SUPPLEMENTARY TO: SCALABLE GROUP LEVEL PROBABILISTIC SPARSE FACTOR ANALYSIS

Jesper L. Hinrich[⋆]*, Søren F. V. Nielsen*[⋆]*, Nicolai A. B. Riis*[⋆]*, Casper T. Eriksen*[⋆]*, Jacob Frøsig*[⋆]*, Marco D. F. Kristensen*[⋆]*, Mikkel N. Schmidt*[⋆]*, Kristoffer H. Madsen*[⋆,†] *and Morten Mørup*[⋆]

[⋆] DTU Compute, Technical University of Denmark, Denmark
[†] Danish Research Centre for Magnetic Resonance, Copenhagen University Hospital Hvidovre, Denmark

ABSTRACT

The derived details for probabilistic sparse factor analysis model for group level analysis are given. The model was first proposed by the authors and applied to functional magnetic resonance imaging[1].

A. PROBABILISTIC SPARSE FACTOR ANALYSIS

The model's likelihood function, for a data array of size $V \times T \times B$ with all relevant parameters collected in $\boldsymbol{\theta}$, can be written as,

$$\mathcal{L}\left(\mathbf{X}|\boldsymbol{\theta}\right) = \prod_{b=1}^{B}\prod_{t=1}^{T}\mathcal{N}\left(\mathbf{x}_t^{(b)}|\mathbf{A}\mathbf{s}_t^{(b)} + \boldsymbol{\mu}^{(b)}, \operatorname{diag}\left(\boldsymbol{\tau}^{(b)}\right)^{-1}\right), \tag{1}$$

in which $\mathbf{x}_t^{(b)}$ is a vector of length V, $\mathbf{A}$ is a matrix of size $V \times D$, where D is the size of the latent space, $\mathbf{s}_t^{(b)}$ is a vector of size D and $\boldsymbol{\tau}^{(b)}$ is a vector of length V. The distributions of the parameters in the model are,

$$\begin{aligned}
P(\mathbf{A}|\boldsymbol{\alpha}) &= \prod_{v=1}^{V}\mathcal{N}\left(\mathbf{a}_v|\mathbf{0}, \operatorname{diag}\left(\boldsymbol{\alpha}_v\right)^{-1}\right)\\
P(\mathbf{S}|\boldsymbol{\gamma}) &= \prod_{b=1}^{B}\prod_{t=1}^{T}\mathcal{N}\left(\mathbf{s}_t^{(b)}|0, \operatorname{diag}(\boldsymbol{\gamma})^{-1}\right)\\
P(\boldsymbol{\mu}) &= \prod_{b=1}^{B}\mathcal{N}\left(\boldsymbol{\mu}^{(b)}|\mathbf{0}, \beta^{-1}\mathbf{I}_V\right)\\
P(\boldsymbol{\tau}) &= \prod_{b=1}^{B}\prod_{v=1}^{V}\mathcal{G}\left(\tau_v^{(b)}|a_\tau, b_{\tau_v^{(b)}}\right)\\
P(\boldsymbol{\gamma}) &= \prod_{d=1}^{D}\mathcal{G}\left(\gamma_d|a_\gamma, b_{\gamma_d}\right)\\
P(\boldsymbol{\alpha}) &= \prod_{v=1}^{V}\prod_{d=1}^{D}\mathcal{G}\left(\alpha_{vd}|a_\alpha, b_{\alpha_{vd}}\right).
\end{aligned}$$

Corresponding author: sfvn@dtu.dk. This work was supported by the Lundbeck Foundation, grant no. R105-9813. The Tesla K40 GPU used was donated by the NVIDIA Corporation.

Finding the parameters $\boldsymbol{\theta}$ from observed data $\mathbf{X}$ can be done by inferring the posterior distribution,

$$P(\boldsymbol{\theta}|\mathbf{X}) = \mathcal{L}(\mathbf{X}|\boldsymbol{\theta})P(\boldsymbol{\theta})/P(\mathbf{X}).$$

Unfortunately an exact inference is unfeasible for all but the simplest problems. Therefore an approximate solution is sought. While there are numerous ways to tackle this, we use variational Bayesian (VB) inference as used by [2] for VB principal components analysis. We use a mean field approximation, and find the following variational distribution to approximate the posterior.

