

Derivation of Maximum Likelihood Factor Analysis using EM*

September 26, 2000

1 The Factor Analysis Model

$$\mathbf{t} = \mathbf{W}\mathbf{x} + \boldsymbol{\epsilon} + \boldsymbol{\mu} \tag{1}$$

where

- \mathbf{t} → observed variables
- \mathbf{W} → matrix of factor loadings
- \mathbf{x} → hidden variables with distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$
- $\boldsymbol{\mu}$ → mean of observed variables (assume zero w.l.g.)
- $\boldsymbol{\epsilon}$ → noise with distribution $\mathcal{N}(\mathbf{0}, \Psi)$

2 Estimation of \mathbf{W} and Ψ using EM

Given a set of data points $\mathcal{D} = \{\mathbf{t}_i\}$ we wish to estimate the parameters \mathbf{W} and Ψ .

The complete data log-likelihood is given by

$$\begin{aligned} l_c(\mathbf{W}, \Psi) &= \log \prod_i^N p(\mathbf{t}_i, \mathbf{x}_i | \mathbf{W}, \Psi) \\ &= \sum_i^N \log p(\mathbf{t}_i, \mathbf{x}_i | \mathbf{W}, \Psi) \\ &= \sum_i^N \log p(\mathbf{t}_i | \mathbf{x}_i, \mathbf{W}, \Psi) p(\mathbf{x}_i | \mathbf{W}, \Psi) \\ &= \sum_i^N \log p(\mathbf{t}_i | \mathbf{x}_i, \mathbf{W}, \Psi) + \sum_i^N \log p(\mathbf{x}_i | \mathbf{W}, \Psi) \end{aligned}$$

*Compiled by Aaron D'Souza. Direct comments, bug reports, raspberries, etc. to adsouza@rubens.usc.edu

but since the distribution of \mathbf{x} is independent of \mathbf{W} and Ψ

$$l_c(\mathbf{W}, \Psi) = \sum_i^N \log p(\mathbf{t}_i | \mathbf{x}_i, \mathbf{W}, \Psi) + \sum_i^N \log p(\mathbf{x}_i) \quad (2)$$

Since the second term in eq. (2) is independent of \mathbf{W} and Ψ , it suffices (for the purpose of estimating \mathbf{W} and Ψ) to only deal with the term

$$L = \sum_i^N \log p(\mathbf{t}_i | \mathbf{x}_i, \mathbf{W}, \Psi)$$

Now for the distribution $p(\mathbf{t} | \mathbf{x})$

$$\mathcal{E}\{\mathbf{t} | \mathbf{x}\} = \mathcal{E}\{(\mathbf{W}\mathbf{x} + \boldsymbol{\epsilon}) | \mathbf{x}\} = \mathbf{W}\mathbf{x} \quad (3)$$

$$\text{Cov}(\mathbf{t} | \mathbf{x}) = \mathcal{E}\{(\mathbf{t} - \mathbf{W}\mathbf{x})(\mathbf{t} - \mathbf{W}\mathbf{x})^T | \mathbf{x}\} = \mathcal{E}\{\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T | \mathbf{x}\} = \Psi \quad (4)$$

Hence we can expand L as

$$\begin{aligned} L &= \sum_i^N \log \frac{1}{(2\pi)^{d/2} |\Psi|^{1/2}} \exp \left\{ -\frac{1}{2} (\mathbf{t}_i - \mathbf{W}\mathbf{x}_i)^T \Psi^{-1} (\mathbf{t}_i - \mathbf{W}\mathbf{x}_i) \right\} \\ &= -\frac{Nd}{2} \log(2\pi) - \frac{N}{2} \log |\Psi| - \frac{1}{2} \sum_i^N (\mathbf{t}_i^T \Psi^{-1} \mathbf{t}_i - 2\mathbf{t}_i^T \Psi^{-1} \mathbf{W}\mathbf{x}_i + \mathbf{x}_i^T \mathbf{W}^T \Psi^{-1} \mathbf{W}\mathbf{x}_i) \\ &= -\frac{Nd}{2} \log(2\pi) - \frac{N}{2} \log |\Psi| - \frac{1}{2} \sum_i^N (\mathbf{t}_i^T \Psi^{-1} \mathbf{t}_i - 2\mathbf{t}_i^T \Psi^{-1} \mathbf{W}\mathbf{x}_i + \text{Tr}[\mathbf{W}^T \Psi^{-1} \mathbf{W}\mathbf{x}_i \mathbf{x}_i^T]) \end{aligned}$$

