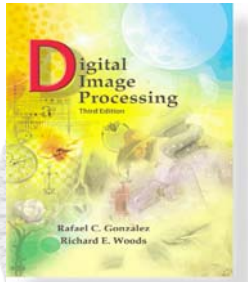
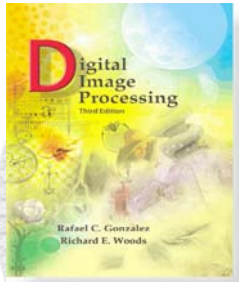


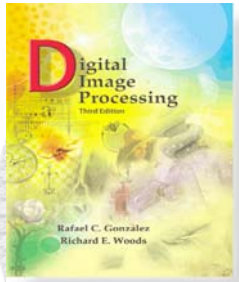
Lecture #25



Sets and Set Operations

Probability events are modeled as sets, so it is customary to begin a study of probability by defining sets and some simple operations among sets.

A ***set*** is a collection of objects, with each object in a set often referred to as an ***element*** or ***member*** of the set. Familiar examples include the set of all image processing books in the world, the set of prime numbers, and the set of planets circling the sun. Typically, sets are represented by uppercase letters, such as *A*, *B*, and *C*, and members of sets by lowercase letters, such as *a*, *b*, and *c*.



Sets and Set Operations (Con't)

We denote the fact that an ***element*** *a* ***belongs*** to set *A* by

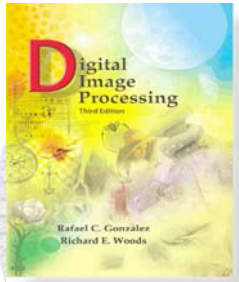
$$a \in A$$

If *a* is not an element of *A*, then we write

$$a \notin A$$

A set can be specified by listing all of its elements, or by listing properties common to all elements. For example, suppose that *I* is the set of all integers. A set *B* consisting the first five nonzero integers is specified using the notation

$$B = \{1, 2, 3, 4, 5\}$$



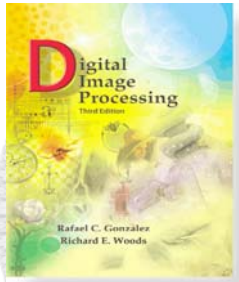
Sets and Set Operations (Con't)

The set of all integers less than 10 is specified using the notation

$$C = \{c \in I \mid c < 10\}$$

which we read as "C is the set of integers such that each members of the set is less than 10." The "such that" condition is denoted by the symbol " | " . As shown in the previous two equations, the elements of the set are enclosed by curly brackets.

The set with no elements is called the ***empty*** or ***null set***, denoted in this review by the symbol Ø.



Sets and Set Operations (Con't)

Two sets A and B are said to be ***equal*** if and only if they contain the same elements. Set equality is denoted by

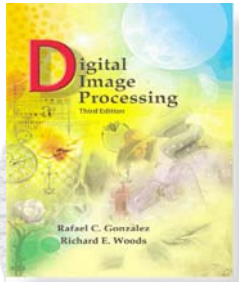
$$A = B$$

If the elements of two sets are not the same, we say that the sets are ***not equal***, and denote this by

$$A \neq B$$

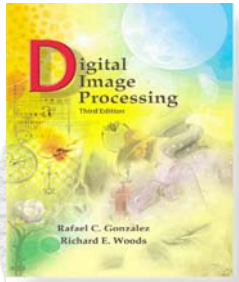
If every element of B is also an element of A, we say that B is a ***subset*** of A:

$$B \subseteq A$$



Sets and Set Operations (Con't)

Finally, we consider the concept of a ***universal set***, which we denote by U and define to be the set containing all elements of interest in a given situation. For example, in an experiment of tossing a coin, there are two possible (realistic) outcomes: heads or tails. If we denote heads by H and tails by T, the universal set in this case is $\{H,T\}$. Similarly, the universal set for the experiment of throwing a single die has six possible outcomes, which normally are denoted by the face value of the die, so in this case $U = \{1,2,3,4,5,6\}$. For obvious reasons, the universal set is frequently called the ***sample space***, which we denote by S. It then follows that, for any set A, we assume that $\emptyset \subseteq A \subseteq S$, and for any element a, $a \in$ S and $a \notin \emptyset$.



Some Basic Set Operations

The operations on sets associated with basic probability theory are straightforward. The ***union*** of two sets A and B, denoted by

$$A \cup B$$

is the set of elements that are either in A or in B, or in both. In other words,

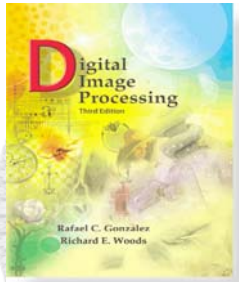
$$A \cup B = \{z \mid z \in A \text{ or } z \in B\}$$

Similarly, the ***intersection*** of sets A and B, denoted by

$$A \cap B$$

is the set of elements common to both A and B; that is,

$$A \cap B = \{z \mid z \in A \text{ and } z \in B\}$$



Set Operations (Con't)

Two sets having no elements in common are said to be ***disjoint*** or ***mutually exclusive***, in which case

$$A \cap B = \emptyset$$

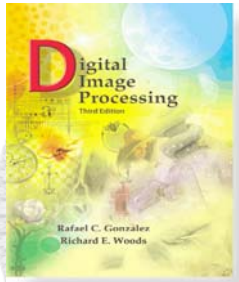
The ***complement*** of set A is defined as

$$A^c = \{z \mid z \notin A\}$$

Clearly, $(A^c)^c = A$. Sometimes the complement of A is denoted as $\overline{A}$.

The ***difference*** of two sets A and B, denoted $A - B$, is the set of elements that belong to A, but not to B. In other words,

$$A - B = \{z \mid z \in A,\ z \notin B\}$$

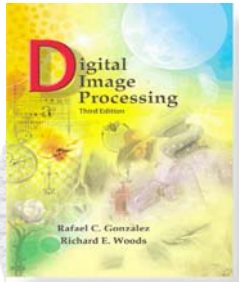


Set Operations (Con't)

It is easily verified that $(A - B) = A \cap B^c$.

The union operation is applicable to multiple sets. For example the union of sets $A_1, A_2, \ldots, A_n$ is the set of points that belong to at least one of these sets. Similar comments apply to the intersection of multiple sets.

The following table summarizes several important relationships between sets. Proofs for these relationships are found in most books dealing with elementary set theory.



Set Operations (Con't)

Some Important Set Relationships

$$S^c = \emptyset;\ \ \emptyset^c = S;$$

$$A \cup A^c = S;\ \ A \cap A^c = \emptyset$$

$$A \cup \emptyset = A;\ \ A \cap \emptyset = \emptyset;\ \ S \cup \emptyset = S;\ \ S \cap \emptyset = \emptyset$$

$$A \cup A = A;\ \ A \cap A = A;\ \ A \cup S = S;\ \ A \cap S = A$$

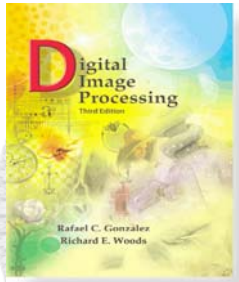
$$A \cup B = B \cup A;\ \ A \cap B = B \cap A$$

$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$$

$$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$$

$$(A \cup B) \cup C = A \cup (B \cup C) = A \cup B \cup C$$

$$(A \cap B) \cap C = A \cap (B \cap C) = A \cap B \cap C$$



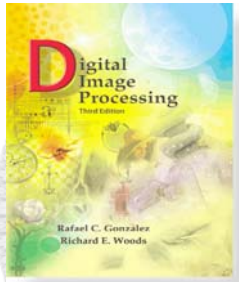
Set Operations (Con't)

It often is quite useful to represent sets and sets operations in a so-called ***Venn diagram***, in which S is represented as a rectangle, sets are represented as areas (typically circles), and points are associated with elements. The following example shows various uses of Venn diagrams.

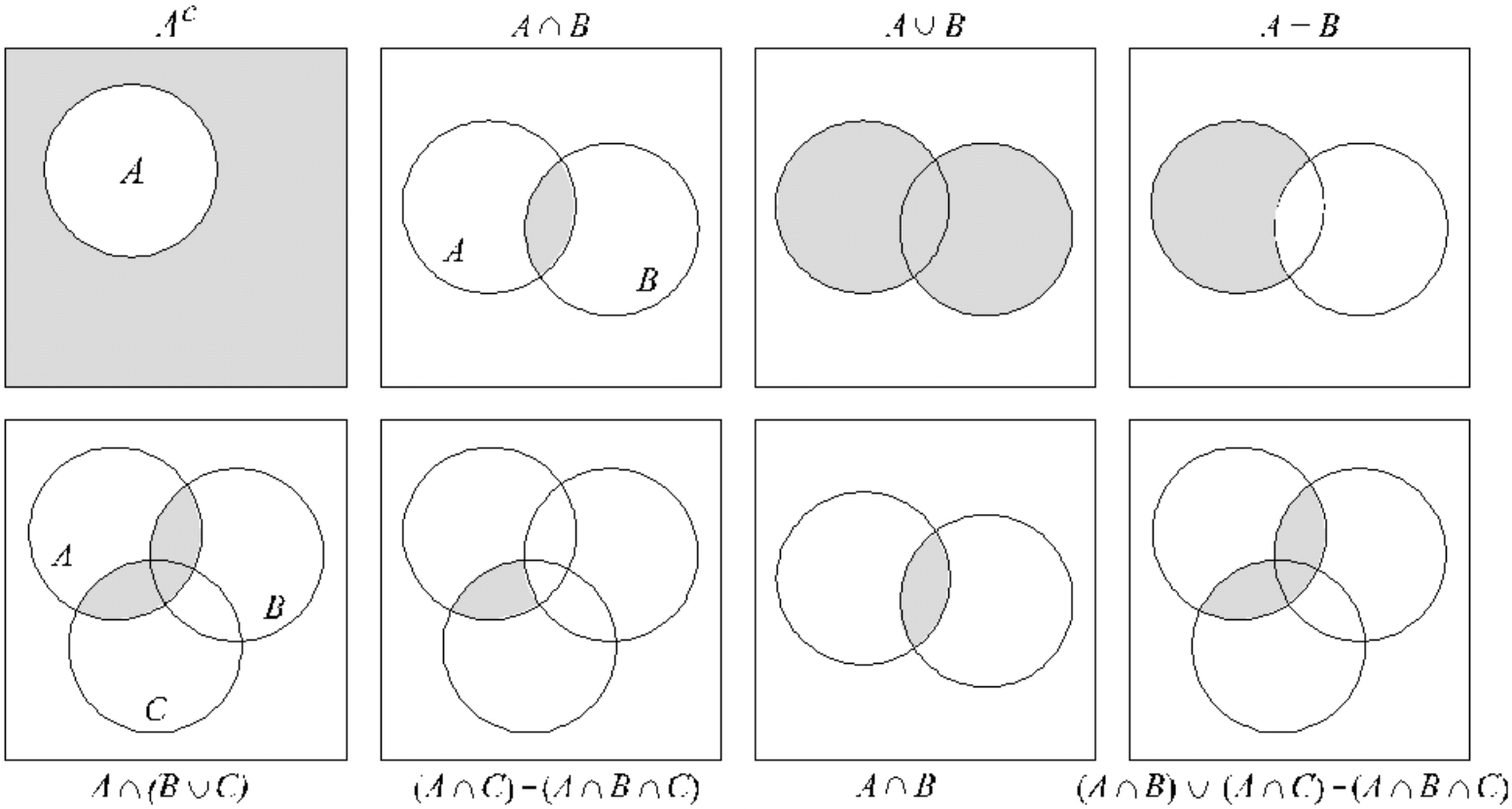
Example: The following figure shows various examples of Venn diagrams. The shaded areas are the result (sets of points) of the operations indicated in the figure. The diagrams in the top row are self explanatory. The diagrams in the bottom row are used to prove the validity of the expression

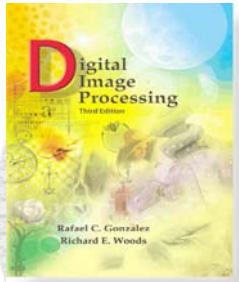
$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C) - A \cap B \cap C$$

which is used in the proof of some probability relationships.