$$\begin{aligned}
Q(\mathbf{A}) &= \prod_{v=1}^{V}\mathcal{N}(\mathbf{a}_v|\boldsymbol{\mu}_{\mathbf{A}_v}, \boldsymbol{\Sigma}_{\mathbf{A}}^{(v)})\\
Q(\mathbf{S}) &= \prod_{b,t=1}^{B,T}\mathcal{N}(\mathbf{s}_t^{(b)}|\boldsymbol{\mu}_{\mathbf{S}_t}^{(b)}, \boldsymbol{\Sigma}_{\mathbf{S}}^{(b)})\\
Q(\boldsymbol{\mu}) &= \prod_{b=1}^{B}\mathcal{N}\left(\boldsymbol{\mu}^{(b)}|\boldsymbol{\mu}_{\boldsymbol{\mu}}^{(b)}, \boldsymbol{\Sigma}_{\boldsymbol{\mu}}^{(b)}\right)\\
Q(\boldsymbol{\tau}) &= \prod_{b,v=1}^{B,V}\mathcal{G}(\tau_v^{(b)}|\tilde{a}_\tau, \tilde{b}_{\tau_v^{(b)}})\\
Q(\boldsymbol{\gamma}) &= \prod_{d=1}^{D}\mathcal{G}(\gamma_d|\tilde{a}_\gamma, \tilde{b}_{\gamma_d})\\
Q(\boldsymbol{\alpha}) &= \prod_{d,v=1}^{D,v}\mathcal{G}(\alpha_{vd}|\tilde{a}_\alpha, \tilde{b}_{\alpha_{vd}})
\end{aligned}$$

The moments of a distribution are then found by conditioning them on all other distributions and using free-form optimization (see [2]).

B. UPDATE RULES

The found moments of the distributions are updated cyclically using Expectation-Maximization. In each iteration, after

all distributions have been updated, the evidence lowerbound (ELBO) is calculated. After a number of iterations, when the relative change in ELBO is below a given threshold a set of local optimal parameters is identified.

$$
\begin{aligned}
\boldsymbol{\Sigma}_{\mathbf{A}}^{v} &= \left(\text{diag} \left\langle \boldsymbol{\alpha}_v \right\rangle + \sum_{b=1}^{B} \left\langle \tau_v^{(b)} \right\rangle \left\langle \mathbf{S}^{(b)} \mathbf{S}^{(b)^\top} \right\rangle \right)^{-1} \\
\boldsymbol{\mu}_{\mathbf{A}}^{v} &= \boldsymbol{\Sigma}_{\mathbf{A}}^{v} \left(\sum_{b=1}^{B} \left\langle \tau_v^{(b)} \right\rangle \sum_{t=1}^{T} \left\langle \mathbf{s}_t^{(b)} \right\rangle \left(x_{tv}^{(b)} - \left\langle \boldsymbol{\mu}_v^{(b)} \right\rangle \right) \right) \\
\boldsymbol{\Sigma}_{\mathbf{S}}^{(b)} &= \left(\text{diag} \left\langle \boldsymbol{\gamma} \right\rangle + \left\langle \mathbf{A}^\top \text{diag}(\boldsymbol{\tau}^{(b)}) \mathbf{A} \right\rangle \right)^{-1} \\
\boldsymbol{\mu}_{\mathbf{S}_t}^{(b)} &= \boldsymbol{\Sigma}_{\mathbf{S}}^{(b)} \left\langle \mathbf{A}^\top \right\rangle \text{diag} \left\langle \boldsymbol{\tau}^{(b)} \right\rangle \left(\mathbf{x}_t^{(b)} - \left\langle \boldsymbol{\mu}^{(b)} \right\rangle \right) \\
\boldsymbol{\Sigma}_{\boldsymbol{\mu}}^{(b)} &= \left(\beta \mathbf{I}_V + \text{diag} \left\langle \boldsymbol{\tau}^{(b)} \right\rangle \right)^{-1} \\
\boldsymbol{\mu}_{\boldsymbol{\mu}}^{(b)} &= \boldsymbol{\Sigma}_{\boldsymbol{\mu}}^{(b)} \text{diag} \left\langle \boldsymbol{\tau}^{(b)} \right\rangle \sum_{t=1}^{T} \left(\mathbf{x}_t^{(b)} - \left\langle \mathbf{A} \right\rangle \left\langle \mathbf{s}_t^{(b)} \right\rangle \right) \\
\tilde{a}_\alpha &= a_\alpha + \frac{1}{2} \quad , \quad \tilde{b}_{\alpha_{vd}} = b_{\alpha_{vd}} + \left\langle a_{vd}^2 \right\rangle \\
\tilde{a}_\gamma &= a_\gamma + \frac{1}{2} \sum_{b=1}^{B} \mathrm{T}^{(\mathrm{b})} \quad , \quad \tilde{b}_{\gamma_d} = b_{\gamma_d} + \frac{1}{2} \sum_{b=1}^{B} \text{trace} \left(\left\langle \mathbf{s}_d \mathbf{s}_d^\top \right\rangle \right) \\
\tilde{a}_{\tau^{(b)}} &= a_\tau + \frac{T^{(b)}}{2} \\
\tilde{b}_{\tau_v^{(b)}} &= b_{\tau_v^{(b)}} + \frac{1}{2} \Big[||\mathbf{x}_v^{(b)}||_{\text{Fro}}^2 + T^{(b)} \left\langle \mu_v^{(b)^2} \right\rangle \\
&\quad - 2 \left(\left\langle \mathbf{a}_v \right\rangle \left\langle \mathbf{S}^{(b)} \right\rangle + \left\langle \mu_v^{(b)} \right\rangle \right) \mathbf{x}_v^\top \\
&\quad + 2 \left\langle \mathbf{a}_v \right\rangle \left\langle \mathbf{S}^{(b)} \right\rangle \mathbf{1}_{T^{(b)}} \left\langle \mu_v^{(b)} \right\rangle \\
&\quad + \text{trace} \left(\left\langle \mathbf{a}_v^\top \mathbf{S}^{(b)} \mathbf{S}^{(b)^\top} \mathbf{a}_v \right\rangle \right) \Big]
\end{aligned}
$$

