Exercise 1. *Factor analysis*

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

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

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

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

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

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

From the basic result on the distribution of linear transformations of Gaussians on FA and ICA lecture slide 11 (Barber Result 8.3), it follows that $\mathbf{v}$ is Gaussian with mean $\boldsymbol{\mu}$ and covariance $\mathbf{\Sigma}$,

$$\begin{aligned} \boldsymbol{\mu} &= \mathbf{F}\underbrace{\mathbb{E}[\mathbf{h}]}_{\mathbf{0}} + \mathbf{c} + \underbrace{\mathbb{E}[\boldsymbol{\epsilon}]}_{\mathbf{0}} && \text{(S.2)}\\ &= \mathbf{c} && \text{(S.3)}\\ \mathbf{\Sigma} &= \mathbf{F}\mathbb{V}[\mathbf{h}]\mathbf{F}^\top + \mathbb{V}[\boldsymbol{\epsilon}] && \text{(S.4)}\\ &= \mathbf{F}\mathbf{C}\mathbf{F}^\top + \mathbf{\Psi}. && \text{(S.5)} \end{aligned}$$

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

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

where $\mathbf{\Lambda} = diag(\lambda_1,\ldots,\lambda_H)$ *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*

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

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

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

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

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

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. Note we do not make a distributional assumption on* $\mathbf{x}$*, in particular* $\mathbf{x}$ *may or may not be Gaussian.*

Given the eigenvalue decomposition $\mathbf{C} = \mathbf{E}\boldsymbol{\Lambda}\mathbf{E}^\top$*, show that*

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

is a whitening matrix.

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

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

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

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

we further have

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

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

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

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

$$\mathbb{V}[\mathbf{RVx}] = \mathbf{R}\mathbb{V}[\mathbf{Vx}]\mathbf{R}^\top = \mathbf{RIR}^\top = \mathbf{I}$$

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

(b) *Consider the ICA model*

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

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 $\mathbf{z} = \mathbf{Vv}$ 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

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

with $\tilde{\mathbf{A}} = \mathbf{VA}$. 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, see e.g. the slides on "Factor and Independent Component Analysis", also simplifies for whitened data because the inverse of an orthonormal matrix is its transpose and because the determinant of an orthonormal matrix equals one.

Exercise 3. *Score matching for the exponential family*

In the lecture, we have derived the objective function $J(\boldsymbol{\theta})$ for score matching,

$$J(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m}\left[\partial_j\psi_j(\mathbf{x}_i;\boldsymbol{\theta}) + \frac{1}{2}\psi_j(\mathbf{x}_i;\boldsymbol{\theta})^2\right], \tag{9}$$

where ψ_j is the partial derivative of the log model-pdf $\log p(\mathbf{x};\boldsymbol{\theta})$ with respect to the j-th coordinate (slope) and $\partial_j\psi_j$ its second partial derivative (curvature). The observed data are denoted by $\mathbf{x}_1,\ldots,\mathbf{x}_n$ and $\mathbf{x}\in\mathbb{R}^m$.

The goal of this exercise is to show that for statistical models of the form

$$\log p(\mathbf{x};\boldsymbol{\theta}) = \sum_{k=1}^{K}\theta_k F_k(\mathbf{x}) - \log Z(\boldsymbol{\theta}), \qquad \mathbf{x}\in\mathbb{R}^m, \tag{10}$$

the score matching objective function becomes a quadratic form, which can be optimised efficiently (see e.g. Barber Appendix A.5.3).

The set of models above are called the (continuous) exponential family, or also log-linear models because the models are linear in the parameters θ_k. Since the exponential family generally includes probability mass functions as well, the qualifier "continuous" may be used to highlight that we are here considering continuous random variables only. The functions $F_k(\mathbf{x})$ are assumed to be known; they are the sufficient statistics (see e.g. Barber Section 8.5).

