Mathematics for Machine Learning

— Probability & Distributions (Supplementary): Gaussian Distribution & Change of Variables/Inverse Transform

Joseph Chuang-Chieh Lin

Department of Computer Science & Information Engineering,
Tamkang University

Fall 2023

Credits for the resource

- The slides are based on the textbooks:
  - *Marc Peter Deisenroth, A. Aldo Faisal, and Cheng Soon Ong: Mathematics for Machine Learning. Cambridge University Press. 2020.*
  - *Howard Anton, Chris Rorres, Anton Kaul: Elementary Linear Algebra. Wiley. 2019.*
- We could partially refer to the monograph:
  Francesco Orabona: A Modern Introduction to Online Learning.
  https://arxiv.org/abs/1912.13213

Outline

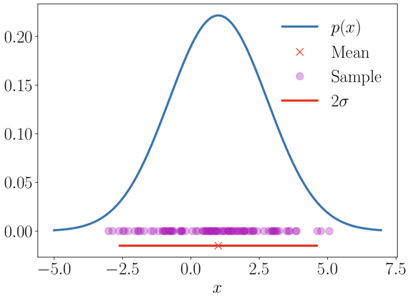
1. Gaussian Distribution
   - Marginals and Conditionals of Gaussians
   - Sums and Linear Transformations
2. Change of Variables
   - Distribution Function Technique
   - Change of Variables

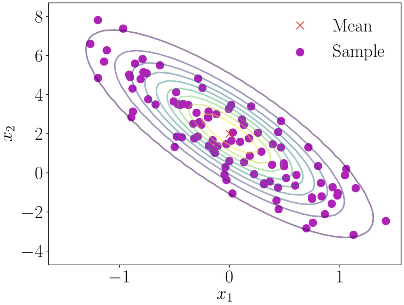
Outline

Introduction

- The Gaussian distribution (a.k.s. normal distribution) is the most well-studied probability distribution for continuous-valued random variables.
- Widely used in statistics and machine learning.

Gaussian Distributions Overlaid with Samples





Univariate & Multivariate Gaussian

The probability density functions.

Univariate

$$p(x \mid \mu, \sigma^2) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

$\boldsymbol{\Sigma} = \mathbb{V}_X[\mathbf{x}] = \mathrm{Cov}_X[\mathbf{x}, \mathbf{x}]$.

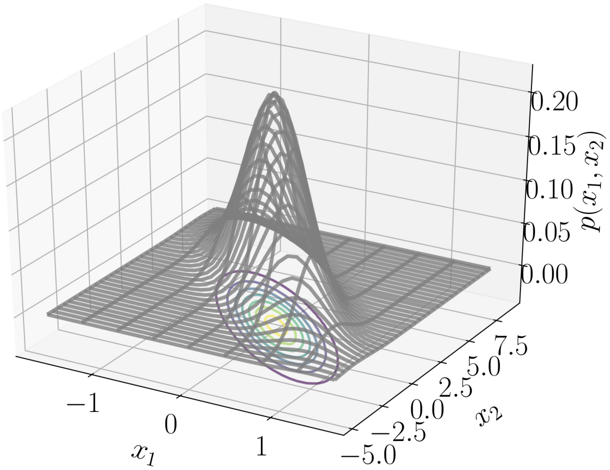
Multivariate

$$p(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-\frac{D}{2}} \det(\boldsymbol{\Sigma})^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right),$$

for $\mathbf{x} \in \mathbb{R}^D$.

We write $p(\mathbf{x}) = \mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})$ or $X \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

Gaussian distribution of two random variables x_1, x_2.



Outline

Marginals and Conditionals of Gaussians

- Let X, Y be two multivariate random variables.
- Concatenate their states to be $[\mathbf{x}^\top, \mathbf{y}]^\top$.

$$p(\mathbf{x}, \mathbf{y}) = \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\mu}_x \\ \boldsymbol{\mu}_y \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{xx} & \boldsymbol{\Sigma}_{xy} \\ \boldsymbol{\Sigma}_{yx} & \boldsymbol{\Sigma}_{yy} \end{bmatrix}\right).$$

where $\boldsymbol{\Sigma}_{xx} = \mathrm{Cov}[\mathbf{x}, \mathbf{x}]$, $\boldsymbol{\Sigma}_{yy} = \mathrm{Cov}[\mathbf{y}, \mathbf{y}]$, $\boldsymbol{\Sigma}_{xy} = \mathrm{Cov}[\mathbf{x}, \mathbf{y}]$.

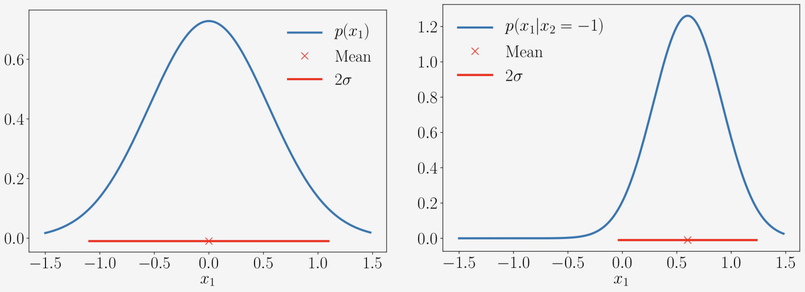
- By [Bishop 2006], the conditional distribution $p(\mathbf{x} \mid \mathbf{y})$ is also Gaussian.

$$\begin{aligned} p(\mathbf{x} \mid \mathbf{y}) &= \mathcal{N}(\boldsymbol{\mu}_{x|y}, \boldsymbol{\Sigma}_{x|y}) \\ \boldsymbol{\mu}_{x|y} &= \boldsymbol{\mu}_x + \boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_{yy}^{-1}(\mathbf{y} - \boldsymbol{\mu}_y) \\ \boldsymbol{\Sigma}_{x|y} &= \boldsymbol{\Sigma}_{xx} - \boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_{yy}^{-1}\boldsymbol{\Sigma}_{yy}. \end{aligned}$$

$$p(\mathbf{x}) = \int p(\mathbf{x}, \mathbf{y})\mathrm{d}\mathbf{y} = \mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}_x, \boldsymbol{\Sigma}_{xx}).$$

