The Expectation-Maximization Algorithm

version: 2020-02-25 · 00:40:12

Motivation: Inference with Missing Data

- **Regression Model:** $y_i = \alpha x_i + \beta z_i + \sigma \varepsilon_i, \quad \varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$.
- Suppose some of the z_i are **missing**:
  - Let $\delta_i = 0$ if z_i missing and $\delta_i = 1$ if z_i observed.
  - Observed data:

$$\boldsymbol{\mathcal{D}} = \begin{bmatrix} y_1 & x_1 & z_1 & \delta_1 = 1 \\ y_2 & x_2 & \texttt{NA} & \delta_2 = 0 \\ y_3 & x_3 & \texttt{NA} & \delta_3 = 0 \\ \vdots & \vdots & \vdots & \vdots \\ y_n & x_n & z_n & \delta_n = 1 \end{bmatrix}.$$

- **Problem:** How to estimate $\boldsymbol{\theta} = (\alpha, \beta, \sigma)$ from $\boldsymbol{\mathcal{D}}$?

Inference with Missing Data

Solution 1: Use only complete observations $\mathcal{S}_1 = \{i : \delta_i = 1\}$.

Inference with Missing Data

Solution 1: Use only complete observations $\mathcal{S}_1 = \{i : \delta_i = 1\}$.

1. Inefficient (throws out data)

Inference with Missing Data

Solution 1: Use only complete observations $\boldsymbol{\mathcal{S}}_1 = \{i : \delta_i = 1\}$.

1. Inefficient (throws out data)
2. Potentially misleading, as in the following example:
   - $x, z \overset{\text{iid}}{\sim} \mathcal{N}(0, 1)$
   - True parameters $\alpha = \beta = \sigma = 1$.
   - Missing data mechanism: $P(\delta = 0 \,|\, y \le 2) = 5\%$, $P(\delta = 0 \,|\, y > 2) = 90\%$ (overall 15% missing)
   - Parameter estimates for $n = 10^6$:

	$\hat{\alpha}$(se)	$\hat{\beta}$(se)
No missing data ($\delta \equiv 1$)	.997(.002)	1.001(.002)
Using only $\boldsymbol{\mathcal{S}}_1$ (85% of sample)	.901(.001)	.900(.001)

Inference with Missing Data

Solution 2: Maximize likelihood over $\boldsymbol{\theta}$ and $\boldsymbol{z}_0 = \{z_i : \delta_i = 0\}$

$$(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{z}}_0) = \underset{(\boldsymbol{\theta}, \boldsymbol{z}_0)}{\arg\max} \left\{ -\frac{n}{2} \log \sigma^2 - \frac{1}{2} \sum_{i=1}^{n} \frac{(y_i - \alpha x_i - \beta z_i)^2}{\sigma^2} \right\}.$$

Inference with Missing Data

Solution 2: Maximize likelihood over $\boldsymbol{\theta}$ and $\boldsymbol{z}_0 = \{z_i : \delta_i = 0\}$

$$(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{z}}_0) = \underset{(\boldsymbol{\theta}, \boldsymbol{z}_0)}{\arg\max} \left\{ -\frac{n}{2}\log\sigma^2 - \frac{1}{2}\sum_{i=1}^{n} \frac{(y_i - \alpha x_i - \beta z_i)^2}{\sigma^2} \right\}.$$

- **Profile likelihood:**

$$\hat{z}_i(\boldsymbol{\theta}) = \beta^{-1}(y_i - \alpha x_i) \implies \big(y_i - \alpha x_i - \beta \hat{z}_i(\boldsymbol{\theta})\big)^2 = 0$$

$$\implies \hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\arg\max} \left\{ -\frac{n}{2}\log(\sigma^2) - \frac{1}{2}\sum_{i \in \boldsymbol{\mathcal{S}}_1} \frac{(y_i - \alpha x_i - \beta z_i)^2}{\sigma^2} \right\}$$

Inference with Missing Data

Solution 2: Maximize likelihood over $\boldsymbol{\theta}$ and $\boldsymbol{z}_0 = \{z_i : \delta_i = 0\}$

$$(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{z}}_0) = \underset{(\boldsymbol{\theta}, \boldsymbol{z}_0)}{\arg\max} \left\{ -\frac{n}{2} \log \sigma^2 - \frac{1}{2} \sum_{i=1}^{n} \frac{(y_i - \alpha x_i - \beta z_i)^2}{\sigma^2} \right\}.$$

- **Profile likelihood:**

$$\hat{z}_i(\boldsymbol{\theta}) = \beta^{-1}(y_i - \alpha x_i) \implies \big(y_i - \alpha x_i - \beta \hat{z}_i(\boldsymbol{\theta})\big)^2 = 0$$

$$\implies \hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\arg\max} \left\{ -\frac{n}{2} \log(\sigma^2) - \frac{1}{2} \sum_{i \in \boldsymbol{\mathcal{S}}_1} \frac{(y_i - \alpha x_i - \beta z_i)^2}{\sigma^2} \right\}$$

  - $(\hat{\alpha}, \hat{\beta})$ exactly the same as using complete data $\boldsymbol{\mathcal{S}}_1$ only!
  - $\hat{\sigma} = \hat{\sigma}_1 \cdot n_1/n$, where $\hat{\sigma}_1$ is the estimator from $\boldsymbol{\mathcal{S}}_1$, so confidence intervals even narrower!

Inference with Missing Data

Solution 2: Maximize likelihood over $\boldsymbol{\theta}$ and $\boldsymbol{z}_0 = \{z_i : \delta_i = 0\}$

$$(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{z}}_0) = \underset{(\boldsymbol{\theta}, \boldsymbol{z}_0)}{\arg\max} \left\{ -\frac{n}{2} \log \sigma^2 - \frac{1}{2} \sum_{i=1}^n \frac{(y_i - \alpha x_i - \beta z_i)^2}{\sigma^2} \right\}.$$

- **Profile likelihood:**

$$\hat{z}_i(\boldsymbol{\theta}) = \beta^{-1}(y_i - \alpha x_i) \implies \big(y_i - \alpha x_i - \beta \hat{z}_i(\boldsymbol{\theta})\big)^2 = 0$$

$$\implies \hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\arg\max} \left\{ -\frac{n}{2} \log(\sigma^2) - \frac{1}{2} \sum_{i \in \mathcal{S}_1} \frac{(y_i - \alpha x_i - \beta z_i)^2}{\sigma^2} \right\}$$

  - $(\hat{\alpha}, \hat{\beta})$ exactly the same as using complete data $\mathcal{S}_1$ only!
  - $\hat{\sigma} = \hat{\sigma}_1 \cdot n_1/n$, where $\hat{\sigma}_1$ is the estimator from $\mathcal{S}_1$, so confidence intervals even narrower!

- **Problem:** $\boldsymbol{z}_0$ is a random variable, not a parameter.

Inference with Missing Data

Solution 3: Model the missing data.

Inference with Missing Data

Solution 3: Model the missing data.

If nothing is known about the missing data mechanism, consider the following model:

$$
\begin{aligned}
x &\sim p(x \mid \boldsymbol{\eta}) & \boldsymbol{\theta} &= (\alpha, \beta, \sigma) : \text{original parameters} \\
z \mid x &\sim p(z \mid x, \boldsymbol{\varphi}) & \boldsymbol{\varphi} &: \text{nuisance parameters} \\
y \mid z, x &\sim \mathcal{N}(\alpha x + \beta z, \sigma^2) & \boldsymbol{\eta} &: \text{ignorable parameters} \\
\delta \mid y, z, x &\sim \text{Bernoulli}\{r(y, x, \boldsymbol{\eta})\} & \boldsymbol{\Theta} &= (\boldsymbol{\theta}, \boldsymbol{\varphi}, \boldsymbol{\eta}) : \text{all parameters}
\end{aligned}
$$

Inference with Missing Data

Solution 3: Model the missing data.

$$x \sim p(x \mid \eta) \qquad \boldsymbol{\theta} = (\alpha, \beta, \sigma) : \text{original parameters}$$

$$z \mid x \sim p(z \mid x, \varphi) \qquad \varphi : \text{nuisance parameters}$$

$$y \mid z, x \sim \mathcal{N}(\alpha x + \beta z, \sigma^2) \qquad \eta : \text{ignorable parameters}$$

$$\delta \mid y, z, x \sim \text{Bernoulli}\{r(y, x, \eta)\} \qquad \boldsymbol{\Theta} = (\boldsymbol{\theta}, \varphi, \eta) : \text{all parameters}$$

- **Likelihood:**

$$\begin{aligned}
\mathcal{L}(\boldsymbol{\Theta} \mid \mathcal{D}) &= \prod_{i \in \mathcal{S}_1} p(\delta_i = 1, y_i, z_i, x_i \mid \boldsymbol{\Theta}) \times \prod_{i \in \mathcal{S}_0} p(\delta_i = 0, y_i, x_i \mid \boldsymbol{\Theta}) \\
&= \prod_{i \in \mathcal{S}_1} r(y_i, x_i, \eta) \cdot p(y_i \mid z_i, x_i, \boldsymbol{\theta}) \cdot p(z_i \mid x_i, \varphi) \cdot p(x_i \mid \eta) \\
&\quad \times \prod_{i \in \mathcal{S}_0} [1 - r(y_i, x_i, \eta)] \cdot \underbrace{p(y_i \mid x_i, \boldsymbol{\Theta})}_{\int p(y_i \mid x_i, z_i, \boldsymbol{\theta}) \cdot (z_i \mid x_i, \varphi) \, \mathrm{d}z_i} \cdot p(x_i \mid \eta)
\end{aligned}$$

Inference with Missing Data

Solution 3: Model the missing data.

$$\begin{aligned} x &\sim p(x \mid \eta) \\ z \mid x &\sim p(z \mid x, \varphi) \\ y \mid z, x &\sim \mathcal{N}(\alpha x + \beta z, \sigma^2) \\ \delta \mid y, z, x &\sim \text{Bernoulli}\{r(y, x, \eta)\} \end{aligned}$$

$\boldsymbol{\theta} = (\alpha, \beta, \sigma)$: original parameters

φ : nuisance parameters

η : ignorable parameters

$\boldsymbol{\Theta} = (\boldsymbol{\theta}, \varphi, \eta)$: all parameters

- **Likelihood:**

$$\begin{aligned} \mathcal{L}(\boldsymbol{\Theta} \mid \boldsymbol{\mathcal{D}}) &= \prod_{i=1}^{n} r(y_i, x_i, \eta)_i^{\delta} \cdot [1 - r(y_i, x_i, \eta)]^{1-\delta_i} \cdot p(x_i \mid \eta) \\ &\quad \times \prod_{i \in \boldsymbol{\mathcal{S}}_1} p(y_i \mid z_i, x_i, \boldsymbol{\theta}) \cdot p(z_i \mid x_i, \varphi) \times \prod_{i \in \boldsymbol{\mathcal{S}}_0} \int p(y_i \mid x_i, z_i, \boldsymbol{\theta}) \cdot (z_i \mid x_i, \varphi) \, \mathrm{d}z_i \\ &= \mathcal{L}(\eta \mid \boldsymbol{\mathcal{D}}) \cdot \mathcal{L}(\boldsymbol{\theta}, \varphi \mid \boldsymbol{\mathcal{D}}) \quad \Longrightarrow \quad \boxed{\max_{\boldsymbol{\Theta}} \mathcal{L}(\boldsymbol{\Theta} \mid \boldsymbol{\mathcal{D}}) = \max_{\eta} \mathcal{L}(\eta \mid \boldsymbol{\mathcal{D}}) \cdot \max_{\boldsymbol{\theta}, \varphi} \mathcal{L}(\boldsymbol{\theta}, \varphi \mid \boldsymbol{\mathcal{D}}).} \end{aligned}$$

Inference with Missing Data

Solution 3: Model the missing data.

$$
\begin{aligned}
x &\sim p(x \mid \eta) \\
z \mid x &\sim p(z \mid x, \varphi) \\
y \mid z, x &\sim \mathcal{N}(\alpha x + \beta z, \sigma^2) \\
\delta \mid y, z, x &\sim \text{Bernoulli}\{r(y, x, \eta)\}
\end{aligned}
$$

$\boldsymbol{\theta} = (\alpha, \beta, \sigma)$: original parameters

φ : nuisance parameters

η : ignorable parameters

$\boldsymbol{\Theta} = (\boldsymbol{\theta}, \varphi, \eta)$: all parameters

- **Likelihood:**

$$
\begin{aligned}
\mathcal{L}(\boldsymbol{\Theta} \mid \boldsymbol{\mathcal{D}}) &= \prod_{i=1}^{n} r(y_i, x_i, \eta)_i^{\delta} \cdot [1 - r(y_i, x_i, \eta)]^{1-\delta_i} \cdot p(x_i \mid \eta) \\
&\quad \times \prod_{i \in \boldsymbol{\mathcal{S}}_1} p(y_i \mid z_i, x_i, \boldsymbol{\theta}) \cdot p(z_i \mid x_i, \varphi) \times \prod_{i \in \boldsymbol{\mathcal{S}}_0} \int p(y_i \mid x_i, z_i, \boldsymbol{\theta}) \cdot (z_i \mid x_i, \varphi) \, \mathrm{d}z_i \\
&= \mathcal{L}(\eta \mid \boldsymbol{\mathcal{D}}) \cdot \mathcal{L}(\boldsymbol{\theta}, \varphi \mid \boldsymbol{\mathcal{D}}) \quad \Longrightarrow \quad \boxed{\max_{\boldsymbol{\Theta}} \mathcal{L}(\boldsymbol{\Theta} \mid \boldsymbol{\mathcal{D}}) = \max_{\eta} \mathcal{L}(\eta \mid \boldsymbol{\mathcal{D}}) \cdot \max_{\boldsymbol{\theta}, \varphi} \mathcal{L}(\boldsymbol{\theta}, \varphi \mid \boldsymbol{\mathcal{D}}).}
\end{aligned}
$$

$\Longrightarrow$ $\hat{\boldsymbol{\theta}}_{\text{ML}}$ does not depend on $p(x \mid \eta)$ and $p(\delta \mid y, z, x, \eta)$. The missing data mechanism and covariate distribution of x are thus said to be **ignorable**.