Note $\langle \cdot \rangle$ is the expected value under the variational distributions. Further, using the properties of the trace operator, the expected value of the expression in $\boldsymbol{\Sigma}_{\mathbf{S}}^{(b)}$ and $\tilde{b}_{\tau_v^{(b)}}$ are determined to be,

$$
\begin{aligned}
\text{trace} \left(\left\langle \mathbf{a}_v^\top \mathbf{S}^{(b)} \mathbf{S}^{(b)^\top} \mathbf{a}_v \right\rangle \right) &= \text{trace} \left(\left\langle \mathbf{S}^{(b)} \mathbf{S}^{(b)^\top} \right\rangle \boldsymbol{\Sigma}_{\mathbf{A}}^{v} \right) \\
&\quad + \left\langle \mathbf{a}_v^\top \right\rangle \left\langle \mathbf{S}^{(b)} \mathbf{S}^{(b)^\top} \right\rangle \left\langle \mathbf{a}_v \right\rangle
\end{aligned}
$$

and

$$
\begin{aligned}
\left\langle \mathbf{A}^\top \text{diag}(\boldsymbol{\tau}^{(b)}) \mathbf{A} \right\rangle &= \left(\sum_{v=1}^{V} \boldsymbol{\Sigma}_A^v \left\langle \tau_v^{(b)} \right\rangle \right) \\
&\quad + \left\langle \mathbf{A}^\top \right\rangle \text{diag} \left\langle \boldsymbol{\tau}^{(b)} \right\rangle \left\langle \mathbf{A} \right\rangle .
\end{aligned}
$$

C. EVIDENCE LOWERBOUND (ELBO)

The evidence lowerbound is the sum of all the expression in this section. It can be divided into to categories; 1) the expected value of the P-distributions under the Q-distributions (i.e. substituting the moments of the P-distributions for the moments of the Q-distributions). 2) The entropy of the Q-distributions.

For each P-distribution the expected value the corresponding Q-distribution is given below,

