


Pattern Classiflcation (2nd ed.)

Richard Duda

Related papers

[Download a PDF Pack](#) of the best related papers 



[Pattern Classification \(2nd ed](#)

[Richard O. Duda, Peter E. Hart, David G. Stork Pattern classification Wiley \(2001\)](#)

Gourav Panwar

[Duda-et al-00](#)

maede abbasi

Pattern Classification (2nd ed.)

Richard O. Duda, Peter E. Hart and David G. Stork
September 3, 1997

NOT FOR GENERAL DISTRIBUTION; for use only by students of designated faculty. This is a pre-publication print of material to appear in Duda, Hart and Stork: **Pattern Classification and Scene Analysis: Part I Pattern Classification**, to be published in 1998 by John Wiley & Sons, Inc. This is a preliminary version and may contain errors; comments and suggestions are heartily encouraged.

Contact: Dr. David G. Stork
Ricoh California Research Center
2882 Sand Hill Road, Suite 115
Menlo Park, CA 94025-7022 USA
stork@crc.ricoh.com

©1997 R. O. Duda, P. E. Hart and D. G. Stork
All rights reserved.

Contents

A.1	Notation	5
A.2	Linear algebra	8
A.2.1	Notation and preliminaries	8
A.2.2	Outer product	9
A.2.3	Derivatives of matrices	10
A.2.4	Determinant and trace	11
A.2.5	Eigenvectors and eigenvalues	12
A.2.6	Matrix inversion	12
A.3	Lagrange optimization	13
A.4	Probability Theory	13
A.4.1	Discrete random variables	13
A.4.2	Expected values	14
A.4.3	Pairs of discrete random variables	15
A.4.4	Statistical independence	16
A.4.5	Expected values of functions of two variables	16
A.4.6	Conditional probability	18
A.4.7	The Law of Total Probability and Bayes' rule	18
A.4.8	Vector random variables	19
A.4.9	Expectations, mean vectors and covariance matrices	20
A.4.10	Continuous random variables	21
A.4.11	Distributions of sums of independent random variables	23
A.4.12	Univariate normal density	24
A.5	Gaussian derivatives and integrals	25
A.5.1	Multivariate normal densities	27
A.5.2	Bivariate normal densities	28
A.6	Information theory	31
A.6.1	Entropy and information	31
A.6.2	Relative entropy	32
A.6.3	Mutual information	32
A.7	Computational complexity	33
	Bibliography	35
	Index	38

Mathematical foundations

Our goal here is to present the basic results and definitions from linear algebra, probability theory, information theory and computational complexity that serve as the mathematical foundations for the pattern recognition techniques discussed throughout this book. We will try to give intuition whenever appropriate, but we do not attempt to prove these results; systematic expositions can be found in the references.

A.1 Notation

Here are the terms and notation used throughout the book. In addition, there are numerous specialized variables and functions whose definitions are usage should be clear from the text.

variables, symbols and operations

\simeq	approximately equal to
\approx	approximately equal to (in an expansion)
\equiv	equivalent to (or defined to be)
\propto	proportional to
∞	infinity
$x \rightarrow a$	x approaches a
$t \leftarrow t + 1$	in an algorithm: assign to variable t the new value $t + 1$
$\lim_{x \rightarrow a} f(x)$	the value of $f(x)$ in the limit x approaching a
$\arg \max_x f(x)$	the value of x that leads to the maximum value of $f(x)$
$\arg \min_x f(x)$	the value of x that leads to the minimum value of $f(x)$
$\ln(x)$	logarithm base e , or natural logarithm of x
$\log(x)$	logarithm base 10 of x
$\log_2(x)$	logarithm base 2 of x
$\exp[x]$ or e^x	exponential of x
$\partial f(x) / \partial x$	partial derivative
$\int_a^b f(x) dx$	the integral of $f(x)$ between a and b . If no limits are written, the full space is assumed
■	Q.E.D., quod erat demonstrandum (“which was to be proved”) — used to signal the end of a proof

mathematical operations

$\mathcal{E}[f(x)]$	the expected value of function $f(x)$
$\mathcal{E}_y[f(x, y)]$	the expected value of function over several variables, $f(x, y)$, taken over a subset y of them
$\text{Var}_f[\cdot]$	$\mathcal{E}_f[(x - \mathcal{E}[x])^2]$
$\langle x \rangle$	expected value of random variable
$\sum_{i=1}^n a_i$	the sum from $i = 1$ to n : $a_1 + a_2 + \dots + a_n$
$\prod_{i=1}^n a_i$	the product from $i = 1$ to n : $a_1 \times a_2 \times \dots \times a_n$

vectors and matrices

\mathbf{R}^d	d -dimensional Euclidean space
\mathbf{x}, Σ	boldface for (column) vectors and matrices
\mathbf{I}	identity matrix, square matrix having 1s on the diagonal and 0 everywhere else
$\text{diag}(a_1, a_2, \dots, a_d)$	matrix whose diagonal elements are a_1, a_2, \dots, a_d , and off-diagonal elements zero
\mathbf{x}^t	the transpose of vector \mathbf{x}
$\ \mathbf{x}\ $	the Euclidean norm of vector \mathbf{x} .
Σ	covariance matrix
$\text{tr}[\mathbf{A}]$	the transpose of \mathbf{A} , with ij entry changed to ji
\mathbf{A}^{-1}	the inverse of matrix \mathbf{A}
\mathbf{A}^\dagger	pseudoinverse of matrix \mathbf{A}
$ \mathbf{A} $ or $\text{Det}[\mathbf{A}]$	determinant of \mathbf{A}
λ	eigenvalue
\mathbf{e}	eigenvector
\mathbf{e}_i	unit vector in the i direction in Euclidean space

probability and distributions

ω	state of nature
$P(\cdot)$	probability
$p(\cdot)$	probability density
$P(a, b)$	the joint probability, i.e., of having both a and b
$p(a, b)$	the joint probability density, i.e., of having both a and b
$p(\mathbf{x} \boldsymbol{\theta})$	the conditional probability density of \mathbf{x} given that $\boldsymbol{\theta}$
$F(x; \theta)$	function of x , with implied (nonexplicit) dependence upon θ
\mathbf{w}	weight
$\lambda(\cdot, \cdot)$	loss function
$\Delta = \begin{pmatrix} \frac{d}{dx_1} \\ \frac{d}{dx_2} \\ \vdots \\ \frac{d}{dx_d} \end{pmatrix}$	gradient operator in \mathbf{R}^d
$\Delta_{\boldsymbol{\theta}} = \begin{pmatrix} \frac{d}{d\theta_1} \\ \frac{d}{d\theta_2} \\ \vdots \\ \frac{d}{d\theta_d} \end{pmatrix}$	gradient operator in $\boldsymbol{\theta}$ coordinates
$\hat{\boldsymbol{\theta}}$	maximum likelihood value of $\boldsymbol{\theta}$
\sim	“has the distribution” e.g., $p(x) \sim N(\mu, \sigma^2)$ means that the density of x is normal, with mean μ and variance σ^2
$N(\mu, \sigma^2)$	normal or Gaussian distribution with mean μ and variance σ^2
$N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$	multidimensional normal or Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$
$U(x_l, x_u)$	a one-dimensional uniform distribution between x_l and x_u .
$U(\mathbf{x}_l, \mathbf{x}_u)$	a d -dimensional uniform density, having the smallest axes-aligned bounding box containing both \mathbf{x}_l and \mathbf{x}_u
$T(\mu, \delta)$	triangle distribution, having center μ and full half-width δ
$\delta(x)$	Dirac delta function
$\Gamma(\cdot)$	Gamma function
$n!$	n factorial = $n \times (n-1) \times (n-2) \times \dots \times 1$
$\binom{a}{b} = \frac{a!}{b!(a-b)!}$	binomial coefficient, a choose b
$O(h(x))$	big oh order of $h(x)$
$\Theta(h(x))$	big theta order of $h(x)$
\bar{x}	mean or average value of x
$\lim_{n \rightarrow y} f(x)$	the value of $f(x)$ in the limit x approaches y
$\sup_x f(x)$	the supremum value of $f(x)$

sets

$\mathcal{A}, \mathcal{B}, \mathcal{C}, \mathcal{D}, \dots$	“Calligraphic” font generally denotes sets or lists, e.g., data set $\mathcal{D} = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$
$\mathbf{x} \in \mathcal{D}$	\mathbf{x} is an element of set \mathcal{D}
$\mathbf{x} \notin \mathcal{D}$	\mathbf{x} is not an element of set \mathcal{D}
$\mathcal{A} \cup \mathcal{B}$	union of two sets, i.e., the set containing all elements of \mathcal{A} and \mathcal{B}
$ \mathcal{D} $	the cardinality of set \mathcal{D} , i.e., the number of (possibly non-distinct) elements in it
$\max_x[\mathcal{D}]$	the x value in set \mathcal{D} that is maximum

A.2 Linear algebra

A.2.1 Notation and preliminaries

A d -dimensional (column) vector \mathbf{x} and its (row) transpose \mathbf{x}^t can be written as

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{pmatrix} \quad \text{and} \quad \mathbf{x}^t = (x_1 \ x_2 \ \dots \ x_d), \quad (1)$$

where here and below, all components take on real values. We denote an $n \times d$ (rectangular) matrix \mathbf{M} and its $d \times n$ transpose \mathbf{M}^t as

$$\mathbf{M} = \begin{pmatrix} m_{11} & m_{12} & m_{13} & \dots & m_{1d} \\ m_{21} & m_{22} & m_{23} & \dots & m_{2d} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ m_{n1} & m_{n2} & m_{n3} & \dots & m_{nd} \end{pmatrix} \quad \text{and} \quad (2)$$

$$\mathbf{M}^t = \begin{pmatrix} m_{11} & m_{21} & \dots & m_{n1} \\ m_{12} & m_{22} & \dots & m_{n2} \\ m_{13} & m_{23} & \dots & m_{n3} \\ \vdots & \vdots & \ddots & \vdots \\ m_{1d} & m_{2d} & \dots & m_{nd} \end{pmatrix}. \quad (3)$$

In other words, the ij^{th} entry of \mathbf{M}^t is the ji^{th} entry of \mathbf{M} .

A square ($d \times d$) matrix is called symmetric if its entries obey $m_{ij} = m_{ji}$; it is called skew-symmetric (or anti-symmetric) if $m_{ij} = -m_{ji}$. An general matrix is called non-negative matrix if $m_{ij} \geq 0$ for all i and j . A particularly important matrix is the *identity matrix*, \mathbf{I} — a $d \times d$ (square) matrix whose diagonal entries are 1’s, and all other entries 0. The *Kronecker delta* function or Kronecker symbol, defined as

IDENTITY
MATRIX

KRONECKER
DELTA

$$\delta_{ij} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{otherwise,} \end{cases} \quad (4)$$

can function as an identity matrix. A general diagonal matrix (i.e., one having 0 for all off diagonal entries) is denoted $\text{diag}(m_{11}, m_{22}, \dots, m_{dd})$, the entries being the successive elements $m_{11}, m_{22}, \dots, m_{dd}$. Addition of vectors and of matrices is component by component.

We can multiply a vector by a matrix, $\mathbf{M}\mathbf{x} = \mathbf{y}$, i.e.,

$$\begin{pmatrix} m_{11} & m_{12} & \dots & m_{1d} \\ m_{21} & m_{22} & \dots & m_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ m_{n1} & m_{n2} & \dots & m_{nd} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad (5)$$

where

$$y_j = \sum_{i=1}^d m_{ji} x_i. \quad (6)$$

Note that if \mathbf{M} is not square, the dimensionality of \mathbf{y} differs from that of \mathbf{x} .

