Sifakis Emmanouil, csdp1288 (16/03/2024)

Problem 1. *Change of variable formula.*
Let $X \sim Exp(\lambda)$ be an exponential random variable with parameter λ and let $Y = X^2$.

(a) Calculate analytically the probability density function (pdf), $f_Y(y)$, of Y using the change of variable formula.

Solution
We define the transformation $g(x) = x^2$ that has derivative $g'(x) = \frac{d}{dx}g(x)$, which is positive for $x > 0$ and negative for $x < 0$. This means that $g(x)$ strictly increases or decreases based on the sign of the input variable x accordingly. For this reason, this $g(x)$ cannot be considered monotonic which is necessary for continuing with the change of variable transformation. However, since our random variable X follows an exponential distribution, $X \sim \text{Exp}(\lambda)$, where λ is the rate parameter, and the distribution is defined for $x \geq 0$. This declaration inherently restricts our domain of interest to non-negative values of x. For this reason we are consered for the monotonicity of $g(x)$ in the positive domain.

Based on the change of variables rule if we apply a tranformation $Y = X^2 : g(x)$, $x \in \mathbb{R}^+$ to a random variable X with pdf: $f_X(x)$, then the pdf pf the transformed random variable, $f_Y(y)$ can be computed as follows:

$$f_Y(y) = f_X(g^{-1}(y)) \cdot \left| \frac{d}{dy} g^{-1}(y) \right| \tag{1}$$

as stated before:

$$\frac{d}{dx}g(x) = \frac{d}{dx}x^2 = 2x, \quad x \geq 0$$

We want to find $g^{-1}(x)$ such that $g^{-1}(g(x)) = x$. Let $y = g(x) = x^2$, solve for x in terms of y, we take the square root of both sides:

$$y = x^2 \implies \sqrt{y} = |x|$$

$$g^{-1}(y) = \sqrt{y}, \quad x \geq 0 \tag{2}$$

Having the $g^{-1}(y)$ (2), we can calculate the second factor of the main equation (1), $\left| \frac{d}{dy} g^{-1}(y) \right|$.

$$\left| \frac{d}{dy} g^{-1}(y) \right| = \left| \frac{d}{dy} \sqrt{y} \right| = \frac{1}{2} \left| y^{-1/2} \right|$$

$$\left| \frac{d}{dy} g^{-1}(y) \right| = \frac{1}{2} y^{-1/2}, \quad \text{since} \quad y^{-1/2} \quad \text{is always positive} \tag{3}$$

Next step is to calculate the first factor of the main equation (1), $f_X(g^{-1}(y))$. $f_X(x)$ represents the probability density function (pdf) of the random variable X , which follows an exponential distribution with parameter λ.

$$f_X(x) = \lambda e^{-\lambda x}, \quad x \geq 0$$

$$f_X(g^{-1}(y)) = f_X(\sqrt{y}) = \lambda e^{-\lambda\sqrt{y}} \tag{4}$$

Next step is to substituting these results (3,4) into Equation (1).

$$f_Y(y) = f_X(g^{-1}(y)) \cdot \left| \frac{d}{dy} g^{-1}(y) \right|$$

$$\boxed{f_Y(y) = \lambda e^{-\lambda\sqrt{y}} \cdot \frac{1}{2} y^{-1/2}} \tag{5}$$

(b) Compute the histogram of the dataset $\{y_i = g(x_i) : x_i \sim \text{Exp}(0.5)\}_{i=1}^{n}$ with $n = 100, 1000,$ and 10000. Plot in the same figure and compare the estimated histogram with $f_Y(y)$ from (a). What do you observe as n increases?

Solution

⇒ Utilization as a python script (notebook **CS673 ex1 - question b**)

(c) Repeat (b) using the dataset $\{y_i = F_Y^{-1}(u_i) : u_i \sim U(0,1)\}_{i=1}^{n}$ where $F_Y(y) = \int_{-\infty}^{y} f_Y(z)\,dz$ is the cumulative distribution function. You are allowed to use the function 'integrate()' of SymPy Python library for the estimation of the indefinite integral.

Solution

⇒ Utilization as a python script (notebook **CS673 ex1 - question c**)

Problem 2. *Multivariate Gaussian.*
Assume that $X = [X_1, X_2, X_3]^T \sim N(\mu, \Sigma)$ where μ is the mean vector and Σ is the covariance matrix.

(a) *Compute the pdf of $Y = X_2 + X_3$ and the pdf of $Z = [X_1, Y]$ assuming that both pdfs are Gaussians.*

Solution

In order to find the PDF of $Y = X_2 + X_3$, we need to find the mean and variance of Y, because we know that Gaussian functions are characterized by theri mean and covariance value.

$$\rightarrow \text{Mean of} \quad Y : E[Y] = E[X_2 + X_3] = E[X_2] + E[X_3] = \mu_2 + \mu_3 \tag{6}$$

(For the, above, Mean of Y the liniarity of expectation is used.)

