Introduction to Probability, Statistics and Random Processes

Chapter 6: Multiple Random Variables

Hossein Pishro-Nik
University of Massachusetts, Amherst
Email: pishro@ecs.umass.edu

Joint Distributions and Independence

- The joint PMF of n discrete random variables X_1, X_2,..., X_n is

$$P_{X_1,X_2,...,X_n}(x_1,x_2,...,x_n) = P(X_1 = x_1, X_2 = x_2, ..., X_n = x_n).$$

- The joint PDF of n continuous random variables is a function $f_{X_1X_2...X_n}(x_1, x_2, ..., x_n)$ such that probability of $A \subset \mathbb{R}^n$ is given by

$$P\Big((X_1,X_2,\cdots,X_n) \in A\Big) = \int_A f_{X_1X_2\cdots X_n}(x_1, x_2, \cdots, x_n)dx_1 dx_2 \cdots dx_n.$$

Joint Distributions and Independence

- Marginal PDF of X_1 can be obtained from the joint PDF by

$$f_{X_1}(x_1) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{X_1 X_2 \ldots X_n}(x_1, x_2, \ldots, x_n) \, dx_2 \cdots dx_n.$$

- Joint CDF of n random variables $X_1, X_2, \ldots, X_n$ is given by

$$F_{X_1, X_2, \ldots, X_n}(x_1, x_2, \ldots, x_n) = P(X_1 \leq x_1, X_2 \leq x_2, \ldots, X_n \leq x_n).$$

Independence

- Random variables X_1, X_2, ... , X_n are independent, if for all $(x_1, x_2, ..., x_n) \in \mathbb{R}^n$,

$$F_{X_1,X_2,...,X_n}(x_1,x_2,...,x_n) = F_{X_1}(x_1)F_{X_2}(x_2)\cdots F_{X_n}(x_n).$$

- If X_1, X_2, ..., X_n are discrete, then they are independent if for all $(x_1, x_2, ..., x_n) \in \mathbb{R}^n$, we have

$$P_{X_1,X_2,...,X_n}(x_1,x_2,...,x_n) = P_{X_1}(x_1)P_{X_2}(x_2)\cdots P_{X_n}(x_n).$$

Independence

- If X_1, X_2, ..., X_n are continuous, then they are independent if for all $(x_1, x_2, ..., x_n) \in \mathbb{R}^n$, we have

$$f_{X_1,X_2,...,X_n}(x_1,x_2, ..., x_n) = f_{X_1}(x_1)f_{X_2}(x_2)\cdots f_{X_n}(x_n).$$

- If random variables X_1, X_2, ..., X_n are independent, then we have

$$E[X_1X_2\cdots X_n] = E[X_1]E[X_2]\cdots E[X_n].$$

Independence

Random variables X_1, X_2, ..., X_n are said to be **independent and identically distributed (i.i.d.)** if they are *independent*, and they have the *same marginal distributions*:

$$F_{X_1}(x) = F_{X_2}(x) = \ldots = F_{X_n}(x), \text{ for all } x \in \mathbb{R}.$$

Sums of Random Variables

- A random variable Y is given by

$$Y = X_1 + X_2 + ... + X_n$$

- The linearity of expectation gives

$$EY = EX_1 + EX_2 + \cdots + EX_n.$$

- We can find the variance of Y

$$\begin{aligned} Var(Y) &= Cov\left(\sum_{i=1}^{n} X_i, \sum_{j=1}^{n} X_j\right) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} Cov(X_i, X_j) \\ &= \sum_{i=1}^{n} Var(X_i) + 2\sum_{i<j} Cov(X_i, X_j). \end{aligned}$$

Sums of Random Variables

In general

$$Var\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} Var(X_i) + 2\sum_{i<j} Cov(X_i, X_j)$$

If X_1, X_2,...,X_n are independent,

$$Var\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} Var(X_i).$$

Sums of Random Variables

- For random variable $Y = X_1 + X_2 + ... + X_n$, we can find the PDF of Y by using the formula below

$$f_Y(y) = f_{X_1}(y) * f_{X_2}(y) * ... * f_{X_n}(y).$$

- This quickly becomes computationally difficult. Thus, we often resort to other methods like moment generating functions.

Moment Generating Functions

- **Definition:** The **nth moment** of a random variable X is defined to be $E[X^n]$. The **nth central moment** of X is defined to be $E[(X - EX)^n]$.
- The first moment is the expected value $E[X]$. The second central moment is the variance of X.

The moment generating function (MGF) of a random variable X is a function $M_X(s)$ defined as

$$M_X(s) = E\left[e^{sX}\right].$$

We say that MGF of X exists, if there exists a positive constant a such that $M_X(s)$ is finite for all $s \in [-a, a]$.