(a) *Denote by $\mathbf{K}(\mathbf{x})$ the matrix with elements $K_{kj}(\mathbf{x})$,*

$$K_{kj}(\mathbf{x}) = \frac{\partial F_k(\mathbf{x})}{\partial x_j}, \qquad k = 1\ldots K, \quad j = 1\ldots m, \tag{11}$$

and by $\mathbf{H}(\mathbf{x})$ the matrix with elements $H_{kj}(\mathbf{x})$,

$$H_{kj}(\mathbf{x}) = \frac{\partial^2 F_k(\mathbf{x})}{\partial x_j^2}, \qquad k = 1\ldots K, \quad j = 1\ldots m. \tag{12}$$

Furthermore, let $\mathbf{h}_j(\mathbf{x}) = (H_{1j}(\mathbf{x}), \ldots, H_{Kj}(\mathbf{x}))^\top$ *be the* j*–th column vector of* $\mathbf{H}(\mathbf{x})$.

Show that for the continuous exponential family, the score matching objective in Equation (9) becomes

$$J(\boldsymbol{\theta}) = \boldsymbol{\theta}^\top \mathbf{r} + \frac{1}{2}\boldsymbol{\theta}^\top \mathbf{M}\boldsymbol{\theta}, \tag{13}$$

where

$$\mathbf{r} = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m} \mathbf{h}_j(\mathbf{x}_i), \qquad \mathbf{M} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{K}(\mathbf{x}_i)\mathbf{K}(\mathbf{x}_i)^\top. \tag{14}$$

Solution. For

$$\log p(\mathbf{x};\boldsymbol{\theta}) = \sum_{k=1}^{K} \theta_k F_k(\mathbf{x}) - \log Z(\boldsymbol{\theta}) \tag{S.31}$$

the first derivative with respect to x_j, the j-th element of $\mathbf{x}$, is

$$\begin{aligned} \psi_j(\mathbf{x};\boldsymbol{\theta}) &= \frac{\partial \log p(\mathbf{x};\boldsymbol{\theta})}{\partial x_j} && \text{(S.32)}\\ &= \sum_{k=1}^{K} \theta_k \frac{\partial F_k(\mathbf{x})}{\partial x_j} && \text{(S.33)}\\ &= \sum_{k=1}^{K} \theta_k K_{kj}(\mathbf{x}). && \text{(S.34)} \end{aligned}$$

The second derivative is

$$\begin{aligned} \partial_j \psi_j(\mathbf{x};\boldsymbol{\theta}) &= \frac{\partial^2 \log p(\mathbf{x};\boldsymbol{\theta})}{\partial x_j^2} && \text{(S.35)}\\ &= \sum_{k=1}^{K} \theta_k \frac{\partial^2 F_k(\mathbf{x})}{\partial x_j^2} && \text{(S.36)}\\ &= \sum_{k=1}^{K} \theta_k H_{kj}(\mathbf{x}), && \text{(S.37)} \end{aligned}$$

which we can write more compactly as

$$\partial_j \psi_j(\mathbf{x};\boldsymbol{\theta}) = \boldsymbol{\theta}^\top \mathbf{h}_j(\mathbf{x}). \tag{S.38}$$

The score matching objective in Equation (9) features the sum $\sum_j \psi_j(\mathbf{x};\boldsymbol{\theta})^2$. The term $\psi_j(\mathbf{x};\boldsymbol{\theta})^2$ equals

$$\begin{aligned} \psi_j(\mathbf{x};\boldsymbol{\theta})^2 &= \left[\sum_{k=1}^{K} \theta_k K_{kj}(\mathbf{x})\right]^2 && \text{(S.39)}\\ &= \sum_{k=1}^{K}\sum_{k'=1}^{K} K_{kj}(\mathbf{x})K_{k'j}(\mathbf{x})\theta_k\theta_{k'}, && \text{(S.40)} \end{aligned}$$

so that

$$\sum_{j=1}^{m} \psi_j(\mathbf{x};\boldsymbol{\theta})^2 = \sum_{j=1}^{m}\sum_{k=1}^{K}\sum_{k'=1}^{K} K_{kj}(\mathbf{x})K_{k'j}(\mathbf{x})\theta_k\theta_{k'} \tag{S.41}$$

$$= \sum_{k=1}^{K}\sum_{k'=1}^{K} \theta_k\theta_{k'} \left[\sum_{j=1}^{m} K_{kj}(\mathbf{x})K_{k'j}(\mathbf{x})\right], \tag{S.42}$$

which can be more compactly expressed using matrix notation. Noting that

$$\sum_{j=1}^{m} K_{kj}(\mathbf{x}_i)K_{k'j}(\mathbf{x}_i)$$

equals the (k,k') element of the matrix-matrix product $\mathbf{K}(\mathbf{x}_i)\mathbf{K}(\mathbf{x}_i)^\top$,