Ignorable vs Nuisance Parameters

- **True Data-Generating Process:** $(\boldsymbol{Y}, \boldsymbol{X}) \sim p_0(\boldsymbol{Y}, \boldsymbol{X})$.
- **Conditional Inference Model:**

$$M_C : \boldsymbol{Y} \mid \boldsymbol{X} \sim p(\boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{\theta}).$$

  - Suppose M_C is *correct*, i.e., exists $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ such that $p(\boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{\theta}_0) = p_0(\boldsymbol{Y} \mid \boldsymbol{X})$.
  - Let $\hat{\boldsymbol{\theta}}_C = \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{\theta})$. If M_C is correct, then $\hat{\boldsymbol{\theta}}_C \to \boldsymbol{\theta}_0$ as sample size $N \to \infty$.

Ignorable vs Nuisance Parameters

- **True Data-Generating Process:** $(\boldsymbol{Y}, \boldsymbol{X}) \sim p_0(\boldsymbol{Y}, \boldsymbol{X})$.
- **Conditional Inference Model:** $M_C : \boldsymbol{Y} \mid \boldsymbol{X} \sim p(\boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{\theta})$.
- **Full Inference Model:**

$$M_F : (\boldsymbol{Y}, \boldsymbol{X}) \sim p(\boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{\theta}) \times p(\boldsymbol{X} \mid \boldsymbol{\theta}, \boldsymbol{\eta}).$$

- Let $(\hat{\boldsymbol{\theta}}_F, \hat{\boldsymbol{\eta}}_F) = \arg\max_{(\boldsymbol{\theta}, \boldsymbol{\eta})} p(\boldsymbol{Y} \mid \boldsymbol{X} \mid \boldsymbol{\theta}) \cdot p(\boldsymbol{X} \mid \boldsymbol{\theta}, \boldsymbol{\eta})$.
- If marginal model $M_X : \boldsymbol{X} \sim p(\boldsymbol{X} \mid \boldsymbol{\eta})$ does not depend on $\boldsymbol{\theta}$, then $\hat{\boldsymbol{\theta}}_F = \hat{\boldsymbol{\theta}}_C$, i.e., $p(\boldsymbol{X} \mid \boldsymbol{\eta})$ is ignorable.
- If $M_X : \boldsymbol{X} \sim p(\boldsymbol{X} \mid \boldsymbol{\theta}, \boldsymbol{\eta})$ does depend on $\boldsymbol{\theta}$, then $\hat{\boldsymbol{\theta}}_F \neq \hat{\boldsymbol{\theta}}_C$.
  - If M_X is correct, then $\hat{\boldsymbol{\theta}}_F \to \boldsymbol{\theta}_0$, and $\mathrm{var}(\hat{\boldsymbol{\theta}}_F) < \mathrm{var}(\hat{\boldsymbol{\theta}}_C)$. Since we need to maximize over $\boldsymbol{\eta}$ to get $\hat{\boldsymbol{\theta}}_F$, $\boldsymbol{\eta}$ are called nuisance parameters.
  - If M_X is incorrect, then generally $\hat{\boldsymbol{\theta}}_F \not\to \boldsymbol{\theta}_0$, even if M_C is correct.

Inference with Missing Data (Continued)

- **Missing Data Setup:** $y_i = \alpha x_i + \beta z_i + \sigma \varepsilon_i, \quad \varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0,1).$

  $\delta_i = 1$ if z_i is observed and $\delta_i = 0$ if it is missing.

- **Complete Data Model:** $M: \ (y, x, z) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$

  $\delta \mid y, x, z \sim \text{Bernoulli}\{r(y, x, \boldsymbol{\eta})\}$

  Note that under M, we have $y \mid x, z \sim \mathcal{N}(\alpha x + \beta z, \sigma^2)$, where

$$\begin{bmatrix} \alpha \\ \beta \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Sigma}_{xx} & \boldsymbol{\Sigma}_{xz} \\ \boldsymbol{\Sigma}_{zx} & \boldsymbol{\Sigma}_{zz} \end{bmatrix}^{-1} \begin{bmatrix} \boldsymbol{\Sigma}_{xy} \\ \boldsymbol{\Sigma}_{zy} \end{bmatrix}, \qquad \sigma^2 = \boldsymbol{\Sigma}_{yy} - \begin{bmatrix} \boldsymbol{\Sigma}_{yx} & \boldsymbol{\Sigma}_{yz} \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix}.$$

Inference with Missing Data (Continued)

- **Missing Data Setup:** $y_i = \alpha x_i + \beta z_i + \sigma \varepsilon_i, \quad \varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0,1).$

  $\delta_i = 1$ if z_i is observed and $\delta_i = 0$ if it is missing.

- **Complete Data Model:** $M : (y, x, z) \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$

  (and ignorable missing data), with $M : y \mid x, z \sim \mathcal{N}(\alpha x + \beta z, \sigma^2)$.

- **Observed Data Likelihood:**

$$\ell(\mathbf{\Sigma} \mid \mathcal{D}) = -\frac{1}{2} \sum_{i \in \mathcal{S}_1} \left\{ \begin{bmatrix} y_i & x_i & z_i \end{bmatrix} \mathbf{\Sigma}^{-1} \begin{bmatrix} y_i \\ x_i \\ z_i \end{bmatrix} + \log |\mathbf{\Sigma}| \right\}$$
$$- \frac{1}{2} \sum_{i \in \mathcal{S}_0} \left\{ \begin{bmatrix} y_i & x_i \end{bmatrix} \begin{bmatrix} \mathbf{\Sigma}_{yy} & \mathbf{\Sigma}_{yx} \\ \mathbf{\Sigma}_{xy} & \mathbf{\Sigma}_{xx} \end{bmatrix}^{-1} \begin{bmatrix} y_i \\ x_i \end{bmatrix} + \log \begin{vmatrix} \mathbf{\Sigma}_{yy} & \mathbf{\Sigma}_{yx} \\ \mathbf{\Sigma}_{xy} & \mathbf{\Sigma}_{xx} \end{vmatrix} \right\}$$

Inference with Missing Data (Continued)

- **Missing Data Setup:** $y_i = \alpha x_i + \beta z_i + \sigma \varepsilon_i, \quad \varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0,1).$

  $\delta_i = 1$ if z_i is observed and $\delta_i = 0$ if it is missing.

- **Complete Data Model:** $M : (y, x, z) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$

  (and ignorable missing data), with $M : y \mid x, z \sim \mathcal{N}(\alpha x + \beta z, \sigma^2)$.

- **Observed Data Likelihood:**

$$\ell(\boldsymbol{\Sigma} \mid \boldsymbol{\mathcal{D}}) = -\frac{1}{2} \sum_{i \in \boldsymbol{\mathcal{S}}_1} \left\{ \begin{bmatrix} y_i & x_i & z_i \end{bmatrix} \boldsymbol{\Sigma}^{-1} \begin{bmatrix} y_i \\ x_i \\ z_i \end{bmatrix} + \log |\boldsymbol{\Sigma}| \right\}$$

$$- \frac{1}{2} \sum_{i \in \boldsymbol{\mathcal{S}}_0} \left\{ \begin{bmatrix} y_i & x_i \end{bmatrix} \begin{bmatrix} \boldsymbol{\Sigma}_{yy} & \boldsymbol{\Sigma}_{yx} \\ \boldsymbol{\Sigma}_{xy} & \boldsymbol{\Sigma}_{xx} \end{bmatrix}^{-1} \begin{bmatrix} y_i \\ x_i \end{bmatrix} + \log \begin{vmatrix} \boldsymbol{\Sigma}_{yy} & \boldsymbol{\Sigma}_{yx} \\ \boldsymbol{\Sigma}_{xy} & \boldsymbol{\Sigma}_{xx} \end{vmatrix} \right\}$$

- **Inference:** $\hat{\boldsymbol{\Sigma}} = \arg\max_{\boldsymbol{\Sigma}} \ell(\boldsymbol{\Sigma} \mid \boldsymbol{\mathcal{D}})$

  Difficult to calculate directly, but simple when $\boldsymbol{z}_0 = \{z_i : \delta_i = 0\}$ is observed!
  That is, if $\boldsymbol{Y}_{n \times 3} = (\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{z})$, then $\hat{\boldsymbol{\Sigma}} = \frac{1}{n} \boldsymbol{Y}' \boldsymbol{Y}$.

Inference with Missing Data (Continued)

- **Observed Data Likelihood:**

$$\ell(\boldsymbol{\Sigma} \mid \mathcal{D}) = -\frac{1}{2}\sum_{i \in \mathcal{S}_1}\left\{\begin{bmatrix} y_i & x_i & z_i \end{bmatrix}\boldsymbol{\Sigma}^{-1}\begin{bmatrix} y_i \\ x_i \\ z_i \end{bmatrix} + \log|\boldsymbol{\Sigma}|\right\}$$
$$-\frac{1}{2}\sum_{i \in \mathcal{S}_0}\left\{\begin{bmatrix} y_i & x_i \end{bmatrix}\begin{bmatrix} \boldsymbol{\Sigma}_{yy} & \boldsymbol{\Sigma}_{yx} \\ \boldsymbol{\Sigma}_{xy} & \boldsymbol{\Sigma}_{xx} \end{bmatrix}^{-1}\begin{bmatrix} y_i \\ x_i \end{bmatrix} + \log\begin{vmatrix} \boldsymbol{\Sigma}_{yy} & \boldsymbol{\Sigma}_{yx} \\ \boldsymbol{\Sigma}_{xy} & \boldsymbol{\Sigma}_{xx} \end{vmatrix}\right\}$$

- **Inference:** $\hat{\boldsymbol{\Sigma}} = \arg\max_{\boldsymbol{\Sigma}} \ell(\boldsymbol{\Sigma} \mid \mathcal{D})$

  Difficult to calculate directly, but simple when $\boldsymbol{z}_0 = \{z_i : \delta_i = 0\}$ is observed. That is, if $\boldsymbol{Y}_{n\times 3} = (\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{z})$, then $\hat{\boldsymbol{\Sigma}} = \frac{1}{n}\boldsymbol{Y}'\boldsymbol{Y}$.

- **Strategy:** Iterative algorithm $(\hat{\boldsymbol{\Sigma}}^{(1)}, \hat{\boldsymbol{z}}_0^{(1)}), \ldots, (\hat{\boldsymbol{\Sigma}}^{(m)}, \hat{\boldsymbol{z}}_0^{(m)})$
  - $\hat{\boldsymbol{\Sigma}}^{(m+1)} = \hat{\boldsymbol{\Sigma}}(\boldsymbol{y}, \boldsymbol{x}, \hat{\boldsymbol{z}}_0^{(m)}, \boldsymbol{z}_1)$
  - $\hat{\boldsymbol{z}}_0^{(m+1)} = ???$

Inference with Missing Data (Continued)

- **Observed Data Likelihood:**

$$\ell(\boldsymbol{\Sigma} \mid \mathcal{D}) = -\frac{1}{2} \sum_{i \in \mathcal{S}_1} \left\{ \begin{bmatrix} y_i & x_i & z_i \end{bmatrix} \boldsymbol{\Sigma}^{-1} \begin{bmatrix} y_i \\ x_i \\ z_i \end{bmatrix} + \log |\boldsymbol{\Sigma}| \right\}$$
$$- \frac{1}{2} \sum_{i \in \mathcal{S}_0} \left\{ \begin{bmatrix} y_i & x_i \end{bmatrix} \begin{bmatrix} \boldsymbol{\Sigma}_{yy} & \boldsymbol{\Sigma}_{yx} \\ \boldsymbol{\Sigma}_{xy} & \boldsymbol{\Sigma}_{xx} \end{bmatrix}^{-1} \begin{bmatrix} y_i \\ x_i \end{bmatrix} + \log \begin{vmatrix} \boldsymbol{\Sigma}_{yy} & \boldsymbol{\Sigma}_{yx} \\ \boldsymbol{\Sigma}_{xy} & \boldsymbol{\Sigma}_{xx} \end{vmatrix} \right\}$$

- **Inference:** $\hat{\boldsymbol{\Sigma}} = \arg\max_{\boldsymbol{\Sigma}} \ell(\boldsymbol{\Sigma} \mid \mathcal{D})$

  Difficult to calculate directly, but simple when $\boldsymbol{z}_0 = \{z_i : \delta_i = 0\}$ is observed. That is, if $\boldsymbol{Y}_{n \times 3} = (\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{z})$, then $\hat{\boldsymbol{\Sigma}} = \frac{1}{n} \boldsymbol{Y}' \boldsymbol{Y}$.

- **Strategy:** Iterative algorithm $(\hat{\boldsymbol{\Sigma}}^{(1)}, \hat{\boldsymbol{z}}_0^{(1)}), \ldots, (\hat{\boldsymbol{\Sigma}}^{(m)}, \hat{\boldsymbol{z}}_0^{(m)})$
  - $\hat{\boldsymbol{\Sigma}}^{(m+1)} = \hat{\boldsymbol{\Sigma}}(\boldsymbol{y}, \boldsymbol{x}, \hat{\boldsymbol{z}}_0^{(m)}, \boldsymbol{z}_1)$
  - $\hat{\boldsymbol{z}}_0^{(m+1)} = \arg\max_{\boldsymbol{z}_0} p(\boldsymbol{z}_0 \mid \mathcal{D}, \hat{\boldsymbol{\Sigma}}^{(m+1)})$?

