STAT/MTHE 353: 5 – Moment Generating Functions and Multivariate Normal Distribution

T. Linder

Queen's University

Winter 2017

Moment Generating Function

Definition Let $\boldsymbol{X} = (X_1, \ldots, X_n)^T$ be a random vector and $\boldsymbol{t} = (t_1, \ldots, t_n)^T \in \mathbb{R}^n$. The *moment generating function* (MGF) is defined by

$$M_{\boldsymbol{X}}(\boldsymbol{t}) = E\big(e^{\boldsymbol{t}^T \boldsymbol{X}}\big)$$

for all $\boldsymbol{t}$ for which the expectation exists (i.e., finite).

Remarks:

- $M_{\boldsymbol{X}}(\boldsymbol{t}) = E\big(e^{\sum_{i=1}^n t_i X_i}\big)$
- For $\boldsymbol{0} = (0, \ldots, 0)^T$, we have $M_{\boldsymbol{X}}(\boldsymbol{0}) = 1$.
- If $\boldsymbol{X}$ is a discrete random variable with finitely many values, then $M_{\boldsymbol{X}}(\boldsymbol{t}) = E\big(e^{\boldsymbol{t}^T \boldsymbol{X}}\big)$ is always finite for all $\boldsymbol{t} \in \mathbb{R}^n$.
- We will *always* assume that the distribution of $\boldsymbol{X}$ is such that $M_{\boldsymbol{X}}(\boldsymbol{t})$ is finite for all $\boldsymbol{t} \in (-t_0, t_0)^n$ for some $t_0 > 0$.

The single most important property of the MGF is that is uniquely determines the distribution of a random vector:

Theorem 1

Assume $M_{\boldsymbol{X}}(\boldsymbol{t})$ and $M_{\boldsymbol{Y}}(\boldsymbol{t})$ are the MGFs of the random vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$ and such that $M_{\boldsymbol{X}}(\boldsymbol{t}) = M_{\boldsymbol{Y}}(\boldsymbol{t})$ for all $\boldsymbol{t} \in (-t_0, t_0)^n$. Then

$$F_{\boldsymbol{X}}(\boldsymbol{z}) = F_{\boldsymbol{Y}}(\boldsymbol{z}) \text{ for all } \boldsymbol{z} \in \mathbb{R}^n$$

where $F_{\boldsymbol{X}}$ and $F_{\boldsymbol{Y}}$ are the joint cdfs of $\boldsymbol{X}$ and $\boldsymbol{Y}$.

Remarks:

- $F_{\boldsymbol{X}}(\boldsymbol{z}) = F_{\boldsymbol{Y}}(\boldsymbol{z})$ for all $\boldsymbol{z} \in \mathbb{R}^n$ clearly implies $M_{\boldsymbol{X}}(\boldsymbol{t}) = M_{\boldsymbol{Y}}(\boldsymbol{t})$. Thus $M_{\boldsymbol{X}}(\boldsymbol{t}) = M_{\boldsymbol{Y}}(\boldsymbol{t}) \iff F_{\boldsymbol{X}}(\boldsymbol{z}) = F_{\boldsymbol{Y}}(\boldsymbol{z})$
- Most often we will use the theorem for random variables instead of random vectors. In this case, $M_X(t) = M_Y(t)$ for all $t \in (-t_0, t_0)$ implies $F_X(z) = F_Y(z)$ for all $z \in \mathbb{R}$.

Connection with moments

- Let $k_1, \ldots, k_n$ be nonnegative integers and $k = k_1 + \cdots + k_n$. Then

$$\begin{aligned}
\frac{\partial^k}{\partial t_1^{k_1} \cdots \partial t_n^{k_n}} M_{\boldsymbol{X}}(\boldsymbol{t}) &= \frac{\partial^k}{\partial t_1^{k_1} \cdots \partial t_n^{k_n}} E\big(e^{t_1 X_1 + \cdots + t_n X_n}\big) \\
&= E\left(\frac{\partial^k}{\partial t_1^{k_1} \cdots \partial t_n^{k_n}} e^{t_1 X_1 + \cdots + t_n X_n}\right) \\
&= E\big(X_1^{k_1} \cdots X_n^{k_n} \big(e^{t_1 X_1 + \cdots + t_n X_n}\big)\big)
\end{aligned}$$

Setting $\boldsymbol{t} = \boldsymbol{0} = (0, \ldots, 0)^T$, we get

$$\boxed{\frac{\partial^k}{\partial t_1^{k_1} \cdots \partial t_n^{k_n}} M_{\boldsymbol{X}}(\boldsymbol{t})\big|_{\boldsymbol{t}=\boldsymbol{0}} = E\big(X_1^{k_1} \cdots X_n^{k_n}\big)}$$

- For a (scalar) random variable X we obtain the k*th moment* of X:

$$\boxed{\frac{d^k}{dt^k} M_X(t)\big|_{t=0} = E\big(X^k\big)}$$

Theorem 2

Assume $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_m$ are independent random vectors in $\mathbb{R}^n$ and let $\boldsymbol{X} = \boldsymbol{X}_1 + \cdots + \boldsymbol{X}_m$. Then

$$M_{\boldsymbol{X}}(\boldsymbol{t}) = \prod_{i=1}^{m} M_{\boldsymbol{X}_i}(\boldsymbol{t})$$

Proof:

$$\begin{aligned} M_{\boldsymbol{X}}(\boldsymbol{t}) &= E\big(e^{\boldsymbol{t}^T \boldsymbol{X}}\big) = E\big(e^{\boldsymbol{t}^T(\boldsymbol{X}_1+\cdots+\boldsymbol{X}_m)}\big) \\ &= E\big(e^{\boldsymbol{t}^T \boldsymbol{X}_1} \cdots e^{\boldsymbol{t}^T \boldsymbol{X}_m}\big) \\ &= E\big(e^{\boldsymbol{t}^T \boldsymbol{X}_1}\big) \cdots E\big(e^{\boldsymbol{t}^T \boldsymbol{X}_m}\big) \\ &= M_{\boldsymbol{X}_1}(\boldsymbol{t}) \cdots M_{\boldsymbol{X}_m}(\boldsymbol{t}) \qquad \square \end{aligned}$$

Note: This theorem gives us a powerful tool for determining the distribution of the sum of independent random variables.

