ECEN 649 Pattern Recognition

Probability Review

Ulisses Braga-Neto

ECE Department

Texas A&M University

Sample Spaces and Events

- A *sample space* S is the collection of all outcomes of an experiment.
- An *event* E is a subset $E \subseteq S$.
- Event E is said to *occur* if it contains the outcome of the experiment.
- Example 1: if the experiment consists of flipping two coins, then

$$S = \{(H, H), \, (H, T), \, (T, H), \, (T, T)\}$$

  The event E that the first coin lands tails is: $E = \{(T, H), \, (T, T)\}$.

- Example 2: If the experiment consists in measuring the lifetime of a lightbulb, then

$$S = \{t \in \mathbb{R} \mid t \geq 0\}$$

  The event that the lightbulb will fail at or earlier than 2 time units is the real interval $E = [0, 2]$.

Special Events

- Inclusion: $E \subseteq F$ iff the occurrence of E implies the occurrence of F.
- Union: $E \cup F$ occurs iff E, F, or both E and F occur.
- Intersection: $E \cap F$ occurs iff both E and F occur. If $E \cap F = \emptyset$, then E and F are mutually-exclusive.
- Complement: E^c occurs iff E does not occur.
- If $E_1, E_2, \ldots$ is an *increasing* sequence of events, that is, $E_1 \subseteq E_2 \subseteq \ldots$ then

$$\lim_{n\to\infty} E_n = \bigcup_{n=1}^{\infty} E_n$$

- If $E_1, E_2, \ldots$ is a *decreasing* sequence of events, that is, $E_1 \supseteq E_2 \supseteq \ldots$ then

$$\lim_{n\to\infty} E_n = \bigcap_{n=1}^{\infty} E_n$$

Special Events - II

- If $E_1, E_2, \ldots$ is *any* sequence of events, then

$$[E_n \; i.o.] = \limsup_{n \to \infty} E_n = \bigcap_{n=1}^{\infty} \bigcup_{i=n}^{\infty} E_i$$

We can see that $\limsup_{n \to \infty} E_n$ occurs iff E_n occurs for an infinite number of n, that is, E_n *occurs infinitely often*.

- If $E_1, E_2, \ldots$ is *any* sequence of events, then

$$\liminf_{n \to \infty} E_n = \bigcup_{n=1}^{\infty} \bigcap_{i=n}^{\infty} E_i$$

We can see that $\liminf_{n \to \infty} E_n$ occurs iff E_n occurs for all but a finite number of n, that is, E_n *eventually occurs for all* n.

Modern Definition of Probability

Probability assigns to each event $E \subseteq S$ a number $P(E)$ such that

1. $0 \leq P(E) \leq 1$
2. $P(S) = 1$
3. For a sequence of mutually-exclusive events $E_1, E_2, \ldots$

$$P\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} P(E_i)$$

Technically speaking, for a non-countably infinite sample space S, probability cannot be assigned to all possible subsets of S in a manner consistent with these axioms; one has to restrict the definition to a σ-*algebra* of events in S.

Direct Consequences

- $P(E^c) = 1 - P(E)$
- If $E \subseteq F$ then $P(E) \leq P(F)$
- $P(E \cup F) = P(E) + P(F) - P(E \cap F)$
- (*Boole's Inequality*) For *any* sequence of events $E_1, E_2, \ldots$

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

- (Continuity of Probability Measure) If $E_1, E_2, \ldots$ is an increasing or decreasing sequence of events, then

$$P\left(\lim_{n\to\infty} E_n\right) = \lim_{n\to\infty} P(E_n)$$

1st Borel-Cantelli Lemma

For *any* sequence of events $E_1, E_2, \ldots$

$$\sum_{n=1}^{\infty} P(E_n) < \infty \;\Rightarrow\; P([E_n \; i.o.]) = 0$$

Proof.

