Machine Learning 2 – Group ESHG
Assignment 03

Willi Gierke, Arik Elimelech, Mehmed Halilovic, Leon Sixt

May 11, 2017

Exercise 1: Maximum Entropy Distributions (10+10+5+5 P)

The differential entropy $H(x)$ for a random variable $x \in \mathbb{R}$ with probability density function $p(x)$ is given by

$$H(x) = -\int p(x) \log p(x) \, dx$$

We would like to find the probability density function $p(x)$ that maximizes the differential entropy under the following constraints

$$\forall x : p(x) \geq 0 \quad , \quad \int p(x) \, dx = 1 \quad , \quad \mathrm{E}[x] = 0 \quad , \quad \mathrm{Var}[x] = \sigma^2 .$$

The first set of inequality constraints is handled by rewriting the unknown density function as $p(x) = \exp(s(x))$ and searching for a function $s(x)$ that maximizes the objective. Here, we view the function $s(x)$ as an infinite-dimensional vector, and therefore can write

$$\frac{\partial \int f(s(x)) \, dx}{\partial s(x)} = f'(s(x)).$$

(a) *Write* the Lagrange function $\Lambda(s(x), \lambda_1, \lambda_2, \lambda_3)$ corresponding to the constrained optimization problem above, where $\lambda_1, \lambda_2, \lambda_3$ are used to incorporate the three equality constraints.

(b) *Show* that the function $s(x)$ that maximizes the objective $H(x)$ is quadratic in x.

(c) *Show* that the function $p(x)$ that maximizes the objective $H(x)$ is a Gaussian probability density function with mean 0 and variance σ^2.

(d) *Show* that for every univariate random variable x of mean 0 and variance σ^2, and assuming $x^* \sim \mathcal{N}(0, \sigma^2)$, the negentropy $J(x) = H(x^*) - H(x)$ that independent component analysis seeks to maximize is always greater or equal to 0, and is equal to 0 when x is Gaussian-distributed.

Exercise 2: Finding Independent Components (10+15+15 P)

We consider the joint probability distribution $p(x, y) = p(x)\, p(y|x)$ with

$$p(x) \sim \mathcal{N}(0, 1),$$
$$p(y|x) = \frac{1}{2}\, \delta(y - x) + \frac{1}{2}\, \delta(y + x),$$

where $\delta(\cdot)$ denotes the Dirac delta function. A useful property of linear component analysis for two-dimensional probability distributions is that the set of all possible directions to look for in $\mathbb{R}^2$ is directly given by

$$\left\{ \begin{pmatrix} \cos\theta \\ \sin\theta \end{pmatrix}, \quad 0 \leq \theta < 2\pi \right\}.$$

The projection of the random vector (x, y) on a particular component can therefore be expressed as a function of θ:

$$z(\theta) = x \cos(\theta) + y \sin(\theta).$$

As a result, linear component analysis such as PCA or ICA in the two-dimensional space is reduced to finding the parameters $\theta \in [0, 2\pi[$ that maximize a certain objective $J(z(\theta))$.

(a) *Sketch* the joint probability distribution $p(x, y)$, along with the projections $z(\theta)$ of this distribution for angles $\theta = 0$, $\theta = \pi/8$ and $\theta = \pi/4$.

(b) *Find* the principal components of $p(x, y)$. That is, find the parameters $\theta \in [0, 2\pi[$ that maximize the variance of the projected data $z(\theta)$.