In the last step we have exploited the relation $\mathbf{x}^T \mathbf{A} \mathbf{x} = \text{Tr}[\mathbf{A} \mathbf{x} \mathbf{x}^T]$, where $\text{Tr}[\cdot]$ is the trace operator.

Taking the expectation of L according to $p(\mathbf{x}_i | \mathbf{t}_i, \mathbf{W}, \Psi)$ we get

$$\mathcal{E}\{L\} = k - \frac{N}{2} \log |\Psi| - \frac{1}{2} \sum_i^N (\mathbf{t}_i^T \Psi^{-1} \mathbf{t}_i - 2\mathbf{t}_i^T \Psi^{-1} \mathbf{W} \mathcal{E}\{\mathbf{x}_i | \mathbf{t}_i\} + \text{Tr}[\mathbf{W}^T \Psi^{-1} \mathbf{W} \mathcal{E}\{\mathbf{x}_i \mathbf{x}_i^T | \mathbf{t}_i\}]) \quad (5)$$

Maximizing eq. (5) w.r.t. \mathbf{W}

$$\frac{\partial \mathcal{E}\{L\}}{\partial \mathbf{W}} = -\frac{1}{2} \sum_i^N \left(-2\Psi^{-1} \mathbf{t}_i \mathcal{E}\{\mathbf{x}_i | \mathbf{t}_i\}^T + 2\Psi^{-1} \mathbf{W} \mathcal{E}\{\mathbf{x}_i \mathbf{x}_i^T | \mathbf{t}_i\} \right) = \mathbf{0}$$

Where we have used the relations $\frac{\partial \mathbf{A}^T \mathbf{X} \mathbf{B}}{\partial \mathbf{X}} = \mathbf{A} \mathbf{B}^T$, and $\frac{\partial \text{Tr}[\mathbf{X}^T \mathbf{A} \mathbf{X} \mathbf{B}]}{\partial \mathbf{X}} = \mathbf{A} \mathbf{X} \mathbf{B} + \mathbf{A}^T \mathbf{X} \mathbf{B}^T$.

Hence

$$\mathbf{W} \sum_i^N \mathcal{E}\{\mathbf{x} \mathbf{x}^T | \mathbf{t}\} = \sum_i^N \mathbf{t}_i \mathcal{E}\{\mathbf{x}_i | \mathbf{t}_i\}^T$$

and we arrive at the update equation

$$\mathbf{W} = \left(\sum_i^N \mathbf{t}_i \mathcal{E}\{\mathbf{x}_i | \mathbf{t}_i\}^T \right) \left(\sum_i^N \mathcal{E}\{\mathbf{x}_i \mathbf{x}_i^T | \mathbf{t}_i\} \right)^{-1} \quad (6)$$

Maximizing eq. (5) w.r.t. Ψ^{-1}

$$\begin{aligned}
\frac{\partial \mathcal{E}\{L\}}{\partial \Psi^{-1}} &= \frac{N}{2} \Psi - \frac{1}{2} \sum_i^N \left(\mathbf{t}_i \mathbf{t}_i^T - 2 \mathbf{t}_i \mathcal{E}\{\mathbf{x}_i | \mathbf{t}_i\}^T \mathbf{W}^T + \mathbf{W} \mathcal{E}\{\mathbf{x}_i \mathbf{x}_i^T | \mathbf{t}_i\} \mathbf{W}^T \right) \\
&= \frac{N}{2} \Psi - \frac{1}{2} \sum_i^N \mathbf{t}_i \mathbf{t}_i^T + \left(\sum_i^N \mathbf{t}_i \mathcal{E}\{\mathbf{x}_i | \mathbf{t}_i\}^T \right) \mathbf{W}^T - \frac{1}{2} \mathbf{W} \left(\sum_i^N \mathcal{E}\{\mathbf{x}_i \mathbf{x}_i^T | \mathbf{t}_i\} \right) \mathbf{W}^T \\
&= \mathbf{0}
\end{aligned} \tag{7}$$