We can compute the variance of $Y = X_2 + X_3$ using the definition of variance and properties of expectations.

$$\rightarrow \text{Variance of } Y : \text{Var}(Y) = E[(Y - E[Y])^2] \tag{7}$$

Now, we can expand $Y - E[Y]$ and compute its square:

$$Y - E[Y] = (X_2 + X_3) - (\mu_2 + \mu_3) = X_2 + X_3 - \mu_2 - \mu_3 \tag{8}$$

$$\begin{aligned} (Y - E[Y])^2 &= (X_2 + X_3 - \mu_2 - \mu_3)^2 \\ &= (X_2 - \mu_2)^2 + 2(X_2 - \mu_2)(X_3 - \mu_3) + (X_3 - \mu_3)^2 \end{aligned} \tag{9}$$

In this last equation $(X_2 - \mu_2)^2$ and $(X_3 - \mu_3)^2$ represent the squared deviation of X_2 and X_3 from their means μ_2 and μ_3. Also the third element $2(X_2 - \mu_2)(X_3 - \mu_3)$ represents the cross-product term, which captures the covariance between X_2 and X_3. By taking the expectation of this squared expression gives us the variance:

$$\text{Var}(Y) = E[(X_2 - \mu_2)^2] + E[(X_3 - \mu_3)^2] + 2E[(X_2 - \mu_2)(X_3 - \mu_3)] \tag{10}$$

Using the properties of expectations for a multivariate Gaussian distribution:
$E[(X_2 - \mu_2)^2] = \Sigma_{22}$, $E[(X_3 - \mu_3)^2] = \Sigma_{33}$, and
$E[(X_2 - \mu_2)(X_3 - \mu_3)] = \Sigma_{23}$ (or Σ_{32} since Σ is symmetric).
Therefore, the variance of Y simplifies to:

$$\text{Var}(Y) = \Sigma_{22} + 2\Sigma_{23} + \Sigma_{33} \tag{11}$$

$$\begin{aligned} \text{Var}(Y) &= \text{Var}(X_2 + X_3)\text{Var}(X_2) + \text{Var}(X_3) + 2\text{Cov}(X_2, X_3) \\ &= \Sigma_{22} + \Sigma_{33} + 2\Sigma_{23} \end{aligned} \tag{12}$$

$\rightarrow$ The following matrix represents the mean vector μ and the covariance matrix of Y.

$$\mu_Y = \begin{bmatrix} \mu_2 \\ \mu_3 \end{bmatrix}, \qquad \Sigma_Y = \begin{bmatrix} \Sigma_{22} & \Sigma_{23} \\ \Sigma_{32} & \Sigma_{33} \end{bmatrix}$$

Since Y is a linear combination of Gaussian random variables X_2 and X_3, it will also follow a Gaussian distribution.
So, the PDF of Y will be:

$$f_Y(y) = \frac{1}{\sqrt{2\pi \text{Var}(Y)}} \exp\left(-\frac{(y - E[Y])^2}{2\text{Var}(Y)}\right)$$

$$\boxed{f_Y(y) = \frac{1}{\sqrt{2\pi(\Sigma_{22} + 2\Sigma_{23} + \Sigma_{33})}} \exp\left(-\frac{(y - (\mu_2 + \mu_3))^2}{2(\Sigma_{22} + 2\Sigma_{23} + \Sigma_{33})}\right)} \tag{13}$$

For the joint distribution $Z = [X_1, Y]$, we know that X_1 and Y are jointly Gaussian. Since the covariance matrix Σ captures the covariance between X_1 and X_2, and X_2 and X_3, but not between X_1 and X_3, we can say that X_1 and Y are also jointly Gaussian with mean vector μ and covariance matrix Σ:

$$\mu_Z = \begin{bmatrix} \mu_1 \\ \mu_2 + \mu_3 \end{bmatrix}, \qquad \Sigma_Z = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} + 2\Sigma_{23} + \Sigma_{33} \end{bmatrix} \tag{14}$$

Thus, the joint PDF of $Z = [X_1, Y]$ is also Gaussian.

$$f_Z(z) = \frac{1}{\sqrt{(2\pi)^2 |\Sigma_Z|}} \exp\left(-\frac{1}{2}(z - \mu_Z)^T \Sigma_Z^{-1} (z - \mu_Z)\right)$$

$$\boxed{f_Z(z) = \frac{1}{2\pi\sqrt{(\Sigma_{11})(\Sigma_{22} + 2\Sigma_{23} + \Sigma_{33})}} \exp\left(-\frac{1}{2}\left[\frac{(x_1 - \mu_1)^2}{\Sigma_{11}} + \frac{(y - (\mu_2 + \mu_3))^2}{\Sigma_{22} + 2\Sigma_{23} + \Sigma_{33}}\right]\right)} \tag{15}$$

(b) Compute the conditional pdf:

$$p(x1|x2 + x3 = 0)$$

Solution

For the computation of the conditional PDF $p(x_1|x_2 + x_3 = 0)$, the properties of multivariate Gaussian distribution will be used. Given that $X = [X_1, X_2, X_3]^T$ follows a multivariate Gaussian distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, the joint PDF of X_1, X_2, and X_3 is a multivariate Gaussian.