Example: MGF for $X \sim \text{Gamma}(r, \lambda)$ and $X_1 + \cdots + X_m$ where the X_i are independent and $X_i \sim \text{Gamma}(r_i, \lambda)$.

Example: MGF for $X \sim \text{Poisson}(\lambda)$ and $X_1 + \cdots + X_m$ where the X_i are independent and $X_i \sim \text{Gamma}(\lambda_i)$. Also, use the MGF to find $E(X)$, $E(X^2)$, and $\text{Var}(X)$.

Theorem 3

Assume $\boldsymbol{X}$ is a random vector in $\mathbb{R}^n$, $\boldsymbol{A}$ is an $m \times n$ real matrix and $\boldsymbol{b} \in \mathbb{R}^m$. Then the MGF of $\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{X} + \boldsymbol{b}$ is given at $\boldsymbol{t} \in \mathbb{R}^m$ by

$$M_{\boldsymbol{Y}}(\boldsymbol{t}) = e^{\boldsymbol{t}^T \boldsymbol{b}} M_{\boldsymbol{X}}(\boldsymbol{A}^T \boldsymbol{t})$$

Proof:

$$\begin{aligned} M_{\boldsymbol{Y}}(\boldsymbol{t}) &= E\big(e^{\boldsymbol{t}^T \boldsymbol{Y}}\big) = E\big(e^{\boldsymbol{t}^T(\boldsymbol{A}\boldsymbol{X}+\boldsymbol{b})}\big) \\ &= e^{\boldsymbol{t}^T \boldsymbol{b}} E\big(e^{\boldsymbol{t}^T \boldsymbol{A}\boldsymbol{X}}\big) = e^{\boldsymbol{t}^T \boldsymbol{b}} E\big(e^{(\boldsymbol{A}^T \boldsymbol{t})^T \boldsymbol{X}}\big) \\ &= e^{\boldsymbol{t}^T \boldsymbol{b}} M_{\boldsymbol{X}}(\boldsymbol{A}^T \boldsymbol{t}) \qquad \square \end{aligned}$$

Note: In the scalar case $Y = aX + b$ we obtain

$$\boxed{M_Y(t) = e^{tb} M_X(at)}$$

Applications to Normal Distribution

Let $X \sim N(0,1)$. Then

$$\begin{aligned} M_X(t) &= E(e^{tX}) = \int_{-\infty}^{\infty} e^{tx} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx \\ &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x^2 - 2tx)}\, dx = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\left[(x-t)^2 - t^2\right]}\, dx \\ &= e^{t^2/2} \underbrace{\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-t)^2}}_{N(t,1) \text{ pdf}}\, dx \\ &= e^{t^2/2} \end{aligned}$$

We obtain that for all $t \in \mathbb{R}$

$$\boxed{M_X(t) = e^{t^2/2}}$$

Moments of a standard normal random variable

Recall the power series expansion $e^z = \sum_{k=0}^{\infty} \frac{z^k}{k!}$ valid for all $z \in \mathbb{R}$. Apply this to $z = tX$

$$\begin{aligned} M_X(t) &= E(e^{tX}) = E\left(\sum_{k=0}^{\infty} \frac{(tX)^k}{k!}\right) \\ &= \sum_{k=0}^{\infty} \frac{E\big[(tX)^k\big]}{k!} = \sum_{k=0}^{\infty} \frac{E(X^k)}{k!} t^k \end{aligned}$$

and to $z = t^2/2$

$$M_X(t) = e^{t^2/2} = \sum_{i=0}^{\infty} \frac{\left(t^2/2\right)^i}{i!}$$

Matching the coefficient of t^k, for $k = 1, 2, \ldots$ we obtain

$$\boxed{E(X^k) = \begin{cases} \dfrac{k!}{2^{k/2}(k/2)!}, & k \text{ even,} \\ 0, & k \text{ odd.} \end{cases}}$$

Sum of independent normal random variables

- Recall that if $X \sim N(0,1)$ and $Y = \sigma X + \mu$, where $\sigma > 0$, then $Y \sim N(\mu, \sigma^2)$. Thus the MGF of a $N(\mu, \sigma^2)$ random variable is

$$\begin{aligned} M_Y(t) &= e^{t\mu} M_X(\sigma t) = e^{t\mu} e^{(\sigma t)^2/2} \\ &= \boxed{e^{t\mu + t^2\sigma^2/2}} \end{aligned}$$

- Let $X_1, \ldots, X_m$ be independent r.v.'s with $X_i \sim N(\mu_i, \sigma_i^2)$ and set $X = X_1 + \cdots + X_m$. Then

$$M_X(t) = \prod_{i=1}^{m} M_{X_i}(t) = \prod_{i=1}^{m} e^{t\mu_i + t^2\sigma_i^2/2} = e^{t\left(\sum_{i=1}^m \mu_i\right) + \left(\sum_{i=1}^m \sigma_i^2\right)t^2/2}$$

This implies

$$\boxed{X \sim N\left(\sum_{i=1}^{m} \mu_i, \sum_{i=1}^{m} \sigma_i^2\right)}$$

i.e., $X_1 + \cdots + X_m$ is normal with mean $\mu_X = \sum_{i=1}^m \mu_i$ and variance $\sigma_X^2 = \sum_{i=1}^m \sigma_i^2$.

Multivariate Normal Distributions

Linear Algebra Review

- Recall that an $n \times n$ real matrix $\boldsymbol{C}$ is called *nonnegative definite* if it is symmetric and

$$\boldsymbol{x}^T \boldsymbol{C} \boldsymbol{x} \geq 0 \text{ for all } \boldsymbol{x} \in \mathbb{R}^n$$

  and *positive definite* if it is symmetric and

$$\boldsymbol{x}^T \boldsymbol{C} \boldsymbol{x} > 0 \text{ for all } \boldsymbol{x} \in \mathbb{R}^n \text{ such that } \boldsymbol{x} \neq \boldsymbol{0}$$

- Let $\boldsymbol{A}$ be an arbitrary $n \times n$ real matrix. Then $\boldsymbol{C} = \boldsymbol{A}^T\boldsymbol{A}$ is nonnegative definite. If $\boldsymbol{A}$ is nonsingular (invertible), then $\boldsymbol{C}$ is positive definite.