Set Operations (Con't)

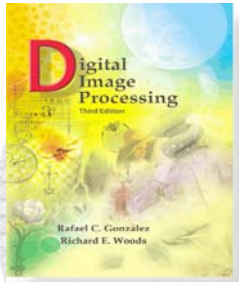




Relative Frequency & Probability

A ***random experiment*** is an experiment in which it is not possible to predict the outcome. Perhaps the best known random experiment is the tossing of a coin. Assuming that the coin is not biased, we are used to the concept that, on average, half the tosses will produce heads (H) and the others will produce tails (T). This is intuitive and we do not question it. In fact, few of us have taken the time to verify that this is true. If we did, we would make use of the concept of relative frequency. Let n denote the total number of tosses, n_H the number of heads that turn up, and n_T the number of tails. Clearly,

$$n_H + n_T = n.$$

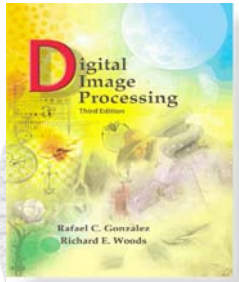


Relative Frequency & Probability

Dividing both sides by n gives

$$\frac{n_H}{n} + \frac{n_T}{n} = 1.$$

The term n_H/n is called the ***relative frequency*** of the event we have denoted by H, and similarly for n_T/n. If we performed the tossing experiment a large number of times, we would find that each of these relative frequencies tends toward a stable, limiting value. We call this value the ***probability of the event***, and denoted it by P(event).



Relative Frequency & Probability

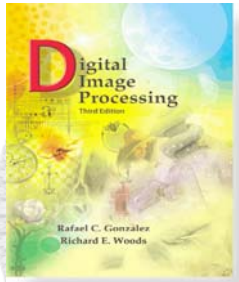
In the current discussion the probabilities of interest are $P(H)$ and $P(T)$. We know in this case that $P(H) = P(T) = 1/2$. Note that the event of an experiment need not signify a single outcome. For example, in the tossing experiment we could let D denote the event "heads or tails," (note that the event is now a set) and the event E, "neither heads nor tails." Then, $P(D) = 1$ and $P(E) = 0$.

The first important property of P is that, for an event A,

$$0 \leq P(A) \leq 1.$$

That is, the probability of an event is a positive number bounded by 0 and 1. For the certain event, S,

$$P(S) = 1.$$

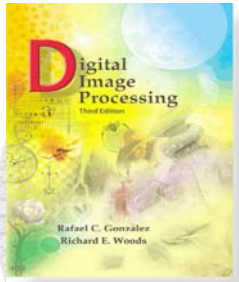


Relative Frequency & Probability

Here the certain event means that the outcome is from the universal or sample set, S. Similarly, we have that for the impossible event, S^c

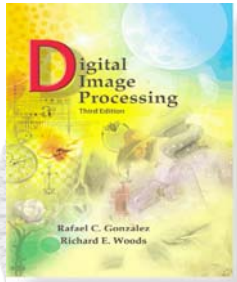
$$P(S^c) = 0.$$

This is the probability of an event being outside the sample set. In the example given at the end of the previous paragraph, $S = D$ and $S^c = E$.



Relative Frequency & Probability

The ***event*** that either events *A* ***or*** *B* ***or both*** have occurred is simply the union of *A* and *B* (recall that events can be sets). Earlier, we denoted the union of two sets by $A \cup B$. One often finds the equivalent notation $A+B$ used interchangeably in discussions on probability. Similarly, the event that ***both*** *A* ***and*** *B* occurred is given by the intersection of *A* and *B*, which we denoted earlier by $A \cap B$. The equivalent notation AB is used much more frequently to denote the occurrence of both events in an experiment.



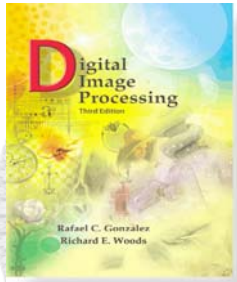
Relative Frequency & Probability

Suppose that we conduct our experiment n times. Let n_1 be the number of times that only event A occurs; n_2 the number of times that B occurs; n_3 the number of times that AB occurs; and n_4 the number of times that neither A nor B occur. Clearly, $n_1+n_2+n_3+n_4=n$. Using these numbers we obtain the following relative frequencies:

$$\frac{n_A}{n} = \frac{n_1 + n_3}{n}$$

$$\frac{n_B}{n} = \frac{n_2 + n_3}{n}$$

$$\frac{n_{AB}}{n} = \frac{n_3}{n}$$



Relative Frequency & Probability

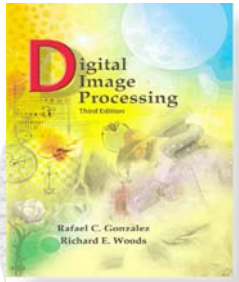
and

$$\begin{aligned}\frac{n_{A\cup B}}{n} &= \frac{n_1 + n_2 + n3}{n} \\ &= \frac{(n_1 + n_3) + (n_2 + n_3) - n_3}{n} \\ &= \frac{n_A}{n} + \frac{n_B}{n} - \frac{n_{AB}}{n}.\end{aligned}$$

Using the previous definition of probability based on relative frequencies we have the important result

$$P(A \cup B) = P(A) + P(B) - P(AB).$$

If A and B are ***mutually exclusive*** it follows that the set AB is empty and, consequently, $P(AB) = 0$.

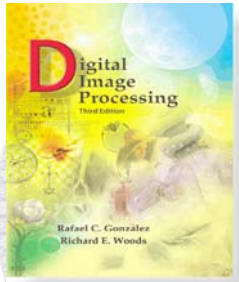


Conditional Probability

The relative frequency of event A occurring, ***given that*** event B has occurred, is given by

$$\frac{n_{A/B}}{n} = \frac{\frac{n_{AB}}{n}}{\frac{n_B}{n}} = \frac{n_3}{n_2 + n_3}.$$

This ***conditional probability*** is denoted by $P(A/B)$, where we note the use of the symbol " / " to denote conditional occurrence. It is common terminology to refer to $P(A/B)$ as the ***probability of*** A ***given*** B.

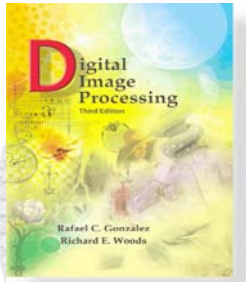


Conditional Probability

Similarly, the relative frequency of B occurring, given that A has occurred is

$$\frac{n_{B/A}}{n} = \frac{\frac{n_{AB}}{n}}{\frac{n_A}{n}} = \frac{n_3}{n_1 + n_3}.$$

We call this relative frequency ***the probability of*** B ***given*** A, and denote it by $P(B/A)$.



Bayes Theorem

$$P(A/B) = \frac{P(B/A)P(A)}{P(B)}$$

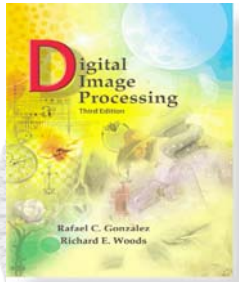
and

$$P(AB) = P(A)P(B/A) = P(B)P(A/B).$$

The second expression may be written as

$$P(B/A) = \frac{P(A/B)P(B)}{P(A)}$$

which is known as ***Bayes' theorem***, so named after the 18th century mathematician Thomas Bayes.



Bayes Theorem

Example: Suppose that we want to extend the expression

$$P(A \cup B) = P(A) + P(B) - P(AB)$$

to three variables, A, B, and C. Recalling that AB is the same as $A \cap B$, we replace B by $B \cup C$ in the preceding equation to obtain

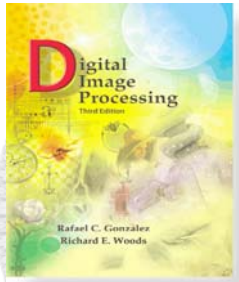
$$P(A \cup B \cup C) = P(A) + P(B \cup C) - P(A \cap [B \cup C]).$$

The second term in the right can be written as

$$P(B \cup C) = P(B) + P(C) - P(BC).$$

From the Table discussed earlier, we know that

$$A \cap [B \cup C] = (A \cap B) \cup (A \cap C)$$



Bayes Theorem

so,

$$\begin{aligned} P(A \cap [B \cup C]) &= P([A \cap B] \cup [A \cap C]) \\ &= P(AB \cup AC) \\ &= P(AB) + P(AC) - P(ABC). \end{aligned}$$

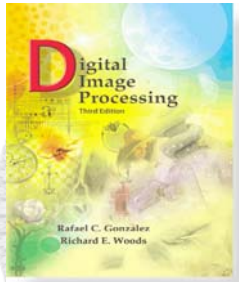
Collecting terms gives us the final result

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(AB) - P(AC) - P(BC) + P(ABC).$$

Proceeding in a similar fashion gives

$$P(ABC) = P(A)P(B/A)P(C/AB).$$

The preceding approach can be used to generalize these expressions to N events.



Statistical Independence

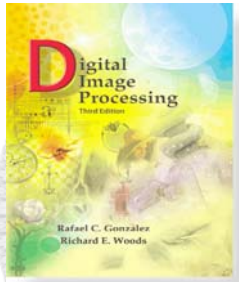
If A and B are ***statistically independent***, then $P(B/A) = P(B)$ and it follows that

$$P(A/B) = P(A)$$
$$P(B/A) = P(B)$$

and

$$P(AB) = P(A)P(B).$$

It was stated earlier that if sets (events) A and B are ***mutually exclusive***, then $A \cap B = \emptyset$ from which it follows that $P(AB) = P(A \cap B) = 0$. As was just shown, the two sets are statistically independent if $P(AB)=P(A)P(B)$, which we assume to be nonzero in general. ***Thus, we conclude that for two events to be statistically independent, they cannot be mutually exclusive***.



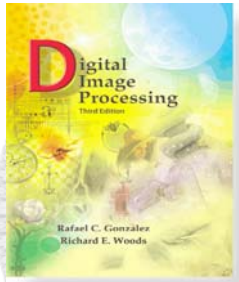
Statistical Independence

For three events A, B, and C to be independent, it must be true that

$$P(AB) = P(A)P(B)$$
$$P(AC) = P(A)P(C)$$
$$P(BC) = P(B)P(C)$$

and

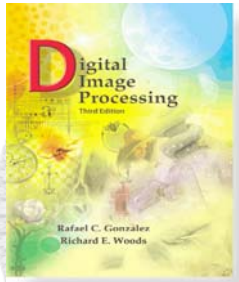
$$P(ABC) = P(A)P(B)P(C).$$



Statistical Independence

In general, for N events to be statistically independent, it must be true that, for all combinations $1 \leq i \leq j \leq k \leq \ldots \leq N$

$$P(A_iA_j) = P(A_i)P(A_j)$$
$$P(A_iA_jA_k) = P(A_i)P(A_j)P(A_k)$$
$$\vdots$$
$$P(A_1A_2\cdots A_N) = P(A_1)P(A_2)\cdots P(A_N).$$

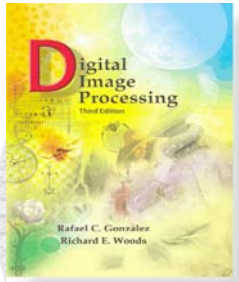


Statistical Independence

Example: (a) An experiment consists of throwing a single die twice. The probability of any of the six faces, 1 through 6, coming up in either experiment is 1/6. Suppose that we want to find the probability that a 2 comes up, followed by a 4. These two events are statistically independent (the second event does not depend on the outcome of the first). Thus, letting *A* represent a 2 and *B* a 4,

$$P(AB) = P(A)P(B) = \frac{1}{6} \times \frac{1}{6} = \frac{1}{36}.$$

We would have arrived at the same result by defining "2 followed by 4" to be a single event, say *C*. The sample set of all possible outcomes of two throws of a die is 36. Then, $P(C)=1/36$.