The conditional distribution $p(x_1|x_2 + x_3 = 0)$ is derived by focusing only on the points in the joint distribution where the sum of $x_2 + x_3 = 0$.

We can represent this condition in terms of the joint distribution as follows. Using Bayes' theorem $p(A|B) = \frac{p(B|A) \cdot p(A)}{p(B)}$ allows us to compute conditional probabilities. In this case, to find the conditional probability density function $p(x_1|x_2 + x_3 = 0)$.

$$p(x_1|x_2 + x_3 = 0) = \frac{p(x_2 + x_3 = 0|x_1) \cdot p(x_1)}{p(x_2 + x_3 = 0)}$$

However, $p(x_2 + x_3 = 0|x_1)$ is not directly available, but we can compute it using the joint distribution $p(x_1, x_2, x_3)$ and marginal distributions. By rearranging terms, we obtain:

$$p(x_1|x_2 + x_3 = 0) = \frac{p(x_1, x_2, x_3)}{p(x_2 + x_3 = 0)} \tag{16}$$

Where:
$\rightarrow p(x_1, x_2, x_3)$ is the joint PDF of X_1, X_2, and X_3.
$\rightarrow p(x_2 + x_3 = 0)$ is the marginal PDF of $X_2 + X_3$ evaluated at 0.

Since $X_2 + X_3$ follows a Gaussian distribution, and its mean is $\mu_2 + \mu_3$ and variance is $\Sigma_{22} + 2\Sigma_{23} + \Sigma_{33}$ the conditional distribution becomes:

$$\boxed{p(x_1|x_2 + x_3 = 0) = \frac{p(x_1, x_2, x_3)}{\sqrt{2\pi(\Sigma_{22} + 2\Sigma_{23} + \Sigma_{33})}} \exp\left(-\frac{(x_2 + x_3 - (\mu_2 + \mu_3))^2}{2(\Sigma_{22} + 2\Sigma_{23} + \Sigma_{33})}\right)} \tag{17}$$

This above involves the joint PDF $p(x_1, x_2, x_3)$, typically a multivariate Gaussian distribution.

(c)*Solution*
$\Rightarrow$ Utilization as a python script (notebook **CS673 ex2 - question c**)

Problem 3. *Maximum likelihood estimation. Generate and infer the parameters of an autoregressive (AR) process.*

(a)*Simulate an AR(1) process which is given by the formula*

$$x_t = a_0 + a_1 x_{t-1} + w_t, \quad t = 0, 1, 2, \ldots, T-1$$

where w_t is white noise (i.e., $w_t \sim N(0, \sigma^2)$ for all t and w_t is independent of $w_{t'}$ for all t, t' with $t \neq t'$), $\sigma = 1.0$, $a_0 = 2.0$, $a_1 = -0.9$, $x_{-1} = 0$, and $T = 1000$.

Solution

$\Rightarrow$ Utilization as a python script (notebook **CS673_ex3 - question a**)

(b) Write down the log-likelihood of the above AR(1) process for the parameter vector $\theta = [a_0, a_1]^T$.

Solution

To write down the log-likelihood of the AR(1) process for the parameter vector $\theta = [a_0, a_1]^T$, the probability density function (pdf) of the white noise w_t must be defined. Given that w_t follows a normal distribution with mean 0 and variance σ^2, its probability density function (pdf) is:

$$f(w_t) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{w_t^2}{2\sigma^2}\right)$$

The log-likelihood of the AR(1) process is the logarithm of the joint probability density function of the observed data $x_0, x_1, \ldots, x_{T-1}$ given the parameters $\theta = [a_0, a_1]^T$. Since x_t is generated by the AR(1) process, the conditional probability density function of x_t given x_{t-1} θ can be expressed as:

$$f(x_t | x_{t-1}, \theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_t - a_0 - a_1 x_{t-1})^2}{2\sigma^2}\right)$$

Given that $x_0 = 0$, the joint probability density function of the observed data can be written as:

$$L(\theta) = \prod_{t=1}^{T-1} f(x_t | x_{t-1}, \theta)$$

And the log-likelihood can be written as:

$$\mathcal{L}(\theta) = \log L(\theta) = \sum_{t=1}^{T-1} \log f(x_t | x_{t-1}, \theta)$$

