Mathematics for Machine Learning

— Probability & Distributions

Gaussian Distribution & Change of Variables/Inverse Transform

Joseph Chuang-Chieh Lin

Department of Computer Science & Engineering,
National Taiwan Ocean University

Fall 2025

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.*
  - *Arnold J. Insel, Lawrence E. Spence, Stephen H. Friedberg: Linear Algebra, 4th Edition. Prentice Hall. 2013.*
  - *Howard Anton, Chris Rorres, Anton Kaul: Elementary Linear Algebra, 12th Edition. Wiley. 2019.*
- We could partially refer to the monograph:
  Francesco Orabona: A Modern Introduction to Online Learning. https://arxiv.org/abs/1912.13213

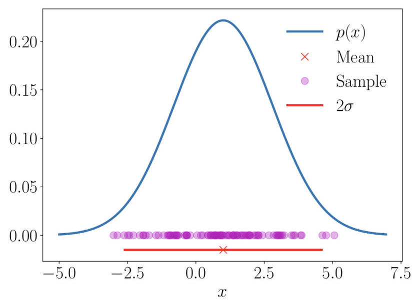
Outline

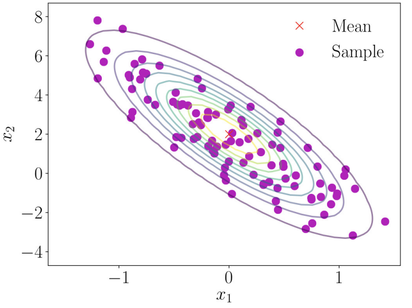
Outline

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

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

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

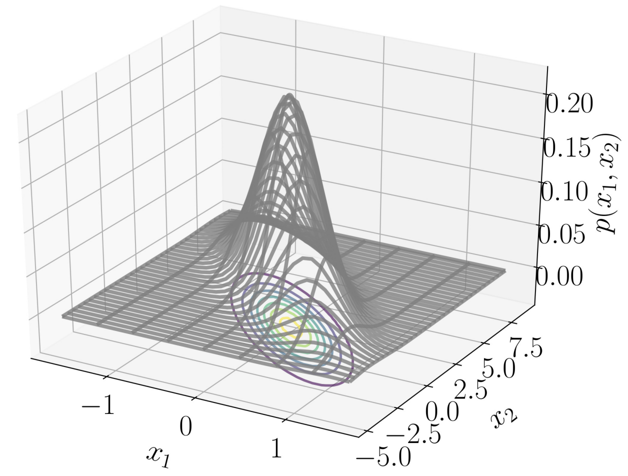
Multivariate

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

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

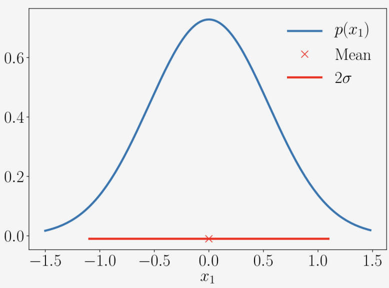
- By [Bishop 2006], the conditional distribution $p(\mathbf{x} \mid \mathbf{y})$ is also Gaussian.

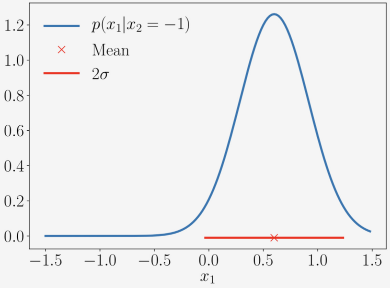
$$\begin{aligned} p(\mathbf{x} \mid \mathbf{y}) &= \mathcal{N}(\boldsymbol{\mu}_{x|y}, \boldsymbol{\Sigma}_{x|y}) \\ \boldsymbol{\mu}_{x|y} &= \boldsymbol{\mu}_x + \boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_{yy}^{-1}(\mathbf{y} - \boldsymbol{\mu}_y) \\ \boldsymbol{\Sigma}_{x|y} &= \boldsymbol{\Sigma}_{xx}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

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

Example

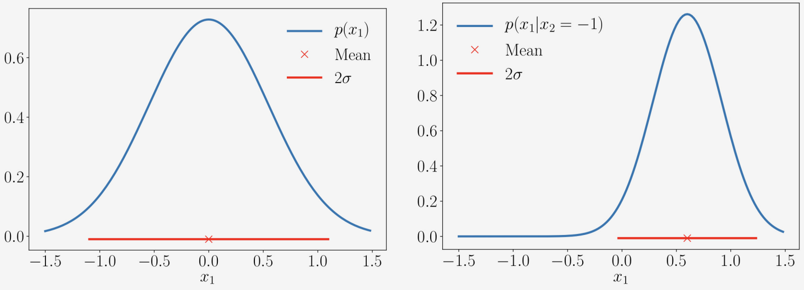
Consider $$p(x_1, x_2) = \mathcal{N}\left(\begin{bmatrix} 0 \\ 2 \end{bmatrix}, \begin{bmatrix} 0.3 & -1 \\ -1 & 5 \end{bmatrix} \right).$$