$$\sum_{j=1}^{m} K_{kj}(\mathbf{x}_i)K_{k'j}(\mathbf{x}_i) = \left[\mathbf{K}(\mathbf{x}_i)\mathbf{K}(\mathbf{x}_i)^\top\right]_{k,k'}, \tag{S.43}$$

we can write

$$\sum_{j=1}^{m} \psi_j(\mathbf{x};\boldsymbol{\theta})^2 = \sum_{k=1}^{K}\sum_{k'=1}^{K} \theta_k\theta_{k'} \left[\mathbf{K}(\mathbf{x}_i)\mathbf{K}(\mathbf{x}_i)^\top\right]_{k,k'} \tag{S.44}$$

$$= \boldsymbol{\theta}^\top \mathbf{K}(\mathbf{x}_i)\mathbf{K}(\mathbf{x}_i)^\top \boldsymbol{\theta} \tag{S.45}$$

where we have used that for some matrix $\mathbf{A}$

$$\boldsymbol{\theta}^\top \mathbf{A}\boldsymbol{\theta} = \sum_{k,k'} \theta_k\theta_{k'}[\mathbf{A}]_{k,k'} \tag{S.46}$$

where $[\mathbf{A}]_{k,k'}$ is the (k,k') element of the matrix $\mathbf{A}$.
Inserting the expressions into Equation (9) gives

$$J(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m}\left[\partial_j\psi_j(\mathbf{x}_i;\boldsymbol{\theta}) + \frac{1}{2}\psi_j(\mathbf{x}_i;\boldsymbol{\theta})^2\right] \tag{S.47}$$

$$= \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m}\partial_j\psi_j(\mathbf{x}_i;\boldsymbol{\theta}) + \frac{1}{2}\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m}\psi_j(\mathbf{x}_i;\boldsymbol{\theta})^2 \tag{S.48}$$

$$= \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m}\boldsymbol{\theta}^\top\mathbf{h}_j(\mathbf{x}_i) + \frac{1}{2}\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\theta}^\top\mathbf{K}(\mathbf{x}_i)\mathbf{K}(\mathbf{x}_i)^\top\boldsymbol{\theta} \tag{S.49}$$

$$= \boldsymbol{\theta}^\top\left[\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m}\mathbf{h}_j(\mathbf{x}_i)\right] + \frac{1}{2}\boldsymbol{\theta}^\top\left[\frac{1}{n}\sum_{i=1}^{n}\mathbf{K}(\mathbf{x}_i)\mathbf{K}(\mathbf{x}_i)^\top\right]\boldsymbol{\theta} \tag{S.50}$$

$$= \boldsymbol{\theta}^\top\mathbf{r} + \frac{1}{2}\boldsymbol{\theta}^\top\mathbf{M}\boldsymbol{\theta}, \tag{S.51}$$

which is the desired result.

(b) *The pdf of a zero mean Gaussian parametrised by the variance σ^2 is*

$$p(x;\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{x^2}{2\sigma^2}\right), \qquad x \in \mathbb{R}. \tag{15}$$

The (multivariate) Gaussian is a member of the exponential family. By comparison with Equation (10), we can re-parametrise the statistical model $\{p(x;\sigma^2)\}_{\sigma^2}$ and work with

$$p(x;\theta) = \frac{1}{Z(\theta)} \exp\left(\theta x^2\right), \qquad \theta < 0, \qquad x \in \mathbb{R}, \tag{16}$$

instead. The two parametrisations are related by $\theta = -1/(2\sigma^2)$. Using the previous result on the (continuous) exponential family, determine the score matching estimate $\hat{\theta}$, and show that the corresponding $\hat{\sigma}^2$ is the same as the maximum likelihood estimate. This result is noteworthy because unlike in maximum likelihood estimation, score matching does not need the partition function $Z(\theta)$ for the estimation.

Solution. By comparison with Equation (10), the sufficient statistics $F(x)$ is x^2.