A closed-form expression for the log-likelihood of the AR(1) process, starts by substituting the expression for $f(x_t | x_{t-1}, \theta)$ into the log-likelihood equation and simplifying the expression:

$$\mathcal{L}(\theta) = \sum_{t=1}^{T-1} \log\left(\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_t - a_0 - a_1 x_{t-1})^2}{2\sigma^2}\right)\right)$$

To reach a closed-form expression for the log-likelihood $\mathcal{L}(\theta)$, the logarithm of the exponential term is simplified.

$$\mathcal{L}(\theta) = \sum_{t=1}^{T-1} \log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) + \sum_{t=1}^{T-1}\left(-\frac{(x_t - a_0 - a_1 x_{t-1})^2}{2\sigma^2}\right)$$

The first term inside the logarithm is a constant and so can be excluded from the sum:

$$\mathcal{L}(\theta) = (T-1)\log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) + \sum_{t=1}^{T-1}\left(-\frac{(x_t - a_0 - a_1 x_{t-1})^2}{2\sigma^2}\right)$$

We can further simplify the constant term:

$$\mathcal{L}(\theta) = (T-1)\left(-\frac{1}{2}\log(2\pi\sigma^2)\right) + \sum_{t=1}^{T-1}\left(-\frac{(x_t - a_0 - a_1 x_{t-1})^2}{2\sigma^2}\right)$$

The final closed-form expression for the log-likelihood $\mathcal{L}(\theta)$ of the AR(1) process in terms of the observed data x_t and the parameters $\theta = [a_0, a_1]^T$:

$$\boxed{\mathcal{L}(\theta) = -\frac{T-1}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{t=1}^{T-1}(x_t - a_0 - a_1 x_{t-1})^2}$$

(c) Compute analytically and then numerically using the simulated process from (a), the maximum likelihood estimator. Plot the mean squared error between the numerically estimated $\hat{\theta}_{MLE}$ and the ground truth as a function of T.

In order to find ,analyticaly, the maximum likelihood estimators of a_0 and a_1 using the log-likelihood function derived above, the partial derivatives of $\mathcal{L}(\theta)$ with respect to a_0 and a_1 should be taken:

⇒ Partial derivative of $\mathcal{L}(\theta)$ with respect to a_0:

$$\frac{\partial\mathcal{L}(\theta)}{\partial a_0} = \frac{\partial}{\partial a_0}\left[-\frac{T-1}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{t=1}^{T-1}(x_t - a_0 - a_1 x_{t-1})^2\right]$$

$$= 0 - \frac{1}{\sigma^2}\sum_{t=1}^{T-1}(1)(x_t - a_0 - a_1 x_{t-1}) \tag{18}$$

$$= -\frac{1}{\sigma^2}\sum_{t=1}^{T-1}(x_t - a_0 - a_1 x_{t-1})$$

⇒ Partial derivative of $\mathcal{L}(\theta)$ with respect to a_1:

$$\frac{\partial \mathcal{L}(\theta)}{\partial a_1} = \frac{\partial}{\partial a_1}\left[-\frac{T-1}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{t=1}^{T-1}(x_t - a_0 - a_1 x_{t-1})^2\right]$$

$$= 0 - \frac{1}{\sigma^2}\sum_{t=1}^{T-1}(-x_{t-1})(x_t - a_0 - a_1 x_{t-1}) \tag{19}$$

$$= -\frac{1}{\sigma^2}\sum_{t=1}^{T-1} x_{t-1}(x_t - a_0 - a_1 x_{t-1})$$

To find the values of a_0 and a_1 that maximize the log-likelihood function the partial derivatives mast be set equal to zero:

$$\frac{\partial \mathcal{L}(\theta)}{\partial a_0} = \frac{1}{\sigma^2}\sum_{t=1}^{T-1}(x_t - a_0 - a_1 x_{t-1}) = 0$$

$$\frac{\partial \mathcal{L}(\theta)}{\partial a_1} = \frac{1}{\sigma^2}\sum_{t=1}^{T-1} x_{t-1}(x_t - a_0 - a_1 x_{t-1}) = 0$$

⇒ For the first equation (18):

$$\frac{1}{\sigma^2}\sum_{t=1}^{T-1}(x_t - a_0 - a_1 x_{t-1}) = 0$$

$$\frac{1}{\sigma^2}\left(\sum_{t=1}^{T-1} x_t - \sum_{t=1}^{T-1} a_0 - \sum_{t=1}^{T-1} a_1 x_{t-1}\right) = 0$$

$$\frac{1}{\sigma^2}\left(S_x - (T-1)a_0 - \sum_{t=1}^{T-1} a_1 x_{t-1}\right) = 0$$

