CS 189 Introduction to Machine Learning
Fall 2017 DIS4

1 Multivariate Gaussians: A review

(a) Consider a two dimensional random variable $Z \in \mathbb{R}^2$. In order for the random variable to be jointly Gaussian, a necessary and sufficient condition is that

- Z_1 and Z_2 are each marginally Gaussian, and
- $Z_1|Z_2 = z$ is Gaussian, and $Z_2|Z_1 = z$ is Gaussian.

A second characterization of a jointly Gaussian RV Z is that it can be written as $Z = AX$, where X is a collection of i.i.d. standard normal RVs and $A \in \mathbb{R}^{2\times 2}$ is a matrix.

Let X_1 and X_2 be i.i.d. standard normal RVs. Let U denote a random variable uniformly distributed on $\{-1, 1\}$, independent of everything else. Verify if the conditions of the first characterization hold for the following random variables, and calculate the covariance matrix Σ_Z.

- $Z_1 = X_1$ and $Z_2 = X_2$.
- $Z_1 = X_1$ and $Z_2 = X_1 + X_2$. (Use the second characterization to argue joint Gaussianity.)
- $Z_1 = X_1$ and $Z_2 = -X_1$.
- $Z_1 = X_1$ and $Z_2 = UX_1$.

Solution: Before diving into the solution, recall that the covariance matrix of a vector random variable X with mean (vector) μ is given by $\Sigma = \mathbb{E}[(X-\mu)(X-\mu)^\top]$. In other words, entry i, j of the covariance matrix denotes the covariance between the random variables X_i and X_j, i.e., $\Sigma_{ij} = \mathsf{cov}(X_i, X_j) = \mathbb{E}[(X_i - \mu_i)(X_j - \mu_j)]$.

Additionally, two random variables U and V are said to be uncorrelated if $\mathsf{cov}(U,V) = 0$

- Z_1 and Z_2 are i.i.d. standard Gaussian, and so $(Z_1|Z_2 = z) \sim N(0,1)$. Also, $Z_2|Z_1 = z \sim N(0,1)$. Hence, the RVs are jointly Gaussian. We also have $\Sigma_Z = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$.
- $Z_1 \sim N(0,1)$, and $Z_2 \sim N(0,2)$, but these RVs are not independent. Also, we have $(Z_2|Z_1 = z) \sim N(z,1)$. In order to calculate the distribution of $(Z_1|Z_2 = z)$, see part (e). Using the second characterization of joint Gaussianity, it is clear that Z is jointly Gaussian, with $A = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}$. The covariance matrix is given by $\Sigma_Z = \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix}$.
- We have $Z_1 \sim N(0,1)$ and $Z_2 \sim N(0,1)$ marginally. However, we have $(Z_1|Z_2 = z) \sim N(-z, 0)$, which is a degenerate Gaussian. The other conditional distribution is identical. Hence, the RVs are jointly Gaussian. The covariance matrix is given by $\Sigma_Z = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}$.

- As before, we have $Z_1 \sim N(0,1)$ and $Z_2 \sim N(0,1)$ marginally. In order to see this, write

$$\begin{aligned} p(Z_2 = z_2) &= p(Z_2 = z_2 | U = 1) p(U = 1) + p(Z_2 = z_2 | U = -1) p(U = -1) \\ &= \frac{1}{2} p(X_1 = z_2 | U = 1) + \frac{1}{2} p(X_2 = -z_2 | U = -1). \end{aligned}$$

The random variable $(Z_2 | Z_1 = z)$ is uniformly distributed on $\{-z, z\}$, and is therefore not Gaussian. The RVs are therefore not jointly Gaussian. The covariance matrix is given by $\Sigma_Z = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$.

(b) Use the above example to show that two Gaussian random variables can be uncorrelated, but not independent. On the other hand, show that two uncorrelated, jointly Gaussian RVs are independent.

Solution: The last example in the previous part shows uncorrelated Gaussians that are not independent. In order to show that jointly Gaussian RVs (with individual variances σ_1^2 and σ_2^2) that are uncorrelated are also independent, assume without loss of generality that the RVs have zero mean, and notice that one can write the joint pdf as

$$\begin{aligned} f_Z(z_1, z_2) &= \frac{1}{(2\pi)\det(\Sigma_Z^{1/2})} \exp\left(-\frac{1}{2} \begin{bmatrix} z_1 & z_2 \end{bmatrix} (\Sigma_Z)^{-1} \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} \right) \\ &= \frac{1}{\sqrt{2\pi\sigma_1^2}} \frac{1}{\sqrt{2\pi\sigma_2^2}} \exp\left(-\frac{1}{2\sigma_1^2} z_1^2 \right) \exp\left(-\frac{1}{2\sigma_1^2} z_2^2 \right) \\ &= f_{Z_1}(z_1) f_{Z_2}(z_2). \end{aligned}$$

The decomposition follows since Σ_Z is a diagonal matrix when the RVs are uncorrelated. Since we have expressed the joint PDF as a product of the individual PDFs, the RVs are independent.