Moment Generating Functions

- The MGF is very useful because of the following reasons:
  - The MGF of X gives us all the moments of X.
  - The MGF, if it exists uniquely determines the distribution.
- With the MGF of a random variable, we have determined it's distribution. This method is very useful when we work on sums of several independent random variables.

Finding Moments from MGF

- Using the Taylor series for e^x, we can write

$$e^{sX} = \sum_{k=0}^{\infty} \frac{(sX)^k}{k!} = \sum_{k=0}^{\infty} \frac{X^k s^k}{k!}.$$

We can obtain all moments of X^k from its MGF:

$$M_X(s) = \sum_{k=0}^{\infty} E[X^k] \frac{s^k}{k!}$$

$$E[X^k] = \frac{d^k}{ds^k} M_X(s) \Big|_{s=0}$$

Moment Generating Functions

- **Theorem:** Consider two random variables X and Y. Suppose that there exists a positive constant c such that MGFs of X and Y are finite and identical for all values of s in $[-c, c]$. Then,

$$F_X(t) = F_Y(t), \text{ for all } t \in \mathbb{R}.$$

Sum on Independent Random Variables

If X_1, X_2, ..., X_n are n independent random variables, then

$$M_{X_1+X_2+\cdots+X_n}(s) = M_{X_1}(s)M_{X_2}(s)\cdots M_{X_n}(s).$$

Characteristic Functions

- The MGF does not exist for all random variables.
- If a random variable does have a well-defined MGF, we can use the characteristic function defined as

$$\phi_X(\omega) = E[e^{j\omega X}]$$

- $j = \sqrt{-1}$ and ω is a real number.
- The characteristic function is defined for all real-valued random variables.
- $|\phi_X(\omega)| \leq 1$.
- It has similar properties to the MGF. If X_1, X_2, ..., X_n are n independent random variables, then

$$\phi_{X_1+X_2+\cdots+X_n}(\omega) = \phi_{X_1}(\omega)\phi_{X_2}(\omega)\cdots\phi_{X_n}(\omega).$$

Random Vectors

- When we have n random variables X_1, X_2, ..., X_n we can put them in a (column) vector $\mathbf{X}$:

$$\mathbf{X} = \begin{bmatrix} X_1 \\ X_2 \\ . \\ . \\ X_n \end{bmatrix}.$$

- $\mathbf{X}$ is a n-dimensional random vector.
- The CDF of the random vector $\mathbf{X}$ is

$$\begin{aligned} F_{\mathbf{X}}(\mathbf{x}) &= F_{X_1,X_2,...,X_n}(x_1, x_2, ..., x_n) \\ &= P(X_1 \leq x_1, X_2 \leq x_2, ..., X_n \leq x_n). \end{aligned}$$

- If the X_i's are jointly continuous, the PDF of $\mathbf{X}$ can be written as

$$f_{\mathbf{X}}(\mathbf{x}) = f_{X_1,X_2,...,X_n}(x_1, x_2, ..., x_n).$$

Expectation

- **Expected value vector** or **mean vector** of random vector $\mathbf{X}$ is defined as

$$E\mathbf{X} = \begin{bmatrix} EX_1 \\ EX_2 \\ . \\ . \\ EX_n \end{bmatrix}.$$

- A **random matrix** is a matrix whose elements are random variables. We have an m by n random matrix $\mathbf{M}$ as

$$\mathbf{M} = \begin{bmatrix} X_{11} & X_{12} & \dots & X_{1n} \\ X_{21} & X_{22} & \dots & X_{2n} \\ . & . & . & . \\ . & . & . & . \\ X_{m1} & X_{m2} & \dots & X_{mn} \end{bmatrix}.$$

- Sometimes it is written as $\mathbf{M} = [X_{ij}]$.

Expectation

- The mean matrix of $\mathbf{M}$ is given as

$$E\mathbf{M} = \begin{bmatrix} EX_{11} & EX_{12} & \dots & EX_{1n} \\ EX_{21} & EX_{22} & \dots & EX_{2n} \\ . & . & . & . \\ . & . & . & . \\ EX_{m1} & EX_{m2} & \dots & EX_{mn} \end{bmatrix}.$$

