_{uv}, \ \text{Cov}(v, u) = \Sigma_{vu}$$

These parts of Σ are easy to find from the instantaneous Σ, given the missingness indicators I.

Then $p(u, v) = N((u, v)|\mu, \Sigma)$ and we want to find $p(u|v)$. We know that $p(u|v)$ is Normal so compute its mean and covariance. From the iteration formulas given as hint we get

$$\text{E}(u|v) = \text{E}(u) + \text{Cov}(v, u)\text{Var}(v)^{-1}(v - \text{E}(v)) = \mu_u + \Sigma_{vu}\Sigma_v^{-1}(v - \mu_v)$$

$$\text{Var}(u|v) = \text{Var}(u) - \text{Cov}(v, u)\text{Var}(v)^{-1}\text{Cov}(u, v) = \Sigma_u - \Sigma_{vu}\Sigma_v^{-1}\Sigma_{uv}$$

Now for each $y(i)$, the simulation of the missing components amounts to drawing a vector from the Normal distribution $N(\text{E}(u|v), \text{Var}(u|v))$.