Where we have used the relations $\frac{\partial \log |\mathbf{X}|}{\partial \mathbf{X}} = (\mathbf{X}^{-1})^T$, and $\frac{\partial \mathbf{A}^T \mathbf{X} \mathbf{B}}{\partial \mathbf{X}} = \mathbf{A} \mathbf{B}^T$.

Hence

$$\Psi = \frac{1}{N} \left[\sum_i^N \mathbf{t}_i \mathbf{t}_i^T - 2 \left(\sum_i^N \mathbf{t}_i \mathcal{E}\{\mathbf{x}_i | \mathbf{t}_i\}^T \right) \mathbf{W}^T + \mathbf{W} \left(\sum_i^N \mathcal{E}\{\mathbf{x}_i \mathbf{x}_i^T | \mathbf{t}_i\} \right) \mathbf{W}^T \right] \tag{8}$$

Substituting the value of \mathbf{W} from the update equation we get

$$\begin{aligned}
\Psi &= \frac{1}{N} \left[\sum_i^N \mathbf{t}_i \mathbf{t}_i^T - 2 \left(\sum_i^N \mathbf{t}_i \mathcal{E}\{\mathbf{x}_i | \mathbf{t}_i\}^T \right) \mathbf{W}^T \right. \\
&\quad \left. + \left(\sum_i^N \mathbf{t}_i \mathcal{E}\{\mathbf{x}_i | \mathbf{t}_i\}^T \right) \left(\sum_i^N \mathcal{E}\{\mathbf{x}_i \mathbf{x}_i^T | \mathbf{t}_i\} \right)^{-1} \left(\sum_i^N \mathcal{E}\{\mathbf{x}_i \mathbf{x}_i^T | \mathbf{t}_i\} \right) \mathbf{W}^T \right]
\end{aligned}$$

giving us the update equation

$$\Psi = \frac{1}{N} \text{diag} \left[\sum_i^N \mathbf{t}_i \mathbf{t}_i^T - \left(\sum_i^N \mathbf{t}_i \mathcal{E}\{\mathbf{x}_i | \mathbf{t}_i\}^T \right) \mathbf{W}^T \right] \tag{9}$$

in which the $\text{diag}[\cdot]$ operator constrains Ψ to be a diagonal matrix.

All that remains now is to determine the values of $\mathcal{E}\{\mathbf{x}_i | \mathbf{t}_i\}$ and $\mathcal{E}\{\mathbf{x}_i \mathbf{x}_i^T | \mathbf{t}_i\}$.

3 Calculation of the sufficient statistics

For the distribution of the observed variable $p(\mathbf{t})$

$$\begin{aligned}
\mathcal{E}\{\mathbf{t}\} &= \mathcal{E}\{\mathbf{W}\mathbf{x} + \boldsymbol{\epsilon}\} \\
&= \mathbf{W} \mathcal{E}\{\mathbf{x}\} + \mathcal{E}\{\boldsymbol{\epsilon}\} \\
&= \mathbf{0}
\end{aligned} \tag{10}$$