- Linearity of expectation is valid for random vectors and matrices.
- For the random vector $\mathbf{Y} = \mathbf{A}\mathbf{X} + \mathbf{b}$ with $\mathbf{A}$, a fixed (non-random) m by n matrix and $\mathbf{b}$ a fixed m-dimensional vector, we have

$$E\mathbf{Y} = \mathbf{A}E\mathbf{X} + \mathbf{b}.$$

- If $\mathbf{X}_1$, $\mathbf{X}_2$, ..., $\mathbf{X}_k$ are n-dimensional random vectors, then we have

$$E[\mathbf{X}_1 + \mathbf{X}_2 + \cdots + \mathbf{X}_k] = E\mathbf{X}_1 + E\mathbf{X}_2 + \cdots + E\mathbf{X}_k.$$

Correlation and Covariance Matrix

- For a random vector $\mathbf{X}$, we define the **correlation** and **covariance** matrix as

Correlation matrix of $\mathbf{X}$:

$$\mathbf{R_X} = E[\mathbf{X}\mathbf{X}^T]$$

Covariance matrix of $\mathbf{X}$:

$$\mathbf{C_X} = E[(\mathbf{X} - E\mathbf{X})(\mathbf{X} - E\mathbf{X})^T] = \mathbf{R_X} - E\mathbf{X}E\mathbf{X}^T$$

Correlation and Covariance Matrix

$$\mathbf{R_X} = E[\mathbf{X}\mathbf{X}^T] = \begin{bmatrix} EX_1^2 & E[X_1X_2] & \ldots & E[X_1X_n] \\ EX_2X_1 & E[X_2^2] & \ldots & E[X_2X_n] \\ . & . & . & . \\ . & . & . & . \\ E[X_nX_1] & E[X_nX_2] & \ldots & E[X_n^2] \end{bmatrix},$$

$$\mathbf{C_X} = E[(\mathbf{X} - E\mathbf{X})(\mathbf{X} - E\mathbf{X})^T]$$

$$= \begin{bmatrix} Var(X_1) & Cov(X_1, X_2) & \ldots & Cov(X_1, X_n) \\ Cov(X_2, X_1) & Var(X_2) & \ldots & Cov(X_2, X_n) \\ . & . & . & . \\ . & . & . & . \\ Cov(X_n, X_1) & Cov(X_nX_2) & \ldots & Var(X_n) \end{bmatrix}.$$

Properties of the Covariance Matrix

- The covariance matrix $\mathbf{C_X}$ is a symmetric matrix.

$$c_{ij} = Cov(X_i, X_j) = Cov(X_j, X_i) = c_{ji}.$$

- $\mathbf{C_X}$ can be diagonalized and all the eigenvalues of $\mathbf{C_X}$ are real.
- A symmetrix matrix $\mathbf{M}$ is
  - **positive semi-definite**(PSD) if, for all vectors b, $\mathbf{b}^T\mathbf{M}\mathbf{b} \geq 0$.
  - **positive definite** (PD), if for all vectors $\mathbf{b} \neq 0$, $\mathbf{b}^T\mathbf{M}\mathbf{b} > 0$.
- The covariance matrix is positive semi-definite.

Properties of the Covariance Matrix

- **Theorem:** Let X be a random vector with n elements. Then, its covariance matrix $\mathbf{C_X}$ is positive semi-definite(PSD).
- **Theorem:** Let X be a random vector with n elements. Then its covariance matrix $\mathbf{C_X}$ is positive definite (PD), if and only if all its eigenvalues are larger than zero. Equivalently, $\mathbf{C_X}$ is positive definite (PD), if and only if $\det(\mathbf{C_X}) > 0$.
- For random vectors $\mathbf{X}$ and $\mathbf{Y}$, the **cross correlation matrix** is given by

$$\mathbf{R_{XY}} = E[\mathbf{X}\mathbf{Y}^T]$$

- The **cross covariance matrix** of $\mathbf{X}$ and $\mathbf{Y}$ is

$$\mathbf{C_{XY}} = E[(\mathbf{X} - E\mathbf{X})(\mathbf{Y} - E\mathbf{Y})^t]$$

Functions of Random Vectors: The Method of Transformations

- A function of a random vector is a random vector. We can use the method of transformations to find distributions of random vectors.
- If $\mathbf{X}$ is an n-dimensional random vector with joint PDF $f_{\mathbf{X}}(\mathbf{x})$. Let $G : \mathbb{R}^n \mapsto \mathbb{R}^n$ be a continuous and invertible function with continuous partial derivatives and let $H = G^{-1}$.
- Let the random vector $\mathbf{Y}$ be given by $\mathbf{Y} = G(\mathbf{X})$ and thus $\mathbf{X} = G^{-1}(\mathbf{Y}) = H(\mathbf{Y})$.

