HY673 – Tutorial 3

Thomas Marchioro

February 2023

Exercise 1 (Change of Variable)**.** Consider a random variable with normal distribution $X \sim \mathcal{N}(0, \sigma^2)$.

1. Compute the PDF of the r.v. $Y = aX^2 + c$, with $a > 0$. *Hint*: When a function g admits multiple inverses $h_1, \ldots, h_k$, the change of variable formula for the PDF is
$$p_Y(y) = \sum_{i=1}^{n} p_X(h_k(y)) \left| \frac{d}{dy} h_k(y) \right|.$$
2. Compute the value of $\mathbb{E}\left[X^3 + aX^2 + bX + c\right]$.

Solution.

1. We can calculate the PDF p_Y of Y using the change of variable formula
$$p_Y(y) = \sum_{i=1}^{n} p_X(h_k(y)) \left| \frac{d}{dy} h_k(y) \right|. \tag{1}$$
where $h_1, \ldots, h_n$ are the inverse functions of $g(x) = ax^2 + c$. Let's do it step by step:
   (a) Find the inverse function $x = g^{-1}(y)$:
$$y = ax^2 + c \Rightarrow x = \pm\sqrt{\frac{y-c}{a}} \Rightarrow g^{-1}(y) = \pm\sqrt{\frac{y-c}{a}}. \tag{2}$$
   The function g admits two possible inverses,
$$h_1(y) = \sqrt{\frac{y-c}{a}} \text{ and } h_2(y) = -\sqrt{\frac{y-c}{a}}. \tag{3}$$
   In this case, you need to sum all of them in the change of variable formula, i.e.,
$$p_Y(y) = \left| \frac{d}{dy} h_1(y) \right| p_X(h_1(y)) + \left| \frac{d}{dy} h_2(y) \right| p_X(h_2(y)). \tag{4}$$
   Notice that since $a > 0$, the argument of the square root is non-negative when $y > c$. This is the domain for both h_1 and h_2, i.e., the inverse functions are defined only for $y \geq c$, which means that the change of variable cannot be applied in the region $y < c$. However, one can easily see that $ax^2 + c$ is never less than c, meaning that the PDF for $y < c$ must be 0.
   (b) Compute the absolute derivative for $h_1(y)$ and $h_2(y)$:
$$\left| \frac{d}{dy} h_1(y) \right| = \frac{1/a}{2\sqrt{\frac{y-c}{a}}}, \quad \left| \frac{d}{dy} h_2(y) \right| = \left| -\frac{1/a}{2\sqrt{\frac{y-c}{a}}} \right| = \frac{1/a}{2\sqrt{\frac{y-c}{a}}} \tag{5}$$
   The derivatives are equal in absolute value.
   (c) Compute $p_X(h_1(y))$ and $p_X(h_2(y))$ by the inverses in the PDF of X. Notice that they both yield the same result due to the symmetry of X's PDF:
$$p_X(h_i(y)) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(h_i(y))^2}{2\sigma^2} \right) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{|\frac{y-c}{a}|}{2\sigma^2} \right) \tag{6}$$
$$= \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{y-c}{2a\sigma^2} \right), \tag{7}$$
   where the last step is due to $y - c$ being non-negative for $y \geq c$.

Thus, at the end of the day the PDF of Y is

$$p_Y(y) = \begin{cases} \frac{1}{a\sqrt{\frac{y-c}{a}}} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{y-c}{2a\sigma^2}\right) & \text{for } y \geq c \\ 0 & \text{for } y < c \end{cases}. \tag{8}$$

2. You don't need to use the change of variable formula to compute the expectation for a function $g(X)$ of the r.v. X. Typically, it's simpler to evaluate the integral

$$\mathbb{E}[g(X)] = \int_{-\infty}^{+\infty} g(x) p_X(x) dx \tag{9}$$

Let's solve also this one step by step:

(a) Use the linearity of expectation:

$$\mathbb{E}\left[X^3 + aX^2 + bX + c\right] = \mathbb{E}[X^3] + a\mathbb{E}[X^2] + b\mathbb{E}[X] + c. \tag{10}$$

(b) The PDF of X is an even function, so the expectation of any odd function of X (such as X^3 and X) is 0:

$$\mathbb{E}[X^3] + a\mathbb{E}[X^2] + b\mathbb{E}[X] + c = 0 \tag{11}$$

(c) To compute $\mathbb{E}[X^2]$, you can use the relationship between variance, second moment, and mean, i.e. $\mathrm{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$. In this case, since the mean of X is 0, the second moment $\mathbb{E}[X^2]$ is equal to the variance, which is σ^2.