The *inner product* of two vectors having the same dimensionality will be denoted here as $\mathbf{x}^t \mathbf{y}$ and yields a scalar:

INNER
PRODUCT

$$\mathbf{x}^t \mathbf{y} = \sum_{i=1}^d x_i y_i = \mathbf{y}^t \mathbf{x}. \quad (7)$$

It is sometimes also called the *scalar product* or *dot product* and denoted $\mathbf{x} \bullet \mathbf{y}$. The *Euclidean norm* or length of the vector is denoted

EUCLIDEAN
NORM

$$\|\mathbf{x}\| = \sqrt{\mathbf{x}^t \mathbf{x}}; \quad (8)$$

we call a vector “normalized” if $\|\mathbf{x}\| = 1$. The angle between two d -dimensional vectors obeys

$$\cos \theta = \frac{\mathbf{x}^t \mathbf{y}}{\|\mathbf{x}\| \|\mathbf{y}\|}, \quad (9)$$

and thus the inner product is a measure of the colinearity of two vectors — a natural indication of their similarity. In particular, if $\mathbf{x}^t \mathbf{y} = 0$, then the vectors are orthogonal, and if $|\mathbf{x}^t \mathbf{y}| = \|\mathbf{x}\| \|\mathbf{y}\|$, the vectors are colinear. From Eq. 9, we have immediately the Cauchy-Schwarz inequality, which states

$$\|\mathbf{x}^t \mathbf{y}\| \leq \|\mathbf{x}\| \|\mathbf{y}\|. \quad (10)$$

We say a set of vectors $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$ is *linearly independent* if no vector in the set can be written as a linear combination of any of the others. Informally, a set of d linearly independent vectors spans an d -dimensional vector space, i.e., any vector in that space can be written as a linear combination of such spanning vectors.

LINEAR
INDEPEND-
ENCE

A.2.2 Outer product

The outer product (sometimes called *matrix product*) of two column vectors yields a matrix

MATRIX
PRODUCT

$$\mathbf{M} = \mathbf{x} \mathbf{y}^t = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{pmatrix} (y_1 \ y_2 \ \dots \ y_n) = \begin{pmatrix} x_1 y_1 & x_1 y_2 & \dots & x_1 y_n \\ x_2 y_1 & x_2 y_2 & \dots & x_2 y_n \\ \vdots & \vdots & \ddots & \vdots \\ x_d y_1 & x_d y_2 & \dots & x_d y_n \end{pmatrix}, \quad (11)$$

that is, the components of \mathbf{M} are $m_{ij} = x_i y_j$. Of course, if $d \neq n$, then \mathbf{M} is not square. Any matrix that can be written as the product of two vectors as in Eq. 11, is called *separable*.

SEPARABLE

A.2.3 Derivatives of matrices

Suppose $f(\mathbf{x})$ is a scalar function of d variables x_i which we represent as the vector \mathbf{x} . Then the derivative or gradient of f with respect to this parameter vector is computed component by component, i.e.,

$$\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} = \begin{pmatrix} \frac{\partial f(\mathbf{x})}{\partial x_1} \\ \frac{\partial f(\mathbf{x})}{\partial x_2} \\ \vdots \\ \frac{\partial f(\mathbf{x})}{\partial x_d} \end{pmatrix}. \quad (12)$$

If we have an n -dimensional vector valued function \mathbf{f} , of a d -dimensional vector \mathbf{x} , we calculate the derivatives and represent them as the *Jacobian matrix*

JACOBIAN
MATRIX

$$\mathbf{J}(\mathbf{x}) = \frac{\partial \mathbf{f}(\mathbf{x})}{\partial \mathbf{x}} = \begin{pmatrix} \frac{\partial f_1(\mathbf{x})}{\partial x_1} & \cdots & \frac{\partial f_1(\mathbf{x})}{\partial x_d} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_n(\mathbf{x})}{\partial x_1} & \cdots & \frac{\partial f_n(\mathbf{x})}{\partial x_d} \end{pmatrix}. \quad (13)$$

If this matrix is square, its determinant (Sect. A.2.4) is called simply the *Jacobian*.

If the entries of \mathbf{M} depend upon a scalar parameter θ , we can take the derivative of \mathbf{M} component by component, to get another matrix, as

$$\frac{\partial \mathbf{M}}{\partial \theta} = \begin{pmatrix} \frac{\partial m_{11}}{\partial \theta} & \frac{\partial m_{12}}{\partial \theta} & \cdots & \frac{\partial m_{1d}}{\partial \theta} \\ \frac{\partial m_{21}}{\partial \theta} & \frac{\partial m_{22}}{\partial \theta} & \cdots & \frac{\partial m_{2d}}{\partial \theta} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial m_{n1}}{\partial \theta} & \frac{\partial m_{n2}}{\partial \theta} & \cdots & \frac{\partial m_{nd}}{\partial \theta} \end{pmatrix}. \quad (14)$$

In Sect. A.2.6 we shall discuss matrix inversion, but for convenience we give here the derivative of the inverse of a matrix, \mathbf{M}^{-1} :

$$\frac{\partial}{\partial \theta} \mathbf{M}^{-1} = -\mathbf{M}^{-1} \frac{\partial \mathbf{M}}{\partial \theta} \mathbf{M}^{-1}. \quad (15)$$

The following vector derivative identities can be verified by writing out the components:

$$\frac{\partial}{\partial \mathbf{x}} [\mathbf{M}\mathbf{x}] = \mathbf{M} \quad (16)$$

$$\frac{\partial}{\partial \mathbf{x}} [\mathbf{y}^t \mathbf{x}] = \frac{\partial}{\partial \mathbf{x}} [\mathbf{x}^t \mathbf{y}] = \mathbf{y} \quad (17)$$

$$\frac{\partial}{\partial \mathbf{x}} [\mathbf{x}^t \mathbf{M} \mathbf{x}] = [\mathbf{M} + \mathbf{M}^t] \mathbf{x}. \quad (18)$$

In the case where \mathbf{M} is symmetric (as for instance a covariance matrix, cf. Sect. A.4.10), then Eq. 18 simplifies to

$$\frac{\partial}{\partial \mathbf{x}}[\mathbf{x}^t \mathbf{M} \mathbf{x}] = 2\mathbf{M} \mathbf{x}. \quad (19)$$

We use the second derivative of a scalar function $f(\mathbf{x})$ to write a Taylor series (or Taylor expansion) about a point \mathbf{x}_0 :

$$f(\mathbf{x}) = f(\mathbf{x}_0) + \underbrace{\left[\frac{\partial f}{\partial \mathbf{x}} \right]_{\mathbf{x}=\mathbf{x}_0}}_{\mathbf{J}} (\mathbf{x} - \mathbf{x}_0) + \frac{1}{2} (\mathbf{x} - \mathbf{x}_0)^t \underbrace{\left[\frac{\partial^2 f}{\partial \mathbf{x}^2} \right]_{\mathbf{x}=\mathbf{x}_0}}_{\mathbf{H}} (\mathbf{x} - \mathbf{x}_0) + O(\|\mathbf{x}\|^3), \quad (20)$$

where \mathbf{H} is the *Hessian* matrix, the matrix of second-order derivatives of $f(\cdot)$ with respect to the parameters, here evaluated at \mathbf{x}_0 . (We shall return in Sect. A.7 to consider the $O(\cdot)$ notation and the order of a function used in Eq. 20 and below.)

HESSIAN
MATRIX

For a vector valued function we write the first-order expansion in terms of the Jacobian as:

$$\mathbf{f}(\mathbf{x}) = \mathbf{f}(\mathbf{x}_0) + \left[\frac{\partial \mathbf{f}}{\partial \mathbf{x}} \right]_{\mathbf{x}=\mathbf{x}_0}^t (\mathbf{x} - \mathbf{x}_0) + O(\|\mathbf{x}\|^2). \quad (21)$$

A.2.4 Determinant and trace

The determinant of a $d \times d$ (square) matrix is a scalar, denoted $|\mathbf{M}|$. If \mathbf{M} is itself a scalar (i.e., a 1×1 matrix M), then $|M| = M$. If \mathbf{M} is 2×2 , then $|\mathbf{M}| = m_{11}m_{22} - m_{21}m_{12}$. The determinant of a general square matrix can be computed by a method called *expansion by minors*, and this leads to a recursive definition. If \mathbf{M} is our $d \times d$ matrix, we define $\mathbf{M}_{i|j}$ to be the $(d-1) \times (d-1)$ matrix obtained by deleting the i^{th} row and the j^{th} column of \mathbf{M} :

EXPANSION
BY MINORS

$$i \begin{pmatrix} m_{11} & m_{12} & \cdots & \overset{j}{\otimes} & \cdots & \cdots & m_{1d} \\ m_{21} & m_{22} & \cdots & \otimes & \cdots & \cdots & m_{2d} \\ \vdots & \vdots & \ddots & \otimes & \cdots & \cdots & \vdots \\ \vdots & \vdots & \cdots & \otimes & \cdots & \cdots & \vdots \\ \otimes & \otimes & \otimes & \otimes & \otimes & \otimes & \otimes \\ \vdots & \vdots & \cdots & \otimes & \cdots & \ddots & \vdots \\ m_{d1} & m_{d2} & \cdots & \otimes & \cdots & \cdots & m_{dd} \end{pmatrix} = \mathbf{M}_{i|j}. \quad (22)$$

Given this definition, we can now compute the determinant of \mathbf{M} the expansion by minors on the first column giving

$$|\mathbf{M}| = m_{11}|\mathbf{M}_{1|1}| - m_{21}|\mathbf{M}_{2|1}| + m_{31}|\mathbf{M}_{3|1}| - \cdots \pm m_{d1}|\mathbf{M}_{d|1}|, \quad (23)$$

where the signs alternate. This process can be applied recursively to the successive (smaller) matrixes in Eq. 23.

For a 3×3 matrix, this determinant calculation can be represented by “sweeping” the matrix — taking the sum of the products of matrix terms along a diagonal, where products from upper-left to lower-right are added with a positive sign, and those from the lower-left to upper-right with a minus sign. That is,

$$\begin{aligned}
|\mathbf{M}| &= \begin{vmatrix} m_{11} & m_{12} & m_{13} \\ m_{21} & m_{22} & m_{23} \\ m_{31} & m_{32} & m_{33} \end{vmatrix} \\
&= m_{11}m_{22}m_{33} + m_{13}m_{21}m_{32} + m_{12}m_{23}m_{31} \\
&\quad - m_{13}m_{22}m_{31} - m_{11}m_{23}m_{32} - m_{12}m_{21}m_{33}.
\end{aligned} \tag{24}$$

For two square matrices \mathbf{M} and \mathbf{N} , we have $|\mathbf{MN}| = |\mathbf{M}| |\mathbf{N}|$, and furthermore $|\mathbf{M}| = |\mathbf{M}^t|$. The determinant of any matrix is a measure of the d -dimensional hypervolume it “subtends.” For the particular case of a covariance matrix $\mathbf{\Sigma}$ (Sect. A.4.10), $|\mathbf{\Sigma}|$ is a measure of the hypervolume of the data taht yielded $\mathbf{\Sigma}$.

The *trace* of a $d \times d$ (square) matrix, denoted $\text{tr}[\mathbf{M}]$, is the sum of its diagonal elements:

$$\text{tr}[\mathbf{M}] = \sum_{i=1}^d m_{ii}. \tag{25}$$

Both the determinant and trace of a matrix are invariant with respect to rotations of the coordinate system.

A.2.5 Eigenvectors and eigenvalues

Given a $d \times d$ matrix \mathbf{M} , a very important class of linear equations is of the form

$$\mathbf{M}\mathbf{x} = \lambda\mathbf{x}, \tag{26}$$

which can be rewritten as

$$(\mathbf{M} - \lambda\mathbf{I})\mathbf{x} = \mathbf{0}, \tag{27}$$

where λ is a scalar, \mathbf{I} the identity matrix, and $\mathbf{0}$ the zero vector. This equation seeks the set of d (possibly non-distinct) solution vectors $\{\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_d\}$ — the eigenvectors — and their associated eigenvalues $\{\lambda_1, \lambda_2, \dots, \lambda_d\}$. Under multiplication by \mathbf{M} the eigenvectors are changed only in magnitude — not direction:

$$\mathbf{M}\mathbf{e}_j = \lambda_j\mathbf{e}_j. \tag{28}$$

CHARACTER- One method of finding the eigenvectors and eigenvalues is to solve the *characteristic equation* (or *secular equation*),

$$|\mathbf{M} - \lambda\mathbf{I}| = \lambda^d + a_1\lambda^{d-1} + \dots + a_{d-1}\lambda + a_d = 0, \tag{29}$$

SECULAR for each of its d (possibly non-distinct) roots λ_j . For each such root, we then solve a set of linear equations to find its associated eigenvector \mathbf{e}_j .