Statistical Independence

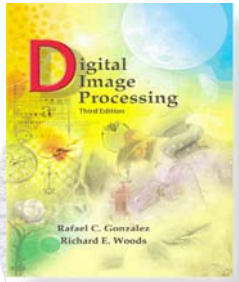
Example (Con't): (b) Consider now an experiment in which we draw one card from a standard card deck of 52 cards. Let *A* denote the event that a king is drawn, *B* denote the event that a queen or jack is drawn, and *C* the event that a diamond-face card is drawn. A brief review of the previous discussion on relative frequencies would show that

$$P(A) = \frac{4}{52},$$

$$P(B) = \frac{8}{52},$$

and

$$P(C) = \frac{13}{52}.$$



Statistical Independence

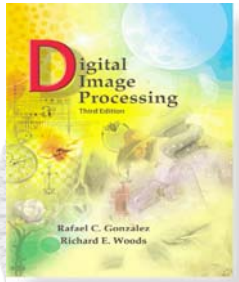
Example (Con't): Furthermore,

$$P(AC) = P(A \cap C) = P(A)P(C) = \frac{1}{52}$$

and

$$P(BC) = P(B \cap C) = P(B)P(C) = \frac{2}{52}.$$

Events A and B are mutually exclusive (we are drawing only one card, so it would be impossible to draw a king and a queen or jack simultaneously). Thus, it follows from the preceding discussion that $P(AB) = P(A \cap B) = 0$ [and also that $P(AB) \neq P(A)P(B)$].

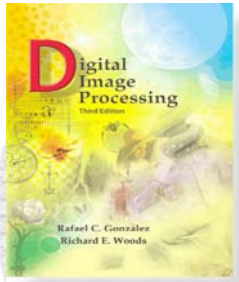


Statistical Independence

Example (Con't): (c) As a final experiment, consider the deck of 52 cards again, and let A_1, A_2, A_3, and A_4 represent the events of drawing an ace in each of four successive draws. If we replace the card drawn before drawing the next card, then the events are statistically independent and it follows that

$$P(A_1A_2A_3A_4) = P(A_1)P(A_2)P(A_3)P(A_4)$$

$$= \left[\frac{4}{52}\right]^4 \approx 3.5 \times 10^{-5}.$$

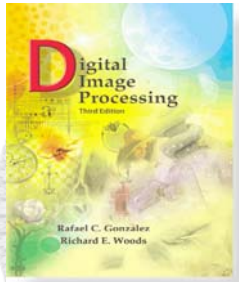


Statistical Independence

Example (Con't): Suppose now that we do not replace the cards that are drawn. The events then are no longer statistically independent. With reference to the results in the previous example, we write

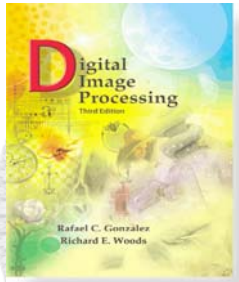
$$\begin{aligned} P(A_1A_2A_3A_4) &= P(A_1)P(A_2A_3A_4/A_1) \\ &= P(A_1)P(A_2/A_1)P(A_3A_4/A_1A_2) \\ &= P(A_1)P(A_2/A_1)P(A_3/A_1A_2)P(A_4/A_1A_2A_3) \\ &= \frac{4}{52}\cdot\frac{3}{51}\cdot\frac{2}{50}\cdot\frac{1}{49} \approx 3.7\times 10^{-6}. \end{aligned}$$

Thus we see that not replacing the drawn card reduced our chances of drawing fours successive aces by a factor of close to 10. This significant difference is perhaps larger than might be expected from intuition.



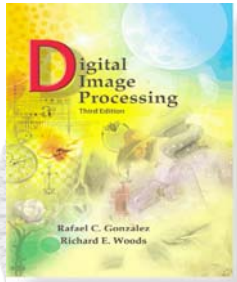
Random Variables

Random variables often are a source of confusion when first encountered. This need not be so, as the concept of a random variable is in principle quite simple. A ***random variable***, x, is a real-valued function ***defined*** on the events of the sample space, S. In words, for each event in S, there is a real number that is the corresponding value of the random variable. Viewed yet another way, a random variable maps each event in S onto the real line. That is it. A simple, straightforward definition.



Random Variables

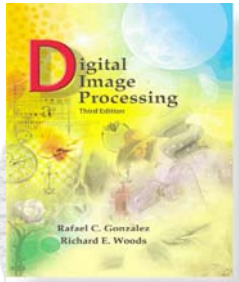
Part of the confusion often found in connection with random variables is the fact that they are ***functions***. The notation also is partly responsible for the problem. In other words, although typically the notation used to denote a random variable is as we have shown it here, x, or some other appropriate variable, to be strictly formal, a random variable should be written as a function $x(\cdot)$ where the argument is a specific event being considered. However, this is seldom done, and, in our experience, trying to be formal by using function notation complicates the issue more than the clarity it introduces. Thus, we will opt for the less formal notation, with the warning that it must be keep clearly in mind that random variables are functions.



Random Variables

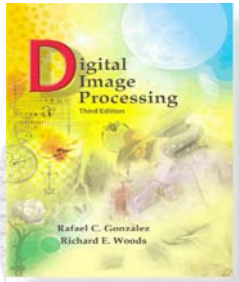
Example: Consider again the experiment of drawing a single card from a standard deck of 52 cards. Suppose that we define the following events. *A*: a heart; *B*: a spade; *C*: a club; and *D*: a diamond, so that $S = \{A, B, C, D\}$. A random variable is easily defined by letting $x = 1$ represent event *A*, $x = 2$ represent event *B*, and so on.

As a second illustration, consider the experiment of throwing a single die and observing the value of the up-face. We can define a random variable as the numerical outcome of the experiment (i.e., 1 through 6), but there are many other possibilities. For example, a binary random variable could be defined simply by letting $x = 0$ represent the event that the outcome of throw is an even number and $x = 1$ otherwise.



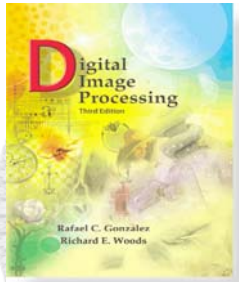
Random Variables

Note the important fact in the examples just given that the probability of the events have not changed; all a random variable does is map events onto the real line.



Random Variables

Thus far we have been concerned with random variables whose values are discrete. To handle ***continuous random variables*** we need some additional tools. In the discrete case, the probabilities of events are numbers between 0 and 1. When dealing with continuous quantities (which are not denumerable) we can no longer talk about the "probability of an event" because that probability is zero. This is not as unfamiliar as it may seem. For example, given a continuous function we know that the area of the function between two limits *a* and *b* is the integral from *a* to *b* of the function. However, the area ***at a point*** is zero because the integral from,say, *a* to *a* is zero. We are dealing with the same concept in the case of continuous random variables.

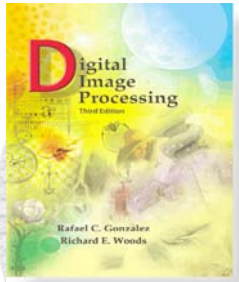


Random Variables

Thus, instead of talking about the probability of a specific value, we talk about the probability that the value of the random variable lies in a specified ***range***. In particular, we are interested in the probability that the random variable is less than or equal to (or, similarly, greater than or equal to) a specified constant a. We write this as

$$F(a) = P(x \leq a).$$

If this function is given for all values of a (i.e., $-\infty < a < \infty$), then the values of random variable x have been defined. Function F is called the ***cumulative probability distribution function*** or simply the ***cumulative distribution function*** (cdf). The shortened term ***distribution function*** also is used.

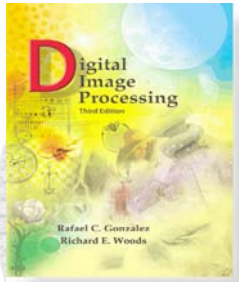


Random Variables

Observe that the notation we have used makes no distinction between a random variable and the values it assumes. If confusion is likely to arise, we can use more formal notation in which we let capital letters denote the random variable and lowercase letters denote its values. For example, the cdf using this notation is written as

$$F_X(x) = P(X \leq x).$$

When confusion is not likely, the cdf often is written simply as $F(x)$. This notation will be used in the following discussion when speaking generally about the cdf of a random variable.

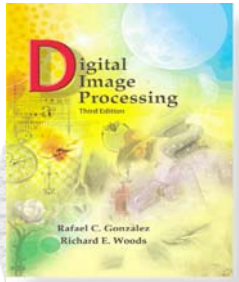


Random Variables

Due to the fact that it is a probability, the cdf has the following properties:

1. $F(-\infty) = 0$
2. $F(\infty) = 1$
3. $0 \leq F(x) \leq 1$
4. $F(x_1) \leq F(x_2)$ if $x_1 < x_2$
5. $P(x_1 < x \leq x_2) = F(x_2) - F(x_1)$
6. $F(x^+) = F(x)$,

where $x^+ = x + \varepsilon$, with ε being a positive, infinitesimally small number.



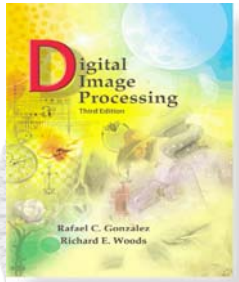
Random Variables

The ***probability density function*** (pdf) of random variable x is defined as the derivative of the cdf:

$$p(x) = \frac{dF(x)}{dx}.$$

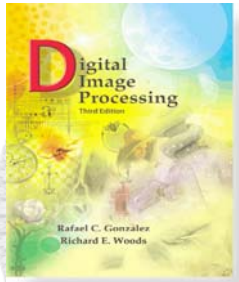
The term ***density function*** is commonly used also. The pdf satisfies the following properties:

1. $p(x) \geq 0$ for all x
2. $\int_{-\infty}^{\infty} p(x)dx = 1$
3. $F(x) = \int_{-\infty}^{x} p(\alpha)d\alpha$, where α is a dummy variable
4. $P(x_1 < x \leq x_2) = \int_{x_1}^{x_2} p(x)dx.$



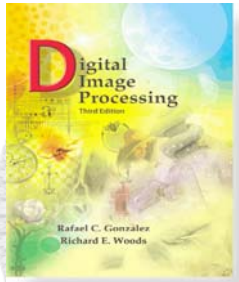
Random Variables

The preceding concepts are applicable to discrete random variables. In this case, there is a finite no. of events and we talk about ***probabilities***, rather than probability density functions. Integrals are replaced by summations and, sometimes, the random variables are subscripted. For example, in the case of a discrete variable with N possible values we would denote the probabilities by $P(x_i)$, i=1, 2,…, N.



Random Variables

In Sec. 3.3 of the book we used the notation $p(r_k)$, $k = 0,1,\ldots, L - 1$, to denote the ***histogram*** of an image with L possible gray levels, r_k, $k = 0,1,\ldots, L - 1$, where $p(r_k)$ is the probability of the kth gray level (random event) occurring. The discrete random variables in this case are gray levels. It generally is clear from the context whether one is working with continuous or discrete random variables, and whether the use of subscripting is necessary for clarity. Also, uppercase letters (e.g., P) are frequently used to distinguish between probabilities and probability density functions (e.g., p) when they are used together in the same discussion.

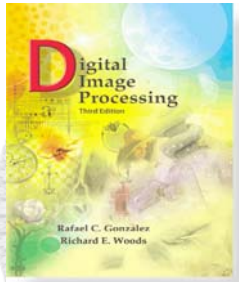


Random Variables

If a random variable x is ***transformed*** by a monotonic transformation function $T(x)$ to produce a new random variable y, the probability density function of y can be obtained from knowledge of $T(x)$ and the probability density function of x, as follows:

$$p_y(y) = p_x(x)\left|\frac{dx}{dy}\right|$$

where the subscripts on the p's are used to denote the fact that they are different functions, and the vertical bars signify the absolute value. A function $T(x)$ is ***monotonically increasing*** if $T(x_1) < T(x_2)$ for $x_1 < x_2$, and ***monotonically decreasing*** if $T(x_1) > T(x_2)$ for $x_1 < x_2$. The preceding equation is valid if $T(x)$ is an increasing or decreasing monotonic function.



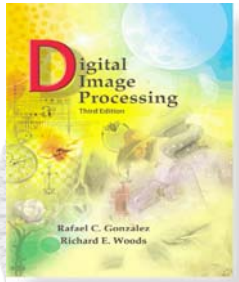
Expected Value & Moments

The ***expected value*** of a function $g(x)$ of a ***continuos*** random variable is defined as

$$E[g(x)] = \int_{-\infty}^{\infty} g(x)p(x)dx.$$

If the random variable is ***discrete*** the definition becomes

$$E[g(x)] = \sum_{i=1}^{N} g(x_i)P(x_i).$$



Expected Value & Moments

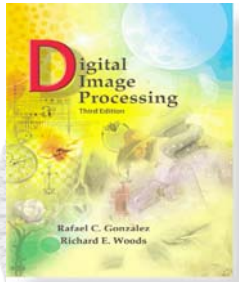
The expected value is one of the operations used most frequently when working with random variables. For example, the expected value of random variable x is obtained by letting $g(x) = x$:

$$E[x] = \overline{x} = m = \int_{-\infty}^{\infty} xp(x)dx$$

when x is continuos and

$$E[x] = \overline{x} = m = \sum_{i=1}^{N} x_i P(x_i)$$

when x is discrete. The expected value of x is equal to its ***average*** (or ***mean***) ***value***, hence the use of the equivalent notation $\overline{x}$ and m.