  Proof: $\boldsymbol{A}^T\boldsymbol{A}$ is symmetric since $(\boldsymbol{A}^T\boldsymbol{A})^T = (\boldsymbol{A}^T)(\boldsymbol{A}^T)^T = \boldsymbol{A}^T\boldsymbol{A}$. Thus it is nonnegative definite since

$$\boldsymbol{x}^T(\boldsymbol{A}^T\boldsymbol{A})\boldsymbol{x} = \boldsymbol{x}^T\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{x} = (\boldsymbol{A}\boldsymbol{x})^T(\boldsymbol{A}\boldsymbol{x}) = \|\boldsymbol{A}\boldsymbol{x}\|^2 \geq 0 \qquad \square$$

For any nonnegative definite $n \times n$ matrix $\boldsymbol{C}$ the following hold:

(1) $\boldsymbol{C}$ has n *nonnegative* eigenvalues $\lambda_1, \ldots, \lambda_n$ (counting multiplicities) and corresponding n *orthogonal* unit-length eigenvectors $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_n$:

$$\boldsymbol{C}\boldsymbol{b}_i = \lambda_i \boldsymbol{b}_i, \quad i = 1, \ldots, n$$

where $\boldsymbol{b}_i^T\boldsymbol{b}_i = 1$, $i = 1, \ldots, n$ and $\boldsymbol{b}_i^T\boldsymbol{b}_j = 0$ if $i \neq j$.

(2) (Spectral decomposition) $\boldsymbol{C}$ can be written as

$$\boldsymbol{C} = \boldsymbol{B}\boldsymbol{D}\boldsymbol{B}^T$$

where $\boldsymbol{D} = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ is the diagonal matrix of the eigenvalues of $\boldsymbol{C}$, and $\boldsymbol{B}$ is the orthogonal matrix whose ith column is $\boldsymbol{b}_i$, i.e., $\boldsymbol{B} = [\boldsymbol{b}_1 \ldots \boldsymbol{b}_n]$.

(3) $\boldsymbol{C}$ is *positive definite* $\iff$ $\boldsymbol{C}$ is *nonsingular* $\iff$ all the eigenvalues λ_i are *positive*

(4) $\boldsymbol{C}$ has a unique nonnegative definite square root $\boldsymbol{C}^{1/2}$, i.e., there exists a unique nonnegative definite $\boldsymbol{A}$ such that

$$\boldsymbol{C} = \boldsymbol{A}\boldsymbol{A}$$

Proof: We only prove the existence of $\boldsymbol{A}$. Let $\boldsymbol{D}^{1/2} = \mathrm{diag}(\lambda_1^{1/2}, \ldots, \lambda_n^{1/2})$ and note that $\boldsymbol{D}^{1/2}\boldsymbol{D}^{1/2} = \boldsymbol{D}$. Let $\boldsymbol{A} = \boldsymbol{B}\boldsymbol{D}^{1/2}\boldsymbol{B}^T$. Then $\boldsymbol{A}$ is nonnegative definite and

$$\begin{aligned} \boldsymbol{A}^2 &= \boldsymbol{A}\boldsymbol{A} = (\boldsymbol{B}\boldsymbol{D}^{1/2}\boldsymbol{B}^T)(\boldsymbol{B}\boldsymbol{D}^{1/2}\boldsymbol{B}^T) \\ &= \boldsymbol{B}\boldsymbol{D}^{1/2}\boldsymbol{B}^T\boldsymbol{B}\boldsymbol{D}^{1/2}\boldsymbol{B}^T = \boldsymbol{B}\boldsymbol{D}^{1/2}\boldsymbol{D}^{1/2}\boldsymbol{B}^T \\ &= \boldsymbol{C} \qquad \Box \end{aligned}$$

Remarks:

- If $\boldsymbol{C}$ is positive definite, then so is $\boldsymbol{A}$.
- If we don't require that $\boldsymbol{A}$ be nonnegative definite, then in general there are infinitely many solutions $\boldsymbol{A}$ for $\boldsymbol{A}\boldsymbol{A}^T = \boldsymbol{C}$.

Lemma 4

If $\boldsymbol{\Sigma}$ is the covariance matrix of some random vector $\boldsymbol{X} = (X_1, \ldots, X_n)^T$, then it is nonnegative definite.

Proof: We know that $\boldsymbol{\Sigma} = \mathrm{Cov}(\boldsymbol{X})$ is symmetric. Let $\boldsymbol{b} \in \mathbb{R}^n$ be arbitrary. Then

$$\boldsymbol{b}^T\boldsymbol{\Sigma}\boldsymbol{b} = \boldsymbol{b}^T\,\mathrm{Cov}(\boldsymbol{X})\boldsymbol{b} = \mathrm{Cov}(\boldsymbol{b}^T\boldsymbol{X}) = \mathrm{Var}(\boldsymbol{b}^T\boldsymbol{X}) \geq 0$$

so $\boldsymbol{\Sigma}$ is nonnegative definite

Remark: It can be shown that an $n \times n$ matrix $\boldsymbol{\Sigma}$ is nonnegative definite if and only if there exists a random vector $\boldsymbol{X} = (X_1, \ldots, X_n)^T$ such that $\mathrm{Cov}(\boldsymbol{X}) = \boldsymbol{\Sigma}$.

Defining the Multivariate Normal Distribution

Let $Z_1, \ldots, Z_n$ be independent r.v.'s with $Z_i \sim N(0,1)$. The multivariate MGF of $\boldsymbol{Z} = (Z_1, \ldots, Z_n)^T$ is

$$\begin{aligned} M_{\boldsymbol{Z}}(\boldsymbol{t}) &= E\big(e^{\boldsymbol{t}^T\boldsymbol{Z}}\big) = E\big(e^{\sum_{i=1}^n t_i Z_i}\big) = \prod_{i=1}^n E\big(e^{t_i Z_i}\big) \\ &= \prod_{i=1}^n e^{t_i^2/2} = e^{\sum_{i=1}^n t_i^2/2} = \boxed{e^{\frac{1}{2}\boldsymbol{t}^T\boldsymbol{t}}} \end{aligned}$$