Inference with Missing Data (Continued)

- **Observed Data Likelihood:**

$$\ell(\boldsymbol{\Sigma} \mid \mathcal{D}) = -\frac{1}{2} \sum_{i \in \mathcal{S}_1} \left\{ \begin{bmatrix} y_i & x_i & z_i \end{bmatrix} \boldsymbol{\Sigma}^{-1} \begin{bmatrix} y_i \\ x_i \\ z_i \end{bmatrix} + \log |\boldsymbol{\Sigma}| \right\}$$
$$- \frac{1}{2} \sum_{i \in \mathcal{S}_0} \left\{ \begin{bmatrix} y_i & x_i \end{bmatrix} \begin{bmatrix} \boldsymbol{\Sigma}_{yy} & \boldsymbol{\Sigma}_{yx} \\ \boldsymbol{\Sigma}_{xy} & \boldsymbol{\Sigma}_{xx} \end{bmatrix}^{-1} \begin{bmatrix} y_i \\ x_i \end{bmatrix} + \log \begin{vmatrix} \boldsymbol{\Sigma}_{yy} & \boldsymbol{\Sigma}_{yx} \\ \boldsymbol{\Sigma}_{xy} & \boldsymbol{\Sigma}_{xx} \end{vmatrix} \right\}$$

- **Inference:** $\hat{\boldsymbol{\Sigma}} = \arg\max_{\boldsymbol{\Sigma}} \ell(\boldsymbol{\Sigma} \mid \mathcal{D})$

  Difficult to calculate directly, but simple when $\boldsymbol{z}_0 = \{z_i : \delta_i = 0\}$ is observed. That is, if $\boldsymbol{Y}_{n \times 3} = (\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{z})$, then $\hat{\boldsymbol{\Sigma}} = \frac{1}{n} \boldsymbol{Y}' \boldsymbol{Y}$.

- **Strategy:** Iterative algorithm $(\hat{\boldsymbol{\Sigma}}^{(1)}, \hat{\boldsymbol{z}}_0^{(1)}), \ldots, (\hat{\boldsymbol{\Sigma}}^{(m)}, \hat{\boldsymbol{z}}_0^{(m)})$
  - $\hat{\boldsymbol{\Sigma}}^{(m+1)} = \hat{\boldsymbol{\Sigma}}(\boldsymbol{y}, \boldsymbol{x}, \hat{\boldsymbol{z}}_0^{(m)}, \boldsymbol{z}_1)$
  - $\hat{\boldsymbol{z}}_0^{(m+1)} = \arg\max_{\boldsymbol{z}_0} p(\boldsymbol{z}_0 \mid \mathcal{D}, \hat{\boldsymbol{\Sigma}}^{(m+1)})$?

    No, converges to $\arg\max_{\boldsymbol{\Sigma}, \boldsymbol{z}_0} \ell(\boldsymbol{\Sigma} \mid \mathcal{D}, \boldsymbol{z}_0) \neq \arg\max_{\boldsymbol{\Sigma}} \ell(\boldsymbol{\Sigma} \mid \mathcal{D})$.

Inference with Missing Data (Continued)

- **Observed Data Likelihood:**

$$\ell(\boldsymbol{\Sigma} \mid \mathcal{D}) = -\frac{1}{2} \sum_{i \in \mathcal{S}_1} \left\{ \begin{bmatrix} y_i & x_i & z_i \end{bmatrix} \boldsymbol{\Sigma}^{-1} \begin{bmatrix} y_i \\ x_i \\ z_i \end{bmatrix} + \log |\boldsymbol{\Sigma}| \right\}$$
$$- \frac{1}{2} \sum_{i \in \mathcal{S}_0} \left\{ \begin{bmatrix} y_i & x_i \end{bmatrix} \begin{bmatrix} \boldsymbol{\Sigma}_{yy} & \boldsymbol{\Sigma}_{yx} \\ \boldsymbol{\Sigma}_{xy} & \boldsymbol{\Sigma}_{xx} \end{bmatrix}^{-1} \begin{bmatrix} y_i \\ x_i \end{bmatrix} + \log \begin{vmatrix} \boldsymbol{\Sigma}_{yy} & \boldsymbol{\Sigma}_{yx} \\ \boldsymbol{\Sigma}_{xy} & \boldsymbol{\Sigma}_{xx} \end{vmatrix} \right\}$$

- **Inference:** $\hat{\boldsymbol{\Sigma}} = \arg\max_{\boldsymbol{\Sigma}} \ell(\boldsymbol{\Sigma} \mid \mathcal{D})$

  Difficult to calculate directly, but simple when $\boldsymbol{z}_0 = \{z_i : \delta_i = 0\}$ is observed. That is, if $\boldsymbol{Y}_{n \times 3} = (\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{z})$, then $\hat{\boldsymbol{\Sigma}} = \frac{1}{n} \boldsymbol{Y}' \boldsymbol{Y}$.

- **Strategy:** Iterative algorithm $(\hat{\boldsymbol{\Sigma}}^{(1)}, \hat{\boldsymbol{z}}_0^{(1)}), \ldots, (\hat{\boldsymbol{\Sigma}}^{(m)}, \hat{\boldsymbol{z}}_0^{(m)})$
  - $\hat{\boldsymbol{\Sigma}}^{(m+1)} = \hat{\boldsymbol{\Sigma}}(\boldsymbol{y}, \boldsymbol{x}, \hat{\boldsymbol{z}}_0^{(m)}, \boldsymbol{z}_1)$
  - $\hat{\boldsymbol{z}}_0^{(m+1)} \sim p(\boldsymbol{z}_0 \mid \mathcal{D}, \hat{\boldsymbol{\Sigma}}^{(m+1)})$?

Inference with Missing Data (Continued)

- **Observed Data Likelihood:**

$$\ell(\boldsymbol{\Sigma} \mid \mathcal{D}) = -\frac{1}{2}\sum_{i \in \mathcal{S}_1}\left\{\begin{bmatrix} y_i & x_i & z_i \end{bmatrix}\boldsymbol{\Sigma}^{-1}\begin{bmatrix} y_i \\ x_i \\ z_i \end{bmatrix} + \log|\boldsymbol{\Sigma}|\right\}$$
$$-\frac{1}{2}\sum_{i \in \mathcal{S}_0}\left\{\begin{bmatrix} y_i & x_i \end{bmatrix}\begin{bmatrix} \boldsymbol{\Sigma}_{yy} & \boldsymbol{\Sigma}_{yx} \\ \boldsymbol{\Sigma}_{xy} & \boldsymbol{\Sigma}_{xx} \end{bmatrix}^{-1}\begin{bmatrix} y_i \\ x_i \end{bmatrix} + \log\begin{vmatrix} \boldsymbol{\Sigma}_{yy} & \boldsymbol{\Sigma}_{yx} \\ \boldsymbol{\Sigma}_{xy} & \boldsymbol{\Sigma}_{xx} \end{vmatrix}\right\}$$

- **Inference:** $\hat{\boldsymbol{\Sigma}} = \arg\max_{\boldsymbol{\Sigma}} \ell(\boldsymbol{\Sigma} \mid \mathcal{D})$

  Difficult to calculate directly, but simple when $\boldsymbol{z}_0 = \{z_i : \delta_i = 0\}$ is observed. That is, if $\boldsymbol{Y}_{n\times 3} = (\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{z})$, then $\hat{\boldsymbol{\Sigma}} = \frac{1}{n}\boldsymbol{Y}'\boldsymbol{Y}$.

- **Strategy:** Iterative algorithm $(\hat{\boldsymbol{\Sigma}}^{(1)}, \hat{\boldsymbol{z}}_0^{(1)}), \ldots, (\hat{\boldsymbol{\Sigma}}^{(m)}, \hat{\boldsymbol{z}}_0^{(m)})$
  - $\hat{\boldsymbol{\Sigma}}^{(m+1)} = \hat{\boldsymbol{\Sigma}}(\boldsymbol{y}, \boldsymbol{x}, \hat{\boldsymbol{z}}_0^{(m)}, \boldsymbol{z}_1)$
  - $\hat{\boldsymbol{z}}_0^{(m+1)} \sim p(\boldsymbol{z}_0 \mid \mathcal{D}, \hat{\boldsymbol{\Sigma}}^{(m+1)})$?

    This produces a stationary stochastic process $\hat{\boldsymbol{\Sigma}}^{(1)}, \hat{\boldsymbol{\Sigma}}^{(2)}, \ldots$, for which the expectation $\tilde{\boldsymbol{\Sigma}} = E[\hat{\boldsymbol{\Sigma}}^{(t)}] \to \boldsymbol{\Sigma}_0$ as $n \to \infty$. However, $\tilde{\boldsymbol{\Sigma}}$ is less efficient than the MLE...

Inference with Missing Data (Continued)

- **Observed Data Likelihood:**

$$\ell(\boldsymbol{\Sigma} \mid \mathcal{D}) = -\frac{1}{2} \sum_{i \in \mathcal{S}_1} \left\{ \begin{bmatrix} y_i & x_i & z_i \end{bmatrix} \boldsymbol{\Sigma}^{-1} \begin{bmatrix} y_i \\ x_i \\ z_i \end{bmatrix} + \log |\boldsymbol{\Sigma}| \right\}$$

$$- \frac{1}{2} \sum_{i \in \mathcal{S}_0} \left\{ \begin{bmatrix} y_i & x_i \end{bmatrix} \begin{bmatrix} \boldsymbol{\Sigma}_{yy} & \boldsymbol{\Sigma}_{yx} \\ \boldsymbol{\Sigma}_{xy} & \boldsymbol{\Sigma}_{xx} \end{bmatrix}^{-1} \begin{bmatrix} y_i \\ x_i \end{bmatrix} + \log \begin{vmatrix} \boldsymbol{\Sigma}_{yy} & \boldsymbol{\Sigma}_{yx} \\ \boldsymbol{\Sigma}_{xy} & \boldsymbol{\Sigma}_{xx} \end{vmatrix} \right\}$$

- **Inference:** $\hat{\boldsymbol{\Sigma}} = \arg\max_{\boldsymbol{\Sigma}} \ell(\boldsymbol{\Sigma} \mid \mathcal{D})$

  Difficult to calculate directly, but simple when $\boldsymbol{z}_0 = \{z_i : \delta_i = 0\}$ is observed. That is, if $\boldsymbol{Y}_{n \times 3} = (\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{z})$, then $\hat{\boldsymbol{\Sigma}} = \frac{1}{n} \boldsymbol{Y}' \boldsymbol{Y}$.

- **Strategy:** Iterative algorithm $(\hat{\boldsymbol{\Sigma}}^{(1)}, \hat{\boldsymbol{z}}_0^{(1)}), \ldots, (\hat{\boldsymbol{\Sigma}}^{(m)}, \hat{\boldsymbol{z}}_0^{(m)})$
  - $\hat{\boldsymbol{\Sigma}}^{(m+1)} = \hat{\boldsymbol{\Sigma}}(\boldsymbol{y}, \boldsymbol{x}, \hat{\boldsymbol{z}}_0^{(m)}, \boldsymbol{z}_1)$
  - $\hat{\boldsymbol{z}}_0^{(m+1)} = E[\boldsymbol{z}_0 \mid \mathcal{D}, \hat{\boldsymbol{\Sigma}}^{(m+1)}]$?

Inference with Missing Data (Continued)

- **Observed Data Likelihood:**

$$\ell(\boldsymbol{\Sigma} \mid \mathcal{D}) = -\frac{1}{2} \sum_{i \in \mathcal{S}_1} \left\{ \begin{bmatrix} y_i & x_i & z_i \end{bmatrix} \boldsymbol{\Sigma}^{-1} \begin{bmatrix} y_i \\ x_i \\ z_i \end{bmatrix} + \log |\boldsymbol{\Sigma}| \right\}$$
$$- \frac{1}{2} \sum_{i \in \mathcal{S}_0} \left\{ \begin{bmatrix} y_i & x_i \end{bmatrix} \begin{bmatrix} \boldsymbol{\Sigma}_{yy} & \boldsymbol{\Sigma}_{yx} \\ \boldsymbol{\Sigma}_{xy} & \boldsymbol{\Sigma}_{xx} \end{bmatrix}^{-1} \begin{bmatrix} y_i \\ x_i \end{bmatrix} + \log \begin{vmatrix} \boldsymbol{\Sigma}_{yy} & \boldsymbol{\Sigma}_{yx} \\ \boldsymbol{\Sigma}_{xy} & \boldsymbol{\Sigma}_{xx} \end{vmatrix} \right\}$$

- **Inference:** $\hat{\boldsymbol{\Sigma}} = \arg\max_{\boldsymbol{\Sigma}} \ell(\boldsymbol{\Sigma} \mid \mathcal{D})$

  Difficult to calculate directly, but simple when $\boldsymbol{z}_0 = \{z_i : \delta_i = 0\}$ is observed. That is, if $\boldsymbol{Y}_{n \times 3} = (\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{z})$, then $\hat{\boldsymbol{\Sigma}} = \frac{1}{n} \boldsymbol{Y}' \boldsymbol{Y}$.

