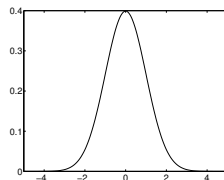


TOOLS: GAUSSIAN DISTRIBUTIONS

Gaussian density in one dimension

$$p(x; \mu, \sigma) := \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right)$$

- μ = expected value of x , σ^2 = variance, σ = standard deviation
- The quotient $\frac{x - \mu}{\sigma}$ measures deviation of x from its expected value in units of σ (i.e. σ defines the length scale)

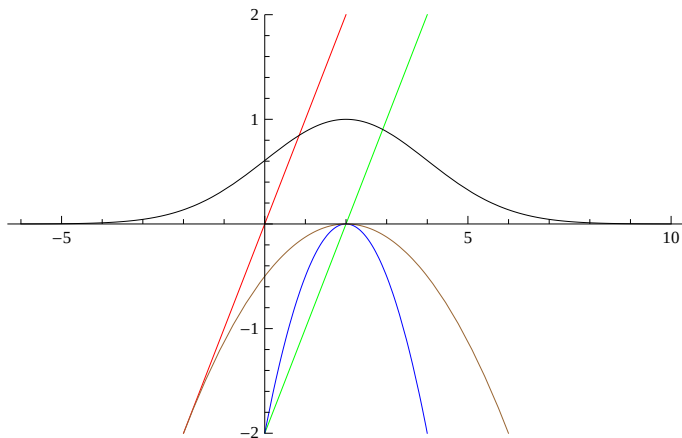


Recall: Standard deviation around the mean

- Recall that the interval $[\mu - \sigma, \mu + \sigma]$ (“one standard deviation”) always contains the same amount of probability mass (ca. 68.27%), regardless of the choice of μ and σ .
- Similarly, the interval $[\mu - 2\sigma, \mu + 2\sigma]$ contains $\sim 95.45\%$ of the mass, and $[\mu - 3\sigma, \mu + 3\sigma]$ contains $\sim 99.73\%$.

COMPONENTS OF A 1d GAUSSIAN

$$\mu = 2, \sigma = 2$$



- Red: $x \mapsto x$

- Green: $x \mapsto x - \mu$

- Blue: $x \mapsto -\frac{1}{2}(x - \mu)^2$

- Brown: $x \mapsto -\frac{1}{2}\left(\frac{x - \mu}{\sigma}\right)^2$

- Black: $x \mapsto \exp\left(-\frac{1}{2}\left(\frac{x - \mu}{\sigma}\right)^2\right)$

Recall: Covariance

The covariance of two random variables X_1, X_2 is

$$\text{Cov}[X_1, X_2] = \mathbb{E}[(X_1 - \mathbb{E}[X_1])(X_2 - \mathbb{E}[X_2])] .$$

If $X_1 = X_2$, the covariance is the variance: $\text{Cov}[X, X] = \text{Var}[X]$.

Covariance matrix

If $\mathbf{X} = (X_1, \dots, X_d)$ is a random vector with values in $\mathbb{R}^d$, the matrix of all covariances

$$\text{Cov}[\mathbf{X}] := (\text{Cov}[X_i, X_j])_{i,j} = \begin{pmatrix} \text{Cov}[X_1, X_1] & \cdots & \text{Cov}[X_1, X_d] \\ \vdots & & \vdots \\ \text{Cov}[X_d, X_1] & \cdots & \text{Cov}[X_d, X_d] \end{pmatrix}$$

is called the **covariance matrix** of $\mathbf{X}$.

Notation

It is customary to denote the covariance matrix $\text{Cov}[\mathbf{X}]$ by Σ .

GAUSSIAN IN MULTIPLE DIMENSIONS

Gaussian density in d dimensions

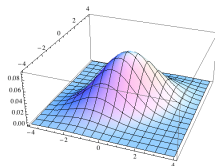
The quadratic function

$$-\frac{(x - \mu)^2}{2\sigma^2} = -\frac{1}{2}(x - \mu)(\sigma^2)^{-1}(x - \mu)$$

is replaced by a quadratic form:

$$p(\mathbf{x}; \mu, \Sigma) := \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} \exp\left(-\frac{1}{2} \left\langle (\mathbf{x} - \mu), \Sigma^{-1}(\mathbf{x} - \mu) \right\rangle\right)$$

for a positive definite matrix Σ .