Now let $\boldsymbol{\mu} \in \mathbb{R}^n$ and $\boldsymbol{A}$ an $n \times n$ real matrix. Then the MGF of $\boldsymbol{X} = \boldsymbol{A}\boldsymbol{Z} + \boldsymbol{\mu}$ is

$$\begin{aligned} M_{\boldsymbol{X}}(\boldsymbol{t}) &= e^{\boldsymbol{t}^T\boldsymbol{\mu}} M_{\boldsymbol{Z}}(\boldsymbol{A}^T\boldsymbol{t}) = e^{\boldsymbol{t}^T\boldsymbol{\mu}} e^{\frac{1}{2}(\boldsymbol{A}^T\boldsymbol{t})^T(\boldsymbol{A}^T\boldsymbol{t})} \\ &= e^{\boldsymbol{t}^T\boldsymbol{\mu}} e^{\frac{1}{2}\boldsymbol{t}^T\boldsymbol{A}\boldsymbol{A}^T\boldsymbol{t}} = \boxed{e^{\boldsymbol{t}^T\boldsymbol{\mu} + \frac{1}{2}\boldsymbol{t}^T\boldsymbol{\Sigma}\boldsymbol{t}}} \end{aligned}$$

where $\boldsymbol{\Sigma} = \boldsymbol{A}\boldsymbol{A}^T$. Note that $\boldsymbol{\Sigma}$ is nonnegative definite.

Definition Let $\boldsymbol{\mu} \in \mathbb{R}^n$ and let $\boldsymbol{\Sigma}$ be an $n \times n$ *nonnegative definite* matrix. A random vector $\boldsymbol{X} = (X_1, \ldots, X_n)$ is said to have a *multivariate normal* distribution with parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ if its multivariate MGF is

$$\boxed{M_{\boldsymbol{X}}(\boldsymbol{t}) = e^{\boldsymbol{t}^T\boldsymbol{\mu} + \frac{1}{2}\boldsymbol{t}^T\boldsymbol{\Sigma}\boldsymbol{t}}}$$

Notation: $\boldsymbol{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

Remarks:

- If $\boldsymbol{Z} = (Z_1, \ldots, Z_n)^T$ with $Z_i \sim N(0,1)$, $i = 1, \ldots, n$, then $\boldsymbol{Z} \sim N(\boldsymbol{0}, \boldsymbol{I})$, where $\boldsymbol{I}$ is the $n \times n$ identity matrix.
- We saw that if $\boldsymbol{Z} \sim N(\boldsymbol{0}, \boldsymbol{I})$, then $\boldsymbol{X} = \boldsymbol{A}\boldsymbol{Z} + \boldsymbol{\mu} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma} = \boldsymbol{A}\boldsymbol{A}^T$. One can show the following:

$\boldsymbol{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ *if and only if* $\boldsymbol{X} = \boldsymbol{A}\boldsymbol{Z} + \boldsymbol{\mu}$ for a random n-vector $\boldsymbol{Z} \sim N(\boldsymbol{0}, \boldsymbol{I})$ and some $n \times n$ matrix $\boldsymbol{A}$ with $\boldsymbol{\Sigma} = \boldsymbol{A}\boldsymbol{A}^T$.

Mean and covariance for multivariate normal distribution

Consider first $\boldsymbol{Z} \sim N(\boldsymbol{0}, \boldsymbol{I})$, i.e., $\boldsymbol{Z} = (Z_1, \ldots, Z_n)^T$, where the Z_i are independent $N(0,1)$ random variables. Then

$$E(\boldsymbol{Z}) = \big(E(Z_1), \ldots, E(Z_n)\big)^T = (0, \cdots, 0)^T$$

and

$$E\big((Z_i - E(Z_i))(Z_j - E(Z_j))\big) = E(Z_i Z_j) = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{if } i \neq j \end{cases}$$

Thus

$$\boxed{E(\boldsymbol{Z}) = \boldsymbol{0}, \qquad \operatorname{Cov}(\boldsymbol{Z}) = \boldsymbol{I}}$$

If $\boldsymbol{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then $\boldsymbol{X} = \boldsymbol{A}\boldsymbol{Z} + \boldsymbol{\mu}$ for a random n-vector $\boldsymbol{Z} \sim N(\boldsymbol{0}, \boldsymbol{I})$ and some $n \times n$ matrix $\boldsymbol{A}$ with $\boldsymbol{\Sigma} = \boldsymbol{A}\boldsymbol{A}^T$.

We have

$$E(\boldsymbol{A}\boldsymbol{Z} + \boldsymbol{\mu}) = \boldsymbol{A}E(\boldsymbol{Z}) + \boldsymbol{\mu} = \boldsymbol{\mu}$$

Also,

$$\operatorname{Cov}(\boldsymbol{A}\boldsymbol{Z} + \boldsymbol{\mu}) = \operatorname{Cov}(\boldsymbol{A}\boldsymbol{Z}) = \boldsymbol{A}\operatorname{Cov}(\boldsymbol{Z})\boldsymbol{A}^T = \boldsymbol{A}\boldsymbol{A}^T = \boldsymbol{\Sigma}$$

Thus

$$\boxed{E(\boldsymbol{X}) = \boldsymbol{\mu}, \qquad \operatorname{Cov}(\boldsymbol{X}) = \boldsymbol{\Sigma}}$$

Joint pdf for multivariate normal distribution

Lemma 5

If a random vector $\boldsymbol{X} = (X_1, \ldots, X_n)^T$ has covariance matrix $\boldsymbol{\Sigma}$ that is not of full rank (i.e., singular), then $\boldsymbol{X}$ does not have a joint pdf.