Example

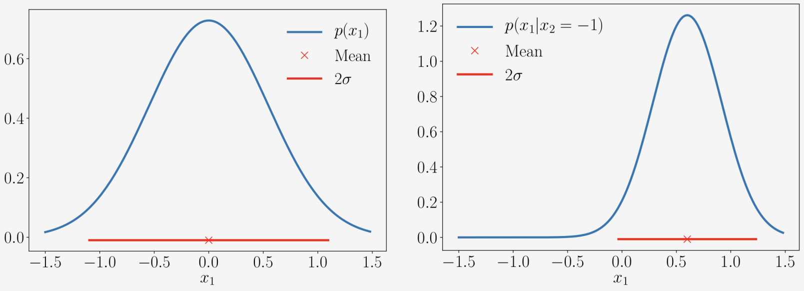
Consider $$p(x_1, x_2) = \mathcal{N}\left(\begin{bmatrix} 0 \\ 2 \end{bmatrix}, \begin{bmatrix} 0.3 & -1 \\ -1 & 5 \end{bmatrix}\right).$$



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

Example

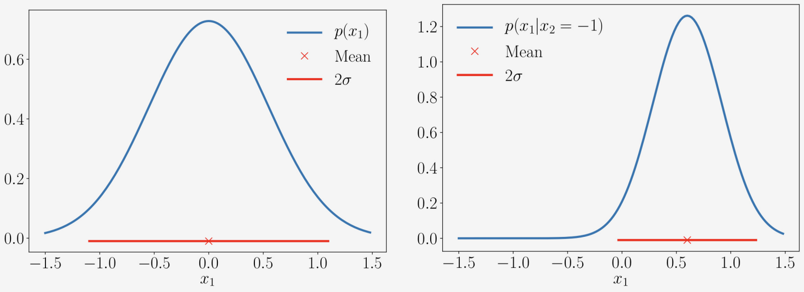
Consider $$p(x_1, x_2) = \mathcal{N}\left(\begin{bmatrix} 0 \\ 2 \end{bmatrix}, \begin{bmatrix} 0.3 & -1 \\ -1 & 5 \end{bmatrix} \right).$$



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

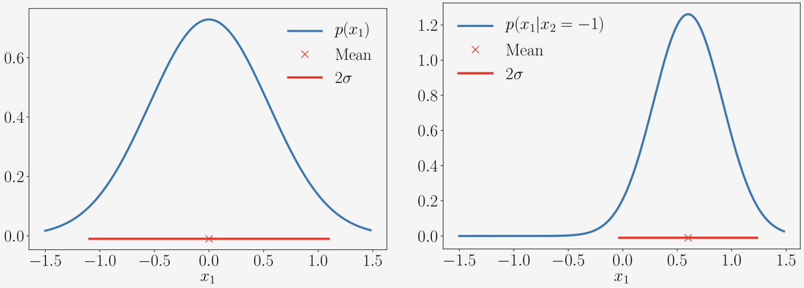


Conditioned on $x_2 = -1$, $\mu_{x_1|x_2=-1} = 0 + (-1) \cdot 0.2 \cdot (-1 - 2) = 0.6$ and $\sigma^2_{x_1|x_2=-1} = 0.3 - (-1) \cdot 0.2 \cdot (-1) = 0.1$.

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

Example

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

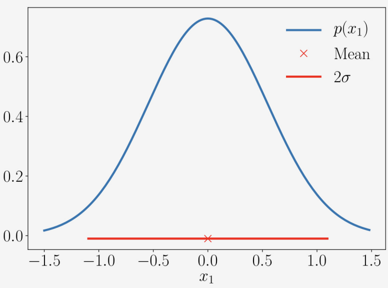


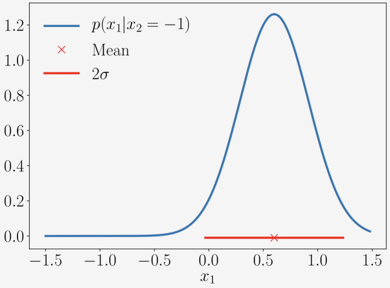
Conditioned on $x_2 = -1$, $\mu_{x_1|x_2=-1} = 0 + (-1) \cdot 0.2 \cdot (-1 - 2) = 0.6$ and $\sigma^2_{x_1|x_2=-1} = 0.3 - (-1) \cdot 0.2 \cdot (-1) = 0.1$.

