Missing data lecture 3: More ML and Ignorability

Maximum likelihood for multivariate normal distribution

Let $y_i \in \mathbb{R}^K$, $y_i \overset{\text{iid}}{\sim} \text{Normal}(\mu, \Sigma)$ for n samples so that the density for y_i is

$$f_Y(y_i \mid \mu, \Sigma) = (2\pi)^{-\frac{K}{2}} (\det \Sigma)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(y_i - \mu)^T \Sigma^{-1} (y_i - \mu)\right)$$

The log-likelihood is:

$$\ell_Y(\mu, \Sigma \mid y_i) = \frac{1}{2} \log \det \Sigma - \frac{1}{2}(y_i - \mu)^T \Sigma^{-1} (y_i - \mu)$$

Differentials and matrix differentiation

The book states the MLEs for the multivariate normal distribution without details. Going through the algebra can be useful for other more complicated problems. But in order to do so, we'll need a slight change to how we're used to thinking about partial differentiation.

It all starts with the familiar derivative:

$$f'(c) = \lim_{u \to 0} \frac{f(c+u) - f(c)}{u}$$

We can rearrange this to get a linear approximation to f at the point c:

$$f(c+u) = f(c) + f'(c)u + r_c(u)$$

where $r_c(u) = o(u)$ or $\lim_{u \to 0} \frac{r_c(u)}{u} = 0$.

We can rearrange this expression to get:

$f(c+u) - f(u) = f'(c)u + r_c(u)$

We can now define something call the differential of f:

$$\mathrm{d}f(c;u) = uf'(c)$$

Which is called the first differential of f at c with increment u.

This quantity is just the change in the linear approximation from c to $u+c$:

$$f(c+u) = f(c) + \mathrm{d}f(c;u) + r_c(u)$$

We can identify the differential by finding the linear approximation to a function at c:

$$f(c+u) = f(c) + \alpha u + r_c(u).$$

If we can find an α that depends on c but not on u such that $r_c(u) = o(u)$ we can say that $\alpha = f'(c)$.

Let f now be a function $\mathbb{R}^m \to \mathbb{R}$ and let the differential be constructed via the same argument as above, but now let $c, u \in \mathbb{R}^m$, and define $r_c(u)$ such that $\lim_{u\to 0} \frac{r_c(u)}{\|u\|} = 0$:

$$f(c+u) = f(c) + A(c)u + r_c(u)$$

If we equate $A(c)$ with the partial derivative of f with respect to u, we can recognize this as the Taylor expansion of $f(c+u)$ around $f(c)$.

We can generalize to vector functions: Let $f(x) : \mathbb{R}^m \to \mathbb{R}^n$:

$$f(c+u) = f(c) + A(c)u + r_c(u).$$

for $\lim_{u\to 0} r_c(u)/\mathrm{Y}u\mathrm{Y} = 0$. Then $\mathrm{d}f(c;) = A(c)u$ is the differential of f evaluated at c of increment u.

Again, we can think of this as the linear approximation to f at c.

We can apply the same ideas to matrices, when combined with the vec function, which concatenates an $n \times p$ matrix column by column into an $n \times p$-length vector. Let F be a matrix function $\mathbb{R}^{n\times q} \to \mathbb{R}^{m\times p}$. Let C and U be in $\mathbb{R}^{m\times q}$. If $A(C) \in \mathbb{R}^{mp\times nq}$ such that:

$$\mathrm{vec}(F(C+U)) = \mathrm{vec}(F(C)) + A(C)\mathrm{vec}(U) + \mathrm{vec}(R_c(U)).$$

Then the $m \times p$ matrix $\mathrm{d}F(C;U)$ is defined by $\mathrm{vec}(\mathrm{d}F(C;U)) = A(C)\mathrm{vec}(U)$.

The reason to do this is because the differential generalizes to matrices a bit easier than do partial derivatives. This is because it isn't clear along which dimensions the partial derivatives should lie: Should the partials of a matrix function become a third dimension, like a 3-d array?