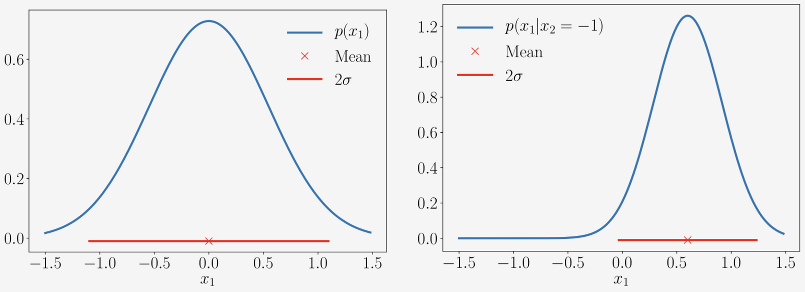
Example

Consider $$p(x_1, x_2) = \mathcal{N}\left(\begin{bmatrix} 0 \\ 2 \end{bmatrix}, \begin{bmatrix} 0.3 & -1 \\ -1 & 5 \end{bmatrix}\right).$$



Example

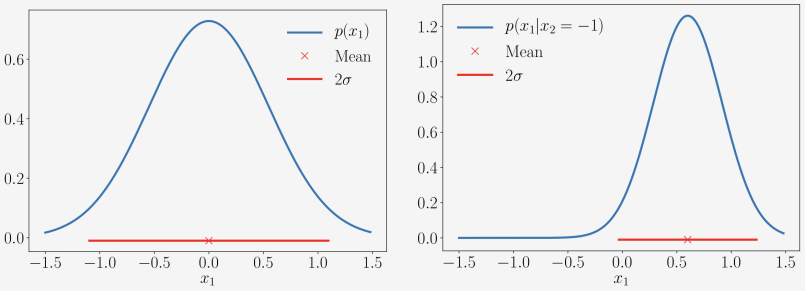
Consider $$p(x_1, x_2) = \mathcal{N}\left(\begin{bmatrix} 0 \\ 2 \end{bmatrix}, \begin{bmatrix} 0.3 & -1 \\ -1 & 5 \end{bmatrix} \right).$$



Conditioned on $x_2 = -1$, $\mu_{x_1|x_2=-1} = 0 + (-1) \cdot 0.2 \cdot (-1 - 2) = 0.6$

Example

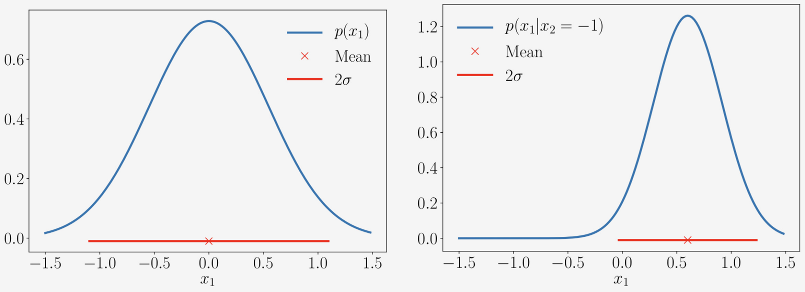
Consider $$p(x_1, x_2) = \mathcal{N}\left(\begin{bmatrix} 0 \\ 2 \end{bmatrix}, \begin{bmatrix} 0.3 & -1 \\ -1 & 5 \end{bmatrix}\right).$$



Conditioned on $x_2 = -1$, $\mu_{x_1|x_2=-1} = 0 + (-1) \cdot 0.2 \cdot (-1 - 2) = 0.6$ and $\sigma^2_{x_1|x_2=-1} =$

Example

Consider $$p(x_1, x_2) = \mathcal{N}\left(\begin{bmatrix} 0 \\ 2 \end{bmatrix}, \begin{bmatrix} 0.3 & -1 \\ -1 & 5 \end{bmatrix}\right).$$

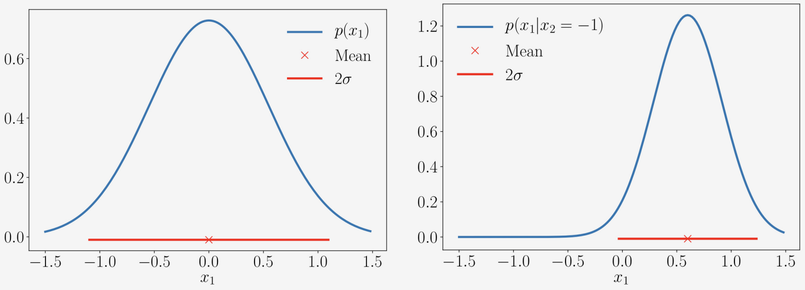


Conditioned on $x_2 = -1$, $\mu_{x_1|x_2=-1} = 0 + (-1) \cdot 0.2 \cdot (-1 - 2) = 0.6$ and $\sigma^2_{x_1|x_2=-1} = 0.3 - (-1) \cdot 0.2 \cdot (-1) = 0.1$.

Thus, $p(x_1 \mid x_2 = -1) =$

Example

Consider $$p(x_1, x_2) = \mathcal{N}\left(\begin{bmatrix} 0 \\ 2 \end{bmatrix}, \begin{bmatrix} 0.3 & -1 \\ -1 & 5 \end{bmatrix} \right).$$

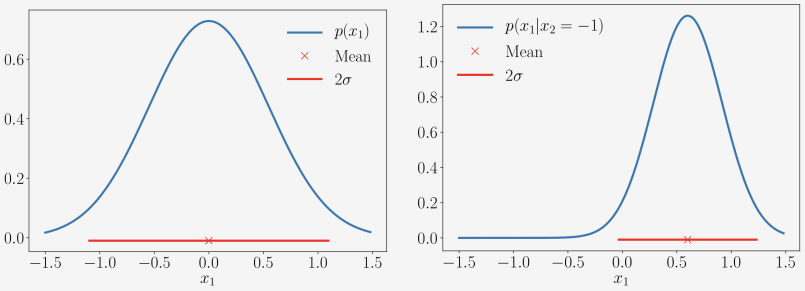


Conditioned on $x_2 = -1$, $\mu_{x_1|x_2=-1} = 0 + (-1) \cdot 0.2 \cdot (-1 - 2) = 0.6$ and $\sigma^2_{x_1|x_2=-1} = 0.3 - (-1) \cdot 0.2 \cdot (-1) = 0.1$.