Proof sketch: If $\boldsymbol{\Sigma}$ is singular, then there exists $\boldsymbol{b} \in \mathbb{R}^n$ such that $\boldsymbol{b} \neq \boldsymbol{0}$ and $\boldsymbol{\Sigma}\boldsymbol{b} = \boldsymbol{0}$. Consider the random variable $\boldsymbol{b}^T\boldsymbol{X} = \sum_{i=1}^n b_i X_i$:

$$\operatorname{Var}(\boldsymbol{b}^T\boldsymbol{X}) = \operatorname{Cov}(\boldsymbol{b}^T\boldsymbol{X}) = \boldsymbol{b}^T\operatorname{Cov}(\boldsymbol{X})\boldsymbol{b} = \boldsymbol{b}^T\boldsymbol{\Sigma}\boldsymbol{b} = 0$$

Therefore $P(\boldsymbol{b}^T\boldsymbol{X} = c) = 1$ for some constant c. If $\boldsymbol{X}$ had a joint pdf $f(\boldsymbol{x})$, then for $B = \{\boldsymbol{x} : \boldsymbol{b}^T\boldsymbol{x} = c\}$ we should have

$$1 = P(\boldsymbol{b}^T\boldsymbol{X} = c) = P(\boldsymbol{X} \in B) = \int \cdots \int_B f(x_1, \ldots, x_n)\, dx_1 \cdots dx_n$$

But this is impossible since B is an $(n-1)$-*dimensional hyperplane* whose n-dimensional volume is *zero*, so the integral must be zero. □

Theorem 6

If $\boldsymbol{X} = (X_1, \ldots, X_n)^T \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is nonsingular, then it has a joint pdf given by

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = \frac{1}{\sqrt{(2\pi)^n \det \boldsymbol{\Sigma}}}\, e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}, \quad \boldsymbol{x} \in \mathbb{R}^n$$

Proof: We know that $\boldsymbol{X} = \boldsymbol{A}\boldsymbol{Z} + \boldsymbol{\mu}$ where $\boldsymbol{Z} = (Z_1, \ldots, Z_n)^T \sim N(\boldsymbol{0}, \boldsymbol{I})$ and $\boldsymbol{A}$ is an $n \times n$ matrix such that $\boldsymbol{A}\boldsymbol{A}^T = \boldsymbol{\Sigma}$. Since $\boldsymbol{\Sigma}$ is nonsingular, $\boldsymbol{A}$ must be nonsingular with inverse $\boldsymbol{A}^{-1}$. Thus the mapping

$$h(\boldsymbol{z}) = \boldsymbol{A}\boldsymbol{z} + \boldsymbol{\mu}$$

is invertible with inverse $g(\boldsymbol{x}) = \boldsymbol{A}^{-1}(\boldsymbol{x} - \boldsymbol{\mu})$ whose Jacobian is

$$J_g(\boldsymbol{x}) = \det \boldsymbol{A}^{-1}$$

By the multivariate transformation theorem

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = f_{\boldsymbol{Z}}(g(\boldsymbol{x}))|J_g(\boldsymbol{x})| = f_{\boldsymbol{Z}}\big(\boldsymbol{A}^{-1}(\boldsymbol{x} - \boldsymbol{\mu})\big)|\det \boldsymbol{A}^{-1}|$$

Proof cont'd: Since $\boldsymbol{Z} = (Z_1, \ldots, Z_n)^T$, where the Z_i are independent $N(0,1)$ random variables, we have

$$f_{\boldsymbol{Z}}(\boldsymbol{z}) = \prod_{i=1}^{n} \left(\frac{1}{\sqrt{2\pi}} \right) e^{-z_i^2/2} = \frac{1}{\sqrt{(2\pi)^n}} \, e^{-\frac{1}{2}\sum_{i=1}^n z_i^2} = \boxed{\frac{1}{\sqrt{(2\pi)^n}} \, e^{-\frac{1}{2}\boldsymbol{z}^T\boldsymbol{z}}}$$

so we get

$$\begin{aligned} f_{\boldsymbol{X}}(\boldsymbol{x}) &= f_{\boldsymbol{Z}}\big(\boldsymbol{A}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\big) |\det \boldsymbol{A}^{-1}| \\ &= \frac{1}{\sqrt{(2\pi)^n}} \, e^{-\frac{1}{2}(\boldsymbol{A}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}))^T(\boldsymbol{A}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}))} |\det \boldsymbol{A}^{-1}| \\ &= \frac{1}{\sqrt{(2\pi)^n}} \, e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T(\boldsymbol{A}^{-1})^T\boldsymbol{A}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})} |\det \boldsymbol{A}^{-1}| \\ &= \frac{1}{\sqrt{(2\pi)^n \det \boldsymbol{\Sigma}}} \, e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})} \end{aligned}$$

since $|\det \boldsymbol{A}^{-1}| = \dfrac{1}{\sqrt{\det \boldsymbol{\Sigma}}}$ and $(\boldsymbol{A}^{-1})^T\boldsymbol{A}^{-1} = \boldsymbol{\Sigma}^{-1}$ (exercise!) $\square$

Special case: bivariate normal

For $n = 2$ we have

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_1 \end{bmatrix} \quad \text{and} \quad \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}$$

where $\mu_i = E(X_i)$, $\sigma_i^2 = \mathrm{Var}(X_i)$, $i = 1, 2$, and

$$\rho = \rho(X_1, X_2) = \frac{\mathrm{Cov}(X_1, X_2)}{\sigma_1\sigma_2}$$

Thus the bivariate normal distribution is determined by five scalar parameters μ_1, μ_2, σ_1^2, σ_2^2, and ρ.

$\boldsymbol{\Sigma}$ is positive definite $\Longleftrightarrow$ $\boldsymbol{\Sigma}$ is invertible $\Longleftrightarrow$ $\det \boldsymbol{\Sigma} > 0$:

$$\det \boldsymbol{\Sigma} = (1-\rho^2)\sigma_1^2\sigma_2^2 > 0 \quad \Longleftrightarrow \quad |\rho| < 1 \text{ and } \sigma_1^2\sigma_2^2 > 0$$

so a bivariate normal random variable (X_1, X_2) has a pdf if and only if the components X_1 and X_2 have positive variances and $|\rho| < 1$.

