UNIVERSITY OF TORONTO | **Statistical Sciences**

DoSS Summer Bootcamp Probability Module 6

Ichiro Hashimoto

University of Toronto

July 17, 2025

Recap

Learnt in last module:

- Moments
  - ▷ Expectation, Raw moments, central moments
  - ▷ Moment-generating functions
- Change-of-variables using MGF
  - ▷ Gamma distribution
  - ▷ Chi square distribution
- Conditional expectation
  - ▷ Law of total expectation
  - ▷ Law of total variance

Outline

- Covariance
  - Covariance as an inner product
  - Correlation
  - Cauchy-Schwarz inequality
  - Uncorrelatedness and Independence
- Concentration
  - Markov's inequality
  - Chebyshev's inequality
  - Chernoff bounds

Covariance

Recall the property of expectation:

$$\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y).$$

Covariance

Recall the property of expectation:

$$\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y).$$

What about the variance?

$$\begin{aligned} Var(X+Y) &= \mathbb{E}(X + Y - \mathbb{E}(X) - \mathbb{E}(Y))^2 \\ &= \mathbb{E}(X - \mathbb{E}(X))^2 + \mathbb{E}(Y - \mathbb{E}(Y))^2 + 2\mathbb{E}((X - \mathbb{E}(X))(Y - \mathbb{E}(Y))) \\ &= Var(X) + Var(Y) + 2\,\underbrace{\mathbb{E}((X - \mathbb{E}(X))(Y - \mathbb{E}(Y)))}_{?} \end{aligned}$$

Covariance

Intuition:

A measure of how much X, Y change together.

Covariance

$Var(X+Y) = Var(X) + Var(Y) + 2\,Cov(X,Y)$

Intuition:

A measure of how much X, Y change together. $= Var(X) + 2\,Cov(X,Y) + Var(Y)$

Covariance

For two jointly distributed real-valued random variables X, Y with finite second moments, the covariance is defined as

$$Cov(X, Y) = \mathbb{E}((X - \mathbb{E}(X))(Y - \mathbb{E}(Y))).$$

Similar to $Var(X) = \mathbb{E}(X-\mathbb{E}X)^2 = \mathbb{E}X^2 - (\mathbb{E}X)^2$.

Simplification:

$$Cov(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y).$$

(Pf) $Cov(X,Y) = \mathbb{E}[(X-\mathbb{E}X)(Y-\mathbb{E}Y)] = \mathbb{E}XY - \mathbb{E}((\mathbb{E}X)\cdot Y) - \mathbb{E}(X\cdot\mathbb{E}Y) + \mathbb{E}((\mathbb{E}X)\cdot(\mathbb{E}Y))$

Constants.

$= \mathbb{E}(XY) - (\mathbb{E}X)\cdot(\mathbb{E}Y) - (\mathbb{E}X)\cdot(\mathbb{E}Y) + (\mathbb{E}X)\cdot(\mathbb{E}Y)$

$= \mathbb{E}(XY) - (\mathbb{E}X)\cdot(\mathbb{E}Y).$

Covariance

Properties:

- $Cov(X, X) = Var(X) \geq 0$;
- $Cov(X, a) = 0, \quad a$ is a constant; $\Rightarrow$ $Cov(X, a) = \mathbb{E}\Big[(X - \mathbb{E}X) \cdot \underbrace{(a - \mathbb{E}a)}_{\mathbb{E}a = a}\Big] = \mathbb{E}[0] = 0.$
- $Cov(X, Y) = Cov(Y, X)$;
- $Cov(X + a, Y + b) = Cov(X, Y)$; $\Rightarrow$ $Cov(X+a, Y+b)$
- $Cov(aX, bY) = abCov(X, Y)$.

$$= \mathbb{E}\Big[\{X + \cancel{a} - \mathbb{E}(X + \cancel{a})\} \cdot \{Y + \cancel{b} - \mathbb{E}(Y + \cancel{b})\}\Big]$$

$$= \mathbb{E}\Big[(X - \mathbb{E}X) \cdot (Y - \mathbb{E}Y)\Big] = Cov(X, Y).$$

Covariance

Properties:

- $Cov(X, X) = Var(X) \geq 0$;
- $Cov(X, a) = 0, \quad a$ is a constant;
- $Cov(X, Y) = Cov(Y, X)$;
- $Cov(X + a, Y + b) = Cov(X, Y)$;
- $Cov(aX, bY) = abCov(X, Y)$.