$$P\left(\bigcap_{n=1}^{\infty} \bigcup_{i=n}^{\infty} E_i\right) = P\left(\lim_{n\to\infty} \bigcup_{i=n}^{\infty} E_i\right)$$

$$= \lim_{n\to\infty} P\left(\bigcup_{i=n}^{\infty} E_i\right)$$

$$\leq \lim_{n\to\infty} \sum_{i=n}^{\infty} P(E_i)$$

$$= 0$$

Example

Consider a sequence of binary random variables $X_1, X_2, \ldots$ that take on values in $\{0, 1\}$, such that

$$P(\{X_n = 0\}) = \frac{1}{2^n}, \quad n = 1, 2, \ldots$$

Since

$$\sum_{n=1}^{\infty} P(\{X_n = 0\}) = 2 < \infty$$

The 1st Borel-Cantelli lemma implies that $P([\{X_n = 0\}\ i.o.]) = 0$, that is, for n sufficiently large, $X_n = 1$ with probability one. In other words,

$$\lim_{n \to \infty} X_n = 1 \quad \textit{with probability 1.}$$

2nd Borel-Cantelli Lemma

For an *independent* sequence of events $E_1, E_2, \ldots$,

$$\sum_{n=1}^{\infty} P(E_n) = \infty \;\Rightarrow\; P([E_n \; i.o.]) = 1$$

Proof. Exercise.

- This is the converse to the 1st lemma; here the stronger assumption of independence is needed.
- As immediate consequence of the 2nd Borel-Cantelli lemma is that *any* event with positive probability, no matter how small, will occur infinitely often in independent repeated trials.

Example: “The infinite typist monkey"

Consider a monkey that sits at a typewriter banging away randomly for an infinite amount of time. It will produce Shakespeare’s complete works, and in fact, the entire Library of Congress, not just once, but an infinite number of times.

Proof. Let L be the length in characters of the desired work. Let E_n be the event that the n-th sequence of characters produced by the monkey matches, character by character, the desired work (we are making it even harder for the monkey, as we are ruling out overlapping frames). Clearly $P(E_n) = 27^{-L} > 0$. It is a very small number, but still positive. Now, since our monkey never gets disappointed nor tired, the events E_n are independent. It follows by the 2nd Borel-Cantelli lemma that E_n will occur, and infinitely often. Q.E.D. (NOTE: In practice, if all the atoms in the universe were typist monkeys banging away billions of characters a second since the big-bang, the probability of getting Shakespeare within the age of the universe would still be vanishingly small.)

Tail Events and Kolmogorov’s 0-1 Law

- Given a sequence of events $E_1, E_2, \ldots$, a *tail event* is an event whose occurrence depends on the whole sequence, but is probabilistically independent of any finite subsequence.
- Examples of tail events: $\lim_{n\to\infty} E_n$ (if $\{E_n\}$ is monotone), $\limsup_{n\to\infty} E_n$, $\liminf_{n\to\infty} E_n$.
- Kolmogorov’s zero-one law: given a sequence of *independent* events $E_1, E_2, \ldots$, all its tail events have either probability 0 or probability 1. That is, tail events are either almost-surely impossible or occur almost surely.
- In practice, it may be difficult to conclude one way or the other. The Borel-Cantelli lemmas together give a sufficient condition to decide on the 0-1 probability of the tail event $\limsup_{n\to\infty} E_n$, with $\{E_n\}$ an independent sequence.

Conditional Probability

- One of the most important concepts in Pattern Recognition, Engineering, and in Probability Theory in general.
- Given that an event F has occurred, for E to occur, $E \cap F$ has to occur. In addition, the sample space gets *restricted* to those outcomes in F, so a normalization factor $P(F)$ has to be introduced. Therefore,

$$P(E|F) = \frac{P(E \cap F)}{P(F)}$$

- By the same token, one can condition on any number of events:

$$P(E|F_1, F_2, \ldots, F_n) = \frac{P(E \cap F_1 \cap F_2 \cap \ldots \cap F_n)}{P(F_1 \cap F_2 \cap \ldots \cap F_n)}$$

Note: For simplicity, it is usual to write $P(E \cap F) = P(E, F)$ to indicate the *joint probability* of E and F.

Very Useful Formulas

- $P(E, F) = P(E|F)P(F)$

- $P(E_1, E_2, \ldots, E_n) =$
  $P(E_n|E_1, \ldots, E_{n-1})P(E_{n-1}|E_1, \ldots, E_{n-2}) \cdots P(E_2|E_1)P(E_1)$

- $P(E) = P(E, F) + P(E, F^c) = P(E|F)P(F) + P(E|F^c)(1 - P(F))$

- $P(E) = \sum_{i=1}^{n} P(E, F_i) = \sum_{i=1}^{n} P(E|F_i)P(F_i)$, whenever $\bigcup F_i \supseteq E$

