STAT 341 | Lab #11 |
Spring 2022 | Tue, May 3rd | Name: Solutions

In this lab, we are going to explore a special multivariate continuous distribution called the multivariate Normal distribution. As its name implies, the multivariate Normal distribution is the multivariate version of the Normal distribution we studied in Unit 3. In order to visualize this distribution, we will look at its bivariate version with random variables Y_1 and Y_2. The extension to the multivariate version is discussed at the end of the lab.

Activity 1: *The Bivariate Normal Distribution*

While looking for resources on the bivariate Normal distribution, I found a great article with excellent graphics that explains the basics of this distribution. The link for this article can be found with this lab in Canvas. Please read this article and use it to answer the questions below.

1. The bivariate Normal distribution has five parameters: μ_1, μ_2, σ_{11}, σ_{22}, and σ_{12} (or ρ). Describe each of these parameters.

   μ_1 is the marginal mean of Y_1, μ_2 is the marginal mean of Y_2, σ_{11} is the marginal variance of Y_1, σ_{22} is the marginal variance of Y_2, and ρ is the correlation between Y_1 and Y_2.

2. Explain why for the bivariate Normal distribution, the following statement is true: Y_1 and Y_2 are independent if and only if their covariance (or correlation) is 0.

   If the correlation is zero, then the joint density for Y_1 and Y_2 is the product of their marginal densities, which is the definition of independence.

3. What is the significance of the point (μ_1, μ_2) in the bivariate Normal distribution?

   This would be the tallest point of the "hill" representing the density function.

4. What is the role of the values σ_{11} and σ_{22} in the shape of the bivariate Normal distribution?

   If you're looking at the contour plots, σ_{11} determines the "width" of the ellipse and σ_{22} determines the "height" of the ellipse. In terms of the bivariate density, they control the widths of the "hill".

5. What is the role of the value ρ in the shape of the bivariate Normal distribution?

   Looking at the contour plots, increasing ρ causes the ellipse to tilt more.

For the next three questions, suppose Y_1 and Y_2 have a bivariate Normal distribution with parameters $\mu_1 = 5$, $\mu_2 = 3$, $\sigma_{11} = 0.25$, $\sigma_{22} = 1$ and $\rho = 0.4$ ($\sigma_{12} = 0.2$).

6. Give the marginal distributions for both Y_1 and Y_2. Which parameter of the bivariate Normal distribution does not appear in either marginal distribution?

   The correlation ρ.

7. Give the conditional distribution of Y_1 if the value of $Y_2 = 2$.

$$Y_1|Y_2 = 2 \sim N\left(\mu_1 + \frac{\sigma_{12}}{\sigma_{22}}(2-\mu_2), \sigma_{11} - \frac{\sigma_{12}^2}{\sigma_{22}}\right)$$

8. Give the conditional distribution of Y_2 if the value of $Y_1 = 6$.

$$Y_2|Y_1 = 6 \sim N\left(\mu_2 + \frac{\sigma_{12}}{\sigma_{11}}(6-\mu_1), \sigma_{22} - \frac{\sigma_{12}^2}{\sigma_{11}}\right)$$

Activity 2: *The Multivariate Normal Distribution*

A multivariate Normal distribution is an extension of the bivariate Normal distribution to three or more Normal random variables. Each of these variables has a mean μ_i and the variances and covariances between the random variables are summarized in a matrix of the form

$$V(Y_1, Y_2, \ldots, Y_n) = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} & \cdots & \cdots & \sigma_{1n} \\ \sigma_{12} & \sigma_{22} & \sigma_{23} & \cdots & \cdots & \sigma_{2n} \\ \sigma_{13} & \sigma_{23} & \sigma_{33} & \cdots & \cdots & \sigma_{3n} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \sigma_{1n} & \sigma_{2n} & \sigma_{3n} & \cdots & \cdots & \sigma_{nn} \end{bmatrix}$$

The marginal distribution of each variable Y_i is a Normal distribution with mean μ_i and variance σ_{ii} and the bivariate distribution of each pair of variables Y_i and Y_j is a bivariate Normal distribution with means μ_i and μ_j, variances σ_{ii} and σ_{jj} and covariance σ_{ij}.