$$\begin{aligned}
\text{Cov}(\mathbf{t}) &= \mathcal{E}\{\mathbf{t} \mathbf{t}^T\} = \mathcal{E}\{(\mathbf{W}\mathbf{x} + \boldsymbol{\epsilon})(\mathbf{W}\mathbf{x} + \boldsymbol{\epsilon})^T\} \\
&= \mathbf{W} \mathcal{E}\{\mathbf{x} \mathbf{x}^T\} \mathbf{W}^T + \mathbf{W} \mathcal{E}\{\mathbf{x} \boldsymbol{\epsilon}^T\} + \mathcal{E}\{\boldsymbol{\epsilon} \mathbf{x}^T\} \mathbf{W}^T + \mathcal{E}\{\boldsymbol{\epsilon} \boldsymbol{\epsilon}^T\} \\
&= \mathbf{W} \mathbf{W}^T + \mathbf{0} + \mathbf{0} + \Psi \\
&= \mathbf{W} \mathbf{W}^T + \Psi \equiv \mathbf{C}
\end{aligned} \tag{11}$$

For the complete data distribution $p(\mathbf{t}, \mathbf{x})$

$$\begin{aligned} \text{Let } \mathbf{y} &= \begin{bmatrix} \mathbf{t} \\ \mathbf{x} \end{bmatrix} \\ \mathcal{E}\{\mathbf{y}\} &= \mathcal{E}\left\{\begin{bmatrix} \mathbf{t} \\ \mathbf{x} \end{bmatrix}\right\} = \mathbf{0} \end{aligned} \quad (12)$$

$$\begin{aligned} \text{Cov}(\mathbf{y}) &= \mathcal{E}\{\mathbf{y}\mathbf{y}^T\} = \mathcal{E}\left\{\begin{bmatrix} \mathbf{t} \\ \mathbf{x} \end{bmatrix} \begin{bmatrix} \mathbf{t}^T & \mathbf{x}^T \end{bmatrix}\right\} \\ &= \mathcal{E}\left\{\begin{bmatrix} \mathbf{t}\mathbf{t}^T & \mathbf{t}\mathbf{x}^T \\ \mathbf{x}\mathbf{t}^T & \mathbf{x}\mathbf{x}^T \end{bmatrix}\right\} \\ &= \begin{bmatrix} \Psi + \mathbf{W}\mathbf{W}^T & \mathbf{W} \\ \mathbf{W}^T & \mathbf{I} \end{bmatrix} \equiv \Lambda \end{aligned} \quad (13)$$

For the distribution $p(\mathbf{x}|\mathbf{t})$

$$\begin{aligned} p(\mathbf{x}|\mathbf{t}) &= \frac{p(\mathbf{t}, \mathbf{x})}{p(\mathbf{t})} \\ &= \frac{(2\pi)^{-(d+q)/2} |\Lambda|^{-1/2} \exp\left(-\frac{1}{2}\mathbf{y}^T \Lambda^{-1} \mathbf{y}\right)}{(2\pi)^{-d/2} |\mathbf{C}|^{-1/2} \exp\left(-\frac{1}{2}\mathbf{t}^T \mathbf{C}^{-1} \mathbf{t}\right)} \\ &= (\text{blah}) \exp\left(-\frac{1}{2}(\mathbf{y}^T \Lambda^{-1} \mathbf{y} - \mathbf{t}^T \mathbf{C}^{-1} \mathbf{t})\right) \\ &= (\text{blah}) \exp\left(-\frac{1}{2}\alpha\right) \end{aligned} \quad (14)$$

where we define

$$\alpha \equiv \mathbf{y}^T \Lambda^{-1} \mathbf{y} - \mathbf{t}^T \mathbf{C}^{-1} \mathbf{t} \quad (15)$$

In the above eq.s $\mathbf{C} = \text{Cov}(\mathbf{t})$ as defined in eq. (11), and $\Lambda = \text{Cov}(\mathbf{t}, \mathbf{x})$ as defined in eq. (13).