Where the new term S_x is defined as $S_x = \sum_{t=1}^{T-1} x_t$. If a_0 be isolated:

$$a_0 = \frac{1}{T-1}\left(S_x - \sum_{t=1}^{T-1} a_1 x_{t-1}\right) \tag{20}$$

⇒ For the second equation (19), if the expression derived for a_0 is used:

$$\frac{1}{\sigma^2}\sum_{t=1}^{T-1} x_{t-1}(x_t - a_0 - a_1 x_{t-1}) = 0$$

$$\frac{1}{\sigma^2}\sum_{t=1}^{T-1} x_{t-1}\left(x_t - \frac{1}{T-1}\left(S_x - \sum_{t=1}^{T-1} a_1 x_{t-1}\right) - a_1 x_{t-1}\right) = 0$$

Analysis further one gets:

$$\frac{1}{\sigma^2}\sum_{t=1}^{T-1} x_{t-1}\left(x_t - \frac{S_x}{T-1} + \frac{1}{T-1}\sum_{t=1}^{T-1} a_1 x_{t-1} - a_1 x_{t-1}\right) = 0$$

Combining terms with a_1 and rearranging:

$$\frac{1}{\sigma^2}\sum_{t=1}^{T-1} x_{t-1}\left(x_t - \frac{S_x}{T-1} + \frac{a_1}{T-1}\sum_{t=1}^{T-1} x_{t-1}\right) = 0$$

Expanding the sum:

$$\frac{1}{\sigma^2}\left(\sum_{t=1}^{T-1} x_{t-1}x_t - \frac{S_x}{T-1}\sum_{t=1}^{T-1} x_{t-1} + \frac{a_1}{T-1}\sum_{t=1}^{T-1} x_{t-1}\sum_{t=1}^{T-1} x_{t-1}\right) = 0 \tag{21}$$

Knowing that:

$$\Rightarrow \sum_{t=1}^{T-1} x_{t-1}x_t = \sum_{t=1}^{T-1} x_{t-1}x_t \quad \text{(Autocovariance)}$$

$$\Rightarrow \sum_{t=1}^{T-1} x_{t-1} = \sum_{t=0}^{T-2} x_t = S_x - x_{T-1} \quad \text{(Summation)}$$

The equation (21) becomes:

$$\frac{1}{\sigma^2}\left(\sum_{t=1}^{T-1} x_{t-1}x_t - \frac{S_x(S_x - x_{T-1})}{T-1} + \frac{a_1}{T-1}\left(S_x - x_{T-1}\right)\left(S_x - x_{T-1}\right)\right) = 0$$

After some algebraic simplification:

$$\sum_{t=1}^{T-1} x_{t-1}x_t - \frac{S_x^2 - S_x x_{T-1}}{T-1} + \frac{a_1}{T-1}\left(S_x^2 - 2S_x x_{T-1} + x_{T-1}^2\right) = 0$$

$$(T-1)\sum_{t=1}^{T-1} x_{t-1}x_t - (S_x^2 - S_x x_{T-1}) + a_1\left(S_x^2 - 2S_x x_{T-1} + x_{T-1}^2\right) = 0$$

$$a_1\left(S_x^2 - 2S_x x_{T-1} + x_{T-1}^2\right) + (T-1)\sum_{t=1}^{T-1} x_{t-1}x_t + S_x x_{T-1} - S_x^2 = 0$$

$$\boxed{a_1 = \frac{S_x^2 - S_x x_{T-1} - (T-1)\sum_{t=1}^{T-1} x_{t-1}x_t}{S_x^2 - 2S_x x_{T-1} + x_{T-1}^2}} \tag{22}$$

This expression gives the maximum likelihood estimator for a_1.
⇒ The expression of a_1 will be used to find a_0. Essentially by substituting a_1 into the equation derived for a_0,(20) gives:

$$a_0 = \frac{1}{T-1}\left(S_x - \sum_{t=1}^{T-1}\frac{S_x^2 - S_x x_{T-1} - (T-1)\sum_{t=1}^{T-1} x_{t-1}x_t}{S_x^2 - 2S_x x_{T-1} + x_{T-1}^2} x_{t-1}\right)$$

Distribute the summation and factor out the common terms in the denominator.

