Supplementary Materials for Multilingual Factor Analysis

Anonymous ACL submission

A Joint Distribution

We show the form of the joint distribution for 2 views. Concatenating our data and parameters as below, we can use Equation (3) of (Ghahramani et al., 1996) to write

$$\boldsymbol{m} = \begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{bmatrix}, \boldsymbol{W} = \begin{bmatrix} \boldsymbol{W}_x \\ \boldsymbol{W}_y \end{bmatrix}$$

$$\boldsymbol{\Psi} = \begin{bmatrix} \boldsymbol{\Psi_x} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\Psi_y} \end{bmatrix}, \boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu_x} \\ \boldsymbol{\mu_y} \end{bmatrix}$$

$$p(\boldsymbol{m}, \boldsymbol{z}|\boldsymbol{\theta}) = \mathcal{N}\left(\begin{bmatrix} \boldsymbol{m} \\ \boldsymbol{z} \end{bmatrix}; \begin{bmatrix} \boldsymbol{\mu} \\ \boldsymbol{0} \end{bmatrix}, \boldsymbol{\Sigma}_{m,z}\right) \quad (1)$$

$$\boldsymbol{\Sigma}_{m,z} = \begin{bmatrix} \boldsymbol{W}\boldsymbol{W}^\top + \boldsymbol{\Psi} & \boldsymbol{W} \\ \boldsymbol{W}^T & \mathbb{I} \end{bmatrix}$$

It is clear that this generalises to any number of views of any dimension, as the concatenation operation does not make any assumptions.

B Projections to Latent Space $\mathbb{E}_{p(\boldsymbol{z}|\boldsymbol{x})}[\boldsymbol{z}]$

We can query the joint Gaussian in 1 using rules from (Petersen et al., 2008) Sections (8.1.2, 8.1.3) and we get

$$p(\boldsymbol{z}|\boldsymbol{x}) = \mathcal{N}\left(\boldsymbol{z}; \boldsymbol{W}_x^\top \boldsymbol{\Sigma}_x^{-1}\tilde{\boldsymbol{x}}, \mathbb{I} - \boldsymbol{W}_x^\top \boldsymbol{\Sigma}_x^{-1}\boldsymbol{W}_x\right)$$

$$\mathbb{E}[\boldsymbol{z}|\boldsymbol{x}] = \boldsymbol{W}_x^\top \boldsymbol{\Sigma}_x^{-1}\tilde{\boldsymbol{x}}$$

C Derivation for the Marginal Likelihood

We want to compute $p(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{\theta})$ so that we can then learn the parameters $\boldsymbol{\theta} = \{\boldsymbol{\theta}_x, \boldsymbol{\theta}_y\}$, $\boldsymbol{\theta}_i = \{\boldsymbol{\mu}_i, \boldsymbol{W}_i, \boldsymbol{\Psi}_i, \}$ by maximising the marginal likelihood as is done in Factor Analysis.

From the joint $p(\boldsymbol{m}, \boldsymbol{z}|\boldsymbol{\theta})$, again using rules from (Petersen et al., 2008) Sections (8.1.2) we get

$$\begin{aligned} p(\boldsymbol{m}|\boldsymbol{\theta}) &= p(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{\theta}) \\ &= \mathcal{N}\left(\begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{bmatrix}; \begin{bmatrix} \boldsymbol{\mu}_x \\ \boldsymbol{\mu}_y \end{bmatrix}, \boldsymbol{W}\boldsymbol{W}^T + \boldsymbol{\Psi}\right) \end{aligned}$$

For the case of two views, the joint probability can be factored as

$$\begin{aligned} p(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{\theta}) &= p(\boldsymbol{x}|\boldsymbol{\theta}_x)p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta}) \\ p(\boldsymbol{x}|\boldsymbol{\theta}_x) &= \mathcal{N}\left(\boldsymbol{x}; \boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x\right) \\ p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta}) &= \mathcal{N}\left(\boldsymbol{y}; \boldsymbol{W}_y\boldsymbol{W}_x^\top \boldsymbol{\Sigma}_x^{-1}\tilde{\boldsymbol{x}} + \boldsymbol{\mu}_y, \boldsymbol{\Sigma}_{\boldsymbol{y}|\boldsymbol{x}}\right) \\ &= \mathcal{N}\left(\boldsymbol{y}; \boldsymbol{W}_y E[\boldsymbol{z}|\boldsymbol{x}] + \boldsymbol{\mu}_y, \boldsymbol{\Sigma}_{\boldsymbol{y}|\boldsymbol{x}}\right), \end{aligned}$$

where

$$\begin{aligned} \boldsymbol{\Sigma}_x &= \boldsymbol{W}_x\boldsymbol{W}_x^T + \boldsymbol{\Psi}_x \\ \boldsymbol{\Sigma}_{\boldsymbol{y}|\boldsymbol{x}} &= \boldsymbol{\Sigma}_y - \boldsymbol{W}_y\boldsymbol{W}_x^\top \boldsymbol{\Sigma}_x^{-1}\boldsymbol{W}_x\boldsymbol{W}_y^\top \end{aligned}$$