Thus, $p(x_1 \mid x_2 = -1) = \mathcal{N}(0.6, 0.1)$,

Example

Consider $$p(x_1, x_2) = \mathcal{N}\left(\begin{bmatrix} 0 \\ 2 \end{bmatrix}, \begin{bmatrix} 0.3 & -1 \\ -1 & 5 \end{bmatrix}\right).$$

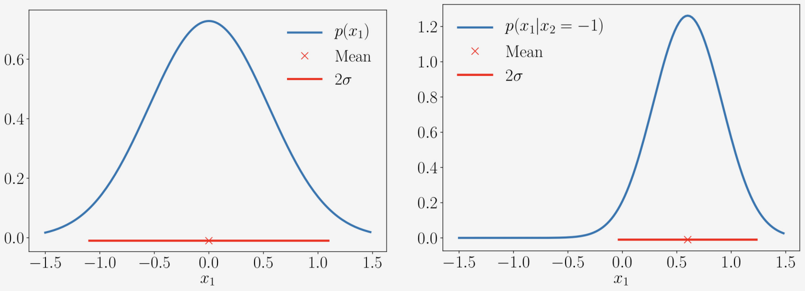


Conditioned on $x_2 = -1$, $\mu_{x_1|x_2=-1} = 0 + (-1) \cdot 0.2 \cdot (-1 - 2) = 0.6$ and $\sigma^2_{x_1|x_2=-1} = 0.3 - (-1) \cdot 0.2 \cdot (-1) = 0.1$.

Thus, $p(x_1 \mid x_2 = -1) = \mathcal{N}(0.6, 0.1)$, $\quad p(x_1) =$

Example

Consider $$p(x_1, x_2) = \mathcal{N}\left(\begin{bmatrix} 0 \\ 2 \end{bmatrix}, \begin{bmatrix} 0.3 & -1 \\ -1 & 5 \end{bmatrix} \right).$$



Conditioned on $x_2 = -1$, $\mu_{x_1|x_2=-1} = 0 + (-1) \cdot 0.2 \cdot (-1 - 2) = 0.6$ and $\sigma^2_{x_1|x_2=-1} = 0.3 - (-1) \cdot 0.2 \cdot (-1) = 0.1$.

Thus, $p(x_1 \mid x_2 = -1) = \mathcal{N}(0.6, 0.1)$, $\quad p(x_1) = \mathcal{N}(0, 0.3)$.

Outline

Sum of Gaussians

Say X, Y are two independent Gaussian random variables with

$$X \sim \mathcal{N}(\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x) \text{ and } Y \sim \mathcal{N}(\boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y).$$

Sum of Gaussians

Say X, Y are two independent Gaussian random variables with

$$X \sim \mathcal{N}(\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x) \text{ and } Y \sim \mathcal{N}(\boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y).$$

- independency: $p(\mathbf{x}, \mathbf{y}) = p(\mathbf{x})p(\mathbf{y})$.

Sum of Gaussians

Say X, Y are two independent Gaussian random variables with

$$X \sim \mathcal{N}(\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x) \text{ and } Y \sim \mathcal{N}(\boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y).$$

- independency: $p(\mathbf{x}, \mathbf{y}) = p(\mathbf{x})p(\mathbf{y})$.

Then $X + Y$ is also a Gaussian distribution with

$$X + Y \sim \mathcal{N}(\boldsymbol{\mu}_x + \boldsymbol{\mu}_y, \boldsymbol{\Sigma}_x + \boldsymbol{\Sigma}_y)$$

Sum of Gaussians

Say X, Y are two independent Gaussian random variables with

$$X \sim \mathcal{N}(\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x) \text{ and } Y \sim \mathcal{N}(\boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y).$$

- independency: $p(\mathbf{x}, \mathbf{y}) = p(\mathbf{x})p(\mathbf{y})$.

Then $X + Y$ is also a Gaussian distribution with

$$X + Y \sim \mathcal{N}(\boldsymbol{\mu}_x + \boldsymbol{\mu}_y, \boldsymbol{\Sigma}_x + \boldsymbol{\Sigma}_y)$$

Please recall $\mathbb{E}[\mathbf{x} + \mathbf{y}]$ and $\mathbb{V}[\mathbf{x} + \mathbf{y}]$.

Example

Linear Combination of Gaussians

$$p(a\mathbf{x} + b\mathbf{y}) =$$

Example

Linear Combination of Gaussians

$$p(a\mathbf{x} + b\mathbf{y}) = \mathcal{N}(a\boldsymbol{\mu}_x + b\boldsymbol{\mu}_y,$$

Example

Linear Combination of Gaussians

$$p(a\mathbf{x} + b\mathbf{y}) = \mathcal{N}(a\boldsymbol{\mu}_x + b\boldsymbol{\mu}_y,\ a^2\boldsymbol{\Sigma}_x + b^2\boldsymbol{\Sigma}_y).$$

Example

Linear Combination of Gaussians

$$p(a\mathbf{x} + b\mathbf{y}) = \mathcal{N}(a\boldsymbol{\mu}_x + b\boldsymbol{\mu}_y,\ a^2\boldsymbol{\Sigma}_x + b^2\boldsymbol{\Sigma}_y).$$

Theorem [Mixture of Two Univariate Gaussian Densities]

Consider a mixture of two univariate Gaussian densities

$$p(x) = \alpha p_1(x) + (1-\alpha)p_2(x_2)$$

for the mixture weight $0 < \alpha < 1$ and $(\mu_1, \sigma_1^2) \neq (\mu_2, \sigma_2^2)$. Then,

$$\begin{aligned} \mathbb{E}[x] &= \alpha\mu_1 + (1-\alpha)\mu_2 \\ \mathbb{V}[x] &= [\alpha\sigma_1^2 + (1-\alpha)\sigma_2^2] \\ &\quad + \left([\alpha\mu_1^2 + (1-\alpha)\mu_2^2] - [\alpha\mu_1 + (1-\alpha)\mu_2]^2\right). \end{aligned}$$