Under this framework, the rows of the matrices $A(c)$ and $A(C)$ correspond to a dimension of the range of the function $f(c)$ or $F(C)$, while the columns correspond to a dimension of the domain.

The power of the differentials is clear from the chain rule, which is called Cauchy's rule of invariance in differential-land. This just means that if $b = f(c)$ and $h = g(b)$ the differential of h is:

$$\begin{aligned} \mathrm{d}(h;\, u) &= \mathrm{d}(h;\, \mathrm{d}(f;\, c)) \\ &= A_g(b) A_f(c) u \end{aligned}$$

if $f : \mathbb{R}^m \to \mathbb{R}^p$ and $g : \mathbb{R}^p \to \mathbb{R}^n$ so $h : \mathbb{R}^m \to \mathbb{R}^n$, then $A_g(b) \in R^{n \times p}$ and $A_f(c) \in R^{p \times m}$ and $u \in \mathbb{R}^m$.

When we have a natural partition of variables u into u_1 and u_2 we can write the differential for $f(u)$ more easily in terms of two differentials:

$$\begin{aligned} \mathrm{d}f(c;\, u) &= A(c) u \\ &= A(c_1) u_1 + A(c_2) u_2 \end{aligned}$$

which just differentiates between the two sets of variables. This is important for thinking about differentials of log-likelihoods like the multivariate normal where we'll have two sets of parameters that we'd like to find the partial derivatives with respect to, Σ and μ:

$$\ell_Y(\mu, \Sigma \mid y) = \frac{1}{2} \log \det \Sigma - \frac{1}{2} (y_i - \mu)^T \Sigma^{-1} (y_i - \mu)$$

$$\mathrm{d}\ell_Y(\mu, \Sigma \mid y) = \frac{1}{2} \mathrm{d}(\log \det \Sigma) - \frac{1}{2} \mathrm{d}((y_i - \mu)^T \Sigma^{-1} (y_i - \mu))$$

Differential with respect to μ

First we'll ignore the differential with respect to Σ. We'll expand out that quadratic form into the parts that depend only on μ:

$$\mathrm{d}\ell_Y(\mu, \Sigma \mid y_i) = y_i^T \Sigma^{-1} \mathrm{d}\mu - \frac{1}{2} \mathrm{d}(\mu^T \Sigma^{-1} \mu)$$

Taking the gradient with respect to μ we get:

$$\begin{aligned} \mathrm{d}\ell_Y(\mu, \Sigma \mid y_i) &= y_i^T \Sigma^{-1} \mathrm{d}\mu - \frac{1}{2} \mathrm{d}(\mu^T) \Sigma^{-1} \mu - \frac{1}{2} \mu^T \mathrm{d}(\Sigma^{-1} \mu) \\ &= y_i^T \Sigma^{-1} \mathrm{d}\mu - \frac{1}{2} \mathrm{d}(\mu)^T \Sigma^{-1} \mu - \frac{1}{2} \mu^T \Sigma^{-1} \mathrm{d}\mu \\ &= y_i^T \Sigma^{-1} \mathrm{d}\mu - \frac{1}{2} \mu^T \Sigma^{-1} \mathrm{d}\mu - \frac{1}{2} \mu^T \Sigma^{-1} \mathrm{d}\mu \\ &= y_i^T \Sigma^{-1} \mathrm{d}\mu - \mu^T \Sigma^{-1} \mathrm{d}\mu \\ &= (y_i - \mu)^T \Sigma^{-1} \mathrm{d}\mu \end{aligned}$$

If we sum over the n terms of the log-likelihood we get:

$$\frac{\partial \ell_Y(\mu, \Sigma \mid y_i)}{\partial \mu} = (\sum_i y_i - n\mu)^T \Sigma^{-1}$$

leading to the MLE for μ:

$$\hat{\mu} = \frac{1}{n}\sum_i y_i$$

It'll be useful to write the log-likelihood a bit differently to find the MLE for \$. Remember that $\det A^{-1} = (\det A)^{-1}$. This will enable us to write everything in terms of Σ^{-1} instead of Σ:

$$\ell_Y(\mu, \Sigma \mid y_i) = \frac{1}{2}\log(\det \Sigma^{-1}) - \frac{1}{2}(y_i - \mu)^T\Sigma^{-1}(y_i - \mu)$$