Suppose we have four random variables Y_1, Y_2, Y_3, and Y_4, with means 3, 2, 4, 5, respectively. The variance/covariance matrix is:

$$V(Y_1, Y_2, Y_3, Y_4) = \begin{bmatrix} 1 & 0 & -1.4 & -0.5 \\ 0 & 3 & 0.4 & 2.2 \\ -1.4 & 0.4 & 4 & 1.4 \\ -0.5 & 2.2 & 1.4 & 2 \end{bmatrix}$$

9. Which variable has the largest variance? the smallest variance?

   Y_3 has the largest variance and Y_1 has the smallest variance.

10. Which pair(s) of variables are independent?

    Y_1 and Y_2 are independent because their covariance is 0.

11. Which pair(s) of variables have a negative linear relationship? a positive linear relationship?

    Y_1 & Y_3 and Y_1 & Y_4 have negative linear relationships.
    Y_2 & Y_3, Y_2 & Y_4, and Y_3 & Y_4 have positive linear relationships.

12. Give the marginal distribution for the variable Y_3.

$$Y_3 \sim N(\mu_3 = 4, \sigma_{33} = 4)$$

13. Give the bivariate distribution for the variables Y_1 and Y_4.

$$(Y_1, Y_4) \sim N(\mu_1 = 3, \mu_4 = 5, \sigma_{11} = 1, \sigma_{44} = 2, \sigma_{14} = -0.5)$$

Activity 3: *Final Exam Review*

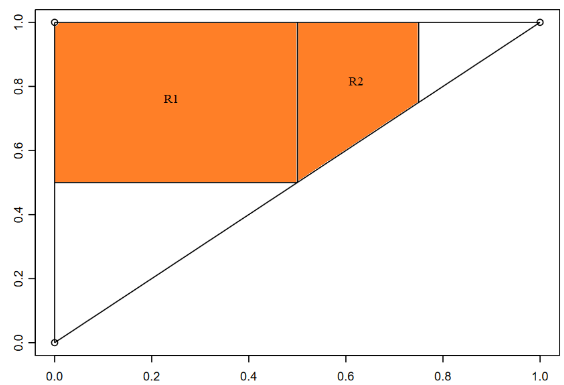
This activity is optional (it will not be graded).

1. Let Y_1 and Y_2 have the joint p.d.f. given by

$$f(y_1, y_2) = \begin{cases} 6(1-y_2) & 0 \le y_1 \le y_2 \le 1 \\ 0 & \text{elsewhere} \end{cases}$$

(a) Find $P(Y_1 \le 3/4, Y_2 \ge 1/2)$.

The region where $Y_1 \le 0.75$ and $Y_2 \ge 0.5$ within the support of Y_1 and Y_2 is shown in the figure below. This region must be split into two parts. The two parts I chose are shown in the graph. You could choose to break this region up in a different manner.



First region

$$\begin{aligned} \int_{0.5}^{1}\int_{0}^{0.5} 6(1-y_2)dy_1dy_2 &= \int_{0.5}^{1} 6(1-y_2)\left(y_1|_0^{0.5}\right)dy_2 \\ &= \int_{0.5}^{1} 3(1-y_2)dy_2 \\ &= \left(3y_2 - 1.5y_2^2\right)\Big|_{0.5}^{1} \\ &= 1.5 - 1.125 = 0.375 \end{aligned}$$

Second region

$$\begin{aligned} \int_{0.5}^{0.75}\int_{y_1}^{1} 6(1-y_2)dy_2dy_1 &= \int_{0.5}^{0.75} \left(6y_2 - 3y_2^2|_{y_1}^{1}\right)dy_1 \\ &= \int_{0.5}^{0.75} 3 - 6y_1 + 3y_1^2 dy_1 \\ &= \left(3y_1 - 3y_1^2 + y_1^3\right)\Big|_{0.5}^{0.75} \\ &= 0.984375 - 0.875 = 0.109375 \end{aligned}$$

Adding the two regions together gives $0.375 + 0.109375 = 0.484375$.

(b) Find the marginal distributions for both Y_1 and Y_2.

The marginal distribution of Y_1 is

$$\begin{aligned} f_1(y_1) &= \int_{y_1}^{1} 6(1-y_2)dy_2 \\ &= (6y_2 - 3y_2^2|_{y_1}^{1}) \\ &= 3y_1^2 - 6y_1 + 3 = 3(1-y_1)^2 \end{aligned}$$

for $0 \le y_1 \le 1$ and 0 otherwise.