(c) With the setup above, let $Z = VX$, where $V \in \mathbb{R}^{2\times 2}$, and $Z, X \in \mathbb{R}^2$. What is the covariance matrix Σ_Z?

Solution: The covariance matrix of a random vector Z (by definition) is given by $\mathbb{E}(Z - \mathbb{E}[Z])(Z - \mathbb{E}[Z])^\top$. Since the mean $\mathbb{E}[Z]$ is 0, we may write $\Sigma_Z = \mathbb{E}[VXX^\top V^\top] = V\mathbb{E}[XX^\top]V^\top = VV^\top$. This follows by linearity of expectation applied to vector random variables (write it out to convince yourself!)

(d) Use the above setup to show that $X_1 + X_2$ and $X_1 - X_2$ are independent. Give another example pair of linear combinations that are independent.

Solution: By our previous arguments, it is sufficient to show that these are uncorrelated. Calculating the covariance matrix, we have $\Sigma = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$, which is diagonal. Any linear combination $Z = VX$ with $VV^\top = D$ for a diagonal matrix D results in uncorrelated random variables.

(e) Given a jointly Gaussian RV $Z \in \mathbb{R}^2$ with covariance matrix $\Sigma_Z = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12} & \Sigma_{22} \end{bmatrix}$, how would you derive the distribution of $Z_1 | Z_2 = z$?

Hint: The following identity may be useful

$$\begin{bmatrix} a & b \\ b & c \end{bmatrix}^{-1} = \begin{bmatrix} 1 & 0 \\ -\frac{b}{c} & 1 \end{bmatrix} \begin{bmatrix} \left(a - \frac{b^2}{c}\right)^{-1} & 0 \\ 0 & \frac{1}{c} \end{bmatrix} \begin{bmatrix} 1 & -\frac{b}{c} \\ 0 & 1 \end{bmatrix}.$$

Solution: One can do this from first principles, by manipulating the densities themselves. However, we will show a linear algebraic method to derive the density. Using the hint, we begin by writing

$$\Sigma^{-1} = \begin{bmatrix} 1 & 0 \\ -\frac{\Sigma_{12}}{\Sigma_{22}} & 1 \end{bmatrix} \begin{bmatrix} \left(\Sigma_{11} - \frac{\Sigma_{12}^2}{\Sigma_{22}}\right)^{-1} & 0 \\ 0 & \frac{1}{\Sigma_{22}} \end{bmatrix} \begin{bmatrix} 1 & -\frac{\Sigma_{12}}{\Sigma_{22}} \\ 0 & 1 \end{bmatrix}.$$

We can now plug this into the density function. Recall that

$$\begin{aligned} f_{Z_1,Z_2}(z_1,z_2) &\propto \exp\left(-\frac{1}{2} \begin{bmatrix} z_1 & z_2 \end{bmatrix} \Sigma^{-1} \begin{bmatrix} z_1 \\ z_2 \end{bmatrix}\right) \\ &\propto \exp\left(-\frac{1}{2} \begin{bmatrix} z_1 & z_2 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ -\frac{\Sigma_{12}}{\Sigma_{22}} & 1 \end{bmatrix} \begin{bmatrix} \left(\Sigma_{11} - \frac{\Sigma_{12}^2}{\Sigma_{22}}\right)^{-1} & 0 \\ 0 & \frac{1}{\Sigma_{22}} \end{bmatrix} \begin{bmatrix} 1 & -\frac{\Sigma_{12}}{\Sigma_{22}} \\ 0 & 1 \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \end{bmatrix}\right) \\ &\propto \exp\left(-\frac{1}{2} \begin{bmatrix} z_1 - \frac{\Sigma_{12}}{\Sigma_{22}} z_2 & z_2 \end{bmatrix} \begin{bmatrix} \left(\Sigma_{11} - \frac{\Sigma_{12}^2}{\Sigma_{22}}\right)^{-1} & 0 \\ 0 & \frac{1}{\Sigma_{22}} \end{bmatrix} \begin{bmatrix} z_1 - \frac{\Sigma_{12}}{\Sigma_{22}} z_2 \\ z_2 \end{bmatrix}\right). \end{aligned}$$

Now see that since the square matrix is diagonal, our density decomposes to yield

$$f_{Z_1,Z_2}(z_1,z_2) \propto \exp\left(-\frac{1}{2}(z_1 - \frac{\Sigma_{12}}{\Sigma_{22}} z_2)^2 \left(\Sigma_{11} - \frac{\Sigma_{12}^2}{\Sigma_{22}}\right)^{-1}\right) \exp\left(-\frac{1}{2\Sigma_{22}} z_2^2\right).$$

Conditional on $Z_2 = z_2$, we see that $Z_1 | Z_2 = z_2 \sim N(\frac{\Sigma_{12}}{\Sigma_{22}} z_2, \Sigma_{11} - \frac{\Sigma_{12}^2}{\Sigma_{22}})$.