EQUATION

A.2.6 Matrix inversion

The inverse of a $n \times d$ matrix \mathbf{M} , denoted \mathbf{M}^{-1} , is the $d \times n$ matrix such that

$$\mathbf{M}\mathbf{M}^{-1} = \mathbf{I}. \tag{30}$$

COFACTOR Suppose first that \mathbf{M} is square. We call the scalar $C_{ij} = (-1)^{i+j}|M_{i|j}|$ the i, j *cofactor*

or equivalently the cofactor of the i, j entry of \mathbf{M} . As defined in Eq. 22, $\mathbf{M}_{i|j}$ is the $(d-1) \times (d-1)$ matrix formed by deleting the i^{th} row and j^{th} column of \mathbf{M} . The *adjoint* of \mathbf{M} , written $\text{Adj}[\mathbf{M}]$, is the matrix whose i, j entry is the j, i cofactor of \mathbf{M} . Given these definitions, we can write the inverse of a matrix as

ADJOINT

$$\mathbf{M}^{-1} = \frac{\text{Adj}[\mathbf{M}]}{|\mathbf{M}|}. \quad (31)$$

If \mathbf{M}^{-1} does not exist — because the columns of \mathbf{M} are not linearly independent or \mathbf{M} is not square — one typically uses instead the *pseudoinverse* \mathbf{M}^\dagger , defined as

PSEUDO-INVERSE

$$\mathbf{M}^\dagger = [\mathbf{M}^t \mathbf{M}]^{-1} \mathbf{M}^t, \quad (32)$$

which insures $\mathbf{M}^\dagger \mathbf{M} = \mathbf{I}$. Again, note especially that here \mathbf{M} need not be square.

A.3 Lagrange optimization

Suppose we seek the position \mathbf{x}_0 of an extremum of a scalar-valued function $f(\mathbf{x})$, subject to some constraint. For the following method to work, such a constraint must be expressible in the form $g(\mathbf{x}) = 0$. To find the extremum, we first form the Lagrangian function

$$L(\mathbf{x}, \lambda) = f(\mathbf{x}) + \underbrace{\lambda g(\mathbf{x})}_{=0}, \quad (33)$$

where λ is a scalar called the Lagrange *undetermined multiplier*. To find the extremum, we take the derivative

UNDETERMINED MULTIPLIER

$$\frac{\partial L(\mathbf{x}, \lambda)}{\partial \mathbf{x}} = \frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} + \underbrace{\lambda \frac{\partial g(\mathbf{x})}{\partial \mathbf{x}}}_{\neq 0 \text{ in gen.}} = 0, \quad (34)$$

and solve the resulting equations for λ and \mathbf{x}_0 — the position of the extremum.

A.4 Probability Theory

A.4.1 Discrete random variables

Let x be a random variable that can assume only a finite number m of different values in the set $\mathcal{X} = \{v_1, v_2, \dots, v_m\}$. We denote p_i as the probability that x assumes the value v_i :

$$p_i = \Pr\{x = v_i\}, \quad i = 1, \dots, m. \quad (35)$$

Then the probabilities p_i must satisfy the following two conditions:

$$\begin{aligned} p_i &\geq 0 \quad \text{and} \\ \sum_{i=1}^m p_i &= 1. \end{aligned} \quad (36)$$

Sometimes it is more convenient to express the set of probabilities $\{p_1, p_2, \dots, p_m\}$ in terms of the *probability mass function* $P(x)$. To distinguish between a random variable and the values that it can assume, it is sometime convenient to use an upper-case letter for the random variable and the corresponding lower-case letter for the value. The mass function would then be written $P_X(x)$. While this avoids the possible confusion in Eq. 37 and elsewhere (where x denotes a value, not a random variable), it also significantly complicates our the notation. Since it is usually clear from context whether one is referring to a random variable or its value, we will use the simpler notation as whenever possible.

The probability mass function must satisfy the following two conditions:

$$\begin{aligned} P(x) &\geq 0 \quad \text{and} \\ \sum_{x \in \mathcal{X}} P(x) &= 1. \end{aligned} \tag{37}$$

A.4.2 Expected values

MEAN

The *mean* or *expected value* or *average* of x is defined by

$$\mathcal{E}[x] = \mu = \sum_{x \in \mathcal{X}} xP(x) = \sum_{i=1}^m v_i p_i. \tag{38}$$

If one thinks of the probability mass function as defining a set of point masses, with p_i being the mass concentrated at $x = v_i$, then the expected value μ is just the center of mass. Alternatively, we can interpret μ as the arithmetic average of the values in a large random sample. More generally, if $f(x)$ is any function of x , the expected value of f is defined by

$$\mathcal{E}[f(x)] = \sum_{x \in \mathcal{X}} f(x)P(x). \tag{39}$$

Note that the process of forming an expected value is *linear*, in that if α_1 and α_2 are arbitrary constants,

$$\mathcal{E}[\alpha_1 f_1(x) + \alpha_2 f_2(x)] = \alpha_1 \mathcal{E}[f_1(x)] + \alpha_2 \mathcal{E}[f_2(x)]. \tag{40}$$

It is sometimes convenient to think of \mathcal{E} as an operator — the (linear) *expectation operator*. Two important special-case expectations are the *second moment* and the *variance*:

SECOND

MOMENT

$$\mathcal{E}[x^2] = \sum_{x \in \mathcal{X}} x^2 P(x) \tag{41}$$

VARIANCE

$$\text{Var}[x] = \mathcal{E}[(x - \mu)^2] = \sigma^2 = \sum_{x \in \mathcal{X}} (x - \mu)^2 P(x), \tag{42}$$

STANDARD
DEVIATION

where σ is the *standard deviation* of x . Physically, if we think of x as a random signal, the second moment is its total average power and the variance is its AC power.

PROBABILITY
MASS
FUNCTION

Alternatively, the variance can be viewed as the moment of inertia of the probability mass function. The variance is never negative, and is zero if and only if all of the probability mass is concentrated at one point.

The standard deviation is a simple but valuable measure of how far values of x are likely to depart from the mean. Its very name suggests that it is the standard or typical amount one should expect a randomly drawn value for x to deviate or differ from μ . *Chebyshev's inequality* provides a mathematical relation between the standard deviation and $|x - \mu|$:

CHEBYSHEV'S
INEQUALITY

$$\Pr\{|x - \mu| > n\sigma\} \leq \frac{1}{n^2}. \quad (43)$$

This inequality is not a tight bound (and it is useless for $n < 1$); a more practical rule of thumb, which strictly speaking is true only for the normal distribution, is that 68% of the values will lie within one, 95% within two, and 99.7% within three standard deviations of the mean (Fig. A.1). Nevertheless, Chebyshev's inequality shows the strong link between the standard deviation and the spread of the distribution. In addition, it suggests that $|x - \mu|/\sigma$ is a meaningful normalized measure of the distance from x to the mean (cf. Sect. A.4.12).

By expanding the quadratic in Eq. 42, it is easy to prove the useful formula

$$\text{Var}[x] = \mathcal{E}[x^2] - (\mathcal{E}[x])^2. \quad (44)$$

Note that, unlike the mean, the variance is *not* linear. In particular, if $y = \alpha x$, where α is a constant, then $\text{Var}[y] = \alpha^2 \text{Var}[x]$. Moreover, the variance of the sum of two random variables is usually *not* the sum of their variances. However, as we shall see below, variances do add when the variables involved are statistically independent.

In the simple but important special case in which x is binary valued (say, $v_1 = 0$ and $v_2 = 1$), we can obtain simple formulas for μ and σ . If we let $p = \Pr\{x = 1\}$, then it is easy to show that

$$\begin{aligned} \mu &= p \quad \text{and} \\ \sigma &= \sqrt{p(1-p)}. \end{aligned} \quad (45)$$

A.4.3 Pairs of discrete random variables

Let x be a random variable whose domain is $\mathcal{X} = \{v_1, v_2, \dots, v_m\}$, and let y be a random variable whose domain is $\mathcal{Y} = \{w_1, w_2, \dots, w_n\}$. We can think of (x, y) as a vector or a point in the *product space* of x and y . For each possible pair of values (v_i, w_j) we have a *joint probability* $p_{ij} = \Pr\{x = v_i, y = w_j\}$. These mn joint probabilities p_{ij} are non-negative and sum to 1. Alternatively, we can define a *joint probability mass function* $P(x, y)$ for which

PRODUCT
SPACE

$$\begin{aligned} P(x, y) &\geq 0 \quad \text{and} \\ \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} P(x, y) &= 1. \end{aligned} \quad (46)$$

The joint probability mass function is a complete characterization of the pair of random variables (x, y) ; that is, everything we can compute about x and y , individually

or together, can be computed from $P(x, y)$. In particular, we can obtain the separate *marginal distributions* for x and y by summing over the unwanted variable:

MARGINAL
DISTRIBUTION

$$\begin{aligned} P_x(x) &= \sum_{y \in \mathcal{Y}} P(x, y) \\ P_y(y) &= \sum_{x \in \mathcal{X}} P(x, y). \end{aligned} \quad (47)$$

As mentioned above, although the notation is more precise when we use subscripts as in Eq. 47, it is common to omit them and write simply $P(x)$ and $P(y)$ whenever the context makes it clear that these are in fact two different functions — rather than the same function merely evaluated with different variables.

A.4.4 Statistical independence

Variables x and y are said to be *statistically independent* if and only if

$$P(x, y) = P_x(x)P_y(y). \quad (48)$$

We can understand such independence as follows. Suppose that $p_i = \Pr\{x = v_i\}$ is the fraction of the time that $x = v_i$, and $q_j = \Pr\{y = w_j\}$ is the fraction of the time that $y = w_j$. Consider those situations where $x = v_i$. If it is still true that the fraction of those situations in which $y = w_j$ is the same value q_j , it follows that knowing the value of x did not give us any additional knowledge about the possible values of y ; in that sense y is independent of x . Finally, if x and y are statistically independent, it is clear that the fraction of the time that the specific pair of values (v_i, w_j) occurs must be the product of the fractions $p_i q_j = P_x(v_i)P_y(w_j)$.

A.4.5 Expected values of functions of two variables

In the natural extension of Sect. A.4.2, we define the expected value of a function $f(x, y)$ of two random variables x and y by

$$\mathcal{E}[f(x, y)] = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} f(x, y) P(x, y), \quad (49)$$

and as before the expectation operator \mathcal{E} is linear:

$$\mathcal{E}[\alpha_1 f_1(x, y) + \alpha_2 f_2(x, y)] = \alpha_1 \mathcal{E}[f_1(x, y)] + \alpha_2 \mathcal{E}[f_2(x, y)]. \quad (50)$$

The means and variances are:

$$\begin{aligned} \mu_x = \mathcal{E}[x] &= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} x P(x, y) \\ \mu_y = \mathcal{E}[y] &= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} y P(x, y) \\ \sigma_x^2 = V[x] = \mathcal{E}[(x - \mu_x)^2] &= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} (x - \mu_x)^2 P(x, y) \\ \sigma_y^2 = V[y] = \mathcal{E}[(y - \mu_y)^2] &= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} (y - \mu_y)^2 P(x, y). \end{aligned} \quad (51)$$

COVAR-
IANCE

An important new “cross-moment” can now be defined, the *covariance* of x and y :

$$\sigma_{xy} = \mathcal{E}[(x - \mu_x)(y - \mu_y)] = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} (x - \mu_x)(y - \mu_y) P(x, y). \quad (52)$$

We can summarize Eqs. 51 & 52 using vector notation as:

$$\boldsymbol{\mu} = \mathcal{E}[\mathbf{x}] = \sum_{\mathbf{x} \in \{\mathcal{X}\mathcal{Y}\}} \mathbf{x} P(\mathbf{x}) \quad (53)$$

$$\boldsymbol{\Sigma} = \mathcal{E}[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^t], \quad (54)$$

where $\boldsymbol{\Sigma}$ is the covariance matrix (cf., Sect. A.4.9).

The covariance is one measure of the degree of statistical dependence between x and y . If x and y are statistically independent, then $\sigma_{xy} = 0$. If α is a constant and $y = \alpha x$, which is a case of strong statistical dependence, it is also easy to show that $\sigma_{xy} = \alpha \sigma_x^2$. Thus, the covariance is positive if x and y both increase or decrease together, and is negative if y decreases when x increases. If $\sigma_{xy} = 0$, the variables x and y are said to be *uncorrelated*. It does *not* follow that uncorrelated variables must be statistically independent — covariance is just one measure of independence. However, it is a fact that uncorrelated variables are statistically independent if they have a multivariate normal distribution, and in practice statisticians often treat uncorrelated variables as if they were statistically independent.