- **Strategy:** Iterative algorithm $(\hat{\boldsymbol{\Sigma}}^{(1)}, \hat{\boldsymbol{z}}_0^{(1)}), \ldots, (\hat{\boldsymbol{\Sigma}}^{(m)}, \hat{\boldsymbol{z}}_0^{(m)})$
  - $\hat{\boldsymbol{\Sigma}}^{(m+1)} = \hat{\boldsymbol{\Sigma}}(\boldsymbol{y}, \boldsymbol{x}, \hat{\boldsymbol{z}}_0^{(m)}, \boldsymbol{z}_1)$
  - $\hat{\boldsymbol{z}}_0^{(m+1)} = E[\boldsymbol{z}_0 \mid \mathcal{D}, \hat{\boldsymbol{\Sigma}}^{(m+1)}]$?

  Almost!

The Expectation-Maximization Algorithm (EM)

- **Setup:**
  - $\boldsymbol{y}_{\text{obs}}$: observed data
  - $\boldsymbol{y}_{\text{miss}}$: missing data
  - $\boldsymbol{y}_{\text{comp}} = \boldsymbol{y}_{\text{obs}} \cup \boldsymbol{y}_{\text{miss}}$: complete data
- **Goal:** Find $\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta} \mid \boldsymbol{y}_{\text{obs}})$.
- **Problem:** $\mathcal{L}(\boldsymbol{\theta} \mid \boldsymbol{y}_{\text{comp}})$ is tractable but

$$\mathcal{L}(\boldsymbol{\theta} \mid \boldsymbol{y}_{\text{obs}}) = \int p(\boldsymbol{y}_{\text{obs}}, \boldsymbol{y}_{\text{miss}} \mid \boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{y}_{\text{miss}}$$

  is not.

The Expectation-Maximization Algorithm (EM)

- **Setup:**
  - $\boldsymbol{y}_{\text{obs}}$: observed data
  - $\boldsymbol{y}_{\text{miss}}$: missing data
  - $\boldsymbol{y}_{\text{comp}} = \boldsymbol{y}_{\text{obs}} \cup \boldsymbol{y}_{\text{miss}}$: complete data
- **Goal:** Find $\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta} \mid \boldsymbol{y}_{\text{obs}})$.
- **EM Algorithm:** An *iterative* algorithm $\hat{\boldsymbol{\theta}}^{(1)}, \hat{\boldsymbol{\theta}}^{(2)}, \ldots$ alternating between two steps:
  - **E-Step:** Construct function $Q_t(\boldsymbol{\theta}) = E[\ell(\boldsymbol{\theta} \mid \boldsymbol{y}_{\text{comp}}) \mid \boldsymbol{y}_{\text{obs}}, \hat{\boldsymbol{\theta}}^{(t)}]$

$$= \int \ell(\boldsymbol{\theta} \mid \boldsymbol{y}_{\text{obs}}, \boldsymbol{y}_{\text{miss}}) \cdot p(\boldsymbol{y}_{\text{miss}} \mid \boldsymbol{y}_{\text{obs}}, \hat{\boldsymbol{\theta}}^{(t)})\, \mathrm{d}\boldsymbol{y}_{\text{miss}}$$

The Expectation-Maximization Algorithm (EM)

- **Setup:**
  - $\boldsymbol{y}_{\text{obs}}$: observed data
  - $\boldsymbol{y}_{\text{miss}}$: missing data
  - $\boldsymbol{y}_{\text{comp}} = \boldsymbol{y}_{\text{obs}} \cup \boldsymbol{y}_{\text{miss}}$: complete data
- **Goal:** Find $\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta} \mid \boldsymbol{y}_{\text{obs}})$.
- **EM Algorithm:** An *iterative* algorithm $\hat{\boldsymbol{\theta}}^{(1)}, \hat{\boldsymbol{\theta}}^{(2)}, \ldots$ alternating between two steps:
  - **E-Step:** Construct function $Q_t(\boldsymbol{\theta}) = E[\ell(\boldsymbol{\theta} \mid \boldsymbol{y}_{\text{comp}}) \mid \boldsymbol{y}_{\text{obs}}, \hat{\boldsymbol{\theta}}^{(t)}]$

$$= \int \ell(\boldsymbol{\theta} \mid \boldsymbol{y}_{\text{obs}}, \boldsymbol{y}_{\text{miss}}) \cdot p(\boldsymbol{y}_{\text{miss}} \mid \boldsymbol{y}_{\text{obs}}, \hat{\boldsymbol{\theta}}^{(t)}) \, \mathrm{d}\boldsymbol{y}_{\text{miss}}$$

  - **M-Step:** Maximize to find next value of $\boldsymbol{\theta}$: $\quad \hat{\boldsymbol{\theta}}^{(t+1)} = \arg\max_{\boldsymbol{\theta}} Q_t(\boldsymbol{\theta})$.

The EM Algorithm: Exponential Families

- **Model:**

$$p(\boldsymbol{y}_{\text{comp}} \mid \boldsymbol{\eta}) = \exp\left\{ \boldsymbol{T}'\boldsymbol{\eta} - \Psi(\boldsymbol{\eta}) \right\} h(\boldsymbol{y}_{\text{comp}})$$

- **E-step:**

$$\begin{aligned} Q_t(\boldsymbol{\eta}) &= E[\ell(\boldsymbol{\eta} \mid \boldsymbol{y}_{\text{comp}}) \mid \boldsymbol{y}_{\text{obs}}, \hat{\boldsymbol{\eta}}^{(t)}] \\ &= \bar{\boldsymbol{T}}_t{}'\boldsymbol{\eta} - \Psi(\boldsymbol{\eta}), \qquad \bar{\boldsymbol{T}}_t = E[\boldsymbol{T} \mid \boldsymbol{y}_{\text{obs}}, \hat{\boldsymbol{\eta}}^{(t)}], \end{aligned}$$

  and $\bar{\boldsymbol{T}}_t$ often easy to compute.

- **M-step:** Convex optimization!

The EM Algorithm: Monotonicity

Theorem. If $\hat{\boldsymbol{\theta}}^{(t)}$ and $\hat{\boldsymbol{\theta}}^{(t+1)}$ are successive steps of the EM algorithm, then

$$\ell(\hat{\boldsymbol{\theta}}^{(t)} \mid \boldsymbol{y}_{\mathsf{obs}}) \le \ell(\hat{\boldsymbol{\theta}}^{(t+1)} \mid \boldsymbol{y}_{\mathsf{obs}}).$$

The EM Algorithm: Rate of Convergence

- Convergence of EM to (local) mode $\boldsymbol{\theta}^\star$ is *linear*:

$$|\hat{\boldsymbol{\theta}}^{(t+1)} - \boldsymbol{\theta}^\star| < K \times |\hat{\boldsymbol{\theta}}^{(t)} - \boldsymbol{\theta}^\star|.$$

- Convergence of Newton-Raphson to (local) model is *quadratic*:

$$|\hat{\boldsymbol{\theta}}^{(t+1)} - \boldsymbol{\theta}^\star| < K \times |\hat{\boldsymbol{\theta}}^{(t)} - \boldsymbol{\theta}^\star|^2.$$

- In practice:
  - Whichever is easier to implement will work better.
  - EM can be used to find a better starting point for NR.

Example: Multivariate Normal

- **Model:** $\boldsymbol{y} = (\boldsymbol{x}, \boldsymbol{z}) \sim \mathcal{N}\left\{\mathbf{0}, \begin{bmatrix} \boldsymbol{\Sigma}_{xx} & \boldsymbol{\Sigma}_{xz} \\ \boldsymbol{\Sigma}_{zx} & \boldsymbol{\Sigma}_{zz} \end{bmatrix}\right\}, \qquad \begin{matrix} \boldsymbol{x} = (x_1, \ldots, x_p), \\ \boldsymbol{z} = (z_1, \ldots, z_q). \end{matrix}$
- **Missing Data:** $\boldsymbol{x}_i$ always observed, but $\delta_i = \begin{cases} 1 & \boldsymbol{z}_i \text{ observed} \\ 0 & \boldsymbol{z}_i \text{ missing} \end{cases}$
- **Observed Data:**
  - Let $\boldsymbol{\mathcal{S}}_k = \{i : \delta_i = k\}, \quad \boldsymbol{Z}_k = \{\boldsymbol{z}_i : i \in \boldsymbol{\mathcal{S}}_k\}, \quad k = 0, 1.$
  - $\boldsymbol{y}_{\text{obs}} = \boldsymbol{\mathcal{D}} = (\boldsymbol{X}, \boldsymbol{Z}_1, \boldsymbol{\delta})$.
- **Complete Data:**
  - $\boldsymbol{y}_{\text{comp}} = (\boldsymbol{Y}, \boldsymbol{\delta}), \qquad \boldsymbol{Y}_{n\times(p+q)} = (\boldsymbol{X}, \boldsymbol{Z}), \qquad \begin{matrix} \boldsymbol{X}_{n\times p} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n), \\ \boldsymbol{Z}_{n\times q} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n). \end{matrix}$
  - Previous example $y \sim \mathcal{N}(\alpha x + \beta z, \sigma^2)$ is a special case with $p = 2$ and $q = 1$:

$$\boldsymbol{x} \leftarrow (y, x), \qquad \boldsymbol{z} \leftarrow z.$$

Example: Multivariate Normal

- **Model:** $\boldsymbol{y} = (\boldsymbol{x}, \boldsymbol{z}) \sim \mathcal{N}\left\{\boldsymbol{0}, \begin{bmatrix} \boldsymbol{\Sigma}_{xx} & \boldsymbol{\Sigma}_{xz} \\ \boldsymbol{\Sigma}_{zx} & \boldsymbol{\Sigma}_{zz} \end{bmatrix}\right\}, \qquad \begin{matrix} \boldsymbol{x} = (x_1, \ldots, x_p), \\ \boldsymbol{z} = (z_1, \ldots, z_q). \end{matrix}$
- **Missing Data:** $\boldsymbol{x}_i$ always observed, but $\delta_i = 1$ (0) if $\boldsymbol{z}_i$ is observed (missing)
- **Observed Data:** $\boldsymbol{y}_{\text{obs}} = \mathcal{D} = (\boldsymbol{X}, \boldsymbol{Z}_1, \boldsymbol{\delta}), \quad \boldsymbol{Z}_1 = \{\boldsymbol{z}_i : \delta_i = 1\}$.
- **Complete Data:** $\boldsymbol{y}_{\text{comp}} = (\boldsymbol{Y}, \boldsymbol{\delta}), \quad \boldsymbol{Y} = (\boldsymbol{X}, \boldsymbol{Z})$.
- **Complete Data Likelihood:** Assuming an ignorable missing data mechanism $\delta \mid \boldsymbol{x}, \boldsymbol{z} \sim \text{Bernoulli}\{r(\boldsymbol{x}, \boldsymbol{\eta})\}$,

$$\begin{aligned} \ell(\boldsymbol{\Sigma} \mid \boldsymbol{y}_{\text{comp}}) &= -\tfrac{1}{2}\left\{n \log|\boldsymbol{\Sigma}| + \textstyle\sum_{i=1}^n \boldsymbol{y}_i' \boldsymbol{\Sigma}^{-1} \boldsymbol{y}_i\right\} \\ &= -\tfrac{1}{2}\left\{n \log|\boldsymbol{\Sigma}| + \textstyle\sum_{i=1}^n \operatorname{tr}(\boldsymbol{\Sigma}^{-1} \boldsymbol{y}_i \boldsymbol{y}_i')\right\}. \end{aligned}$$

Multivariate Normal: EM Algorithm

- **Model:** $\boldsymbol{y} = (\boldsymbol{x}, \boldsymbol{z}) \sim \mathcal{N}\left\{\mathbf{0}, \begin{bmatrix} \boldsymbol{\Sigma}_{xx} & \boldsymbol{\Sigma}_{xz} \\ \boldsymbol{\Sigma}_{zx} & \boldsymbol{\Sigma}_{zz} \end{bmatrix}\right\}, \qquad \begin{matrix} \boldsymbol{x} = (x_1, \ldots, x_p), \\ \boldsymbol{z} = (z_1, \ldots, z_q). \end{matrix}$
- **Observed Data:** $\boldsymbol{y}_{\text{obs}} = \mathcal{D} = (\boldsymbol{X}, \boldsymbol{Z}_1, \delta), \quad \boldsymbol{Z}_1 = \{\boldsymbol{z}_i : \delta_i = 1\}.$
- **Complete Data Likelihood:** For $\boldsymbol{y}_{\text{comp}} = (\boldsymbol{Y}, \delta), \quad \boldsymbol{Y} = (\boldsymbol{X}, \boldsymbol{Z}),$

$$\ell(\boldsymbol{\Sigma} \mid \boldsymbol{y}_{\text{comp}}) = -\tfrac{1}{2}\left\{n \log |\boldsymbol{\Sigma}| + \textstyle\sum_{i=1}^n \text{tr}(\boldsymbol{\Sigma}^{-1} \boldsymbol{y}_i \boldsymbol{y}_i')\right\}.$$

- **E-Step:**
  - **Q-Function:**
    $$Q_t(\boldsymbol{\Sigma}) = -\frac{1}{2}\left\{n \log |\boldsymbol{\Sigma}| + \text{tr}(\boldsymbol{\Sigma}^{-1} \boldsymbol{Y}_1' \boldsymbol{Y}_1) + \sum_{i \in \boldsymbol{\mathcal{S}}_0} \text{tr}\left(\boldsymbol{\Sigma}^{-1} E\left[\boldsymbol{y}_i \boldsymbol{y}_i' \mid \boldsymbol{x}_i, \hat{\boldsymbol{\Sigma}}^{(t)}\right]\right)\right\},$$ where
    $\boldsymbol{Y}_1 = \{\boldsymbol{y}_i : i \in \boldsymbol{\mathcal{S}}_1\}.$