Expected Value & Moments

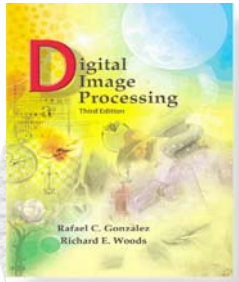
The ***variance*** of a random variable, denoted by σ^2, is obtained by letting $g(x) = x^2$ which gives

$$\sigma^2 = E[x^2] = \int_{-\infty}^{\infty} x^2 p(x) dx$$

for continuous random variables and

$$\sigma^2 = E[x^2] = \sum_{i=1}^{N} x_i^2 P(x_i)$$

for discrete variables.



Expected Value & Moments

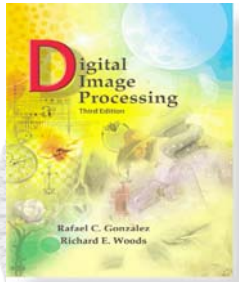
Of particular importance is the variance of random variables that have been ***normalized*** by subtracting their mean. In this case, the variance is

$$\sigma^2 = E[(x-m)^2] = \int_{-\infty}^{\infty}(x-m)^2 p(x)dx$$

and

$$\sigma^2 = E[(x-m)^2] = \sum_{i=1}^{N}(x_i-m)^2 P(x_i)$$

for continuous and discrete random variables, respectively. The square root of the variance is called the ***standard deviation***, and is denoted by σ.



Expected Value & Moments

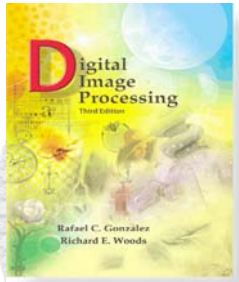
We can continue along this line of thought and define the *n*th ***central moment*** of a continuous random variable by letting $g(x) = (x - m)^n$:

$$\mu_n = E[(x-m)^n] = \int_{-\infty}^{\infty} (x-m)^n p(x)dx$$

and

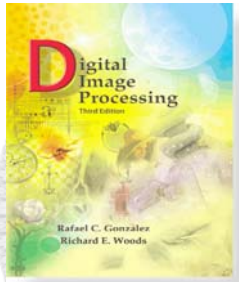
$$\mu_n = E[(x-m)^n] = \sum_{i=1}^{N} (x_i - m)^n P(x_i)$$

for discrete variables, where we assume that $n \geq 0$. Clearly, $\mu_0=1$, $\mu_1=0$, and $\mu_2=\sigma^2$. The term ***central*** when referring to moments indicates that the mean of the random variables has been subtracted out. The moments defined above in which the mean is not subtracted out sometimes are called ***moments about the origin***.



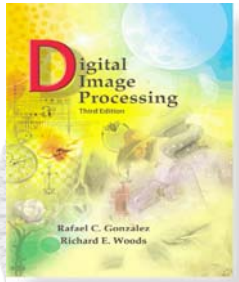
Expected Value & Moments

In image processing, moments are used for a variety of purposes, including histogram processing, segmentation, and description. In general, moments are used to characterize the probability density function of a random variable. For example, the second, third, and fourth central moments are intimately related to the ***shape*** of the probability density function of a random variable. The second central moment (the centralized variance) is a measure of ***spread*** of values of a random variable about its mean value, the third central moment is a measure of ***skewness*** (bias to the left or right) of the values of x about the mean value, and the fourth moment is a relative measure of ***flatness***. In general, knowing all the moments of a density specifies that density.



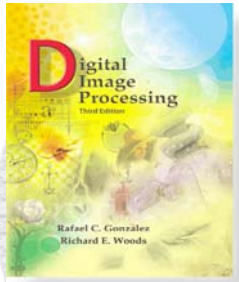
Expected Value & Moments

Example: Consider an experiment consisting of repeatedly firing a rifle at a target, and suppose that we wish to characterize the behavior of bullet impacts on the target in terms of whether we are shooting high or low.. We divide the target into an upper and lower region by passing a horizontal line through the bull's-eye. The events of interest are the vertical distances from the center of an impact hole to the horizontal line just described. Distances above the line are considered positive and distances below the line are considered negative. The distance is zero when a bullet hits the line.



Expected Value & Moments

In this case, we define a random variable directly as the value of the distances in our sample set. Computing the mean of the random variable indicates whether, ***on average***, we are shooting high or low. If the mean is zero, we know that the average of our shots are on the line. However, the mean does not tell us how far our shots deviated from the horizontal. The variance (or standard deviation) will give us an idea of the ***spread of the shots***. A small variance indicates a tight grouping (with respect to the mean, and in the vertical position); a large variance indicates the opposite. Finally, a third moment of zero would tell us that the spread of the shots is symmetric about the mean value, a positive third moment would indicate a high bias, and a negative third moment would tell us that we are shooting low more than we are shooting high with respect to the mean location.



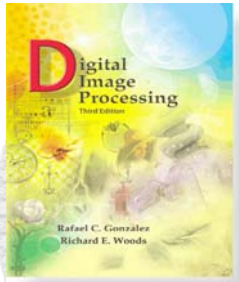
Gaussian Probability Density Function

Because of its importance, we will focus in this tutorial on the ***Gaussian probability density function*** to illustrate many of the preceding concepts, and also as the basis for generalization to more than one random variable. The reader is referred to Section 5.2.2 of the book for examples of other density functions.

A random variable is called ***Gaussian*** if it has a probability density of the form

$$p(x) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-(x-m)^2/\sigma^2}$$

where m and σ are as defined in the previous section. The term ***normal*** also is used to refer to the Gaussian density. A plot and properties of this density function are given in Section 5.2.2 of the book.

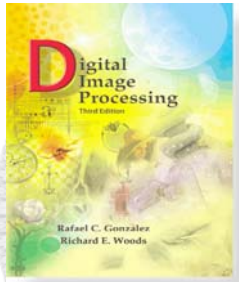


Gaussian Probability Density Function

The cumulative distribution function corresponding to the Gaussian density is

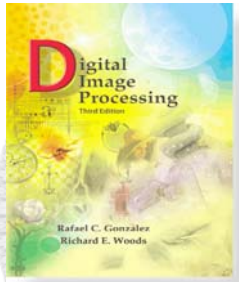
$$F(x) = \int_{-\infty}^{x} p(x)dx$$
$$= \frac{1}{\sqrt{2\pi}\,\sigma} \int_{-\infty}^{x} e^{-(x-m)^2/\sigma^2} dx.$$

which, as before, we interpret as the probability that the random variable lies between minus infinite and an arbitrary value x. This integral has no known closed-form solution, and it must be solved by numerical or other approximation methods. Extensive tables exist for the Gaussian cdf.



Multiple Random Variables

In the previous example, we used a single random variable to describe the behavior of rifle shots with respect to a horizontal line passing through the bull's-eye in the target. Although this is useful information, it certainly leaves a lot to be desired in terms of telling us how well we are shooting with respect to the center of the target. In order to do this we need two random variables that will map our events onto the *xy*-plane. It is not difficult to see how if we wanted to describe events in 3-D space we would need three random variables. In general, we consider in this section the case of *n* random variables, which we denote by x_1, x_2,…, x_n (the use of *n* here is not related to our use of the same symbol to denote the *n*th moment of a random variable).



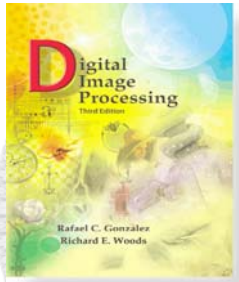
Multiple Random Variables

It is convenient to use vector notation when dealing with several random variables. Thus, we represent a ***vector random variable*** **x** as

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}.$$

Then, for example, the cumulative distribution function introduced earlier becomes

$$\begin{aligned} F(\mathbf{a}) &= F(a_1, a_2, \ldots, a_n) \\ &= P\{x_1 \leq a_1, x_2 \leq a_2, \ldots, x_n \leq a_n\} \end{aligned}$$

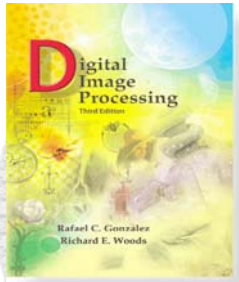


Multiple Random Variables

when using vectors. As before, when confusion is not likely, the ***cdf of a random variable vector*** often is written simply as $F(\mathbf{x})$. This notation will be used in the following discussion when speaking generally about the cdf of a random variable vector.

As in the single variable case, the ***probability density function of a random variable vector*** is defined in terms of derivatives of the cdf; that is,

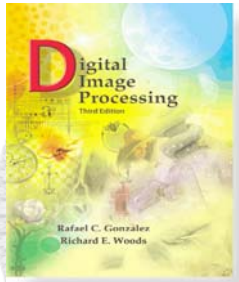
$$p(\mathbf{x}) = p(x_1, x_2, \ldots, x_n) = \frac{\partial^n F(x_1, x_2, \ldots, x_n)}{\partial x_1 \partial x_2 \cdots \partial x_n}.$$



Multiple Random Variables

The ***expected value*** of a function of **x** is defined basically as before:

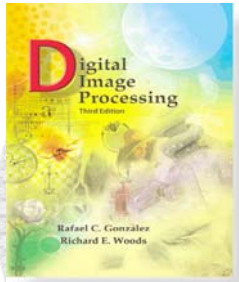
$$E[g(\mathbf{x})] = E[g(x_1, x_2, \ldots, x_n)]$$
$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(x_1, x_2, \ldots, x_n) p(x_1, x_2, \ldots, x_n) dx_1 dx_2 \cdots dx_n.$$



Multiple Random Variables

Cases dealing with expectation operations involving pairs of elements of **x** are particularly important. For example, the joint moment (about the origin) of order *kq* between variables x_i and x_j

$$\eta_{kq}(i,j) = E[x_i^k x_j^q] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x_i^k x_j^q p(x_i,x_j)dx_i dx_j.$$

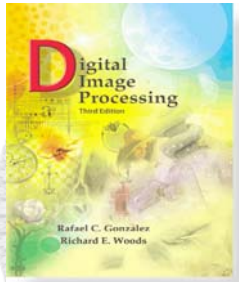


Multiple Random Variables

When working with any two random variables (any two elements of **x**) it is common practice to simplify the notation by using x and y to denote the random variables. In this case the joint moment just defined becomes

$$\eta_{kq} = E[x^k y^q] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x^k y^q p(x,y)\,dx\,dy.$$

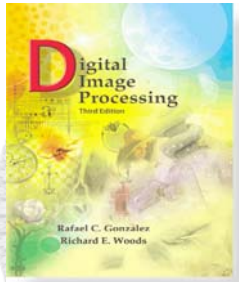
It is easy to see that η_{k0} is the kth moment of x and η_{0q} is the qth moment of y, as defined earlier.



Multiple Random Variables

The moment $\eta_{11} = E[xy]$ is called the ***correlation*** of x and y. As discussed in Chapters 4 and 12 of the book, correlation is an important concept in image processing. In fact, it is important in most areas of signal processing, where typically it is given a special symbol, such as R_{xy}:

$$R_{xy} = \eta_{11} = E[xy] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xyp(x,y)dxdy.$$



Multiple Random Variables

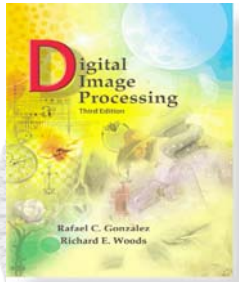
If the condition

$$R_{xy} = E[x]E[y]$$

holds, then the two random variables are said to be ***uncorrelated***. From our earlier discussion, we know that if x and y are ***statistically independent***, then $p(x, y) = p(x)p(y)$, in which case we write

$$R_{xy} = \int_{-\infty}^{\infty} xp(x)dx \int_{-\infty}^{\infty} yp(y)dy = E[x]E[y].$$

Thus, we see that ***if two random variables are statistically independent then they are also uncorrelated***. The converse of this statement is ***not*** true in general.