$$
\begin{aligned}
\left\langle \log P(\mathbf{A}|\boldsymbol{\alpha}) \right\rangle &= \sum_{v=1}^{V} -\frac{1}{2} \log(2\pi) + \frac{1}{2} \left\langle \log \alpha_{vd} \right\rangle - \frac{1}{2} \left\langle \alpha_{vd} \right\rangle \left\langle a_{vd}^2 \right\rangle \\
\left\langle \log P(\mathbf{S}|\boldsymbol{\gamma}) \right\rangle &= \sum_{b=1}^{B} \sum_{t=1}^{T^{(b)}} -\frac{D}{2} \log(2\pi) + \left(\frac{1}{2} \sum_{d=1}^{D} \left\langle \log \gamma_d \right\rangle \right) \\
&\quad - \frac{1}{2} \text{trace} \left(\text{diag} \left\langle \boldsymbol{\gamma} \right\rangle \left\langle \mathbf{s}_t^{(b)} \mathbf{s}_t^{(b)^\top} \right\rangle \right) \\
\left\langle \log P(\boldsymbol{\mu}) \right\rangle &= \sum_{b=1}^{B} -\frac{V}{2} \log(2\pi) + \frac{V}{2} \log(\beta) - \frac{1}{2} \beta \left\langle \boldsymbol{\mu}_{\boldsymbol{\mu}}^{(b)^\top} \boldsymbol{\mu}_{\boldsymbol{\mu}}^{(b)} \right\rangle \\
\left\langle \log P(\boldsymbol{\alpha}) \right\rangle &= \sum_{d=1}^{D} \sum_{v=1}^{V} -\log(\Gamma(a_\alpha)) + a_\alpha \log(b_{\alpha_{vd}}) \\
&\quad + (a_\alpha - 1) \left\langle \log \alpha_{vd} \right\rangle - b_{\alpha_{vd}} \left\langle \alpha_{vd} \right\rangle \\
\left\langle \log P(\boldsymbol{\gamma}) \right\rangle &= \sum_{d=1}^{D} -\log(\Gamma(a_\gamma)) + a_\gamma \log(b_{\gamma_d}) \\
&\quad + (a_\gamma - 1) \left\langle \log \gamma_d \right\rangle - b_{\gamma_d} \left\langle \gamma_d \right\rangle \\
\left\langle \log P(\boldsymbol{\tau}) \right\rangle &= \sum_{b=1}^{B} \sum_{v=1}^{V} -\log(\Gamma(a_\tau)) + a_\tau \log \left(b_{\tau_v^{(b)}} \right) \\
&\quad + (a_\tau - 1) \left\langle \log \tau_v^{(b)} \right\rangle - b_{\tau_v^{(b)}} \left\langle \tau_v^{(b)} \right\rangle
\end{aligned}
$$

Finally, the entropy for each Q-distribution is,

$$
\begin{aligned}
-\left\langle \log Q(\mathbf{A}) \right\rangle &= \sum_{v=1}^{V} \left[\frac{1}{2} \log |\boldsymbol{\Sigma}_{\mathbf{A}}^{v}| + \frac{D}{2} (1 + \log(2\pi)) \right] \\
-\left\langle \log Q(\mathbf{S}) \right\rangle &= \sum_{b=1}^{B} \sum_{t=1}^{T^{(b)}} \left[\frac{1}{2} \log |\boldsymbol{\Sigma}_{\mathbf{S}}^{(b)}| + \frac{D}{2} (1 + \log(2\pi)) \right] \\
-\left\langle \log Q(\boldsymbol{\mu}) \right\rangle &= \sum_{b=1}^{B} \left[\frac{1}{2} \log |\boldsymbol{\Sigma}_{\boldsymbol{\mu}}^{(b)}| + \frac{V}{2} (1 + \log(2\pi)) \right] \\
-\left\langle \log Q(\boldsymbol{\alpha}) \right\rangle &= \sum_{d=1}^{D} \sum_{v=1}^{V} \log(\Gamma(\tilde{a}_{\alpha_{vd}})) - (\tilde{a}_\alpha - 1) \psi(\tilde{a}_\alpha) \\
&\quad - \log(\tilde{b}_{\alpha_{vd}}) + \tilde{a}_\alpha \\
-\left\langle \log Q(\boldsymbol{\gamma}) \right\rangle &= \sum_{d=1}^{D} \log(\Gamma(\tilde{a}_\gamma)) - (\tilde{a}_\gamma - 1) \psi(\tilde{a}_\gamma) - \log(\tilde{b}_{\gamma_d}) + \tilde{a}_\gamma \\
-\left\langle \log Q(\boldsymbol{\tau}) \right\rangle &= \sum_{b=1}^{B} \sum_{v=1}^{V} \log(\Gamma(\tilde{a}_{\tau^{(b)}})) - (\tilde{a}_{\tau^{(b)}} - 1) \psi(\tilde{a}_\tau) \\
&\quad - \log(\tilde{b}_{\tau_v^{(b)}}) + \tilde{a}_{\tau^{(b)}}
\end{aligned}
$$

D. IMPLEMENTATION

A MATLAB implementation is available[1]. The implementation is limited to the use of a single GPU card, as well as analysis with $T^{(i)} = T^{(j)}, \forall i, j$. These limitations were deemed acceptable for the work in [1], as subjects with differing timesteps are not widespread in the field.

E. REFERENCES

[1] Jesper L. Hinrich, Søren F.V. Nielsen, Nicolai A. B. Riis, Casper T. Eriksen, Jacob Frøsig, Marco D. F. Kristensen, Mikkel N. Schmidt, Kristoffer Hougaard Madsen, and Morten Mørup. Scalable group level probabilistic sparse factor analysis. In *(in review)*.

[2] C M Bishop. Variational principal components. In *Proceedings of the 1999 the 9th International Conference on 'Artificial Neural Networks (ICANN99)*, pages 509–514. IEEE, 1999.

[1] `https://....`