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

Example

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



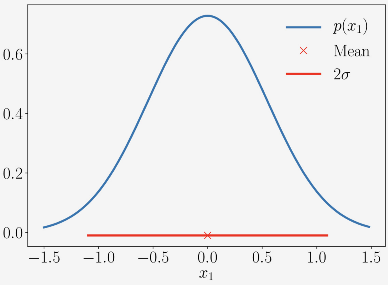


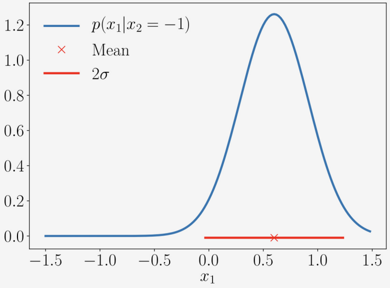
Conditioned on $x_2 = -1$, $\mu_{x_1|x_2=-1} = 0 + (-1) \cdot 0.2 \cdot (-1 - 2) = 0.6$ and $\sigma^2_{x_1|x_2=-1} = 0.3 - (-1) \cdot 0.2 \cdot (-1) = 0.1$.

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 Two Independent Gaussians

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

Example

Linear Combination of Two Independent Gaussians

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

Example

Linear Combination of Two Independent Gaussians

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

Example

Linear Combination of Two Independent Gaussians

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

Theorem [Mixture of Two Univariate Gaussian Densities]

Consider a mixture of two univariate Gaussian densities

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

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. $$\mathbb{E}[x] = \int_{-\infty}^{\infty} xp(x)\mathrm{d}x = \int_{-\infty}^{\infty} (\alpha x p_1(x) + (1-\alpha) x p_2(x))\mathrm{d}x$$
   $$= \alpha\mu_1 + (1-\alpha)\mu_2.$$
2. $\mathbb{E}[x^2] =$

Proof of the Theorem

Sketch:

1. $$\mathbb{E}[x] = \int_{-\infty}^{\infty} xp(x)\mathrm{d}x = \int_{-\infty}^{\infty} (\alpha x p_1(x) + (1-\alpha) x p_2(x))\mathrm{d}x = \alpha\mu_1 + (1-\alpha)\mu_2.$$
2. $$\mathbb{E}[x^2] = \int_{-\infty}^{\infty} x^2 p(x)\mathrm{d}x = \int_{-\infty}^{\infty} (\alpha x^2 p_1(x) + (1-\alpha) x^2 p_2(x))\mathrm{d}x = \alpha(\mu_1^2 + \sigma_1^2) + (1-\alpha)(\mu_2^2 + \sigma_2^2).$$

Proof of the Theorem

Sketch:

1. $$\mathbb{E}[x] = \int_{-\infty}^{\infty} xp(x)\mathrm{d}x = \int_{-\infty}^{\infty} (\alpha x p_1(x) + (1-\alpha) x p_2(x))\mathrm{d}x$$
$$= \alpha\mu_1 + (1-\alpha)\mu_2.$$
2. $$\mathbb{E}[x^2] = \int_{-\infty}^{\infty} x^2 p(x)\mathrm{d}x = \int_{-\infty}^{\infty} (\alpha x^2 p_1(x) + (1-\alpha) x^2 p_2(x))\mathrm{d}x$$
$$= \alpha(\mu_1^2 + \sigma_1^2) + (1-\alpha)(\mu_2^2 + \sigma_2^2).$$
   - **Recall:** $\mathbb{V}_X[x] = \mathbb{E}_X[x^2] - (\mathbb{E}_X[x])^2$.

Proof of the Theorem

Sketch:

1. $$\mathbb{E}[x] = \int_{-\infty}^{\infty} xp(x)\mathrm{d}x = \int_{-\infty}^{\infty} (\alpha x p_1(x) + (1-\alpha) x p_2(x))\mathrm{d}x$$
$$= \alpha\mu_1 + (1-\alpha)\mu_2.$$
2. $$\mathbb{E}[x^2] = \int_{-\infty}^{\infty} x^2 p(x)\mathrm{d}x = \int_{-\infty}^{\infty} (\alpha x^2 p_1(x) + (1-\alpha) x^2 p_2(x))\mathrm{d}x$$
$$= \alpha(\mu_1^2 + \sigma_1^2) + (1-\alpha)(\mu_2^2 + \sigma_2^2).$$
   - **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)

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

Linear Transformation by a Matrix (2/2)

Let's consider the reverse transformation.

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

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

Linear Transformation by a Matrix (2/2)

Let's consider the reverse transformation.

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

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

Linear Transformation by a Matrix (2/2)

Let's consider the reverse transformation.

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

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

Linear Transformation by a Matrix (2/2)

Let's consider the reverse transformation.

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

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

Linear Transformation by a Matrix (2/2)

Let's consider the reverse transformation.