D Scaled Reconstruction Errors

$$\begin{aligned} \log p(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{\theta}) &= \log p^*(\boldsymbol{x}|\boldsymbol{\theta}_x) + \log p^*(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta}) + C \\ C &= -\frac{1}{2}(\log|2\pi\boldsymbol{\Sigma}_{\boldsymbol{y}|\boldsymbol{x}}| + \log|2\pi\boldsymbol{\Sigma}_{\boldsymbol{x}}|) \\ \log p^*(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta}) &= -\frac{1}{2}||\tilde{\boldsymbol{y}} - \boldsymbol{W}_y E[\boldsymbol{z}|\boldsymbol{x}]||^2_{\boldsymbol{\Sigma}_{y|x}} \quad (2) \\ \log p^*(\boldsymbol{x}|\boldsymbol{\theta}_x) &= -\frac{1}{2}||\boldsymbol{x} - \boldsymbol{\mu}_x||^2_{\boldsymbol{\Sigma}_x} \\ &= -\frac{1}{2}||\boldsymbol{\Sigma}_x^{-\frac{1}{2}}\tilde{\boldsymbol{x}}||^2 \end{aligned}$$

Setting $\boldsymbol{A} = \boldsymbol{\Psi}_x\boldsymbol{\Sigma}_x^{-1}\boldsymbol{\Psi}_x$, we can re-parametrise as

$$\begin{aligned} \log p^*(\boldsymbol{x}|\boldsymbol{\theta}_x) &= -\frac{1}{2}||\boldsymbol{\Psi}_x\boldsymbol{\Sigma}_x^{-1}\tilde{\boldsymbol{x}}||^2_{\boldsymbol{A}} \\ &= -\frac{1}{2}||(\boldsymbol{\Sigma}_x - \boldsymbol{W}_x\boldsymbol{W}_x^\top)\boldsymbol{\Sigma}_x^{-1}\tilde{\boldsymbol{x}}||^2_{\boldsymbol{A}} \\ &= -\frac{1}{2}||\tilde{\boldsymbol{x}} - \boldsymbol{W}_x\boldsymbol{W}_x^\top\boldsymbol{\Sigma}_x^{-1}\tilde{\boldsymbol{x}}||^2_{\boldsymbol{A}} \\ &= -\frac{1}{2}||\tilde{\boldsymbol{x}} - \boldsymbol{W}_x E[\boldsymbol{z}|\boldsymbol{x}]||^2_{\boldsymbol{A}} \end{aligned}$$

E Expectation Maximisation for MBFA

Define

$$\tilde{\boldsymbol{x}} = \begin{bmatrix} \boldsymbol{x}_1 - \boldsymbol{\mu}_1 \\ \vdots \\ \boldsymbol{x}_v - \boldsymbol{\mu}_1 \end{bmatrix}, \boldsymbol{W} = \begin{bmatrix} \boldsymbol{W}_1 \\ \vdots \\ \boldsymbol{W}_v \end{bmatrix}$$

$$\boldsymbol{\Psi} = \begin{bmatrix} \boldsymbol{\Psi}_1 & & 0 \\ & \ddots & \\ 0 & & \boldsymbol{\Psi}_v \end{bmatrix} = \text{Bdiag}(\boldsymbol{\Psi}_1, \ldots, \boldsymbol{\Psi}_v)$$

Hence

$$p(\tilde{\boldsymbol{x}}|\boldsymbol{z}; \boldsymbol{\Psi}, \boldsymbol{W}) = \mathcal{N}(\tilde{\boldsymbol{x}}|\boldsymbol{W}\boldsymbol{z}, \boldsymbol{\Psi})$$

This follows the same form as regular factor analysis, but with a block-diagonal constraint on $\boldsymbol{\Psi}$. Thus by Equations (5) and (6) of (Ghahramani et al., 1996), we apply EM as follows.

