Missing Data and the EM Algorithm

Charlie Nash

The University of Edinburgh

March 2, 2016

Where my data at?

x_1	x_2	x_3	x_4
0.10	0.17	-	0.98
0.02	0.32	-	1.32
-	-	-0.70	0.52
0.05	0.20	-0.38	0.97
-	-	-0.37	1.01

Missing data indicator variables

Let $\mathbf{X}$ be a data matrix with missing data. Define the missing data indicator matrix $\mathbf{M}$ as:

- $M_{ij} = 0$ if dimension j of data case i is missing
- $M_{ij} = 1$ if the value is present

Now define the visible $\mathbf{X}_v$ and missing data $\mathbf{X}_m$ as:

- $\mathbf{X}_v = \{x_{ij} : M_{ij} = 1\}$ be the visible data,
- and let $\mathbf{X}_m = \{x_{ij} : M_{ij} = 0\}$ be the hidden data.

Missing data mechanisms

For a particular dataset we may have an understanding of how missing data arises. More specifically we may have a model:

$$p(\mathbf{M}|\mathbf{X}, \theta) = p(\mathbf{M}|\mathbf{X}_v, \mathbf{X}_m, \theta).$$

Depending on the conditional independence structure of this model the data can be:

- Missing completely at random (MCAR)
- Missing at random (MAR)
- Not missing at random (NMAR)

MCAR

The data is missing completely at random (MCAR) if the missingness does not depend on any of the variables in the dataset, whether observed or unobserved. That is:

$$p(\mathbf{M}|\mathbf{X}_v, \mathbf{X}_m, \theta) = p(\mathbf{M}|\theta).$$

Example: Weather measurements across time are randomly corrupted due to faulty equipment.

MAR

The data is missing at random (MAR) if the missingness does not depend on the values of the *missing* variables in the dataset. That is:

$$p(\mathbf{M}|\mathbf{X}_v, \mathbf{X}_m, \theta) = p(\mathbf{M}|\mathbf{X}_v, \theta).$$

Example: A dataset records people's height and weight. Those with high BMI also have their blood pressure recorded.

NMAR

The data is not missing at random (NMAR) if the missingness does depend on the values of the missing data.

$$p(\mathbf{M}|\mathbf{X}_v, \mathbf{X}_m, \boldsymbol{\theta}) = p(\mathbf{M}|\mathbf{X}_v, \mathbf{X}_m, \boldsymbol{\theta}).$$

Example: A temperature sensor can only record between 0-100 degrees. If the temperature is above that then it is recorded as missing.

Fitting generative models with missing data

Now assume that we have a probability model $p(\mathbf{X}, \mathbf{M}|\theta)$, then we want to compute:

$$\hat{\theta} = \text{argmax}_\theta \; p(\mathbf{X}_v, \mathbf{M}|\theta),$$

where

$$\begin{aligned} p(\mathbf{X}_v, \mathbf{M}|\theta) &= \prod_i p(\mathbf{x}_{iv}, \mathbf{m}_i|\theta) \\ &= \prod_i \int_{\mathbf{x}_{im}} p(\mathbf{x}_{iv}, \mathbf{x}_{im}, \mathbf{m}_i|\theta) \; d\mathbf{x}_{im} \\ &= \prod_i \int_{\mathbf{x}_{im}} p(\mathbf{m}_i|\mathbf{x}_{iv}, \mathbf{x}_{im}, \theta) p(\mathbf{x}_{iv}, \mathbf{x}_{im}|\theta) \; d\mathbf{x}_{im} \end{aligned}$$

