These are exercises for self-study and exam preparation. All material is examinable unless otherwise mentioned.

Exercise 1. *Factor analysis*

A friend proposes to improve the factor analysis model by working with correlated latent variables. The proposed model is

$$p(\mathbf{h}; \mathbf{C}) = \mathcal{N}(\mathbf{h}; \mathbf{0}, \mathbf{C}) \qquad p(\mathbf{v}|\mathbf{h}; \mathbf{F}, \boldsymbol{\Psi}, \mathbf{c}) = \mathcal{N}(\mathbf{v}; \mathbf{F}\mathbf{h} + \mathbf{c}, \boldsymbol{\Psi}) \tag{1}$$

where $\mathbf{C}$ *is some covariance matrix, and the other variables are defined as in the lecture slides.* $\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ *denotes the pdf of a Gaussian with mean* $\boldsymbol{\mu}$ *and covariance matrix* $\boldsymbol{\Sigma}$*.*

(a) What is marginal distribution of the visibles $p(\mathbf{v}; \boldsymbol{\theta})$ *where* $\boldsymbol{\theta}$ *stands for the parameters* $\mathbf{C}, \mathbf{F}, \mathbf{c}, \boldsymbol{\Psi}$*?*

Solution. The model specifications are equivalent to the following data generating process:

$$\mathbf{h} \sim \mathcal{N}(\mathbf{h}; \mathbf{0}, \mathbf{C}) \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{\epsilon}; \mathbf{0}, \boldsymbol{\Psi}) \qquad \mathbf{v} = \mathbf{F}\mathbf{h} + \mathbf{c} + \boldsymbol{\epsilon} \tag{S.1}$$

Recall the basic result on the distribution of linear transformations of Gaussians: if $\mathbf{x}$ has density $\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_x, \mathbf{C}_x)$, $\mathbf{z}$ density $\mathcal{N}(\mathbf{z}; \boldsymbol{\mu}_z, \mathbf{C}_z)$, and $\mathbf{x} \perp\!\!\!\perp \mathbf{z}$ then $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{z}$ has density

$$\mathcal{N}(\mathbf{y}; \mathbf{A}\boldsymbol{\mu}_x + \boldsymbol{\mu}_z, \mathbf{A}\mathbf{C}_x\mathbf{A}^\top + \mathbf{C}_z).$$

It thus follows that $\mathbf{v}$ is Gaussian with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$,

$$\boldsymbol{\mu} = \mathbf{F}\underbrace{\mathbb{E}[\mathbf{h}]}_{\mathbf{0}} + \mathbf{c} + \underbrace{\mathbb{E}[\boldsymbol{\epsilon}]}_{\mathbf{0}} \tag{S.2}$$

$$= \mathbf{c} \tag{S.3}$$

$$\boldsymbol{\Sigma} = \mathbf{F}\mathbb{V}[\mathbf{h}]\mathbf{F}^\top + \mathbb{V}[\boldsymbol{\epsilon}] \tag{S.4}$$

$$= \mathbf{F}\mathbf{C}\mathbf{F}^\top + \boldsymbol{\Psi}. \tag{S.5}$$

(b) Assume that the singular value decomposition of $\mathbf{C}$ *is given by*

$$\mathbf{C} = \mathbf{E}\boldsymbol{\Lambda}\mathbf{E}^\top \tag{2}$$

where $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_D)$ *is a diagonal matrix containing the eigenvalues, and* $\mathbf{E}$ *is a orthonormal matrix containing the corresponding eigenvectors. The matrix square root of* $\mathbf{C}$ *is the matrix* $\mathbf{M}$ *such that*

$$\mathbf{M}\mathbf{M} = \mathbf{C}, \tag{3}$$

and we denote it by $\mathbf{C}^{1/2}$*. Show that the matrix square root of* $\mathbf{C}$ *equals*

$$\mathbf{C}^{1/2} = \mathbf{E}\,\mathrm{diag}(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_D})\mathbf{E}^\top. \tag{4}$$

