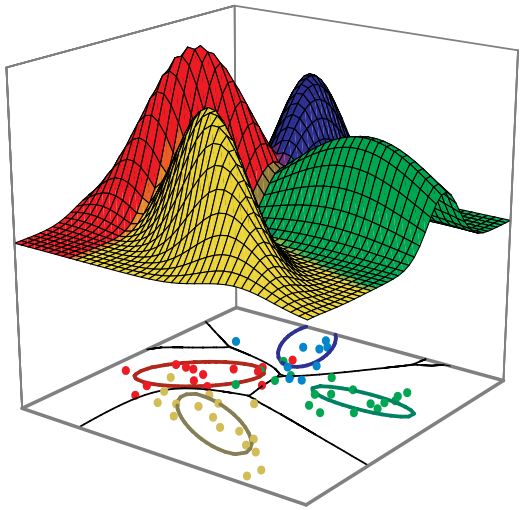
Part 2: Review of fundamental math



Notation

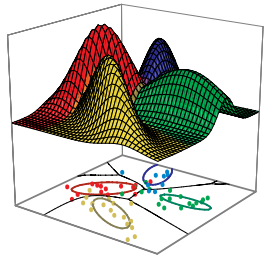
Linear Algebra

Probability Theory

Gaussian Distribution

Some materials in these slides were taken from Pattern Classification (2nd ed) **by R. O. Duda, P. E. Hart and D. G. Stork, John Wiley & Sons, 2000; Appendix A**

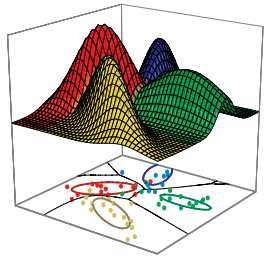
Notation



- Variables, Symbols, and Operations

$\simeq$	approximately equal to
$\equiv$	equivalent to (or defined to be)
$\propto$	proportional to
∞	infinity
$x \to a$	x approaches a
$t \leftarrow t + 1$	in an algorithm: assign to variable t the new value $t + 1$
$\lim_{x \to a} f(x)$	the value of $f(x)$ in the limit as x approaches 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 $f(x)$
$\lceil x \rceil$	ceiling of x—that is, the least integer not smaller than x (e.g., $\lceil 3.5 \rceil = 4$)

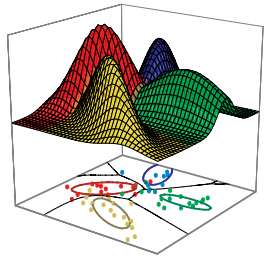
Notation



- Variables, Symbols, and Operations

$\lfloor x \rfloor$	floor of x—that is, the greatest integer not larger than x (e.g., $\lfloor 3.5 \rfloor = 3$)
m mod n	m modulo n—that is, the remainder when m is divided by n (e.g., 7 mod 5 = 2)
Rand$[l, u)$	in a computer program, a routine that returns a real number x, randomly chosen in the range $l \leq x < u$
$\ln(x)$	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—that is, e raised to the power of x
$\partial f(x)/\partial x$	partial derivative of f with respect to x
$\int_a^b f(x)dx$	the integral of $f(x)$ between a and b. If no limits are written, the full space is assumed
$F(x;\ \theta)$	function of x, with implied dependence upon θ
■	Q.E.D., quod erat demonstrandum ("which was to be proved")—used to signal the end of a proof

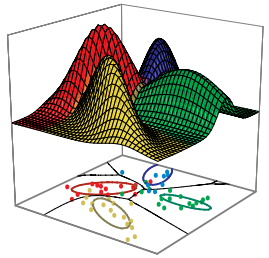
Notation



- Mathematical Operations

$\bar{x}$	mean or average value of x
$\mathcal{E}[f(x)]$	the expected value of function $f(x)$ where x is a random variable
$\mathcal{E}_y[f(x, y)]$	the expected value of function over several variables, $f(x, y)$, taken over a subset y of them
$\mathrm{Var}[f(\cdot)]$	the variance—that is, $\mathcal{E}[(f(x) - \mathcal{E}[f(x)])^2]$
$\mathrm{Var}_f[\cdot]$	the variance—that is, $\mathcal{E}_f[(x - \mathcal{E}_f[x])^2]$
$\sum_{i=1}^{n} a_i$	the sum from $i = 1$ to n—that is, $a_1 + a_2 + \cdots + a_n$
$\prod_{i=1}^{n} a_i$	the product from $i = 1$ to n—that is, $a_1 \times a_2 \times \cdots \times a_n$
$f(t) \star g(t)$	convolution of $f(t)$ and $g(t)$, $\int_{-\infty}^{\infty} f(\tau) g(t - \tau)\, d\tau$

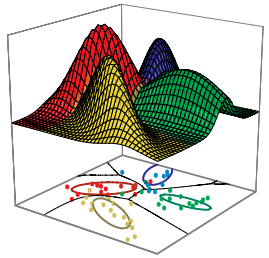
Notation



- Vectors and Matrices

$\mathbf{R}^d$	d-dimensional Euclidean space
$\mathbf{x}, \mathbf{A}, \ldots$	boldface is used for (column) vectors and matrices
$\mathbf{f}(x)$	vector-valued function (note the boldface) of a scalar argument
$\mathbf{f}(\mathbf{x})$	vector-valued function (note the boldface) of a vector argument
$\mathbf{I}$	identity matrix, a square matrix having 1's on the diagonal and 0 everywhere else
$\mathbf{1}_i$	vector of length i consisting solely of 1's
$diag(a_1, a_2, \ldots, a_d)$	matrix whose diagonal elements are $a_1, a_2, \ldots, a_d$, and off-diagonal elements are 0

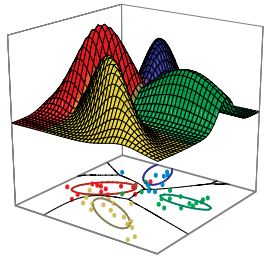
Notation



- Vectors and Matrices

$\mathbf{x}^t$	transpose of vector $\mathbf{x}$
$\|\mathbf{x}\|$	Euclidean norm of vector $\mathbf{x}$
$\boldsymbol{\Sigma}$	covariance matrix
tr[$\mathbf{A}$]	the trace of $\mathbf{A}$—that is, the sum of its diagonal elements
$\mathbf{A}^{-1}$	the inverse of matrix $\mathbf{A}$
$\mathbf{A}^{\dagger}$	pseudoinverse of matrix $\mathbf{A}$
$\|\mathbf{A}\|$ or Det[$\mathbf{A}$]	determinant of $\mathbf{A}$
λ	eigenvalue
$\mathbf{e}$	eigenvector
$\mathbf{u}_i$	unit vector in the ith direction in Euclidean space

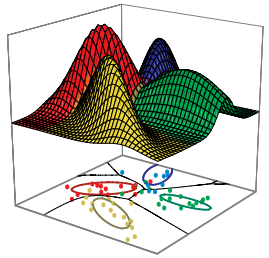
Notation



- Sets

$\mathcal{A}, \mathcal{B}, \mathcal{C}, \mathcal{D}, \ldots$	"Calligraphic" font generally denotes sets or lists—for example, a data set $\mathcal{D} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$
$\mathbf{x} \in \mathcal{D}$	$\mathbf{x}$ 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—that is, the set containing all elements in either $\mathcal{A}$ or $\mathcal{B}$
$\mathcal{A} \cap \mathcal{B}$	intersection of two sets—that is, the set containing all elements that are in both $\mathcal{A}$ and $\mathcal{B}$
$\lvert\mathcal{D}\rvert$	the cardinality of set $\mathcal{D}$—that is, the number of (possibly nondistinct) discrete elements in it