- (*Bayes Theorem*):

$$P(E|F) = \frac{P(F|E)P(E)}{P(F)} = \frac{P(F|E)P(E)}{P(F|E)P(E) + P(F|E^c)(1 - P(E)))}$$

Independent Events

- Events E and F are independent if the occurrence of one does not carry information as to the occurrence of the other. That is:

  $$P(E|F) = P(E) \quad \text{and} \quad P(F|E) = P(F).$$

  It is easy to see that this is equivalent to the condition

  $$P(E, F) = P(E)P(F)$$

- If E and F are independent, so are the pairs (E,F^c), (E^c,F), and (E^c,F^c).
- **Caution 1**: E being independent of F and G *does not* imply that E is independent of $F \cap G$.
- **Caution 2**: Three events E, F, G are independent if $P(E, F, G) = P(E)P(F)P(G)$ *and each pair* of events is independent.

Random Variables

- A *random variable* X is a (measurable) function $X : S \to \mathbb{R}$, that is, it assigns to each outcome of the experiment a real number.

- Given a set of real numbers A, we define an event

$$\{X \in A\} = X^{-1}(A) \subseteq S.$$

- It can be shown that all probability questions about a r.v. X can be phrased in terms of the probabilities of a simple set of events:

$$\{X \in (-\infty, a]\}, \quad a \in \mathbb{R}.$$

These events can be written more simply as $\{X \leq a\}$, for $a \in \mathbb{R}$.

Probability Distribution Functions

- The probability of the special events $\{X \leq a\}$ gets a special name.
- The *probability distribution function* (PDF) of a r.v. X is the function $F_X : \mathbb{R} \to [0, 1]$ given by

$$F_X(a) = P(\{X \leq a\}), \quad a \in \mathbb{R}.$$

- Properties of a PDF:
  1. F_X is non-decreasing: $a \leq b \Rightarrow F_X(a) \leq F_X(b)$.
  2. $\lim_{a \to -\infty} F_X(a) = 0$ and $\lim_{a \to +\infty} F_X(a) = 1$
  3. F_X is right-continuous: $\lim_{b \to a_+} F_X(b) = F_X(a)$.
- Everything we know about events applies here, since a PDF is nothing more than the probability of certain events.

Joint and Conditional PDFs

- These are crucial elements in PR and Engineering. Once again, these concepts involve only the probabilities of certain special events.

- The *joint PDF* of two r.v.'s X and Y is the joint probability of the events $\{X \leq a\}$ and $\{Y \leq b\}$, for $a, b \in \mathbb{R}$. Formally, we define a function $F_{XY} : \mathbb{R} \times \mathbb{R} \to [0, 1]$ given by

  $$F_{XY}(a, b) = P(\{X \leq a\}, \{Y \leq b\}) = P(\{X \leq a, Y \leq B\}), \quad a, b \in \mathbb{R}$$

  This is the probability of the "lower-left quadrant" with corner at (a, b).

- Note that $F_{XY}(a, \infty) = F_X(a)$ and $F_{XY}(\infty, b) = F_Y(b)$. These are called the *marginal PDFs*.

Joint and Conditional PDFs - II

- The *conditional PDF* of X given Y is just the conditional probability of the corresponding special events. Formally, it is a function $F_{X|Y} : \mathbb{R} \times \mathbb{R} \to [0, 1]$ given by

$$\begin{aligned} F_{X|Y}(a,b) &= P(\{X \leq a\}|\{Y \leq b\}) \\ &= \frac{P(\{X \leq a, Y \leq B\})}{P(\{Y \leq b\})} = \frac{F_{XY}(a,b)}{F_Y(b)}, \quad a, b \in \mathbb{R} \end{aligned}$$

- The r.v.'s X and Y are *indepedent r.v.s* if $F_{X|Y}(a,b) = F_X(a)$ and $F_{Y|X}(b,a) = F_Y(b)$, that is, if $F_{XY}(a,b) = F_X(a)F_Y(b)$, for $a, b \in \mathbb{R}$.