We have

$$\boldsymbol{\Sigma}^{-1} = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}^{-1} = \frac{1}{\det \boldsymbol{\Sigma}} \begin{bmatrix} \sigma_2^2 & -\rho\sigma_1\sigma_2 \\ -\rho\sigma_1\sigma_2 & \sigma_1^2 \end{bmatrix}$$

and

$$\begin{aligned} &(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) \\ &= \begin{bmatrix} x_1-\mu_1, & x_2-\mu_1 \end{bmatrix} \frac{1}{(1-\rho^2)\sigma_1^2\sigma_2^2} \begin{bmatrix} \sigma_2^2 & -\rho\sigma_1\sigma_2 \\ -\rho\sigma_1\sigma_2 & \sigma_1^2 \end{bmatrix} \begin{bmatrix} x_1-\mu_1 \\ x_2-\mu_1 \end{bmatrix} \\ &= \frac{1}{(1-\rho^2)\sigma_1^2\sigma_2^2} \begin{bmatrix} x_1-\mu_1, & x_2-\mu_1 \end{bmatrix} \begin{bmatrix} \sigma_2^2(x_1-\mu_1) - \rho\sigma_1\sigma_2(x_2-\mu_2) \\ \sigma_1^2(x_2-\mu_2) - \rho\sigma_1\sigma_2(x_1-\mu_1) \end{bmatrix} \\ &= \frac{1}{(1-\rho^2)\sigma_1^2\sigma_2^2} \left(\sigma_2^2(x_1-\mu_1)^2 - 2\rho\sigma_1\sigma_2(x_1-\mu_1)(x_2-\mu_2) + \sigma_1^2(x_2-\mu_2)^2 \right) \\ &= \frac{1}{(1-\rho^2)} \left(\frac{(x_1-\mu_1)^2}{\sigma_1^2} + \frac{(x_2-\mu_2)^2}{\sigma_2^2} - \frac{2\rho(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2} \right) \end{aligned}$$

Thus the joint pdf of $(X_1, X_2)^T \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is

$$\boxed{f(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \, e^{\frac{1}{2(1-\rho^2)}\left(\frac{(x_1-\mu_1)^2}{\sigma_1^2} + \frac{(x_2-\mu_2)^2}{\sigma_2^2} - \frac{2\rho(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2} \right)}}$$

Remark: If $\rho = 0$, then

$$\begin{aligned} f(x_1, x_2) &= \frac{1}{2\pi\sigma_1\sigma_2} \, e^{\frac{1}{2}\left(\frac{(x_1-\mu_1)^2}{\sigma_1^2} + \frac{(x_2-\mu_2)^2}{\sigma_2^2} \right)} \\ &= \frac{1}{\sigma_1\sqrt{2\pi}} e^{\frac{(x_1-\mu_1)^2}{2\sigma_1^2}} \cdot \frac{1}{\sigma_2\sqrt{2\pi}} e^{\frac{(x_2-\mu_2)^2}{2\sigma_2^2}} \\ &= f_{X_1}(x_1) f_{X_2}(x_2) \end{aligned}$$

Therefore X_1 and X_2 are independent. It is also easy to see that $f(x_1, x_2) = f_{X_1}(x_1) f_{X_2}(x_2)$ for all x_1 and x_2 implies $\rho = 0$. Thus we obtain

> Two jointly normal random variables X_1 and X_2 are independent if and only if they are *uncorrelated*.

In general, the following important facts can be proved using the multivariate MGF:

(i) If $\boldsymbol{X} = (X_1, \ldots, X_n)^T \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then $X_1, X_2, \ldots X_n$ are independent if and only if they are uncorrelated, i.e., $\mathrm{Cov}(X_i, X_j) = 0$ if $i \neq j$, i.e., $\boldsymbol{\Sigma}$ is a diagonal matrix.

(ii) Assume $\boldsymbol{X} = (X_1, \ldots, X_n)^T \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and let

$$\boldsymbol{X}_1 = (X_1, \ldots, X_k)^T, \qquad \boldsymbol{X}_2 = (X_{k+1}, \ldots, X_n)^T$$

Then $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are independent if and only if $\mathrm{Cov}(\boldsymbol{X}_1, \boldsymbol{X}_2) = \mathbf{0}_{k\times(n-k)}$, the $k \times (n-k)$ matrix of zeros, i.e., $\boldsymbol{\Sigma}$ can be partitioned as

$$\boldsymbol{\Sigma} = \left[\begin{array}{c:c} \boldsymbol{\Sigma}_{11} & \mathbf{0}_{k\times(n-k)} \\ \hdashline \mathbf{0}_{(n-k)\times k} & \boldsymbol{\Sigma}_{22} \end{array}\right]$$

where $\boldsymbol{\Sigma}_{11} = \mathrm{Cov}(\boldsymbol{X}_1)$ and $\boldsymbol{\Sigma}_{22} = \mathrm{Cov}(\boldsymbol{X}_2)$.

Marginals of multivariate normal distributions

Let $\boldsymbol{X} = (X_1, \ldots, X_n)^T \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. If $\boldsymbol{A}$ is an $m \times n$ matrix and $\boldsymbol{b} \in \mathbb{R}^m$, then

$$\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{X} + \boldsymbol{b}$$

is a random m-vector. Its MGF at $\boldsymbol{t} \in \mathbb{R}^m$ is

$$M_{\boldsymbol{Y}}(\boldsymbol{t}) = e^{\boldsymbol{t}^T\boldsymbol{b}} M_{\boldsymbol{X}}(\boldsymbol{A}^T\boldsymbol{t})$$

Since $M_{\boldsymbol{X}}(\boldsymbol{\tau}) = e^{\boldsymbol{\tau}^T\boldsymbol{\mu} + \frac{1}{2}\boldsymbol{\tau}^T\boldsymbol{\Sigma}\boldsymbol{\tau}}$ for all $\boldsymbol{\tau} \in \mathbb{R}^n$, we obtain

$$\begin{aligned} M_{\boldsymbol{Y}}(\boldsymbol{t}) &= e^{\boldsymbol{t}^T\boldsymbol{b}} e^{(\boldsymbol{A}^T\boldsymbol{t})^T\boldsymbol{\mu} + \frac{1}{2}(\boldsymbol{A}^T\boldsymbol{t})^T\boldsymbol{\Sigma}(\boldsymbol{A}^T\boldsymbol{t})} \\ &= e^{\boldsymbol{t}^T(\boldsymbol{b}+\boldsymbol{A}\boldsymbol{\mu}) + \frac{1}{2}\boldsymbol{t}^T\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^T\boldsymbol{t}} \end{aligned}$$

This means that $\boldsymbol{Y} \sim N(\boldsymbol{b} + \boldsymbol{A}\boldsymbol{\mu}, \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^T)$, i.e., $\boldsymbol{Y}$ is multivariate normal with mean $\boldsymbol{b} + \boldsymbol{A}\boldsymbol{\mu}$ and covariance $\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^T$.