The marginal distribution of Y_2 is

$$\begin{aligned} f_2(y_2) &= \int_{0}^{y_2} 6(1-y_2)dy_1 \\ &= (6(1-y_2)y_1|_{0}^{y_2}) \\ &= 6y_2(1-y_2) \end{aligned}$$

for $0 \le y_2 \le 1$ and 0 otherwise.

(c) Find the conditional density function of Y_1 when $Y_2 = y_2$.

$$f(y_1|Y_2 = y_2) = \frac{f(y_1, y_2)}{f_2(y_2)} = \frac{6(1-y_2)}{6y_2(1-y_2)} = \frac{1}{y_2}$$

for $0 \le y_1 \le y_2$ and 0 elsewhere.

(d) Find the conditional density function of Y_2 when $Y_1 = y_1$.

$$f(y_2|Y_1 = y_1) = \frac{f(y_1, y_2)}{f_1(y_1)} = \frac{6(1-y_2)}{3(1-y_1)^2} = \frac{2(1-y_2)}{(1-y_1)^2}$$

for $y_1 \le y_2 \le 1$ and 0 elsewhere.

(e) Are Y_1 and Y_2 independent random variables? Explain your answer.

We have found the marginal distributions as

$$\begin{aligned} f_1(y_1) &= 3(1-y_1)^2 \qquad \text{for } 0 \le y_1 \le 1 \\ f_2(y_2) &= 6y_2(1-y_2) \qquad \text{for } 0 \le y_2 \le 1 \end{aligned}$$

Since the product of the marginal distributions $(3(1-y_1)^2 * 6y_2(1-y_2))$ does not equal to joint distribution $(6(1-y_2))$, these random variables are not independent.

(f) Find $E(Y_1)$ and $V(Y_1)$.

$$\begin{aligned}
E(Y_1) &= \int_0^1 y_1 \cdot 3(1-y_1)^2 \; dy_1 = \int_0^1 3y_1(1-2y_1+y_1^2) \; dy_1 \\
&= \int_0^1 3y_1 - 6y_1^2 + 3y_1^3 \; dy_1 = \frac{3}{2}y_1^2 - \frac{6}{3}y_1^3 + \frac{3}{4}y_1^4 \Big|_0^1 \\
&= \frac{3}{2} - 2 + \frac{3}{4} = \frac{6-8+3}{4} = \frac{1}{4} = 0.25 \\
E(Y_1^2) &= \int_0^1 y_1^2 \cdot 3(1-y_1)^2 \; dy_1 = \int_0^1 3y_1^2(1-2y_1+y_1^2) \; dy_1 \\
&= \int_0^1 3y_1^2 - 6y_1^3 + 3y_1^4 \; dy_1 = \frac{3}{3}y_1^3 - \frac{6}{4}y_1^4 + \frac{3}{5}y_1^5 \Big|_0^1 \\
&= 1 - \frac{3}{2} + \frac{3}{5} = \frac{10-15+6}{10} = \frac{1}{10} = 0.1 \\
V(Y_1) &= E(Y_1^2) - [E(Y_1)]^2 = 0.1 - 0.25^2 = 0.1 - 0.0625 = 0.0375
\end{aligned}$$

(g) Find $E(Y_2)$ and $V(Y_2)$.

$$\begin{aligned}
E(Y_2) &= \int_0^1 y_2 \cdot 6y_2(1-y_2) \; dy_2 = \int_0^1 6y_2^2 - 6y_2^3 \; dy_2 \\
&= \frac{6}{3}y_2^3 - \frac{6}{4}y_2^4 \Big|_0^1 = 2 - \frac{3}{2} = \frac{1}{2} = 0.5 \\
E(Y_2^2) &= \int_0^1 y_2^2 \cdot 6y_2(1-y_2) \; dy_2 = \int_0^1 6y_2^3 - 6y_2^4 \; dy_2 \\
&= \frac{6}{4}y_2^4 - \frac{6}{5}y_2^5 \Big|_0^1 = \frac{3}{2} - \frac{6}{5} = \frac{15-12}{10} = \frac{3}{10} = 0.3 \\
V(Y_2) &= E(Y_2^2) - [E(Y_2)]^2 = 0.3 - 0.5^2 = 0.3 - 0.25 = 0.05
\end{aligned}$$