Corollary about variance:

$$Var(aX + b) = a^2 Var(X).$$

Covariance

Relate covariance to inner product:

Inner product (not rigorous)

Inner product is a operator from a vector space V to a field F (use $\mathbb{R}$ here as an example): $< \cdot, \cdot >: V \times V \to \mathbb{R}$ that satisfies:

- Symmetry: $< x, y > = < y, x >$;
- Linearity in the first argument: $< ax + by, z > = a < x, z > + b < y, z >$;
- Positive-definiteness: $< x, x > \geq 0$, and $< x, x > = 0 \Leftrightarrow x = 0$

Covariance satisfies the conditions above EXCEPT

$(\Omega, \mathcal{F}, \mathbb{P}) \to \underbrace{L^2(\Omega, \mathcal{F}, \mathbb{P})}_{\text{space of random variables with finite second moment.}}$ defined by $\mathbb{E} X^2 < \infty$

Covariance

Relate covariance to inner product:

Inner product (not rigorous)

Inner product is a operator from a vector space V to a field F (use $\mathbb{R}$ here as an example): $< \cdot, \cdot >$: $V \times V \to \mathbb{R}$ that satisfies:

- Symmetry: $< x, y > = < y, x >$;
- Linearity in the first argument: $< ax + by, z > = a < x, z > + b < y, z >$;
- Positive-definiteness: $< x, x > \geq 0$, and $< x, x > = 0 \Leftrightarrow x = 0$

Remark:
Covariance defines an inner product over the quotient vector space obtained by taking the subspace of random variables with finite second moment and identifying any two that differ by a constant.

Covariance

Properties inherited from the inner product space

Recall in Euclidean vector space:

- $< x, y >= x^\top y = \sum_{i=1}^n x_i y_i$;
- $||x||_2 = \sqrt{< x, x >}$;
- $< x, y >= ||x||_2 \cdot ||y||_2 \cos(\theta)$.

Respectively:

- $< X, Y >= Cov(X, Y)$;
- $||X|| = \sqrt{Var(X)}$;

UNIVERSITY OF TORONTO

Covariance

A substitute for $\cos(\theta)$**:**

Correlation

For two jointly distributed real-valued random variables X, Y with finite second moments, the correlation is defined as

$$Corr(X, Y) = \rho_{XY} = \frac{Cov(X, Y)}{\sqrt{Var(X) \cdot Var(Y)}}. = \frac{\langle X, Y \rangle}{\|X\| \cdot \|Y\|}$$

Covariance

A substitute for $\cos(\theta)$**:**

Correlation

For two jointly distributed real-valued random variables X, Y with finite second moments, the correlation is defined as

$$Corr(X,Y) = \rho_{XY} = \frac{Cov(X,Y)}{\sqrt{Var(X) \cdot Var(Y)}}.$$

Uncorrelatedness:

$$X, Y \text{ uncorrelated} \quad \Leftrightarrow \quad Corr(X,Y) = 0.$$

Covariance

$|\langle X, Y\rangle| \leq \|X\| \|Y\|$

Cauchy-Schwarz inequality

$$|Cov(X,Y)| \leq \sqrt{Var(X)Var(Y)}.$$

Proof: Let $\hat{X} = X - \mathbb{E}X$, $\hat{Y} = Y - \mathbb{E}Y$.

$$0 \leq \mathbb{E}(\hat{X} + \underset{\in \mathbb{R}}{t}\hat{Y})^2 = \underbrace{\mathbb{E}\hat{X}^2}_{Var(X)} + 2t\underbrace{\mathbb{E}(\hat{X}\cdot\hat{Y})}_{Cov(X,Y)} + t^2\underbrace{\mathbb{E}\hat{Y}^2}_{Var(Y)}$$

$$= Var(Y)t^2 + 2\,Cov(X,Y)\cdot t + Var(X)$$

Since this quadratic inequality holds for any $t \in \mathbb{R}$,

we must have $D/4 = Cov(X,Y)^2 - Var(X)Var(Y) \leq 0$.

$$\therefore |Cov(X,Y)| \leq \sqrt{Var(X)Var(Y)}$$

Covariance

Uncorrelatedness and Independence:

Observe the relationship:

$$Corr(X, Y) = 0 \quad \Leftrightarrow \quad Cov(X, Y) = 0 \quad \Leftrightarrow \quad \mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(\overset{Y}{\cancel{X}})$$

$$\Uparrow$$

X and Y are independent

$\therefore$ If X and Y are independent, then $Corr(X, Y) = 0$.