Example: Let $a_1, \ldots, a_n \in \mathbb{R}$ and determine the distribution of $Y = a_1X_1 + \cdots + a_nX_n$.

For some $1 \le m < n$ let $\{i_1, \ldots, i_m\} \subset \{1, \ldots, n\}$ such that $i_1 < i_2 < \cdots < i_m$. Let $\boldsymbol{e}_j = (0, \ldots, 0, 1, 0, \ldots, 0)^t$ be the jth unit vector in $\mathbb{R}^n$ and define the $m \times n$ matrix $\boldsymbol{A}$ by

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{e}_{i_1}^T \\ \vdots \\ \boldsymbol{e}_{i_m}^T \end{bmatrix}$$

Then

$$\boldsymbol{A}\boldsymbol{X} = \begin{bmatrix} \boldsymbol{e}_{i_1}^T \\ \vdots \\ \boldsymbol{e}_{i_m}^T \end{bmatrix} \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix} = \begin{bmatrix} X_{i_1} \\ \vdots \\ X_{i_m} \end{bmatrix}$$

Thus $(X_{i_1}, \ldots, X_{i_m})^T \sim N(\boldsymbol{A}\boldsymbol{\mu}, \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^T)$.

Note the following:

$$\boldsymbol{A}\boldsymbol{\mu} = \begin{bmatrix} \mu_{i_1} \\ \vdots \\ \mu_{i_m} \end{bmatrix}$$

and the (j,k)th entry of $\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^T$ is

$$\begin{aligned} (\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^T)_{jk} &= \big(\boldsymbol{A} \times (i_k\text{th column of } \boldsymbol{\Sigma})\big)_j \\ &= (\boldsymbol{\Sigma})_{i_j i_k} = \mathrm{Cov}(X_{i_j}, X_{i_k}) \end{aligned}$$

Thus if $\boldsymbol{X} = (X_1, \ldots, X_n)^T \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then $(X_{i_1}, \ldots, X_{i_m})^T$ is multivariate normal whose mean and covariance are obtained by picking out the corresponding elements of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$.

Special case: For $m = 1$ we obtain that $X_i \sim N(\mu_i, \sigma_i^2)$, where $\mu_i = E(X_i)$ and $\sigma_i^2 = \mathrm{Var}(X_i)$, for all $i = 1, \ldots, n$.

Conditional distributions

Let $\boldsymbol{X} = (X_1, \ldots, X_n)^T \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and for $1 \le m < n$ define

$$\boldsymbol{X}_1 = (X_1, \ldots, X_m)^T, \qquad \boldsymbol{X}_2 = (X_{m+1}, \ldots, X_n)^T$$

We know that $\boldsymbol{X}_1 \sim N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})$ and $\boldsymbol{X}_2 \sim N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22})$ where $\boldsymbol{\mu}_i = E(\boldsymbol{X}_i)$, $\boldsymbol{\Sigma}_{ii} = \mathrm{Cov}(\boldsymbol{X}_i)$, $i = 1, 2$.

Then $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ can be partitioned as

$$\boldsymbol{\mu} = \left[\begin{array}{c} \boldsymbol{\mu}_1 \\ \hdashline \boldsymbol{\mu}_2 \end{array}\right], \qquad \boldsymbol{\Sigma} = \left[\begin{array}{c:c} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \hdashline \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{array}\right]$$

where $\boldsymbol{\Sigma}_{ij} = \mathrm{Cov}(\boldsymbol{X}_i, \boldsymbol{X}_j)$, $i, j = 1, 2$. Note that $\boldsymbol{\Sigma}_{11}$ is $m \times m$, $\boldsymbol{\Sigma}_{22}$ is $(n-m) \times (n-m)$, $\boldsymbol{\Sigma}_{12}$ is $m \times (n-m)$, and $\boldsymbol{\Sigma}_{21}$ is $(n-m) \times m$. Also, $\boldsymbol{\Sigma}_{21} = \boldsymbol{\Sigma}_{12}^T$.

We assume that $\boldsymbol{\Sigma}_{11}$ is *nonsingular* and we want to determine the conditional distribution of $\boldsymbol{X}_2$ given $\boldsymbol{X}_1 = \boldsymbol{x}_1$.

Recall that $\boldsymbol{X} = \boldsymbol{A}\boldsymbol{Z} + \boldsymbol{\mu}$ for some $\boldsymbol{Z} = (Z_1, \ldots, Z_n)^T$ where the Z_i are independent $N(0,1)$ random variables and $\boldsymbol{A}$ is such that $\boldsymbol{A}\boldsymbol{A}^T = \boldsymbol{\Sigma}$.

Let $\boldsymbol{Z}_1 = (Z_1, \ldots, Z_m)^T$ and $\boldsymbol{Z}_2 = (Z_{m+1}, \ldots, Z_n)^T$. We want to determine such $\boldsymbol{A}$ in a partitioned form with dimensions corresponding to the partitioning of $\boldsymbol{\Sigma}$:

$$\boldsymbol{A} = \left[\begin{array}{c:c} \boldsymbol{B} & \boldsymbol{0}_{m \times (n-m)} \\ \hdashline \boldsymbol{C} & \boldsymbol{D} \end{array}\right]$$

We can write $\boldsymbol{\Sigma} = \boldsymbol{A}\boldsymbol{A}^T$ as

$$\begin{aligned} \left[\begin{array}{c:c} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \hdashline \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{array}\right] &= \left[\begin{array}{c:c} \boldsymbol{B} & \boldsymbol{0}_{m \times (n-m)} \\ \hdashline \boldsymbol{C} & \boldsymbol{D} \end{array}\right] \left[\begin{array}{c:c} \boldsymbol{B}^T & \boldsymbol{C}^T \\ \hdashline \boldsymbol{0}_{(n-m) \times m} & \boldsymbol{D}^T \end{array}\right] \\ &= \left[\begin{array}{c:c} \boldsymbol{B}\boldsymbol{B}^T & \boldsymbol{B}\boldsymbol{C}^T \\ \hdashline \boldsymbol{C}\boldsymbol{B}^T & \boldsymbol{C}\boldsymbol{C}^T + \boldsymbol{D}\boldsymbol{D}^T \end{array}\right] \end{aligned}$$