Also, since

$$\begin{aligned} \Lambda &= \begin{bmatrix} \Psi + \mathbf{W}\mathbf{W}^T & \mathbf{W} \\ \mathbf{W}^T & \mathbf{I} \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix} \end{aligned}$$

we have

$$\begin{aligned} \Lambda^{-1} &= \begin{bmatrix} \Lambda^{-1,11} & \Lambda^{-1,12} \\ \Lambda^{-1,21} & \Lambda^{-1,22} \end{bmatrix} \\ &= \begin{bmatrix} (\mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21})^{-1} & \mathbf{A}_{11}^{-1}\mathbf{A}_{12}(\mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12} - \mathbf{A}_{22})^{-1} \\ (\mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12} - \mathbf{A}_{22})^{-1}\mathbf{A}_{21}\mathbf{A}_{11}^{-1} & (\mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12})^{-1} \end{bmatrix} \end{aligned}$$

Consider the term inside the exponent of eq. (14)

$$\begin{aligned}
\alpha &= \mathbf{y}^T \Lambda^{-1} \mathbf{y} - \mathbf{t}^T \mathbf{C}^{-1} \mathbf{t} \\
&= [\mathbf{t}^T \quad \mathbf{x}^T] \Lambda^{-1} \begin{bmatrix} \mathbf{t} \\ \mathbf{x} \end{bmatrix} - \mathbf{t}^T \mathbf{C}^{-1} \mathbf{t} \\
&= \mathbf{t}^T \Lambda^{-1,11} \mathbf{t} + \mathbf{t}^T \Lambda^{-1,12} \mathbf{x} + \mathbf{x}^T \Lambda^{-1,21} \mathbf{t} + \mathbf{x}^T \Lambda^{-1,22} \mathbf{x} - \mathbf{t}^T \mathbf{C}^{-1} \mathbf{t} \\
&= \mathbf{t}^T (\Lambda^{-1,11} - \mathbf{C}^{-1}) \mathbf{t} + \mathbf{t}^T \Lambda^{-1,12} \mathbf{x} + \mathbf{x}^T \Lambda^{-1,21} \mathbf{t} + \mathbf{x}^T \Lambda^{-1,22} \mathbf{x} \\
&= \mathbf{t}^T (\Lambda^{-1,11} - \mathbf{C}^{-1}) \mathbf{t} + 2\mathbf{t}^T \Lambda^{-1,12} \mathbf{x} + \mathbf{x}^T \Lambda^{-1,22} \mathbf{x}
\end{aligned} \tag{16}$$

Consider the term

$$\begin{aligned}
\Lambda^{-1,11} - \mathbf{C}^{-1} &= (\mathbf{A}_{11} - \mathbf{A}_{12} \mathbf{A}_{22}^{-1} \mathbf{A}_{21}) - \mathbf{A}_{11}^{-1} \\
&= \mathbf{A}_{11}^{-1} + \mathbf{A}_{11}^{-1} \mathbf{A}_{12} (\mathbf{A}_{22} - \mathbf{A}_{21} \mathbf{A}_{11}^{-1} \mathbf{A}_{12}) \mathbf{A}_{21} \mathbf{A}_{11}^{-1} - \mathbf{A}_{11}^{-1} \\
&= \mathbf{A}_{11}^{-1} \mathbf{A}_{12} (\mathbf{A}_{22} - \mathbf{A}_{21} \mathbf{A}_{11}^{-1} \mathbf{A}_{12}) \mathbf{A}_{21} \mathbf{A}_{11}^{-1} \\
&= \boldsymbol{\beta}^T \Lambda^{-1,22} \boldsymbol{\beta}
\end{aligned} \tag{17}$$

where we define

$$\begin{aligned}
\boldsymbol{\beta} &= \mathbf{A}_{21} \mathbf{A}_{11}^{-1} \\
&= \mathbf{W}^T (\boldsymbol{\Psi} + \mathbf{W} \mathbf{W}^T)^{-1}
\end{aligned} \tag{18}$$

which we can also write as (refer to appendix C)

$$\boldsymbol{\beta} = (\mathbf{I} + \mathbf{W}^T \boldsymbol{\Psi}^{-1} \mathbf{W})^{-1} \mathbf{W}^T \boldsymbol{\Psi}^{-1} \tag{19}$$