Covariance

Uncorrelatedness and Independence:

Observe the relationship:

$$Corr(X, Y) = 0 \quad \Leftrightarrow \quad Cov(X, Y) = 0 \quad \Leftrightarrow \quad \mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(X)$$

Conclusions:

- Independence $\Rightarrow$ Uncorrelatedness
- Uncorrelatedness $\not\Longrightarrow$ Independence

Remark:
Independence is a very strong assumption/property on the distribution.

Covariance

Special case: multivariate normal

Multivariate normal

A k-dimensional random vector $\mathbf{X} = (X_1, X_2, \cdots, X_k)^\top$ follows a multivariate normal distribution $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, if

covariance matrix

$$f_{\mathbf{X}}(x_1, \ldots, x_k) = \frac{\exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)}{\sqrt{(2\pi)^k|\boldsymbol{\Sigma}|}},$$

where $\boldsymbol{\mu} = \mathbb{E}[\mathbf{X}] = (\mathbb{E}[X_1], \mathbb{E}[X_2], \ldots, \mathbb{E}[X_k])^\top$, and $[\boldsymbol{\Sigma}]_{i,j} = \Sigma_{i,j} = Cov(X_i, X_j)$.

Observation:

The distribution is decided by the covariance structure.

Covariance

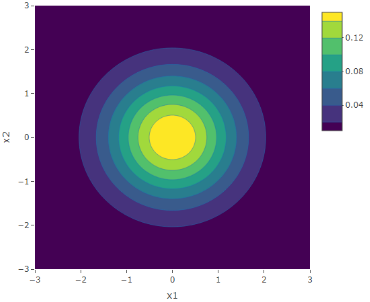
$$X_i, i = 1, \cdots k \text{ independent} \Leftrightarrow f_{\mathbf{X}}(x_1, \ldots, x_k) = \prod_{i=1}^{k} f_{X_i}(x_i)$$

$$\Leftrightarrow \mathbf{\Sigma} = \cancel{I_k} \Leftrightarrow Cov(X_i, X_j) = 0, i \neq j.$$

diagonal

Example:

- $Corr(X, Y) = 0$



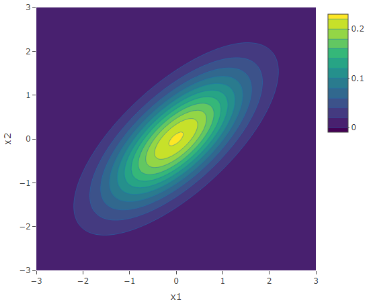
Covariance

$$X_i, i = 1, \cdots k \text{ independent} \Leftrightarrow f_{\mathbf{X}}(x_1, \ldots, x_k) = \prod_{i=1}^{k} f_{X_i}(x_i)$$

$$\Leftrightarrow \boldsymbol{\Sigma} = I_k \Leftrightarrow Cov(X_i, X_j) = 0, i \neq j.$$

Example:

- $Corr(X, Y) = 0.7$



UNIVERSITY OF TORONTO

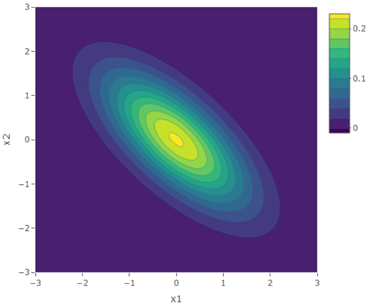
Covariance

$$X_i, i = 1, \cdots k \text{ independent} \Leftrightarrow f_{\mathbf{X}}(x_1, \ldots, x_k) = \prod_{i=1}^{k} f_{X_i}(x_i)$$

$$\Leftrightarrow \mathbf{\Sigma} = I_k \Leftrightarrow Cov(X_i, X_j) = 0, i \neq j.$$

Example:

- $Corr(X, Y) = -0.7$



Since Σ is PSD, in particular symmetric,

$\exists$ orthogonal matrix V and diagonal $\Lambda = \begin{pmatrix} \lambda_1^2 & & 0 \\ & \ddots & \\ 0 & & \lambda_d^2 \end{pmatrix}$

$(V^T V = V V^T = I)$.