(c) *Find* the independent components of $p(x, y)$. That is, find the parameters $\theta \in [0, 2\pi[$ that maximize the non-Gaussianity of $z(\theta)$. We use as a measure of non-Gaussianity the excess kurtosis defined as

$$\mathrm{kurt}\,[z(\theta)] = \frac{\mathrm{E}\big[(z(\theta) - \mathrm{E}[z(\theta)])^4\big]}{(\mathrm{Var}[z(\theta)])^2} - 3.$$

Task 1

a

$$E[x] = \int p(x)x\mathrm{d}x = 0$$
$$Var[x] = \int p(x)(x-\cancel{\mu})^2\mathrm{d}x = \int p(x)x^2\mathrm{d}x = \sigma^2$$
$$\int p(x)\mathrm{d}x = 1 \quad \Rightarrow \quad \sigma^2 = \int p(x)\sigma^2\mathrm{d}x$$

$$\begin{aligned}\nabla(s(x), \lambda_1, \lambda_2, \lambda_3) = &- \int \exp(s(x))s(x)\mathrm{d}x \\ &+ \lambda_1(\int \exp(s(x))\mathrm{d}x - 1) \\ &+ \lambda_2 \int \exp(s(x))x\mathrm{d}x \\ &+ \lambda_3(\int \exp(s(x))(x^2 - \sigma^2)\mathrm{d}x)\end{aligned}$$

b

$$\begin{aligned}\frac{\partial \nabla(s(x), \lambda_1, \lambda_2, \lambda_3)}{\partial s(x)} &= - \int \exp(s(x))(1 + s(x))\mathrm{d}x \\ &+ \lambda_1 \int \exp(s(x))\mathrm{d}x \\ &+ \lambda_2 \int \exp(s(x))x\mathrm{d}x \\ &+ \lambda_3(\int \exp(s(x))(x^2 - \sigma^2)\mathrm{d}x) \\ &= \int \exp(s(x)) \left(-(1 + s(x)) + \lambda_1 + \lambda_2 x + \lambda_3(x^2 - \sigma^2)\right) \mathrm{d}x\end{aligned}$$

By setting this derivative to zero, we can show that at the extremum, $s(x)$ is a quadratic function.

$$\begin{aligned}& \frac{\partial \nabla(s(x), \lambda_1, \lambda_2, \lambda_3)}{\partial s(x)} \stackrel{!}{=} 0 \\ \Leftrightarrow \quad & \int \exp(s(x)) \left(-(1 + s(x)) + \lambda_1 + \lambda_2 x + \lambda_3(x^2 - \sigma^2)\right) \mathrm{d}x = 0 \\ \Leftrightarrow \quad & \int \exp(s(x)) \left(-1 + \lambda_1 + \lambda_2 x + \lambda_3(x^2 - \sigma^2)\right) \mathrm{d}x = \int \exp(s(x))s(x)\mathrm{d}x \\ \Rightarrow \quad & s(x) = -1 + \lambda_1 + \lambda_2 x + \lambda_3(x^2 - \sigma^2)\end{aligned}$$

c

$$p(x) = \exp(s(x)) = \exp(-1 + \lambda_1 + \lambda_2 x + \lambda_3(x^2 - \sigma^2)) \tag{1}$$

Using the constraints, we can determine the λ-values to show that this is actually a Gaussian with zero mean and variance σ^2.

$$\begin{aligned} 1 &= \int p(x)\mathrm{d}x = \int \exp(-1 + \lambda_1 + \lambda_2 x + \lambda_3(x^2 - \sigma^2))\mathrm{d}x \\ &= \sqrt{\frac{\pi}{-\lambda_3}} \exp\left(\frac{\lambda_2^2}{-4\lambda_3} + (\lambda_1 - 1 - \lambda_3\sigma^2)\right) \end{aligned}$$

We can then use the variance constraint

$$\begin{aligned} & \int \exp(s(x))x^2\mathrm{d}x = \sigma^2 \\ \Leftrightarrow \quad & \frac{1}{-2\lambda_3} \underbrace{\sqrt{\frac{\pi}{-\lambda_3}} \exp\left(\frac{\lambda_2^2}{-4\lambda_3} + (\lambda_1 - 1 - \lambda_3\sigma^2)\right)}_{=1} = \sigma^2 \\ \Leftrightarrow \quad & \lambda_3 = -\frac{1}{2\sigma^2} \end{aligned}$$