$$\begin{aligned}
a_0 &= \frac{1}{T-1}\left(S_x - \sum_{t=1}^{T-1} \frac{S_x^2 x_{t-1} - S_x x_{T-1} x_{t-1} - (T-1)\sum_{t=1}^{T-1} x_{t-1}^2 x_t}{(S_x - x_{T-1})^2}\right) \\
&= \frac{1}{T-1}\left(S_x - \sum_{t=1}^{T-1} \frac{S_x^2 x_{t-1} - S_x x_{T-1} x_{t-1} - (T-1)\sum_{t=1}^{T-1} x_{t-1}^2 x_t}{(S_x - x_{T-1})^2}\right)
\end{aligned}$$

$$\begin{aligned}
a_0 &= \frac{1}{T-1}\left(S_x - \frac{S_x^2 \sum_{t=1}^{T-1} x_{t-1} - S_x x_{T-1} \sum_{t=1}^{T-1} x_{t-1} - (T-1)\sum_{t=1}^{T-1} x_{t-1}^2 x_t}{(S_x - x_{T-1})^2}\right) \\
&= \frac{1}{T-1}\left(S_x - \frac{S_x^2 (S_x - x_{T-1}) - S_x (S_x - x_{T-1})^2 - (T-1)\sum_{t=1}^{T-1} x_{t-1}^2 x_t}{(S_x - x_{T-1})^2}\right) \\
&= \frac{1}{T-1}\left(S_x - \frac{S_x^3 - 2S_x^2 x_{T-1} + S_x x_{T-1}^2 - S_x^3 + 2S_x^2 x_{T-1} - S_x x_{T-1}^2 - (T-1)\sum_{t=1}^{T-1} x_{t-1}^2 x_t}{(S_x - x_{T-1})^2}\right) \\
&= \frac{1}{T-1}\left(S_x - \frac{S_x + (T-1)\sum_{t=1}^{T-1} x_{t-1}^2 x_t}{(S_x - x_{T-1})^2}\right)
\end{aligned}$$

The final expression for a_0 is:

$$\boxed{a_0 = \frac{1}{T-1}\left(S_x - \frac{S_x + (T-1)\sum_{t=1}^{T-1} x_{t-1}^2 x_t}{(S_x - x_{T-1})^2}\right)}$$

$\Rightarrow$ Numerical utilization as a python script (notebook **CS673 ex3 - question c**)

Problem 4. *Gaussian Mixture Model (GMM) with prior.*
(a) *You will derive the Expectation-Maximization (EM) algorithm when prior knowledge regarding the mean values is available. Let π, $\{\mu_k\}_{k=1}^K$, $\{\Sigma_k\}_{k=1}^K$ be the parameters of a Gaussian Mixture Model (GMM) with K Gaussians and data dimension d. Moreover, assume that each μ_k is independently sampled from a Gaussian prior, $\mu_k \sim \mathcal{N}(\mu_{0k}, \lambda^{-1}I_d)$, $k = 1, \ldots, K$, where μ_{0k} is the prior mean vector while λ is the inverse variance and it is interpreted as the strength of the prior (e.g., larger values for λ implies stronger prior). We assume no prior information regarding the weights, π, and the covariance matrices, $\{\Sigma_k\}_{k=1}^K$. Repeat the derivation steps of the EM algorithm starting from the maximization of the logarithm of the posterior distribution:*

$$p(\pi, \{\mu_k\}_{k=1}^K, \{\Sigma_k\}_{k=1}^K|x) \propto p(x|\pi, \{\mu_k\}_{k=1}^K, \{\Sigma_k\}_{k=1}^K) \times p(\{\mu_k\}_{k=1}^K)$$

where $p(\{\mu_k\}_{k=1}^K)$ is the Gaussian prior distribution for the mean vectors.
Hint: Only the formula for the mean vectors will be different.

Solution Given the prior knowledge on the mean vectors, the maximization step involves finding the parameters that maximize the expected complete-data log-likelihood, which includes the prior term. For a single Gaussian component k, the log of the posterior distribution, taking the prior into account, can be written as:

$$\log p(\pi, \mu_k, \Sigma_k|x) = \log p(x|\pi, \mu_k, \Sigma_k) + \log p(\mu_k)$$

The $Q(\theta|\theta^{(t)})$ will describe the complete-data log-likelihood, where θ represents the parameters of interest (π, $\{\mu_k\}$, $\{\Sigma_k\}$), and $\theta^{(t)}$ represents the current parameter estimates at the t-th iteration. The maximization step involves maximizing $Q(\theta|\theta^{(t)})$ with respect to μ_k.

$$\frac{\partial Q}{\partial \mu_k} = \frac{\partial}{\partial \mu_k}\left(\sum_{i=1}^{N}\sum_{j=1}^{K}\gamma_{ij}^{(t)} \log\left(\mathcal{N}(x_i|\mu_j, \Sigma_j)\right) + \log p(\mu_k)\right)$$