E-Step: Compute $\mathbb{E}[\boldsymbol{z}|\boldsymbol{x}]$ and $\mathbb{E}[\boldsymbol{z}\boldsymbol{z}^\top|\boldsymbol{x}]$ given the parameters $\boldsymbol{\theta}_t = \{\boldsymbol{W}_t, \boldsymbol{\Psi}_t\}$.

$$\begin{aligned} \mathbb{E}[\boldsymbol{z}^{(i)}|\tilde{\boldsymbol{x}}^{(i)}] &= \boldsymbol{B}_t\tilde{\boldsymbol{x}}^{(i)} \\ \mathbb{E}[\boldsymbol{z}^{(i)}\boldsymbol{z}^{(i)\top}|\tilde{\boldsymbol{x}}^{(i)}] &= \mathbb{I} - \boldsymbol{B}_t\boldsymbol{W}_t + \boldsymbol{B}_t\tilde{\boldsymbol{x}}^{(i)}\tilde{\boldsymbol{x}}^{(i)\top}\boldsymbol{B}_t^\top \\ &= \boldsymbol{M}_t + \boldsymbol{B}_t\tilde{\boldsymbol{x}}^{(i)}\tilde{\boldsymbol{x}}^{(i)\top}\boldsymbol{B}_t^\top \qquad (3) \end{aligned}$$

where

$$\begin{aligned} \boldsymbol{M}_t &= \left(\mathbb{I} + \boldsymbol{W}_t^\top\boldsymbol{\Psi}_t^{-1}\boldsymbol{W}_t\right)^{-1} \\ \boldsymbol{B}_t &= \boldsymbol{W}_t^\top(\boldsymbol{\Psi}_t + \boldsymbol{W}_t\boldsymbol{W}_t^\top)^{-1} \\ &= \boldsymbol{M}_t\boldsymbol{W}_t^\top\boldsymbol{\Psi}_t^{-1}. \qquad (4) \end{aligned}$$

Equation 3 is obtained by applying the Woodbury identity, and Equation 4 by applying the closely related push-through identity, as found in Section 3.2 of (Petersen et al., 2008).

M-Step: Update parameters $\boldsymbol{\theta}_{t+1} = \{\boldsymbol{W}_{t+1}, \boldsymbol{\Psi}_{t+1}\}$.

Define

$$\boldsymbol{S} = \frac{1}{m}\sum_{i=1}^{m}\tilde{\boldsymbol{x}}^{(i)}\tilde{\boldsymbol{x}}^{(i)\top}$$

By first observing

$$\begin{aligned} \frac{1}{m}\sum_{i=1}^{m}\tilde{\boldsymbol{x}}^{(i)}\mathbb{E}[\boldsymbol{z}^{(i)}|\tilde{\boldsymbol{x}}^{(i)}]^\top &= \boldsymbol{S}\boldsymbol{B}_t^\top \\ \frac{1}{m}\sum_{j=1}^{m}\mathbb{E}[\boldsymbol{z}^{(j)}\boldsymbol{z}^{(j)\top}|\tilde{\boldsymbol{x}}^{(j)}] &= \boldsymbol{M}_t + \boldsymbol{B}_t\boldsymbol{S}\boldsymbol{B}_t^\top, \end{aligned}$$

update the parameters as follows.

$$\begin{aligned} \boldsymbol{W}_{t+1} &= \boldsymbol{S}\boldsymbol{B}_t^\top\left(\mathbb{I} - \boldsymbol{B}_t\boldsymbol{W}_t + \boldsymbol{B}_t\boldsymbol{S}\boldsymbol{B}_t^\top\right)^{-1} \\ &= \boldsymbol{S}\boldsymbol{B}_t^\top\left(\boldsymbol{M}_t + \boldsymbol{B}_t\boldsymbol{S}\boldsymbol{B}_t^\top\right)^{-1} \\ \widetilde{\boldsymbol{\Psi}}_{t+1} &= \frac{1}{m}\left(\sum_{i=1}^{m}\tilde{\boldsymbol{x}}^{(i)}\tilde{\boldsymbol{x}}^{(i)\top} - \boldsymbol{W}_{t+1}\mathbb{E}[\boldsymbol{z}^{(i)}|\tilde{\boldsymbol{x}}^{(i)}]\tilde{\boldsymbol{x}}^{(i)\top}\right) \\ &= \boldsymbol{S} - \frac{1}{m}\sum_{i=1}^{m}\boldsymbol{W}_{t+1}\boldsymbol{B}_t\tilde{\boldsymbol{x}}^{(i)}\tilde{\boldsymbol{x}}^{(i)\top} \\ &= \boldsymbol{S} - \boldsymbol{W}_{t+1}\boldsymbol{B}_t\boldsymbol{S} \\ &= \boldsymbol{S} - \boldsymbol{S}\boldsymbol{B}_t^\top\boldsymbol{W}_{t+1}^\top \end{aligned}$$

Imposing the block diagonal constraint,

$$\boldsymbol{\Psi}_{t+1} = \text{Bdiag}\left((\widetilde{\boldsymbol{\Psi}}_{t+1})_{11}, \ldots, (\widetilde{\boldsymbol{\Psi}}_{t+1})_{vv}\right)$$

where $(\tilde{\boldsymbol{\Psi}})_{ii} = \boldsymbol{\Psi}_i$.