Fitting generative models with missing data

Now if we assume that the data is missing at random, that is:

$$p(\mathbf{m}_i|\mathbf{x}_{iv}, \mathbf{x}_{im}, \theta) = p(\mathbf{m}_i|\mathbf{x}_{iv}, \theta),$$
$$p(\mathbf{m}_i|\mathbf{x}_{iv}, \mathbf{x}_{im}, \theta)p(\mathbf{x}_{iv}, \mathbf{x}_{im}|\theta) = p(\mathbf{m}_i|\mathbf{x}_{iv}, \mathbf{x}_{im}, \theta_1)p(\mathbf{x}_{im}, \mathbf{x}_{im}|\theta_2),$$

for all $i = 1, \ldots, N$. Then we have:

$$\begin{aligned} p(\mathbf{X}_v, \mathbf{M}|\theta) &= \prod_i \int_{\mathbf{x}_{im}} p(\mathbf{m}_i|\mathbf{x}_{iv}, \mathbf{x}_{im}, \theta)p(\mathbf{x}_{iv}, \mathbf{x}_{im}|\theta)\ d\mathbf{x}_{im} \\ &= \prod_i p(\mathbf{m}_i|\mathbf{x}_{iv}, \theta_1) \int_{\mathbf{x}_{im}} p(\mathbf{x}_{iv}, \mathbf{x}_{im}|\theta_2)\ d\mathbf{x}_{im}, \end{aligned}$$

Fitting generative models with missing data

Therefore we have

$$\begin{aligned}\hat{\theta} &= \text{argmax}_\theta \; p(\mathbf{X}_v, \mathbf{M}|\theta) \\ &= \text{argmax}_\theta \; \log \; p(\mathbf{X}_v, \mathbf{M}|\theta) \\ &= \text{argmax}_\theta \; \sum_i \log \; p(\mathbf{m}_i|\mathbf{x}_{iv}, \theta_1) + \log \int_{\mathbf{x}_{im}} p(\mathbf{x}_{iv}, \mathbf{x}_{im}|\theta_2) \; d\mathbf{x}_{im},\end{aligned}$$

If we are just interested in θ_2 then the relationship between the missingness of the data and the data values is ignorable and we can simply find the MLE of the parameters for the visible data.

Optimizing the likelihood

The trouble is that the likelihood of the parameters for the visible data is difficult to optimise:

$$\log\, p(\mathbf{X}_v|\theta) = \sum_i \log \int_{\mathbf{x}_{im}} p(\mathbf{x}_{iv}, \mathbf{x}_{im}|\theta)\; d\mathbf{x}_{im}.$$

For example, for a multivariate Gaussian with missing data we have:

$$\begin{aligned} p(\mathbf{x}_v, \mathbf{x}_m|\theta) &= \mathcal{N}\left(\begin{bmatrix} \mathbf{x}_v \\ \mathbf{x}_m \end{bmatrix} \middle| \begin{bmatrix} \boldsymbol{\mu}_v \\ \boldsymbol{\mu}_m \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_v & \boldsymbol{\Sigma}_{vm} \\ \boldsymbol{\Sigma}_{mv} & \boldsymbol{\Sigma}_m \end{bmatrix} \right) \\ \Rightarrow p(\mathbf{x}_v|\theta) &= \int_{\mathbf{x}_m} p(\mathbf{x}_v, \mathbf{x}_m|\theta)\; d\mathbf{x}_{im} \\ &= \mathcal{N}(\mathbf{x}_v|\boldsymbol{\mu}_v, \boldsymbol{\Sigma}_v), \end{aligned}$$

Optimizing the likelihood

For every datapoint the contribution to the log likelihood is the log of the marginal Gaussian for the visible dimensions of that example. The log likelihood ignoring constants is

$$\log\, p(\mathbf{X}_v|\boldsymbol{\theta}) = \sum_i -\frac{1}{2}\log|[\boldsymbol{\Sigma}]_{iv}| - \frac{1}{2}(\mathbf{x}_{iv} - [\boldsymbol{\mu}]_{iv})^\intercal [\boldsymbol{\Sigma}]_{iv}^{-1}(\mathbf{x}_{iv} - [\boldsymbol{\mu}]_{iv}),$$

so that the gradient with respect to the j'th dimension of $\boldsymbol{\mu}$ is:

$$\nabla_{\mu_j}\log\, p(\mathbf{X}_v|\boldsymbol{\theta}) = \sum_i \left[[\boldsymbol{\Sigma}]_{iv}^{-1}\mathbf{x}_{iv} - [\boldsymbol{\Sigma}]_{iv}^{-1}[\boldsymbol{\mu}]_{iv}\right]_j .$$

Therefore an analytic solution for the optimal $\boldsymbol{\mu}$ is not possible. One way to solve for $\boldsymbol{\mu}$ is to compute gradients and use a gradient-based optimiser. Alternatively the EM-algorithm can be used to iteratively find a local minimum.