We want to solve for $\boldsymbol{B}$, $\boldsymbol{C}$ and $\boldsymbol{D}$. First consider $\boldsymbol{B}\boldsymbol{B}^T = \boldsymbol{\Sigma}_{11}$. We choose $\boldsymbol{B}$ to be the unique positive definite square root of $\boldsymbol{\Sigma}_{11}$:

$$\boldsymbol{B} = \boldsymbol{\Sigma}_{11}^{1/2}$$

Recall that $\boldsymbol{B}$ is symmetric and it is invertible since $\boldsymbol{\Sigma}_{11}$ is. Then $\boldsymbol{\Sigma}_{21} = \boldsymbol{C}\boldsymbol{B}^T =$ implies

$$\boldsymbol{C} = \boldsymbol{\Sigma}_{21}(\boldsymbol{B}^T)^{-1} = \boldsymbol{\Sigma}_{21}\boldsymbol{B}^{-1}$$

Then $\boldsymbol{\Sigma}_{22} = \boldsymbol{C}\boldsymbol{C}^T + \boldsymbol{D}\boldsymbol{D}^T$ gives

$$\begin{aligned} \boldsymbol{D}\boldsymbol{D}^T &= \boldsymbol{\Sigma}_{22} - \boldsymbol{C}\boldsymbol{C}^T = \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{B}^{-1}\boldsymbol{B}^{-1}(\boldsymbol{\Sigma}_{21})^T \\ &= \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}(\boldsymbol{B}\boldsymbol{B})^{-1}\boldsymbol{\Sigma}_{12} = \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12} \end{aligned}$$

Now note that $\boldsymbol{X} = \boldsymbol{A}\boldsymbol{Z} + \boldsymbol{\mu}$ gives

$$\boldsymbol{X}_1 = \boldsymbol{B}\boldsymbol{Z}_1 + \boldsymbol{\mu}_1, \qquad \boldsymbol{X}_2 = \boldsymbol{C}\boldsymbol{Z}_1 + \boldsymbol{D}\boldsymbol{Z}_2 + \boldsymbol{\mu}_2$$

Since $\boldsymbol{B}$ is invertible, given $\boldsymbol{X}_1 = \boldsymbol{x}_1$, we have $\boldsymbol{Z}_1 = \boldsymbol{B}^{-1}(\boldsymbol{x}_1 - \boldsymbol{\mu}_1)$. So given $\boldsymbol{X}_1 = \boldsymbol{x}_1$, we have that the conditional distribution of $\boldsymbol{X}_2$ and the conditional distribution of

$$\boldsymbol{C}\boldsymbol{B}^{-1}(\boldsymbol{x}_1 - \boldsymbol{\mu}_1) + \boldsymbol{D}\boldsymbol{Z}_2 + \boldsymbol{\mu}_2$$

are the same.

But $\boldsymbol{Z}_2$ is independent of $\boldsymbol{X}_1$, so given $\boldsymbol{X}_1 = \boldsymbol{x}_1$, the conditional distribution of $\boldsymbol{C}\boldsymbol{B}^{-1}(\boldsymbol{x}_1 - \boldsymbol{\mu}_1) + \boldsymbol{D}\boldsymbol{Z}_2 + \boldsymbol{\mu}_2$ is the same as its unconditional distribution.

We conclude that the conditional distribution of $\boldsymbol{X}_2$ given $\boldsymbol{X}_1 = \boldsymbol{x}_1$ is *multivariate normal* with mean

$$\begin{aligned} E(\boldsymbol{X}_2 | \boldsymbol{X}_1 = \boldsymbol{x}_1) &= \boldsymbol{\mu}_2 + \boldsymbol{C}\boldsymbol{B}^{-1}(\boldsymbol{x}_1 - \boldsymbol{\mu}_1) \\ &= \boldsymbol{\mu}_2 + \boldsymbol{\Sigma}_{21}\boldsymbol{B}^{-1}\boldsymbol{B}^{-1}(\boldsymbol{x}_1 - \boldsymbol{\mu}_1) \\ &= \boxed{\boldsymbol{\mu}_2 + \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}(\boldsymbol{x}_1 - \boldsymbol{\mu}_1)} \end{aligned}$$

and covariance matrix $\boxed{\boldsymbol{\Sigma}_{22|1} = \boldsymbol{D}\boldsymbol{D}^T = \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}}$

Special case: bivariate normal

Suppose $\boldsymbol{X} = (X_1, X_2)^T \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_1 \end{bmatrix} \quad \text{and} \quad \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}$$

We have

$$\mu_2 + \Sigma_{21}\Sigma_{11}^{-1}(x_1 - \mu_1) = \mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x_1 - \mu_1)$$

and

$$\Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12} = \sigma_2^2 - \frac{\rho^2\sigma_1^2\sigma_2^2}{\sigma_1^2} = \sigma_2^2(1 - \rho^2)$$

Thus the conditional distribution of X_2 given $X_1 = x_1$ is normal with (conditional) mean

$$\boxed{E(X_2|X_1 = x_1) = \mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x_1 - \mu_1)}$$

and variance

$$\boxed{\mathrm{Var}(X_2|X_1 = x_1) = \sigma_2^2(1 - \rho^2)}$$

Equivalently, the conditional distribution of X_2 given $X_1 = x_1$ is

$$N\big(\mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x_1 - \mu_1), \sigma_2^2(1 - \rho^2)\big)$$

If $|\rho| < 1$, then the conditional pdf exists and is given by

$$\boxed{f_{X_2|X_1}(x_2|x_1) = \frac{1}{\sigma_2\sqrt{2\pi(1-\rho^2)}}e^{-\frac{\left(x_2 - \mu_2 - \rho\frac{\sigma_2}{\sigma_1}(x_1 - \mu_1)\right)^2}{2\sigma_2^2(1-\rho^2)}}}$$

Remark: Note that $E(X_2|X_1 = x_1) = \mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x_1 - \mu_1)$ is a linear (affine) function of x_1.

Example: Recall the MMSE estimate problem for $X \sim N(0, \sigma_X^2)$ from the observation $Y = X + Z$, where $Z \sim N(0, \sigma_Z^2)$ and X and Z are independent. Use the above the find $g^*(y) = E[X|Y = y]$ and compute the minimum mean square error $E\big[(X - g^*(Y))^2\big]$.