We first determine the score matching objective function. For that, we need to determine the quantities $\mathbf{r}$ and $\mathbf{M}$ in Equation (14). Here, both $\mathbf{r}$ and $\mathbf{M}$ are scalars, and so are the matrices $\mathbf{K}$ and $\mathbf{H}$ that define $\mathbf{r}$ and $\mathbf{M}$. By their definitions, we obtain

$$K(x) = \frac{\partial F(x)}{\partial x} = 2x \tag{S.52}$$

$$H(x) = \frac{\partial^2 F(x)}{\partial x^2} = 2 \tag{S.53}$$

$$r = 2 \tag{S.54}$$

$$M = \frac{1}{n}\sum_{i=1}^{n} K(x_i)^2 \tag{S.55}$$

$$= 4m_2 \tag{S.56}$$

where m_2 denotes the second empirical moment,

$$m_2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2. \tag{S.57}$$

With Equation (9), the score matching objective thus is

$$J(\theta) = 2\theta + \frac{1}{2} 4m_2\theta^2 \tag{S.58}$$

$$= 2\theta + 2m_2\theta^2 \tag{S.59}$$

A necessary condition for the minimiser to satisfy is

$$\frac{\partial J(\theta)}{\partial \theta} = 2 + 4\theta m_2 \tag{S.60}$$

$$= 0 \tag{S.61}$$

The only parameter value that satisfies the condition is

$$\hat{\theta} = -\frac{1}{2m_2}. \tag{S.62}$$

The second derivative of $J(\theta)$ is

$$\frac{\partial^2 J(\theta)}{\theta^2} = m_2, \tag{S.63}$$

which is positive (as long as all data points are non-zero). Hence $\hat{\theta}$ is a minimiser.

From the relation $\theta = -1/(2\sigma^2)$, we obtain that the score matching estimate of the variance σ^2 is

$$\hat{\sigma}^2 = -\frac{1}{2\hat{\theta}} = m_2. \tag{S.64}$$

We can obtain the score matching estimate $\hat{\sigma}^2$ from $\hat{\theta}$ in this manner for the same reason that we were able to work with transformed parameters in maximum likelihood estimation.

For zero mean Gaussians, the second moment m_2 is the maximum likelihood estimate of the variance, which shows that the score matching and maximum likelihood estimate are here the same. While the two methods generally yield different estimates, the result also holds for multivariate Gaussians where the score matching estimates also equal the maximum likelihood estimates (see the original article on score matching http://jmlr.org/papers/volume6/hyvarinen05a/hyvarinen05a.pdf).

Exercise 4. *Inverse transform sampling*

The cumulative distribution function (cdf) $F_x(\alpha)$ of a (continuous or discrete) random variable x indicates the probability that x takes on values smaller or equal to α,

$$F_x(\alpha) = \mathbb{P}(x \leq \alpha). \tag{17}$$

For continuous random variables, the cdf is defined via the integral

$$F_x(\alpha) = \int_{-\infty}^{\alpha} p_x(u) \mathrm{d}u, \tag{18}$$

where p_x denotes the pdf of the random variable x (u is here a dummy variable). Note that F_x maps the domain of x to the interval $[0, 1]$. For simplicity, we here assume that F_x is invertible.

(a) *For a continuous random variable x with cdf F_x show that the random variable $y = F_x(x)$ is uniformly distributed on $[0, 1]$.*

Importantly, this implies that the random variable $F_x^{-1}(y)$ has cdf F_x if y is uniformly distributed on $[0, 1]$, which gives rise to a method called "inverse transform sampling": In order to generate n iid samples of a random variable x with cdf F_x, we

- *calculate the inverse F_x^{-1}*
- *sample n iid random variables uniformly distributed on $[0, 1]$: $y_i \sim \mathcal{U}(0, 1)$, $i = 1, \ldots, n$.*
- *transform each sample by F_x^{-1}: $x_i = F_x^{-1}(y_i)$, $i = 1, \ldots, n$.*

By construction of the method, the x_i are n iid samples of x.

Solution. We start with the cumulative distribution function (cdf) F_y for y,

$$F_y(\beta) = \mathbb{P}(y \leq \beta). \tag{S.65}$$

Since $F_x(x)$ maps x to $[0, 1]$, $F_y(\beta)$ is zero for $\beta < 0$ and one for $\beta > 1$. We next consider $\beta \in [0, 1]$.