Proof of the Theorem

Sketch:

1. $\mathbb{E}[x] = \int_{-\infty}^{\infty} xp(x)\mathrm{d}x = \int_{-\infty}^{\infty} (\alpha x p_1(x) + (1-\alpha) x p_2(x))\mathrm{d}x.$
2. $\mathbb{E}[x^2] =$

Proof of the Theorem

Sketch:

1. $\mathbb{E}[x] = \int_{-\infty}^{\infty} xp(x)\mathrm{d}x = \int_{-\infty}^{\infty} (\alpha x p_1(x) + (1-\alpha) x p_2(x))\mathrm{d}x.$
2. $\mathbb{E}[x^2] = \int_{-\infty}^{\infty} x^2 p(x)\mathrm{d}x = \int_{-\infty}^{\infty} (\alpha x^2 p_1(x) + (1-\alpha) x^2 p_2(x))\mathrm{d}x.$

Proof of the Theorem

Sketch:

1. $\mathbb{E}[x] = \int_{-\infty}^{\infty} xp(x)\mathrm{d}x = \int_{-\infty}^{\infty} (\alpha x p_1(x) + (1-\alpha) x p_2(x))\mathrm{d}x.$
2. $\mathbb{E}[x^2] = \int_{-\infty}^{\infty} x^2 p(x)\mathrm{d}x = \int_{-\infty}^{\infty} (\alpha x^2 p_1(x) + (1-\alpha) x^2 p_2(x))\mathrm{d}x.$
   - **Recall:** $\mathbb{V}_X[x] = \mathbb{E}_X[x^2] - (\mathbb{E}_X[x])^2$.

Proof of the Theorem

Sketch:

1. $\mathbb{E}[x] = \int_{-\infty}^{\infty} xp(x)\mathrm{d}x = \int_{-\infty}^{\infty} (\alpha x p_1(x) + (1-\alpha) x p_2(x))\mathrm{d}x.$
2. $\mathbb{E}[x^2] = \int_{-\infty}^{\infty} x^2 p(x)\mathrm{d}x = \int_{-\infty}^{\infty} (\alpha x^2 p_1(x) + (1-\alpha) x^2 p_2(x))\mathrm{d}x.$
   - **Recall:** $\mathbb{V}_X[x] = \mathbb{E}_X[x^2] - (\mathbb{E}_X[x])^2$.

Using 1 & 2 we can prove the theorem.

Linear Transformation by a Matrix (1/2)

$X \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\mathbf{y} = \boldsymbol{A}\mathbf{x}$

- The expectation: $\mathbb{E}[\mathbf{y}] = \mathbb{E}[\boldsymbol{A}\mathbf{x}] =$

Linear Transformation by a Matrix (1/2)

$X \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\mathbf{y} = \boldsymbol{A}\mathbf{x}$

- The expectation: $\mathbb{E}[\mathbf{y}] = \mathbb{E}[\boldsymbol{A}\mathbf{x}] = \boldsymbol{A}\mathbb{E}[\mathbf{x}] = \boldsymbol{A}\boldsymbol{\mu}$.

Linear Transformation by a Matrix (1/2)

$X \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\mathbf{y} = \boldsymbol{A}\mathbf{x}$

- The expectation: $\mathbb{E}[\mathbf{y}] = \mathbb{E}[\boldsymbol{A}\mathbf{x}] = \boldsymbol{A}\mathbb{E}[\mathbf{x}] = \boldsymbol{A}\boldsymbol{\mu}$.
- The variance: $\mathbb{V}[\mathbf{y}] = \mathbb{V}[\boldsymbol{A}\mathbf{x}] =$

Linear Transformation by a Matrix (1/2)

$X \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\mathbf{y} = \boldsymbol{A}\mathbf{x}$

- The expectation: $\mathbb{E}[\mathbf{y}] = \mathbb{E}[\boldsymbol{A}\mathbf{x}] = \boldsymbol{A}\mathbb{E}[\mathbf{x}] = \boldsymbol{A}\boldsymbol{\mu}$.
- The variance: $\mathbb{V}[\mathbf{y}] = \mathbb{V}[\boldsymbol{A}\mathbf{x}] = \boldsymbol{A}\mathbb{V}[\mathbf{x}]\boldsymbol{A}^\top = \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^\top$.

Linear Transformation by a Matrix (1/2)

$X \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\mathbf{y} = \boldsymbol{A}\mathbf{x}$

- The expectation: $\mathbb{E}[\mathbf{y}] = \mathbb{E}[\boldsymbol{A}\mathbf{x}] = \boldsymbol{A}\mathbb{E}[\mathbf{x}] = \boldsymbol{A}\boldsymbol{\mu}$.
- The variance: $\mathbb{V}[\mathbf{y}] = \mathbb{V}[\boldsymbol{A}\mathbf{x}] = \boldsymbol{A}\mathbb{V}[\mathbf{x}]\boldsymbol{A}^\top = \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^\top$.
- Thus, we have

Linear Transformation by a Matrix (1/2)

$X \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$ and $\mathbf{y} = \boldsymbol{A}\mathbf{x}$

- The expectation: $\mathbb{E}[\mathbf{y}] = \mathbb{E}[\boldsymbol{A}\mathbf{x}] = \boldsymbol{A}\mathbb{E}[\mathbf{x}] = \boldsymbol{A}\boldsymbol{\mu}$.
- The variance: $\mathbb{V}[\mathbf{y}] = \mathbb{V}[\boldsymbol{A}\mathbf{x}] = \boldsymbol{A}\mathbb{V}[\mathbf{x}]\boldsymbol{A}^\top = \boldsymbol{A}\Sigma\boldsymbol{A}^\top$.
- Thus, we have
$$Y \sim \mathcal{N}(\boldsymbol{A}\boldsymbol{\mu}, \boldsymbol{A}\Sigma\boldsymbol{A}^\top).$$

Linear Transformation by a Matrix (2/2)

Let's consider the *reverse transformation*.