Summing up all the steps, we obtain that the expectation evaluates to

$$\mathbb{E}\left[X^3 + aX^2 + bX + c\right] = a\sigma^2 + c. \tag{12}$$

Exercise 2 (Maximum Likelihood Estimation)**.** A random variable X is described by the following parametric piecewise-uniform distribution

$$p_\theta(x) = \begin{cases} 1-\theta, & 0 \leq x \leq 1 \\ 2\theta, & -\frac{1}{2} \leq x < 0 \end{cases} \tag{13}$$

The parameter θ is unknown and needs to be estimated based on the observations $X_1, \ldots, X_n$.

1. Let N_1, N_2 be random variables describing the number of observations falling in the regions $[0, 1]$ and $[-1/2, 0)$, respectively, for a sample of n observations. Compute the expectation of both N_1 and N_2.
2. Let n_1, n_2 be the observed values of N_1 and N_2 in the sample $x_1, \ldots, x_n$. Derive the maximum likelihood estimator $\hat{\theta}_{\mathrm{MLE}}$ (*Hint*: n_1 and n_2 are all you need to compute the likelihood).
3. Compute the Fisher information $\mathcal{I}(\theta) = -\mathbb{E}_{X \sim p_\theta}\left[\frac{\partial^2}{\partial \theta^2} \log p_\theta(X)\right]$.

Solution.

1. Let χ be the indicator function. The expectations of N_1 and N_2 for n observations $X_1, \ldots, X_n$ are

$$\mathbb{E}[N_1] = \mathbb{E}\left[\sum_{i=1}^{n} \chi\{X_i \in [0,1]\}\right] = \sum_{i=1}^{n} \Pr[X_i \in [0,1]] \tag{14}$$

$$= n \Pr[X_i \in [0,1]] = n \int_0^1 p_\theta(x) dx = (1-\theta)n \tag{15}$$

and, likewise,

$$\mathbb{E}[N_2] = n \int_{-1/2}^{0} p_\theta(x) dx = \theta n. \tag{16}$$

Notice that, as expected (pun intended), $\mathbb{E}[N_1 + N_2] = n$.

2. The likelihood of the observed sample can be expressed as

$$\prod_{i=1}^{n} p_\theta(x_i) = (1-\theta)^{n_1}(2\theta)^{n_2}. \tag{17}$$

The best estimator according to MLE is the one that maximizes the log-likelihood, i.e.,

$$\hat{\theta}_{\text{MLE}} = \arg\max_\theta \sum_{i=1}^{n} \log p_\theta(x_i) = \arg\max_\theta n_1 \log(1-\theta) + n_2 \log(2\theta). \tag{18}$$

Notice that using the properties of the logarithm, $n_2 \log(2\theta) = n_2 \log\theta + n_2 \log(2)$. The term $n_2 \log(2)$ is constant, meaning that it does not affect the result of the argmax and can thus be dropped. At this point we can easily find the maximum by taking the derivative w.r.t. θ and looking for critical points.

$$\frac{\partial}{\partial\theta}\left[n_1 \log(1-\theta) + n_2 \log(2\theta)\right] = -\frac{n_1}{1-\theta} + \frac{n_2}{\theta} = 0, \tag{19}$$

which is solved for

$$(1-\theta)n_2 - \theta n_1 = 0 \Rightarrow \hat{\theta}_{\text{MLE}} = \frac{n_2}{n_1+n_2} = \frac{n_2}{n}. \tag{20}$$