Solution. We verify that $\mathbf{C}^{1/2}\mathbf{C}^{1/2} = \mathbf{C}$:

$$\begin{aligned}
\mathbf{C}^{1/2}\mathbf{C}^{1/2} &= \mathbf{E}\,\mathrm{diag}(\sqrt{\lambda_1},\ldots,\sqrt{\lambda_D})\mathbf{E}^\top\mathbf{E}\,\mathrm{diag}(\sqrt{\lambda_1},\ldots,\sqrt{\lambda_D})\mathbf{E}^\top && \text{(S.6)}\\
&= \mathbf{E}\,\mathrm{diag}(\sqrt{\lambda_1},\ldots,\sqrt{\lambda_D})\;\mathbf{I}\;\mathrm{diag}(\sqrt{\lambda_1},\ldots,\sqrt{\lambda_D})\mathbf{E}^\top && \text{(S.7)}\\
&= \mathbf{E}\,\mathrm{diag}(\sqrt{\lambda_1},\ldots,\sqrt{\lambda_D})\,\mathrm{diag}(\sqrt{\lambda_1},\ldots,\sqrt{\lambda_D})\mathbf{E}^\top && \text{(S.8)}\\
&= \mathbf{E}\,\mathrm{diag}(\lambda_1,\ldots,\lambda_D)\mathbf{E}^\top && \text{(S.9)}\\
&= \mathbf{E}\boldsymbol{\Lambda}\mathbf{E}^\top && \text{(S.10)}\\
&= \mathbf{C} && \text{(S.11)}
\end{aligned}$$

(c) Show that the proposed factor analysis model is equivalent to the original factor analysis model

$$p(\mathbf{h};\mathbf{I}) = \mathcal{N}(\mathbf{h};\mathbf{0},\mathbf{I}) \qquad p(\mathbf{v}|\mathbf{h};\tilde{\mathbf{F}},\boldsymbol{\Psi},\mathbf{c}) = \mathcal{N}(\mathbf{v};\tilde{\mathbf{F}}\mathbf{h}+\mathbf{c},\boldsymbol{\Psi}) \tag{5}$$

with $\tilde{\mathbf{F}} = \mathbf{F}\mathbf{C}^{1/2}$, so that the extra parameters given by the covariance matrix $\mathbf{C}$ are actually redundant and nothing is gained with the richer parametrisation.

Solution. We verify that the model has the same distribution for the visibles. As before $\mathbb{E}[\mathbf{v}] = \mathbf{c}$, and the covariance matrix is

$$\begin{aligned}
\mathbb{V}[\mathbf{v}] &= \tilde{\mathbf{F}}\mathbf{I}\tilde{\mathbf{F}}^\top + \boldsymbol{\Psi} && \text{(S.12)}\\
&= \mathbf{F}\mathbf{C}^{1/2}\mathbf{C}^{1/2}\mathbf{F}^\top + \boldsymbol{\Psi} && \text{(S.13)}\\
&= \mathbf{F}\mathbf{C}\mathbf{F}^\top + \boldsymbol{\Psi} && \text{(S.14)}
\end{aligned}$$

where we have used that $\mathbf{C}^{1/2}$ is a symmetric matrix. This means that the correlation between the $\mathbf{h}$ can be absorbed into the factor matrix $\mathbf{F}$ and the set of pdfs defined by the proposed model equals the set of pdfs of the original factor analysis model.