$Y \sim \mathcal{N}(\boldsymbol{\mu}_y, \boldsymbol{\Sigma})$ and $\mathbf{y} = \boldsymbol{A}\mathbf{x}$ for $\mathbf{x}, \mathbf{y} \in \mathbb{R}^M$

- $p(\mathbf{y}) = \mathcal{N}(\mathbf{y} \mid \boldsymbol{A}\mathbf{x}, \boldsymbol{\Sigma})$.
  - **Note:** $\boldsymbol{A}$ might not be invertible...

Linear Transformation by a Matrix (2/2)

Let's consider the *reverse transformation*.

$Y \sim \mathcal{N}(\boldsymbol{\mu}_y, \boldsymbol{\Sigma})$ and $\mathbf{y} = \boldsymbol{A}\mathbf{x}$ for $\mathbf{x}, \mathbf{y} \in \mathbb{R}^M$

- $p(\mathbf{y}) = \mathcal{N}(\mathbf{y} \mid \boldsymbol{A}\mathbf{x}, \boldsymbol{\Sigma})$.
  - **Note:** $\boldsymbol{A}$ might not be invertible...
- $\mathbf{y} = \boldsymbol{A}\mathbf{x}$

Linear Transformation by a Matrix (2/2)

Let's consider the *reverse transformation*.

$Y \sim \mathcal{N}(\boldsymbol{\mu}_y, \boldsymbol{\Sigma})$ and $\mathbf{y} = \boldsymbol{A}\mathbf{x}$ for $\mathbf{x}, \mathbf{y} \in \mathbb{R}^M$

- $p(\mathbf{y}) = \mathcal{N}(\mathbf{y} \mid \boldsymbol{A}\mathbf{x}, \boldsymbol{\Sigma})$.
  - **Note:** $\boldsymbol{A}$ might not be invertible...
- $\mathbf{y} = \boldsymbol{A}\mathbf{x} \iff \boldsymbol{A}^\top \mathbf{y} = \boldsymbol{A}^\top \boldsymbol{A}\mathbf{x}$

Linear Transformation by a Matrix (2/2)

Let's consider the *reverse transformation*.

$Y \sim \mathcal{N}(\boldsymbol{\mu}_y, \boldsymbol{\Sigma})$ and $\mathbf{y} = \boldsymbol{A}\mathbf{x}$ for $\mathbf{x}, \mathbf{y} \in \mathbb{R}^M$

- $p(\mathbf{y}) = \mathcal{N}(\mathbf{y} \mid \boldsymbol{A}\mathbf{x}, \boldsymbol{\Sigma})$.
  - **Note:** $\boldsymbol{A}$ might not be invertible...
- $\mathbf{y} = \boldsymbol{A}\mathbf{x} \iff \boldsymbol{A}^\top \mathbf{y} = \boldsymbol{A}^\top \boldsymbol{A}\mathbf{x} \iff (\boldsymbol{A}^\top \boldsymbol{A})^{-1} \boldsymbol{A}^\top \mathbf{y} = \mathbf{x}$.

Linear Transformation by a Matrix (2/2)

Let's consider the *reverse transformation*.

$Y \sim \mathcal{N}(\boldsymbol{\mu}_y, \boldsymbol{\Sigma})$ and $\mathbf{y} = \boldsymbol{A}\mathbf{x}$ for $\mathbf{x}, \mathbf{y} \in \mathbb{R}^M$

- $p(\mathbf{y}) = \mathcal{N}(\mathbf{y} \mid \boldsymbol{A}\mathbf{x}, \boldsymbol{\Sigma})$.
  - **Note:** $\boldsymbol{A}$ might not be invertible...
- $\mathbf{y} = \boldsymbol{A}\mathbf{x} \iff \boldsymbol{A}^\top \mathbf{y} = \boldsymbol{A}^\top \boldsymbol{A}\mathbf{x} \iff (\boldsymbol{A}^\top \boldsymbol{A})^{-1} \boldsymbol{A}^\top \mathbf{y} = \mathbf{x}$.
  - This works even for non-invertible $\boldsymbol{A}$!.

Linear Transformation by a Matrix (2/2)

Let's consider the *reverse transformation*.

$Y \sim \mathcal{N}(\boldsymbol{\mu}_y, \boldsymbol{\Sigma})$ and $\mathbf{y} = \boldsymbol{A}\mathbf{x}$ for $\mathbf{x}, \mathbf{y} \in \mathbb{R}^M$

- $p(\mathbf{y}) = \mathcal{N}(\mathbf{y} \mid \boldsymbol{A}\mathbf{x}, \boldsymbol{\Sigma})$.
  - **Note:** $\boldsymbol{A}$ might not be invertible...
- $\mathbf{y} = \boldsymbol{A}\mathbf{x} \iff \boldsymbol{A}^\top \mathbf{y} = \boldsymbol{A}^\top \boldsymbol{A}\mathbf{x} \iff (\boldsymbol{A}^\top \boldsymbol{A})^{-1}\boldsymbol{A}^\top \mathbf{y} = \mathbf{x}$.
  - This works even for non-invertible $\boldsymbol{A}$!.
- The variance: $\mathbb{V}[\mathbf{x}] = \mathbb{V}[(\boldsymbol{A}^\top \boldsymbol{A})^{-1}\boldsymbol{A}^\top \mathbf{y}] =$

Linear Transformation by a Matrix (2/2)

Let's consider the *reverse transformation*.

$Y \sim \mathcal{N}(\boldsymbol{\mu}_y, \boldsymbol{\Sigma})$ and $\mathbf{y} = \boldsymbol{A}\mathbf{x}$ for $\mathbf{x}, \mathbf{y} \in \mathbb{R}^M$

- $p(\mathbf{y}) = \mathcal{N}(\mathbf{y} \mid \boldsymbol{A}\mathbf{x}, \boldsymbol{\Sigma})$.
  - **Note:** $\boldsymbol{A}$ might not be invertible...
- $\mathbf{y} = \boldsymbol{A}\mathbf{x} \iff \boldsymbol{A}^\top \mathbf{y} = \boldsymbol{A}^\top \boldsymbol{A}\mathbf{x} \iff (\boldsymbol{A}^\top \boldsymbol{A})^{-1} \boldsymbol{A}^\top \mathbf{y} = \mathbf{x}$.
  - This works even for non-invertible $\boldsymbol{A}$!.
- The variance: $\mathbb{V}[\mathbf{x}] = \mathbb{V}[(\boldsymbol{A}^\top \boldsymbol{A})^{-1} \boldsymbol{A}^\top \mathbf{y}] = (\boldsymbol{A}^\top \boldsymbol{A})^{-1} \boldsymbol{A}^\top \boldsymbol{\Sigma} \boldsymbol{A} (\boldsymbol{A}^\top \boldsymbol{A})^{-1}$.