$$\text{s.t.} \quad \underbrace{\Sigma = V \Lambda V^T \Longleftrightarrow V^T \Sigma V = \Lambda}_{\text{spectral decomposition}}$$

$$\Sigma^{-1} = V \Lambda^{-1} V^T$$

$$(x-\mu)^T \Sigma^{-1} (x-\mu) = \left[V^T (x-\mu) \right]^T \Lambda^{-1} \left[\underbrace{V^T (x-\mu)}_{\text{Let } z \text{ and use change of variable}} \right] = z^T \Lambda^{-1} z$$

$$= \sum_{i=1}^{n} \lambda_i^{-2} z_i^2$$

Also, since V is orthogonal, $|V| = 1$

Therefore, $|\Sigma| = |\Lambda| = \prod_{i=1}^{n} \lambda_i^2$.

Thus change of variable implies

$$f(z) = \frac{\exp\left(-\frac{1}{2} \sum_{i=1}^{n} \lambda_i^{-2} z_i^2\right)}{\sqrt{(2\pi)^n \prod_{i=1}^{n} \lambda_i^2}}$$

$$= \prod_{i=1}^{n} \underbrace{\frac{1}{\sqrt{2\pi}\, \lambda_i} \exp\left(-\frac{z_i^2}{2\lambda_i^2}\right)}_{\mathcal{N}(0,\, \lambda_i^2) \text{ in one-dimension.}}$$

This expression implies z_i's are independent.

In terms of z_i's $\operatorname{Cov}(z) = \Lambda$: diagonal

Concentration

Measures of a distribution:

- $\mathbb{E}(X^k)$, $\mathbb{E}(X)$, $Var(X)$;
- $Cov(X, Y)$ and $Corr(X, Y)$.

Concentration

Measures of a distribution:

- $\mathbb{E}(X^k)$, $\mathbb{E}(X)$, $Var(X)$;
- $Cov(X, Y)$ and $Corr(X, Y)$.

Tail probability: $\mathbf{P}(|X| > t)$

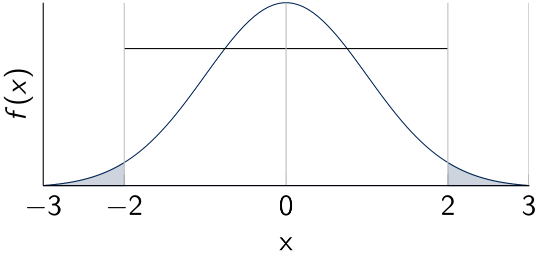


Figure: Probability density function of $\mathcal{N}(0, 1)$

Concentration

Concentration inequalities:

- Markov inequality
- Chebyshev inequality
- Chernoff bounds

Concentration

Concentration inequalities:

- Markov inequality
- Chebyshev inequality
- Chernoff bounds

Markov inequality

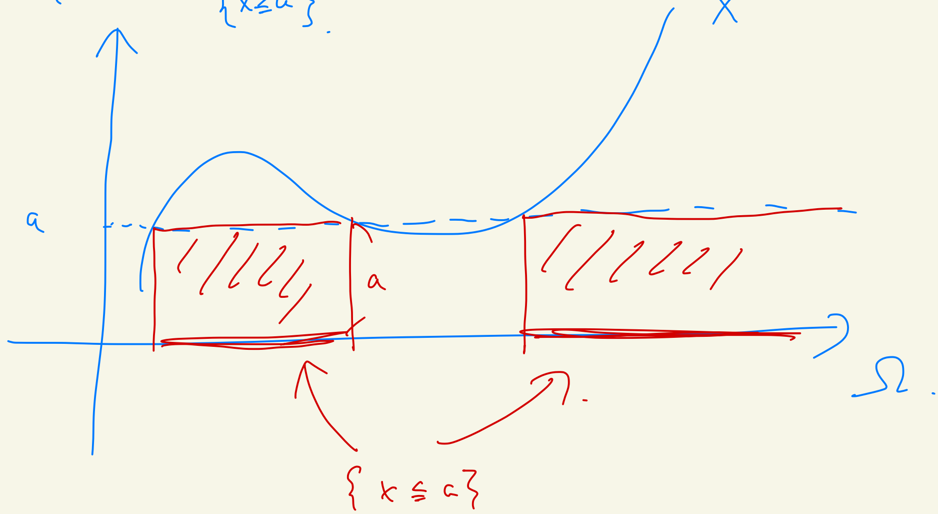
Let X be a random variable that is non-negative (almost surely). Then, for every constant $a > 0$,

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}(X)}{a}.$$

Proof: We use monotonicity of expectation, i.e.

if $X \geq Y$ then $\mathbb{E}X \geq \mathbb{E}Y$.