UNCORRE-
LATED

There is an important *Cauchy-Schwarz inequality* for the variances σ_x and σ_y and the covariance σ_{xy} . It can be derived by observing that the variance of a random variable is never negative, and thus the variance of $\lambda x + y$ must be non-negative no matter what the value of the scalar λ . This leads to the famous inequality

CAUCHY-
SCHWARZ
INEQUALITY

$$\sigma_{xy}^2 \leq \sigma_x^2 \sigma_y^2, \quad (55)$$

which is analogous to the vector inequality $(\mathbf{x}^t \mathbf{y})^2 \leq \|\mathbf{x}\|^2 \|\mathbf{y}\|^2$ (Eq. 9).

The *correlation coefficient*, defined as

CORRELA-
TION COEF-
FICIENT

$$\rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y}, \quad (56)$$

is a normalized covariance, and must always be between -1 and $+1$. If $\rho = +1$, then x and y are maximally positively correlated, while if $\rho = -1$, they are maximally negatively correlated. If $\rho = 0$, the variables are uncorrelated. It is common for statisticians to consider variables to be uncorrelated for practical purposes if the magnitude of their correlation coefficient is below some threshold, such as .05, although the threshold that makes sense does depend on the actual situation.

If x and y are statistically independent, then for any two functions f and g

$$\mathcal{E}[f(x)g(y)] = \mathcal{E}[f(x)]\mathcal{E}[g(y)], \quad (57)$$

a result which follows from the definition of statistical independence and expectation. Note that if $f(x) = x - \mu_x$ and $g(y) = y - \mu_y$, this theorem again shows that $\sigma_{xy} = \mathcal{E}[(x - \mu_x)(y - \mu_y)]$ is zero if x and y are statistically independent.

A.4.6 Conditional probability

When two variables are statistically dependent, knowing the value of one of them lets us get a better estimate of the value of the other one. This is expressed by the following definition of the *conditional probability* of x given y :

$$\Pr\{x = v_i | y = w_j\} = \frac{\Pr\{x = v_i, y = w_j\}}{\Pr\{y = w_j\}}, \quad (58)$$

or, in terms of mass functions,

$$P(x|y) = \frac{P(x, y)}{P_y(y)}. \quad (59)$$

Note that if x and y are statistically independent, this gives $P(x|y) = P_x(x)$. That is, when x and y are independent, knowing the value of y gives you no information about x that you didn't already know from its marginal distribution $P_x(x)$.

To gain intuition about this definition of conditional probability, consider a simple two-variable binary case where both x and y are either 0 or 1. Suppose that a large number n of pairs of xy -values are randomly produced. Let n_{ij} be the number of pairs in which we find $x = i$ and $y = j$, i.e., we see the $(0, 0)$ pair n_{00} times, the $(0, 1)$ pair n_{01} times, and so on, where $n_{00} + n_{01} + n_{10} + n_{11} = n$. Suppose we pull out those pairs where $y = 1$, i.e., the $(0, 1)$ pairs and the $(1, 1)$ pairs. Clearly, the fraction of those cases in which x is also 1 is

$$\frac{n_{11}}{n_{01} + n_{11}} = \frac{n_{11}/n}{(n_{01} + n_{11})/n}. \quad (60)$$

Intuitively, this is what we would like to get for $P(x|y)$ when $y = 1$ and n is large. And, indeed, this is what we do get, because n_{11}/n is approximately $P(x, y)$ and $\frac{n_{11}/n}{(n_{01} + n_{11})/n}$ is approximately $P_y(y)$ for large n .

A.4.7 The Law of Total Probability and Bayes' rule

The expression

$$P_y(y) = \sum_{x \in \mathcal{X}} P(x, y) \quad (61)$$

is an instance of the *Law of Total Probability*. This law says that if an event A can occur in m different ways A_1, A_2, \dots, A_m , and if these m subevents are *mutually exclusive* — that is, cannot occur at the same time — then the probability of A occurring is the sum of the probabilities of the subevents A_i . In particular, the random variable y can assume the value y in m different ways — with $x = v_1$, with $x = v_2, \dots$, and with $x = v_m$. Because these possibilities are mutually exclusive, it follows from the Law of Total Probability that $P_y(y)$ is the sum of the joint probability $P(x, y)$ over all possible values for x . But from the definition of the conditional probability $P(y|x)$ we have

$$P(x, y) = P(y|x)P_x(x), \quad (62)$$

and thus, we obtain

$$P(x|y) = \frac{P(y|x)P_x(x)}{\sum_{x \in \mathcal{X}} P(y|x)P_x(x)}, \quad (63)$$

or in words,

$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{evidence}},$$

where these terms are discussed more fully in Chapt. ??.

Equation 63 is usually called *Bayes' rule*. Note that the denominator, which is just $P_y(y)$, is obtained by summing the numerator over all x values. By writing the denominator in this form we emphasize the fact that everything on the right-hand side of the equation is conditioned on x . If we think of x as the important variable, then we can say that the shape of the distribution $P(x|y)$ depends only on the numerator $P(y|x)P_x(x)$; the denominator is just a normalizing factor, sometimes called the *evidence*, needed to insure that $P(x|y)$ sums to one.

EVIDENCE

The standard interpretation of Bayes' rule is that it “inverts” statistical connections, turning $P(y|x)$ into $P(x|y)$. Suppose that we think of x as a “cause” and y as an “effect” of that cause. That is, we assume that if the cause x is present, it is easy to determine the probability of the effect y being observed, where the conditional probability function $P(y|x)$ — the *likelihood* — specifies this probability explicitly. If we observe the effect y , it might not be so easy to determine the cause x , because there might be several different causes, each of which could produce the same observed effect. However, Bayes' rule makes it easy to determine $P(x|y)$, provided that we know both $P(y|x)$ and the so-called *prior probability* $P_x(x)$, the probability of x before we make any observations about y . Said slightly differently, Bayes' rule shows how the probability distribution for x changes from the *prior distribution* $P_x(x)$ before anything is observed about y to the *posterior* $P(x|y)$ once we have observed the value of y .

LIKELIHOOD

PRIOR

POSTERIOR

A.4.8 Vector random variables

To extend these results from two variables x and y to d variables x_1, x_2, \dots, x_d , it is convenient to employ vector notation. The joint probability mass function $P(\mathbf{x})$ satisfies $P(\mathbf{x}) \geq 0$ and $\sum P(\mathbf{x}) = 1$ (Eq. 46), where the sum extends over all possible values for the vector \mathbf{x} . Note that $P(\mathbf{x})$ is a function of d variables, x_1, x_2, \dots, x_d , and can be a very complicated, multi-dimensional function. However, if the random variables x_i are statistically independent, it reduces to the product

$$\begin{aligned} P(\mathbf{x}) &= P_{x_1}(x_1)P_{x_2}(x_2) \cdots P_{x_d}(x_d) \\ &= \prod_{i=1}^d P_{x_i}(x_i). \end{aligned} \quad (64)$$

Here the separate marginal distributions $P_{x_i}(x_i)$ can be obtained by summing the joint distribution over the other variables. In addition to these univariate marginals, other marginal distributions can be obtained by this use of the Law of Total Probability. For example, suppose that we have $P(x_1, x_2, x_3, x_4, x_5)$ and we want $P(x_1, x_4)$, we merely calculate

$$P(x_1, x_4) = \sum_{x_2} \sum_{x_3} \sum_{x_5} P(x_1, x_2, x_3, x_4, x_5). \quad (65)$$

One can define many different conditional distributions, such as $P(x_1, x_2|x_3)$ or $P(x_2|x_1, x_4, x_5)$. For example,

$$P(x_1, x_2|x_3) = \frac{P(x_1, x_2, x_3)}{P(x_3)}, \quad (66)$$

where all of the joint distributions can be obtained from $P(\mathbf{x})$ by summing out the unwanted variables. If instead of scalars we have vector variables, then these conditional distributions can also be written as

$$P(\mathbf{x}_1|\mathbf{x}_2) = \frac{P(\mathbf{x}_1, \mathbf{x}_2)}{P(\mathbf{x}_2)}, \quad (67)$$

and likewise, in vector form, Bayes' rule becomes

$$P(\mathbf{x}_1|\mathbf{x}_2) = \frac{P(\mathbf{x}_2|\mathbf{x}_1)P(\mathbf{x}_1)}{\sum_{\mathbf{x}_1} P(\mathbf{x}_2|\mathbf{x}_1)P(\mathbf{x}_1)}. \quad (68)$$

A.4.9 Expectations, mean vectors and covariance matrices

The expected value of a vector is defined to be the vector whose components are the expected values of the original components. Thus, if $\mathbf{f}(\mathbf{x})$ is an n -dimensional, vector-valued function of the d -dimensional random vector \mathbf{x} ,

$$\mathbf{f}(\mathbf{x}) = \begin{bmatrix} f_1(\mathbf{x}) \\ f_2(\mathbf{x}) \\ \vdots \\ f_n(\mathbf{x}) \end{bmatrix}, \quad (69)$$

then the expected value of \mathbf{f} is defined by

$$\mathcal{E}[\mathbf{f}] = \begin{bmatrix} \mathcal{E}[f_1(\mathbf{x})] \\ \mathcal{E}[f_2(\mathbf{x})] \\ \vdots \\ \mathcal{E}[f_n(\mathbf{x})] \end{bmatrix} = \sum_{\mathbf{x}} \mathbf{f}(\mathbf{x})P(\mathbf{x}). \quad (70)$$

MEAN
VECTOR

In particular, the d -dimensional *mean vector* $\boldsymbol{\mu}$ is defined by

$$\boldsymbol{\mu} = \mathcal{E}[\mathbf{x}] = \begin{bmatrix} \mathcal{E}[x_1] \\ \mathcal{E}[x_2] \\ \vdots \\ \mathcal{E}[x_d] \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_d \end{bmatrix} = \sum_{\mathbf{x}} \mathbf{x}P(\mathbf{x}). \quad (71)$$

COVARIANCE
MATRIX

Similarly, the *covariance matrix* $\boldsymbol{\Sigma}$ is defined as the (square) matrix whose ij^{th} element σ_{ij} is the covariance of x_i and x_j :

$$\sigma_{ij} = \mathcal{E}[(x_i - \mu_i)(x_j - \mu_j)] \quad i, j = 1 \dots d, \quad (72)$$

as we saw in the two-variable case of Eq. 52. Therefore, in expanded form we have

$$\begin{aligned}
\mathbf{\Sigma} &= \begin{bmatrix} \mathcal{E}[(x_1 - \mu_1)(x_1 - \mu_1)] & \mathcal{E}[(x_1 - \mu_1)(x_2 - \mu_2)] & \dots & \mathcal{E}[(x_1 - \mu_1)(x_d - \mu_d)] \\ \mathcal{E}[(x_2 - \mu_2)(x_1 - \mu_1)] & \mathcal{E}[(x_2 - \mu_2)(x_2 - \mu_2)] & \dots & \mathcal{E}[(x_2 - \mu_2)(x_d - \mu_d)] \\ \vdots & \vdots & \ddots & \vdots \\ \mathcal{E}[(x_d - \mu_d)(x_1 - \mu_1)] & \mathcal{E}[(x_d - \mu_d)(x_2 - \mu_2)] & \dots & \mathcal{E}[(x_d - \mu_d)(x_d - \mu_d)] \end{bmatrix} \\
&= \begin{bmatrix} \sigma_{11} & \sigma_{12} & \dots & \sigma_{1d} \\ \sigma_{21} & \sigma_{22} & \dots & \sigma_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{d1} & \sigma_{d2} & \dots & \sigma_{dd} \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \dots & \sigma_{1d} \\ \sigma_{21} & \sigma_2^2 & \dots & \sigma_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{d1} & \sigma_{d2} & \dots & \sigma_d^2 \end{bmatrix}. \tag{73}
\end{aligned}$$

We can use the vector product $(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^t$, to write the covariance matrix as

$$\mathbf{\Sigma} = \mathcal{E}[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^t]. \tag{74}$$

Thus, the diagonal elements of $\mathbf{\Sigma}$ are just the variances of the individual elements of \mathbf{x} , which can never be negative; the off-diagonal elements are the covariances, which can be positive or negative. If the variables are statistically independent, the covariances are zero, and the covariance matrix is diagonal. The analog to the Cauchy-Schwarz inequality comes from recognizing that if \mathbf{w} is any d -dimensional vector, then the variance of $\mathbf{w}^t \mathbf{x}$ can never be negative. This leads to the requirement that the quadratic form $\mathbf{w}^t \mathbf{\Sigma} \mathbf{w}$ never be negative. Matrices for which this is true are said to be *positive semi-definite*; thus, the covariance matrix $\mathbf{\Sigma}$ must be positive semi-definite. It can be shown that this is equivalent to the requirement that none of the eigenvalues of $\mathbf{\Sigma}$ can ever be negative.