(h) Find $Cov(Y_1, Y_2)$ and ρ. Interpret the value of ρ.

$$\begin{aligned}
E(Y_1Y_2) &= \int_{y_1=0}^{y_1=1} \int_{y_2=y_1}^{y_2=1} 6y_1y_2 - 6y_1y_2^2 \; dy_2dy_1 \\
&= \int_{y_1=0}^{y_1=1} 3y_1y_2^2 - 2y_1y_2^3 \Big|_{y_2=y_1}^{y_2=1} \; dy_1 \\
&= \int_{y_1=0}^{y_1=1} y_1 - 3y_1^3 + 2y_1^4 \; dy_1 = \frac{1}{2}y_1^2 - \frac{3}{4}y_1^4 + \frac{2}{5}y_1^5 \Big|_{y_1=0}^{y_1=1} \\
&= \frac{1}{2} - \frac{3}{4} + \frac{2}{5} = \frac{10-15+8}{20} = \frac{3}{20}
\end{aligned}$$

$$\begin{aligned}
Cov(Y_1, Y_2) &= E(Y_1Y_2) - E(Y_1)E(Y_2) = \frac{3}{20} - \frac{1}{4}\left(\frac{1}{2}\right) \\
&= \frac{6-5}{40} = \frac{1}{40} = 0.025 \\
\rho &= \frac{Cov(Y_1, Y_2)}{\sqrt{V(Y_1)V(Y_2)}} = \frac{0.025}{\sqrt{0.0375 \times 0.05}} = 0.5774
\end{aligned}$$

Y_1 and Y_2 have a moderately-strong, positive linear relationship. So, on average an increase in Y_1 results in an increase in Y_2.

(i) Find $E(Y_1|Y_2 = y_2)$ and $V(Y_1|Y_2 = y_2)$.

$$\begin{aligned}
E(Y_1|Y_2 = y_2) &= \int_{y_1=0}^{y_1=y_2} \frac{y_1}{y_2}\, dy_1 = \frac{y_1^2}{2y_2}\Big|_{y_1=0}^{y_1=y_2} = \frac{y_2^2}{2y_2} = \frac{y_2}{2} \\
E(Y_1^2|Y_2 = y_2) &= \int_{y_1=0}^{y_1=y_2} \frac{y_1^2}{y_2}\, dy_1 = \frac{y_1^3}{3y_2}\Big|_{y_1=0}^{y_1=y_2} = \frac{y_2^3}{3y_2} = \frac{y_2^2}{3} \\
V(Y_1|Y_2 = y_2) &= E(Y_1^2|Y_2 = y_2) - [E(Y_1|Y_2 = y_2)]^2 \\
&= \frac{y_2^2}{3} - \left(\frac{y_2}{2}\right)^2 = \frac{4y_2^2 - 3y_2^2}{12} = \frac{y_2^2}{12}
\end{aligned}$$

(j) Find $E(Y_2|Y_1 = y_1)$ and $V(Y_2|Y_1 = y_1)$.

$$\begin{aligned}
E(Y_2|Y_1 = y_1) &= \int_{y_2=y_1}^{y_2=1} \frac{2y_2 - 2y_2^2}{(1-y_1)^2}\, dy_2 = \frac{y_2^2 - \frac{2}{3}y_2^3}{(1-y_1)^2}\Big|_{y_2=y_1}^{y_2=1} = \frac{\frac{1}{3} - y_1^2 + \frac{2}{3}y_1^3}{(1-y_1)^2} \\
E(Y_2^2|Y_1 = y_1) &= \int_{y_2=y_1}^{y_2=1} \frac{2y_2^2 - 2y_2^3}{(1-y_1)^2}\, dy_2 = \frac{\frac{2}{3}y_2^3 - \frac{2}{4}y_2^4}{(1-y_1)^2}\Big|_{y_2=y_1}^{y_2=1} = \frac{\frac{1}{6} - \frac{2}{3}y_1^3 + \frac{1}{2}y_1^4}{(1-y_1)^2} \\
V(Y_2|Y_1 = y_1) &= E(Y_2^2|Y_1 = y_1) - [E(Y_2|Y_1 = y_1)]^2 \\
&= \frac{\frac{1}{6} - \frac{2}{3}y_1^3 + \frac{1}{2}y_1^4}{(1-y_1)^2} - \left(\frac{\frac{1}{3} - y_1^2 + \frac{2}{3}y_1^3}{(1-y_1)^2}\right)^2
\end{aligned}$$