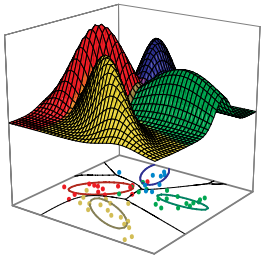
Notation



- Probability, Distributions, and Complexity

ω	state of nature (i.e. class of object)
$P(\cdot)$	probability mass
$p(\cdot)$	probability density
$P(a, b)$	the joint probability—that is, the probability of having both a and b
$p(a, b)$	the joint probability density—that is, the probability density of having both a and b
$\Pr[\cdot]$	the probability of a condition being met—for example, $\Pr[x < x_0]$ means the probability that x is less than x_0
$p(\mathbf{x}\|\boldsymbol{\theta})$	the conditional probability density of $\mathbf{x}$ given $\boldsymbol{\theta}$
$\mathbf{w}$	weight vector
$\lambda(\cdot, \cdot)$	loss function
$\nabla = \begin{pmatrix} \frac{\partial}{\partial x_1} \\ \frac{\partial}{\partial x_2} \\ \vdots \\ \frac{\partial}{\partial x_d} \end{pmatrix}$	gradient operator in $\mathbf{R}^d$, sometimes written $grad[\cdot]$

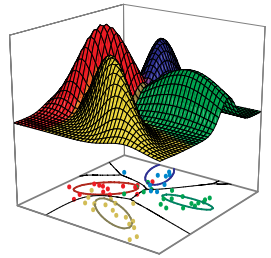
Notation



- Probability, Distributions, and Complexity

$\nabla_{\boldsymbol{\theta}} = \begin{pmatrix} \frac{\partial}{\partial\theta_1} \\ \frac{\partial}{\partial\theta_2} \\ \vdots \\ \frac{\partial}{\partial\theta_d} \end{pmatrix}$	gradient operator in $\boldsymbol{\theta}$ coordinates, sometimes written $grad_{\boldsymbol{\theta}}$
$\hat{\boldsymbol{\theta}}$	maximum-likelihood estimate of $\boldsymbol{\theta}$
$\sim$	"has the distribution"—for example, $p(x) \sim N(\mu, \sigma^2)$ 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 vector $\boldsymbol{\mu}$ 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—that is, uniform density within the smallest axes-aligned bounding box that contains both $\mathbf{x}_l$ and $\mathbf{x}_u$, and 0 elsewhere
$T(\mu, \delta)$	triangle distribution, having center μ and full half-width δ

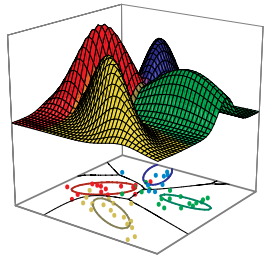
Notation



- Probability, Distributions, and Complexity

$\delta(x)$	Dirac delta function, which has value 0 for $x \neq 0$, and integrates to unity
δ_{ij}	Kronecker delta symbol, which has value 1 if its two indexes match, and 0 otherwise
$\Gamma(\cdot)$	Gamma function
$n!$	n factorial—that is, $n \times (n-1) \times (n-2) \times \cdots \times 1$
$\binom{n}{k} = \frac{n!}{k!\,(n-k)!}$	binomial coefficient, read "n choose k," for n and k integers
$O(h(x))$	big oh order of $h(x)$
$\Theta(h(x))$	big theta order of $h(x)$
$\Omega(h(x))$	big omega order of $h(x)$
$\sup_x f(x)$	the supremum value of $f(x)$—the least upper bound or global maximum of $f(x)$ over all values of x

Linear Algebra - Notation



- **x** = d-dimensional column vector

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{pmatrix}$$

- $\mathbf{x}^t$ = transpose of **x**

$$\mathbf{x}^t = (x_1 \; x_2 \; \ldots \; x_d).$$

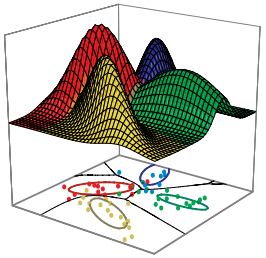
- n x d matrix **M**

$$\mathbf{M} = \begin{pmatrix} m_{11} & m_{12} & m_{13} & \ldots & m_{1d} \\ m_{21} & m_{22} & m_{23} & \ldots & m_{2d} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ m_{n1} & m_{n2} & m_{n3} & \ldots & m_{nd} \end{pmatrix}$$

- $\mathbf{M}^t$ = transpose of **M**

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

Linear Algebra - Notation



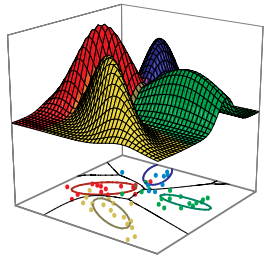
- Symmetric: $m_{ij} = m_{ji}$
- Screw-symmetric: $m_{ij} = -m_{ji}$
- Nonnegative: $m_{ij} > 0$ for all i,j
- Diagonal: $m_{ij} = 0$ for all $i \neq j$
- Identity, $\mathbf{I}$: diagonal with $m_{ij} = 1$ for all i=j
- Addition/subtraction element by element

- Multiplication: **Mx=y**

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

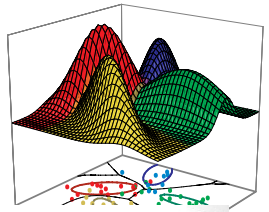
$$y_i = \sum_{j=1}^{d} m_{ij} x_j$$

Linear Algebra



- Scalar / Inner Product: **x•y** or $\mathbf{x}^t\mathbf{y}$: $\mathbf{x}^t\mathbf{y} = \sum_{i=1}^{d} x_i y_i = \mathbf{y}^t\mathbf{x}$
- Euclidean norm: $\|\mathbf{x}\| = \sqrt{\mathbf{x}^t\mathbf{x}}$
- Normalized if: $\|\mathbf{x}\| = 1$
- Angle between 2 vectors obeys: $\cos\theta = \frac{\mathbf{x}^t\mathbf{y}}{||\mathbf{x}||\ ||\mathbf{y}||}$
- 2 vectors orthogonal if $x^ty=0$; colinear if: $||\mathbf{x}^t\mathbf{y}|| = ||\mathbf{x}||\ ||\mathbf{y}||$
- Set of vectors $\{x_1, x_2, \ldots, x_n\}$ linearly independent if no vector in set can be written as combination of other vectors.
  - Set of d linearly independent vectors spans a d-dimensional space -- any vector in space linear combo of set.

Linear Algebra



- Outer Product: $\mathbf{M} = \mathbf{x}\mathbf{y}^t = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{pmatrix} (y_1\ y_2\ \cdots\ y_n) = \begin{pmatrix} x_1y_1 & x_1y_2 & \cdots & x_1y_n \\ x_2y_1 & x_2y_2 & \cdots & x_2y_n \\ \vdots & \vdots & \ddots & \vdots \\ x_dy_1 & x_dy_2 & \cdots & x_dy_n \end{pmatrix}$

- Derivative of scalar function of a vector:
  - aka gradient

$$\nabla f(\mathbf{x}) = \text{grad}\, f(\mathbf{x}) = \frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} = \begin{pmatrix} \frac{\partial f(\mathbf{x})}{\partial x_1} \\ \frac{\partial f(\mathbf{x})}{\partial x_2} \\ \vdots \\ \frac{\partial f(\mathbf{x})}{\partial x_d} \end{pmatrix}$$

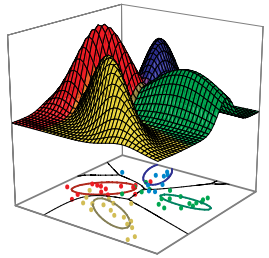
- Derivative of a vector function of a vector:
  - aka Jacobian matrix

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

- Derivative of a matrix M is element-by-element

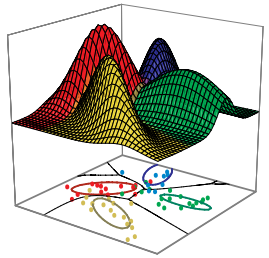
$$\frac{\partial}{\partial \mathbf{x}}[\mathbf{M}\mathbf{x}] = \mathbf{M} \qquad \frac{\partial}{\partial \mathbf{x}}[\mathbf{y}^t\mathbf{x}] = \frac{\partial}{\partial \mathbf{x}}[\mathbf{x}^t\mathbf{y}] = \mathbf{y} \qquad \frac{\partial}{\partial \mathbf{x}}[\mathbf{x}^t\mathbf{M}\mathbf{x}] = [\mathbf{M} + \mathbf{M}^t]\mathbf{x}$$