A.4.10 Continuous random variables

When the random variable x can take values in the continuum, it no longer makes sense to talk about the probability that x has a particular value, such as 2.5136, because the probability of any particular exact value will almost always be zero. Rather, we talk about the probability that x falls in some interval (a, b) ; instead of having a probability mass function $P(x)$ we have a *probability mass density function* $p(x)$. The mass density has the property that

MASS
DENSITY

$$\Pr\{x \in (a, b)\} = \int_a^b p(x) dx. \tag{75}$$

The name *density* comes by analogy with material density. If we consider a small interval $(a, a + \Delta x)$ over which $p(x)$ is essentially constant, having value $p(a)$, we see that $p(a) = \Pr\{x \in (a, a + \Delta x)\} / \Delta x$. That is, the probability mass density at $x = a$ is the probability mass $\Pr\{x \in (a, a + \Delta x)\}$ per unit distance. It follows that the probability density function must satisfy

$$\begin{aligned}
p(x) &\geq 0 \quad \text{and} \\
\int_{-\infty}^{\infty} p(x) dx &= 1. \tag{76}
\end{aligned}$$

In general, most of the definitions and formulas for discrete random variables carry over to continuous random variables with sums replaced by integrals. In particular, the expected value, mean and variance for a continuous random variable are defined by

$$\begin{aligned}\mathcal{E}[f(x)] &= \int_{-\infty}^{\infty} f(x)p(x) dx \\ \mu = \mathcal{E}[x] &= \int_{-\infty}^{\infty} xp(x) dx \\ \text{Var}[x] = \sigma^2 = \mathcal{E}[(x - \mu)^2] &= \int_{-\infty}^{\infty} (x - \mu)^2 p(x) dx,\end{aligned}\tag{77}$$

and, as in Eq. 44, we have $\sigma^2 = \mathcal{E}[x^2] - (\mathcal{E}[x])^2$.

The multivariate situation is similarly handled with continuous random vectors \mathbf{x} . The probability density function $p(\mathbf{x})$ must satisfy

$$\begin{aligned}p(\mathbf{x}) &\geq 0 \quad \text{and} \\ \int_{-\infty}^{\infty} p(\mathbf{x}) d\mathbf{x} &= 1,\end{aligned}\tag{78}$$

where the integral is understood to be a d -fold, multiple integral, and where $d\mathbf{x}$ is the element of d -dimensional volume $d\mathbf{x} = dx_1 dx_2 \cdots dx_d$. The corresponding moments for a general n -dimensional vector-valued function are

$$\mathcal{E}[\mathbf{f}(\mathbf{x})] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \mathbf{f}(\mathbf{x}) p(\mathbf{x}) dx_1 dx_2 \cdots dx_d = \int_{-\infty}^{\infty} \mathbf{f}(\mathbf{x}) p(\mathbf{x}) d\mathbf{x}\tag{79}$$

and for the particular d -dimensional functions as above, we have

$$\begin{aligned}\boldsymbol{\mu} = \mathcal{E}[\mathbf{x}] &= \int_{-\infty}^{\infty} \mathbf{x} p(\mathbf{x}) d\mathbf{x} \\ \boldsymbol{\Sigma} = \mathcal{E}[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^t] &= \int_{-\infty}^{\infty} (\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^t p(\mathbf{x}) d\mathbf{x}.\end{aligned}\tag{80}$$

If the components of \mathbf{x} are statistically independent, then the joint probability density function factors as

$$p(\mathbf{x}) = \prod_{i=1}^d p_i(x_i)\tag{81}$$

and the covariance matrix is diagonal.

Conditional probability density functions are defined just as conditional mass functions. Thus, for example, the density for x given y is given by

$$p(x|y) = \frac{p(x, y)}{p_y(y)} \quad (82)$$

and Bayes' rule for density functions is

$$p(x|y) = \frac{p(y|x)p_x(x)}{\int_{-\infty}^{\infty} p(y|x)p_x(x) dx}, \quad (83)$$

and likewise for the vector case.

Occasionally we will need to take the expectation with respect to a subset of the variables, and in that case we must show this as a subscript, for instance

$$E_{x_1}[f(x_1, x_2)] = \int_{-\infty}^{\infty} f(x_1, x_2)p(x_1) dx_1. \quad (83)$$

A.4.11 Distributions of sums of independent random variables

It frequently happens that we know the distributions for two independent random variables x and y , and we need to know the distribution for their sum $z = x + y$. It is easy to obtain the mean and the variance of the sum:

$$\begin{aligned} \mu_z &= \mathcal{E}[z] = \mathcal{E}[x + y] = \mathcal{E}[x] + \mathcal{E}[y] = \mu_x + \mu_y, \\ \sigma_z^2 &= \mathcal{E}[(z - \mu_z)^2] = \mathcal{E}[(x + y - (\mu_x + \mu_y))^2] = \mathcal{E}[(x - \mu_x + y - \mu_y)^2] \\ &= \mathcal{E}[(x - \mu_x)^2] + 2 \underbrace{\mathcal{E}[(x - \mu_x)(y - \mu_y)]}_{=0} + \mathcal{E}[(y - \mu_y)^2] \\ &= \sigma_x^2 + \sigma_y^2, \end{aligned} \quad (84)$$

where we have used the fact that the cross-term factors into $\mathcal{E}[x - \mu_x]\mathcal{E}[y - \mu_y]$ when x and y are independent; in this case the product is manifestly zero, since each of the expectations vanishes. Thus, in words, the mean of the sum of two independent random variables is the sum of their means, and the variance of their sum is the sum of their variances. If the variables are random yet not independent — for instance $y = -x$, where x is randomly distribution — then the variance is not the sum of the component variances.

It is only slightly more difficult to work out the exact probability density function for $z = x + y$ from the separate density functions for x and y . The probability that z is between ζ and $\zeta + \Delta z$ can be found by integrating the joint density $p(x, y) = p_x(x)p_y(y)$ over the thin strip in the xy -plane between the lines $x + y = \zeta$ and $x + y = \zeta + \Delta z$. It follows that, for small Δz ,

$$\Pr\{\zeta < z < \zeta + \Delta z\} = \left\{ \int_{-\infty}^{\infty} p_x(x)p_y(\zeta - x) dx \right\} \Delta z, \quad (85)$$

and hence that the probability density function for the sum is the *convolution* of the probability density functions for the components: CONVOLUTION

$$p_z(z) = p_x \star p_y = \int_{-\infty}^{\infty} p_x(x)p_y(z - x) dx. \quad (86)$$

As one would expect, these results generalize. It is not hard to show that:

- The mean of the sum of d independent random variables x_1, x_2, \dots, x_d is the sum of their means. (In fact the variables need not be independent for this to hold.)
- The variance of the sum is the sum of their variances.
- The probability density function for the sum is the convolution of the separate density functions:

$$p_z(z) = p_{x_1} \star p_{x_2} \star \dots \star p_{x_d}. \quad (87)$$

A.4.12 Univariate normal density

CENTRAL
LIMIT
THEOREM
GAUSSIAN

One of the most important results of probability theory is the *Central Limit Theorem*, which states that, under various conditions, the distribution for the sum of d independent random variables approaches a particular limiting form known as the *normal distribution*. As such, the *normal* or *Gaussian* probability density function is very important, both for theoretical and practical reasons. In one dimension, it is defined by

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2} \left(\frac{x-\mu}{\sigma} \right)^2}. \quad (88)$$

The normal density is traditionally described as a “bell-shaped curve”; it is completely determined by the numerical values for two parameters, the mean μ and the variance σ^2 . This is often emphasized by writing $p(x) \sim N(\mu, \sigma^2)$, which is read as “ x is distributed normally with mean μ and variance σ^2 .” The distribution is symmetrical about the mean, the peak occurring at $x = \mu$ and the width of the “bell” is proportional to the standard deviation σ . The normal density satisfies the following equations:

$$\begin{aligned} \mathcal{E}[1] &= \int_{-\infty}^{\infty} p(x) dx = 1 \\ \mathcal{E}[x] &= \int_{-\infty}^{\infty} x p(x) dx = \mu \\ \mathcal{E}[(x-\mu)^2] &= \int_{-\infty}^{\infty} (x-\mu)^2 p(x) dx = \sigma^2. \end{aligned} \quad (89)$$

Normally distributed data points tend to cluster about the mean. Numerically, the probabilities obey

$$\begin{aligned} \Pr\{|x-\mu| \leq \sigma\} &\approx 0.68 \\ \Pr\{|x-\mu| \leq 2\sigma\} &\approx 0.95 \\ \Pr\{|x-\mu| \leq 3\sigma\} &\approx 0.997, \end{aligned} \quad (90)$$

as shown in Fig. A.1.

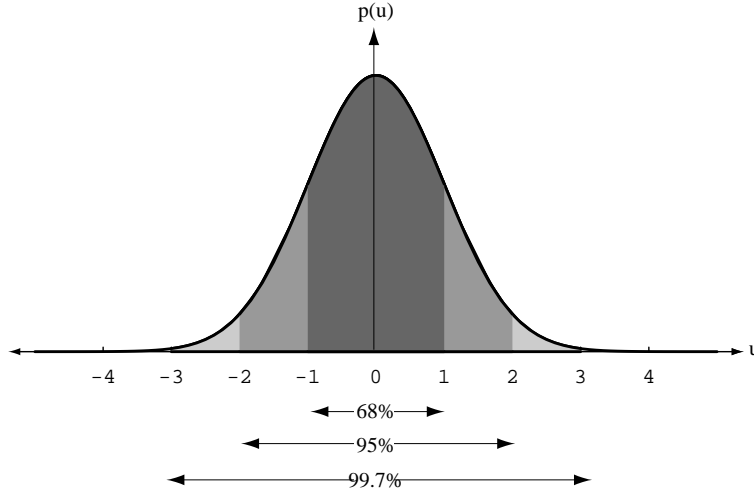


Figure A.1: A one-dimensional Gaussian distribution, $p(u) \sim N(0, 1)$, has 68% of its probability mass in the range $|u| \leq 1$, 95% in the range $|u| \leq 2$, and 99.7% in the range $|u| \leq 3$.

A natural measure of the distance from x to the mean μ is the distance $|x - \mu|$ measured in units of standard deviations:

$$r = \frac{|x - \mu|}{\sigma}, \quad (91)$$

the *Mahalanobis distance* from x to μ . Thus, the probability is .95 that the Mahalanobis distance from x to μ will be less than 2. If a random variable x is modified by (a) subtracting its mean and (b) dividing by its standard deviation, it is said to be *standardized*. Clearly, a standardized normal random variable $u = (x - \mu)/\sigma$ has zero mean and unit standard deviation, that is,

MAHALANOBIS
DISTANCE

STANDARDIZED

$$p(u) = \frac{1}{\sqrt{2\pi}} e^{-u^2/2}, \quad (92)$$

which can be written as $p(u) \sim N(0, 1)$.

A.5 Gaussian derivatives and integrals

Because of the prevalence of Gaussian functions throughout pattern recognition, we often have occasion to integrate and differentiate them. The first three derivatives of a one-dimensional (normalized) Gaussian are

$$\begin{aligned} \frac{\partial}{\partial x} \left[\frac{1}{\sqrt{2\pi}\sigma} e^{-x^2/(2\sigma^2)} \right] &= \frac{-x}{\sqrt{2\pi}\sigma^3} e^{-x^2/(2\sigma^2)} \\ \frac{\partial^2}{\partial x^2} \left[\frac{1}{\sqrt{2\pi}\sigma} e^{-x^2/(2\sigma^2)} \right] &= \frac{1}{\sqrt{2\pi}\sigma^5} (-\sigma^2 + x^2) e^{-x^2/(2\sigma^2)} \\ \frac{\partial^3}{\partial x^3} \left[\frac{1}{\sqrt{2\pi}\sigma} e^{-x^2/(2\sigma^2)} \right] &= \frac{1}{\sqrt{2\pi}\sigma^7} (3x\sigma^2 - x^3) e^{-x^2/(2\sigma^2)}, \end{aligned} \quad (93)$$

and are shown in Fig. A.2.