Another way to see the result is to consider the data generating process and noting that we can sample $\mathbf{h}$ from $\mathcal{N}(\mathbf{h};\mathbf{0},\mathbf{C})$ by first sampling $\mathbf{h}'$ from $\mathcal{N}(\mathbf{h}';\mathbf{0},\mathbf{I})$ and then transforming the sample by $\mathbf{C}^{1/2}$,

$$\mathbf{h} \sim \mathcal{N}(\mathbf{h};\mathbf{0},\mathbf{C}) \qquad \Longleftrightarrow \qquad \mathbf{h} = \mathbf{C}^{1/2}\mathbf{h}' \qquad \mathbf{h}' \sim \mathcal{N}(\mathbf{h}';\mathbf{0},\mathbf{I}). \tag{S.15}$$

This follows again from the basic properties of linear transformations of Gaussians, i.e.

$$\mathbb{V}(\mathbf{C}^{1/2}\mathbf{h}') = \mathbf{C}^{1/2}\mathbb{V}(\mathbf{h}')(\mathbf{C}^{1/2})^\top = \mathbf{C}^{1/2}\mathbf{I}\mathbf{C}^{1/2} = \mathbf{C}$$

and $\mathbb{E}(\mathbf{C}^{1/2}\mathbf{h}') = \mathbf{C}^{1/2}\mathbb{E}(\mathbf{h}') = \mathbf{0}$.

To generate samples from the proposed factor analysis model, we would thus proceed as follows:

$$\mathbf{h}' \sim \mathcal{N}(\mathbf{h}';\mathbf{0},\mathbf{I}) \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{\epsilon};\mathbf{0},\boldsymbol{\Psi}) \qquad \mathbf{v} = \mathbf{F}(\mathbf{C}^{1/2}\mathbf{h}') + \mathbf{c} + \boldsymbol{\epsilon} \tag{S.16}$$

But the term

$$\mathbf{v} = \mathbf{F}(\mathbf{C}^{1/2}\mathbf{h}') + \mathbf{c} + \boldsymbol{\epsilon}$$

can be written as

$$\mathbf{v} = (\mathbf{F}\mathbf{C}^{1/2})\mathbf{h}' + \mathbf{c} + \boldsymbol{\epsilon} = \tilde{\mathbf{F}}\mathbf{h}' + \mathbf{c} + \boldsymbol{\epsilon}$$

and since $\mathbf{h}'$ follows $\mathcal{N}(\mathbf{h}';\mathbf{0},\mathbf{I})$, we are back at the original factor analysis model.

Exercise 2. ***Independent component analysis***

(a) Whitening corresponds to linearly transforming a random variable $\mathbf{x}$ *(or the corresponding data) so that the resulting random variable* $\mathbf{z}$ *has an identity covariance matrix, i.e.*

$$\mathbf{z} = \mathbf{V}\mathbf{x} \quad \textit{with} \quad \mathbb{V}[\mathbf{x}] = \mathbf{C} \quad \textit{and} \quad \mathbb{V}[\mathbf{z}] = \mathbf{I}.$$

The matrix $\mathbf{V}$ *is called the whitening matrix. We do not make a distributional assumption on* $\mathbf{x}$*, in particular* $\mathbf{x}$ *may or may not be Gaussian.*
Given the eigenvalue decomposition $\mathbf{C} = \mathbf{E}\mathbf{\Lambda}\mathbf{E}^\top$*, show that*

$$\mathbf{V} = \operatorname{diag}(\lambda_1^{-1/2}, \ldots, \lambda_d^{-1/2})\mathbf{E}^\top \tag{6}$$

is a whitening matrix.

Solution. From $\mathbb{V}[\mathbf{z}] = \mathbb{V}[\mathbf{V}\mathbf{x}] = \mathbf{V}\mathbb{V}[\mathbf{x}]\mathbf{V}^\top$, it follows that