$Y = a \, \mathbb{1}_{\{X \geq a\}}$.



Since $Y \leq X$, we have

$$\mathbb{E}X \geq \mathbb{E}Y = \mathbb{E}\left[a \cdot \mathbb{1}_{(X \leq a)}\right] = a \cdot \mathbb{P}(X \leq a)$$

$$\therefore \quad \mathbb{P}(X \leq a) \leq \frac{\mathbb{E}X}{a}.$$

Concentration

Markov inequality (continued)

Let X be a random variable, then for every constant $a > 0$,

$$\mathbb{P}(|X| \geq a) \leq \frac{\mathbb{E}(|X|)}{a}.$$

A more general conclusion:

Markov inequality (continued)

Let X be a random variable, if $\Phi(x)$ is monotonically increasing on $[0, \infty)$, then for every constant $a > 0$,

apply Markov inequality

$$\mathbb{P}(|X| \geq a) = \mathbb{P}(\Phi(|X|) \geq \Phi(a)) \leq \frac{\mathbb{E}(\Phi(|X|))}{\Phi(a)}.$$

Concentration

$\phi(x) = x^2$

Chebyshev inequality

Let X be a random variable with finite expectation $\mathbb{E}(X)$ and variance $Var(X)$, then for every constant $a > 0$,

$$\mathbb{P}(|X - \mathbb{E}(X)| \geq a) \leq \frac{Var(X)}{a^2},$$

or equivalently,

$$\mathbb{P}(|X - \mathbb{E}(X)| \geq a\sqrt{Var(X)}) \leq \frac{1}{a^2}.$$

Example:

Take $a = 2$,

$$\mathbb{P}(|X - \mathbb{E}(X)| \geq 2\sqrt{Var(X)}) \leq \frac{1}{4}.$$

Concentration

$\phi(x) = \exp(x)$

Chernoff bound (general)

Let X be a random variable, then for $t \geq 0$,

$$\mathbb{P}(X \geq a) = \mathbb{P}(e^{t \cdot X} \geq e^{t \cdot a}) \leq \boxed{\frac{\mathbb{E}\left[e^{t \cdot X}\right]}{e^{t \cdot a}}},$$

→ this is only useful when $\mathbb{E}\, e^{tx} < \infty$ for some $t > 0$.

and

$$\mathbb{P}(X \geq a) \leq \inf_{t \geq 0} \frac{\mathbb{E}\left[e^{t \cdot X}\right]}{e^{t \cdot a}}.$$

find optimal t.

Remark:

This is especially useful when considering $X = \sum_{i=1}^{n} X_i$ with X_i's independent,

$$\mathbb{P}(X \geq a) \leq \inf_{t \geq 0} \frac{\mathbb{E}\left[\prod_i e^{t \cdot X_i}\right]}{e^{t \cdot a}} = \inf_{t \geq 0} e^{-t \cdot a} \prod_i \mathbb{E}\left[e^{t \cdot X_i}\right].$$

Suppose $X_i \overset{\text{i.i.d.}}{\sim} Z$.

$$\mathbb{P}\left(\sum_{i=1}^{n} X_i \geq a \right) \leq \inf_{t>0} \frac{\left(\mathbb{E}\left[e^{tZ} \right] \right)^n}{e^{t \cdot a}}$$

e.g. $X_i \overset{\text{i.i.d.}}{\sim} \text{Bern}(1/2)$

$$\mathbb{E}\left[e^{tX_i} \right] = \frac{e^t + e^{-t}}{2} \leq e^t.$$

can make a tighter bound.

$$\mathbb{P}\left(\sum_{i=1}^{n} X_i \geq a \right) \leq \inf_{t>0} \frac{e^{nt}}{e^{ta}}$$

$$= \inf_{t>0} \exp\left(t(n-a) \right)$$

Problem Set

Problem 1: Let

$$f_{X,Y}(x,y) = \begin{cases} 2 & 0 \leq y \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases},$$

compute $Cov(X, Y)$.

Problem 2: For $X \sim \mathcal{N}(0,1)$, compute the Chernoff bound.