Linear Algebra



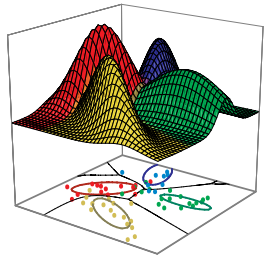
- Determinant of a square matrix, $|\mathbf{M}|$, is a scalar value
  - Reveals properties of **M**
  - $|\mathbf{M}| = 0$ if rows not linearly indep
  - Must be non-zero for $\mathbf{M}^{-1}$ to exist
  - For 1D, $|\mathbf{M}| = m_{11}$
  - For 2D, $|\mathbf{M}| = m_{11}m_{22}-m_{12}m_{21}$
  - For 3D, $|\mathbf{M}| = m_{11}m_{22}m_{33}+m_{13}m_{21}m_{32}+m_{12}m_{23}m_{31}$
    $-m_{13}m_{22}m_{31} -m_{11}m_{23}m_{32} -m_{12}m_{21}m_{33}$
  - For higher dimension, use expansion by minors
  - M,N eqal size: $|\mathbf{MN}| = |\mathbf{M}|\,|\mathbf{N}|$; $|\mathbf{M}| = |\mathbf{M}^t|$
  - Trace = sum of all diagonal values: $\mathrm{tr}[\mathbf{M}] = \sum_{i=1}^{d} m_{ii}$

Linear Algebra



- Matrix inversion: $\mathbf{M}\mathbf{M}^{-1}=\mathbf{I}$
  - Compute through adjoint method (p609) or linear operations
  - $[\mathbf{MN}]^{-1}=\mathbf{N}^{-1}\mathbf{M}^{-1}$
- If **M** not square, use Pseudo inverse: $\mathbf{M}^{\dagger} = [\mathbf{M}^t\mathbf{M}]^{-1}\mathbf{M}^t$
  - Useful for least-squares methods since ensures: $\mathbf{M}^{\dagger}\mathbf{M} = \mathbf{I}$

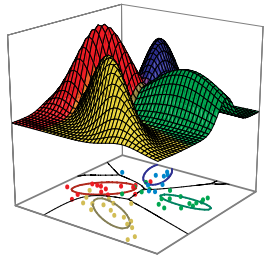
Linear Algebra



- Eigenvectors & Eigenvalues
  - For dxd matrix **M**, let $\mathbf{Mx}=\lambda\mathbf{x}$ or $(\mathbf{M}-\lambda\mathbf{I})\mathbf{x} = 0$
  - Up to d (possibly non-distinct) solutions for real symmetric **M**
    - λ_i is i^{th} eigenvalue
    - $x = e_i$ is i^{th} eigenvector
  - Multiplication of eigenvector by **M** only scales, does not change direction: $\mathbf{Me}_j = \lambda_j\mathbf{e}_j$
  - If M is diagonal, λ_i are non-zero diagonal values & eigenvectors are unit vectors parallel to coordinate axes.

$$\mathrm{tr}[\mathbf{M}] = \sum_{i=1}^{d} \lambda_i \qquad \text{and} \qquad |\mathbf{M}| = \prod_{i=1}^{d} \lambda_i$$

Linear Algebra

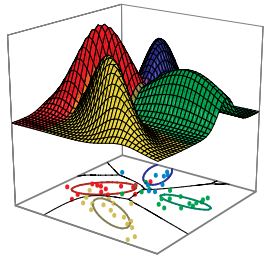


- Exercise: Find the eigenvectors and eigenvalues of the following matrix

$$A = \begin{pmatrix} 2 & 1 \\ 1 & 4 \end{pmatrix}$$

- 1) Solve $det(A-\lambda I) = 0$ to get 2 eigenvalues
- 2) Plug in each eigenvalue and calculate corresponding eigenvector

Solution of Exercise 1: Get eigenvalues, then plug in:



$$\textit{For } \lambda_1 = 3+\sqrt{2} \textit{ using} \begin{pmatrix} 2-\lambda & 1 \\ 1 & 4-\lambda \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = 0, \textit{we obtain :}$$

$$\begin{cases} (-1-\sqrt{2})x_1 + x_2 = 0 \\ x_1 + (1-\sqrt{2})x_2 = 0 \end{cases} \Leftrightarrow (-1-\sqrt{2})x_1 + x_2 = 0$$

Both eq'n describe same vector…

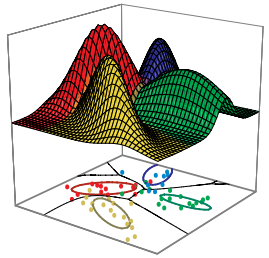
This latter equation is a straight line colinear to the vector:

$$\vec{V}_1 = (1, 1+\sqrt{2})^T$$

$$For\ \lambda_2 = 3-\sqrt{2}\ \text{using} \begin{pmatrix} 2-\lambda & 1 \\ 1 & 4-\lambda \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = 0, \text{we obtain :}$$

$$\begin{cases} (\sqrt{2}-1)x_1 + x_2 = 0 \\ x_1 + (1+\sqrt{2})x_2 = 0 \end{cases} \Leftrightarrow (\sqrt{2}-1)x_1 + x_2 = 0 \qquad \vec{V}_2 = (1, 1-\sqrt{2})^T$$

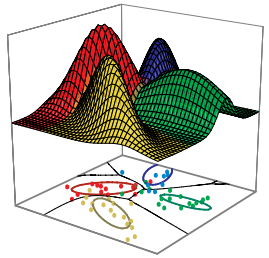
Definite Matrices



- A square matrix A is:
  1. *positive definite if* $x^T Ax>0$ for all nonzero column vectors x.
  2. *negative definite* if $x^T Ax < 0$ for all nonzero x.
  3. *positive semi-definite* if $x^T Ax \geq 0$.
  4. *negative semi-definite* if $x^T Ax \leq 0$ for all x.

These definitions are hard to check directly and you might as well forget them for all practical purposes.

Definite Matrices

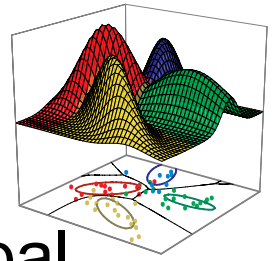


More useful in practice are the following properties, which hold when the matrix A is symmetric and which are easier to check.

The *i*th *principal minor* of A is the matrix A_i formed by the first *i* rows and columns of A. So, the first principal minor of A is the matrix $A_i = (a_{11})$, the second principal minor is the matrix:

$$A_2 = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}, \textit{and so on.}$$

Definite Matrices



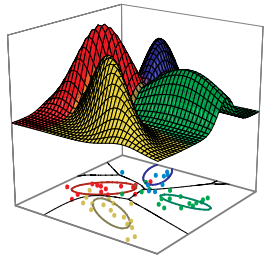
- The matrix A is *positive definite* if all its principal minors $A_1, A_2, \ldots, A_n$ have strictly positive determinants

- If these determinants are non-zero and alternate in signs, starting with $det(A_1)<0$, then the matrix A is *negative definite*

- If the determinants are all non-negative, then the matrix is *positive semi-definite*

- If the determinant alternate in signs, starting with $det(A_1)\leq 0$, then the matrix is *negative semi-definite*

Definite Matrices

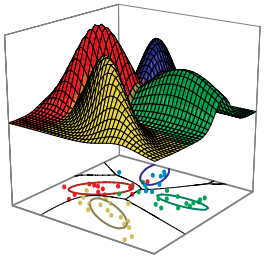


To fix ideas, consider a 2x2 symmetric matrix:

$$A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}.$$

- It is positive definite if:
  a) $det(A_1) = a_{11} > 0$
  b) $det(A_2) = a_{11}a_{22} - a_{12}a_{12} > 0$
- It is negative definite if:
  a) $det(A_1) = a_{11} < 0$
  b) $det(A_2) = a_{11}a_{22} - a_{12}a_{12} > 0$
- It is positive semi-definite if:
  a) $det(A_1) = a_{11} \geq 0$
  b) $det(A_2) = a_{11}a_{22} - a_{12}a_{12} \geq 0$
- And it is negative semi-definite if:
  a) $det(A_1) = a_{11} \leq 0$
  b) $det(A_2) = a_{11}a_{22} - a_{12}a_{12} \geq 0.$

Definite Matrices - Exercise



Check whether the following matrices are positive definite, negative definite, positive semi-definite, negative semi-definite or none of the above.