Also remember that $\text{tr}(A) = \sum_i A_{ii}$, $\text{tr}(A + B) = \text{tr}(A) + \text{tr}(B)$, and that $f(x) = \text{tr}(f(x))$ for a univariate function $f(x)$. Finally, recall that $\text{tr}(ABC) = \text{tr}(CAB) = \text{tr}(BCA)$. This will let us rewrite the

Putting all this together allows us to write the log-likelihood for the multivariate normal as such:

$$\ell_Y(\mu, \Sigma \mid y_i) = \frac{1}{2}\log(\det \Sigma^{-1}) - \frac{1}{2}\text{tr}\left((y_i - \mu)(y_i - \mu)^T\Sigma^{-1}\right)$$

For the partial derivative of $\det \Sigma^{-1}$ with respect to Σ^{-1}, we get

$$\frac{\partial \det \Sigma^{-1}}{\partial \Sigma^{-1}} = \det \Sigma^{-1}((\Sigma^{-1})^{-1})^T$$

and for the partial derivative of $\text{tr}(AB)$ with respect to B we get A^T, so the partial derivative with respect to Σ^{-1} of the log-likelihood gives us:

$$\frac{\partial \ell_Y(\mu, \Sigma \mid y_i)}{\partial \Sigma^{-1}} = \frac{1}{2}\Sigma - \frac{1}{2}(y_i - \mu)(y_i - \mu)^T$$

Summing over the n terms gives:

$$\frac{\partial \ell_Y(\mu, \Sigma \mid Y)}{\partial \Sigma^{-1}} = \frac{n}{2}\Sigma - \frac{1}{2}\sum_i (y_i - \mu)(y_i - \mu)^T$$

$$\hat{\Sigma} = \frac{1}{n}\sum_i (y_i - \hat{\mu})(y_i - \hat{\mu})^T$$

Normal repeated measures modesl

In many longitudinal studies where some outcome of interest is measured for participants K times, the following model may describe the data generating process well, where $y_i \in \mathbb{R}^K$ and X_i is a $K \times m$ design matrix:

$$y_i \mid X_i \sim \text{Normal}(X_i\beta, \Sigma(\psi))$$

The textbook lists several models that could describe different scenarios. 1. Independent-but-not-identically-distributed observations within groups:

$$\begin{aligned} y_{ik} \mid (X_i)_k &= (X_i)_k\beta + \epsilon_{ik} \\ \epsilon_{ik} &\sim \text{Normal}(0, \sigma_k^2) \\ \epsilon_{ik} &\perp\!\!\!\perp \epsilon_{jl} \forall i \neq j \cup k \neq l \end{aligned}$$

This implies the following simple structure for $\Sigma(\psi)$ above:

$\Sigma(\psi) = \text{diag}(\sigma_1^2, ..., \sigma_K^2)$.

2. Compound symmetry (I'll call this a random intercept model):

$$\begin{aligned} y_{ik} \mid (X_i)_k &= (X_i)_k\beta + \gamma_i + \epsilon_{ik} \\ \epsilon_{ik} &\sim \text{Normal}(0, \sigma^2) \\ \epsilon_{ik} &\perp\!\!\!\perp \epsilon_{jl} \forall i \neq j \cup k \neq l \\ \gamma_i &\sim \text{Normal}(0, \tau^2) \\ \gamma_i &\perp\!\!\!\perp \gamma_j \forall i \neq j \\ \gamma_i &\perp\!\!\!\perp \epsilon_{ij} \forall j \end{aligned}$$

Conditional on X_i, the covariance between y_{ik} and y_{ij} is:

$$\begin{aligned} \text{Cov}(y_{ik}, y_{ij} \mid X_i) &= \text{Cov}((X_i)_k\beta + \gamma_i + \epsilon_{ik}, (X_i)_j\beta + \gamma_i + \epsilon_{ij} \mid X_i) \\ &= \tau^2 \end{aligned}$$