Multivariate Normal: EM Algorithm

- **Complete Data Likelihood:** For $\boldsymbol{y}_{\text{comp}} = (\boldsymbol{Y}, \delta)$, $\quad \boldsymbol{Y} = (\boldsymbol{X}, \boldsymbol{Z})$,

$$\ell(\boldsymbol{\Sigma} \mid \boldsymbol{y}_{\text{comp}}) = -\tfrac{1}{2}\Big\{ n \log |\boldsymbol{\Sigma}| + \textstyle\sum_{i=1}^{n} \operatorname{tr}(\boldsymbol{\Sigma}^{-1} \boldsymbol{y}_i \boldsymbol{y}_i') \Big\}.$$

- **E-Step:**
  - **Q-Function:**
    $Q_t(\boldsymbol{\Sigma}) = -\dfrac{1}{2}\Big\{ n \log |\boldsymbol{\Sigma}| + \operatorname{tr}(\boldsymbol{\Sigma}^{-1} \boldsymbol{Y}_1' \boldsymbol{Y}_1) + \sum_{i \in \boldsymbol{\mathcal{S}}_0} \operatorname{tr}\Big(\boldsymbol{\Sigma}^{-1} E\Big[\boldsymbol{y}_i \boldsymbol{y}_i' \mid \boldsymbol{x}_i, \hat{\boldsymbol{\Sigma}}^{(t)} \Big] \Big) \Big\}$, where
    $\boldsymbol{Y}_1 = \{ \boldsymbol{y}_i : i \in \boldsymbol{\mathcal{S}}_1 \}$.
  - **Conditional Expectation:**

$$\boldsymbol{z}_i \mid \boldsymbol{x}_i, \hat{\boldsymbol{\Sigma}}^{(t)} \sim \mathcal{N}(\hat{\boldsymbol{\mu}}_i^{(t)}, \hat{\boldsymbol{\Omega}}^{(t)}), \qquad \begin{aligned} \hat{\boldsymbol{\mu}}_i^{(t)} &= \hat{\boldsymbol{\Sigma}}_{zx}^{(t)} [\hat{\boldsymbol{\Sigma}}_{xx}^{(t)}]^{-1} \boldsymbol{x}_i \\ \hat{\boldsymbol{\Omega}}^{(t)} &= \hat{\boldsymbol{\Sigma}}_{zz}^{(t)} - \hat{\boldsymbol{\Sigma}}_{zx}^{(t)} [\hat{\boldsymbol{\Sigma}}_{xx}^{(t)}]^{-1} \hat{\boldsymbol{\Sigma}}_{xz}^{(t)} \end{aligned}$$

$$\implies E\Big[\boldsymbol{y}_i \boldsymbol{y}_i' \mid \boldsymbol{x}_i, \hat{\boldsymbol{\Sigma}}^{(t)} \Big] = E\left\{ \begin{bmatrix} \boldsymbol{x}_i \boldsymbol{x}_i' & \boldsymbol{x}_i \boldsymbol{z}_i' \\ \boldsymbol{z}_i \boldsymbol{x}_i' & \boldsymbol{z}_i \boldsymbol{z}_i' \end{bmatrix} \mid \boldsymbol{x}_i, \hat{\boldsymbol{\Sigma}}^{(t)} \right\} = \underbrace{\begin{bmatrix} \boldsymbol{x}_i \boldsymbol{x}_i' & \boldsymbol{x}_i [\hat{\boldsymbol{\mu}}_i^{(t)}]' \\ [\hat{\boldsymbol{\mu}}_i^{(t)}] \boldsymbol{x}_i' & \hat{\boldsymbol{\Omega}}^{(t)} + [\hat{\boldsymbol{\mu}}_i^{(t)}][\hat{\boldsymbol{\mu}}_i^{(t)}]' \end{bmatrix}}_{\hat{\boldsymbol{T}}_i^{(t)}}$$

$$\implies Q_t(\boldsymbol{\Sigma}) = -\frac{1}{2}\left\{ n \log |\boldsymbol{\Sigma}| + \operatorname{tr}\Big[\boldsymbol{\Sigma}^{-1} \big(\boldsymbol{Y}_1' \boldsymbol{Y}_1 + \textstyle\sum_{i \in \boldsymbol{\mathcal{S}}_0} \hat{\boldsymbol{T}}_i^{(t)} \big) \Big] \right\}.$$

Multivariate Normal: EM Algorithm

- **Model:** $\boldsymbol{y} = (\boldsymbol{x}, \boldsymbol{z}) \sim \mathcal{N}\left\{\mathbf{0}, \begin{bmatrix} \boldsymbol{\Sigma}_{xx} & \boldsymbol{\Sigma}_{xz} \\ \boldsymbol{\Sigma}_{zx} & \boldsymbol{\Sigma}_{zz} \end{bmatrix}\right\}, \qquad \begin{matrix} \boldsymbol{x} = (x_1, \ldots, x_p), \\ \boldsymbol{z} = (z_1, \ldots, z_q). \end{matrix}$
- **Observed Data:** $\boldsymbol{y}_{\text{obs}} = \boldsymbol{\mathcal{D}} = (\boldsymbol{X}, \boldsymbol{Z}_1, \delta), \quad \boldsymbol{Z}_1 = \{\boldsymbol{z}_i : \delta_i = 1\}.$
- **E-Step:**

$$Q_t(\boldsymbol{\Sigma}) = -\frac{1}{2}\left\{n \log|\boldsymbol{\Sigma}| + \operatorname{tr}\left[\boldsymbol{\Sigma}^{-1}(Y_1' Y_1 + \textstyle\sum_{i \in \boldsymbol{\mathcal{S}}_0} \hat{\boldsymbol{T}}_i^{(t)})\right]\right\}.$$

- **M-Step:**

$$\hat{\boldsymbol{\Sigma}}^{(t+1)} = \frac{1}{n}\left(\boldsymbol{Y}_1' \boldsymbol{Y}_1 + \textstyle\sum_{i \in \boldsymbol{\mathcal{S}}_0} \hat{\boldsymbol{T}}_i^{(t)}\right)$$

(Since $Q_t(\boldsymbol{\Sigma})$ has same shape as loglikelihood of $\boldsymbol{y}_i \overset{\text{iid}}{\sim} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$)

Example: Mixture of Exponential Families

- **Exponential Family:** $\boldsymbol{y} \sim g(\boldsymbol{y} \mid \boldsymbol{\eta}) = \exp\{\, \boldsymbol{T}'\boldsymbol{\eta} - \Psi(\boldsymbol{\eta})\} \cdot h(\boldsymbol{y})$.
- **Mixture Model:** The K-component mixture model is

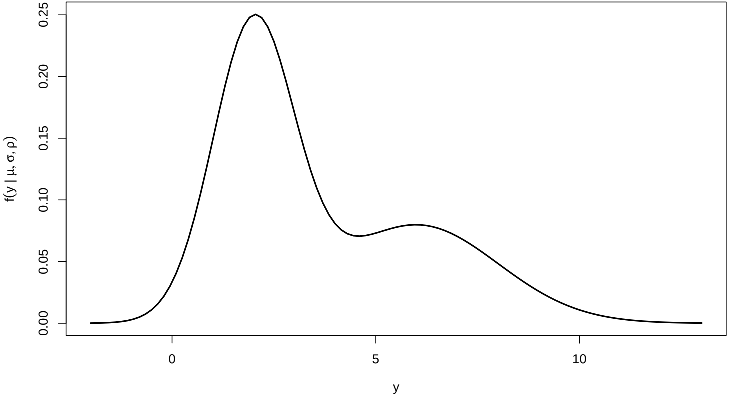
$$f(\boldsymbol{y} \mid \boldsymbol{\theta}) = \sum_{k=1}^{K} \rho_k \cdot g(\boldsymbol{y} \mid \boldsymbol{\eta}_k),$$

  where $\boldsymbol{\Lambda} = (\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_K)$, $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_K)$, and $\rho_k \geq 0$, $\sum_{k=1}^{K} \rho_k = 1$.

Example: Mixture of Exponential Families

- **Model:** $f(\boldsymbol{y} \mid \boldsymbol{\Lambda}, \boldsymbol{\rho}) = \sum_{k=1}^{K} \rho_k \cdot g(\boldsymbol{y} \mid \boldsymbol{\eta}_k), \qquad g(\boldsymbol{y} \mid \boldsymbol{\eta}) = \exp\{\boldsymbol{T}'\boldsymbol{\eta} - \Psi(\boldsymbol{\eta})\} h(\boldsymbol{y}).$
- **Example:** $K = 2, \qquad g(y \mid \boldsymbol{\eta}_k) \cong \mathcal{N}(\mu_k, \sigma_k^2),$

$$\boldsymbol{\mu} = (2, 6), \qquad \boldsymbol{\sigma} = (1, 2), \qquad \boldsymbol{\rho} = (.6, .4)$$



Example: Mixture of Exponential Families

- **Model:** $f(\boldsymbol{y} \mid \boldsymbol{\Lambda}, \boldsymbol{\rho}) = \sum_{k=1}^{K} \rho_k \cdot g(\boldsymbol{y} \mid \boldsymbol{\eta}_k), \qquad g(\boldsymbol{y} \mid \boldsymbol{\eta}) = \exp\{\boldsymbol{T}' \boldsymbol{\eta} - \Psi(\boldsymbol{\eta})\} h(\boldsymbol{y}).$

- **Applications:**

  1. **Density Estimation:** For large enough K, mixture model is arbitrarily accurate approximate to any data-generating process $\boldsymbol{y} \sim f_0(\boldsymbol{y})$ with same support.

  2. **Classification:** To simulate $\boldsymbol{y} \sim f(\boldsymbol{y} \mid \boldsymbol{\Lambda}, \boldsymbol{\rho})$:

$$\boldsymbol{z} = (z_1, \ldots, z_K) \overset{\text{iid}}{\sim} \text{Multinomial}(1, \boldsymbol{\rho})$$

$$\boldsymbol{y} \mid \boldsymbol{z} \overset{\text{ind}}{\sim} g(\boldsymbol{y} \mid \boldsymbol{\eta}_{\boldsymbol{z}}), \qquad \boldsymbol{\eta}_{\boldsymbol{z}} \text{ is } \boldsymbol{\eta}_k \text{ for which } z_k = 1$$

$$\implies \Pr(\boldsymbol{y} \text{ is in group } k \mid \boldsymbol{y}, \boldsymbol{\Lambda}, \boldsymbol{\rho}) = \Pr(z_k = 1 \mid \boldsymbol{y}, \boldsymbol{\Lambda}, \boldsymbol{\rho})$$

by Bayes Formula: $\Pr(A \mid B) = \dfrac{\Pr(B \mid A)\Pr(A)}{\Pr(B)}$

$$= \frac{\Pr(\boldsymbol{y} \mid z_k = 1, \boldsymbol{\Lambda}, \boldsymbol{\rho}) \Pr(z_k = 1, \boldsymbol{\Lambda}, \boldsymbol{\rho})}{f(\boldsymbol{y} \mid \boldsymbol{\Lambda}, \boldsymbol{\rho})}$$

$$= \frac{\rho_k \cdot g(\boldsymbol{y} \mid \boldsymbol{\eta}_k)}{\sum_{j=1}^{K} \rho_j \cdot g(\boldsymbol{y} \mid \boldsymbol{\eta}_j)}$$

Example: Mixture of Exponential Families

- **Model:** $f(\boldsymbol{y} \mid \boldsymbol{\Lambda}, \rho) = \sum_{k=1}^{K} \rho_k \cdot g(\boldsymbol{y} \mid \boldsymbol{\eta}_k), \qquad g(\boldsymbol{y} \mid \boldsymbol{\eta}) = \exp\{\boldsymbol{T}'\boldsymbol{\eta} - \Psi(\boldsymbol{\eta})\}h(\boldsymbol{y}).$
- **Inference:** Estimate $\boldsymbol{\theta} = (\boldsymbol{\Lambda}, \boldsymbol{\rho})$ given $\boldsymbol{Y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)$, $\boldsymbol{y}_i \overset{\text{iid}}{\sim} f(\boldsymbol{y} \mid \boldsymbol{\Lambda}, \boldsymbol{\rho})$.