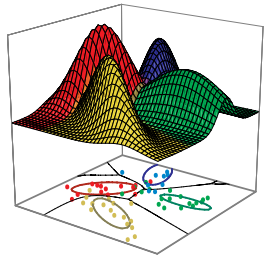
$$(a)\ A = \begin{pmatrix} 2 & 1 \\ 1 & 4 \end{pmatrix}$$

$$(b)\ A = \begin{pmatrix} -2 & 4 \\ 4 & -8 \end{pmatrix}$$

$$(c)\ A = \begin{pmatrix} -2 & 2 \\ 2 & -4 \end{pmatrix}$$

$$(d)\ A = \begin{pmatrix} 2 & 4 \\ 4 & 3 \end{pmatrix}$$

Definite Matrices – Exercise



Solutions:

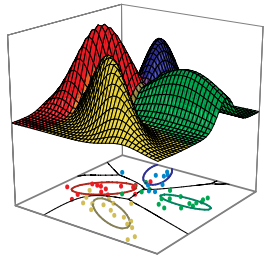
- $A_1 = 2 > 0$
  $A_2 = 8 - 1 = 7 > 0$ $\Rightarrow$ A is positive definite

- $A_1 = -2$
  $A_2 = (-2 \times -8) - 16 = 0$ $\Rightarrow$ A is negative semi-positive

- $A_1 = -2$
  $A_2 = 8 - 4 = 4 > 0$ $\Rightarrow$ A is negative definite

- $A_1 = 2 > 0$
  $A_2 = 6 - 16 = -10 < 0$ $\Rightarrow$ A is none of the above

Probability Theory – Discrete RV



- Let *x* be discrete random variable
  - Can assume any of m values in set $\mathcal{X} = \{v_1, v_2, \ldots, v_m\}$
  - Denote p_i=probability that *x* assumes value v_i:

$$p_i = \Pr[x = v_i], \qquad i = 1, \ldots, m$$

  - Then probabilities p_i must satisfy 2 conditions:

$$P(x) \geq 0, \qquad \text{and} \qquad \sum_{x \in \mathcal{X}} P(x) = 1$$

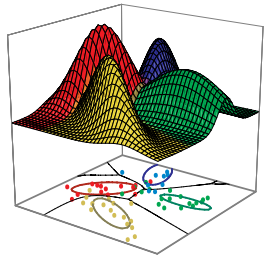
- Expected values
  - aka mean, average of random variable *x* is:

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

  - Linear operation:

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

Probability Theory – Expected Value

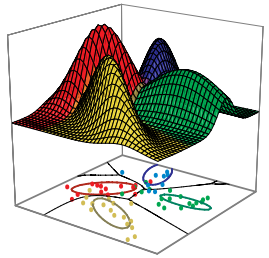


- 2 special expectations:
  - 2nd moment: $\mathcal{E}[x^2] = \sum_{x \in \mathcal{X}} x^2 P(x)$
  - Variance: $\mathrm{Var}[x] = \sigma^2 = \mathcal{E}[(x-\mu)^2] = \sum_{x \in \mathcal{X}} (x-\mu)^2 P(x) = \mathcal{E}[x^2] - (\mathcal{E}[x])^2$
    - Variance is moment of inertia of probability mass func
    - Standard deviation, σ, is measure of how far values of *x* are likely to be from the mean, $\boldsymbol{\mu}$.
    - Chebyshev's innequality: $\Pr[|x-\mu| > n\sigma] \leq \frac{1}{n^2}$
    - Variance is not linear: for y=αx, $\mathrm{Var}[y] = \alpha^2 \mathrm{Var}[x]$

      and Var[**x**+**y**] ≠ Var[**x**] + Var[**y**] *(typically)*
    - When x binary (v_1=0, v_2=1) and *p*=Pr[x=1]

$$\mu = p \qquad \text{and} \qquad \sigma = \sqrt{p(1-p)}$$

Probability Theory – Pairs of discrete RV



- 2 discrete random variables take on values:

$$\mathcal{X} = \{v_1, v_2, \ldots, v_m\} \quad \mathcal{Y} = \{w_1, w_2, \ldots, w_n\}$$

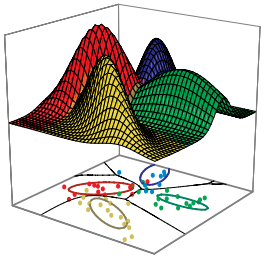
  - Think of (*x*,*y*) pair as a vector or point in product space of x and y.
  - For pair (v_i,w_j) have joint probability: $p_{ij} = \Pr[x = v_i, y = w_j]$
- Joint probability mass function P(*x*,*y*):

$$P(x, y) \geq 0 \quad \text{and} \quad \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} P(x, y) = 1$$

  - P(x,y) is a complete characterization of RV pair (x,y)
- Marginal distributions:

$$P_x(x) = \sum_{y \in \mathcal{Y}} P(x, y) \qquad P_y(y) = \sum_{x \in \mathcal{X}} P(x, y).$$

Probability Theory – Pairs of discrete RV



- *x* and *y* statistically independent iff: $P(x, y) = P_x(x) P_y(y)$
  - i.e. knowing value of x gives no additional info about y
- Expected values of functions of 2 RV's

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

- Still linear: $\mathcal{E}[\alpha_1 f_1(x, y) + \alpha_2 f_2(x, y)] = \alpha_1 \mathcal{E}[f_1(x, y)] + \alpha_2 \mathcal{E}[f_2(x, y)]$

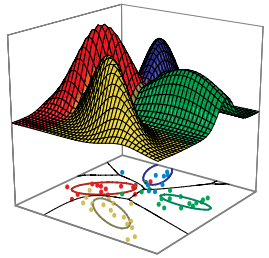
- Means: $\mu_x = \mathcal{E}[x] = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} x P(x, y)$ $\mu_y = \mathcal{E}[y] = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} y P(x, y)$
  - Vector notation: $\boldsymbol{\mu} = \mathcal{E}[\mathbf{x}] = \sum_{\mathbf{x} \in \{\mathcal{X}\mathcal{Y}\}} \mathbf{x} P(\mathbf{x})$

- Variances:

$$\sigma_x^2 = \mathrm{Var}[x] = \mathcal{E}[(x - \mu_x)^2] = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} (x - \mu_x)^2 P(x, y)$$

$$\sigma_y^2 = \mathrm{Var}[y] = \mathcal{E}[(y - \mu_y)^2] = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} (y - \mu_y)^2 P(x, y)$$

Probability Theory – Pairs of discrete RV

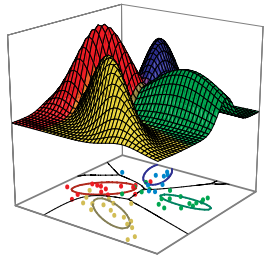


- Covariance:

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

  - Matrix notation: $\boldsymbol{\Sigma} = \mathcal{E}[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^t]$
    - For **x**=(*x*,*y*), $\Sigma = \begin{bmatrix} \sigma_{xx}^2 & \sigma_{xy}^2 \\ \sigma_{xy}^2 & \sigma_{yy}^2 \end{bmatrix}$
  - if σ_{xy}=0, x and y are uncorrelated
    - Not necessarily statistically independent! (true for Normal)
  - Cauchy-Schwartz inneq: $\sigma_{xy}^2 \le \sigma_x^2\sigma_y^2$
  - Correlation coefficient: $\rho = \dfrac{\sigma_{xy}}{\sigma_x\sigma_y}$
    - Normalized -1<*p*<1