$$\begin{aligned}
\mathbb{V}[\mathbf{z}] &= \mathbf{V}\mathbb{V}[\mathbf{x}]\mathbf{V}^\top && \text{(S.17)}\\
&= \mathbf{V}\mathbf{C}\mathbf{V}^\top && \text{(S.18)}\\
&= \mathbf{V}\mathbf{E}\mathbf{\Lambda}\mathbf{E}^\top\mathbf{V}^\top && \text{(S.19)}\\
&= \operatorname{diag}(\lambda_1^{-1/2}, \ldots, \lambda_d^{-1/2})\mathbf{E}^\top\mathbf{E}\mathbf{\Lambda}\mathbf{E}^\top\mathbf{V}^\top && \text{(S.20)}\\
&= \operatorname{diag}(\lambda_1^{-1/2}, \ldots, \lambda_d^{-1/2})\mathbf{\Lambda}\mathbf{E}^\top\mathbf{V}^\top && \text{(S.21)}
\end{aligned}$$

where we have used that $\mathbf{E}^\top\mathbf{E} = \mathbf{I}$. Since

$$\mathbf{V}^\top = \left[\operatorname{diag}(\lambda_1^{-1/2}, \ldots, \lambda_d^{-1/2})\mathbf{E}^\top\right]^\top = \mathbf{E}\operatorname{diag}(\lambda_1^{-1/2}, \ldots, \lambda_d^{-1/2})$$

we further have

$$\begin{aligned}
\mathbb{V}[\mathbf{z}] &= \operatorname{diag}(\lambda_1^{-1/2}, \ldots, \lambda_d^{-1/2})\mathbf{\Lambda}\mathbf{E}^\top\mathbf{E}\operatorname{diag}(\lambda_1^{-1/2}, \ldots, \lambda_d^{-1/2}) && \text{(S.22)}\\
&= \operatorname{diag}(\lambda_1^{-1/2}, \ldots, \lambda_d^{-1/2})\mathbf{\Lambda}\operatorname{diag}(\lambda_1^{-1/2}, \ldots, \lambda_d^{-1/2}) && \text{(S.23)}\\
&= \operatorname{diag}(\lambda_1^{-1/2}, \ldots, \lambda_d^{-1/2})\operatorname{diag}(\lambda_1, \ldots, \lambda_d)\operatorname{diag}(\lambda_1^{-1/2}, \ldots, \lambda_d^{-1/2}) && \text{(S.24)}\\
&= \mathbf{I}, && \text{(S.25)}
\end{aligned}$$

so that $\mathbf{V}$ is indeed a valid whitening matrix. Note that whitening matrices are not unique. For example,

$$\tilde{\mathbf{V}} = \mathbf{E}\operatorname{diag}(\lambda_1^{-1/2}, \ldots, \lambda_d^{-1/2})\mathbf{E}^\top$$

is also a valid whitening matrix. More generally, if $\mathbf{V}$ is a whitening matrix, then $\mathbf{R}\mathbf{V}$ is also a whitening matrix when $\mathbf{R}$ is an orthonormal matrix. This is because

$$\mathbb{V}[\mathbf{R}\mathbf{V}\mathbf{x}] = \mathbf{R}\mathbb{V}[\mathbf{V}\mathbf{x}]\mathbf{R}^\top = \mathbf{R}\mathbf{I}\mathbf{R}^\top = \mathbf{I}$$

where we have used that $\mathbf{V}$ is a whitening matrix so that $\mathbf{V}\mathbf{x}$ has identity covariance matrix.

(b) Consider the ICA model

$$\mathbf{v} = \mathbf{A}\mathbf{h}, \qquad \mathbf{h} \sim p_{\mathbf{h}}(\mathbf{h}), \qquad p_{\mathbf{h}}(\mathbf{h}) = \prod_{i=1}^{D} p_h(h_i), \tag{7}$$

where the matrix $\mathbf{A}$ *is invertible and the* h_i *are independent random variables of mean zero and variance one. Let* $\mathbf{V}$ *be a whitening matrix for* $\mathbf{v}$*. Show that* $\mathbf{z} = \mathbf{V}\mathbf{v}$ *follows the ICA model*