2. An environmental engineer measures the amount (by weight) of particulate pollution in air samples of a certain volume collected over two smokestacks at a coal-operated power plant. One of the stacks is equipped with a cleaning device. Let X_1 denote the amount of pollutant per sample collected above the stack that has no cleaning device and let X_2 denote the amount of pollutant per sample collected above the stack that is equipped with the cleaning device. Suppose that the relative frequency behavior of X_1 and X_2 can be modeled by

$$f(x_1, x_2) = \begin{cases} 1 & 0 \le x_1 \le 2, 0 \le x_2 \le 1, 2x_2 \le x_1 \\ 0 & \text{elsewhere} \end{cases}$$

(a) Find the expected value and variance of the random variable $Y_1 = X_1 - X_2$, which is the amount of pollutant removed by the cleaner.

From the previous homework, we found that the marginal distribution for X_1 is

$$f_{X_1}(x_1) = \int_0^{0.5x_1} dx_2 = 0.5x_1 \qquad \text{for } 0 \le x_1 \le 2$$

The marginal distribution for X_2 is

$$f_{X_2}(x_2) = \int_{2x_2}^{2} dx_1 = 2(1 - x_2) \qquad \text{for } 0 \le x_2 \le 1$$

We can find the values of μ_{X_1} and μ_{X_2} using the marginal distributions:

$$\begin{aligned}
\mu_{X_1} &= \int_0^2 x_1 * 0.5x_1 dx_1 \\
&= \int_0^2 0.5x_1^2 dx_1 \\
&= \frac{1}{6}\left(y_1^3\Big|_0^2\right) \\
&= \frac{1}{6}(8-0) = \frac{4}{3}
\end{aligned}$$

$$\begin{aligned}
\mu_{X_2} &= \int_0^1 x_2 * 2(1 - x_2) dx_2 \\
&= \int_0^1 2x_2 - 2x_2^2 dx_2 \\
&= x_2^2 - \frac{2}{3}x_2^3\Big|_0^1 \\
&= 1 - \frac{2}{3} = \frac{1}{3}
\end{aligned}$$

$Y_1 = X_1 - X_2 = a_1X_1 + a_2X_2$ is a linear combination of X_1 and X_2 with $a_1 = 1$ and $a_2 = -1$. The expected value of Y_1 is:

$$\begin{aligned}
E(Y_1) &= E(X_1 - X_2) \\
&= 1 * E(X_1) + (-1) * E(X_2) \\
&= 1 * \mu_{X_1} + (-1) * \mu_{X_2} \\
&= 1 * \frac{4}{3} + (-1) * \frac{1}{3} \\
&= 1
\end{aligned}$$

The formula for the variance of Y_1 requires the variances of X_1 and X_2. These are:

$$\begin{aligned} E(X_1^2) &= \int_0^2 x_1^2 * 0.5 x_1 dx_1 \\ &= \int_0^2 0.5 x_1^3 dx_1 \\ &= \frac{1}{8}\left(x_1^4 \Big|_0^2 \right) \\ &= \frac{1}{8}(16-0) \\ &= 2 \end{aligned}$$

$$\begin{aligned} \sigma_{X_1}^2 &= E(X_1^2) - \mu_{X_1}^2 \\ &= 2 - \left(\frac{4}{3}\right)^2 \\ &= \frac{2}{9} \end{aligned}$$

$$\begin{aligned} E(X_2^2) &= \int_0^1 x_2^2 * 2(1-x_2) dx_2 \\ &= \int_0^1 2x_2^2 - 2x_2^3 dx_2 \\ &= \frac{2}{3}x_2^3 - \frac{1}{2}x_2^4 \Big|_0^1 \\ &= \frac{2}{3} - \frac{1}{2} \\ &= \frac{1}{6} \end{aligned}$$

$$\begin{aligned} \sigma_{X_2}^2 &= E(X_2^2) - \mu_{X_2}^2 \\ &= \frac{1}{6} - \left(\frac{1}{3}\right)^2 \\ &= \frac{1}{18} \end{aligned}$$

