Probabilistic Modelling and Reasoning
Tutorial 8

Spring 2018
Michael Gutmann

The purpose of the tutorials is twofold: First, they help you better understand the lecture material. Secondly, they provide exam preparation material. You are not expected to complete all questions before the tutorial sessions. Start early and do as many as you have time for.

Exercise 1. *Factor analysis*

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

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

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 $\mathbf{C}, \mathbf{F}, \mathbf{c}, \mathbf{\Psi}$?

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

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

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

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

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.

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.

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

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 $\mathbf{C} = \mathbf{E}\mathbf{\Lambda}\mathbf{E}^\top$, show that

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

is a whitening 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.

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

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

(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.

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.

(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.