Where $\gamma_{ij}^{(t)} = p(z_i = k|x_i, \theta^{(t)})$, and $\mathcal{N}(x_i|\mu_j, \Sigma_j)$ is the Gaussian distribution with mean μ_j and covariance Σ_j.
$\rightarrow$ The likelihood term is denoted as $\mathcal{L}_k = \sum_{i=1}^N \gamma_{ik}^{(t)}(x_i - \mu_k)$
$\rightarrow$ The prior term as $P_k = \lambda(\mu_k - \mu_{0k})$. Where λ is the inverse variance, and μ_{0k} is the prior mean vector.
The update rule for μ_k is obtained by setting the derivative equal to zero and solving for μ_k:

$$\frac{\partial Q}{\partial \mu_k} = 0\frac{\partial}{\partial \mu_k}\left(\mathcal{L}_k + P_k\right) = 0$$

$$\Rightarrow \sum_{i=1}^{N}\gamma_{ik}^{(t)}(x_i - \mu_k) + \lambda(\mu_k - \mu_{0k}) = 0$$

$$\sum_{i=1}^{N} \gamma_{ik}^{(t)} x_i - \sum_{i=1}^{N} \gamma_{ik}^{(t)} \mu_k + \lambda \mu_k - \lambda \mu_{0k} = 0$$

$$\left(\sum_{i=1}^{N} \gamma_{ik}^{(t)} \right) \mu_k + \lambda \mu_k = \sum_{i=1}^{N} \gamma_{ik}^{(t)} x_i + \lambda \mu_{0k}$$

$$\left(\sum_{i=1}^{N} \gamma_{ik}^{(t)} + \lambda \right) \mu_k = \sum_{i=1}^{N} \gamma_{ik}^{(t)} x_i + \lambda \mu_{0k}$$

Solving for μ_k:

$$\boxed{\mu_k = \frac{\sum_{i=1}^{N} \gamma_{ik}^{(t)} x_i + \lambda \mu_{0k}}{\sum_{i=1}^{N} \gamma_{ik}^{(t)} + \lambda}}$$

This is the update rule for μ_k in the presence of prior knowledge about the mean vectors.

(b) Generate $n = 1000$ samples from a GMM with $K = 3$ components using the ancestral sampling algorithm. The mean vectors of the three equiprobable Gaussian components are $\mu_1 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$, $\mu_2 = \begin{bmatrix} 1 \\ -0.5 \end{bmatrix}$, and $\mu_3 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$ while the respective covariance matrices being

$$\Sigma_1 = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 0.9 \end{bmatrix}, \quad \Sigma_2 = \begin{bmatrix} 1 & -0.8 \\ -0.8 & 1.1 \end{bmatrix}, \quad \Sigma_3 = \begin{bmatrix} 1.5 & 1.3 \\ 1.3 & 1 \end{bmatrix}.$$

Solution

⇒ Utilization as a python script (notebook **CS673_ex4 - question b**)

(c) Use the equations derived in (a) and the data from (b) to estimate the parameters of the GMM. Consider three cases:

i) Few data with strong correct prior (e.g., $n \approx 100$ or less, $\mu_{0k} \approx \mu_k$ and $\lambda = O(10^3)$),

ii) Few data with strong wrong prior (e.g., $n \approx 100$ or less, $\mu_{0k} \approx \mu_k + 1$ and $\lambda = O(10^3)$),

iii) Many data with strong wrong prior (e.g., $n \approx 10^4$, $\mu_{0k} \approx \mu_k + 1$ and $\lambda = O(10^3)$).

Solution

⇒ Utilization as a python script (notebook **CS673_ex4 - question c**)

Problem 5. *Evidence lower bound (ELBO).*
(a)*Let $p(x,z)$ be the joint PDF, $p(x)$ be the marginal PDF (or evidence), and $p(z|x)$ be the posterior PDF. Assume also another conditional PDF denoted by $q(z|x)$. For all x , prove that:*

$$\log p(x) = \mathbb{E}_{q(z|x)}\left[\log \frac{p(x|z)}{q(z|x)}\right] + DKL\left(q(z|x)||p(z|x)\right) \tag{23}$$

Solution

In order to prove equation (18) one can rewrite $\text{DKL}\left(q(z|x)||p(z|x)\right)$ in terms of expectations and $\log p(x)$, we can start with the definition of the Kullback-Leibler (KL) divergence:

$$\text{DKL}\left(q(z|x)||p(z|x)\right) = \mathbb{E}_{q(z|x)}\left[\log \frac{q(z|x)}{p(z|x)}\right] = \mathbb{E}_{q(z|x)}\left[\log q(z|x) - \log p(z|x)\right]$$

Becasue of the linearity of expectation:

$$\text{DKL}\left(q(z|x)||p(z|x)\right) = \mathbb{E}_{q(z|x)}\left[\log q(z|x)\right] - \mathbb{E}_{q(z|x)}\left[\log p(z|x)\right]$$