The formula for the variance of Y_1 also requires the covariance, which depends on the value $E(X_1X_2)$:

$$\begin{aligned}
E(X_1X_2) &= \int_0^2 \int_0^{0.5x_1} x_1x_2dx_2dx_1 \\
&= \int_0^2 x_1 \left(\frac{1}{2}x_2^2 \Big|_0^{0.5x_1} \right) dx_1 \\
&= \int_0^2 x_1 \left(\frac{1}{8}x_1^2 \right) dx_1 \\
&= \int_0^2 \frac{1}{8}x_1^3dx_1 \\
&= \frac{1}{32} \left(x_1^4 \Big|_0^2 \right) \\
&= \frac{1}{32}(16-0) \\
&= 0.5
\end{aligned}$$

Next, we can calculate $Cov(X_1, X_2)$ as:

$$\begin{aligned}
Cov(X_1, X_2) &= E(X_1X_2) - \mu_{X_1}\mu_{X_2} \\
&= 0.5 - \left(\frac{4}{3} \right) \left(\frac{1}{3} \right) \\
&= \frac{1}{18}
\end{aligned}$$

Finally, the variance of Y_1 is then:

$$\begin{aligned}
V(Y_1) &= V(X_1 - X_2) \\
&= 1^2V(X_1) + (-1)^2V(X_2) + 2(1)(-1)Cov(X_1, X_2) \\
&= \frac{2}{9} + \frac{1}{18} - 2\left(\frac{1}{18} \right) \\
&= \frac{1}{6}
\end{aligned}$$

(b) Find the pdf of the random variable $Y_1 = X_1 - X_2$.
(Hint: start by finding the joint p.d.f. of Y_1 and $Y_2 = X_2$.)

Let $Y_1 = X_1 - X_2$ and $Y_2 = X_2$. Solve these two equations for X_1 and X_2.

$$x_2 = y_2$$

$$\begin{aligned} y_1 &= x_1 - x_2 \\ y_1 + x_2 &= x_1 \\ y_1 + y_2 &= x_1 \end{aligned}$$

Next, find the Jacobian matrix.

$$J = \begin{bmatrix} \frac{\partial x_1}{y_1} & \frac{\partial x_1}{y_2} \\ \frac{\partial x_2}{y_1} & \frac{\partial x_2}{y_2} \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$$

The determinate of J is just 1.

The joint p.d.f. of Y_1 and Y_2 is:

$$f_{Y_1,Y_2}(y_1, y_2) = f_{X_1,X_2}(x_1 = y_1 + y_2, x_2 = y_2)|J| = 1 * 1 = 1$$

Now, we must find the support of the variables Y_1 and Y_2. Start with the support of X_1 and X_2 and substitute the values for y_1 and y_2.

- $0 \le x_1 \le 2$ implies $0 \le y_1 + y_2 \le 2$ or $-y_2 \le y_1 \le 2 - y_2$
- $0 \le x_2 \le 1$ implies $0 \le y_2 \le 1$
- $2x_2 \le x_1$ implies $2y_2 \le y_1 + y_2$ or $y_2 \le y_1$

If we draw these regions, we find the support of Y_1 and Y_2 consists of two triangular regions with constraints:

- $0 \le y_2 \le y_1 \le 1$
- $0 \le y_2 \le 2 - y_1; 1 \le y_1 \le 2$

The marginal distribution of $Y_1 = X_1 - X_2$ for y_1 between 0 and 1 will then be:

$$f_{Y_1}(y_1) = \int_0^{y_1} 1 dy_2 = y_2\Big|_0^{y_1} = y_1$$

The marginal distribution of $Y_1 = X_1 - X_2$ for y_1 between 1 and 2 will then be:

$$f_{Y_1}(y_1) = \int_0^{2-y_1} 1 dy_2 = y_2\Big|_0^{2-y_1} = 2 - y_1$$

Putting it all together:

$$f_{Y_1}(y_1) = \begin{cases} y_1 & 0 \le y_1 \le 1 \\ 2 - y_1 & 1 \le y_1 \le 2 \\ 0 & \text{elsewhere} \end{cases}$$