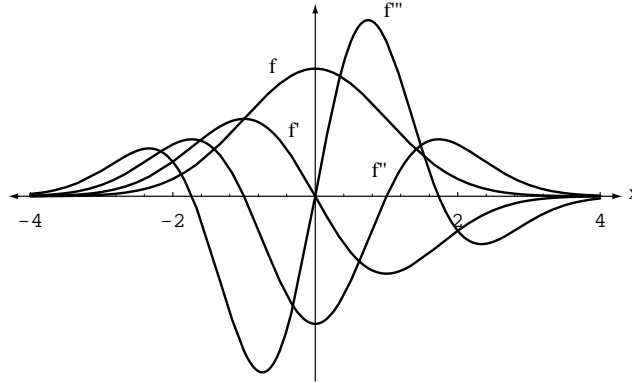


Figure A.2: A one-dimensional Gaussian distribution and its first three derivatives, shown for $f(x) \sim N(0, 1)$.

ERROR FUNCTION

An important finite integral of the Gaussian is the so-called *error function*, defined as

$$\text{erf}(u) = \frac{2}{\sqrt{\pi}} \int_0^u e^{-x^2/2} dx. \quad (94)$$

Note especially the pre-factor of 2 and the lower limit of integration. As can be seen from Fig. A.1, $\text{erf}(0) = 0$, $\text{erf}(1) = .68$ and $\lim_{x \rightarrow \infty} \text{erf}(x) = 1$. There is no closed analytic form for the error function, and thus we typically use tables, approximations or numerical integration for its evaluation (Fig. A.3).

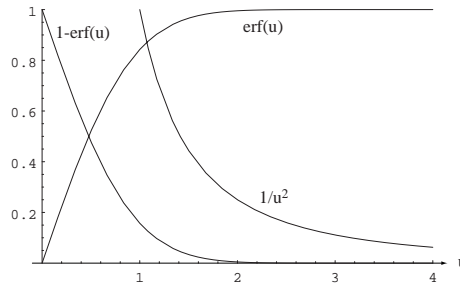


Figure A.3: The error function corresponds to the area under a standardized Gaussian (Eq. 94) between $-u$ and u , i.e., is the probability that a sample is chosen from the Gaussian $|x| \leq u$. Thus, the complementary probability, $1 - \text{erf}(u)$ is the probability that a sample is chosen with $|x| > u$ for the standardized Gaussian. Chebyshev's inequality states that for an *arbitrary* distribution having standard deviation $= 1$, this latter probability is bounded by $1/u^2$. As shown, this bound is quite loose for a Gaussian.

GAMMA FUNCTION

In calculating moments of Gaussians, we need the general integral of powers of x weighted by a Gaussian. Recall first the definition of a *gamma function*

$$\int_0^{\infty} x^n e^{-x} dx = \Gamma(n+1), \quad (95)$$

where the gamma function obeys

$$\Gamma(n) = n\Gamma(n-1) \quad (96)$$

and $\Gamma(1/2) = \sqrt{\pi}$. For n an integer we have $\Gamma(n+1) = n \times (n-1) \times (n-2) \dots 1 = n!$, read “ n factorial.”

Changing variables in Eq. 95, we find the moments of a (normalized) Gaussian distribution as

$$2 \int_0^{\infty} x^n \frac{e^{-x^2/(2\sigma^2)}}{\sqrt{2\pi}\sigma} dx = \frac{2^{n/2}\sigma^n}{\sqrt{\pi}} \Gamma\left(\frac{n+1}{2}\right), \quad (97)$$

where again we have used a pre-factor of 2 and lower integration limit of 0 in order give non-trivial (i.e., non-vanishing) results for odd n .

A.5.1 Multivariate normal densities

Normal random variables have many desirable theoretical properties. For example, it turns out that the convolution of two Gaussian functions is again a Gaussian function, and thus the distribution for the sum of two independent normal random variables is again normal. In fact, sums of dependent normal random variables also have normal distributions. Suppose that each of the d random variables x_i is normally distributed, each with its own mean and variance: $p(x_i) \sim N(\mu_i, \sigma_i^2)$. If these variables are independent, their joint density has the form

$$\begin{aligned} p(\mathbf{x}) &= \prod_{i=1}^d p_{x_i}(x_i) = \prod_{i=1}^d \frac{1}{\sqrt{2\pi}\sigma_i} e^{-\frac{1}{2} \left(\frac{x_i - \mu_i}{\sigma_i} \right)^2} \\ &= \frac{1}{(2\pi)^{d/2} \prod_{i=1}^d \sigma_i} e^{-\frac{1}{2} \sum_{i=1}^d \left(\frac{x_i - \mu_i}{\sigma_i} \right)^2}. \end{aligned} \quad (98)$$

This can be written in a compact matrix form if we observe that for this case the covariance matrix is diagonal, i.e.,

$$\Sigma = \begin{bmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_d^2 \end{bmatrix}, \quad (99)$$

and hence the inverse of the covariance matrix is easily written as

$$\mathbf{\Sigma}^{-1} = \begin{bmatrix} 1/\sigma_1^2 & 0 & \dots & 0 \\ 0 & 1/\sigma_2^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1/\sigma_d^2 \end{bmatrix}. \quad (100)$$

Thus, the quadratic form in Eq. 98 can be written as

$$\sum_{i=1}^d \left(\frac{x_i - \mu_i}{\sigma_i} \right)^2 = (\mathbf{x} - \boldsymbol{\mu})^t \mathbf{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}). \quad (101)$$

Finally, by noting that the determinant of $\mathbf{\Sigma}$ is just the product of the variances, we can write the joint density compactly in the form

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{d/2} |\mathbf{\Sigma}|^{1/2}} e^{-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^t \mathbf{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})}. \quad (102)$$

This is the general form of a *multivariate normal density function*, where the covariance matrix $\mathbf{\Sigma}$ is no longer required to be diagonal. With a little linear algebra, it can be shown that if \mathbf{x} obeys this density function, then

$$\begin{aligned} \boldsymbol{\mu} = \mathcal{E}[\mathbf{x}] &= \int_{-\infty}^{\infty} \mathbf{x} p(\mathbf{x}) d\mathbf{x} \\ \mathbf{\Sigma} = \mathcal{E}[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^t] &= \int_{-\infty}^{\infty} (\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^t p(\mathbf{x}) d\mathbf{x}, \end{aligned} \quad (103)$$

just as one would expect. Multivariate normal data tend to cluster about the mean vector, $\boldsymbol{\mu}$, falling in an ellipsoidally-shaped cloud whose principal axes are the eigenvectors of the covariance matrix. The natural measure of the distance from \mathbf{x} to the mean $\boldsymbol{\mu}$ is provided by the quantity

$$r^2 = (\mathbf{x} - \boldsymbol{\mu})^t \mathbf{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}), \quad (104)$$

which is the square of the Mahalanobis distance from \mathbf{x} to $\boldsymbol{\mu}$. It is not as easy to standardize a vector random variable (reduce it to zero mean and unit covariance matrix) as it is in the univariate case. The expression analogous to $u = (x - \mu)/\sigma$ is $\mathbf{u} = \mathbf{\Sigma}^{-1/2}(\mathbf{x} - \boldsymbol{\mu})$, which involves the “square root” of the inverse of the covariance matrix. The process of obtaining $\mathbf{\Sigma}^{-1/2}$ requires finding the eigenvalues and eigenvectors of $\mathbf{\Sigma}$, and is just a bit beyond the scope of this Appendix.

A.5.2 Bivariate normal densities

It is illuminating to look at the so-called bivariate normal density, that is, the case of two Gaussian random variables x_1 and x_2 . In this case, it is convenient to define $\sigma_1^2 = \sigma_{11}, \sigma_2^2 = \sigma_{22}$, and to introduce the correlation coefficient ρ defined by

$$\rho = \frac{\sigma_{12}}{\sigma_1 \sigma_2}. \quad (105)$$

With this notation, that the covariance matrix becomes

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}, \quad (106)$$

and its determinant simplifies to

$$|\mathbf{\Sigma}| = \sigma_1^2\sigma_2^2(1 - \rho^2). \quad (107)$$

Thus, the inverse covariance matrix is given by

$$\begin{aligned} \mathbf{\Sigma}^{-1} &= \frac{1}{\sigma_1^2\sigma_2^2(1 - \rho^2)} \begin{bmatrix} \sigma_2^2 & -\rho\sigma_1\sigma_2 \\ -\rho\sigma_1\sigma_2 & \sigma_1^2 \end{bmatrix} \\ &= \frac{1}{1 - \rho^2} \begin{bmatrix} \frac{1}{\sigma_1^2} & -\frac{\rho}{\sigma_1\sigma_2} \\ -\frac{\rho}{\sigma_1\sigma_2} & \frac{1}{\sigma_2^2} \end{bmatrix}. \end{aligned} \quad (108)$$

Next we explicitly expand the quadratic form in the normal density:

$$\begin{aligned} &(\mathbf{x} - \boldsymbol{\mu})^t \mathbf{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \\ &= [(x_1 - \mu_1)(x_2 - \mu_2)] \frac{1}{1 - \rho^2} \begin{bmatrix} \frac{1}{\sigma_1^2} & -\frac{\rho}{\sigma_1\sigma_2} \\ -\frac{\rho}{\sigma_1\sigma_2} & \frac{1}{\sigma_2^2} \end{bmatrix} \begin{bmatrix} (x_1 - \mu_1) \\ (x_2 - \mu_2) \end{bmatrix} \\ &= \frac{1}{1 - \rho^2} \left[\left(\frac{x_1 - \mu_1}{\sigma_1} \right)^2 - 2\rho \left(\frac{x_1 - \mu_1}{\sigma_1} \right) \left(\frac{x_2 - \mu_2}{\sigma_2} \right) + \left(\frac{x_2 - \mu_2}{\sigma_2} \right)^2 \right]. \end{aligned} \quad (109)$$

Thus, the general bivariate normal density has the form

$$\begin{aligned} p_{x_1x_2}(x_1, x_2) &= \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1 - \rho^2}} \times \\ &e^{-\frac{1}{2(1 - \rho^2)} \left[\left(\frac{x_1 - \mu_1}{\sigma_1} \right)^2 - 2\rho \left(\frac{x_1 - \mu_1}{\sigma_1} \right) \left(\frac{x_2 - \mu_2}{\sigma_2} \right) + \left(\frac{x_2 - \mu_2}{\sigma_2} \right)^2 \right]}. \end{aligned} \quad (110)$$

As we can see from Fig. A.4, $p(x_1, x_2)$ is a hill-shaped surface over the x_1x_2 plane. The peak of the hill occurs at the point $(x_1, x_2) = (\mu_1, \mu_2)$, i.e., at the mean vector $\boldsymbol{\mu}$. The shape of the hump depends on the two variances σ_1^2 and σ_2^2 , and the correlation coefficient ρ . If we slice the surface with horizontal planes parallel to the x_1x_2 plane, we obtain the so-called *level curves*, defined by the locus of points where the quadratic form

$$\left(\frac{x_1 - \mu_1}{\sigma_1} \right)^2 - 2\rho \left(\frac{x_1 - \mu_1}{\sigma_1} \right) \left(\frac{x_2 - \mu_2}{\sigma_2} \right) + \left(\frac{x_2 - \mu_2}{\sigma_2} \right)^2 \quad (111)$$

is constant. It is not hard to show that $|\rho| \leq 1$, and that this implies that the level curves are ellipses. The x and y extent of these ellipses are determined by the variances σ_1^2 and σ_2^2 , and their eccentricity is determined by ρ . More specifically, the *principal axes* of the ellipse are in the direction of the eigenvectors \mathbf{e}_i of $\mathbf{\Sigma}$, and the different widths in these directions $\sqrt{\lambda_i}$. For instance, if $\rho = 0$, the principal axes of the ellipses are parallel to the coordinate axes, and the variables are statistically independent. In the special cases where $\rho = 1$ or $\rho = -1$, the ellipses collapse to straight lines. Indeed,

PRINCIPAL
AXES

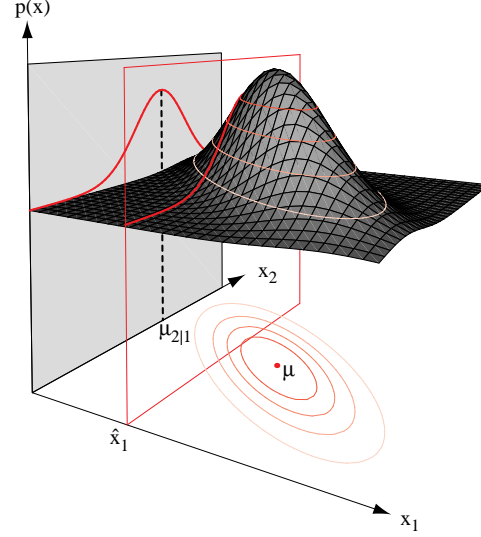


Figure A.4: A two-dimensional Gaussian having mean $\boldsymbol{\mu}$ and non-diagonal covariance $\boldsymbol{\Sigma}$. If the value on one variable is known, for instance $x_1 = \hat{x}_1$, the distribution over the other variable is Gaussian with mean $\mu_{2|1}$.

the joint density becomes singular in this situation, because there is really only one independent variable. We shall avoid this degeneracy by assuming that $|\rho| < 1$.