Probability Theory – Conditional/Total/Bayes



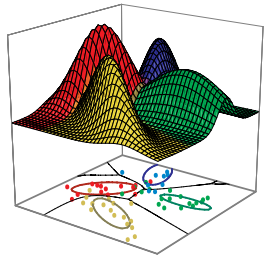
- Conditional Probability
  - Prob $x = v_i$ given that $y = w_i$:

$$\Pr[x = v_i | y = w_j] = \frac{\Pr[x = v_i, y = w_j]}{\Pr[y = w_j]} \quad \text{or} \quad P(x|y) = \frac{P(x, y)}{P(y)}$$

    - If x,y statistically independent, $P(x|y) = P(x)$

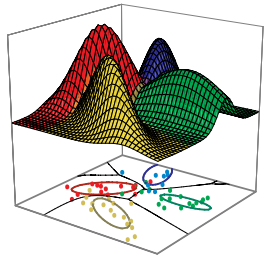
- Total Probability
  - If event A can occur in m mutually exclusive ways A_1, A_2,… A_m, prob of A occurring is sum of prob of all sub-events A_i.
  - For RVs x and y, where x can take on m mutually exclusive values, $P(y) = \sum_{x \in \mathcal{X}} P(x, y)$

Probability Theory – Conditional/Total/Bayes



- Bayes Rule
  - But we already have: $P(x, y) = P(y|x)P(x)$
  - We can write Bayes Rule: $P(x|y) = \frac{P(y|x)P(x)}{\sum_{x \in \mathcal{X}} P(y|x)P(x)}$
    - In other words: $\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{evidence}}$
    - Think of x as a "cause", y as an observable "effect"
    - Easy to compute *likelihood* P(*y*|*x*), easy to compute *prior* P(x), not easy to compute *posterior* P(*x*|*y*)
    - Bayes rule converts from the prior distribution P(*x*) (i.e. before any observations are made) to the posterior distribution P(*x*|*y*).
    - Evidence term P(*y*) in denominator is merely a scaling factor.

Probability Theory – Vector RV's

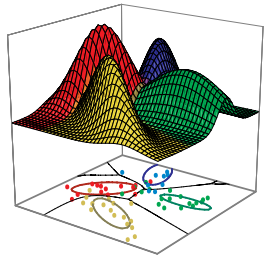


- Direct extension of pairs of RVs
- Use matrix notation. **x** has d RVs $x_1, x_2, \ldots x_d$
- P(**x**) is a complex multidimensional function
- If RVs x_i are indep: $P(\mathbf{x}) = P_{x_1}(x_1)P_{x_2}(x_2)\cdots P_{x_d}(x_d) = \prod_{i=1}^{d} P_{x_i}(x_i)$
- Can sum out unwanted RVs to get marginals
- Bayes holds for Vector RV's $P(\mathbf{x}_1|\mathbf{x}_2) = \dfrac{P(\mathbf{x}_2|\mathbf{x}_1)P(\mathbf{x}_1)}{\sum_{\mathbf{x}_1} P(\mathbf{x}_2|\mathbf{x}_1)P(\mathbf{x}_1)}$

- Expectation is applied to each component of **x**

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

Probability Theory – Vector RV's



- Covariance matrix (square, d x d):

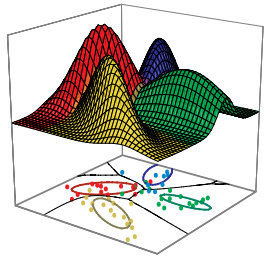
$$\boldsymbol{\Sigma} = \begin{bmatrix} \mathcal{E}[(x_1-\mu_1)(x_1-\mu_1)] & \mathcal{E}[(x_1-\mu_1)(x_2-\mu_2)] & \dots & \mathcal{E}[(x_1-\mu_1)(x_d-\mu_d)] \\ \mathcal{E}[(x_2-\mu_2)(x_1-\mu_1)] & \mathcal{E}[(x_2-\mu_2)(x_2-\mu_2)] & \dots & \mathcal{E}[(x_2-\mu_2)(x_d-\mu_d)] \\ \vdots & \vdots & \ddots & \vdots \\ \mathcal{E}[(x_d-\mu_d)(x_1-\mu_1)] & \mathcal{E}[(x_d-\mu_d)(x_2-\mu_2)] & \dots & \mathcal{E}[(x_d-\mu_d)(x_d-\mu_d)] \end{bmatrix}$$

$$= \begin{bmatrix} \sigma_{11} & \sigma_{12} & \dots & \sigma_{1d} \\ \sigma_{21} & \sigma_{22} & \dots & \sigma_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{d1} & \sigma_{d2} & \dots & \sigma_{dd} \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \dots & \sigma_{1d} \\ \sigma_{21} & \sigma_2^2 & \dots & \sigma_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{d1} & \sigma_{d2} & \dots & \sigma_d^2 \end{bmatrix}.$$

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

- Symmetric; diagonal = variances of individual vars which are nonnegative; off-diagonal are covariances which can be negative or positive.
- *Positive semidefinite* such that all eigenvalues > 0.

Probability Theory – Continuous RV's

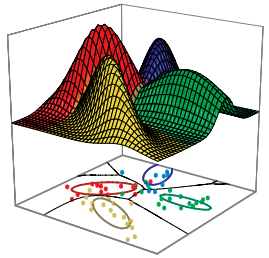


- When x can take on continuous values, no longer talk about prob of x having certain value
- Instead talk about prob of x being in certain range.
  - Prob mass function → *probability density function*

$$\Pr[x \in (a, b)] = \int_a^b p(x)\, dx$$

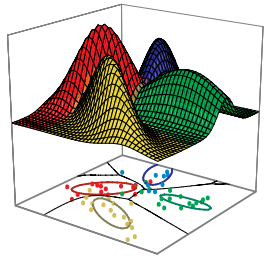
$$p(x) \geq 0 \qquad \text{and} \qquad \int_{-\infty}^{\infty} p(x)\, dx = 1.$$

Probability Theory – Continuous RV's



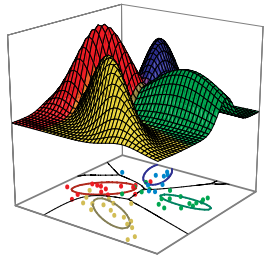
- Most definitions for discrete RV's translate directly with summations replaced by integrals:
  - Expectation $$\mathcal{E}[f(x)] = \int_{-\infty}^{\infty} f(x)p(x)\,dx$$
  - Mean $$\mu = \mathcal{E}[x] = \int_{-\infty}^{\infty} xp(x)\,dx$$
  - Variance $$\mathrm{Var}[x] = \sigma^2 = \mathcal{E}[(x-\mu)^2] = \int_{-\infty}^{\infty} (x-\mu)^2 p(x)\,dx,$$

Probability Theory – Continuous RV's



- Likewise for continuous random vectors **x**
  - Expectation $$\mathcal{E}[\mathbf{f}(\mathbf{x})] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\mathbf{f}(\mathbf{x})p(\mathbf{x})\,dx_1dx_2\cdots dx_d = \int_{-\infty}^{\infty}\mathbf{f}(\mathbf{x})p(\mathbf{x})\,d\mathbf{x}$$
  - Mean $$\boldsymbol{\mu} = \mathcal{E}[\mathbf{x}] = \int_{-\infty}^{\infty}\mathbf{x}p(\mathbf{x})\,d\mathbf{x}$$
  - Variance $$\boldsymbol{\Sigma} = \mathcal{E}[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^t] = \int_{-\infty}^{\infty}(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^t p(\mathbf{x})\,d\mathbf{x}.$$
  - Conditional $$p(x|y) = \frac{p(x,y)}{p(y)}$$
  - Bayes Rule $$p(x|y) = \frac{p(y|x)p(x)}{\int_{-\infty}^{\infty}p(y|x)p(x)\,dx}$$