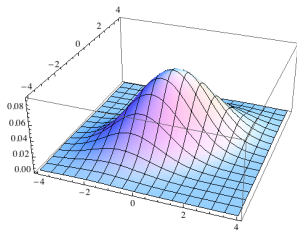
Covariance matrix of a Gaussian

If a random vector $\mathbf{X} \in \mathbb{R}^d$ has Gaussian distribution with density $p(\mathbf{x}; \mu, \Sigma)$, its covariance matrix is $\text{Cov}[\mathbf{X}] = \Sigma$. In other words, a Gaussian is parameterized by its covariance and its mean.

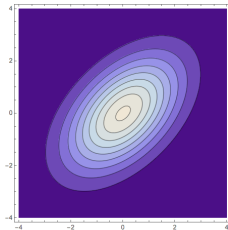
Assuming a multivariate Gaussian model means we assume that all stochastic dependence between dimensions is captured by the covariance.

GAUSSIAN DENSITY: EXAMPLE

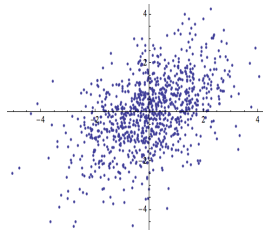
$$p(\mathbf{x}; \mu, \Sigma) \quad \text{with} \quad \mu = (0, 0) \quad \text{and} \quad \Sigma = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$$



density

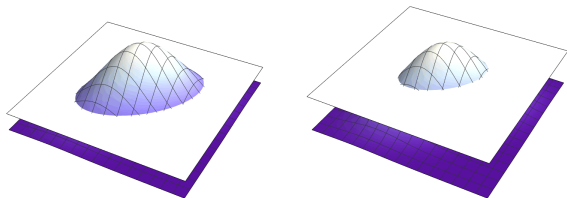


contour lines

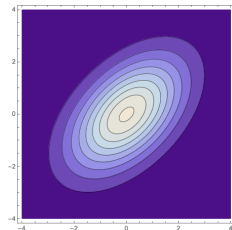


1000 sample points

CONTOUR LINES



Intersect density with a horizontal plane, draw intersection as a curve, and project it down onto the plane.



Each elliptical line is such a contour, for planes at different heights.

Contours and standard deviation

- Each ellipse consists of all points $\mathbf{x} \in \mathbb{R}^2$ that satisfy the equation

$$\langle \mathbf{x}, \Sigma^{-1} \mathbf{x} \rangle = c \quad \text{for some fixed } c > 0 .$$

Changing c changes the size of the ellipse.

- The ellipses play the same role as intervals around the mean for 1d Gaussians: The ellipse with $\langle \mathbf{x}, \Sigma^{-1} \mathbf{x} \rangle = 1$ contains $\sim 39.3\%$ of the probability mass (the value corresponding to ~ 68.26 in 1 dimension), etc.
- That is: The area within the ellipse given by $\langle \mathbf{x}, \Sigma^{-1} \mathbf{x} \rangle = k^2$ corresponds to k standard deviations.

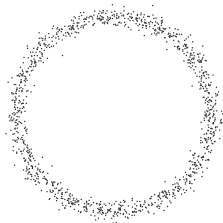
HIGH DIMENSIONS

In high dimensions, there is a counter-intuitive effect. The Gaussians samples we have seen look like clouds that are densest around the mean.

In high dimensions, what happens if we plot distance from the mean looks more like this:



actual Gaussian sample in
 $d = 2$ dimensions



conceptual illustration of Gaussian
sample for d large

The Gaussian density function in high dimensions looks (again conceptually) as in low dimensions, with its mode at the mean. The underlying Euclidean volume, however, scales differently.

EUCLIDEAN VOLUME

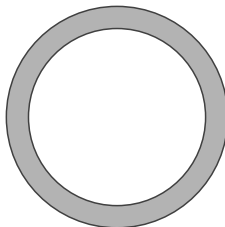
Here are balls of radius 1, in $d = 1, 2$ and 3 dimensions. Shaded in gray is the area between radii 0.8 and 1:

$d = 1$



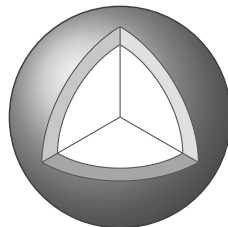
20%

$d = 2$



36%

$d = 3$



48.8%

In a high-dimensional Euclidean ball, almost all mass is close to the surface.