3. The Fisher information is

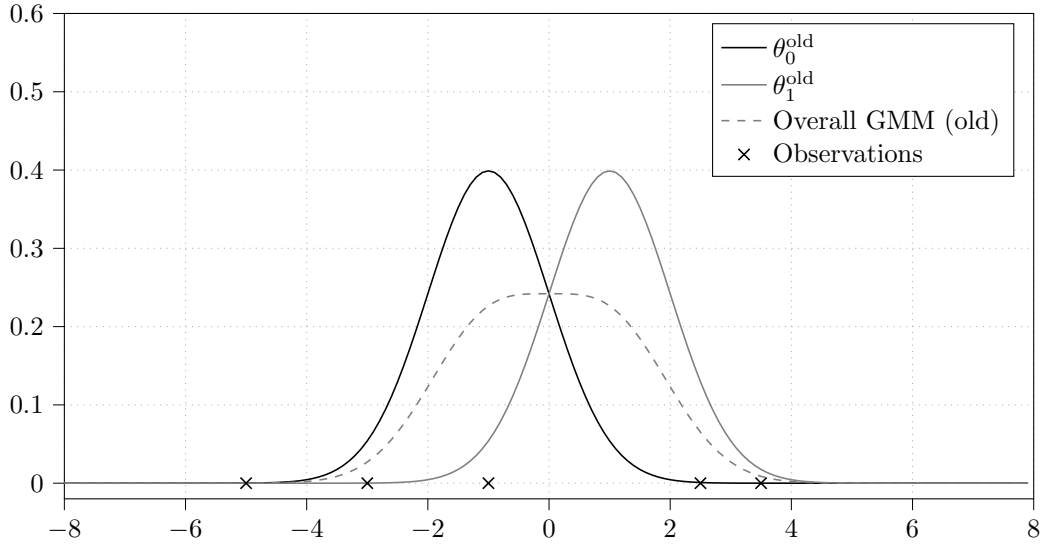
$$\begin{aligned}
\mathcal{I}(\theta) &= -\int_{-\infty}^{\infty} p_\theta(x_i)\left(\frac{\partial^2}{\partial\theta^2}\log p_\theta(x_i)\right)dx && (21)\\
&= -\int_0^1 (1-\theta)\frac{\partial^2}{\partial\theta^2}\log(1-\theta)dx - \int_{-1/2}^0 2\theta\frac{\partial^2}{\partial\theta^2}(2\theta)dx && (22)\\
&= -(1-\theta)\cdot\frac{\partial^2}{\partial\theta^2}\log(1-\theta) - \theta\cdot\frac{\partial^2}{\partial\theta^2}\log(2\theta) && (23)\\
&= -(1-\theta)\frac{\partial}{\partial\theta}\frac{-1}{1-\theta} - \theta\frac{\partial}{\partial\theta}\frac{2}{2\theta} && (24)\\
&= (1-\theta)\frac{1}{(1-\theta)^2} + \theta\frac{1}{\theta^2} && (25)\\
&= \frac{1}{1-\theta} + \frac{1}{\theta} && (26)\\
&= \frac{1}{\theta(1-\theta)}. && (27)
\end{aligned}$$

Notice that this result is equivalent to the Fisher information of a Bernoulli trial (essentially a "coin toss") with probabilities $(\theta, 1-\theta)$. This result makes sense if you think about what each observation tells you about θ. Essentially all that matters is in which region the observation falls: either $[0,1]$ or $[-1/2, 0)$. This happens with probabilities $1-\theta$ and θ.

Exercise 3 (Toy GMM)**.** Consider a simple 1D Gaussian mixture model (GMM) with two modes θ_0, θ_1. The model is initialized with the following parameters:

- θ_0^{old}: prior π_0^{old}, mean $\mu_0^{\text{old}} = -1$, standard deviation $\sigma_0^{\text{old}} = 1$;
- θ_1^{old}: prior π_1^{old}, mean $\mu_1^{\text{old}} = -1$, standard deviation $\sigma_1^{\text{old}} = 1$.

Suppose the observations are $x = [-5, -3, -1, +2.5, +3.5]$. You can see the initial state and the observed data points in the following figure.



While answering the questions below, you can use Python to perform the necessary calculations. If you do that, you should include your code among the submitted files.

1. **E step**: What are the values of the responsibilities for each observed data point in x? Report them in a table.

Data point	-5	-3	-1	$+2.5$	$+3.5$
Responsibilities of θ_0^{old}					
Responsibilities of θ_1^{old}					

2. **M step**: What are the updated parameters $\pi_0^{\text{new}}, \mu_0^{\text{new}}, \sigma_0^{\text{new}}, \pi_1^{\text{new}}, \mu_1^{\text{new}}, \sigma_1^{\text{new}}$ of the GMM? Provide a plot of its PDF.