- Bayes' Theorem for r.v.'s:

$$F_{X|Y}(a,b) = \frac{F_{Y|X}(b,a)F_X(a)}{F_Y(b)}, \quad a, b \in \mathbb{R}.$$

Probability Density Functions

- The notion of a *probability density function* (pdf) is fundamental in probability theory. However, it is a secondary notion to that of a PDF. In fact, all r.v.'s must have a PDF, but not all r.v.'s have a pdf.
- If F_X is everywhere continuous and differentiable, then X is said to be a *continuous* r.v. and the pdf f_X of X is given by:

$$f_X(a) = \frac{dF_X}{dx}(a), \quad a \in \mathbb{R}.$$

- Probability statements about X can then be made in terms of integration of f_X. For example,

$$F_X(a) = \int_{-\infty}^{a} f_X(x)dx, \quad a \in \mathbb{R}$$

$$P(\{a \leq X \leq b\}) = \int_{a}^{b} f_X(x)dx, \quad a, b \in \mathbb{R}$$

Useful Continuous R.V.'s

- Uniform(a,b)

$$f_X(x) = \frac{1}{b-a}, \quad a < x < b.$$

- The univariate Gaussian(μ, $\sigma > 0$):

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \quad x \in \mathbb{R}.$$

- Exponential($\lambda > 0$):

$$f_X(x) = \lambda e^{-\lambda x}, \quad x \geq 0.$$

- Gamma($\lambda > 0$, $t > 0$):

$$f_X(x) = \frac{\lambda e^{-\lambda x}(\lambda x)^{t-1}}{\Gamma(t)}, \quad x \geq 0.$$

- Beta(a,b):

$$f_X(x) = \frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1}, \quad 0 < x < 1.$$

Probability Mass Function

- If F_X is not everywhere differentiable, then X does not have a pdf.
- A useful particular case is that where X assumes countably many values and F_X is a *staircase function*. In this case, X is said to be a *discrete* r.v.
- One defines the *probability mass function* (PMF) of a discrete r.v. X to be:

$$p_X(a) = P(\{X = a\}) = F_X(a) - F_X(a^-), \quad a \in \mathbb{R}.$$

(Note that for any continuous r.v. X, $P(\{X = a\}) = 0$ so there is no PMF to speak of.)

Useful Discrete R.V.'s

- Bernoulli(p):

$$p_X(0) = P(\{X = 0\}) = 1 - p$$

$$p_X(1) = P(\{X = 0\}) = p$$

- Binomial(n,p):

$$p_X(i) = P(\{X = i\}) = \begin{pmatrix} n \\ i \end{pmatrix} p^i (1-p)^{n-i}, \quad i = 0, 1, \ldots, n$$

- Poisson (λ):

$$p_X(i) = P(\{X = i\}) = e^{-\lambda} \frac{\lambda^i}{i!}, \quad i = 0, 1, \ldots$$

Expectation

- Expectation is a fundamental concept in probability theory, which has to do with the intuitive concept of "averages."

- The mean value of a r.v. X is an average of its values weighted by their probabilities. If X is a continuous r.v., this corresponds to:

$$E[X] = \int_{-\infty}^{\infty} x f_X(x)\, dx$$

If X is discrete, then this can be written as:

$$E[X] = \sum_{i:p_X(x_i)>0} x_i p_X(x_i)$$

Expectation of Functions of R.V.'s

- Given a r.v. X, and a (measurable) function $g : \mathbb{R} \to \mathbb{R}$, then $g(X)$ is also a r.v. If there is a pdf, it can be shown that

  $$E[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x)\, dx$$

  If X is discrete, then this becomes:

  $$E[g(X)] = \sum_{i : p_X(x_i) > 0} g(x_i) p_X(x_i)$$

- Immediate corollary: $E[aX + c] = aE[X] + c$.
- (Joint r.v.'s) If $g : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ is measurable, then (continuous case):

  $$E[g(X, Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) f_{XY}(x, y)\, dx dy$$

Other Properties of Expectation

- Linearity: $E[a_1 X_1 + \ldots a_n X_n] = a_1 E[X_1] + \ldots a_n E[X_n]$.
- if X and Y are *uncorrelated*, then $E[XY] = E[X]E[Y]$ (independence always implies uncorrelatedness, but the converse is true only in special cases; e.g. jointly Gaussian or multinomial random variables).
- If $X \geq Y$ then $E[X] > E[Y]$.
- If X is non-negative,

$$E[X] = \int_0^\infty P(\{X > x\})\, dx$$

- Markov's Inequality: If X is non-negative, then for $a > 0$,

$$P(\{X \geq a\}) \leq \frac{E[X]}{a}$$

- Cauchy-Schwarz Inequality: $E[XY] \leq \sqrt{E[X^2]E[Y^2]}$

Conditional Expectation

- If X and Y are jointly continuous r.v.'s and $f_Y(y) > 0$, we define:

$$E[X|Y=y] = \int_{-\infty}^{\infty} x f_{X|Y}(x,y)\,dx \;=\; \int_{-\infty}^{\infty} x\,\frac{f_{XY}(x,y)}{f_Y(y)}\,dx$$