Linear Transformation by a Matrix (2/2)

Let's consider the *reverse transformation*.

$Y \sim \mathcal{N}(\boldsymbol{\mu}_y, \boldsymbol{\Sigma})$ and $\mathbf{y} = \boldsymbol{A}\mathbf{x}$ for $\mathbf{x}, \mathbf{y} \in \mathbb{R}^M$

- $p(\mathbf{y}) = \mathcal{N}(\mathbf{y} \mid \boldsymbol{A}\mathbf{x}, \boldsymbol{\Sigma})$.
  - **Note:** $\boldsymbol{A}$ might not be invertible...
- $\mathbf{y} = \boldsymbol{A}\mathbf{x} \iff \boldsymbol{A}^\top \mathbf{y} = \boldsymbol{A}^\top \boldsymbol{A}\mathbf{x} \iff (\boldsymbol{A}^\top \boldsymbol{A})^{-1} \boldsymbol{A}^\top \mathbf{y} = \mathbf{x}$.
  - This works even for non-invertible $\boldsymbol{A}$!.
- The variance: $\mathbb{V}[\mathbf{x}] = \mathbb{V}[(\boldsymbol{A}^\top \boldsymbol{A})^{-1} \boldsymbol{A}^\top \mathbf{y}] = (\boldsymbol{A}^\top \boldsymbol{A})^{-1} \boldsymbol{A}^\top \boldsymbol{\Sigma} \boldsymbol{A} (\boldsymbol{A}^\top \boldsymbol{A})^{-1}$.
- Thus, we have

Linear Transformation by a Matrix (2/2)

Let's consider the *reverse transformation*.

$Y \sim \mathcal{N}(\boldsymbol{\mu}_y, \boldsymbol{\Sigma})$ and $\mathbf{y} = \boldsymbol{A}\mathbf{x}$ for $\mathbf{x}, \mathbf{y} \in \mathbb{R}^M$

- $p(\mathbf{y}) = \mathcal{N}(\mathbf{y} \mid \boldsymbol{A}\mathbf{x}, \boldsymbol{\Sigma})$.
  - **Note:** $\boldsymbol{A}$ might not be invertible...
- $\mathbf{y} = \boldsymbol{A}\mathbf{x} \iff \boldsymbol{A}^\top \mathbf{y} = \boldsymbol{A}^\top \boldsymbol{A}\mathbf{x} \iff (\boldsymbol{A}^\top \boldsymbol{A})^{-1} \boldsymbol{A}^\top \mathbf{y} = \mathbf{x}$.
  - This works even for non-invertible $\boldsymbol{A}$!.
- The variance: $\mathbb{V}[\mathbf{x}] = \mathbb{V}[(\boldsymbol{A}^\top \boldsymbol{A})^{-1} \boldsymbol{A}^\top \mathbf{y}] = (\boldsymbol{A}^\top \boldsymbol{A})^{-1} \boldsymbol{A}^\top \boldsymbol{\Sigma} \boldsymbol{A} (\boldsymbol{A}^\top \boldsymbol{A})^{-1}$.
- Thus, we have

$$X \sim \mathcal{N}((\boldsymbol{A}^\top \boldsymbol{A})^{-1} \boldsymbol{A}^\top \boldsymbol{\mu}_y, (\boldsymbol{A}^\top \boldsymbol{A})^{-1} \boldsymbol{A}^\top \boldsymbol{\Sigma} \boldsymbol{A} (\boldsymbol{A}^\top \boldsymbol{A})^{-1}).$$

Exercise

Another example of *reverse transformation*.

$Y \sim \mathcal{N}(\boldsymbol{\mu}_y, \boldsymbol{\Sigma})$ and $\mathbf{y} = \boldsymbol{A}\mathbf{x}$ for $\mathbf{x}, \mathbf{y} \in \mathbb{R}^M$, and $\boldsymbol{A}$ is invertible

- $p(\mathbf{y}) = \mathcal{N}(\mathbf{y} \mid \boldsymbol{A}\mathbf{x}, \boldsymbol{\Sigma})$.
- Compute $\mathbb{E}[\mathbf{x}]$.
- Compute $\mathbb{V}[\mathbf{x}]$.
- Derive $X \sim \mathcal{N}(?,\ ?)$.

A Sampling Approach

We want to obtain samples from a multivariate $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

- However, we only have a sampler of $\mathcal{N}(\mathbf{0}, \boldsymbol{I})$ at hand.

A Sampling Approach

We want to obtain samples from a multivariate $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

- However, we only have a sampler of $\mathcal{N}(\mathbf{0}, \boldsymbol{I})$ at hand.

- Assume that we have $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$.
- Then, define $\mathbf{y} = \boldsymbol{A}\mathbf{x} + \boldsymbol{\mu}$, where $\boldsymbol{A}\boldsymbol{A}^\top = \boldsymbol{\Sigma}$.

A Sampling Approach

We want to obtain samples from a multivariate $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

- However, we only have a sampler of $\mathcal{N}(\mathbf{0}, \boldsymbol{I})$ at hand.

- Assume that we have $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$.
- Then, define $\mathbf{y} = \boldsymbol{A}\mathbf{x} + \boldsymbol{\mu}$, where $\boldsymbol{A}\boldsymbol{A}^\top = \boldsymbol{\Sigma}$.
- To derive $\boldsymbol{A}$:

A Sampling Approach

We want to obtain samples from a multivariate $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

- However, we only have a sampler of $\mathcal{N}(\mathbf{0}, \boldsymbol{I})$ at hand.