F Independence to Noise in Direct Methods

We are maximising the following quantity with respect to $\boldsymbol{\theta} = \{\boldsymbol{W}, \boldsymbol{\mu}, \boldsymbol{\Psi}\}$

$$\begin{aligned} p(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\theta}) &= \prod_i p(\boldsymbol{y}^{(i)}|\boldsymbol{x}^{(i)}, \boldsymbol{\theta}) \\ &= \prod_i \mathcal{N}(\boldsymbol{y}^{(i)}; \boldsymbol{W}\boldsymbol{x}^{(i)} + \boldsymbol{\mu}, \boldsymbol{\Psi}) \\ \log p(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\theta}) &= -\frac{1}{2}\left(\sum_i \|\boldsymbol{y}^{(i)} - \boldsymbol{W}\boldsymbol{x}^{(i)}\|_{\boldsymbol{\Psi}}^2 - C\right) \\ \frac{\partial \log p(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\theta})}{\partial \boldsymbol{W}} &\propto \left(\sum_i \boldsymbol{\Psi}^{-1}(\boldsymbol{y}^{(i)} - \boldsymbol{W}\boldsymbol{x}^{(i)})\boldsymbol{x}^{(i)\top}\right) \\ &\propto \boldsymbol{\Psi}^{-1}\left(\sum_i \boldsymbol{y}^{(i)}\boldsymbol{x}^{(i)\top} - \boldsymbol{W}\sum_i \boldsymbol{x}^{(i)}\boldsymbol{x}^{(i)\top}\right) \end{aligned}$$

The maximum likelihood is achieved when

$$\frac{\partial \log p(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\theta})}{\partial \boldsymbol{W}} = \boldsymbol{0},$$

and since $\boldsymbol{\Psi}^{-1}$ has an inverse (namely $\boldsymbol{\Psi}$), this means that

$$\boldsymbol{W}\sum_i \boldsymbol{x}^{(i)}\boldsymbol{x}^{(i)\top} = \sum_i \boldsymbol{y}^{(i)}\boldsymbol{x}^{(i)\top}$$

It is clear from here that the MLE of $\boldsymbol{W}$ does not depend on $\boldsymbol{\Psi}$, thus we can conclude that adding a noise parameter to this directed linear model has no effect on its predictions.

Table 1: Precision @1 between MBFA fitted for 1K iterations and MBFA fitted for 20K iterations.

Method	EN-IT	IT-EN	EN-FR	FR-EN	IT-FR	FR-IT
MBFA-1K	71.9	73.3	76.7	78.2	82.4	77.5
MBFA-20K	71.9	73.4	76.7	78.1	82.6	77.5
MBFA-1K+CSLS	77.5	77.6	81.9	82.0	86.8	82.1
MBFA-20K+CSLS	77.4	77.7	81.9	82.1	86.8	81.9

G Learning curve of EM

Figure 1 shows the negative log-likelihood of the three language model over the first 5,000 iterations. The precision of the learned model is very close when evaluated at iteration 1,000 and at iteration 20,000 as seen in Table 1. This suggests that the model need not be trained to full convergence to work well.

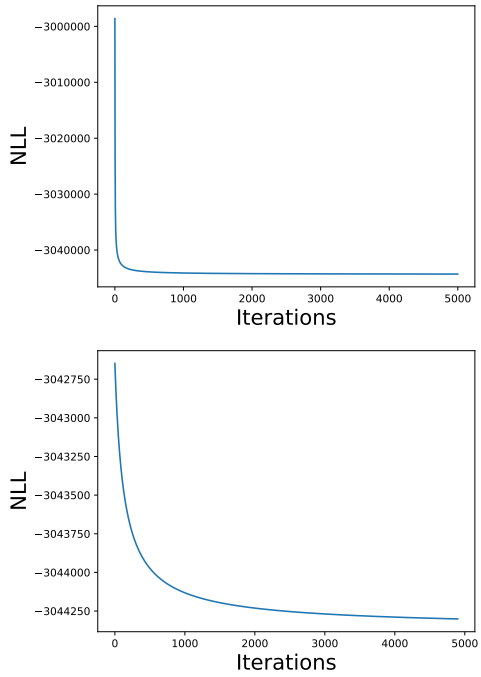

Figure 1: Training curve of EM algorithm over the first 5,000 iterations. It is clear that the procedure quickly finds a good approximation to the optimal parameters and then slowly converges to the real optimum. Top picture shows the entire training curve, while the bottom picture starts from iteration 100.

References

Zoubin Ghahramani, Geoffrey E Hinton, et al. 1996. The em algorithm for mixtures of factor analyzers. Technical report, Technical Report CRG-TR-96-1, University of Toronto.

Kaare Brandt Petersen, Michael Syskind Pedersen, et al. 2008. The matrix cookbook. *Technical University of Denmark*, 7(15):510.