Hence, our distribution becomes

$$p(x) = \exp(-1/2 + \lambda_1 + \lambda_2 x - \frac{x^2}{2\sigma^2})$$

and the first constraint

$$\begin{aligned} & \sqrt{\frac{\pi}{-\lambda_3}} \exp\left(\frac{\lambda_2^2}{-4\lambda_3} + (\lambda_1 - 1 - \lambda_3\sigma^2)\right) = 1 \\ \Leftrightarrow \quad & \sqrt{2\pi\sigma^2} \exp\left(\lambda_2^2\sigma^2/2 + (\lambda_1 - 1/2)\right) = 1 \\ \Leftrightarrow \quad & (\lambda_1 - 1/2) = -\ln\left(\sqrt{2\pi\sigma^2}\right) \\ \Leftrightarrow \quad & \lambda_1 - 1/2 = -\ln\left(\sqrt{2\pi\sigma^2}\right) - \lambda_2^2\sigma^2/2 \end{aligned}$$

which yields

$$p(x) = \exp\left(-\ln\left[\sqrt{2\pi\sigma^2}\right] + \lambda_2 x - \frac{x^2}{2\sigma^2}\right) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\lambda_2 x - \frac{x^2}{2\sigma^2}\right)$$

Since the mean is zero, λ_2 has to disappear (otherwise the Gaussian would be shifted). Therefore, in the end we obtain the expected result

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{x^2}{2\sigma^2}\right) \tag{2}$$

d

We know x^* is Gaussian distributed. From our constraints we know that $E[x] = 0$ and $Var[x] = \sigma^2$. With these constraints the differential entropy $H(x)$ for a random variable is maximzed by the Gaussian distribution. Hence, there doesn't exist a random variable x with $H(x) > H(x^*)$. Therefore, $J(x) >= 0$.

2

2a

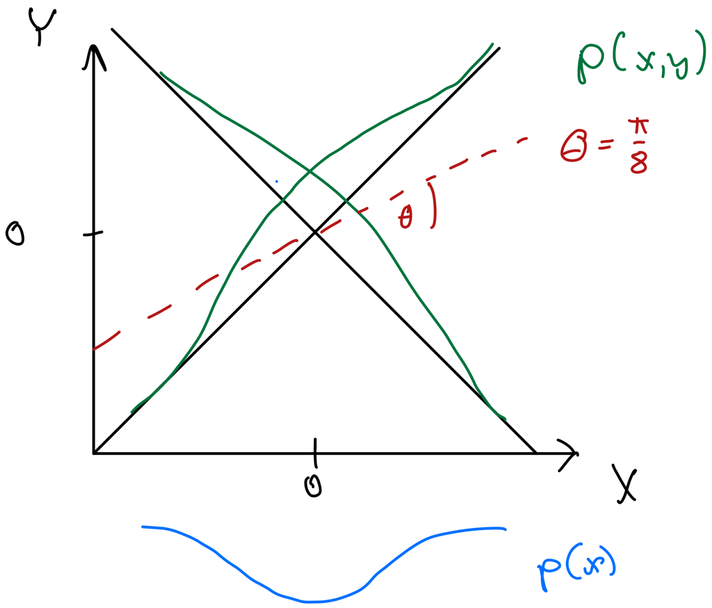

Figure 1: Sketch of $p(x, y)$. The black lines indicate where the mass is located and the green distributions how the mass is distributed.