While the variance is $\tau^2 + \sigma^2$. This implies that we can write the variance-covariance matrix as

$$\Sigma(\psi) = \tau^2 1_K 1_K^T + \sigma^2 I_K.$$

3. Autoregressive (I'll call this a random intercept model):

$$\begin{aligned} y_{ik} \mid (X_i)_k &= (X_i)_k\beta + \epsilon_{ik} \\ \epsilon_{ik} \mid \epsilon_{i,k-1} &\sim \text{Normal}(\rho\epsilon_{i,k-1}, \sigma^2) \\ \epsilon_{i1} &\sim \text{Normal}(0, \frac{\sigma^2}{1-\rho^2}) \\ \epsilon_{ik} &\perp\!\!\!\perp \epsilon_{jl} \forall i \neq j \cap \forall k, l \end{aligned}$$

This implies the following simple structure for $\Sigma(\psi)$:

$$\Sigma(\psi)_{ij} = \frac{\sigma^2}{1-\rho^2}\rho^{|i-j|}$$

4. Random effects model

This is a more general version of the random intercept model. Let $z_k \in \mathbb{R}^q$.

$$
\begin{aligned}
y_{ik} \mid (X_i)_k &= (X_i)_k\beta + z_k^T\gamma_i + \epsilon_{ik} \\
\epsilon_{ik} &\sim \text{Normal}(0, \sigma^2) \\
\epsilon_{ik} &\perp\!\!\!\perp \epsilon_{jl} \forall i \neq j \cup k \neq l \\
\gamma_i &\sim \text{Normal}(0, \Omega) \\
\gamma_i &\perp\!\!\!\perp \gamma_j \forall i \neq j \\
\gamma_i &\perp\!\!\!\perp \epsilon_{ij} \forall j
\end{aligned}
$$

We can write this in matrix form as:

$$
y_i \mid X_i = X_i\beta + Z\gamma + \epsilon_i
$$

The conditional covariance is

$$
\begin{aligned}
\text{Cov}(y_i \mid X_i) &= \text{Cov}(X_i\beta + Z\gamma + \epsilon_i \mid X_i) \\
&= Z\Omega Z^T + \sigma^2 I_K
\end{aligned}
$$

MLEs in repeated measure models

The book suggests the following strategy to find the MLEs in the unstructured case, which is 1 above:

Take $\beta^{(0)}$ and $\Sigma^{(0)}$ as initial guesses. Then for $t = 1$ until some termination criterion iterate:

$$
\beta^{(t+1)} = \sum_i (X_i^T(\Sigma^{(t)})^{-1}X_i)^{-1}X_i^T(\Sigma^{(t)})^{-1}y_i
$$

and

$$
\Sigma^{(t+1)} = \frac{1}{n}\sum_i (y_i - X_i\beta^{(t)})(y_i - X_i\beta^{(t)})^T
$$

We can derive these update rules from the log-likelihood, but we'll need to rewrite the model so that it looks a little more familiar.

The model as written in matrix form by unit i is:

$$
\begin{aligned}
y_i \mid X_i &= X_i\beta + \epsilon_i \\
\epsilon_i &\sim \text{Normal}(0, \Sigma) \\
\epsilon_i &\perp\!\!\!\perp \epsilon_j \forall i \neq j
\end{aligned}
$$

Let $y = (y_1^T, y_2^T, \dots, y_n^T)^T$ and let $X = (X_1^T, X_2^T, \dots, X_n^T)^T$, and let $\epsilon = (\epsilon_1^T, \epsilon_2^T, \dots, \epsilon_n^T)^T$. Then the model can be written:

$$
\begin{aligned}
y \mid X &= X\beta + \epsilon \\
\epsilon &\sim \text{Normal}(0, I_n \otimes \Sigma)
\end{aligned}
$$

so $\text{Cov}(\epsilon)$ is block-diagonal:

$$\begin{bmatrix} \Sigma & 0 & \dots & 0 \\ 0 & \Sigma & \dots & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & \dots & \Sigma \end{bmatrix}$$

The log-likelihood is:

$$\ell_Y(\mu, \Sigma \mid y_i) = -\frac{1}{2}\log\det(I_n \otimes \Sigma) - \frac{1}{2}(y - X\beta)^T(I_n \otimes \Sigma)^{-1}(y - X\beta)$$

The determinant of $I_n \otimes \Sigma$ is $\det(\Sigma)^n$ because it's just block-diagonal, and the inverse of $I_n \otimes \Sigma$ is similarly $I_n \otimes \Sigma^{-1}$.