Denote the inverse of β by α, $F_x^{-1}(\beta) = \alpha$. Since F_x is a non-decreasing function, we have

$$\mathbb{P}(y \leq \beta) = \mathbb{P}(F_x(x) \leq \beta) = \mathbb{P}(x \leq F_x^{-1}(\beta)) = \mathbb{P}(x \leq \alpha) = F_x(\alpha). \tag{S.66}$$

Since $\mathbb{P}(x \leq \alpha) = F_x(\alpha)$, we obtain

$$F_y(\beta) = \mathbb{P}(y \leq \beta) = F_x(\alpha) = F_x(F_x^{-1}(\beta)) = \beta \tag{S.67}$$

The cdf F_y

$$F_y(\beta) = \begin{cases} 0 & \text{if } \beta < 0 \\ \beta & \text{if } \beta \in [0,1] \\ 1 & \text{if } \beta > 1 \end{cases} \tag{S.68}$$

is the cdf of a uniform random variable on $[0,1]$, so that $y = F_x(x)$ is uniformly distributed on $[0,1]$.

(b) *A Laplace random variable x of mean zero and variance one has the density $p(x)$*

$$p(x) = \frac{1}{\sqrt{2}} \exp\left(-\sqrt{2}|x|\right) \qquad x \in \mathbb{R}. \tag{19}$$

Use inverse transform sampling to generate n iid samples from x.

Solution. The main task is to compute the cumulative distribution function (cdf) F_x of x and its inverse. The cdf is by definition

$$F_x(\alpha) = \int_{-\infty}^{\alpha} \frac{1}{\sqrt{2}} \exp\left(-\sqrt{2}|u|\right) \mathrm{d}u. \tag{S.69}$$

We first consider the case where $\alpha \leq 0$. Since $-|u| = u$ for $u \leq 0$, we have

$$\begin{align} F_x(\alpha) &= \int_{-\infty}^{\alpha} \frac{1}{\sqrt{2}} \exp\left(\sqrt{2}u\right) \mathrm{d}u \tag{S.70} \\ &= \frac{1}{2} \exp\left(\sqrt{2}u\right) \Big|_{-\infty}^{\alpha} \tag{S.71} \\ &= \frac{1}{2} \exp\left(\sqrt{2}\alpha\right). \tag{S.72} \end{align}$$

For $\alpha > 0$, we have

$$\begin{align} F_x(\alpha) &= \int_{-\infty}^{\alpha} \frac{1}{\sqrt{2}} \exp\left(-\sqrt{2}|u|\right) \mathrm{d}u \tag{S.73} \\ &= 1 - \int_{\alpha}^{\infty} \frac{1}{\sqrt{2}} \exp\left(-\sqrt{2}|u|\right) \mathrm{d}u \tag{S.74} \end{align}$$

where we have used the fact that the pdf has to integrate to one. For values of $u > 0$, $-|u| = -u$, so that

$$\begin{align} F_x(\alpha) &= 1 - \int_{\alpha}^{\infty} \frac{1}{\sqrt{2}} \exp\left(-\sqrt{2}u\right) \mathrm{d}u \tag{S.75} \\ &= 1 + \frac{1}{2} \exp\left(-\sqrt{2}u\right) \Big|_{\alpha}^{\infty} \tag{S.76} \\ &= 1 - \frac{1}{2} \exp\left(-\sqrt{2}\alpha\right). \tag{S.77} \end{align}$$

In total, for $\alpha \in \mathbb{R}$, we thus have

$$F_x(\alpha) = \begin{cases} \frac{1}{2} \exp\left(\sqrt{2}\alpha\right) & \text{if } \alpha \leq 0 \\ 1 - \frac{1}{2} \exp\left(-\sqrt{2}\alpha\right) & \text{if } \alpha > 0 \end{cases} \tag{S.78}$$

Figure 1 visualises $F_x(\alpha)$.