Multiple Random Variables

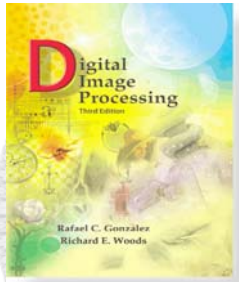
The joint central moment of order kq involving random variables x and y is defined as

$$\mu_{kq} = E[(x - m_x)^k (y - m_y)^q]$$
$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - m_x)^k (y - m_y)^q p(x,y) dx dy$$

where $m_x = E[x]$ and $m_y = E[y]$ are the means of x and y, as defined earlier. We note that

$$\mu_{20} = E[(x - m_x)^2] \quad \text{and} \quad \mu_{02} = E[(y - m_y)^2]$$

are the variances of x and y, respectively.

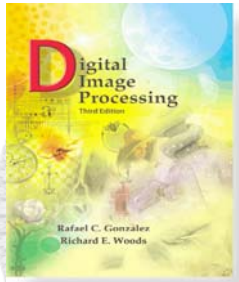


Covariance

The moment μ_{11}

$$\mu_{11} = E[(x - m_x)(y - m_y)]$$
$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - m_x)(y - m_y) p(x,y) dx dy$$

is called the ***covariance*** of x and y. As in the case of correlation, the covariance is an important concept, usually given a special symbol such as C_{xy}.

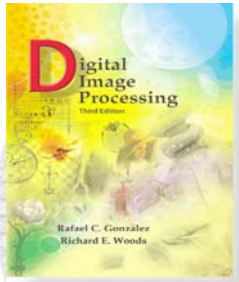


Covariance

By direct expansion of the terms inside the expected value brackets, and recalling the $m_x = E[x]$ and $m_y = E[y]$, it is straightforward to show that

$$\begin{aligned} C_{xy} &= E[xy] - m_y E[x] - m_x E[y] + m_x m_y \\ &= E[xy] - E[x]E[y] \\ &= R_{xy} - E[x]E[y]. \end{aligned}$$

From our discussion on correlation, we see that the covariance is zero if the random variables are either uncorrelated ***or*** statistically independent. This is an important result worth remembering.



Covariance

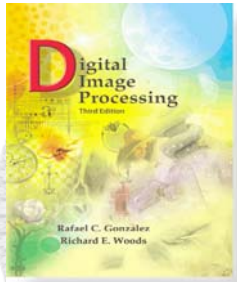
If we divide the covariance by the square root of the product of the variances we obtain

$$\gamma = \frac{\mu_{11}}{\sqrt{\mu_{20}\mu_{02}}}$$

$$= \frac{C_{xy}}{\sigma_x \sigma_y}$$

$$= E\left[\frac{(x - m_x)}{\sigma_x}\frac{(y - m_y)}{\sigma_y}\right].$$

The quantity γ is called the ***correlation coefficient*** of random variables x and y. It can be shown that γ is in the range $-1 \leq \gamma \leq 1$ (see Problem 12.5). As discussed in Section 12.2.1, the correlation coefficient is used in image processing for matching.

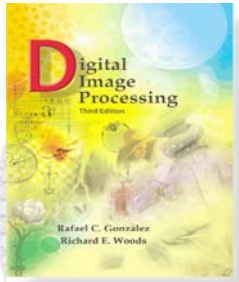


Multivariate Gaussian Distribution

As an illustration of a probability density function of more than one random variable, we consider the ***multivariate Gaussian probability density function***, defined as

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{n/2}|\mathbf{C}|^{1/2}} e^{-\frac{1}{2}\left[(\mathbf{x}-\mathbf{m})^T\mathbf{C}^{-1}(\mathbf{x}-\mathbf{m})\right]}$$

where n is the ***dimensionality*** (number of components) of the random vector $\mathbf{x}$, $\mathbf{C}$ is the ***covariance matrix*** (to be defined below), $|\mathbf{C}|$ is the determinant of matrix $\mathbf{C}$, $\mathbf{m}$ is the ***mean vector*** (also to be defined below) and T indicates transposition (see the review of matrices and vectors).



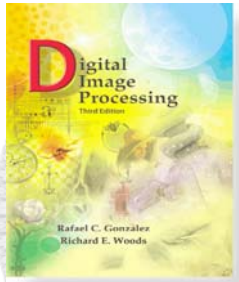
Multivariate Gaussian Distribution

The ***mean vector*** is defined as

$$\mathbf{m} = E[\mathbf{x}] = \begin{bmatrix} E[x_1] \\ E[x_2] \\ \vdots \\ E[x_n] \end{bmatrix}$$

and the ***covariance matrix*** is defined as

$$\mathbf{C} = E[(\mathbf{x} - \mathbf{m})(\mathbf{x} - \mathbf{m})^T].$$

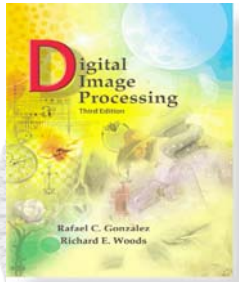


Multivariate Gaussian Distribution

The element of **C** are the covariances of the elements of **x**, such that

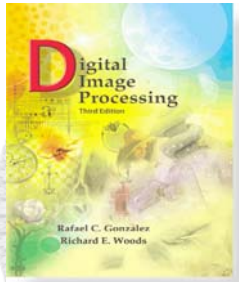
$$c_{ij} = C_{x_i x_j} = E[(x_i - m_i)(x_j - m_j)]$$

where, for example, x_i is the ith component of **x** and m_i is the ith component of **m**.



Multivariate Gaussian Distribution

Covariance matrices are ***real*** and ***symmetric*** (see the review of matrices and vectors). The elements along the main diagonal of **C** are the variances of the elements **x**, such that $c_{ii} = \sigma_{x_i}^2$. When all the elements of **x** are uncorrelated or statistically independent, $c_{ij} = 0$, and the covariance matrix becomes a ***diagonal matrix***. If all the variances are equal, then the covariance matrix becomes proportional to the ***identity matrix***, with the constant of proportionality being the variance of the elements of **x**.



Multivariate Gaussian Distribution

Example: Consider the following ***bivariate*** ($n = 2$) Gaussian probability density function

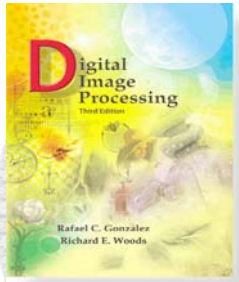
$$p(\mathbf{x}) = \frac{1}{(2\pi)^{n/2}|\mathbf{C}|^{1/2}} e^{-\frac{1}{2}[(\mathbf{x}-\mathbf{m})^T\mathbf{C}^{-1}(\mathbf{x}-\mathbf{m})]}$$

with

$$\mathbf{m} = \begin{bmatrix} m_1 \\ m_2 \end{bmatrix}$$

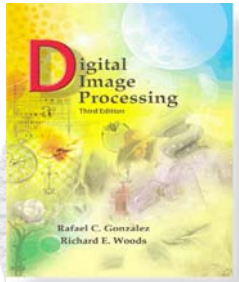
and

$$\mathbf{C} = \begin{bmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{bmatrix}$$

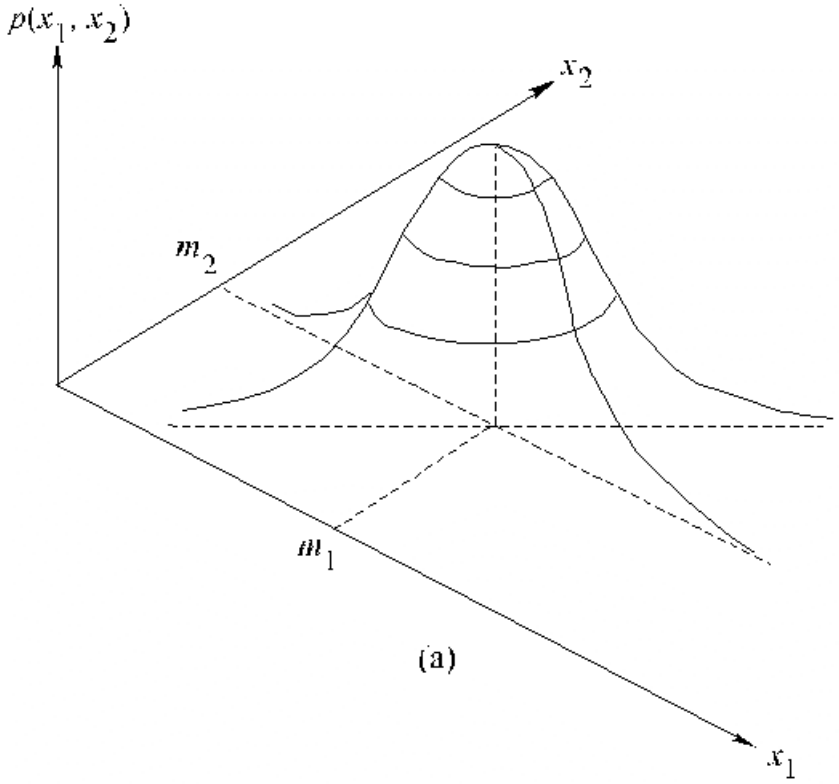


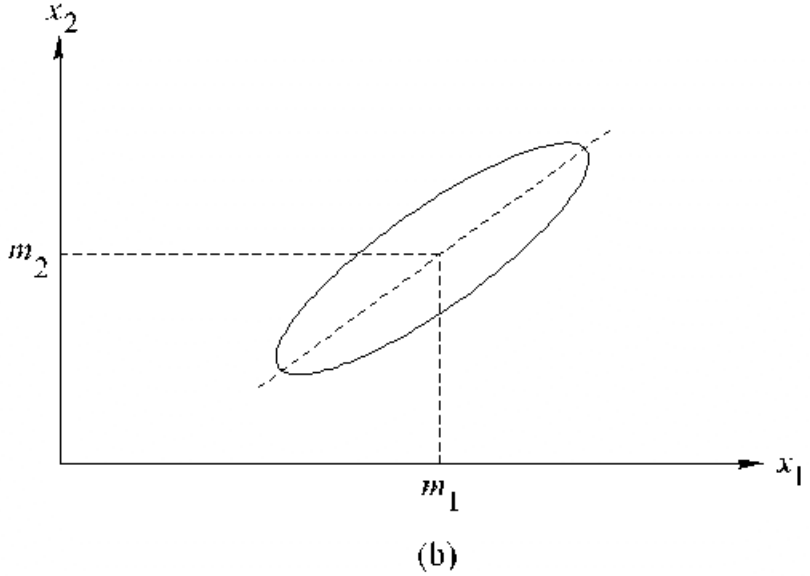
Multivariate Gaussian Distribution

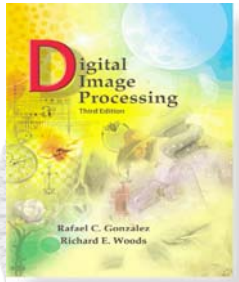
where, because **C** is known to be symmetric, $c_{12} = c_{21}$. A schematic diagram of this density is shown in Part (a) of the following figure. Part (b) is a horizontal slice of Part (a). From the review of vectors and matrices, we know that the main directions of data spread are in the directions of the eigenvectors of **C**. Furthermore, if the variables are uncorrelated or statistically independent, the covariance matrix will be diagonal and the eigenvectors will be in the same direction as the coordinate axes x_1 and x_2 (and the ellipse shown would be oriented along the x_1 - and x_2-axis). If, the variances along the main diagonal are equal, the density would be symmetrical in all directions (in the form of a bell) and Part (b) would be a circle. Note in Parts (a) and (b) that the density is centered at the mean values (m_1, m_2).



Multivariate Gaussian Distribution

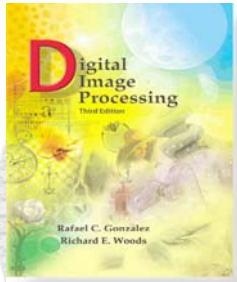




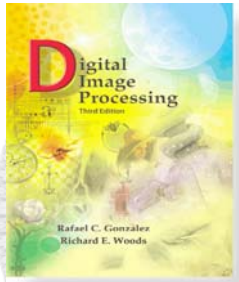


Linear Transformations of RV's

As discussed in the *Review of Matrices and Vectors*, a linear transformation of a vector **x** to produce a vector **y** is of the form **y** = **Ax**. Of particular importance in our work is the case when the rows of **A** are the eigenvectors of the covariance matrix. Because **C** is real and symmetric, we know from the discussion in the *Review of Matrices and Vectors* that it is always possible to find *n* orthonormal eigenvectors from which to form **A**. The implications of this are discussed in considerable detail at the end of the *Review of Matrices and Vectors*, which we recommend should be read again as a conclusion to the present discussion.