One of the important properties of the multivariate normal density is that all conditional and marginal probabilities are also normal. To find such a density explicitly, which we denote $p_{x_2|x_1}(x_2|x_1)$, we substitute our formulas for $p_{x_1x_2}(x_1, x_2)$ and $p_{x_1}(x_1)$ in the defining equation

$$\begin{aligned}
 p_{x_2|x_1}(x_2|x_1) &= \frac{p_{x_1x_2}(x_1, x_2)}{p_{x_1}(x_1)} \\
 &= \left[\frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} e^{-\frac{1}{2(1-\rho^2)} \left[\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x_1-\mu_1}{\sigma_1}\right)\left(\frac{x_2-\mu_2}{\sigma_2}\right) + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2 \right]} \right] \\
 &\quad \times \left[\sqrt{2\pi}\sigma_1 e^{\frac{1}{2}\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2} \right] \\
 &= \frac{1}{\sqrt{2\pi}\sigma_2\sqrt{1-\rho^2}} e^{-\frac{1}{2(1-\rho^2)} \left[\frac{x_2-\mu_2}{\sigma_2} - \rho\frac{x_1-\mu_1}{\sigma_1} \right]^2} \\
 &= \frac{1}{\sqrt{2\pi}\sigma_2\sqrt{1-\rho^2}} e^{-\frac{1}{2} \left(\frac{x_2 - [\mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x_1 - \mu_1)]}{\sigma_2\sqrt{1-\rho^2}} \right)^2}. \tag{112}
 \end{aligned}$$

Thus, we have verified that the conditional density $p_{x_1|x_2}(x_1|x_2)$ is a normal distribution. Moreover, we have explicit formulas for the *conditional mean* $\mu_{2|1}$ and the *conditional variance* $\sigma_{2|1}^2$:

$$\begin{aligned}
 \mu_{2|1} &= \mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x_1 - \mu_1) \quad \text{and} \\
 \sigma_{2|1}^2 &= \sigma_2^2(1 - \rho^2), \tag{113}
 \end{aligned}$$

as illustrated in Fig. A.4.

These formulas provide some insight into the question of how knowledge of the value of x_1 helps us to estimate x_2 . Suppose that we know the value of x_1 . Then a natural estimate for x_2 is the conditional mean, $\mu_{2|1}$. In general, $\mu_{2|1}$ is a linear function of x_1 ; if the correlation coefficient ρ is positive, the larger the value of x_1 , the larger the value of $\mu_{2|1}$. If it happens that x_1 is the mean value μ_1 , then the best we can do is to guess that x_2 is equal to μ_2 . Also, if there is no correlation between x_1 and x_2 , we ignore the value of x_1 , whatever it is, and we always estimate x_2 by μ_2 . Note that in that case the variance of x_2 , given that we know x_1 , is the same as the variance for the marginal distribution, i.e., $\sigma_{2|1}^2 = \sigma_2^2$. If there is correlation, knowledge of the value of x_1 , whatever the value is, reduces the variance. Indeed, with 100% correlation there is no variance left in x_2 when the value of x_1 is known.

A.6 Information theory

A.6.1 Entropy and information

Assume we have a discrete set of symbols $\{v_1 v_2 \dots v_m\}$ with associated probabilities p_i . The entropy of the discrete distribution — a measure of the randomness or unpredictability of a sequence of symbols drawn from it — is

$$H = - \sum_{i=1}^m P_i \log_2 P_i, \quad (114)$$

where here we use the logarithm is base 2. In case any of the probabilities vanish, we use the relation $0 \log 0 = 0$. (For continuous distributions, we often use logarithm base e , denoted \ln .) If we recall the expectation operator (cf. Eq. 39), we can write $H = \mathcal{E}[\log 1/P]$, where we think of P as being a random variable whose possible values are p_1, p_2, \dots, p_m . Note that the entropy does not depend on the symbols, but just on their probabilities. The entropy is non-negative and measured in *bits* when the base of the logarithm is 2. One bit corresponds to the uncertainty that can be resolved by the answer to a single yes/no question. For a given number of symbols m , the uniform distribution in which each symbol is equally likely, is the *maximum entropy distribution* (and $H = \log_2 m$ bits) — we have the maximum uncertainty about the identity of each symbol that will be chosen. Conversely, if all the p_i are 0 except one, we have the *minimum entropy distribution* ($H = 0$ bits) — we are certain as to the symbol that will appear.

BIT

For a continuous distribution, the entropy is

$$H = - \int_{-\infty}^{\infty} p(x) \log p(x) dx, \quad (115)$$

and again $H = \mathcal{E}[\log 1/p]$. It is worth mentioning that among all continuous density functions having a given mean μ and variance σ^2 , it is the Gaussian that has the maximum entropy ($H = .5 + \log_2 (\sqrt{2\pi}\sigma)$ bits). We can let σ approach zero to find that a probability density in the form of a *Dirac delta* function, i.e.,

DIRAC
DELTA

$$\delta(x - a) = \begin{cases} 0 & \text{if } x \neq a \\ \infty & \text{if } x = a, \end{cases} \quad \text{with}$$

$$\int_{-\infty}^{\infty} \delta(x) dx = 1, \quad (116)$$

has the minimum entropy ($H = -\infty$ bits). For a Dirac function, we are sure that the value a will be selected each time.

Our use of entropy in continuous functions, such as in Eq. 115, belies some subtle issues which are worth pointing out. If x had units, such as meters, then the probability density $p(x)$ would have to have units of $1/x$. There is something fundamentally wrong in taking the logarithm of $p(x)$, since the argument of any nonlinear function has to be dimensionless. What we should really be dealing with is a dimensionless quantity, say $p(x)/p_0(x)$, where $p_0(x)$ is some reference density function (cf., Sect. A.6.2).

One of the key properties of the entropy of a *discrete* distribution is that it is invariant to “shuffling” the event labels; no such property is evident for continuous variables. The related question with continuous variables concerns what happens when one makes a change of variables. In general, if we make a change of variables, such as $y = x^3$ or even $y = 10x$, we will get a different value for the integral of $\int q(y) \log q(y) dy$, where q is the induced density for y . If entropy is supposed to measure the intrinsic disorganization, it doesn’t make sense that y would have a different amount of intrinsic disorganization than x .

Fortunately, in practice these concerns do not present important stumbling blocks since relative entropy and differences in entropy are more fundamental than H taken by itself. Nevertheless, questions of the foundations of entropy measures for continuous variables are addressed in books listed in Bibliographical Remarks.

A.6.2 Relative entropy

KULLBACK-
LEIBLER
DISTANCE

Suppose we have two discrete distributions over the same variable x , $p(x)$ and $q(x)$. The relative entropy or *Kullback-Leibler distance* is a measure of the “distance” between these distributions:

$$D_{KL}(p(x), q(x)) = \sum_x q(x) \ln \frac{q(x)}{p(x)}. \quad (117)$$

The continuous version is

$$D_{KL}(p(x), q(x)) = \int_{-\infty}^{\infty} q(x) \ln \frac{q(x)}{p(x)} dx. \quad (118)$$

Although $D_{KL}(p(\cdot), q(\cdot)) \geq 0$ and $D_{KL}(p(\cdot), q(\cdot)) = 0$ if and only if $p(\cdot) = q(\cdot)$, the relative entropy is not a true metric, since D_{KL} is not necessarily symmetric in the interchange $p \leftrightarrow q$ and furthermore the triangle inequality need not be satisfied.

A.6.3 Mutual information

Now suppose we have two distributions over possibly *different* random variables, e.g., $p(x)$ and $q(y)$. The mutual information is the reduction in uncertainty about one variable due to the knowledge of the other variable

$$I(p; q) = H(p) - H(p|q) = \sum_{x,y} r(x,y) \log \frac{r(x,y)}{p(x)q(y)}, \quad (119)$$

where $r(x, y)$ is the probability of finding value x and y . Mutual information is simply the relative entropy between the joint distribution $r(x, y)$ and the product distribution $p(x)q(y)$ and as such it measures how much the distributions of the variables differ from statistical independence. Mutual information does not obey all the properties of a metric. In particular, the metric requirement that if $p(x) = q(y)$ then $I(x; y) = 0$ need not hold, in general. As an example, suppose we have two binary random variables with $r(0, 0) = r(1, 1) = 1/2$, so $r(0, 1) = r(1, 0) = 0$. According to Eq. 119, the mutual information between $p(x)$ and $q(y)$ is $\log 2 = 1$.

The relationships among the entropy, relative entropy and mutual information are summarized in Fig. A.5. The figure shows, for instance, that the joint entropy $H(p, q)$ is generally larger than individual entropies $H(p)$ and $H(q)$; that $H(p) = H(p|q) + I(p; q)$, and so on.

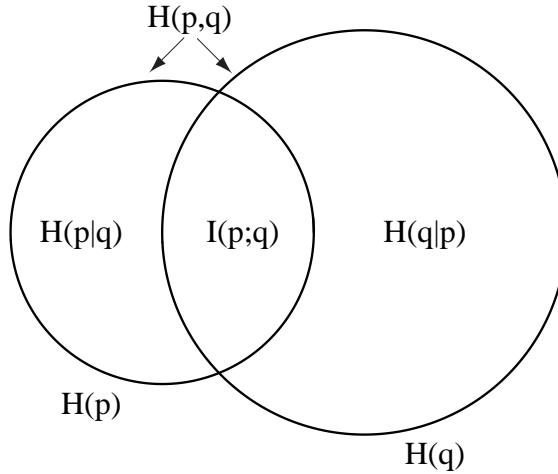


Figure A.5: The relationship among the entropy of distributions p and q , mutual information $I(p, q)$, and conditional entropies $H(p|q)$ and $H(q|p)$. From this figure one can quickly see relationships among the information functions, for instance $I(p; p) = H(p)$; that if $I(p; q) = 0$ then $H(q|p) = H(q)$, and so forth.

A.7 Computational complexity

In analyzing and describing the difficulty of problems and the algorithms designed to solve such problems, we turn now to computational complexity. For instance, calculating the standard deviation of a distribution is somehow “harder” than calculating its mean. Furthermore, some algorithms for computing some function may be faster or take less memory, than another algorithm. How can we specify such differences, independent of the current computer hardware (which is always changing anyway)?

To this end we use the concept of the order of a function and asymptotic notation and “big oh,” “big omega,” and “big theta” asymptotic notations. The three asymptotic bounds most often used are:

Asymptotic upper bound $O(g(x)) = \{f(x): \text{there exist positive constants } c \text{ and } x_0 \text{ such that } 0 \leq f(x) \leq cg(x) \text{ for all } x \geq x_0\}$

Asymptotic lower bound $\Omega(g(x)) = \{f(x): \text{there exist positive constants } c \text{ and } x_0 \text{ such that } 0 \leq cg(x) \leq f(x) \text{ for all } x \geq x_0\}$

Asymptotically tight bound $\Theta(g(x)) = \{f(x): \text{there exist positive constants } c_1, c_2, \text{ and } x_0 \text{ such that } 0 \leq c_1g(x) \leq f(x) \leq c_2g(x) \text{ for all } x \geq x_0\}$

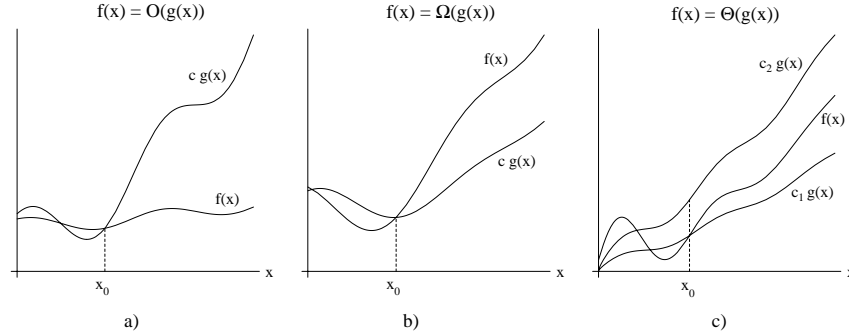


Figure A.6: Three types of order of a function describe the upper, lower and tight asymptotic bounds. a) $f(x) = O(g(x))$. b) $f(x) = \Omega(g(x))$. c) $f(x) = \Theta(g(x))$.