If X, Y are jointly discrete and $p_Y(y_j) > 0$, this can be written as:

$$E[X|Y=y_j] = \sum_{i:p_X(x_i)>0} x_i p_{X|Y}(x_i,y_j) \;=\; \sum_{i:p_X(x_i)>0} x_i\,\frac{p_{XY}(x_i,y_j)}{p_Y(y_j)}$$

- Conditional expectations have all the properties of usual expectations. For example,

$$E\left[\sum_{i=1}^{n} X_i|Y=y\right] = \sum_{i=1}^{n} E[X_i|Y=y]$$

$E[Y|X]$ is a Random Variable

- Given a r.v. X, the mean $E[X]$ is a deterministic parameter.
- Now, $E[X|Y]$ is not random w.r.t. to X, but it is a function of the r.v. Y, so it is also a r.v. One can show that its mean is precisely $E[X]$:

  $$E[E[X|Y]] = E[X]$$

  What this says in the continuous case is:

  $$E[X] = \int_{-\infty}^{\infty} E[X|Y = y] f_Y(y)\, dy$$

  and, in the discrete case,

  $$E[X] = \sum_{i:p_Y(y_i)>0} E[X|Y = y_i] P(\{Y = y_i\})$$

- Computing $E[X|Y = y]$ first is often easier than finding $E[X]$ directly.

Conditional Expectation and Prediction

- One is interested in predicting the value of an unknown r.v. Y. We also want the predictor $\hat{Y}$ to be optimal according to some criterion.
- The criterion most widely used is the *Mean Square Error:*

$$\textit{MSE}\,(\hat{Y}) = E[(\hat{Y} - Y)^2]$$

- It can be shown that the MMSE (minimum MSE) constant (no-information) estimator of Y is just the mean $E[Y]$.
- Given now some partial information through an observed r.v. $X = x$, it can be shown that the overall MMSE estimator is the conditional mean $E[Y|X = x]$. The function $\eta(x) = E[Y|X = x]$ is called the *regression* of Y on X.
- Is this Pattern Recognition yet? Why?

Variance

- The mean $E[X]$ is a good “guess” of the value of a r.v. X, but by itself it can be misleading. The *variance* Var(X) says how the values of X are spread around the mean $E[X]$:

$$\text{Var}(X) = E[(X - E[X])^2] = E[X^2] - (E[X])^2$$

- The MSE of the best constant estimator $\hat{Y} = E[Y]$ of Y is just the variance of Y!

- Property: $\text{Var}(aX + c) = a^2\text{Var}(X)$

- Chebyshev’s Inequality: For any $\epsilon > 0$,

$$P(\{|X - E[X]| \geq \epsilon\}) \leq \frac{\text{Var}(X)}{\epsilon^2}$$

Conditional Variance

- If X and Y are jointly-distributed r.v.'s , we define:

$$\text{Var}(X|Y) = E[(X - E[X|Y])^2|Y] = E[X^2|Y] - (E[X|Y])^2$$

- Conditional Variance Formula:

$$\text{Var}(X) = E[\text{Var}(X|Y)] + \text{Var}(E[X|Y])$$

This breaks down the *total variance* in a "within-rows" component and a "across-rows" component.

Covariance

- Is $\text{Var}(X_1 + X_2) = \text{Var}(X_1) + \text{Var}(X_2)$?
- The covariance of two r.v.'s X and Y is given by:

$$\text{Cov}(X, Y) = E[(X - E[X])(Y - E[Y])] = E[XY] - E[X]E[Y]$$

- Two variables X and Y are uncorrelated if and only if $\text{Cov}(X, Y) = 0$. Jointly Gaussian X and Y are independent if and only if they are uncorrelated (in general, independence implies uncorrelatedness but not vice-versa).
- It can be shown that

$$\text{Var}(X_1 + X_2) = \text{Var}(X_1) + \text{Var}(X_2) + 2\text{Cov}(X_1, X_2)$$

- So the variance is distributive over sums if all variables are *pair-wise uncorrelated*.