3. Compute the overall log-likelihood of the observations x according to the old and new GMM parameters. Did the likelihood increase or decrease after the update? Is this result expected?

4. Sample 1000 points from the new GMM model and make a histogram in the range $[-8, 8]$ with 20 bins. You can use the `plt.hist()` function with the option `density` set to `True`. Compare the histogram with a plot of the PDF for the overall (new) GMM. How do the two graphs compare?

Solution.

1. The responsibility for a value x_n and mode θ_k is computed as follows:

$$\gamma_{nk} = \frac{\pi_k \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left(-\frac{(x_n-\mu_k)^2}{2\sigma_k^2}\right)}{\sum_{j=1}^{2} \pi_j \frac{1}{\sqrt{2\pi}\sigma_j} \exp\left(-\frac{(x_n-\mu_j)^2}{2\sigma_j^2}\right)} \tag{28}$$

$$= \frac{\exp\left(-\frac{(x_n-\mu_k)^2}{2}\right)}{\sum_{j=1}^{2} \exp\left(-\frac{(x_n-\mu_j)^2}{2}\right)} \tag{29}$$

Data point	-5	-3	-1	$+2.5$	$+3.5$
Resp. of θ_0^{old}	0.999	0.997	0.881	6.7×10^{-3}	9.1×10^{-4}
Resp. of θ_1^{old}	4.5×10^{-5}	2.5×10^{-3}	0.119	0.993	0.999

2. Following the M step, the updated parameters are computed as

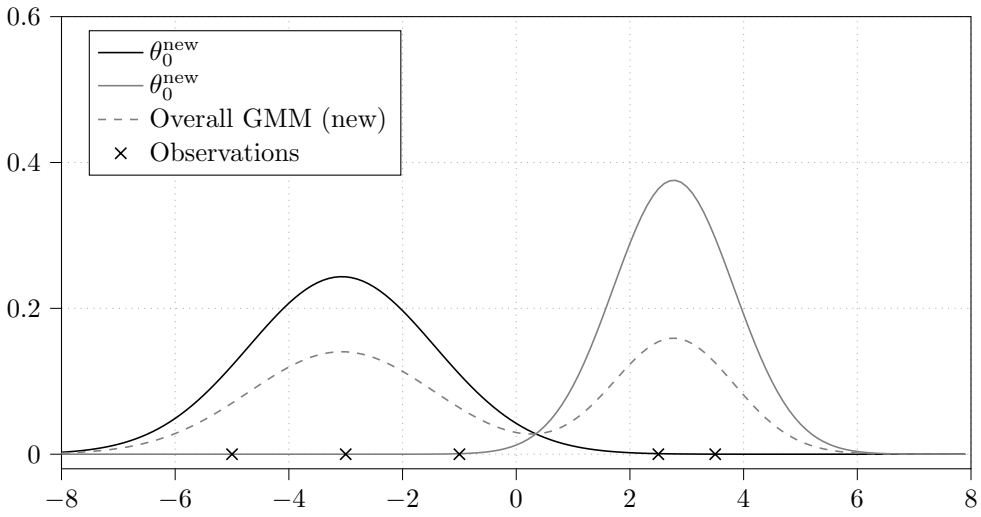
$$\pi_k^{\text{new}} = \frac{N_k}{5} \tag{30}$$

$$\mu_k^{\text{new}} = \frac{1}{N_k} \sum_{n=1}^{5} \gamma_{nk} x_n \tag{31}$$

$$\sigma_k^{\text{new}} = \sqrt{\frac{1}{N_k} \sum_{n=1}^{5} \gamma_{nk} (x_n - \mu_k^{\text{new}})^2} \tag{32}$$