BIG OH

Consider the asymptotic upper bound. We say that $f(x)$ is “of the big oh order of $g(x)$ ” (written $f(x) = O(g(x))$) if there exist constants c_0 and x_0 such that $f(x) \leq c_0g(x)$ for all $x > x_0$. We shall assume that all our functions are positive and dispense with taking absolute values. This means simply that for sufficiently large x , an upper bound on $f(x)$ grows no worse than $g(x)$. For instance, if $f(x) = a + bx + cx^2$ then $f(x) = O(x^2)$ because for sufficiently large x , the constant, linear and quadratic terms can be “overcome” by proper choice of c_0 and x_0 . The generalization to functions of two or more variables is straightforward. It should be clear that by the definition above, the (big oh) order of a function is not unique. For instance, we can describe our particular $f(x)$ as being $O(x^2)$, $O(x^3)$, $O(x^4)$, $O(x^2 \ln x)$, and so forth. We write the tightest asymptotic upper bound $f(x) = o(g(x))$, read “little oh of $g(x)$ ” for the minimum in the class $O(g(x))$. Thus for instance if $f(x) = ax^2 + bx + c$, then $f(x) = o(x^2)$. Conversely, we use big omega notation, $\Omega(\cdot)$, for lower bounds, and little omega, $\omega(\cdot)$, for the tightest lower bound.

LITTLE OH

Of these, the big oh notation has proven to be most useful since we generally want an *upper* bound on the resources needed to solve a problem; it is frequently too difficult to determine the little oh complexity.

Such a rough analysis does not tell us the constants c and x_0 . For a finite size problem it is possible (though not likely) that a particular $O(x^3)$ algorithm is simpler than a particular $O(x^2)$ algorithm, and it is occasionally necessary for us to determine these constants to find which of several implementations is the simplest. Nevertheless, for our purposes the big oh notation as just described is generally the best way to describe the computational complexity of an algorithm.

Suppose we have a set of n vectors, each of which is d -dimensional and we want to calculate the mean vector. Clearly, this requires $O(nd)$ multiplications. Sometimes we stress space and time complexities, which are particularly relevant when contemplating parallel hardware implementations. For instance, the d -dimensional sample mean

could be calculated with d separate processors, each adding n sample values. Thus we can describe this implementation as $O(d)$ in *space* (i.e., the amount of memory or possibly the number of processors) and $O(n)$ in *time* (i.e., number of sequential steps). Of course for any particular algorithm there may be a number of time-space tradeoffs.

SPACE
COMPLEXITY

TIME
COMPLEXITY

Bibliographical Remarks

There are several good books on linear system, such as [13], and matrix computations [9]. Lagrange optimization and related techniques are covered in the definitive book [2]. While [12] is of historic interest and significance, readers seeking clear presentations of the central ideas in probability are [11, 8, 6, 18]. Another book treating the foundations is [3]. A handy reference to terms in probability and statistics is [17]. The definitive collection of papers on information theory is [7], and an excellent textbook, at the level of this one, is [5]; readers seeking a more abstract and formal treatment should consult [10]. The multi-volume [14, 15, 16] contains a description of computational complexity, the big oh and other asymptotic notations. Somewhat more accessible treatments can be found in [4] and [1].

Bibliography

- [1] Alfred V. Aho, John E. Hopcroft, and Jeffrey D. Ullman. *The Design and Analysis of Computer Algorithms*. Addison-Wesley, Reading, MA, 1974.
- [2] Dimitri P. Bertsekas. *Constrained Optimization and Lagrange Multiplier Methods*. Athena Scientific, xxx, xx, 1996.
- [3] Patrick Billingsley. *Probability and Measure*. John Wiley and Sons, New York, NY, 2 edition, 1986.
- [4] Thomas H. Cormen, Charles E. Leiserson, and Ronald L. Rivest. *Introduction to Algorithms*. MIT Press, Cambridge, MA, 1990.
- [5] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley Interscience, New York, NY, 1991.
- [6] Alvin W. Drake. *Fundamentals of Applied Probability Theory*. McGraw-Hill, New York, NY, 1967.
- [7] David Slepian (editor). *Key Papers in The Development of Information Theory*. IEEE Press, New York, NY, 1974.
- [8] William Feller. *An Introduction to Probability Theory and Its Applications*, volume 1. Wiley, New York, NY, 1968.
- [9] Gene H. Golub and Charles F. Van Loan. *Matrix Computations*. Johns Hopkins University Press, Baltimore, MD, 3 edition, 1996.
- [10] Robert M. Gray. *Entropy and Information Theory*. Springer-Verlag, New York, NY, 1990.
- [11] Richard W. Hamming. *The Art of Probability for Scientists and Engineers*. Addison-Wesley, New York, NY, 1991.
- [12] Harold Jeffreys. *Theory of Probability*. Oxford University Press, Oxford, UK, 1961 reprint edition, 1939.
- [13] Thomas Kailath. *Linear Systems*. Prentice-Hall, Englewood Cliffs, NJ, 1980.
- [14] Donald E. Knuth. *The Art of Computer Programming, Volume I*, volume 1. Addison-Wesley, Reading, MA, 1 edition, 1973.
- [15] Donald E. Knuth. *The Art of Computer Programming, Volume III*, volume 3. Addison-Wesley, Reading, MA, 1 edition, 1973.

- [16] Donald E. Knuth. *The Art of Computer Programming, Volume II*, volume 2. Addison-Wesley, Reading, MA, 1 edition, 1981.
- [17] Francis H. C. Marriott. *A Dictionary of Statistical Terms*. Longman Scientific & Technical, Essex, UK, 5 edition, 1990.
- [18] Yuri A. Rozanov. *Probability Theory: A Concise Course*. Dover, New York, NY, 1969.

Index

- \dagger , *see* matrix, pseudoinverse
- ρ , *see* correlation, coefficient
- $\mathcal{E}[\cdot]$, *see* expectation
- adjoint, *see* matrix, adjoint
- asymptotic lower bound, *see* lower bound, asymptotic
- asymptotic notation, 33
- asymptotic tight bound, *see* tight bound, asymptotic
- asymptotic upper bound, *see* upper bound, asymptotic
- average, *see* expected value
- Bayes' rule, 18, 19, 23
 - vector, 20
- big oh, 34
- big omega, 34
- big theta, 34
- Cauchy-Schwarz inequality, 9, 17
 - vector analog, 21
- Central Limit Theorem, 24
- Chebyshev's inequality, 15
- cofactor
 - matrix, *see* matrix, cofactor
- complexity
 - space, 35
 - time, 35
- computational complexity, 33–35
- conditional probability, *see* probability, conditional
- convolution, 23
- correlation
 - coefficient, 17
 - coefficient (ρ), 29, 31
- covariance, 17, 21
 - matrix, *see* matrix, covariance
 - normalized, 17
- cross moment, *see* covariance
- density
 - Gaussian
 - bivariate, 28
 - conditional mean, 30
 - marginal, 30
 - mean, 24
 - univariate, 24
 - variance, 24
 - joint
 - singular, 30
- distance
 - Euclidean, 9
 - Kullback-Leibler, 32
 - Mahalanobis, 25, 28
- distribution
 - Gaussian, 24
 - area, 15
 - covariance, 28
 - eigenvector, 28
 - moment, 27
 - multivariate, 27
 - principal axes, 28
 - joint, 20
 - marginal, 19
 - maximum entropy, 31
 - prior, 19
- dot product, *see* inner product
- eigenvalue, 12
- eigenvector, 12
- entropy, 31
 - continuous distribution, 31
 - discrete, 32
 - relative, 32
- error function ($\text{erf}(\cdot)$), 26
- Euclidean norm, *see* distance, Euclidean
- events
 - mutually exclusive, 18
- evidence, 19
- expectation
 - continuous, 22
 - entropy, 31

- linearity, 14, 16
 - vector, 20
- expected value, 14
 - two variables, 16
- factorial, 27
- function
 - Dirac delta, 31
 - gamma, 26
 - Kronecker, 8
 - vector valued, 22
- gamma function, *see* function, gamma
- Gaussian derivative, 26
- Gaussian derivative, 25
- gradient, 10
- Hessian matrix, *see* matrix, Hessian
- identity matrix, *see* matrix, identity (\mathbf{I})
- independence
 - statistical, 16
- independent variables
 - sum, 23
- information
 - bit, *see* bit
 - mutual, 32–33
- information theory, 31–33
- inner product, 9
- Jacobian matrix, *see* matrix, Jacobean
- Jacobian, 10, 11
- Kronecker delta, *see* function, Kronecker
- Kullback-Leibler, *see* distance, Kullback-Leibler
- Lagrange optimization, *see* optimization, Lagrange
- Lagrange undetermined multiplier, 13
- Law of Total Probability, 18
- level curves, 29
- likelihood, 19
- linear independence, 9
 - matrix columns, 13
- little oh, 34
- little omega, 34
- lower bound
 - asymptotic ($\Omega(\cdot)$), 34
- Mahalanobis distance, *see* distance, Mahalanobis
- marginal, 16
 - distribution, 16
- mass function
 - probability, *see* probability, mass function
- matrix
 - addition, 8
 - adjoint, 13
 - anti-symmetric, 8
 - covariance, 10
 - determinant, 28, 29
 - diagonal, 21, 22, 27
 - eigenvalues, 21
 - inverse, 27, 29
 - derivative, 10–11
 - determinant, 11–12
 - hypervolume, 12
 - Hessian, 11
 - identity (\mathbf{I}), 8
 - inverse
 - derivative, 10
 - inversion, 12–13
 - Jacobian, 10
 - multiplication, 9
 - non-negative, 8
 - positive semi-definite, 21
 - product, *see* outer product
 - pseudoinverse, 13
 - separable, 10
 - skew-symmetric, 8
 - square, 8
 - symmetric, 8, 10
 - trace, 12
- maximum entropy, 31
- mean, *see* expected value
 - calculation
 - computational complexity, 33
 - two variables, 16
- mean vector, *see* vector, mean
- moment
 - cross, *see* covariance
 - second, 14
- multiple integral, 22
- mutual information, *see* information, mutual
- normal, *see* distribution, Gaussian
- optimization
 - Lagrange, 13

- outer product, 9, 21
- principal axes, *see* axes, principal
- prior, 19
- prior distribution, *see* distribution, prior
- probability
 - conditional, 18
 - density, 21
 - joint, 22
 - joint, 15, 18
 - mass, 18, 21
 - joint, 15
 - mass function, 14
 - total
 - law, *see* Bayes' rule
- probability theory, 13–25
- product space, 15
- random variable
 - discrete, 13
 - vector, 19–21
- scalar product, *see* inner product
- second moment, *see* moment, second
- space-time tradeoff, 35
- standard deviation, 14, 25
- statistical
 - independence
 - expectation, 17
- statistical dependence, 17
- statistical independence, *see* independence, statistical, 17, 21
 - Gaussian, 29
 - vector, 19
- Taylor series, 11
- tight bound
 - asymptotic ($\Theta(\cdot)$), 34
- trace, *see* matrix, trace
- transpose, 8
- unpredictability, *see* entropy
- upper bound
 - asymptotic ($O(\cdot)$), 34
- variable
 - random
 - continuous, 21–23
 - discrete, 15
 - standardized, 28
 - standardized, 25
 - variables
 - uncorrelated, 17
- variance, 14
 - nonlinearity, 15
 - two variables, 16
- vector, 8
 - addition, 8
 - colinearity, 9
 - linearly independent, 9
 - mean, 20
 - orthogonal, 9
 - space, 9
 - span, 9
- vector product, *see* outer product