Sum of independent RVs

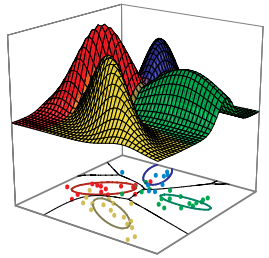


- Know densities for 2 RV's, want to know density of their sum $z=x+y$.
  - Mean: $\mu_z = \mathcal{E}[z] = \mathcal{E}[x+y] = \mathcal{E}[x] + \mathcal{E}[y] = \mu_x + \mu_y,$
  - Var:
$$\begin{aligned}\sigma_z^2 &= \mathcal{E}[(z-\mu_z)^2] = \mathcal{E}[(x+y-(\mu_x+\mu_y))^2] = \mathcal{E}[((x-\mu_x)+(y-\mu_y))^2] \\ &= \mathcal{E}[(x-\mu_x)^2] + 2\underbrace{\mathcal{E}[(x-\mu_x)(y-\mu_y)]}_{=0} + \mathcal{E}[(y-\mu_y)^2] \\ &= \sigma_x^2 + \sigma_y^2,\end{aligned}$$
  - p(z):
$$\Pr[\zeta < z < \zeta + \Delta z] = \left[\int_{-\infty}^{\infty} p(x)\,p(\zeta - x)\,dx\right]\Delta z$$

Density of sum is convolution of densities.

$$p(z) = p_x(x) \star p_y(y) = \int_{-\infty}^{\infty} p_x(x)\,p_y(z-x)\,dx$$

Gaussian Distribution



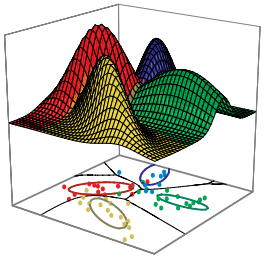
- Central limit theorem:
  - The distribution of the sum of d independent RV's approaches a normal distribution (or Gaussian)
- Important for theoretical & practical reasons

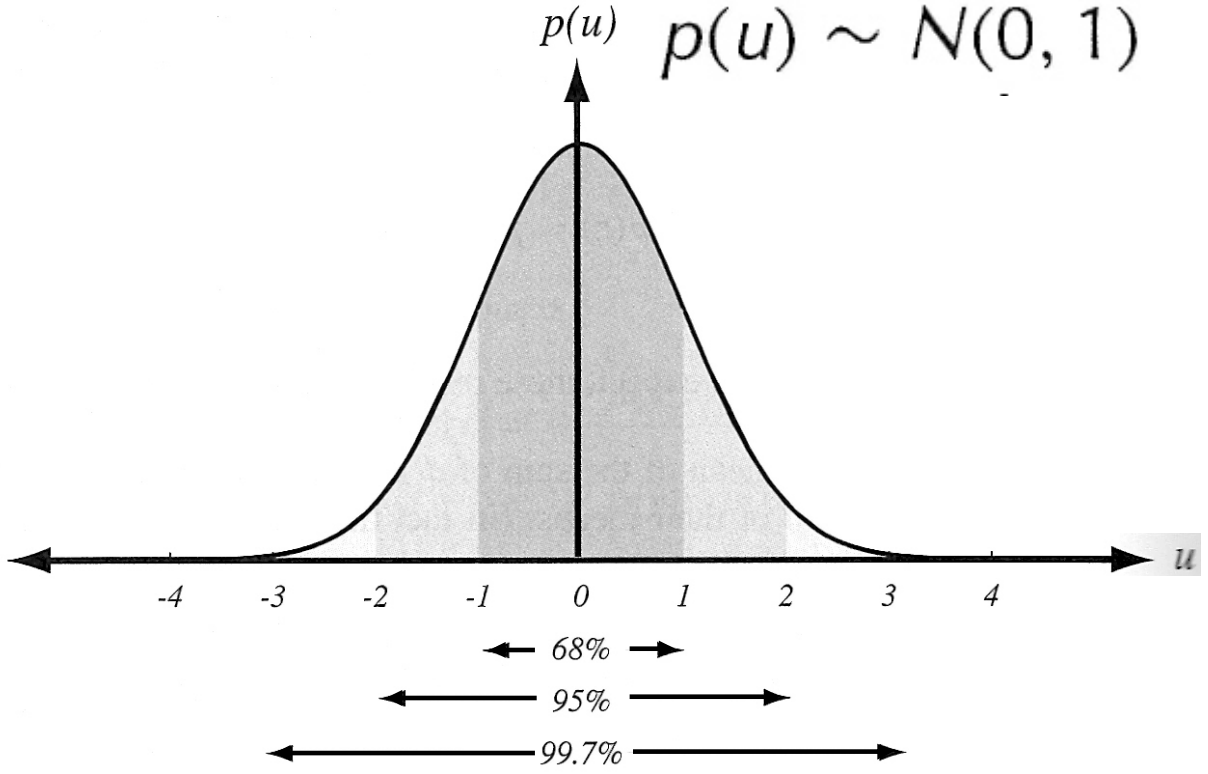
$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-1/2((x-\mu)^2/\sigma^2)} \qquad p(x) \sim N(\mu, \sigma^2)$$

- Bell-shaped curve
- Completely defined by mean μ and variance σ

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

Gaussian Distribution



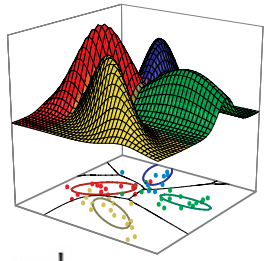


$$\Pr[|x - \mu| \leq \sigma] \simeq 0.68$$

$$\Pr[|x - \mu| \leq 2\sigma] \simeq 0.95$$

$$\Pr[|x - \mu| \leq 3\sigma] \simeq 0.997$$

Gaussian Distribution

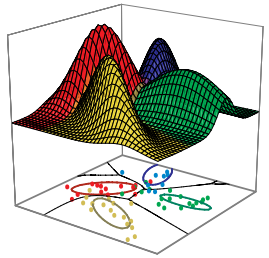


- Mahalanobis distance from x to μ: $r = \frac{|x - \mu|}{\sigma}$
  - Known as z-score for 1D case
  - Distance measured in s.d.'s
- Standardardized RV: $u = \frac{x - \mu}{\sigma} \rightarrow p(u) = \frac{1}{\sqrt{2\pi}} e^{-u^2/2}$
  - RV shifted by mean & scaled by s.d.

Table A.1. The Probability a Sample Drawn from a Standardized Gaussian has Absolute Value Less Than a Criterion (i.e., Pr[$|u| \leq z$])

z	Pr[$\|u\| \leq z$]	z	Pr[$\|u\| \leq z$]	z	Pr[$\|u\| \leq z$]
0.0	0.0	1.0	0.683	2.0	0.954
0.1	0.080	1.1	0.729	2.1	0.964
0.2	0.158	1.2	0.770	2.326	0.980
0.3	0.236	1.3	0.806	2.5	0.989
0.4	0.311	1.4	0.838	2.576	0.990
0.5	0.383	1.5	0.866	3.0	0.9974
0.6	0.452	1.6	0.890	3.090	0.9980
0.7	0.516	1.7	0.911	3.291	0.999
0.8	0.576	1.8	0.928	3.5	0.9995
0.9	0.632	1.9	0.943	4.0	0.99994