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

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

Linear Transformation by a Matrix (2/2)

Let's consider the reverse transformation.

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

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

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

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

Linear Transformation by a Matrix (2/2)

Let's consider the reverse transformation.

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

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

Linear Transformation by a Matrix (2/2)

Let's consider the reverse transformation.

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

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

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

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 $\boldsymbol{A}\boldsymbol{I}\boldsymbol{A}^\top = \boldsymbol{A}\boldsymbol{A}^\top = \boldsymbol{\Sigma}$.

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 $\boldsymbol{A}\boldsymbol{I}\boldsymbol{A}^\top = \boldsymbol{A}\boldsymbol{A}^\top = \boldsymbol{\Sigma}$.
- 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 $\boldsymbol{A}\boldsymbol{I}\boldsymbol{A}^\top = \boldsymbol{A}\boldsymbol{A}^\top = \boldsymbol{\Sigma}$.
- To derive $\boldsymbol{A}$: Use Cholesky decomposition of the covariance matrix $\boldsymbol{\Sigma}$.
  - $\boldsymbol{A}$ will be triangular and efficient for computation.

Outline

Product of Gaussian Densities: Statement

Product of Gaussians

- Let $x \in \mathbb{R}^D$, and consider two Gaussians

$$\mathcal{N}(x \mid a, A), \qquad \mathcal{N}(x \mid b, B),$$

where $A, B \in \mathbb{R}^{D \times D}$ are positive definite.

Product of Gaussian Densities: Statement

Product of Gaussians

- Let $x \in \mathbb{R}^D$, and consider two Gaussians
$$\mathcal{N}(x \mid a, A), \qquad \mathcal{N}(x \mid b, B),$$
where $A, B \in \mathbb{R}^{D\times D}$ are positive definite.
- Their product can be written as
$$\mathcal{N}(x \mid a, A)\,\mathcal{N}(x \mid b, B) = c\,\mathcal{N}(x \mid m, S),$$
with
$$S = (A^{-1} + B^{-1})^{-1}, \qquad m = S(A^{-1}a + B^{-1}b),$$
and
$$c = \mathcal{N}(a \mid b, A+B) = \mathcal{N}(b \mid a, A+B).$$

Proof Step 1: Completing the Square

- Write both Gaussians explicitly:

$$\mathcal{N}(x \mid a, A) = (2\pi)^{-D/2}|A|^{-1/2} \exp\Big(-\tfrac{1}{2}(x-a)^\top A^{-1}(x-a)\Big),$$
$$\mathcal{N}(x \mid b, B) = (2\pi)^{-D/2}|B|^{-1/2} \exp\Big(-\tfrac{1}{2}(x-b)^\top B^{-1}(x-b)\Big).$$

Proof Step 1: Completing the Square

- Write both Gaussians explicitly:

$$\mathcal{N}(x \mid a, A) = (2\pi)^{-D/2}|A|^{-1/2} \exp\Big(-\tfrac{1}{2}(x-a)^\top A^{-1}(x-a)\Big),$$
$$\mathcal{N}(x \mid b, B) = (2\pi)^{-D/2}|B|^{-1/2} \exp\Big(-\tfrac{1}{2}(x-b)^\top B^{-1}(x-b)\Big).$$

- Their product is

$$\mathcal{N}(x \mid a, A)\,\mathcal{N}(x \mid b, B) = (2\pi)^{-D}|A|^{-1/2}|B|^{-1/2} \cdot \exp\Big(-\tfrac{1}{2}\big[(x-a)^\top A^{-1}(x-a) + (x-b)^\top B^{-1}(x-b)\big]\Big).$$

Step 1: Completing the Square (cont.)

- Expand the exponent. Let $P_A = A^{-1}$, $P_B = B^{-1}$:

$$\begin{aligned}&(x-a)^\top P_A(x-a) + (x-b)^\top P_B(x-b)\\ &= x^\top(P_A+P_B)x - 2x^\top(P_Aa + P_Bb) + a^\top P_A a + b^\top P_B b.\end{aligned}$$

Step 1: Completing the Square (cont.)

- Expand the exponent. Let $P_A = A^{-1}$, $P_B = B^{-1}$:

$$\begin{aligned}&(x-a)^\top P_A(x-a) + (x-b)^\top P_B(x-b)\\ &= x^\top (P_A + P_B)x - 2x^\top (P_A a + P_B b) + a^\top P_A a + b^\top P_B b.\end{aligned}$$

- Define

$$P := P_A + P_B = A^{-1} + B^{-1}, \quad S := P^{-1}, \quad h := P_A a + P_B b.$$

- We complete the square by choosing m such that $Pm = h$:

$$m = P^{-1}h = S(A^{-1}a + B^{-1}b).$$

Then

$$x^\top Px - 2x^\top h = (x-m)^\top P(x-m) - m^\top Pm.$$

Step 1: Completing the Square (cont.)