$$\mathbf{X} = \begin{bmatrix} X_1 \\ X_2 \\ . \\ . \\ X_n \end{bmatrix} = \begin{bmatrix} H_1(Y_1, Y_2, ..., Y_n) \\ H_2(Y_1, Y_2, ..., Y_n) \\ . \\ . \\ H_n(Y_1, Y_2, ..., Y_n) \end{bmatrix}.$$

Functions of Random Vectors: The Method of Transformations

- The PDF of $\mathbf{Y}$, $f_{Y_1,Y_2,\ldots,Y_n}(y_1, y_2, \ldots, y_n)$, is given by

$$\boxed{f_{\mathbf{Y}}(\mathbf{y}) = f_{\mathbf{X}}\big(H(\mathbf{y})\big)|J|}$$

where J is the Jacobian of H defined by

$$J = \det \begin{bmatrix} \frac{\partial H_1}{\partial y_1} & \frac{\partial H_1}{\partial y_2} & \cdots & \frac{\partial H_1}{\partial y_n} \\ \frac{\partial H_2}{\partial y_1} & \frac{\partial H_2}{\partial y_2} & \cdots & \frac{\partial H_2}{\partial y_n} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial H_n}{\partial y_1} & \frac{\partial H_n}{\partial y_2} & \cdots & \frac{\partial H_n}{\partial y_n} \end{bmatrix},$$

and evaluated at $(y_1, y_2, \ldots, y_n)$.

Normal (Gaussian) Random Vectors

Random variables X_1, X_2,..., X_n are said to be **jointly normal** if, for all a_1,a_2,..., $a_n \in \mathbb{R}$, the random variable

$$a_1X_1 + a_2X_2 + ... + a_nX_n$$

is a normal random variable.

A random vector

$$\mathbf{X} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix}$$

is said to be **normal** or **Gaussian** if the random variables X_1, X_2,..., X_n are jointly normal.

Normal (Gaussian) Random Vectors

- The **standard normal random vector**

$$\mathbf{Z} = \begin{bmatrix} Z_1 \\ Z_2 \\ . \\ . \\ Z_n \end{bmatrix},$$

where Z_i's are i.i.d. and $Z_i \sim N(0,1)$.

For a standard normal random vector $\mathbf{Z}$, where the Z_i's are i.i.d. and $Z_i \sim N(0,1)$, the PDF is given by

$$f_{\mathbf{Z}}(\mathbf{z}) = \frac{1}{(2\pi)^{\frac{n}{2}}} \exp\left\{-\frac{1}{2}\mathbf{z}^T\mathbf{z}\right\}.$$

Normal (Gaussian) Random Vectors

For a normal random vector $\mathbf{X}$ with mean $\mathbf{m}$ and covariance matrix $\mathbf{C}$, the PDF is given by

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{n}{2}}\sqrt{\det \mathbf{C}}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\mathbf{m})^T \mathbf{C}^{-1}(\mathbf{x}-\mathbf{m})\right\}$$

Normal (Gaussian) Random Vectors

If $\mathbf{X} = [X_1, X_2, ..., X_n]^T$ is a normal random vector, and we know $Cov(X_i, X_j) = 0$ for all $i \neq j$, then X_1,X_2, ..., X_n are independent.

If $\mathbf{X} = [X_1, X_2, ..., X_n]^T$ is a normal random vector, $\mathbf{X} \sim N(\mathbf{m}, \mathbf{C})$, $\mathbf{A}$ is an m by n fixed matrix, and $\mathbf{b}$ is an m-dimensional fixed vector, then the random vector $\mathbf{Y} = \mathbf{AX} + \mathbf{b}$ is a normal random vector with mean $\mathbf{A}E\mathbf{X} + \mathbf{b}$ and covariance matrix $\mathbf{ACA}^T$.

$$\mathbf{Y} \sim N(\mathbf{A}E\mathbf{X} + \mathbf{b}, \mathbf{ACA}^T)$$

Union Bound

- **Union bound** or **Boole's inequality** is applicable when you need to show that the probability of union of some events is less than some value.

The Union Bound

For any events $A_1, A_2, ..., A_n$, we have

$$P\left(\bigcup_{i=1}^{n} A_i\right) \leq \sum_{i=1}^{n} P(A_i)$$

- It is widely used in the area of *random graphs*.

Union Bound and its Extensions

- We can extend the union bounds and obtain the lower and upper bound on probability of union of events.
- These bounds are known as **Bonferroni inequalities**.
- Start writing the inclusion-exclusion formula. If you stop at the first term, you obtain an upper bound on the probability of union. If you stop at the second term, you obtain a lower bound. If you stop at the third term, you obtain an upper bound, etc.
- In general, if you write an odd number of terms, you get an upper bound and if you write an even number of terms, you get a lower bound.