The EM algorithm (refresher)

The EM algorithm alternates the following two steps until convergence:

- **E-step**: Find $p(\mathbf{X}_m|\mathbf{X}_v, \boldsymbol{\theta}_{\text{old}})$ using current parameters $\boldsymbol{\theta}_{\text{old}}$,
- **M-step**: Maximize the expected complete data log-likelihood $Q(\boldsymbol{\theta}, \boldsymbol{\theta}_{\text{old}})$ with respect to $\boldsymbol{\theta}$,

where

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}_{\text{old}}) = \mathbb{E}\left[\sum_i \log\, p(\mathbf{x}_{iv}, \mathbf{x}_{im}, |\boldsymbol{\theta})\right]_{p(\mathbf{X}_m|\mathbf{X}_v, \boldsymbol{\theta}^{\text{old}})}$$

The EM algorithm (justification)

Consider the log-likelihood for one example (using $\mathbf{x}$ for $\mathbf{x}_v$ and $\mathbf{z}$ for $\mathbf{x}_m$):

$$
\begin{aligned}
\log\, p(\mathbf{x}|\boldsymbol{\theta}) &= \int_{\mathbf{z}} q(\mathbf{z}) \log\, p(\mathbf{x}|\boldsymbol{\theta})\; d\mathbf{z} \\
&= \int_{\mathbf{z}} q(\mathbf{z}) \log\, \frac{p(\mathbf{x}|\boldsymbol{\theta}) p(\mathbf{z}|\mathbf{x}, \boldsymbol{\theta}) q(\mathbf{z})}{p(\mathbf{z}|\mathbf{x}, \boldsymbol{\theta}) q(\mathbf{z})}\; d\mathbf{z} \\
&= \mathbb{E}\left[\log\, \frac{1}{q(\mathbf{z})}\right]_{q(\mathbf{z})} + \mathbb{E}\left[\log\, p(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta})\right]_{q(\mathbf{z})} \\
&+ D_{KL}(q(\mathbf{z})||p(\mathbf{z}|\mathbf{x}, \boldsymbol{\theta})) \\
&\geq \mathbb{E}\left[\log\, \frac{1}{q(\mathbf{z})}\right]_{q(\mathbf{z})} + \mathbb{E}\left[\log\, p(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta})\right]_{q(\mathbf{z})},
\end{aligned}
$$

as $D_{KL}(p(\mathbf{x})||q(\mathbf{x})) \geq 0$ for any p, q.

The EM algorithm (justification)

Write the lower bound as

$$L^*(q, \boldsymbol{\theta}) = \mathbb{E}\left[\log \frac{1}{q(\mathbf{z})}\right]_{q(\mathbf{z})} + \mathbb{E}\left[\log p(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta})\right]_{q(\mathbf{z})}$$

The EM algorithm alternates between:

- **E-step**: Optimizing L^* with respect to q, holding $\boldsymbol{\theta}$ fixed (Can show that optimal q is $q(\mathbf{z}) = p(\mathbf{z}|\mathbf{x}, \boldsymbol{\theta})$).
- **M-step**: Optimizing L^* with respect to $\boldsymbol{\theta}$, holding q fixed. (First term constant w.r.t $\boldsymbol{\theta}$ so maximize second term the expected complete data log likelihood.)