Example: Mixture of Exponential Families

- **Model:** $f(\boldsymbol{y} \mid \boldsymbol{\Lambda}, \rho) = \sum_{k=1}^{K} \rho_k \cdot g(\boldsymbol{y} \mid \boldsymbol{\eta}_k), \qquad g(\boldsymbol{y} \mid \boldsymbol{\eta}) = \exp\{\boldsymbol{T}'\boldsymbol{\eta} - \Psi(\boldsymbol{\eta})\} h(\boldsymbol{y}).$
- **Inference:** Estimate $\boldsymbol{\theta} = (\boldsymbol{\Lambda}, \boldsymbol{\rho})$ given $\boldsymbol{Y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)$, $\boldsymbol{y}_i \overset{\text{iid}}{\sim} f(\boldsymbol{y} \mid \boldsymbol{\Lambda}, \boldsymbol{\rho})$.
  - **Simulation:**

$$\boldsymbol{z}_i = (z_{i1}, \ldots, z_{iK}) \overset{\text{iid}}{\sim} \text{Multinomial}(1, \boldsymbol{\rho})$$

$$\boldsymbol{y}_i \mid \boldsymbol{z}_i \overset{\text{ind}}{\sim} g(\boldsymbol{y} \mid \boldsymbol{\eta}_{\boldsymbol{z}_i}), \qquad \boldsymbol{\eta}_{\boldsymbol{z}_i} \text{ is } \boldsymbol{\eta}_k \text{ for which } z_{ik} = 1$$

Example: Mixture of Exponential Families

- **Model:** $f(\boldsymbol{y} \mid \boldsymbol{\Lambda}, \rho) = \sum_{k=1}^{K} \rho_k \cdot g(\boldsymbol{y} \mid \boldsymbol{\eta}_k), \qquad g(\boldsymbol{y} \mid \boldsymbol{\eta}) = \exp\{\boldsymbol{T}'\boldsymbol{\eta} - \Psi(\boldsymbol{\eta})\} h(\boldsymbol{y}).$
- **Inference:** Estimate $\boldsymbol{\theta} = (\boldsymbol{\Lambda}, \boldsymbol{\rho})$ given $\boldsymbol{Y} = (\boldsymbol{y}_1, \dots, \boldsymbol{y}_n)$, $\boldsymbol{y}_i \overset{\text{iid}}{\sim} f(\boldsymbol{y} \mid \boldsymbol{\Lambda}, \boldsymbol{\rho})$.
  - **Simulation:**

$$\boldsymbol{z}_i = (z_{i1}, \dots, z_{iK}) \overset{\text{iid}}{\sim} \text{Multinomial}(1, \boldsymbol{\rho})$$

$$\boldsymbol{y}_i \mid \boldsymbol{z}_i \overset{\text{ind}}{\sim} g(\boldsymbol{y} \mid \boldsymbol{\eta}_{\boldsymbol{z}_i}), \qquad \boldsymbol{\eta}_{\boldsymbol{z}_i} \text{ is } \boldsymbol{\eta}_k \text{ for which } z_{ik} = 1$$

  - Suggests that the EM setup would be
    - $\boldsymbol{y}_{\text{comp}} = (\boldsymbol{Y}, \boldsymbol{Z})$
    - $\boldsymbol{y}_{\text{obs}} = \boldsymbol{Y}$
    - $\boldsymbol{y}_{\text{miss}} = \boldsymbol{Z} = (\boldsymbol{z}_1, \dots, \boldsymbol{z}_n)$.

Mixture of EFs: EM Algorithm

- **Model:** $\boldsymbol{y}_i \overset{\text{iid}}{\sim} f(\boldsymbol{y} \mid \boldsymbol{\Lambda}, \boldsymbol{\rho}) = \sum_{k=1}^{K} \rho_k \cdot g(\boldsymbol{y} \mid \boldsymbol{\eta}_k), \qquad g(\boldsymbol{y} \mid \boldsymbol{\eta}) = \exp\{\boldsymbol{T}'\boldsymbol{\eta} - \Psi(\boldsymbol{\eta})\} h(\boldsymbol{y}).$
- **Complete Data:** $\boldsymbol{y}_i \mid \boldsymbol{z}_i \overset{\text{ind}}{\sim} g(\boldsymbol{y} \mid \boldsymbol{\eta}_{\boldsymbol{z}_i}), \qquad \boldsymbol{z}_i \overset{\text{iid}}{\sim} \text{Multinomial}(1, \boldsymbol{\rho}).$
- **Complete Data Log-likelihood:**

$$\ell(\boldsymbol{\theta} \mid \boldsymbol{Y}, \boldsymbol{Z}) = \sum_{i=1}^{n} \underbrace{\left[\boldsymbol{T}_i'\boldsymbol{\eta}_{z_i} - \Psi(\boldsymbol{\eta}_{\boldsymbol{z}_i})\right]}_{\text{Exponential Family}} + \underbrace{\sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \log(\rho_k)}_{\text{Multinomial}}$$

$$= \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \Big[\boldsymbol{T}_i'\boldsymbol{\eta}_k - \Psi(\boldsymbol{\eta}_k) + \log(\rho_k)\Big]$$

$$= \sum_{k=1}^{K} \sum_{i=1}^{n} z_{ik} \Big[\boldsymbol{T}_i'\boldsymbol{\eta}_k - \Psi(\boldsymbol{\eta}_k) + \log(\rho_k)\Big].$$

Mixture of EFs: EM Algorithm

- **Model:** $\boldsymbol{y}_i \overset{\text{iid}}{\sim} f(\boldsymbol{y} \mid \boldsymbol{\Lambda}, \boldsymbol{\rho}) = \sum_{k=1}^{K} \rho_k \cdot g(\boldsymbol{y} \mid \boldsymbol{\eta}_k), \qquad g(\boldsymbol{y} \mid \boldsymbol{\eta}) = \exp\{\boldsymbol{T}'\boldsymbol{\eta} - \Psi(\boldsymbol{\eta})\}h(\boldsymbol{y}).$
- **Complete Data:** $\boldsymbol{y}_i \mid \boldsymbol{z}_i \overset{\text{ind}}{\sim} g(\boldsymbol{y} \mid \boldsymbol{\eta}_{\boldsymbol{z}_i}), \qquad \boldsymbol{z}_i \overset{\text{iid}}{\sim} \text{Multinomial}(1, \boldsymbol{\rho}).$
- **Complete Data Log-likelihood:**

$$\ell(\boldsymbol{\theta} \mid \boldsymbol{Y}, \boldsymbol{Z}) = \sum_{k=1}^{K} \sum_{i=1}^{n} z_{ik} \Big[\boldsymbol{T}_i' \boldsymbol{\eta}_k - \Psi(\boldsymbol{\eta}_k) + \log(\rho_k) \Big].$$

- **E-Step:** $Q_t(\boldsymbol{\theta}) = \sum_{k=1}^{K} \sum_{i=1}^{n} E[z_{ik} \mid \boldsymbol{y}_i, \hat{\boldsymbol{\theta}}^{(t)}] \Big[\boldsymbol{T}_i' \boldsymbol{\eta}_k - \Psi(\boldsymbol{\eta}_k) + \log(\rho_k) \Big].$

  To calculate the expectation, note that $z_{ik} \in \{0, 1\}$, such that

$$E[z_{ik} \mid \boldsymbol{y}_i, \hat{\boldsymbol{\theta}}^{(t)}] = \Pr(z_{ik} = 1 \mid \boldsymbol{y}_i, \hat{\boldsymbol{\theta}}^{(t)}) = \frac{\hat{\rho}_k^{(t)} g(\boldsymbol{y}_i \mid \hat{\boldsymbol{\eta}}_k^{(t)})}{\sum_{j=1}^{K} \hat{\rho}_j^{(t)} g(\boldsymbol{y}_i \mid \hat{\boldsymbol{\eta}}_j^{(t)})} = \hat{q}_{ik}^{(t)}.$$

Mixture of EFs: EM Algorithm

- **Model:** $\boldsymbol{y}_i \overset{\text{iid}}{\sim} f(\boldsymbol{y} \mid \boldsymbol{\Lambda}, \boldsymbol{\rho}) = \sum_{k=1}^{K} \rho_k \cdot g(\boldsymbol{y} \mid \boldsymbol{\eta}_k), \qquad g(\boldsymbol{y} \mid \boldsymbol{\eta}) = \exp\{\boldsymbol{T}'\boldsymbol{\eta} - \Psi(\boldsymbol{\eta})\} h(\boldsymbol{y}).$
- **Complete Data:** $\boldsymbol{y}_i \mid \boldsymbol{z}_i \overset{\text{ind}}{\sim} g(\boldsymbol{y} \mid \boldsymbol{\eta}_{\boldsymbol{z}_i}), \qquad \boldsymbol{z}_i \overset{\text{iid}}{\sim} \text{Multinomial}(1, \boldsymbol{\rho}).$
- **Complete Data Log-likelihood:**

$$\ell(\boldsymbol{\theta} \mid \boldsymbol{Y}, \boldsymbol{Z}) = \sum_{k=1}^{K} \sum_{i=1}^{n} z_{ik} \Big[\boldsymbol{T}_i' \eta_k - \Psi(\boldsymbol{\eta}_k) + \log(\rho_k) \Big].$$

- **E-Step:** $Q_t(\boldsymbol{\theta}) = \sum_{k=1}^{K} \sum_{i=1}^{n} \hat{q}_{ik}^{(t)} \Big[\boldsymbol{T}_i' \boldsymbol{\eta}_k - \Psi(\boldsymbol{\eta}_k) + \log(\rho_k) \Big]$

$$= \sum_{k=1}^{K} \Big[\hat{\boldsymbol{T}}_k^{(t)\prime} \boldsymbol{\eta}_k - q_k^{(t)} \Psi(\boldsymbol{\eta}_k) + q_k^{(t)} \log(\rho_k) \Big],$$

where $\hat{\boldsymbol{T}}_k^{(t)} = \sum_{i=1}^{n} \hat{q}_{ik}^{(t)} \boldsymbol{T}_i$ and $q_k^{(t)} = \sum_{i=1}^{n} \hat{q}_{ik}^{(t)}$.

Mixture of EFs: EM Algorithm

- **Model:** $\boldsymbol{y}_i \overset{\text{iid}}{\sim} f(\boldsymbol{y} \mid \boldsymbol{\Lambda}, \boldsymbol{\rho}) = \sum_{k=1}^{K} \rho_k \cdot g(\boldsymbol{y} \mid \boldsymbol{\eta}_k), \qquad g(\boldsymbol{y} \mid \boldsymbol{\eta}) = \exp\{\boldsymbol{T}'\boldsymbol{\eta} - \Psi(\boldsymbol{\eta})\} h(\boldsymbol{y}).$
- **Complete Data:** $\boldsymbol{y}_i \mid \boldsymbol{z}_i \overset{\text{ind}}{\sim} g(\boldsymbol{y} \mid \boldsymbol{\eta}_{\boldsymbol{z}_i}), \qquad \boldsymbol{z}_i \overset{\text{iid}}{\sim} \text{Multinomial}(1, \boldsymbol{\rho}).$
- **E-Step:** $Q_t(\boldsymbol{\theta}) = \sum_{k=1}^{K} \left[\hat{\boldsymbol{T}}_k^{(t)\prime} \boldsymbol{\eta}_k - q_k^{(t)} \Psi(\boldsymbol{\eta}_k) + q_k^{(t)} \log(\rho_k) \right].$
- **M-Step:**
  - **EF Parameters:** $\boldsymbol{\eta}_k^{(t+1)} = \underset{\boldsymbol{\eta}}{\arg\max} \left[\hat{\boldsymbol{T}}_k^{(t)\prime} \boldsymbol{\eta} - q_k^{(t)} \Psi(\boldsymbol{\eta}) \right]$, i.e., separable convex optimization problems.
  - **Mixing Parameters:** $\hat{\boldsymbol{\rho}}^{(t+1)} = \underset{\boldsymbol{\rho}}{\arg\max} \sum_{k=1}^{K} q_k^{(t)} \log(\rho_k).$

    Actually a $K-1$ dimensional optimization since $\rho_K = 1 - \sum_{k=1}^{K-1} \rho_k$. Similarly, by definition $q_K^{(t)} = 1 - \sum_{k=1}^{K-1} q_k^{(t)}$, such that with $\boldsymbol{q}^{(t)} = (q_1^{(t)}, \ldots, q_K^{(t)})$,

$$\frac{\partial}{\partial \rho_j} \sum_{k=1}^{K} q_k^{(t)} \log(\rho_k) \bigg|_{\boldsymbol{\rho} = \hat{\boldsymbol{\rho}}^{(t+1)}} = \frac{q_j^{(t)}}{\hat{\rho}_j^{(t+1)}} - \frac{1 - \sum_{k=1}^{K-1} q_k^{(t)}}{1 - \sum_{k=1}^{K-1} \hat{\rho}_k^{(t+1)}} = 0 \iff \hat{\boldsymbol{\rho}}^{(t+1)} = \boldsymbol{q}^{(t)}.$$

Example: Probit Regression

- **Logistic Regression:**

$$y_i \mid \boldsymbol{x}_i \overset{\text{ind}}{\sim} \text{Bernoulli}(\rho_i), \qquad \rho = \frac{1}{1 + \exp(-\boldsymbol{x}_i' \boldsymbol{\beta})}.$$

- **Probit Regression:** $\rho_i = \Phi(\boldsymbol{x}_i' \boldsymbol{\beta})$, where Φ is the CDF of $\mathcal{N}(0, 1)$.

  Can think of this as $z_i \overset{\text{ind}}{\sim} \mathcal{N}(\boldsymbol{x}_i' \boldsymbol{\beta}, 1)$, and $y_i = \mathbb{1}\{z_i > 0\}$ since

$$\Pr(y = 1 \mid \boldsymbol{x}) = \Pr(z > 0 \mid \boldsymbol{x}) = \Pr(\underbrace{z - \boldsymbol{x}' \boldsymbol{\beta}}_{\mathcal{N}(0,1)} > -\boldsymbol{x}' \boldsymbol{\beta} \mid \boldsymbol{x}) = \Phi(\boldsymbol{x}' \boldsymbol{\beta}).$$

  $\implies$ the EM setup is

$$\boldsymbol{y}_{\text{obs}} = (\boldsymbol{y}, \boldsymbol{X}), \qquad \boldsymbol{y}_{\text{comp}} = (\boldsymbol{z}, \boldsymbol{y}, \boldsymbol{X}), \qquad \boldsymbol{y}_{\text{miss}} = \boldsymbol{z}.$$

Probit Regression: EM Algorithm

- **Probit Regression:** $y_i \mid \boldsymbol{x}_i \overset{\text{ind}}{\sim} \text{Bernoulli}(\rho_i)$, $\quad \rho_i = \Pr(Z < \boldsymbol{x}_i'\beta)$, $Z \sim \mathcal{N}(0,1)$.
- **Complete Data:** $z_i \mid \boldsymbol{x}_i \overset{\text{ind}}{\sim} \mathcal{N}(\boldsymbol{x}_i'\beta)$, $\quad y_i = \mathbb{1}\{z_i > 0\}$.
- **Complete Data Likelihood:** $\ell(\beta \mid \boldsymbol{z}, \boldsymbol{y}, \boldsymbol{X}) = -\frac{1}{2}\sum_{i=1}^n (z_i - \boldsymbol{x}_i'\beta)^2$
- **E-Step:**

$$
\begin{aligned}
Q_t(\beta) &= E[\ell(\beta \mid \boldsymbol{z}, \boldsymbol{y}, \boldsymbol{X}) \mid \boldsymbol{y}, \boldsymbol{X}, \hat{\beta}^{(t)}] \\
&= -\frac{1}{2}\sum_{i=1}^n E\left[(z_i - \boldsymbol{x}_i'\beta)^2 \mid y_i, \boldsymbol{x}_i, \hat{\beta}^{(t)}\right].
\end{aligned}
$$

Probit Regression: EM Algorithm

- **Probit Regression:** $y_i \mid \boldsymbol{x}_i \overset{\text{ind}}{\sim} \text{Bernoulli}(\rho_i)$, $\rho_i = \Pr(Z < \boldsymbol{x}_i'\beta)$, $Z \sim \mathcal{N}(0,1)$.