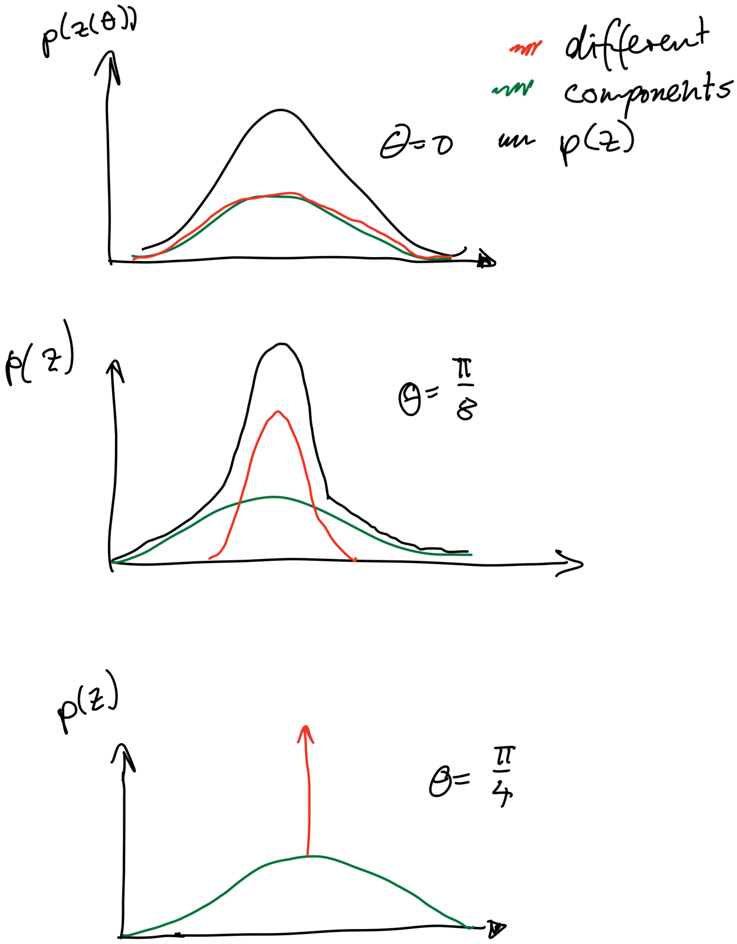

Figure 2: $z(\theta)$ for $\theta = 0, \theta = \frac{\pi}{8}, \theta = \frac{\pi}{4}$. Green and red indicate the different contribution of each of the components. The arrow symbolises the dirac delta function.

2b

As the projection of a gaussian distribution is again a gaussian distribution, we can write the distribution $p(z(\theta))$ as:

$$p(z(\theta)) = \frac{1}{2}\mathcal{N}_1(\theta) + \frac{1}{2}\mathcal{N}_2(\theta) \tag{3}$$

where $\mathcal{N}_1(\theta)$ is the projection of the distribution of first compontent (the one with angle $\frac{\pi}{4}$) and $\mathcal{N}_2(\theta)$ analogously for angle $-\frac{\pi}{4}$.
To figure out $p(z(\theta))$, we just have to find the projection of the distribution of both compontents. The means are 0 and will remain 0 under the projections.

The projection of the simgas are given by:

$$\hat{\sigma}_1^2 = \left|\cos\left(\theta - \frac{\pi}{4}\right)\right| \tag{4}$$

$$\hat{\sigma}_2^2 = \left|\sin\left(\theta - \frac{\pi}{4}\right)\right| \tag{5}$$

We ignore the precise intensities in our calculations as only their relative magnitude matter.
Substituting back into $p(z(\theta))$:

$$p(z(\theta)) = \frac{1}{2}\mathcal{N}_1(0, \hat{\sigma}_1^2) + \frac{1}{2}\mathcal{N}(0, \hat{\sigma}_2^2) \tag{6}$$

$$\mathbb{V}[Z] = \int z^2 \left(\frac{1}{2}\mathcal{N}(0, \hat{\sigma}_1^2) + \frac{1}{2}\mathcal{N}(0, \hat{\sigma}_2^2)\right) dz = \tag{7}$$