- Expand the exponent. Let $P_A = A^{-1}$, $P_B = B^{-1}$:

$$\begin{aligned}(x-a)^\top P_A(x-a) + (x-b)^\top P_B(x-b) \\ = x^\top(P_A+P_B)x - 2x^\top(P_A a + P_B b) + a^\top P_A a + b^\top P_B b.\end{aligned}$$

- Define

$$P := P_A + P_B = A^{-1} + B^{-1}, \quad S := P^{-1}, \quad h := P_A a + P_B b.$$

- We complete the square by choosing m such that $Pm = h$:

$$m = P^{-1}h = S(A^{-1}a + B^{-1}b).$$

  Then

$$x^\top Px - 2x^\top h = (x-m)^\top P(x-m) - m^\top Pm.$$

- Hence

$$\begin{aligned}\mathcal{N}(x \mid a, A)\,\mathcal{N}(x \mid b, B) = C_0 \exp\big(-\tfrac{1}{2}(x-m)^\top P(x-m)\big) \\ \cdot \exp\Big(-\tfrac{1}{2}\big[a^\top P_A a + b^\top P_B b - m^\top Pm\big]\Big),\end{aligned}$$

Proof Step 1: Identifying $\mathcal{N}(x \mid m, S)$

- Using $P = S^{-1}$, we recognize a Gaussian in x:

$$\exp\big(-\tfrac{1}{2}(x-m)^\top P(x-m)\big) = (2\pi)^{D/2}|S|^{1/2}\mathcal{N}(x \mid m, S).$$

Proof Step 1: Identifying $\mathcal{N}(x \mid m, S)$

- Using $P = S^{-1}$, we recognize a Gaussian in x:

$$\exp\big(-\tfrac{1}{2}(x-m)^\top P(x-m)\big) = (2\pi)^{D/2}|S|^{1/2}\mathcal{N}(x \mid m, S).$$

- Therefore

$$\mathcal{N}(x \mid a, A)\,\mathcal{N}(x \mid b, B) = c\,\mathcal{N}(x \mid m, S),$$

where

$$\begin{aligned} c &= (2\pi)^{-D/2}|A|^{-1/2}|B|^{-1/2}|S|^{1/2} \\ &\quad \times \exp\Big(-\tfrac{1}{2}\big[a^\top A^{-1}a + b^\top B^{-1}b - m^\top S^{-1}m\big]\Big), \end{aligned}$$

and we have already identified

$$S = (A^{-1} + B^{-1})^{-1}, \qquad m = S(A^{-1}a + B^{-1}b).$$

- It remains to show that this c is equal to $\mathcal{N}(a \mid b, A + B)$.

Proof Step 2: Determining the Constant c (1/2)

- Integrate both sides over x:

$$\int \mathcal{N}(x \mid a, A)\,\mathcal{N}(x \mid b, B)\,\mathrm{d}x = c \int \mathcal{N}(x \mid m, S)\,\mathrm{d}x = c,$$

since $\mathcal{N}(x \mid m, S)$ is normalized.

Proof Step 2: Determining the Constant c (1/2)

- Integrate both sides over x:

$$\int \mathcal{N}(x \mid a, A)\,\mathcal{N}(x \mid b, B)\,\mathrm{d}x = c\int \mathcal{N}(x \mid m, S)\,\mathrm{d}x = c,$$

  since $\mathcal{N}(x \mid m, S)$ is normalized.

- Give a probabilistic interpretation. Let

$$X \sim \mathcal{N}(b, B), \qquad a = X + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, A),$$

  with X and ε independent.

Proof Step 2: Determining the Constant c (1/2)

- Integrate both sides over x:

$$\int \mathcal{N}(x \mid a, A)\, \mathcal{N}(x \mid b, B)\, \mathrm{d}x = c \int \mathcal{N}(x \mid m, S)\, \mathrm{d}x = c,$$

since $\mathcal{N}(x \mid m, S)$ is normalized.

- Give a probabilistic interpretation. Let

$$X \sim \mathcal{N}(b, B), \qquad a = X + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, A),$$

with X and ε independent.

- The joint density of (X, a) is

$$p(X, a) = p(a \mid X)\, p(X) = \mathcal{N}(a \mid X, A)\, \mathcal{N}(X \mid b, B).$$

As a function of X, this is precisely $\mathcal{N}(X \mid a, A)\, \mathcal{N}(X \mid b, B)$.

Proof Step 2: Determining the Constant c (1/2)

- Integrate both sides over x:

$$\int \mathcal{N}(x \mid a, A)\,\mathcal{N}(x \mid b, B)\,\mathrm{d}x = c \int \mathcal{N}(x \mid m, S)\,\mathrm{d}x = c,$$

  since $\mathcal{N}(x \mid m, S)$ is normalized.