Let's focus on the β terms. Expanding the quadratic form gives:

$$-\frac{1}{2}(y^T(I_n \otimes \Sigma)^{-1}y + y^T(I_n \otimes \Sigma)^{-1}X\beta - \frac{1}{2}\beta^T X^T(I_n \otimes \Sigma)^{-1}X\beta$$

Taking the derivative with respect to β gives:

$$y^T(I_n \otimes \Sigma)^{-1}X\mathrm{d}\beta - \frac{1}{2}\mathrm{d}\beta^T X^T(I_n \otimes \Sigma)^{-1}X\beta - \frac{1}{2}\beta^T X^T(I_n \otimes \Sigma)^{-1}X\mathrm{d}\beta$$

Collecting terms gives:

$$(y^T(I_n \otimes \Sigma^{-1})X - \beta^T X^T(I_n \otimes \Sigma)^{-1}X)\mathrm{d}\beta$$

This looks more daunting than it is, we can use block matrix multiplication to get:

$$(\sum_i y_i^T\Sigma^{-1}X_i - \beta^T\sum_i X_i^T\Sigma^{-1}X)\mathrm{d}\beta$$

If Σ were known, we could solve this equation simply:

$$\hat{\beta} = (\textstyle\sum_i X_i^T\Sigma^{-1}X)^{-1}(\sum_i X_i\Sigma^{-1}y_i)$$

Like we did above, we can rewrite the likelihood in terms of Σ^{-1} to give:

$$\begin{aligned}
\ell_Y(\mu, \Sigma \mid y_i) &= \frac{n}{2}\log\det(\Sigma^{-1}) - \frac{1}{2}\sum_i (y_i - X_i\beta)^T\Sigma^{-1}(y_i - X_i\beta) \\
&= \frac{n}{2}\log\det(\Sigma^{-1}) - \frac{1}{2}\sum_i \mathrm{tr}(y_i - X_i\beta)^T\Sigma^{-1}(y_i - X_i\beta) \\
&= \frac{n}{2}\log\det(\Sigma^{-1}) - \frac{1}{2}\sum_i \mathrm{tr}(y_i - X_i\beta)(y_i - X_i\beta)^T\Sigma^{-1} \\
&= \frac{n}{2}\log\det(\Sigma^{-1}) - \frac{1}{2}\left(\sum_i \mathrm{tr}(y_i - X_i\beta)(y_i - X_i\beta)^T\right)\Sigma^{-1}
\end{aligned}$$

Taking derivatives with respect to Σ^{-1} gives:

$$\begin{aligned} \mathrm{d}\ell_Y(\mu, \Sigma \mid y_i) &= \frac{n}{2}\Sigma\,\mathrm{d}\Sigma^{-1} - \frac{1}{2}\sum_i (y_i - X_i\beta)(y_i - X_i\beta)^T \mathrm{d}\Sigma^{-1} \\ &= \left(\frac{n}{2}\Sigma - \frac{1}{2}\sum_i (y_i - X_i\beta)(y_i - X_i\beta)^T\right)\mathrm{d}\Sigma^{-1} \end{aligned}$$

Both of these derivatives have to be zero at the maximum likeihood estimate (assuming we're not on a boundary of the parameter space), so we'll get two sets of equations:

$$\Sigma^{(t+1)} = \frac{1}{n}\sum_i (y_i - X_i\beta^{(t)})(y_i - X_i\beta)^T$$

$$\beta^{(t+1)} = (\textstyle\sum_i X_i^T(\Sigma^{(t)})^{-1}X_i)^{-1}(\sum_i X_i(\Sigma^{(t)})^{-1}y_i)$$