Substituting eq. (17) into eq. (16) we get

$$\begin{aligned}
\alpha &= \mathbf{t}^T \boldsymbol{\beta}^T \Lambda^{-1,22} \boldsymbol{\beta} \mathbf{t} + 2\mathbf{t}^T \Lambda^{-1,12} \mathbf{x} + \mathbf{x}^T \Lambda^{-1,22} \mathbf{x} \\
&= (\mathbf{x} - \boldsymbol{\beta} \mathbf{t})^T \Lambda^{-1,22} (\mathbf{x} - \boldsymbol{\beta} \mathbf{t}) + 2\mathbf{t}^T \boldsymbol{\beta}^T \Lambda^{-1,22} \mathbf{x} + 2\mathbf{t}^T \Lambda^{-1,12} \mathbf{x} \\
&= (\mathbf{x} - \boldsymbol{\beta} \mathbf{t})^T \Lambda^{-1,22} (\mathbf{x} - \boldsymbol{\beta} \mathbf{t}) + 2\mathbf{t}^T (\boldsymbol{\beta}^T \Lambda^{-1,22} + \Lambda^{-1,12}) \mathbf{x}
\end{aligned} \tag{20}$$

Now

$$\begin{aligned}
\boldsymbol{\beta}^T \Lambda^{-1,22} + \Lambda^{-1,12} &= \mathbf{A}_{11}^{-1} \mathbf{A}_{12} (\mathbf{A}_{22} - \mathbf{A}_{21} \mathbf{A}_{11}^{-1} \mathbf{A}_{12})^{-1} \\
&\quad + \mathbf{A}_{11}^{-1} \mathbf{A}_{12} (\mathbf{A}_{21} \mathbf{A}_{11}^{-1} \mathbf{A}_{12} - \mathbf{A}_{22})^{-1} \\
&= \mathbf{0}
\end{aligned}$$

Hence

$$\alpha = (\mathbf{x} - \boldsymbol{\beta} \mathbf{t})^T \Lambda^{-1,22} (\mathbf{x} - \boldsymbol{\beta} \mathbf{t}) \tag{21}$$

Which from eq. (14) implies that $p(\mathbf{x}|\mathbf{t})$ has the form

$$p(\mathbf{x}|\mathbf{t}) = (\text{blah}) \exp \left(-\frac{1}{2} (\mathbf{x} - \boldsymbol{\beta} \mathbf{t})^T \Lambda^{-1,22} (\mathbf{x} - \boldsymbol{\beta} \mathbf{t}) \right)$$

from which we can deduce

$$\mathcal{E} \{ \mathbf{x} | \mathbf{t} \} = \boldsymbol{\beta} \mathbf{t} = (\mathbf{I} + \mathbf{W}^T \boldsymbol{\Psi}^{-1} \mathbf{W})^{-1} \mathbf{W}^T \boldsymbol{\Psi}^{-1} \mathbf{t} \tag{22}$$

and

$$\text{Cov}(\mathbf{x}|\mathbf{t}) = (\Lambda^{-1,22})^{-1} \quad (23)$$

but

$$\begin{aligned} [\text{Cov}(\mathbf{x}|\mathbf{t})]^{-1} &= \Lambda^{-1,22} \\ &= (\mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12})^{-1} \\ &= \mathbf{A}_{22}^{-1} - \mathbf{A}_{22}^{-1}\mathbf{A}_{21}(-\mathbf{A}_{11} + \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21})^{-1}\mathbf{A}_{12}\mathbf{A}_{22}^{-1} \\ &= \mathbf{I} - \mathbf{W}^T(-\Psi - \mathbf{W}\mathbf{W}^T + \mathbf{W}\mathbf{W}^T)^{-1}\mathbf{W} \\ &= \mathbf{I} + \mathbf{W}^T\Psi^{-1}\mathbf{W} \end{aligned} \quad (24)$$

and hence we can write

$$\text{Cov}(\mathbf{x}|\mathbf{t}) = (\mathbf{I} + \mathbf{W}^T\Psi^{-1}\mathbf{W})^{-1} \quad (25)$$