Case study: Multivariate normal distribution with missing data

For the E-step of the EM-algorithm we compute the expected complete data log likelihood, where the expectation is under the distribution of missing variables given visible variables and parameters:

$$
\begin{aligned}
Q(\theta, \theta^{\text{old}}) &= \mathbb{E}\left[\sum_i \log \mathcal{N}(\mathbf{x}_i|\boldsymbol{\mu}, \boldsymbol{\Sigma})\right]_{p(\mathbf{X}_m|\mathbf{X}_v, \theta^{\text{old}})} \\
&= -\frac{N}{2}\log |2\pi\boldsymbol{\Sigma}| - \frac{1}{2}\sum_i \mathbb{E}[(\mathbf{x}_i - \boldsymbol{\mu})^\intercal \boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \boldsymbol{\mu})] \\
&= -\frac{N}{2}\log |2\pi\boldsymbol{\Sigma}| - \frac{1}{2}\sum_i \mathbb{E}[\text{tr}(\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^\intercal] \\
&= -\frac{N}{2}\log |2\pi\boldsymbol{\Sigma}| - \frac{1}{2}\text{tr}\left(\boldsymbol{\Sigma}^{-1}\sum_i \mathbb{E}[\mathbf{x}_i\mathbf{x}_i^\intercal] - 2\boldsymbol{\mu}\mathbb{E}[\mathbf{x}_i]^\intercal + \boldsymbol{\mu}\boldsymbol{\mu}^\intercal\right)
\end{aligned}
$$

Case study: Multivariate normal distribution with missing data

Computing the gradient with respect to $\boldsymbol{\mu}$ we have

$$\nabla_{\boldsymbol{\mu}} Q(\theta, \theta^{\text{old}}) = \sum_i \left[\boldsymbol{\Sigma}^{-1}\mathbb{E}[\mathbf{x}_i] - \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right]$$

$$\Rightarrow \hat{\boldsymbol{\mu}} = \frac{1}{N}\sum_i \mathbb{E}[\mathbf{x}_i],$$

so the optimal $\boldsymbol{\mu}$ is the sample mean with the missing values imputed with their expectations. Similarly for the covariance we have

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{N}\sum_i \mathbb{E}[\mathbf{x}_i\mathbf{x}_i^{\mathsf{T}}] - \hat{\boldsymbol{\mu}}\hat{\boldsymbol{\mu}}^{\mathsf{T}}$$

Therefore we need to compute $\mathbb{E}[\mathbf{x}_i]$ and $\mathbb{E}[\mathbf{x}_i\mathbf{x}_i^{\mathsf{T}}]$ under $p(\mathbf{X}_m|\mathbf{X}_v, \theta^{\text{old}})$ and then optimise Q with respect to $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$.

Case study: Multivariate normal distribution with missing data

For datapoint i we have a decomposition into visible and hidden variables $\mathbf{x}_i = [\mathbf{x}_{im}, \mathbf{x}_{iv}]$, and as these variables are jointly Gaussian we have the conditional distribution:

$$p(\mathbf{x}_{im}|\mathbf{x}_{iv}, \theta^{\text{old}}) = \mathcal{N}(\mathbf{x}_{im}|\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*), \tag{1}$$

where $\boldsymbol{\mu}^*$ and $\boldsymbol{\Sigma}^*$ can be obtained from the conditional formula for Gaussians. Then we have

$$\mathbb{E}[\mathbf{x}_i] = [\boldsymbol{\mu}^*, \mathbf{x}_{iv}], \tag{2}$$

$$\mathbb{E}[\mathbf{x}_i \mathbf{x}_i^\mathsf{T}] = \begin{pmatrix} \mathbb{E}[\mathbf{x}_{im}\mathbf{x}_{im}^\mathsf{T}] & \boldsymbol{\mu}^* \mathbf{x}_{iv}^\mathsf{T} \\ \mathbf{x}_{iv}\boldsymbol{\mu}^{*\mathsf{T}} & \mathbf{x}_{iv}\mathbf{x}_{iv}^\mathsf{T} \end{pmatrix}, \tag{3}$$

where $\mathbb{E}[\mathbf{x}_{im}\mathbf{x}_{im}^\mathsf{T}] = \mathbb{E}[\mathbf{x}_{im}]\mathbb{E}[\mathbf{x}_{im}]^\mathsf{T} + \boldsymbol{\Sigma}^*$

References

Kevin Murphy (2012)
Machine Learning: A Probabilistic Perspective
Chapter 11, Mixture models and the EM algorithm.

David Barber (2012)
Bayesian Reasoning and Machine Learning
Chapter 11, Learning with hidden variables.

Schafer and Graham (2002)
Missing Data: Our view of the state of the art
Psychological methods