$$\mathbf{z} = \tilde{\mathbf{A}}\mathbf{h}, \qquad \mathbf{h} \sim p_{\mathbf{h}}(\mathbf{h}), \qquad p_{\mathbf{h}}(\mathbf{h}) = \prod_{i=1}^{D} p_h(h_i), \tag{8}$$

where $\tilde{\mathbf{A}}$ *is an orthonormal matrix.*

Solution. If $\mathbf{v}$ follows the ICA model, we have

$$\begin{aligned}\mathbf{z} &= \mathbf{V}\mathbf{v} && \text{(S.26)}\\ &= \mathbf{V}\mathbf{A}\mathbf{h} && \text{(S.27)}\\ &= \tilde{\mathbf{A}}\mathbf{h} && \text{(S.28)}\end{aligned}$$

with $\tilde{\mathbf{A}} = \mathbf{V}\mathbf{A}$. By the whitening operation, the covariance matrix of $\mathbf{z}$ is identity, so that

$$\mathbf{I} = \mathbb{V}(\mathbf{z}) = \tilde{\mathbf{A}}\mathbb{V}(\mathbf{h})\tilde{\mathbf{A}}^\top. \tag{S.29}$$

By the ICA model, $\mathbb{V}(\mathbf{h}) = \mathbf{I}$, so that $\tilde{\mathbf{A}}$ must satisfy

$$\mathbf{I} = \tilde{\mathbf{A}}\tilde{\mathbf{A}}^\top, \tag{S.30}$$

which means that $\tilde{\mathbf{A}}$ is orthonormal.

In the original ICA model, the number of parameters is given by the number of elements of the matrix $\mathbf{A}$, which is D^2 if $\mathbf{v}$ is D-dimensional. An orthogonal matrix contains $D(D-1)/2$ degrees of freedom (see e.g. https://en.wikipedia.org/wiki/Orthogonal_matrix), so that we can think that whitening "solves half of the ICA problem". Since whitening is a relatively simple standard operation, many algorithms, e.g. "fastICA", first reduce the complexity of the estimation problem by whitening the data. Moreover, due to the properties of the orthogonal matrix, the log-likelihood for the ICA model also simplifies for whitened data: The log-likelihood for ICA model without whitening is

$$\ell(\mathbf{B}) = \sum_{i=1}^{n}\sum_{j=1}^{D} \log p_h(\mathbf{b}_j\mathbf{v}_i) + n\log|\det\mathbf{B}| \tag{S.31}$$

where $\mathbf{B} = \mathbf{A}^{-1}$. If we first whiten the data, the log-likelihood becomes

$$\ell(\tilde{\mathbf{B}}) = \sum_{i=1}^{n}\sum_{j=1}^{D} \log p_h(\tilde{\mathbf{b}}_j\mathbf{z}_i) + n\log|\det\tilde{\mathbf{B}}| \tag{S.32}$$

where $\tilde{\mathbf{B}} = \tilde{\mathbf{A}}^{-1} = \tilde{\mathbf{A}}^\top$ since $\tilde{\mathbf{A}}$ is an orthogonal matrix. This means $\tilde{\mathbf{B}}^{-1} = \tilde{\mathbf{A}} = \tilde{\mathbf{B}}^\top$ and $\tilde{\mathbf{B}}$ is an orthogonal matrix. Hence $\det\tilde{\mathbf{B}} = 1$, and the $\log\det$ term is zero. Hence, the log-likelihood on whitened data simplifies to

$$\ell(\tilde{\mathbf{B}}) = \sum_{i=1}^{n}\sum_{j=1}^{D} \log p_h(\tilde{\mathbf{b}}_j\mathbf{z}_i). \tag{S.33}$$

While the log-likelihood takes a simpler form, the optimisation problem is now a constrained optimisation problem: $\tilde{\mathbf{B}}$ is constrained to be orthonormal. For further information, see e.g. Chapter 9 of *Independent Component Analysis* by Hyvärinen, Karhunen, and Oja.