- Give a probabilistic interpretation. Let

$$X \sim \mathcal{N}(b, B), \qquad a = X + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, A),$$

  with X and ε independent.

- The joint density of (X, a) is

$$p(X, a) = p(a \mid X)\,p(X) = \mathcal{N}(a \mid X, A)\,\mathcal{N}(X \mid b, B).$$

  As a function of X, this is precisely $\mathcal{N}(X \mid a, A)\,\mathcal{N}(X \mid b, B)$.

- The marginal density of a is Gaussian with mean b and covariance $A + B$:

$$a \sim \mathcal{N}(b, A + B) \quad \Rightarrow \quad p(a) = \mathcal{N}(a \mid b, A + B).$$

Proof Step 2: Determining the Constant c (2/2)

- But

$$p(a) = \int p(X, a)\,\mathrm{d}X = \int \mathcal{N}(x \mid a, A)\,\mathcal{N}(x \mid b, B)\,\mathrm{d}x.$$

Proof Step 2: Determining the Constant c (2/2)

- But

$$p(a) = \int p(X, a)\,\mathrm{d}X = \int \mathcal{N}(x \mid a, A)\,\mathcal{N}(x \mid b, B)\,\mathrm{d}x.$$

- Hence

$$c = \mathcal{N}(a \mid b, A + B),$$

and, by symmetry in a and b, also $c = \mathcal{N}(b \mid a, A + B)$.

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

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

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

Two Approaches

- So far we considered the discrete case (e.g., $\Pr[X = x]$).
- For continuous distributions, we will consider the 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: $F_X(x) = \Pr[X \leq x]$.

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 $f_Y(y)$:
$$f_Y(y) = \frac{\mathrm{d}}{\mathrm{d}y} F_Y(y).$$

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

Example

Example

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

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

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

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

Example

Example

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

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

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 $f_X : [0,1] \to [0,1]$:

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

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 $f_X : [0,1] \to [0,1]$:

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

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:

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

The Roadmap (2/2)

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

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

  - 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:

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

  - 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:

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

  - 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!

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

- Thus,

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

Remark

For decreasing functions,

$$f_Y(y) = -f_X(U^{-1}(y)) \cdot \left(\frac{\mathrm{d}}{\mathrm{d}y} U^{-1}(y)\right).$$

Remark

For decreasing functions,

$$f_Y(y) \;=\; -f_X(U^{-1}(y)) \cdot \left(\frac{\mathrm{d}}{\mathrm{d}y} U^{-1}(y)\right).$$

So for both increasing and decreasing U,

$$f_Y(y) \;=\; f_X(U^{-1}(y)) \cdot \left|\frac{\mathrm{d}}{\mathrm{d}y} U^{-1}(y)\right|.$$

Remark

For decreasing functions,

$$f_Y(y) \;=\; -f_X(U^{-1}(y)) \cdot \left(\frac{\mathrm{d}}{\mathrm{d}y} U^{-1}(y)\right).$$

So for both increasing and decreasing U,

$$f_Y(y) \;=\; f_X(U^{-1}(y)) \cdot \left|\frac{\mathrm{d}}{\mathrm{d}y} U^{-1}(y)\right|.$$

- The term $\left|\frac{\mathrm{d}}{\mathrm{d}y} U^{-1}(y)\right|$ measures how much a unit volume changes when applying U.

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

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 $\mathbf{y} = \boldsymbol{A}\mathbf{x}$.

- $\mathbf{y} = \boldsymbol{A}\mathbf{x}$

- $\mathbf{y} = \boldsymbol{A}\mathbf{x} \implies \mathbf{x} = \boldsymbol{A}^{-1}\mathbf{y}$.

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \boldsymbol{A}^{-1} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}$$

- $\mathbf{y} = \boldsymbol{A}\mathbf{x} \implies \mathbf{x} = \boldsymbol{A}^{-1}\mathbf{y}$.

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \boldsymbol{A}^{-1} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}.$$

- $\mathbf{y} = \boldsymbol{A}\mathbf{x} \implies \mathbf{x} = \boldsymbol{A}^{-1}\mathbf{y}$.

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \boldsymbol{A}^{-1} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \frac{1}{ad-bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}.$$

- The corresponding pdf is given by

$$f(\mathbf{x}) = f(\boldsymbol{A}^{-1}\mathbf{y}) = \frac{1}{2\pi} \exp\left(-\frac{1}{2}\mathbf{y}^\top (\boldsymbol{A}^{-1})^\top \boldsymbol{A}^{-1}\mathbf{y}\right)$$

- $\mathbf{y} = \boldsymbol{A}\mathbf{x} \implies \mathbf{x} = \boldsymbol{A}^{-1}\mathbf{y}$.