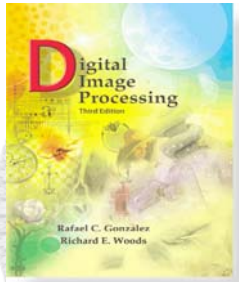
EECS490: Digital Image Processing



Basic Matrices

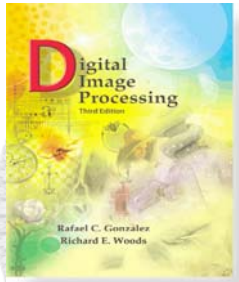
An $m \times n$ (read "m by n") ***matrix***, denoted by **A**, is a rectangular array of entries or elements (numbers, or symbols representing numbers) enclosed typically by square brackets, where m is the number of rows and n the number of columns.

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}$$



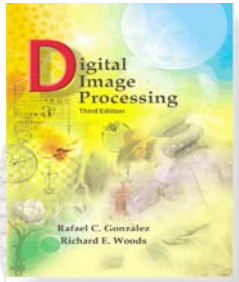
Basic Matrices

- **A** is ***square*** if $m = n$.
- **A** is ***diagonal*** if all off-diagonal elements are 0, and not all diagonal elements are 0.
- **A** is the ***identity matrix*** (**I**) if it is diagonal and all diagonal elements are 1.
- **A** is the ***zero*** or ***null matrix*** (**0**) if all its elements are 0.
- The ***trace*** of **A** equals the sum of the elements along its main diagonal.
- Two matrices **A** and **B** are ***equal*** iff the have the same number of rows and columns, and $a_{ij} = b_{ij}$.



Basic Matrices

- The ***transpose*** $\mathbf{A}^T$ of an $m \times n$ matrix $\mathbf{A}$ is an $n \times m$ matrix obtained by interchanging the rows and columns of $\mathbf{A}$.
- A square matrix for which $\mathbf{A}^T = \mathbf{A}$ is said to be ***symmetric***.
- Any matrix X for which $\mathbf{XA} = \mathbf{I}$ and $\mathbf{AX} = \mathbf{I}$ is called the ***inverse*** of $\mathbf{A}$.
- Let c be a real or complex number (called a ***scalar***). The ***scalar multiple*** of c and matrix $\mathbf{A}$, denoted $c\mathbf{A}$, is obtained by multiplying every elements of $\mathbf{A}$ by c. If $c = -1$, the scalar multiple is called the ***negative*** of $\mathbf{A}$.



Basic Matrices

A ***column vector*** is an $m \times 1$ matrix:

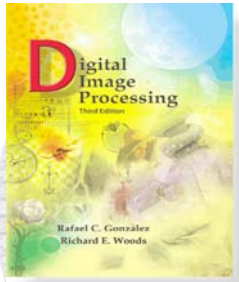
$$\mathbf{a} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_m \end{bmatrix}$$

A ***row vector*** is a $1 \times n$ matrix:

$$\mathbf{b} = [b_1, b_2, \cdots b_n]$$

A column vector can be expressed as a row vector by using the transpose:

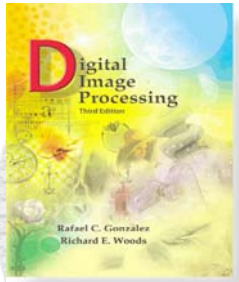
$$\mathbf{a}^T = [a_1, a_2, \cdots, a_m]$$



Basic Matrices

- The ***sum*** of two matrices **A** and **B** (of equal dimension), denoted **A** + **B**, is the matrix with elements $a_{ij} + b_{ij}$.
- The ***difference*** of two matrices, **A**– **B**, has elements $a_{ij} - b_{ij}$.
- The ***product***, **AB**, of $m \times n$ matrix **A** and $p \times q$ matrix **B**, is an $m \times q$ matrix **C** whose (i,j)-th element is formed by multiplying the entries across the ith row of **A** times the entries down the jth column of **B**; that is,

$$c_{ij} = a_{i1}b_{1j} + a_{i2}b_{2j} + \cdots + a_{in}b_{pj}$$



Basic Matrices

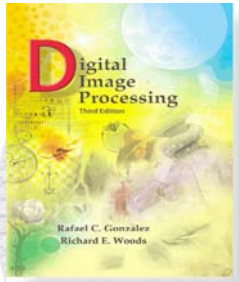
The ***inner product*** (also called ***dot product***) of two vectors

$$\mathbf{a} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_m \end{bmatrix} \qquad \mathbf{b} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix}$$

is defined as

$$\mathbf{a}^T\mathbf{b} = \mathbf{b}^T\mathbf{a} = a_1b_1 + a_2b_2 + \cdots + a_mb_m = \sum_{i=1}^{m} a_ib_i.$$

Note that the inner product is a scalar.

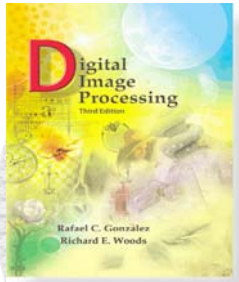


Vectors & Vector Spaces

A ***vector space*** is defined as a nonempty set *V* of entities called *vectors* and associated scalars that satisfy the conditions outlined in A through C below. A vector space is *real* if the scalars are real numbers; it is *complex* if the scalars are complex numbers.

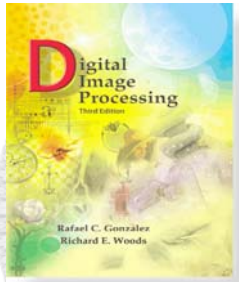
- Condition A: There is in *V* an operation called *vector addition*, denoted **x** + **y**, that satisfies:

  1. **x** + **y** = **y** + **x** for all vectors **x** and **y** in the space.
  2. **x** + (**y** + **z**) = (**x** + **y**) + **z** for all **x**, **y**, and **z**.
  3. There exists in *V* a unique vector, called the *zero vector*, and denoted **0**, such that **x** + **0** = **x** and **0** + **x** = **x** for all vectors **x**.
  4. For each vector **x** in *V*, there is a unique vector in *V*, called the *negation* of **x**, and denoted **–x**, such that **x** + (**– x**) = **0** and (**– x**) + **x** = **0**.



Vectors & Vector Spaces

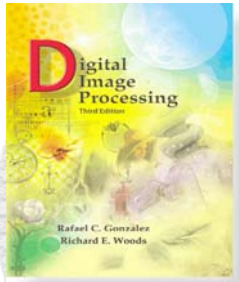
- Condition B: There is in V an operation called *multiplication by a scalar* that associates with each scalar c and each vector $\mathbf{x}$ in V a unique vector called the *product* of c and $\mathbf{x}$, denoted by $c\mathbf{x}$ and $\mathbf{x}c$, and which satisfies:
  1. $c(d\mathbf{x}) = (cd)\mathbf{x}$ for all scalars c and d, and all vectors $\mathbf{x}$.
  2. $(c + d)\mathbf{x} = c\mathbf{x} + d\mathbf{x}$ for all scalars c and d, and all vectors $\mathbf{x}$.
  3. $c(\mathbf{x} + \mathbf{y}) = c\mathbf{x} + c\mathbf{y}$ for all scalars c and all vectors $\mathbf{x}$ and $\mathbf{y}$.
- Condition C: $1\mathbf{x} = \mathbf{x}$ for all vectors $\mathbf{x}$.



Vectors & Vector Spaces

We are interested particularly in real vector spaces of real $m\times 1$ column matrices. We denote such spaces by $\Re^m$, with vector addition and multiplication by scalars being as defined earlier for matrices. Vectors (column matrices) in $\Re^m$ are written as

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{bmatrix}$$



Vectors & Vector Spaces

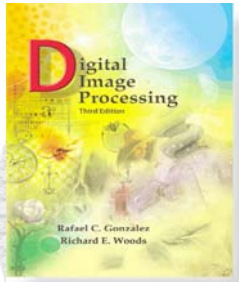
Example

The vector space with which we are most familiar is the two-dimensional real vector space $\Re^2$, in which we make frequent use of graphical representations for operations such as vector addition, subtraction, and multiplication by a scalar. For instance, consider the two vectors

$$\mathbf{a} = \begin{bmatrix} 2 \\ 1 \end{bmatrix} \qquad \mathbf{b} = \begin{bmatrix} 1 \\ 3 \end{bmatrix}$$

Using the rules of matrix addition and subtraction we have

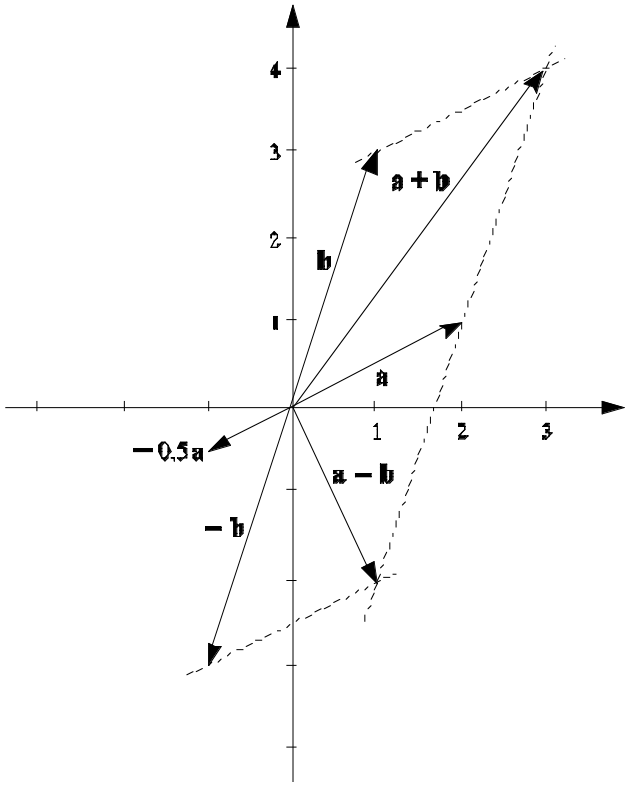
$$\mathbf{a} + \mathbf{b} = \begin{bmatrix} 3 \\ 4 \end{bmatrix} \qquad \mathbf{a} - \mathbf{b} = \begin{bmatrix} 1 \\ -2 \end{bmatrix}$$

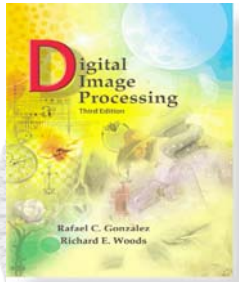


Vectors & Vector Spaces

Example (Con't)

The following figure shows the familiar graphical representation of the preceding vector operations, as well as multiplication of vector **a** by scalar $c = -0.5$.





Vectors & Vector Spaces

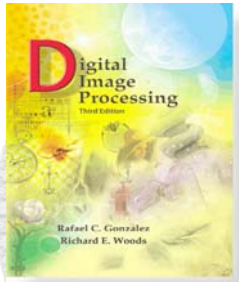
Consider two real vector spaces V_0 and V such that:

- Each element of V_0 is also an element of V (i.e., V_0 is a *subset* of V).
- Operations on elements of V_0 are the same as on elements of V. Under these conditions, V_0 is said to be a ***subspace*** of V.

A ***linear combination*** of $\mathbf{v}_1,\mathbf{v}_2,\ldots,\mathbf{v}_n$ is an expression of the form

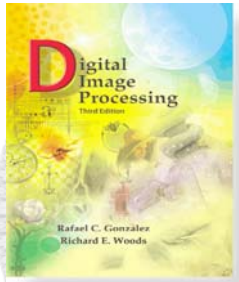
$$\alpha_1\mathbf{v}_1 + \alpha_2\mathbf{v}_2 + \cdots + \alpha_n\mathbf{v}_n$$

where the α's are scalars.



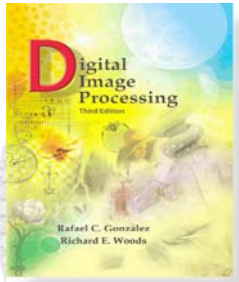
Vectors & Vector Spaces

A vector $\mathbf{v}$ is said to be ***linearly dependent*** on a set, S, of vectors $\mathbf{v}_1,\mathbf{v}_2,\ldots,\mathbf{v}_n$ if and only if $\mathbf{v}$ can be written as a linear combination of these vectors. Otherwise, $\mathbf{v}$ is ***linearly independent*** of the set of vectors $\mathbf{v}_1,\mathbf{v}_2,\ldots,\mathbf{v}_n$.