Also, since

$$\begin{aligned} \text{Cov}(\mathbf{x}|\mathbf{t}) &= \mathcal{E}\{\mathbf{x}\mathbf{x}^T|\mathbf{t}\} - \mathcal{E}\{\mathbf{x}|\mathbf{t}\}\mathcal{E}\{\mathbf{x}|\mathbf{t}\}^T \\ \mathcal{E}\{\mathbf{x}\mathbf{x}^T|\mathbf{t}\} &= \text{Cov}(\mathbf{x}|\mathbf{t}) + \mathcal{E}\{\mathbf{x}|\mathbf{t}\}\mathcal{E}\{\mathbf{x}|\mathbf{t}\}^T \\ &= (\mathbf{I} + \mathbf{W}^T\Psi^{-1}\mathbf{W})^{-1} + \beta\mathbf{t}\mathbf{t}^T\beta^T \\ &= \mathbf{I} - \mathbf{W}^T(\Psi + \mathbf{W}\mathbf{W}^T)^{-1}\mathbf{W} + \beta\mathbf{t}\mathbf{t}^T\beta^T \\ &= \mathbf{I} - \beta\mathbf{W} + \beta\mathbf{t}\mathbf{t}^T\beta^T \end{aligned} \quad (26)$$

A Matrix Inversion Theorem

$$(\mathbf{A} + \mathbf{X}\mathbf{R}\mathbf{Y})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{X}(\mathbf{R}^{-1} + \mathbf{Y}\mathbf{A}^{-1}\mathbf{X})^{-1}\mathbf{Y}\mathbf{A}^{-1}$$

B Partitioned Matrix Inversion Theorem

If we define

$$\Lambda = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix}$$

then

$$\begin{aligned} \Lambda^{-1} &= \begin{bmatrix} \Lambda^{-1,11} & \Lambda^{-1,12} \\ \Lambda^{-1,21} & \Lambda^{-1,22} \end{bmatrix} \\ &= \begin{bmatrix} (\mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21})^{-1} & \mathbf{A}_{11}^{-1}\mathbf{A}_{12}(\mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12} - \mathbf{A}_{22})^{-1} \\ (\mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12} - \mathbf{A}_{22})^{-1}\mathbf{A}_{21}\mathbf{A}_{11}^{-1} & (\mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12})^{-1} \end{bmatrix} \end{aligned}$$

C Another useful result

$$\begin{aligned}\mathbf{A} &= \mathbf{W}^T(\Psi + \mathbf{W}\mathbf{W}^T)^{-1} \\ &= \mathbf{W}^T [\Psi^{-1} - \Psi^{-1}\mathbf{W}(\mathbf{I} + \mathbf{W}^T\Psi^{-1}\mathbf{W})^{-1}\mathbf{W}^T\Psi^{-1}] \\ &= [\mathbf{I} - \mathbf{W}^T\Psi^{-1}\mathbf{W}(\mathbf{I} + \mathbf{W}^T\Psi^{-1}\mathbf{W})^{-1}] \mathbf{W}^T\Psi^{-1} \\ &= [\mathbf{I} + (\mathbf{I} + \mathbf{W}^T\Psi^{-1}\mathbf{W})^{-1} - (\mathbf{I} + \mathbf{W}^T\Psi^{-1}\mathbf{W})(\mathbf{I} + \mathbf{W}^T\Psi^{-1}\mathbf{W})^{-1}] \mathbf{W}^T\Psi^{-1} \\ &= (\mathbf{I} + \mathbf{W}^T\Psi^{-1}\mathbf{W})^{-1}\mathbf{W}^T\Psi^{-1}\end{aligned}$$

In a similar fashion, we can also prove the more general result

$$\mathbf{W}^T(\Psi + \mathbf{W}\mathbf{A}\mathbf{W}^T)^{-1} = \mathbf{A}^{-1}(\mathbf{A}^{-1} + \mathbf{W}^T\Psi^{-1}\mathbf{W})^{-1}\mathbf{W}^T\Psi^{-1}$$