2 Probabilistic model of Weighted Least Squares

Let us now set up a probabilistic model from which weighted least squares arises as the natural solution.

(a) Let $X_1, X_2, \ldots, X_n \in \mathbb{R}^d$ be n random vectors and $Y_1, Y_2, \ldots, Y_n \in \mathbb{R}$ be one-dimensional random variables. Assume $Y_i | X_i$ are independently distributed as

$$Y_i = X_i^T w + z_i, \tag{1}$$

where $z_i \sim N(0, \sigma_i^2)$, for some fixed but unknown parameter vector $w \in \mathbb{R}^d$. What is the conditional distribution of Y_i given X_i?

Solution: We have

$$P(Y_i|X_i) \propto \exp\{-\frac{(Y_i - X_i^T w)^2}{2\sigma_i^2}\}. \tag{2}$$

Therefore, $Y_i|X_i \sim N(X_i^T w, \sigma_i^2)$.

(b) Derive the solution to weighted least square as a maximum likelihood estimator of the above model.

Solution:

The log likelihood function is

$$\begin{aligned}
L(w) &\propto \log \Pi_{i=1}^n P(Y_i|X_i) \\
&\propto \log \Pi_{i=1}^n \exp\{-\frac{(Y_i - X_i^T w)^2}{2\sigma_i^2}\} \\
&\propto \sum_{i=1}^n -\frac{(Y_i - X_i^T w)^2}{2\sigma_i^2} \\
&\propto -\frac{1}{2}\{w^T \sum_{i=1}^n \frac{X_i X_i^T}{\sigma_i^2} w - 2\sum_{i=1}^n \frac{Y_i X_i^T}{\sigma_i^2} w\}.
\end{aligned}$$

Taking gradient with respect to w and setting the gradient to zero, we get

$$\begin{aligned}
w &= (\sum_{i=1}^n \frac{X_i X_i^T}{\sigma_i^2})^{-1} \sum_{i=1}^n \frac{Y_i X_i}{\sigma_i^2} \\
&= (X^T \Lambda X)^{-1} X^T \Lambda Y,
\end{aligned}$$

with $X \in \mathbb{R}^{n \times d}$ whose ith row is X_i^T, $Y \in \mathbb{R}^n$ whose ith entry is Y_i, and Λ is a d-dimensional diagonal matrix with the ith diagonal being $1/\sigma_i^2$.

(c) Define $\tilde{Y}_i = \frac{Y_i}{\sigma_i}$ and $\tilde{X}_i = \frac{X_i}{\sigma_i}$. Suppose we still have

$$Y_i = X_i^T w + z_i, \tag{3}$$

where $z_i \sim N(0, \sigma_i^2)$.

Write out the relationship of $\tilde{X}_i$ and $\tilde{Y}_i$.

Solution: We have

$$\tilde{Y}_i = \tilde{X}_i^T w + z_i/\sigma_i, \tag{4}$$

where $z_i \sim N(0, \sigma_i^2)$.

That is, we have

$$\tilde{Y}_i = \tilde{X}_i^T w + \varepsilon_i, \tag{5}$$

where $\varepsilon_i \sim N(0, 1)$.

(d) Suppose $(\tilde{X}_i, \tilde{Y}_i)$ are observed for $i = 1, \ldots, n$. What is the maximum likelihood estimator of w (as a function of the tuples $(\tilde{X}_i, \tilde{Y}_i)$)?

Solution:

In Part (c), we have the probablistic model for regular least square. Therefore, we have the maximum likelihood estimator of w being

$$w = (\tilde{X}^T \tilde{X})^{-1} \tilde{X}^T \tilde{Y}, \tag{6}$$

with $\tilde{X} \in \mathbb{R}^{n \times d}$ having ith row $\tilde{X}_i^T$, and $Y \in \mathbb{R}^n$ having ith entry is $\tilde{Y}_i$.

(e) You are given training data $\tilde{Y}_i = \frac{Y_i}{\sigma_i}$ and $\tilde{X}_i = \frac{X_i}{\sigma_i}$. Using part (d), derive the solution to the weighted least squares problem.

Solution:

Let $X \in \mathbb{R}^{n \times d}$ whose ith row is X_i^T, $Y \in \mathbb{R}^n$ whose ith entry is Y_i.
We have $\tilde{X} = \sqrt{\Lambda} X$ and $\tilde{Y} = \sqrt{\Lambda} Y$, with

$$\sqrt{\Lambda} = \text{diag}(1/\sigma_1, \cdots, 1/\sigma_n). \tag{7}$$

Plugging to the solution of Part (d), we obtain

$$w = (X^T \Lambda X)^{-1} X^T \Lambda Y.$$