Vectors & Vector Spaces

A set S of vectors $\mathbf{v}_1,\mathbf{v}_2,\ldots,\mathbf{v}_n$ in V is said to ***span*** some subspace V_0 of V if and only if S is a subset of V_0 and every vector $\mathbf{v}_0$ in V_0 is linearly dependent on the vectors in S. The set S is said to be a ***spanning set*** for V_0. A ***basis*** for a vector space V is a linearly independent spanning set for V. The number of vectors in the basis for a vector space is called the ***dimension*** of the vector space. If, for example, the number of vectors in the basis is n, we say that the vector space is n-dimensional.



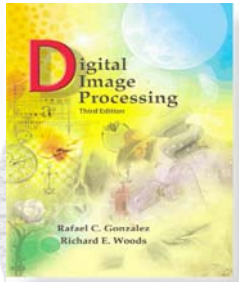
Vectors & Vector Spaces

An important aspect of the concepts just discussed lies in the representation of any vector in $\Re^m$ as a ***linear combination*** of the basis vectors. For example, any vector

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$$

in $\Re^3$ can be represented as a linear combination of the basis vectors

$$\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \quad \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \quad \text{and} \quad \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$



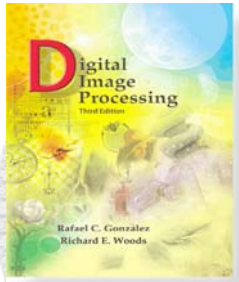
Vector Norms

A ***vector norm*** on a vector space V is a function that assigns to each vector $\mathbf{v}$ in V a nonnegative real number, called the ***norm*** of $\mathbf{v}$, denoted by $\|\mathbf{v}\|$. By definition, the norm satisfies the following conditions:

(1) $\|\mathbf{v}\| > 0$ for $\mathbf{v} \neq \mathbf{0}$; $\|\mathbf{0}\| = 0$,

(2) $\|c\mathbf{v}\| = |c|\|\mathbf{v}\|$ for all scalars c and vectors $\mathbf{v}$, and

(3) $\|\mathbf{u} + \mathbf{v}\| \leq \|\mathbf{u}\| + \|\mathbf{v}\|$.



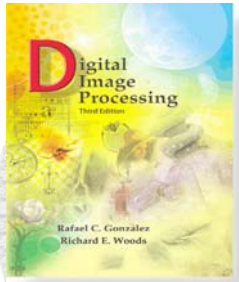
Vector Norms

There are numerous norms that are used in practice. In our work, the norm most often used is the so-called **2-*norm***, which, for a vector $\mathbf{x}$ in real $\Re^m$, space is defined as

$$\|\mathbf{x}\| = [x_1^2 + x_2^2 + \cdots + x_m^2]^{1/2}$$

which is recognized as the *Euclidean distance* from the origin to point $\mathbf{x}$; this gives the expression the familiar name Euclidean norm. The expression also is recognized as the length of a vector $\mathbf{x}$, with origin at point $\mathbf{0}$. From earlier discussions, the norm also can be written as

$$\|\mathbf{x}\| = [\mathbf{x}^T\mathbf{x}]^{1/2}$$



Cauchy-Schwartz Inequality

The ***Cauchy-Schwartz*** inequality states that

$$|\mathbf{x}^T\mathbf{y}| \leq \|\mathbf{x}\|\,\|\mathbf{y}\|$$

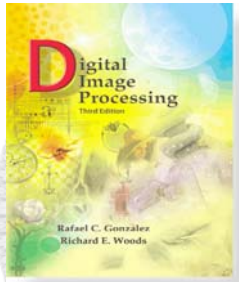
Another well-known result used in the book is the expression

$$\cos\theta = \frac{\mathbf{x}^T\mathbf{y}}{\|\mathbf{x}\|\,\|\mathbf{y}\|}$$

where θ is the angle between vectors **x** and **y**. From these expressions it follows that the inner product of two vectors can be written as

$$\mathbf{x}^T\mathbf{y} = \|\mathbf{x}\|\,\|\mathbf{y}\|\cos\theta$$

Thus, the inner product can be expressed as a function of the norms of the vectors and the angle between the vectors.

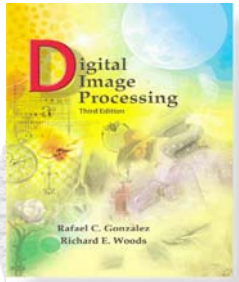


Orthogonality

From the preceding results, two vectors in $\Re^m$ are ***orthogonal*** if and only if their inner product is zero. Two vectors are ***orthonormal*** if, in addition to being orthogonal, the length of each vector is 1.

From the concepts just discussed, we see that an arbitrary vector **a** is turned into a vector $\mathbf{a}_n$ of unit length by performing the operation $\mathbf{a}_n = \mathbf{a}/\|\mathbf{a}\|$. Clearly, then, $\|\mathbf{a}_n\| = 1$.

A ***set of vectors*** is said to be an ***orthogonal*** set if every two vectors in the set are orthogonal. A ***set of vectors*** is ***orthonormal*** if every two vectors in the set are orthonormal.



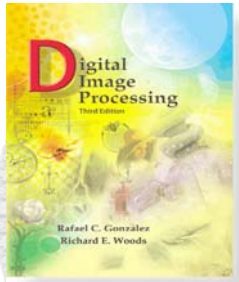
Orthogonality

Let $B = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\}$ be an orthogonal or orthonormal basis in the sense defined in the previous section. Then, an important result in vector analysis is that any vector v can be represented with respect to the orthogonal basis B as

$$\mathbf{v} = \alpha_1 \mathbf{v}_1 + \alpha_2 \mathbf{v}_2 + \cdots + \alpha_n \mathbf{v}_n$$

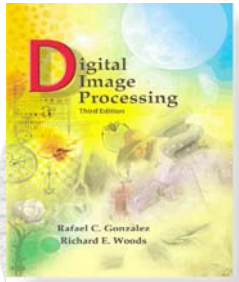
where the coefficients are given by

$$\alpha_i = \frac{\mathbf{v}^T \mathbf{v}_i}{\mathbf{v}_i^T \mathbf{v}_i} = \frac{\mathbf{v}^T \mathbf{v}_i}{\|\mathbf{v}_i\|^2}$$



Orthogonality

The key importance of this result is that, if we represent a vector as a linear combination of orthogonal or orthonormal basis vectors, we can determine the coefficients directly from simple inner product computations. It is possible to convert a linearly dependent spanning set of vectors into an orthogonal spanning set by using the well-known ***Gram-Schmidt*** process. There are numerous programs available that implement the Gram-Schmidt and similar processes, so we will not dwell on the details here.



Discrete Image Transforms*

Definition: The ***linear transform*** of a real matrix **x** is given by

$$\mathbf{y} = \mathbf{T}\mathbf{x}.$$

where

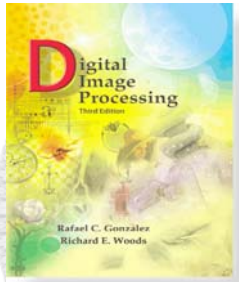
$$y_i = \sum_{j=0}^{N-1} t_{ij} x_j$$

If **T** is unitary then $\mathbf{T}^{-1} = \mathbf{T}^{*\mathrm{T}}$

If **T** is unitary and real then $\mathbf{T}^{-1} = \mathbf{T}^{\mathrm{T}}$

Many important transforms are unitary or unitary and real.

*See for example, Castleman, Digital Image Processing, 2/e, Ch. 13



Orthogonal Transforms

Many transforms used in image processing have only real elements in their kernel matrix

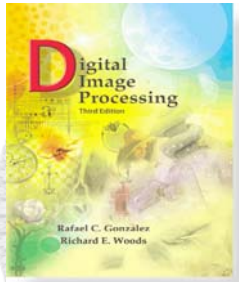
$$G = F\Im$$

where

$$G_{mn} = \sum_{i=0}^{N-1}\sum_{k=0}^{N-1} F_{ik}\Im(i,k,m,n)$$

and $\Im$ represents the kernel function of the transform.

For example, this can be used to represent a Fourier or many other image transforms.



Orthogonal Transforms

Many such image transforms are separable and can be carried out as a rowwise operation following by a column wise operation, or vice versa, i.e.

$$\Im(i,k,m,n) = T_r(i,m)T_c(k,n)$$

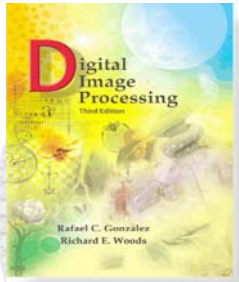
or

$$G_{mn} = \sum_{i=0}^{N-1} T(i,m)\left[\sum_{k=0}^{N-1} F_{ik}T(k,n)\right]$$

This can be written as $G = TFT$

whcih can be simply inverted (inverse transformed) as

$$F = T^{-1}FT^{-1} = T^{*T}GT^{*T}$$



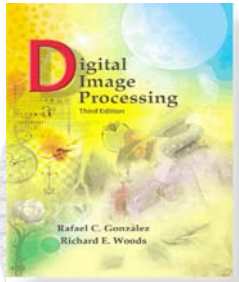
Orthogonal Transforms

The rows of the kernel matrix for a set of basis vectors in a N-dimensional vector space and are orthonormal

$$TT^{*T} = I$$

or

$$\sum_{i=0}^{N-1} T_{ji} T_{ki}^{*} = \delta_{ik}$$



Eigenvalues & Eigenvectors

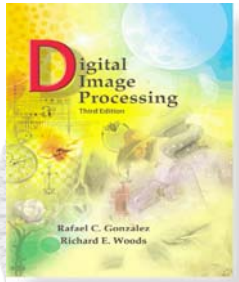
Definition: The ***eigenvalues*** of a real matrix **M** are the real numbers λ for which there is a nonzero vector **e** such that

$$\mathbf{Me} = \lambda\,\mathbf{e}.$$

The ***eigenvectors*** of **M** are the nonzero vectors **e** for which there is a real number λ such that $\mathbf{Me} = \lambda\,\mathbf{e}$.

If $\mathbf{Me} = \lambda\,\mathbf{e}$ for $\mathbf{e} \neq 0$, then **e** is an ***eigenvector*** of **M** associated with ***eigenvalue*** λ, and vice versa. The eigenvectors and corresponding eigenvalues of **M** constitute the ***eigensystem*** of **M**.

Numerous theoretical and truly practical results in the application of matrices and vectors stem from this beautifully simple definition.



Eigenvalues & Eigenvectors

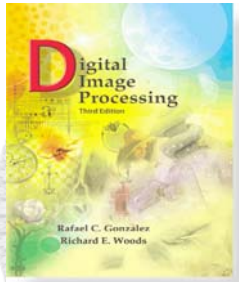
Example: Consider the matrix

$$\mathbf{M} = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}$$

It is easy to verify that $\mathbf{M}\mathbf{e}_1 = \lambda_1\mathbf{e}_1$ and $\mathbf{M}\mathbf{e}_2 = \lambda_2\mathbf{e}_2$ for $\lambda_1 = 1$, $\lambda_2 = 2$ and

$$\mathbf{e}_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \quad \text{and} \quad \mathbf{e}_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

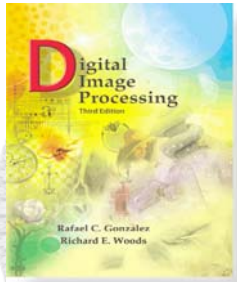
In other words, $\mathbf{e}_1$ is an eigenvector of $\mathbf{M}$ with associated eigenvalue λ_1, and similarly for $\mathbf{e}_2$ and λ_2.



Eigenvalues & Eigenvectors

The following properties, which we give without proof, are essential background in the use of vectors and matrices in digital image processing. In each case, we assume a real matrix of order $m \times m$ although, as stated earlier, these results are equally applicable to complex numbers.