Correlation Coefficient

- The correlation coefficient ρ between two r.v.'s X and Y is given by:

$$\rho(X,Y) = \frac{\text{Cov}(X,Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}$$

- Properties:

1. $-1 \leq \rho(X,Y) \leq 1$
2. X and Y are uncorrelated iff $\rho(X,Y) = 0$.
3. (Perfect linear correlation):

$$\rho(X,Y) = \pm 1 \ \Leftrightarrow \ Y = a \pm bX, \text{ where } b = \frac{\sigma_y}{\sigma_x}$$

Vector Random Variables

- A vector r.v. or *random vector* $\mathbf{X} = (X_1, \ldots, X_d)$ takes values in $\mathbb{R}^d$. Its distribution is the joint distribution of the component r.v.'s.
- The mean of $\mathbf{X}$ is the vector $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_d)$, where $\mu_i = E[X_i]$.
- The *covariance matrix* Σ is a $d \times d$ matrix given by:

$$\Sigma = E[(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^T]$$

where $\Sigma_{ii} = \text{Var}(X_i)$ and $\Sigma_{ij} = \text{Cov}(X_i, X_j)$.

Properties of the Covariance Matrix

- Matrix Σ is real symmetric and thus diagonalizable:

  $$\Sigma = UDU^T$$

  where U is the matrix of eigenvectors and D is the diagonal matrix of eigenvalues.

- All eigenvalues are nonnegative (Σ is *positive semi-definite*). In fact, except for "degenerate" cases, all eigenvalues are positive, and so Σ is invertible (Σ is said to be *positive definite* in this case).

- (*Whitening or Mahalanobis transformation*) It is easy to check that the random vector

  $$\mathbf{Y} = \Sigma^{-\frac{1}{2}}(\mathbf{X} - \boldsymbol{\mu}) = D^{-\frac{1}{2}}U^T(\mathbf{X} - \boldsymbol{\mu})$$

  has zero mean and covariance matrix $\mathbf{I}_d$ (so that all components of $\mathbf{Y}$ are zero-mean, unit-variance, and uncorrelated).

Sample Estimates

- Given n *independent and identically-distributed* (i.i.d.) sample observations $\mathbf{X}_1, \ldots, \mathbf{X}_n$ of the random vector $\mathbf{X}$, then the *sample mean* estimator is given by:

  $$\hat{\boldsymbol{\mu}} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

  It can be shown that this estimator is *unbiased* (that is, $E[\hat{\boldsymbol{\mu}}] = \boldsymbol{\mu}$) and *consistent* (that is, $\hat{\mu}$ converges in probability to $\boldsymbol{\mu}$ as $n \to \infty$).

- The *sample covariance* estimator is given by:

  $$\mathbf{S} = \frac{1}{n-1} \sum_{i=1}^{n} (\mathbf{X}_i - \hat{\boldsymbol{\mu}})(\mathbf{X}_i - \hat{\boldsymbol{\mu}})^T$$

  This is an unbiased and consistent estimator of Σ.

The Multivariate Gaussian

- The multivariate Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance matrix Σ (assuming Σ invertible, so that also det$(\Sigma) > 0$) corresponds to the multivariate pdf

  $$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^p \ \det(\Sigma)}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$$

  We write $X \sim \mathcal{N}_p(\boldsymbol{\mu}, \Sigma)$.

- The multivariate gaussian has elliptical contours of the form

  $$(\mathbf{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu}) = c^2, \quad c > 0$$

  The axes of the ellipsoids are given by the eigenvectors of Σ and the length of the axes are proportional to its eigenvalues.

Properties of The Multivariate Gaussian

- The density of each component X_i is univariate gaussian $\mathcal{N}(\mu_i, \Sigma_{ii})$.
- The components of $\mathbf{X} \sim \mathcal{N}(\mu, \Sigma)$ are independent *if and only if* they are uncorrelated, i.e., Σ is a diagonal matrix.
- The whitening transformation $\mathbf{Y} = \Sigma^{-\frac{1}{2}}(\mathbf{X} - \boldsymbol{\mu})$ produces another multivariate gaussian $\mathbf{Y} \sim \mathcal{N}(0, \mathbf{I}_p)$.
- In general, if $\mathbf{A}$ is a nonsingular $p \times p$ matrix and $\mathbf{c}$ is a p-vector, then $\mathbf{Y} = \mathbf{AX} + \mathbf{c} \sim \mathcal{N}_p(\mathbf{A}\boldsymbol{\mu} + \mathbf{c}, \mathbf{A}\Sigma\mathbf{A}^T)$.
- The r.v.'s $\mathbf{AX}$ and $\mathbf{BX}$ are independent if and only if $\mathbf{A}\Sigma\mathbf{B}^T = 0$.
- If $\mathbf{Y}$ and $\mathbf{X}$ are jointly Gaussian, then the distribution of $\mathbf{Y}$ given $\mathbf{X}$ is again Gaussian.
- The regression $E[\mathbf{Y}|\mathbf{X}]$ is a linear function of $\mathbf{X}$.