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \boldsymbol{A}^{-1} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}.$$

- The corresponding pdf is given by

$$f(\mathbf{x}) = f(\boldsymbol{A}^{-1}\mathbf{y}) = \frac{1}{2\pi} \exp\left(-\frac{1}{2}\mathbf{y}^\top (\boldsymbol{A}^{-1})^\top \boldsymbol{A}^{-1}\mathbf{y} \right)$$

- $\frac{\partial}{\partial \mathbf{y}} \boldsymbol{A}^{-1}\mathbf{y} =$

- $\mathbf{y} = \boldsymbol{A}\mathbf{x} \implies \mathbf{x} = \boldsymbol{A}^{-1}\mathbf{y}$.

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \boldsymbol{A}^{-1} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}.$$

- The corresponding pdf is given by

$$f(\mathbf{x}) = f(\boldsymbol{A}^{-1}\mathbf{y}) = \frac{1}{2\pi} \exp\left(-\frac{1}{2}\mathbf{y}^\top (\boldsymbol{A}^{-1})^\top \boldsymbol{A}^{-1}\mathbf{y}\right)$$

- $\dfrac{\partial}{\partial \mathbf{y}}\boldsymbol{A}^{-1}\mathbf{y} = \boldsymbol{A}^{-1}$. So, $\det\left(\dfrac{\partial}{\partial \mathbf{y}}\boldsymbol{A}^{-1}\mathbf{y}\right) = \det(\boldsymbol{A}^{-1}) =$

- $\mathbf{y} = \boldsymbol{A}\mathbf{x} \implies \mathbf{x} = \boldsymbol{A}^{-1}\mathbf{y}$.

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \boldsymbol{A}^{-1} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}.$$

- The corresponding pdf is given by

$$f(\mathbf{x}) = f(\boldsymbol{A}^{-1}\mathbf{y}) = \frac{1}{2\pi} \exp\left(-\frac{1}{2}\mathbf{y}^\top (\boldsymbol{A}^{-1})^\top \boldsymbol{A}^{-1}\mathbf{y}\right)$$

- $\frac{\partial}{\partial \mathbf{y}} \boldsymbol{A}^{-1}\mathbf{y} = \boldsymbol{A}^{-1}$. So, $\det\left(\frac{\partial}{\partial \mathbf{y}} \boldsymbol{A}^{-1}\mathbf{y}\right) = \det(\boldsymbol{A}^{-1}) = \frac{1}{ad - bc}$.

- $\mathbf{y} = \boldsymbol{A}\mathbf{x} \implies \mathbf{x} = \boldsymbol{A}^{-1}\mathbf{y}$.

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \boldsymbol{A}^{-1} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}.$$

- The corresponding pdf is given by

$$f(\mathbf{x}) = f(\boldsymbol{A}^{-1}\mathbf{y}) = \frac{1}{2\pi} \exp\left(-\frac{1}{2}\mathbf{y}^\top (\boldsymbol{A}^{-1})^\top \boldsymbol{A}^{-1}\mathbf{y}\right)$$

- $\frac{\partial}{\partial \mathbf{y}}\boldsymbol{A}^{-1}\mathbf{y} = \boldsymbol{A}^{-1}$. So, $\det\left(\frac{\partial}{\partial \mathbf{y}}\boldsymbol{A}^{-1}\mathbf{y}\right) = \det(\boldsymbol{A}^{-1}) = \frac{1}{ad - bc}$.
- Thus, $f(\mathbf{y}) = \frac{1}{2\pi} \exp\left(-\frac{1}{2}\mathbf{y}^\top (\boldsymbol{A}^{-1})^\top \boldsymbol{A}^{-1}\mathbf{y}\right) \cdot \left|\frac{1}{ad - bc}\right|$.

Outline

Standard Multivariate Gaussian

- Let $Z = (Z_1, \ldots, Z_D)^\top$ with independent coordinates
$$Z_i \sim \mathcal{N}(0,1), \quad i = 1, 2, \ldots, D.$$
- The 1D standard Gaussian pdf is
$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}.$$

Standard Multivariate Gaussian

- Let $Z = (Z_1, \ldots, Z_D)^\top$ with independent coordinates
$$Z_i \sim \mathcal{N}(0,1), \quad i = 1, 2, \ldots, D.$$
- The 1D standard Gaussian pdf is
$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}.$$
- By independence, the joint density of Z is
$$p_Z(z_1, \ldots, z_D) = \prod_{i=1}^{D} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z_i^2}{2}\right) = (2\pi)^{-D/2} \exp\left(-\frac{1}{2}\sum_{i=1}^{D} z_i^2\right).$$