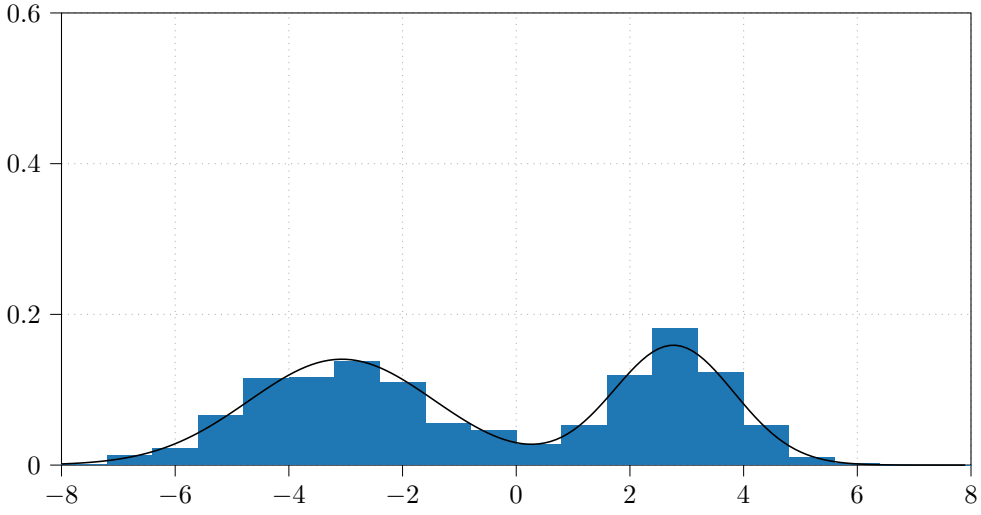
with $N_k = \sum_{n=1}^{5} \gamma_{nk}$.

	Prior	Mean	Standard deviation
θ_0^{new}	0.58	−3.06	1.64
θ_1^{new}	0.42	2.77	1.06



3. The log-likelihood computed on the initial and updated parameters are −1.116 and −0.591, respectively. The likelihood value is increased, which is expected since that is the purpose of each step of the EM algorithm.

4. Being the sample sufficiently large, the histogram with normalized area reflects the PDF of the overall GMM model.



Exercise 4 (Sampling from Multivariate Gaussians)**.** In this exercise you will learn how to sample from any multivariate normal distribution starting from i.i.d. values of a standard normal random variable.

1. Show that if a vector $z \in \mathbb{R}^d$ is a vector with d i.i.d. components $z_i \sim \mathcal{N}(0,1)$, the linear transformation
$$x = Lz + \mu,\ L \in \mathbb{R}^{d\times d}, \mu \in \mathbb{R}^d$$
follows a multivariate normal distribution $\mathcal{N}(\mu, \Sigma)$ with $\Sigma = LL^\top$. Prove also that any normal random vector can be written using this linear transformation with suitably chosen L and μ.

2. Using the above result, show that the sum of two *dependent* normal random variables $X_1 + X_2$, i.e. with $\mathrm{Cov}(X_1, X_2) \neq 0$ can always be written as the sum of two independent normal random variables $Y_1 + Y_2$.

3. The function `np.random.randn()` in Python can be seen as a generator that produces i.i.d. samples z_i according to a standard normal distribution $\mathcal{N}(0,1)$. Use this generator to sample $n = 1000$ values from a multivariate distribution $\mathcal{N}(\mu, \Sigma)$, with
$$\mu = \begin{bmatrix} 6 \\ -10 \end{bmatrix},\ \Sigma = \begin{bmatrix} 4 & -3 \\ -3 & 9 \end{bmatrix}$$
Plot the values on a 2D graph and comment the result.

Solution.

1. We can use the change of variable theorem for multivariate distributions
$$p_X(x) = p_Z(g^{-1}(x)) \left|\det J(g^{-1}(x))\right| \tag{33}$$
and apply it to the linear transformation
$$g(z) = Lz + \mu. \tag{34}$$
Let's do it step by step.