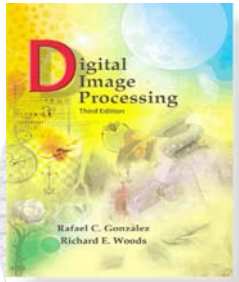
1. If $\{\lambda_1, \lambda_2, \ldots, \lambda_q, q \leq m$, is set of distinct eigenvalues of $\mathbf{M}$, and $\mathbf{e}_i$ is an eigenvector of $\mathbf{M}$ with corresponding eigenvalue λ_i, $i = 1,2,\ldots,q$, then $\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_q\}$ is a linearly independent set of vectors. An important implication of this property: If an $m \times m$ matrix $\mathbf{M}$ has m distinct eigenvalues, its eigenvectors will constitute an orthogonal (orthonormal) set, which means that any m-dimensional vector can be expressed as a linear combination of the eigenvectors of $\mathbf{M}$.



Eigenvalues & Eigenvectors

2. The numbers along the main diagonal of a diagonal matrix are equal to its eigenvalues. It is not difficult to show using the definition $\mathbf{Me} = \lambda\,\mathbf{e}$ that the eigenvectors can be written by inspection when **M** is diagonal.

3. A real, symmetric $m \times m$ matrix **M** has a set of m linearly independent eigenvectors that may be chosen to form an orthonormal set. This property is of particular importance when dealing with covariance matrices (e.g., see Section 11.4 and our review of probability) which are real and symmetric.

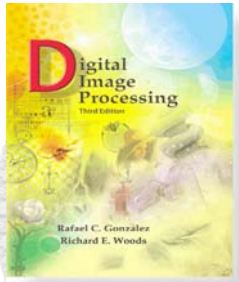


Eigenvalues & Eigenvectors

4. A corollary of Property 3 is that the eigenvalues of an $m \times m$ real symmetric matrix are real, and the associated eigenvectors may be chosen to form an orthonormal set of m vectors.

5. Suppose that **M** is a real, symmetric $m \times m$ matrix, and that we form a matrix **A** whose rows are the m orthonormal eigenvectors of **M**. Then, the product $\mathbf{A}\mathbf{A}^T = \mathrm{I}$ because the rows of **A** are orthonormal vectors. Thus, we see that $\mathbf{A}^{-1} = \mathbf{A}^T$ when matrix **A** is formed in the manner just described.

6. Consider matrices **M** and **A** in 5. The product $\mathbf{D} = \mathbf{A}\mathbf{M}\mathbf{A}^{-1} = \mathbf{A}\mathbf{M}\mathbf{A}^T$ is a diagonal matrix whose elements along the main diagonal are the eigenvalues of **M**. The eigenvectors of **D** are the same as the eigenvectors of **M**.



Eigenvalues & Eigenvectors

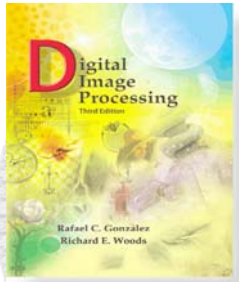
Example

Suppose that we have a random population of vectors, denoted by $\{\mathbf{x}\}$, with covariance matrix (see the review of probability):

$$\mathbf{C}_\mathbf{x} = E\{(\mathbf{x} - \mathbf{m}_\mathbf{x})(\mathbf{x} - \mathbf{m}_\mathbf{x})^T\}$$

Suppose that we perform a transformation of the form $\mathbf{y} = \mathbf{A}\mathbf{x}$ on each vector $\mathbf{x}$, where the rows of $\mathbf{A}$ are the orthonormal eigenvectors of $\mathbf{C}_\mathbf{x}$. The covariance matrix of the population $\{\mathbf{y}\}$ is

$$\begin{aligned}
\mathbf{C}_\mathbf{y} &= E\{(\mathbf{y} - \mathbf{m}_\mathbf{y})(\mathbf{y} - \mathbf{m}_\mathbf{y})^T\} \\
&= E\{(\mathbf{A}\mathbf{x} - \mathbf{A}\mathbf{m}_\mathbf{x})(\mathbf{A}\mathbf{x} - \mathbf{A}\mathbf{m}_\mathbf{x})^T\} \\
&= E\{\mathbf{A}(\mathbf{x} - \mathbf{m}_\mathbf{x})(\mathbf{x} - \mathbf{m}_\mathbf{x})^T\mathbf{A}^T\} \\
&= \mathbf{A}E\{(\mathbf{x} - \mathbf{m}_\mathbf{x})(\mathbf{x} - \mathbf{m}_\mathbf{x})^T\}\mathbf{A}^T \\
&= \mathbf{A}\mathbf{C}_\mathbf{x}\mathbf{A}^T
\end{aligned}$$



Eigenvalues & Eigenvectors

$$\underline{C}_x = E\left\{\left(\underline{x} - \underline{m}_x\right)\left(\underline{x} - \underline{m}_x\right)^T\right\}$$

Perform a transformation of the form $\mathbf{y} = \mathbf{Ax}$ on each vector $\mathbf{x}$, where the rows of $\mathbf{A}$ are the orthonormal eigenvectors of $\mathbf{C_x}$. The covariance matrix of the population $\{\mathbf{y}\}$ is

C_y is non-zero only along the diagonals so we have decoupled the data.

$$\underline{C}_y = E\left\{\left(\underline{y} - \underline{m}_y\right)\left(\underline{y} - \underline{m}_y\right)^T\right\}$$

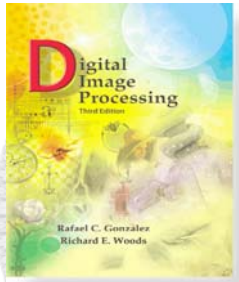
$$\underline{C}_y = E\left\{\left(\underline{A}\underline{x} - \underline{A}\underline{m}_x\right)\left(\underline{A}\underline{x} - \underline{A}\underline{m}_x\right)^T\right\}$$

$$\underline{C}_y = E\left\{\underline{A}\left(\underline{x} - \underline{m}_x\right)\left(\underline{x} - \underline{m}_x\right)^T A^T\right\}$$

$$\underline{C}_y = \underline{A}E\left\{\left(\underline{x} - \underline{m}_x\right)\left(\underline{x} - \underline{m}_x\right)^T\right\}A^T$$

$$\underline{C}_y = \underline{A}\underline{C}_x A^T$$

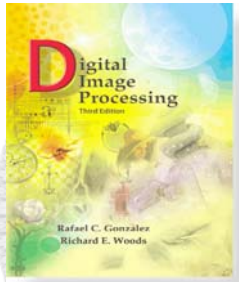
$$\underline{A} = \begin{bmatrix} \underline{e}_1 \\ \underline{e}_2 \\ \vdots \\ \underline{e}_m \end{bmatrix}$$



Eigenvalues & Eigenvectors

From Property 6, we know that $\mathbf{C_y}=\mathbf{AC_xA}^T$ is a diagonal matrix with the eigenvalues of $\mathbf{C_x}$ along its main diagonal. The elements along the main diagonal of a covariance matrix are the variances of the components of the vectors in the population. The off diagonal elements are the covariances of the components of these vectors.

The fact that $\mathbf{C_y}$ is diagonal means that the elements of the vectors in the population $\{\mathbf{y}\}$ are ***uncorrelated*** (their covariances are 0). Thus, we see that application of the linear transformation $\mathbf{y} = \mathbf{Ax}$ involving the eigenvectors of $\mathbf{C_x}$ decorrelates the data, and the elements of $\mathbf{C_y}$ along its main diagonal give the variances of the components of the $\mathbf{y}$'s along the eigenvectors. Basically, what has

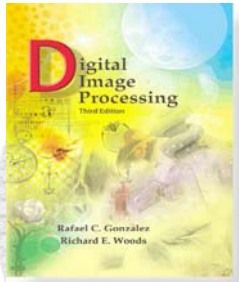


Eigenvalues & Eigenvectors

been accomplished here is a coordinate transformation that aligns the data along the eigenvectors of the covariance matrix of the population.

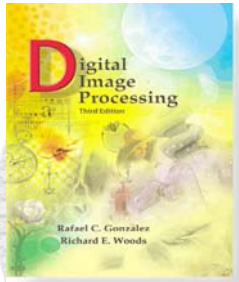
The preceding concepts are illustrated in the following figure. Part (a) shows a data population $\{\mathbf{x}\}$ in two dimensions, along with the eigenvectors of $\mathbf{C}_\mathbf{x}$ (the black dot is the mean). The result of performing the transformation $\mathbf{y}=\mathbf{A}(\mathbf{x}-\mathbf{m}_\mathbf{x})$ on the $\mathbf{x}$'s is shown in Part (b) of the figure.

The fact that we subtracted the mean from the $\mathbf{x}$'s caused the $\mathbf{y}$'s to have zero mean, so the population is centered on the coordinate system of the transformed data. It is important to note that all we have done here is make the eigenvectors the

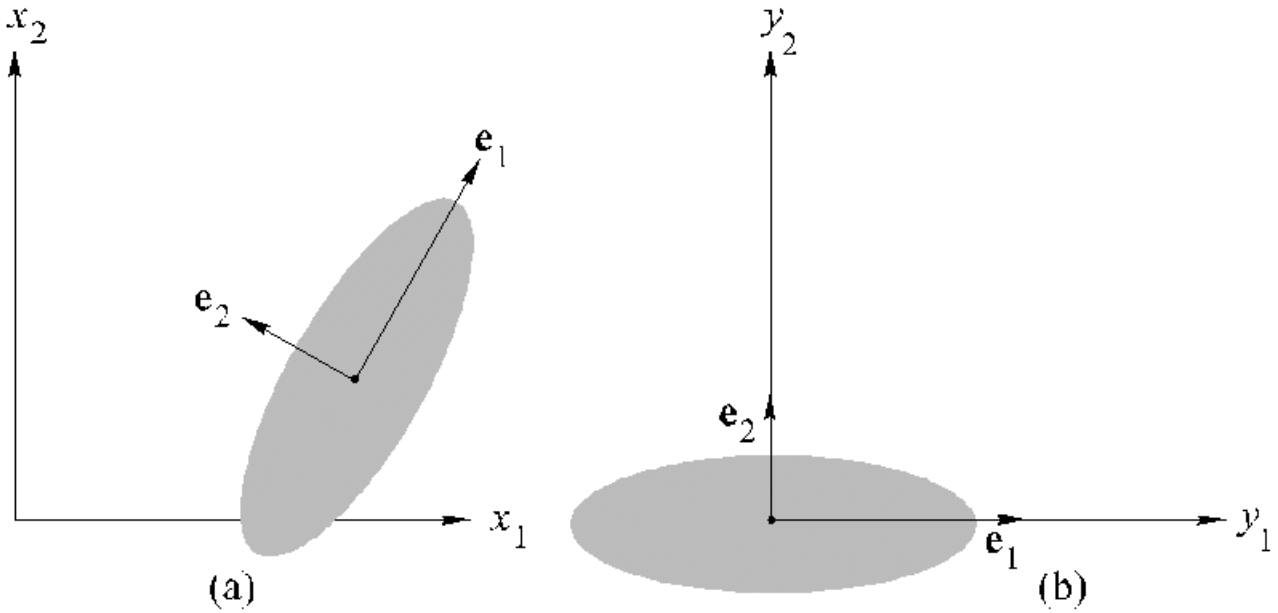


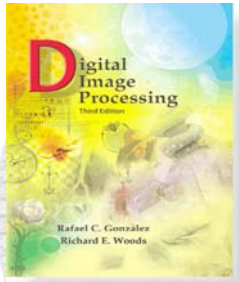
Eigenvalues & Eigenvectors

new coordinate system (y_1,y_2). Because the covariance matrix of the **y**'s is diagonal, this in fact also decorrelated the data. The fact that the main data spread is along $\mathbf{e}_1$ is due to the fact that the rows of the transformation matrix **A** were chosen according the order of the eigenvalues, with the first row being the eigenvector corresponding to the largest eigenvalue.



Eigenvalues & Eigenvectors





Karhunen-Loève Expansion

$$\underline{y} = \underline{A}\underline{x}$$

where the rows of **A** are the orthonormal eigenvectors of $\mathbf{C_x}$. Drop the upper M-N rows of A to give B

$$\underline{B} = \begin{bmatrix} \underline{e}_1 \\ \underline{e}_2 \\ \vdots \\ \underline{e}_M \end{bmatrix}$$

Then $$\hat{\underline{y}} = \underline{B}\underline{x}$$

Inverting $$\hat{\underline{x}} = \underline{B}^T \hat{\underline{y}}$$

This approximates x using only M components with a squared error given by

$$MSE = \sum_{k=M+1}^{N} \lambda_k$$