- Assume that we have $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$.
- Then, define $\mathbf{y} = \boldsymbol{A}\mathbf{x} + \boldsymbol{\mu}$, where $\boldsymbol{A}\boldsymbol{A}^\top = \boldsymbol{\Sigma}$.
- To derive $\boldsymbol{A}$: Use Cholesky decomposition of the covariance matrix $\boldsymbol{\Sigma}$.
  - $\boldsymbol{A}$ will be triangular and efficient for computation.

Outline

Motivation

Consider the following examples.

- Assuming that X is a random variable distributed according to some well-known distribution, then what is the distribution of X^2?
- Assuming that X_1, X_2 are two univariate standard normal distributions, then what is the distribution of $\frac{1}{2}(X_1 + X_2)$?

Motivation

Consider the following examples.

- Assuming that X is a random variable distributed according to some well-known distribution, then what is the distribution of X^2?
- Assuming that X_1, X_2 are two univariate standard normal distributions, then what is the distribution of $\frac{1}{2}(X_1 + X_2)$?
- What if the transformation is nonlinear?

Motivation

Consider the following examples.

- Assuming that X is a random variable distributed according to some well-known distribution, then what is the distribution of X^2?
- Assuming that X_1, X_2 are two univariate standard normal distributions, then what is the distribution of $\frac{1}{2}(X_1 + X_2)$?
- What if the transformation is nonlinear?
  - Closed-form expressions are not readily available.

Straightforward for Discrete Random Variables

Example: Univariate Random Variables

Given

- A discrete random variable X with pmf $\Pr[X = x]$.
- An invertible function $U(x)$.

Consider the transformed random variable $Y := U(X)$. with pmf $\Pr[Y = y]$. Then

$$\begin{aligned} \Pr[Y = y] &= \Pr[U(X) = y] \quad \text{(transformation of interest)} \\ &= \Pr[X = U^{-1}(y)] \quad \text{(inverse)} \end{aligned}$$

where we can observe $x = U^{-1}(y)$.

Two Approaches

- We consider the discrete case (e.g., $\Pr[X = x]$).
- Two approaches:
  1. Cumulative distribution (Distribution Function Technique).
  2. Change-of-variable.

Outline

Distribution Function Technique

Note: a cdf of X: $F_X(x) = \Pr[X \leq x]$.

Goal: Find the cdf of the random variable $Y := U(X)$

1. Find the cdf
$$F_Y(y) = \Pr[Y \leq y].$$
2. Differentiating $F_Y(y)$ to get the pdf $f_Y(y)$:
$$f_Y(y) = \frac{\mathrm{d}}{\mathrm{d}y} F_Y(y).$$

Distribution Function Technique

Note: a cdf of X: $F_X(x) = \Pr[X \le x]$.

Goal: Find the cdf of the random variable $Y := U(X)$

1. Find the cdf
$$F_Y(y) = \Pr[Y \le y].$$

2. Differentiating $F_Y(y)$ to get the pdf $f_Y(y)$:
$$f_Y(y) = \frac{\mathrm{d}}{\mathrm{d}y} F_Y(y).$$

Note: The domain of the random variable may have changed!

Example

Example

Let X be a continuous random variable with pdf $f_X : [0, 1] \mapsto [0, 1]$:

$$f_X(x) = 3x^2.$$

Goal: Find the pdf of $Y = X^2$.

$$F_Y(y) \quad = \quad \Pr[Y \leq y]$$

Example

Example

Let X be a continuous random variable with pdf $f_X : [0,1] \mapsto [0,1]$:

$$f_X(x) = 3x^2.$$

Goal: Find the pdf of $Y = X^2$.

$$\begin{aligned} F_Y(y) &= \Pr[Y \leq y] = \Pr[X^2 \leq y] \\ &= \Pr[X \leq y^{\frac{1}{2}}] \\ &= F_X(y^{\frac{1}{2}}) \end{aligned}$$

Example

Example

Let X be a continuous random variable with pdf $f_X : [0,1] \mapsto [0,1]$:

$$f_X(x) = 3x^2.$$

Goal: Find the pdf of $Y = X^2$.

$$\begin{aligned} F_Y(y) &= \Pr[Y \le y] = \Pr[X^2 \le y] \\ &= \Pr[X \le y^{\frac{1}{2}}] \\ &= F_X(y^{\frac{1}{2}}) = \int_0^{y^{\frac{1}{2}}} 3t^2 \mathrm{d}t \\ &= [t^3]_0^{y^{\frac{1}{2}}} = y^{\frac{3}{2}}, \quad 0 \le y \le 1. \end{aligned}$$

Example

Example

Let X be a continuous random variable with pdf $f_X : [0,1] \mapsto [0,1]$:

$$f_X(x) = 3x^2.$$

Goal: Find the pdf of $Y = X^2$.

$$\begin{aligned} F_Y(y) &= \Pr[Y \le y] = \Pr[X^2 \le y] \\ &= \Pr[X \le y^{\frac{1}{2}}] \\ &= F_X(y^{\frac{1}{2}}) = \int_0^{y^{\frac{1}{2}}} 3t^2 \mathrm{d}t \\ &= [t^3]_0^{y^{\frac{1}{2}}} = y^{\frac{3}{2}}, \quad 0 \le y \le 1. \end{aligned}$$

Thus,

$$f_Y(y) = \frac{\mathrm{d}}{\mathrm{d}y} F_Y(y) = \frac{3}{2} y^{\frac{1}{2}}$$

for $0 \le y \le 1$.

Exercise

Theorem [Casella & Berger (2002)]

Let X be a continuous random variable with a *strictly monotone* cumulative distribution function $F_X(x)$. Then, the random variable Y defined as

$$Y := F_X(X)$$

has a uniform distribution.

Exercise

Consider $f_X(x) = 3x^2$ in the previous example. Show that $Y := F_X(X)$ attains a uniform distribution.

Remark

The first approach relies on the following facts:

- We can transform the cdf of Y into an expression that is a cdf of X.
- We can differentiate the cdf to obtain the pdf.