   (a) Find the inverse transformation $z = g^{-1}(x)$:
$$x = Lz + \mu \Rightarrow x - \mu = Lz \Rightarrow L^{-1}(x-\mu) = L^{-1}Lz \Rightarrow z = L^{-1}(x-\mu). \tag{35}$$

   (b) Calculate the Jacobian matrix for $g^{-1}(x)$
$$J(g^{-1}(x)) = \frac{\partial}{\partial x}\left[L^{-1}x - L^{-1}\mu\right] = L^{-1} \tag{36}$$
   and then its determinant
$$\det J(g^{-1}(x)) = \det L^{-1} = \frac{1}{\det L}. \tag{37}$$

   (c) Remember the PDF of d i.i.d. random variables with distribution $\mathcal{N}(0,1)$
$$p_Z(z) = \prod_{i=1}^{d} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z_i^2}{2}\right) = \frac{1}{\sqrt{2\pi})^d} \exp\left(-\frac{z^\top z}{2}\right) \tag{38}$$

   (d) Plug the inverse transformation $g^{-1}(x)$ inside the PDF of z and do some algebra tricks
$$p_Z(g^{-1}(x)) = \frac{1}{(\sqrt{2\pi})^d} \exp\left(-\frac{(g^{-1}(x))^\top (g^{-1}(x))}{2}\right) \tag{39}$$
$$= \frac{1}{(\sqrt{2\pi})^d} \exp\left(-\frac{(L^{-1}(x-\mu))^\top (L^{-1}(x-\mu))}{2}\right) \tag{40}$$
$$= \frac{1}{(\sqrt{2\pi})^d} \exp\left(-\frac{(x-\mu)^\top (L^{-1})^\top L^{-1}(x-\mu)}{2}\right) \tag{41}$$
$$= \frac{1}{(\sqrt{2\pi})^d} \exp\left(-\frac{(x-\mu)^\top (LL^T)^{-1}(x-\mu)}{2}\right). \tag{42}$$

(e) Now we just need to multiply the absolute value of the Jacobian's determinant and we are done

$$p_X(x) = \frac{1}{(\sqrt{2\pi})^d} \exp\left(-\frac{(x-\mu)^\top (LL^T)^{-1}(x-\mu)}{2}\right) \frac{1}{|\det L|}. \tag{43}$$

Notice that this is the PDF of a multivariate normal distribution

$$p_X(x) = \frac{1}{(\sqrt{2\pi})^d \sqrt{\det \Sigma}} \exp\left(-\frac{(x-\mu)^\top \Sigma^{-1}(x-\mu)}{2}\right) \tag{44}$$

with $\Sigma = LL^\top$, so q.e.d. I guess. Furthermore, as long as we find a matrix L such that $LL^\top = \Sigma$, we can express any normal random vector using a linear transformation of independent random variables. Since the covariance matrix Σ is always positive semidefinite, the matrix L can always be found using the Cholesky decomposition.

2. We have just proven that we can write any normal random vector as a linear transformation of i.i.d. standard normal random variables $\mathcal{N}(0,1)$ as $x = Lz + \mu$. This means that we can write

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} \ell_{11} & \ell_{12} \\ \ell_{21} & \ell_{22} \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} + \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \tag{45}$$

with i.i.d. z_1 and z_2 with distribution $\mathcal{N}(0,1)$ and suitably chosen ℓ_{ij} (if you want to be super nitpicking, you can also write $\ell_{12} = 0$ since the L is always lower-triangular, but we don't really care). This relationship can be written as the following system of equations

$$\begin{cases} x_1 = \ell_{11} z_1 + \ell_{12} z_2 + \mu_1 \\ x_2 = \ell_{21} z_1 + \ell_{22} z_2 + \mu_2 \end{cases}. \tag{46}$$

Finally, we can plug these expressions in the sum $X_1 + X_2$ and rearrange them

$$\begin{aligned} x_1 + x_2 &= \ell_{11} z_1 + \ell_{12} z_2 + \mu_1 + \ell_{21} z_1 + \ell_{22} z_2 + \mu_2 && (47)\\ &= (\ell_{11} + \ell_{21}) z_1 + \mu_1 + (\ell_{12} + \ell_{22}) z_2 + \mu_2 && (48)\\ &y_1 + y_2, && (49) \end{aligned}$$

with

$$\begin{cases} y_1 &= (\ell_{11} + \ell_{21}) z_1 + \mu_1 \\ y_2 &= (\ell_{12} + \ell_{22}) z_2 + \mu_2 \end{cases}. \tag{50}$$

Notice that y_1 and y_2 are independent since they are linear transformations of independent random variables. Furthermore, you can easily see that this result generalizes for the sum of any number of dependent normal random variables.

3. Please, see the notebook `02_multivariate_normal.ipynb`.