Standard Multivariate Gaussian

- Let $Z = (Z_1, \ldots, Z_D)^\top$ with independent coordinates

$$Z_i \sim \mathcal{N}(0,1), \quad i = 1, 2, \ldots, D.$$

- The 1D standard Gaussian pdf is

$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}.$$

- By independence, the joint density of Z is

$$p_Z(z_1, \ldots, z_D) = \prod_{i=1}^{D} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z_i^2}{2}\right) = (2\pi)^{-D/2} \exp\left(-\frac{1}{2}\sum_{i=1}^{D} z_i^2\right).$$

  Writing $\sum_{i=1}^{D} z_i^2 = \|\mathbf{z}\|^2 = \mathbf{z}^\top \mathbf{z}$, we get

$$p_Z(\mathbf{z}) = (2\pi)^{-D/2} \exp\left(-\frac{1}{2}\mathbf{z}^\top \mathbf{z}\right), \quad \mathbf{z} \in \mathbb{R}^D.$$

Introducing Mean and Covariance

- Let $\mathbf{\Sigma}$ be a symmetric positive definite $D \times D$ matrix. Then there exists an invertible L such that

$$\mathbf{\Sigma} = LL^\top \quad \text{(e.g. Cholesky factorization).}$$

- Define $X = \boldsymbol{\mu} + LZ$. Then

$$\mathbb{E}[X] = \boldsymbol{\mu} + L\,\mathbb{E}[Z] = \boldsymbol{\mu}, \quad \text{and}$$

$$\begin{aligned} \mathrm{Cov}(X) &= \mathbb{E}[(X-\boldsymbol{\mu})(X-\boldsymbol{\mu})^\top] = \mathbb{E}[LZZ^\top L^\top] \\ &= L\,\mathbb{E}[ZZ^\top]\,L^\top = LI_D L^\top = \mathbf{\Sigma}. \end{aligned}$$

- Hence X has mean $\boldsymbol{\mu}$ and covariance $\mathbf{\Sigma}$; we write $X \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{\Sigma})$.

Change of Variables (1/2)

- The map from Z to X is affine:

$$T(\mathbf{z}) = \boldsymbol{\mu} + L\mathbf{z}, \qquad X = T(Z).$$

Its inverse is

$$z = T^{-1}(\mathbf{x}) = L^{-1}(\mathbf{x} - \boldsymbol{\mu}).$$

- The Jacobian of T^{-1} $\left(\text{i.e., } \frac{\partial}{\partial \mathbf{x}} T^{-1}(\mathbf{x})\right)$ is $J = L^{-1}$, so

$$|\det(J)| = \left|\det(L^{-1})\right| = (\det(L))^{-1}.$$

Change of Variables (1/2)

- The map from Z to X is affine:

$$T(\mathbf{z}) = \boldsymbol{\mu} + L\mathbf{z}, \qquad X = T(Z).$$

Its inverse is

$$z = T^{-1}(\mathbf{x}) = L^{-1}(\mathbf{x} - \boldsymbol{\mu}).$$

- The Jacobian of T^{-1} $\left(\text{i.e., } \frac{\partial}{\partial \mathbf{x}} T^{-1}(\mathbf{x})\right)$ is $J = L^{-1}$, so

$$|\det(J)| = \left|\det(L^{-1})\right| = (\det(L))^{-1}.$$

- By the change-of-variables formula,

$$p_X(\mathbf{x}) = p_Z\big(T^{-1}(\mathbf{x})\big) \left|\det(J)\right|.$$

Change of Variables (2/2)

- From

$$p_X(\mathbf{x}) = p_Z\big(T^{-1}(\mathbf{x})\big)\,\big|\det(J)\big|,$$

Plugging in p_Z and $\mathbf{z} = L^{-1}(\mathbf{x} - \boldsymbol{\mu})$, we obtain

$$\begin{aligned} p_X(\mathbf{x}) &= (2\pi)^{-D/2} \exp\left(-\frac{1}{2}\mathbf{z}^\top \mathbf{z}\right)(\det(L))^{-1} \\ &= (2\pi)^{-D/2} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top (L^{-1})^\top L^{-1}(\mathbf{x} - \boldsymbol{\mu})\right)(\det(L))^{-1}. \end{aligned}$$

Final Form of the Multivariate Gaussian

- Recall that $(L^{-1})^\top L^{-1} = (LL^\top)^{-1} = \mathbf{\Sigma}^{-1}$, and

$$\det(\mathbf{\Sigma}) = \det(LL^\top) = (\det(L))^2 \implies (\det(L))^{-1} = (\det(\mathbf{\Sigma}))^{-1/2}.$$

- Substituting into the previous expression gives

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

- Thus the pdf of the multivariate Gaussian $X \sim \mathcal{N}(\mu, \mathbf{\Sigma})$ is

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

Discussions