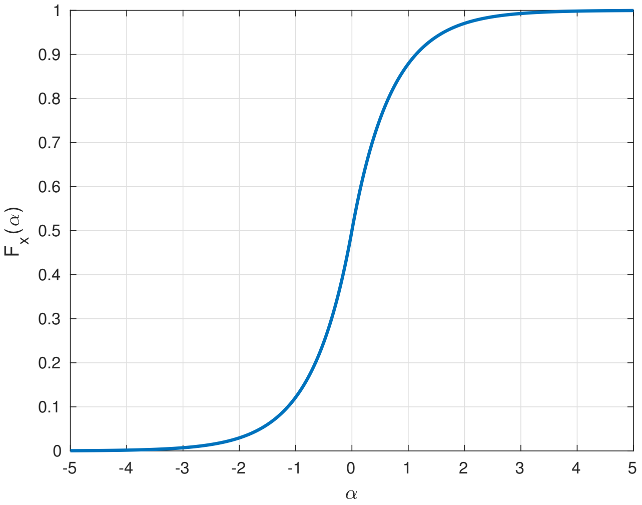


Figure 1: The cumulative distribution function $F_x(\alpha)$ for a Laplace distributed random variable.

As the figure suggests, there is a unique inverse to $y = F_x(\alpha)$. For $y \leq 1/2$, we have

$$y = \frac{1}{2}\exp\left(\sqrt{2}\alpha\right) \tag{S.79}$$

$$\log(2y) = \sqrt{2}\alpha \tag{S.80}$$

$$\alpha = \frac{1}{\sqrt{2}}\log(2y) \tag{S.81}$$

For $y > 1/2$, we have

$$y = 1 - \frac{1}{2}\exp\left(-\sqrt{2}\alpha\right) \tag{S.82}$$

$$-y = -1 + \frac{1}{2}\exp\left(-\sqrt{2}\alpha\right) \tag{S.83}$$

$$1 - y = \frac{1}{2}\exp\left(-\sqrt{2}\alpha\right) \tag{S.84}$$

$$\log(2 - 2y) = -\sqrt{2}\alpha \tag{S.85}$$

$$\alpha = -\frac{1}{\sqrt{2}}\log(2 - 2y) \tag{S.86}$$

The function $y \mapsto g(y)$ that occurs in the log

$$g(y) = \begin{cases} 2y & \text{if } y \leq \frac{1}{2} \\ 2 - 2y & \text{if } y > \frac{1}{2} \end{cases} \tag{S.87}$$

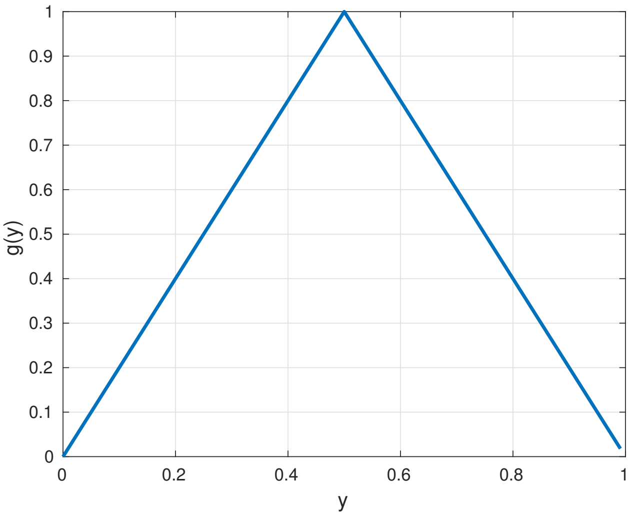
is shown below and can be written as $g(y) = 1 - 2|y - 1/2|$.

We thus can write the inverse $F_x^{-1}(y)$ of the cdf $y = F_x(\alpha)$ as

$$F_x^{-1}(y) = -\text{sign}\left(y - \frac{1}{2}\right)\frac{1}{\sqrt{2}}\log\left[1 - 2\left|y - \frac{1}{2}\right|\right]. \tag{S.88}$$

To generate n iid samples from x, we first generate n iid samples y_i that are uniformly distributed on $[0, 1]$, and then compute for each $F_x^{-1}(y_i)$. The properties of inverse transform sampling guarantee that the x_i,

$$x_i = F_x^{-1}(y_i) \tag{S.89}$$



are independent and Laplace distributed.

Inverse transform sampling can be used to generate samples from many standard distributions. For example, it allows one to generate Gaussian random variables from uniformly distributed random variables. The method is called the Box-Muller transform, see e.g. `https://en.wikipedia.org/wiki/Box-Muller_transform`. How to generate the required samples from the uniform distribution is a research field on its own, see e.g. `https://en.wikipedia.org/wiki/Random_number_generation` and `http://statweb.stanford.edu/~owen/mc/Ch-unifrng.pdf`.