Gaussian Derivatives & Integrals

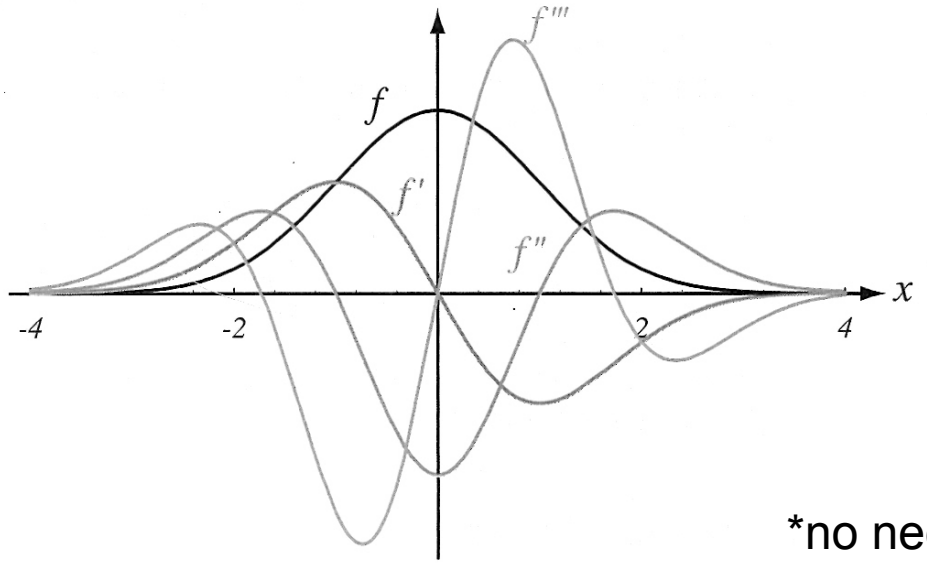


- Used so often in statistical PR, useful to know*:

$$\frac{\partial}{\partial x}\left[\frac{1}{\sqrt{2\pi}\sigma}e^{-x^2/(2\sigma^2)}\right] = \frac{-x}{\sqrt{2\pi}\sigma^3}e^{-x^2/(2\sigma^2)} = \frac{-x}{\sigma^2}p(x)$$

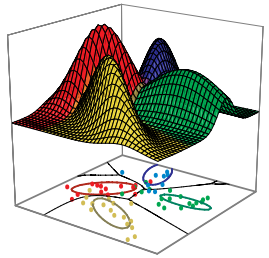
$$\frac{\partial^2}{\partial x^2}\left[\frac{1}{\sqrt{2\pi}\sigma}e^{-x^2/(2\sigma^2)}\right] = \frac{1}{\sqrt{2\pi}\sigma^5}\left(-\sigma^2+x^2\right)e^{-x^2/(2\sigma^2)} = \frac{-\sigma^2+x^2}{\sigma^4}p(x)$$

$$\frac{\partial^3}{\partial x^3}\left[\frac{1}{\sqrt{2\pi}\sigma}e^{-x^2/(2\sigma^2)}\right] = \frac{1}{\sqrt{2\pi}\sigma^7}\left(3x\sigma^2-x^3\right)e^{-x^2/(2\sigma^2)} = \frac{-3x\sigma^2-x^3}{\sigma^6}p(x)$$



*no need to memorize these…

Multivariate Gaussian Distribution



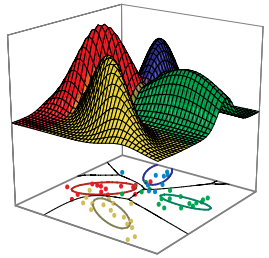
- Convolution of two Gaussians is a Gaussian
  - Therefore distribution of sum of two independent normal RVs is normal.
- Joint density of d indep normal RVs: $p_{x_i}(x_i) \sim N(\mu_i, \sigma_i^2)$

$$p(\mathbf{x}) = \prod_{i=1}^{d} p(x_i) = \prod_{i=1}^{d} \frac{1}{\sqrt{2\pi}\sigma_i} e^{-1/2((x_i-\mu_i)/\sigma_i)^2}$$

$$= \frac{1}{(2\pi)^{d/2}\prod_{i=1}^{d}\sigma_i} \exp\left[-\frac{1}{2}\sum_{i=1}^{d}\left(\frac{x_i-\mu_i}{\sigma_i}\right)^2\right]$$

Continued next slide…

Multivariate Gaussian Distribution



- Joint density of d indep normal RVs: $p_{x_i}(x_i) \sim N(\mu_i, \sigma_i^2)$
  - Indep so covariance matrix is diagonal:

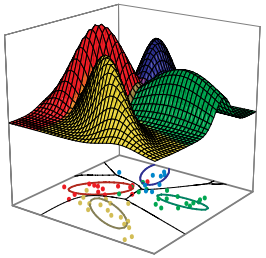
$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_d^2 \end{bmatrix} \qquad \boldsymbol{\Sigma}^{-1} = \begin{bmatrix} 1/\sigma_1^2 & 0 & \dots & 0 \\ 0 & 1/\sigma_2^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1/\sigma_d^2 \end{bmatrix}$$

  - So exponent becomes: $\sum_{i=1}^{d} \left(\frac{x_i - \mu_i}{\sigma_i} \right)^2 = (\mathbf{x} - \boldsymbol{\mu})^t \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})$

  - Noting that |Σ| is product of variances, we get general form for the multivariate normal density function.

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{d/2} |\boldsymbol{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^t \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right]$$

Bivariate Normal Density



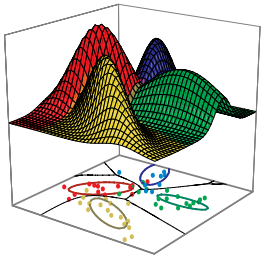
- Let $\rho = \frac{\sigma_{12}}{\sigma_1\sigma_2}$, so $\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}$ and $|\Sigma| = \sigma_1^2\sigma_2^2(1-\rho^2)$
- Inverse covariance matrix is:

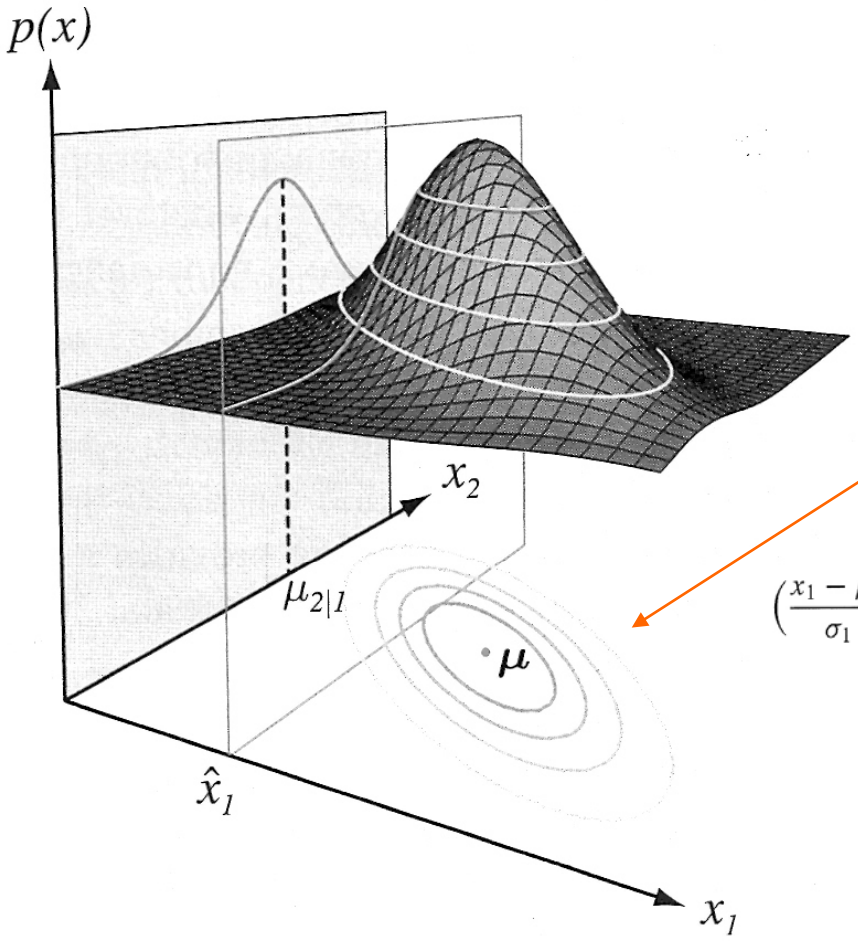
$$\Sigma^{-1} = \frac{1}{\sigma_1^2\sigma_2^2(1-\rho^2)} \begin{bmatrix} \sigma_2^2 & -\rho\sigma_1\sigma_2 \\ -\rho\sigma_1\sigma_2 & \sigma_1^2 \end{bmatrix}$$
$$= \frac{1}{1-\rho^2} \begin{bmatrix} 1/\sigma_1^2 & -\rho/(\sigma_1\sigma_2) \\ -\rho/(\sigma_1\sigma_2) & 1/\sigma_2^2 \end{bmatrix}.$$