- **Complete Data:** $z_i \mid \boldsymbol{x}_i \overset{\text{ind}}{\sim} \mathcal{N}(\boldsymbol{x}_i'\beta)$, $y_i = \mathbb{1}\{z_i > 0\}$.

- **Complete Data Likelihood:** $\ell(\beta \mid \boldsymbol{z}, \boldsymbol{y}, \boldsymbol{X}) = -\frac{1}{2}\sum_{i=1}^n (z_i - \boldsymbol{x}_i'\beta)^2$

- **E-Step:** $Q_t(\beta) = -\frac{1}{2}\sum_{i=1}^n E\left[(z_i - \boldsymbol{x}_i'\beta)^2 \mid y_i, \boldsymbol{x}_i, \hat{\beta}^{(t)}\right]$.

  To calculate the expectation, note that $\pm(z_i - \boldsymbol{x}_i'\hat{\beta}^{(t)}) \sim \mathcal{N}(0,1)$, such that for $y_i = 0$,

$$
\begin{aligned}
E\left[(z_i - \boldsymbol{x}_i'\beta)^2 \mid y_i, \boldsymbol{x}_i, \hat{\beta}^{(t)}\right] &= E\left[(z_i - \boldsymbol{x}_i'\beta)^2 \mid z_i < 0, \boldsymbol{x}_i, \hat{\beta}^{(t)}\right] \\
&= E\left[\{z_i - \boldsymbol{x}_i'\hat{\beta}^{(t)} - \boldsymbol{x}_i'(\beta - \hat{\beta}^{(t)})\}^2 \mid z_i - \boldsymbol{x}_i'\hat{\beta}^{(t)} < -\boldsymbol{x}_i'\hat{\beta}^{(t)}\right] \\
&= E\left[\{Z - \boldsymbol{x}_i'(\beta - \hat{\beta}^{(t)})\}^2 \mid Z < -\boldsymbol{x}_i'\hat{\beta}^{(t)}\right], \qquad Z \sim \mathcal{N}(0,1).
\end{aligned}
$$

Probit Regression: EM Algorithm

- **Probit Regression:** $y_i \mid \boldsymbol{x}_i \overset{\text{ind}}{\sim} \text{Bernoulli}(\rho_i),$ $\quad \rho_i = \Pr(Z < \boldsymbol{x}_i'\beta), \quad Z \sim \mathcal{N}(0,1).$
- **Complete Data:** $z_i \mid \boldsymbol{x}_i \overset{\text{ind}}{\sim} \mathcal{N}(\boldsymbol{x}_i'\beta), \qquad y_i = \mathbb{1}\{z_i > 0\}.$
- **Complete Data Likelihood:** $\ell(\beta \mid \boldsymbol{z}, \boldsymbol{y}, \boldsymbol{X}) = -\frac{1}{2}\sum_{i=1}^{n}(z_i - \boldsymbol{x}_i'\beta)^2$
- **E-Step:** $Q_t(\beta) = -\frac{1}{2}\sum_{i=1}^{n} E\left[(z_i - \boldsymbol{x}_i'\beta)^2 \mid y_i, \boldsymbol{x}_i, \hat{\beta}^{(t)}\right].$

  Similarly for $y_i = 1$,

$$
\begin{aligned}
E\left[(z_i - \boldsymbol{x}_i'\beta)^2 \mid y_i, \boldsymbol{x}_i, \hat{\beta}^{(t)}\right] &= E\left[(z_i - \boldsymbol{x}_i'\beta)^2 \mid z_i > 0, \boldsymbol{x}_i, \hat{\beta}^{(t)}\right] \\
&= E\left[\{Z - \boldsymbol{x}_i'(\beta - \hat{\beta}^{(t)})\}^2 \mid Z > -\boldsymbol{x}_i'\hat{\beta}^{(t)}\right] \\
&= E\left[\{Z - \boldsymbol{x}_i'(\beta - \hat{\beta}^{(t)})\}^2 \mid Z < \boldsymbol{x}_i'\hat{\beta}^{(t)}\right], \qquad Z \sim \mathcal{N}(0,1),
\end{aligned}
$$

  where we can replace Z by $-1 \times Z$ since $\mathcal{N}(0,1)$ is symmetric.

Probit Regression: EM Algorithm

- **Probit Regression:** $y_i \mid \boldsymbol{x}_i \overset{\text{ind}}{\sim} \text{Bernoulli}(\rho_i)$, $\quad \begin{aligned} \rho_i &= \Pr(Z < \boldsymbol{x}_i'\beta), \\ Z &\sim \mathcal{N}(0,1). \end{aligned}$
- **Complete Data:** $z_i \mid \boldsymbol{x}_i \overset{\text{ind}}{\sim} \mathcal{N}(\boldsymbol{x}_i'\beta)$, $\quad y_i = \mathbb{1}\{z_i > 0\}$.
- **Complete Data Likelihood:** $\ell(\beta \mid \boldsymbol{z}, \boldsymbol{y}, \boldsymbol{X}) = -\frac{1}{2}\sum_{i=1}^{n}(z_i - \boldsymbol{x}_i'\beta)^2$
- **E-Step:**

$$\begin{aligned} Q_t(\beta) &= -\frac{1}{2}\sum_{i=1}^{n} E\left[(z_i - \boldsymbol{x}_i'\beta)^2 \mid y_i, \boldsymbol{x}_i, \hat{\beta}^{(t)}\right] \\ &= -\frac{1}{2}\sum_{i=1}^{n} \mathcal{G}\left(\boldsymbol{x}_i'(\beta - \hat{\beta}^{(t)}), (2 \times \mathbb{1}\{y_i = 1\} - 1) \cdot \boldsymbol{x}_i'\hat{\beta}^{(t)}\right) \end{aligned}$$

where $\mathcal{G}(a,b) = E[(Z-a)^2 \mid Z < b]$.

Probit Regression: EM Algorithm

- **Probit Regression:** $y_i \mid \boldsymbol{x}_i \overset{\text{ind}}{\sim} \text{Bernoulli}(\rho_i)$, $\rho_i = \Pr(Z < \boldsymbol{x}_i'\beta)$, $Z \sim \mathcal{N}(0, 1)$.
- **Complete Data:** $z_i \mid \boldsymbol{x}_i \overset{\text{ind}}{\sim} \mathcal{N}(\boldsymbol{x}_i'\beta)$, $y_i = \mathbb{1}\{z_i > 0\}$.
- **Complete Data Likelihood:** $\ell(\beta \mid \boldsymbol{z}, \boldsymbol{y}, \boldsymbol{X}) = -\frac{1}{2}\sum_{i=1}^{n}(z_i - \boldsymbol{x}_i'\beta)^2$
- **E-Step:** $Q_t(\boldsymbol{\beta}) = -\frac{1}{2}\sum_{i=1}^{n}\mathcal{G}\Big(\boldsymbol{x}_i'(\boldsymbol{\beta} - \hat{\beta}^{(t)}), (2 \times \mathbb{1}\{y_i = 1\} - 1) \cdot \boldsymbol{x}_i'\hat{\beta}^{(t)}\Big)$,

  where $\mathcal{G}(a, b) = E[(Z - a)^2 \mid Z < b]$. To calculate $\mathcal{G}(a, b)$

Probit Regression: EM Algorithm

- **Probit Regression:** $y_i \mid \boldsymbol{x}_i \overset{\text{ind}}{\sim} \text{Bernoulli}(\rho_i), \qquad \begin{aligned} \rho_i &= \Pr(Z < \boldsymbol{x}_i'\beta), \\ Z &\sim \mathcal{N}(0,1). \end{aligned}$
- **Complete Data:** $z_i \mid \boldsymbol{x}_i \overset{\text{ind}}{\sim} \mathcal{N}(\boldsymbol{x}_i'\beta), \qquad y_i = \mathbb{1}\{z_i > 0\}.$
- **Complete Data Likelihood:**

$$\ell(\beta \mid \boldsymbol{z}, \boldsymbol{y}, \boldsymbol{X}) = -\frac{1}{2}\sum_{i=1}^n (z_i - \boldsymbol{x}_i'\beta)^2 = -\frac{1}{2}\sum_{i=1}^n z_i^2 - 2(\boldsymbol{x}_i'\beta)\cdot z_i + (\boldsymbol{x}_i'\beta)^2$$

- **E-Step:**

$$\begin{aligned} Q_t(\beta) &= E[\ell(\beta \mid \boldsymbol{z}, \boldsymbol{y}, \boldsymbol{X}) \mid \boldsymbol{y}, \boldsymbol{X}, \hat{\beta}^{(t)}] \\ &= -\frac{1}{2}\sum_{i=1}^n \left\{ E[z_i^2 \mid y_i, \boldsymbol{x}_i, \hat{\beta}^{(t)}] - 2(\boldsymbol{x}_i'\beta)\cdot E[z_i \mid y_i, \boldsymbol{x}_i, \hat{\beta}^{(t)}] + (\boldsymbol{x}_i'\beta)^2 \right\}. \end{aligned}$$

Probit Regression: EM Algorithm

- **Probit Regression:** $y_i \mid \boldsymbol{x}_i \overset{\text{ind}}{\sim} \text{Bernoulli}(\rho_i)$, $\quad \rho_i = \Pr(Z < \boldsymbol{x}_i'\beta)$, $\quad Z \sim \mathcal{N}(0,1)$.

- **Complete Data:** $z_i \mid \boldsymbol{x}_i \overset{\text{ind}}{\sim} \mathcal{N}(\boldsymbol{x}_i'\beta)$, $\qquad y_i = \mathbb{1}\{z_i > 0\}$.

- **E-Step:**

$$Q_t(\beta) = -\frac{1}{2}\sum_{i=1}^{n}\left\{E[z_i^2 \mid y_i, \boldsymbol{x}_i, \hat{\beta}^{(t)}] - 2(\boldsymbol{x}_i'\beta)\cdot E[z_i \mid y_i, \boldsymbol{x}_i, \hat{\beta}^{(t)}] + (\boldsymbol{x}_i'\beta)^2\right\}.$$

To calculate the expectations, note that $\pm(z_i - \boldsymbol{x}_i'\hat{\beta}^{(t)}) \sim \mathcal{N}(0,1)$, such that

$$\begin{aligned} E[z_i \mid y_i, \boldsymbol{x}_i, \hat{\beta}^{(t)}] &= \boldsymbol{x}_i'\hat{\beta}^{(t)} + E[z_i - \boldsymbol{x}_i'\hat{\beta}^{(t)} \mid y_i, \boldsymbol{x}_i, \hat{\beta}^{(t)}] \\ &= \boldsymbol{x}_i'\hat{\beta}^{(t)} + \begin{cases} E[Z \mid Z < \boldsymbol{x}_i'\hat{\beta}^{(t)}] & y_i = 1 \\ E[Z \mid Z < -\boldsymbol{x}_i'\hat{\beta}^{(t)}] & y_i = 0, \end{cases} \end{aligned}$$

where $Z \sim \mathcal{N}(0,1)$, and similarly for $E[z_i^2 \mid y_i, \boldsymbol{x}_i, \hat{\beta}^{(t)}]$.

$\implies$ Need to calculate $g(a) = E[Z|Z < a]$ and $h(a) = E[Z^2 \mid Z < a]$.

Probit Regression: EM Algorithm

- **Probit Regression:** $y_i \mid \boldsymbol{x}_i \overset{\text{ind}}{\sim} \text{Bernoulli}(\rho_i)$, $\quad \rho_i = \Pr(Z < \boldsymbol{x}_i'\beta),\quad Z \sim \mathcal{N}(0,1)$.
- **Complete Data:** $z_i \mid \boldsymbol{x}_i \overset{\text{ind}}{\sim} \mathcal{N}(\boldsymbol{x}_i'\beta)$, $\quad y_i = \mathbb{1}\{z_i > 0\}$.
- **E-Step:** $Q_t(\beta) = -\frac{1}{2}\sum_{i=1}^n \left\{ E[z_i^2 \mid y_i, \boldsymbol{x}_i, \hat{\beta}^{(t)}] - 2(\boldsymbol{x}_i'\beta)\cdot E[z_i \mid y_i, \boldsymbol{x}_i, \hat{\beta}^{(t)}] + (\boldsymbol{x}_i'\beta)^2 \right\}$.
  - Requires $g(a) = E[Z|Z < a]$ and $h(a) = E[Z^2 \mid Z < a]$, where $Z \sim \mathcal{N}(0,1)$.
  - Moment-generating function (MGF) of a truncated normal:

$$M(t) = E[e^{Zt} \mid Z < a] = \frac{\int_{-\infty}^a e^{tz}\cdot e^{-z^2/2}\,\mathrm{d}z}{\int_{-\infty}^a e^{-z^2/2}\,\mathrm{d}z} = \frac{e^{t^2/2}\Phi(a-t)}{\Phi(a)}$$

$$\implies g(a) = \frac{\mathrm{d}M(0)}{\mathrm{d}t} = -1 \times \frac{\phi(a)}{\Phi(a)}, \qquad h(a) = \frac{\mathrm{d}^2 M(0)}{\mathrm{d}t^2} = 1 - a \times \frac{\phi(a)}{\Phi(a)},$$

where $\phi(z)$ and $\Phi(z)$ are the PDF and CDF of $Z \sim \mathcal{N}(0,1)$.