Convergence of Random Sequences

- A *random sequence* $\{X[1], X[2], \ldots\}$ is a countable set of r.v.s.

- *"Sure"* convergence: We say that $X[n] \rightarrow X$ surely if for all outcomes $\xi \in S$ of the experiment we have $\lim_{n\rightarrow\infty} X[n,\xi] = X(\xi)$.

- *Almost-sure* convergence or convergence *with probability one*: The sequence fails to converge only for an event of probability zero, i.e.:

$$P\left(\{\xi \in S | \lim_{n\rightarrow\infty} X[n,\xi] = X(\xi)\}\right) \; = \; 1$$

- *Mean-square* (m.s.) convergence: This uses convergence of an *energy norm* to zero.

$$\lim_{n\rightarrow\infty} E[|X[n] - X|^2] \; = \; 0$$

Convergence of Random Sequences - II

- Convergence in *Probability*: This uses convergence of the "probability of error" to zero.

$$\lim_{n\to\infty} P[|X[n] - X| > \epsilon] = 0, \quad \forall\epsilon > 0\,.$$

- Convergence *in Distribution* : This is just converge of the corresponding PDFs.

$$\lim_{n\to\infty} F_{X_n}(a) = F_X(a)$$

  at all points $a \in \mathbb{R}$ where F_X is continuous.

- Relationship between modes of convergence:

$$\left.\begin{array}{c}\text{Sure} \Rightarrow \text{Almost-sure}\\[1em] \text{Mean-square}\end{array}\right\} \Rightarrow \text{Probability} \Rightarrow \text{Distribution}$$

Uniformly-Bounded Random Sequences

- A random sequence $\{X[1], X[2], \ldots\}$ is said to be *uniformly bounded* if there exists a finite $K > 0$, *which does not depend on* n, such that

$$|X[n]| \,<\, K \text{ with prob. 1}\,, \qquad \text{for all } n = 1, 2, \ldots$$

- **Theorem.** If a random sequence $\{X[1], X[2], \ldots\}$ is uniformly bounded, then

$$X[n] \to X \text{ in m.s.} \;\Leftrightarrow\; X[n] \to X \text{ in probability}\,.$$

  That is, convergence in m.s. and in probability are the same.

- The relationship between modes of convergence becomes:

$$\text{Sure} \Rightarrow \text{Almost-sure} \Rightarrow \left\{ \begin{array}{c} \text{Mean-square} \\ \\ \text{Probability} \end{array} \right\} \Rightarrow \text{Distribution}$$

Example

Let X(n) consist of 0-1 binary random variables.

1. Set X(0) = 1
2. From the next 2 points, pick one randomly and set to 1, the other to zero.
3. From the next 3 points, pick one randomly and set to 1, the rest to zero.
4. From the next 4 points, pick one randomly and set to 1, the rest to zero.
5. $\ldots$

Then $X(n)$ is clearly converging slowly in some sense to zero, but not with probability one! In fact, $P[\{X_n = 1\}\, i.o.] = 1$. However, one can show that $X(n) \to 0$ in mean-square and also in probability. (Exercise.)

Limiting Theorems

- (Weak Law of Large Numbers). Given an i.i.d. random sequence $\{X_1, X_2, \ldots\}$ with common finite mean μ. Then

$$\frac{1}{n}\sum_{i=1}^{n} X_i \to \mu, \quad \textit{in probability}$$

- (Strong Law of Large Numbers) The same convergence holds with probability 1.

- (Central Limit Theorem) Given an i.i.d. random sequence $\{X_1, X_2, \ldots\}$ with common mean μ and variance σ^2. Then

$$\frac{1}{\sigma\sqrt{n}}\left(\sum_{i=1}^{n} X_i - n\mu\right) \to \mathcal{N}(0,1), \quad \textit{in distribution}$$