- Expand to get:

$$p_{x_1x_2}(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \times \exp\left[-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x_1-\mu_1}{\sigma_1}\right)\left(\frac{x_2-\mu_2}{\sigma_2}\right) + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2\right]\right]$$

Bivariate Normal Density



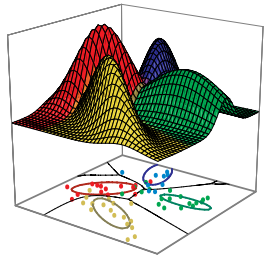


Curves of constant:

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

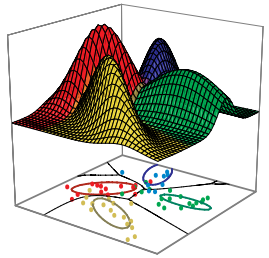
FIGURE A.4. A two-dimensional Gaussian having mean $\boldsymbol{\mu}$ and nondiagonal covariance $\boldsymbol{\Sigma}$. If the value on one variable is known, for instance $x_1 = \hat{x}_1$, the distribution over the other variable is Gaussian with mean $\mu_{2|1}$.

Bivariate Normal Density



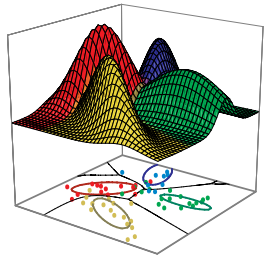
- Peak of hill at $(x_1, x_2) = (\mu_1, \mu_2)$
- Shape of curve determined by σ_{11} and σ_{22} and *p*
  - σ_{11} and σ_{22} determine extent along x_1 and x_2 axis
  - Eccentricity determined by *p*
  - If *p* becomes 1 or -1, curves collapse to lines
    - Distribution is in fact 1D
  - The principle axes are in the directions of the eigenvectors of e_i of Σ, and the widths in these directions are $\sqrt{\lambda_i}$
  - All conditional and marginal probabilities are also normal.

Hypothesis Testing



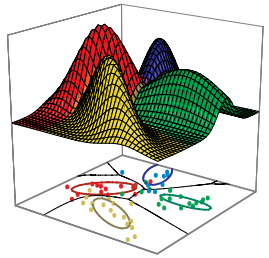
- Provides formal test to decide if results of an experiment are ‘significant’ or ‘accidental’
- Define a Null Hypothesis, H_0
  - E.g. Assume set of n samples drawn from a known distribution D_0.
- Assume H_0 to be true
- Ask whether can ‘reject the null hypothesis’ with some degree of confidence

Hypothesis Testing



- Example:
  - Let D_0 be standardized Gaussian p(x)~N(0,1)
    - H_0 = sample x was drawn from Gaussian with mean μ=0
      - (i.e. assume that x was drawn from D_0)
  - If value of sample is small (e.g. x=0.3), likely came from D_0
  - If value large (e.g. x=5), more confident that x did NOT come from D_0
  - If sample differs from x=0 by some criterion value, then can reject null hypothesis at a given confidence
    - Confidence often 0.01, 0.05
    - Conclude that sample was drawn from Gaussian with $\mu \neq 0$ with some probability/confidence
    - Then say 'difference of sample from 0 is *statistically significant*'
  - For H_0 above, if $|x| \geq 2.576$, can reject H_0 at .01 confidence level
    - (We get this critical value from Table A.1)

Hypothesis Testing

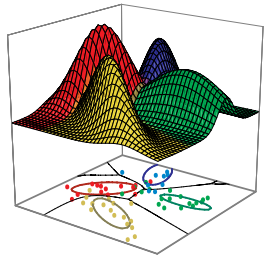


- Several tests available for discrete problems
  - Will see χ^2 applied to testing if 2 variables are indep (Part 3)
  - Will also use hypothesis testing for comparing classifier accuracy to random or to another classifier. (Part 4)

Table A.2. Critical Values of Chi-Square (at Two Confidence levels) for Different Degrees of Freedom (df)

df	.05	.01	df	.05	.01	df	.05	.01
1	3.84	6.64	11	19.68	24.72	21	32.67	38.93
2	5.99	9.21	12	21.03	26.22	22	33.92	40.29
3	7.82	11.34	13	22.36	27.69	23	35.17	41.64
4	9.49	13.28	14	23.68	29.14	24	36.42	42.98
5	11.07	15.09	15	25.00	30.58	25	37.65	44.31
6	12.59	16.81	16	26.30	32.00	26	38.88	45.64
7	14.07	18.48	17	27.59	33.41	27	40.11	46.96
8	15.51	20.09	18	28.87	34.80	28	41.34	48.28
9	16.92	21.67	19	30.14	37.57	29	42.56	49.59
10	18.31	23.21	20	31.41	37.57	30	43.77	50.89

Information Theory

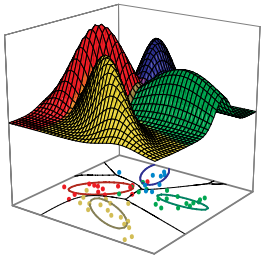


- Entropy – a measure of randomness
  - Assume have a discrete set of symbols $\{v_1, v_2, \ldots v_n\}$
  - Associated probabilities P_i
  - Draw a particular sequence of symbols.
  - Entropy is randomness of unpredictability of that seq:

$$H = -\sum_{i=1}^{m} P_i \log_2 P_i \quad \text{or} \quad H = \mathcal{E}[\log 1/P]$$

    - Where P is random var with values $P_1, P_2, \ldots P_n$
    - Measured in bits for log base 2, 'nats' for *ln*
    - $\log_2(1/P)$ = *'surprise'*
      - If P_i=0 for all but one value, no surprise

Information Theory



- Entropy – a measure of randomness
  - Entropy maximum for uniform distribution ($H=\log_2(n)$ bits)
  - For continuous, Gaussian has max entropy

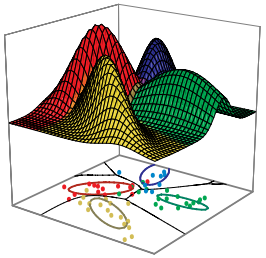
$$(H = 0.5 + \log_2 (\sqrt{2\pi}\sigma) \text{ bits})$$

- Mutual information
  - Reduction in uncertainty about one variable given knowledge of the other variable

$$I(p;\ q) = H(p) - H(p|q) = \sum_{x,y} r(x, y) \log_2 \frac{r(x, y)}{p(x)q(y)},$$

r(x,y) is joint probability of x and y

Information Theory



- Relationship between entropy, conditional entropy, and mutual information:

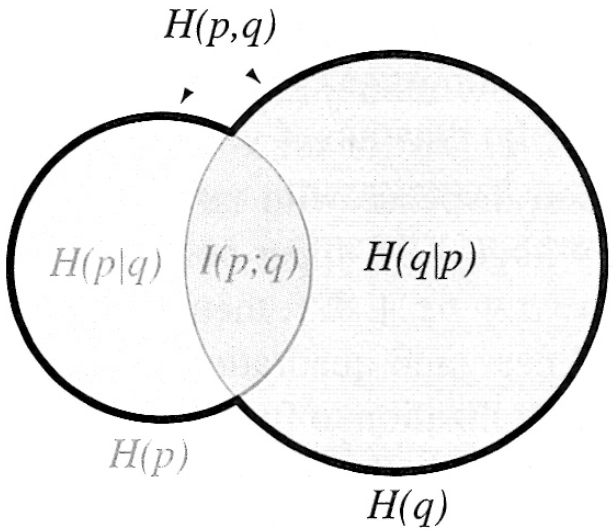


FIGURE A.5. For two distributions p and q, this figure shows the mathematical relationships among the entropy, mutual information $I(p;\ q)$, and conditional entropies $H(p|q)$ and $H(q|p)$. For instance $I(p;\ p) = H(p)$; if $I(p;\ q) = 0$, then $H(q|p) = H(q)$; $H(p, q) = H(p|q) + H(q)$; and so forth.