The aim is to express $\text{DKL}\left(q(z|x)||p(z|x)\right)$ in terms of $\log p(x)$ and expectation. By using Bayes' rule to express $p(z|x)$ in terms of $p(x|z)$:

$$p(z|x) = \frac{p(x|z)\cdot p(z)}{p(x)} \xrightarrow{\text{taking log on both sides}} \log p(z|x) = \log p(x|z) + \log p(z) - \log p(x)$$

Substituting this into the expression for $\text{DKL}\left(q(z|x)||p(z|x)\right)$, we get:

$$\begin{aligned}
\text{DKL}\left(q(z|x)||p(z|x)\right) &= \mathbb{E}_{q(z|x)}\left[\log q(z|x)\right] - \mathbb{E}_{q(z|x)}\left[\log p(x|z) + \log p(z) - \log p(x)\right] \\
&= \mathbb{E}_{q(z|x)}\left[\log q(z|x)\right] - \mathbb{E}_{q(z|x)}\left[\log p(x|z)\right] - \mathbb{E}_{q(z|x)}\left[\log p(z)\right] + \mathbb{E}_{q(z|x)}\left[\log p(x)\right] \\
&= \mathbb{E}_{q(z|x)}\left[\log \frac{q(z|x)}{p(x|z)}\right] + \log p(x) - \underbrace{\mathbb{E}_{q(z|x)}\left[\log p(z)\right]}
\end{aligned}$$

As it can be observed the proof is almost ready apart from the term $\mathbb{E}_{q(z|x)}\left[\log p(z)\right]$. This term represents the expected value of the logarithm of $p(z)$ with respect to the distribution $q(z|x)$. To compute this expectation simply integrate over all possible values of z, weighted by their probabilities under $q(z|x)$:

$$\mathbb{E}_{q(z|x)}\left[\log p(z)\right] = \int q(z|x)\log p(z)\,dz$$

Knowing this it is possible to:

$$\text{DKL}\left(q(z|x)||p(z|x)\right) = \mathbb{E}_{q(z|x)}\left[\log \frac{q(z|x)}{p(x|z)}\right] + \log p(x) - \left(\int q(z|x)\log p(z)\,dz\right)$$

The term $\int q(z|x)\log p(z)\,dz$ is a constant term with respect to z . Therefore, it can be combined with the constant term $\log p(x)$. So, rewriting it in a more clear form:

$$\log p(x) = \mathbb{E}_{q(z|x)}\left[\log \frac{p(x|z)}{q(z|x)}\right] + \text{DKL}\left(q(z|x)||p(z|x)\right)$$

(b) Using the above formula, prove the evidence lower bound for the GMM case, which reads:

$$\log p_\theta(x) \geq \mathbb{E}_{p_{\theta^{\text{old}}}(z|x)}\left[\log p_\theta(x,z)\right] - \mathbb{E}_{p_{\theta^{\text{old}}}(z|x)}\left[p_{\theta^{\text{old}}}(z|x)\right]$$

Solution

To derive the evidence lower bound (ELBO) for the Gaussian Mixture Model (GMM) case, the logical step is to start with the general formula for ELBO:

$$\log p_\theta(x) = \mathbb{E}_{q(z|x)}\left[\log \frac{p(x,z)}{q(z|x)}\right] + \text{DKL}(q(z|x)||p(z|x))$$

For the GMM case the $q(z|x)$ can be defined as $q(z|x) = p_{\theta^{\text{old}}}(z|x)$, representing the posterior in relation to the previous model. Substituting this into the ELBO formula, one gets:

$$\log p_\theta(x) = \mathbb{E}_{p_{\theta^{\text{old}}}(z|x)}\left[\log \frac{p_\theta(x,z)}{p_{\theta^{\text{old}}}(z|x)}\right] + \text{DKL}(p_{\theta^{\text{old}}}(z|x)||p_\theta(z|x))$$

The KL divergence term $\text{DKL}(p_{\theta^{\text{old}}}(z|x)||p_\theta(z|x))$ is non-negative, so it can be dismissed in order to obtain a lower bound:

$$\log p_\theta(x) \geq \mathbb{E}_{p_{\theta^{\text{old}}}(z|x)}\left[\log \frac{p_\theta(x,z)}{p_{\theta^{\text{old}}}(z|x)}\right]$$

$$\log p_\theta(x) \geq \mathbb{E}_{p_{\theta^{\text{old}}}(z|x)}\left[\log p_\theta(x,z)\right] - \mathbb{E}_{p_{\theta^{\text{old}}}(z|x)}\left[p_{\theta^{\text{old}}}(z|x)\right]$$

This lower bound gives an expression involving the joint log-likelihood and the reference model's posterior distribution, providing a bound on the marginal likelihood $\log p_\theta(x)$.