Outline

What We have Learnt From the Calculus Course

$$\int f(g(x))g'(x)\mathrm{d}x = \int f(u)\mathrm{d}u, \text{ where } \ u = g(x).$$

What We have Learnt From the Calculus Course

$$\int f(g(x))g'(x)\mathrm{d}x = \int f(u)\mathrm{d}u, \text{ where } u = g(x).$$

- Intuitively, considering $\mathrm{d}u \approx \Delta u = g'(x)\Delta x$ as the "small changes".

The Roadmap (1/2)

- Consider a univariate random variable X and an invertible function U such that $Y := U(X)$.
- Assume that X has states $x \in [a, b]$.
- By the definition of a cdf, we have

The Roadmap (1/2)

- Consider a univariate random variable X and an invertible function U such that $Y := U(X)$.
- Assume that X has states $x \in [a, b]$.
- By the definition of a cdf, we have

$$F_Y(y) = \Pr[Y \leq y]$$

The Roadmap (1/2)

- Consider a univariate random variable X and an invertible function U such that $Y := U(X)$.
- Assume that X has states $x \in [a, b]$.
- By the definition of a cdf, we have

$$F_Y(y) = \Pr[Y \leq y] = \Pr[U(X) \leq y]$$

The Roadmap (1/2)

- Consider a univariate random variable X and an invertible function U such that $Y := U(X)$.
- Assume that X has states $x \in [a, b]$.
- By the definition of a cdf, we have

$$F_Y(y) = \Pr[Y \leq y] = \Pr[U(X) \leq y]$$

If U is *strictly increasing*, then so is its inverse U^{-1}.

$$\Pr[U(X) \leq y] = \Pr[U^{-1}(U(X)) \leq U^{-1}(y)]$$

The Roadmap (1/2)

- Consider a univariate random variable X and an invertible function U such that $Y := U(X)$.
- Assume that X has states $x \in [a, b]$.
- By the definition of a cdf, we have

$$F_Y(y) = \Pr[Y \leq y] = \Pr[U(X) \leq y]$$

If U is *strictly increasing*, then so is its inverse U^{-1}.

$$\Pr[U(X) \leq y] = \Pr[U^{-1}(U(X)) \leq U^{-1}(y)] = \Pr[X \leq U^{-1}(y)].$$

Then, $\quad F_Y(y) = \Pr[X \leq U^{-1}(y)] = \int_a^{U^{-1}(y)} f_X(x)\mathrm{d}x$

The Roadmap (2/2)

- To obtain the pdf, we differentiate $F_Y(y)$ w.r.t. y:

$$f_Y(y) = \frac{\mathrm{d}}{\mathrm{d}y} F_Y(y) = \frac{\partial}{\partial f} \int_a^{U^{-1}(y)} f_X(x)\mathrm{d}x.$$

The Roadmap (2/2)

- To obtain the pdf, we differentiate $F_Y(y)$ w.r.t. y:

$$f_Y(y) = \frac{\mathrm{d}}{\mathrm{d}y} F_Y(y) = \frac{\partial}{\partial f} \int_a^{U^{-1}(y)} f_X(x) \mathrm{d}x.$$

  - The integral on the right-hand side is w.r.t. x, but we need an integral w.r.t. y ($\because$ we are differentiating w.r.t. y...)
  - Change-of-variable comes to the rescue!

The Roadmap (2/2)

- To obtain the pdf, we differentiate $F_Y(y)$ w.r.t. y:

$$f_Y(y) = \frac{\mathrm{d}}{\mathrm{d}y} F_Y(y) = \frac{\partial}{\partial f} \int_a^{U^{-1}(y)} f_X(x)\mathrm{d}x.$$

  - The integral on the right-hand side is w.r.t. x, but we need an integral w.r.t. y ($\because$ we are differentiating w.r.t. y...)
  - Change-of-variable comes to the rescue!

- $\int f_X(U^{-1}(y)) {U^{-1}}'(y)\mathrm{d}y = \int f_X(x)\mathrm{d}x$, where $x = U^{-1}(y)$.

The Roadmap (2/2)

- To obtain the pdf, we differentiate $F_Y(y)$ w.r.t. y:

$$f_Y(y) = \frac{\mathrm{d}}{\mathrm{d}y} F_Y(y) = \frac{\partial}{\partial f} \int_a^{U^{-1}(y)} f_X(x)\mathrm{d}x.$$

  - The integral on the right-hand side is w.r.t. x, but we need an integral w.r.t. y ($\because$ we are differentiating w.r.t. y...)
  - Change-of-variable comes to the rescue!

- $\displaystyle\int f_X(U^{-1}(y))U^{-1\prime}(y)\mathrm{d}y = \int f_X(x)\mathrm{d}x$, where $x = U^{-1}(y)$.

- Thus,

$$\begin{aligned} f_Y(y) &= \frac{\partial}{\partial y} \int_a^{U^{-1}(y)} f_X(U^{-1}(y))U^{-1\prime}(y)\mathrm{d}y \\ &= f_X(U^{-1}(y)) \cdot \left(\frac{\partial}{\partial y} U^{-1}(y) \right). \end{aligned}$$

The Main Theorem

Theorem [Billingsley (1995)]

Let $f_X(\mathbf{x})$ be the pdf of the multivariate continuous random variable X. If the vector-valued function $\mathbf{y} = U(\mathbf{x})$ is differentiable and invertible for all values within the domain of $\mathbf{x}$, then for corresponding values of $\mathbf{y}$, the pdf of $Y = U(X)$ is given by

$$f(\mathbf{y}) = f_{\mathbf{x}}(U^{-1}(\mathbf{y})) \cdot \left| \det \left(\frac{\partial}{\partial \mathbf{y}} U^{-1}(\mathbf{y}) \right) \right| .$$

Example (On the Backboard)

Example

Consider a bivariate random variable X with states $\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$ and pdf

$$f\left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}\right) = \frac{1}{2\pi} \exp\left(-\frac{1}{2}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}^\top \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}\right).$$

Then, consider a matrix $\boldsymbol{A} \in \mathbb{R}^{2\times 2}$ defined as

$$\boldsymbol{A} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}.$$

Goal: Find the pdf of the random variable Y with states $\mathbf{y} = \boldsymbol{A}\mathbf{x}$.

Discussions