Probit Regression: EM Algorithm

- **Probit Regression:** $y_i \mid \boldsymbol{x}_i \overset{\text{ind}}{\sim} \text{Bernoulli}(\rho_i)$, $\rho_i = \Pr(Z < \boldsymbol{x}_i'\beta)$, $Z \sim \mathcal{N}(0,1)$.
- **Complete Data:** $z_i \mid \boldsymbol{x}_i \overset{\text{ind}}{\sim} \mathcal{N}(\boldsymbol{x}_i'\beta)$, $y_i = \mathbb{1}\{z_i > 0\}$.
- **E-Step:** After some algebra, get

$$Q_t(\beta) = -\frac{1}{2}\sum_{i=1}^{n}(\hat{z}_i^{(t)} - \boldsymbol{x}_i'\beta)^2,$$

- **M-Step:** Equivalent to maximing the likelihood of $\hat{z}_i^{(t)} \overset{\text{ind}}{\sim} \mathcal{N}(\boldsymbol{x}_i'\beta, 1)$

$$\implies \hat{\beta}^{(t+1)} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\hat{\boldsymbol{z}}^{(t)}.$$

Example: Multivariate t-Distribution

- **Definition:** Let $\boldsymbol{z} = (z_1, \ldots, z_d) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}) \quad \amalg \quad x \sim \chi^2_{(\nu)}$. Then

  $$\boldsymbol{y} = \frac{\boldsymbol{z}}{\sqrt{x/\nu}} + \boldsymbol{\mu}$$

  has a multivariate Student-t distribution, denoted $\boldsymbol{y} \sim t_{(\nu)}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

- **EM Setup:** To simulate observations $\boldsymbol{y}_i \overset{\text{iid}}{\sim} t_{(\nu)}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, do

  $$x_i \overset{\text{iid}}{\sim} \chi^2_{(\nu)}$$
  $$\boldsymbol{y}_i \mid x_i \overset{\text{ind}}{\sim} \mathcal{N}(\boldsymbol{\mu}, \nu\boldsymbol{\Sigma}/x_i).$$

  This suggests the setup for EM is

  - $\boldsymbol{y}_{\text{obs}} = \boldsymbol{Y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)$.
  - $\boldsymbol{y}_{\text{comp}} = (\boldsymbol{Y}, \boldsymbol{x})$, where $\boldsymbol{x} = (x_1, \ldots, x_n)$.
  - $\boldsymbol{y}_{\text{miss}} = \boldsymbol{x}$.

Multivariate t: EM Algorithm

- **Model:** $\boldsymbol{y}_i \overset{\text{iid}}{\sim} t_{(\nu)}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \cong \mathcal{N}(0, \boldsymbol{\Sigma}) / \sqrt{\chi^2_{(\nu)}/\nu} + \boldsymbol{\mu}$.
- **Complete Data:** $x_i \overset{\text{iid}}{\sim} \chi^2_{(\nu)}, \qquad \boldsymbol{y}_i \mid x_i \overset{\text{ind}}{\sim} \mathcal{N}(\boldsymbol{\mu}, \nu\boldsymbol{\Sigma}/x_i)$.
- **Complete Data Likelihood:** With $\boldsymbol{\Omega} = \nu\boldsymbol{\Sigma}$ and $\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\Omega}, \nu)$,

$$\begin{aligned}
\ell(\boldsymbol{\theta} \mid \boldsymbol{Y}, \boldsymbol{x}) = &-\frac{1}{2}\left[n \log|\boldsymbol{\Omega}| + \sum_{i=1}^n x_i \cdot (\boldsymbol{y}_i - \boldsymbol{\mu})' \boldsymbol{\Omega}^{-1} (\boldsymbol{y}_i - \boldsymbol{\mu})\right] \\
&- \frac{1}{2}\left[n\nu \log(2) + 2n \log \Gamma(\nu/2) - \nu \sum_{i=1}^n \log(x_i)\right].
\end{aligned}$$

- **E-Step:**

$$\begin{aligned}
Q_t(\boldsymbol{\theta}) = &-\frac{1}{2}\left[n \log|\boldsymbol{\Omega}| + \sum_{i=1}^n E[x_i \mid \boldsymbol{y}_i, \hat{\boldsymbol{\theta}}^{(t)}] \cdot (\boldsymbol{y}_i - \boldsymbol{\mu})' \boldsymbol{\Omega}^{-1} (\boldsymbol{y}_i - \boldsymbol{\mu})\right] \\
&- \frac{1}{2}\left[n\nu \log(2) + 2n \log \Gamma(\nu/2) - \nu \sum_{i=1}^n E[\log(x_i) \mid \boldsymbol{y}_i, \hat{\boldsymbol{\theta}}^{(t)}]\right].
\end{aligned}$$

Multivariate t: EM Algorithm

- **Model:** $\boldsymbol{y}_i \overset{\text{iid}}{\sim} t_{(\nu)}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \cong \mathcal{N}(0, \boldsymbol{\Sigma})/\sqrt{\chi^2_{(\nu)}/\nu} + \boldsymbol{\mu}$.
- **Complete Data:** $x_i \overset{\text{iid}}{\sim} \chi^2_{(\nu)}, \qquad \boldsymbol{y}_i \mid x_i \overset{\text{ind}}{\sim} \mathcal{N}(\boldsymbol{\mu}, \nu\boldsymbol{\Sigma}/x_i)$.
- **E-Step:** Requires $E[x \mid \boldsymbol{y}, \boldsymbol{\theta}]$ and $E[\log(x) \mid \boldsymbol{y}, \boldsymbol{\theta}]$.
  - Conditional distribution of x:

$$
\begin{aligned}
p(x \mid \boldsymbol{y}, \boldsymbol{\theta}) &\propto p(\boldsymbol{y} \mid x, \boldsymbol{\theta}) \cdot p(x \mid \boldsymbol{\theta}) \\
&\propto \exp\left\{ \frac{\nu - 2}{2} \log(x) - \frac{1}{2} x \cdot (\boldsymbol{y} - \boldsymbol{\mu})' \boldsymbol{\Omega}^{-1} (\boldsymbol{y} - \boldsymbol{\mu}) + \frac{d}{2} \log(x) \right\} \\
&= \exp\left\{ (\alpha - 1) \log(x) - \beta \cdot x \right\},
\end{aligned}
$$

where $\alpha = \alpha(\boldsymbol{\theta}) = \frac{1}{2}(\nu + d)$

$\beta = \beta(\boldsymbol{\theta}) = \frac{1}{2}[(\boldsymbol{y} - \boldsymbol{\mu})' \boldsymbol{\Omega}^{-1} (y - \boldsymbol{\mu}) + 1]$.

$\implies x \mid \boldsymbol{y} \sim \text{Gamma}(\alpha, \beta)$.

Multivariate t: EM Algorithm

- **Model:** $\boldsymbol{y}_i \overset{\text{iid}}{\sim} t_{(\nu)}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \cong \mathcal{N}(0, \boldsymbol{\Sigma})/\sqrt{\chi^2_{(\nu)}/\nu} + \boldsymbol{\mu}$.
- **Complete Data:** $x_i \overset{\text{iid}}{\sim} \chi^2_{(\nu)}, \qquad \boldsymbol{y}_i \mid x_i \overset{\text{ind}}{\sim} \mathcal{N}(\boldsymbol{\mu}, \nu\boldsymbol{\Sigma}/x_i)$.
- **E-Step:** Requires $E[x \mid \boldsymbol{y}, \boldsymbol{\theta}]$ and $E[\log(x) \mid \boldsymbol{y}, \boldsymbol{\theta}]$.
  - $x \mid \boldsymbol{y}, \boldsymbol{\theta} \sim \text{Gamma}(\alpha, \beta)$, where $\alpha = \frac{1}{2}(\nu + d)$

    $$\beta = \tfrac{1}{2}[(\boldsymbol{y} - \boldsymbol{\mu})'\boldsymbol{\Omega}^{-1}(\boldsymbol{y} - \boldsymbol{\mu}) + 1].$$
  - Gamma distribution is an exponential family

    $$p(x \mid \boldsymbol{y}, \boldsymbol{\theta}) = \exp\{\alpha \log(x) - \beta \cdot x - \Psi(\alpha, -\beta)\} \cdot h(x),$$

    where $\Psi(\alpha, -\beta) = -\alpha \log(\beta) + \log \Gamma(\alpha)$.

    $\implies$ Sufficient statistics are $\boldsymbol{T} = (\log(x), x)$, such that

    $$E[\boldsymbol{T} \mid y] = \nabla\Psi(\alpha, -\beta) = \left(-\log(\beta) + \frac{\Gamma'(\alpha)}{\Gamma(\alpha)}, \frac{\alpha}{\beta}\right).$$

Multivariate t: EM Algorithm

- **Model:** $\boldsymbol{y}_i \overset{\text{iid}}{\sim} t_{(\nu)}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \cong \mathcal{N}(0, \boldsymbol{\Sigma}) / \sqrt{\chi^2_{(\nu)}/\nu} + \boldsymbol{\mu}$.
- **Complete Data:** $x_i \overset{\text{iid}}{\sim} \chi^2_{(\nu)}, \qquad \boldsymbol{y}_i \mid x_i \overset{\text{ind}}{\sim} \mathcal{N}(\boldsymbol{\mu}, \nu\boldsymbol{\Sigma}/x_i)$.
- **E-Step:**

$$
\begin{aligned}
Q_t(\boldsymbol{\theta}) = &-\frac{1}{2}\left[n \log|\boldsymbol{\Omega}| + \sum_{i=1}^n \hat{x}_i^{(t)} \cdot (\boldsymbol{y}_i - \boldsymbol{\mu})'\boldsymbol{\Omega}^{-1}(\boldsymbol{y}_i - \boldsymbol{\mu})\right] \\
&-\frac{1}{2}\left[n\nu \log(2) + 2n \log\Gamma(\nu/2) - \nu \sum_{i=1}^n \hat{w}_i^{(t)}\right],
\end{aligned}
$$

where

$$
\hat{x}_i^{(t)} = \frac{\hat{\alpha}^{(t)}}{\hat{\beta}_i^{(t)}} \qquad \hat{w}_i^{(t)} = -\log \hat{\beta}_i^{(t)} + \frac{\Gamma'(\hat{\alpha}^{(t)})}{\Gamma(\hat{\alpha}^{(t)})}
$$

$$
\hat{\alpha}^{(t)} = \tfrac{1}{2}(\hat{\nu}^{(t)} + d) \qquad \hat{\beta}_i^{(t)} = \tfrac{1}{2}[(\boldsymbol{y}_i - \hat{\boldsymbol{\mu}}^{(t)})'\boldsymbol{\Omega}^{-1}(\boldsymbol{y}_i - \hat{\boldsymbol{\mu}}^{(t)}) + 1].
$$

Multivariate t: EM Algorithm

- **Model:** $\boldsymbol{y}_i \overset{\text{iid}}{\sim} t_{(\nu)}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \cong \mathcal{N}(0, \boldsymbol{\Sigma}) / \sqrt{\chi^2_{(\nu)}/\nu} + \boldsymbol{\mu}$.
- **Complete Data:** $x_i \overset{\text{iid}}{\sim} \chi^2_{(\nu)}, \qquad \boldsymbol{y}_i \mid x_i \overset{\text{ind}}{\sim} \mathcal{N}(\boldsymbol{\mu}, \nu\boldsymbol{\Sigma}/x_i)$.
- **E-Step:**

$$
\begin{aligned}
Q_t(\boldsymbol{\theta}) = &-\frac{1}{2}\left[n \log|\boldsymbol{\Omega}| + \sum_{i=1}^n \hat{x}_i^{(t)} \cdot (\boldsymbol{y}_i - \boldsymbol{\mu})'\boldsymbol{\Omega}^{-1}(\boldsymbol{y}_i - \boldsymbol{\mu})\right] \\
&- \frac{1}{2}\left[n\nu\log(2) + 2n\log\Gamma(\nu/2) - \nu\sum_{i=1}^n \hat{w}_i^{(t)}\right].
\end{aligned}
$$

- **M-Step:**
  - $$\hat{\boldsymbol{\mu}}^{(t+1)} = \frac{\sum_{i=1}^n \hat{x}_i^{(t)} \boldsymbol{y}_i}{\sum_{i=1}^n \hat{x}_i^{(t)}}, \qquad \hat{\boldsymbol{\Omega}}^{(t+1)} = \frac{\sum_{i=1}^n \hat{x}_i^{(t)} (\boldsymbol{y}_i - \hat{\boldsymbol{\mu}}^{(t+1)})(\boldsymbol{y}_i - \hat{\boldsymbol{\mu}}^{(t+1)})'}{\sum_{i=1}^n \hat{x}_i^{(t)}}.$$
  - $\hat{\nu}^{(t+1)} = \arg\min_\nu \left\{ n\nu\log(2) + 2n\log\Gamma(\nu/2) - \nu\sum_{i=1}^n \hat{w}_i^{(t)} \right\}$, a convex optimization problem.