Union Bound and its Extensions

Generalization of the Union Bound: Bonferroni inequalities For any events $A_1, A_2, ..., A_n$, we have

$$P\left(\bigcup_{i=1}^n A_i\right) \leq \sum_{i=1}^n P(A_i)$$

$$P\left(\bigcup_{i=1}^n A_i\right) \geq \sum_{i=1}^n P(A_i) - \sum_{i<j} P(A_i \cap A_j)$$

$$P\left(\bigcup_{i=1}^{n} A_i\right) \leq \sum_{i=1}^{n} P(A_i) - \sum_{i<j} P(A_i \cap A_j) + \sum_{i<j<k} P(A_i \cap A_j \cap A_k)$$

Expected Value of the Number of Events

- The union bound formula is also equal to the expected value of the number of occurred events.
- Let $A_1, A_2, ..., A_n$ be any events. Define the indicator random variables X_1, X_2,...,X_n as

$$X_i = \begin{cases} 1 & \text{if } A_i \text{ occurs} \\ 0 & \text{otherwise} \end{cases}$$

- If we define $X = X_1 + X_2 + X_3 + ... + X_n$, then X shows the number of events that actually occur. We have

$$\begin{aligned} EX &= EX_1 + EX_2 + EX_3 + ... + EX_n \text{by linearity of expectation} \\ &= P(A_1) + P(A_2) + ... + P(A_n), \end{aligned}$$

which is indeed the righthand-side of the union bound.

Markov and Chebyshev Inequalities

Markov's Inequality

If X is any nonnegative random variable, then

$$P(X \geq a) \leq \frac{EX}{a}, \text{ for any } a > 0.$$

Chebyshev's Inequality

If X is any random variable, then for any $b > 0$ we have

$$P\big(|X - EX| \geq b\big) \leq \frac{Var(X)}{b^2}.$$

Chernoff Bounds

Chernoff Bounds:

$$P(X \geq a) \leq e^{-sa} M_X(s), \ \text{for all } s > 0,$$
$$P(X \leq a) \leq e^{-sa} M_X(s), \ \text{for all } s < 0$$

Comparison between Markov, Chebyshev, and Chernoff Bounds

- We found upper bounds on $P(X \geq \alpha n)$ for $X \sim Binomial(n, p)$ with $p = \frac{1}{4}$ and $\alpha = \frac{3}{4}$:

$$P(X \geq \frac{3n}{4}) \leq \frac{2}{3} \qquad \text{Markov},$$

$$P(X \geq \frac{3n}{4}) \leq \frac{4}{n} \qquad \text{Chebyshev},$$

$$P(X \geq \frac{3n}{4}) \leq \left(\frac{16}{27}\right)^{\frac{n}{4}} \qquad \text{Chernoff}.$$

- The bound given by Markov is the "weakest".
- The bound given by Chebyshev's inequality is "stronger" than the one given by Markov's inequality. Note that $\frac{4}{n}$ goes to zero as n goes to infinity.
- The strongest bound is the Chernoff bound. It goes to zero exponentially fast.

Cauchy-Schwarz Inequality

Cauchy-Schwarz Inequality

For any two random variables X and Y, we have

$$|EXY| \leq \sqrt{E[X^2]E[Y^2]},$$

where equality holds if and only if $X = \alpha Y$, for some constant $\alpha \in \mathbb{R}$.

Jensen's Inequality

Definition: Consider a function $g : I \to \mathbb{R}$, where I is an interval in $\mathbb{R}$. We say that g is a **convex** function if for any two points x and y in I and any $\alpha \in [0,1]$, we have

$$g(\alpha x + (1-\alpha)y) \leq \alpha g(x) + (1-\alpha)g(y).$$

We say that g is **concave** if

$$g(\alpha x + (1-\alpha)y) \geq \alpha g(x) + (1-\alpha)g(y).$$

Jensen's Inequality

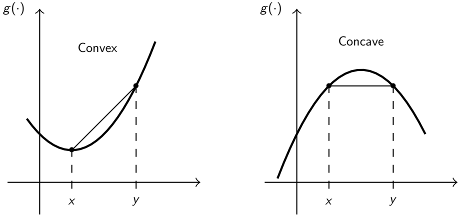

Figure: Pictorial representation of a convex function and a concave function.

Jensen's Inequality

Jensen's Inequality:

If $g(x)$ is a convex function on R_X, and $E[g(X)]$ and $g(E[X])$ are finite, then

$$E[g(X)] \geq g(E[X]).$$

A twice-differentiable function $g : I \to \mathbb{R}$ is convex if and only if $g''(x) \geq 0$ for all $x \in I$.