$$= \frac{1}{2}\hat{\sigma}_1^2 + \frac{1}{2}\hat{\sigma}_2^2 \tag{8}$$

Now, we want to find the directions that maximizes the variances. We project the data and therefore θ and $\theta - \pi$ result in the same variance. Thus we consider only $\cos\left(\theta - \frac{\pi}{4}\right) \geq 0$:

$$\mathbb{V}[Z] = \frac{1}{2}\hat{\sigma}_1^2 + \frac{1}{2}\hat{\sigma}_2^2 \tag{9}$$

$$\propto \left|\cos\left(\theta - \frac{\pi}{4}\right)\right| + \left|\sin\left(\theta - \frac{\pi}{4}\right)\right| \tag{10}$$

$$= \cos\left(\theta - \frac{\pi}{4}\right) + \left|\sin\left(\theta - \frac{\pi}{4}\right)\right| = L \tag{11}$$

Here, we removed the absolute value around cos, as we know it is always positive. For $\theta \geq \frac{\pi}{4}$:

$$\frac{\partial L}{\partial \theta} = -\sin\left(\theta - \frac{\pi}{4}\right) + \cos\left(\theta - \frac{\pi}{4}\right) \stackrel{!}{=} 0 \tag{12}$$

$$\iff \cos\left(\theta - \frac{\pi}{4}\right) = \sin\left(\theta - \frac{\pi}{4}\right) \tag{13}$$

$$\iff \theta = \frac{\pi}{2} \tag{14}$$

For $\theta < \frac{\pi}{4}$:

$$\frac{\partial L}{\partial \theta} = -\sin\left(\theta - \frac{\pi}{4}\right) - \cos\left(\theta - \frac{\pi}{4}\right) \stackrel{!}{=} 0 \tag{15}$$

$$\iff -\cos\left(\theta - \frac{\pi}{4}\right) = \sin\left(\theta - \frac{\pi}{4}\right) \tag{16}$$

$$\iff \theta = 0 \tag{17}$$

Therefore, the principal components are $(1, 0)$ and $(0, 1)$ with angles 0 and $\frac{\pi}{2}$.

2c

$$\frac{\mathbb{E}[(z(\theta) - \mathbb{E}[z(\theta)])^4]}{\mathbb{V}[z(\theta)]^2} = \tag{18}$$

$$= \frac{\mathbb{E}[z(\theta)^4]}{\mathbb{V}[z(\theta)]^2} \tag{19}$$

$$= \frac{\int z^4 \left(\frac{1}{2}\mathcal{N}_1(0, \hat{\sigma}_1^2) + \frac{1}{2}\mathcal{N}(0, \hat{\sigma}_2^2)\right) dz}{(\frac{1}{2}\hat{\sigma}_1^2 + \frac{1}{2}\hat{\sigma}_2^2)^2} \tag{20}$$

$$= \frac{\frac{3}{2}\left(\hat{\sigma}_1^4 + \hat{\sigma}_2^4\right)}{\left(\frac{1}{2}\hat{\sigma}_1^2 + \frac{1}{2}\hat{\sigma}_2^2\right)^2} \tag{21}$$

$$= \frac{\frac{3}{2}\cos(4\theta - \pi)}{\left(\frac{1}{2}\left|\cos\left(\theta - \frac{\pi}{4}\right)\right| + \frac{1}{2}\left|\sin\left(\theta - \frac{\pi}{4}\right)\right|\right)^2} \tag{22}$$

$$= \frac{\frac{3}{2}\cos(4\theta - \pi)}{\frac{1}{4} + \frac{1}{2}\left|\cos\left(\theta - \frac{\pi}{4}\right)\right|\left|\sin\left(\theta - \frac{\pi}{4}\right)\right|} \tag{23}$$

Observing that the denominator is minimum for $\pm\frac{\pi}{2}$ and that the numerator is also maximal for $\pm\frac{\pi}{2}$, yields the two solutions $\theta = \pm\frac{\pi}{2}$.