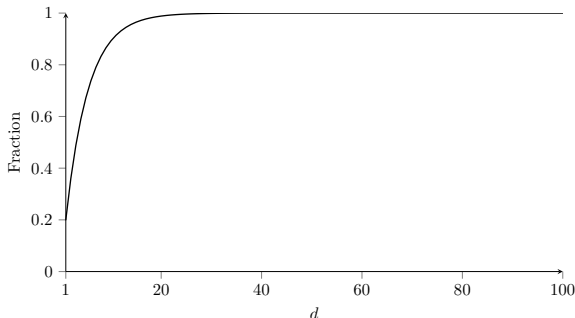
EFFECT ON DENSITIES

- A density is defined with respect to an underlying measure or volume. It rescales this volume locally.
- Although the Gaussian density in high dimensions peaks at the mean, its effect on probability mass gets overpowered by the way volume scales.
- Since this is an effect of the volume, it affects other densities as well, not just Gaussians.

Effect of dimension

To give you an idea, here is how the gray volume ratio on the previous slide scales as dimension increases beyond what we can plot:

d	Gray fraction
1	20.00%
2	36.00%
3	48.80%
4	59.04%
5	67.23%
6	73.79%
7	79.03%
8	83.22%
9	86.58%
10	89.26%



TOOLS:

SPECTRA AND GEOMETRY OF GAUSSIANS

EIGENVALUES

The properties of covariance matrices are summarized by their spectral properties (their eigenvalues and eigenvectors). That makes spectral properties the key to understanding Gaussian distributions.

Recall the definition

We consider a square matrix $A \in \mathbb{R}^{d \times d}$.

A non-zero vector $\xi \in \mathbb{R}^d$ is an **eigenvector** of A if there is a scalar λ such that

$$A\xi = \lambda\xi .$$

λ is called an **eigenvalue** of A for the eigenvector ξ .

In words: The direction of ξ does not change under application of A . Only its length changes, by a factor λ .

The set of eigenvalues of a matrix is called its **spectrum**.

Properties

- In general, eigenvalues are complex numbers $\lambda \in \mathbb{C}$.
- The class of matrices with the nicest eigen-structure are symmetric matrices, for which all eigenvalues are real and the eigenvectors are mutually orthogonal.

EIGENSTRUCTURE OF SYMMETRIC MATRICES

If a matrix is symmetric:

- There are d mutually orthogonal eigendirections.
- Eigenvectors for distinct eigenvalues are orthogonal. Eigenvectors for the same eigenvalue can be chosen orthogonally.
- If $\text{rank}(A) = d$, the eigenvectors form an orthogonal basis of $\mathbb{R}^d$. We can normalize each eigenvector (which again produces an eigenvector) and obtain an *orthonormal* basis of $\mathbb{R}^d$.

Definiteness

type	if ...
positive definite	all eigenvalues > 0
positive semi-definite	all eigenvalues ≥ 0
negative semi-definite	all eigenvalues ≤ 0
negative definite	all eigenvalues < 0
indefinite	none of the above

ORTHONORMAL BASES

Recall that a basis $\{\mathbf{v}_1, \dots, \mathbf{v}_d\}$ of $\mathbb{R}^d$ is called an **orthonormal basis (ONB)** if

$$\langle \mathbf{v}_i, \mathbf{v}_j \rangle = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}$$

That is: The $\mathbf{v}_i$ are pairwise orthogonal and each of length 1.

Orthogonal matrices

A matrix is orthogonal precisely if its rows form an ONB. Any two ONBs can be transformed into each other by an orthogonal matrix.

Transforming between ONBs

If $\mathcal{V} = \{\mathbf{v}_1, \dots, \mathbf{v}_d\}$ and $\mathcal{W} = \{\mathbf{w}_1, \dots, \mathbf{w}_d\}$ are ONBs, there is an orthogonal matrix O such that

$$A_{[\mathcal{V}]} = OA_{[\mathcal{W}]}O^{-1}$$

for any matrix A . By $A_{[\mathcal{V}]}$, we denote the representation of A in $\mathcal{V}$.

EIGENVECTOR ONB

Setting

- Suppose A symmetric, $\xi_1, \dots, \xi_d$ are eigenvectors and form an ONB.
- $\lambda_1, \dots, \lambda_d$ are the corresponding eigenvalues.

How does A act on a vector $\mathbf{v} \in \mathbb{R}^d$?

1. Represent $\mathbf{v}$ in basis $\xi_1, \dots, \xi_d$:

$$\mathbf{v} = \sum_{j=1}^d v_j^A \xi_j \quad \text{where } v_j^A \in \mathbb{R}$$

2. Multiply by A : Eigenvector definition (recall: $A\xi_j = \lambda_j \xi_j$) yields

$$A\mathbf{v} = A\left(\sum_{j=1}^d v_j^A \xi_j\right) = \sum_{j=1}^d v_j^A A\xi_j = \sum_{j=1}^d v_j^A \lambda_j \xi_j$$

A symmetric matrix acts by scaling along the directions of its eigenvectors.

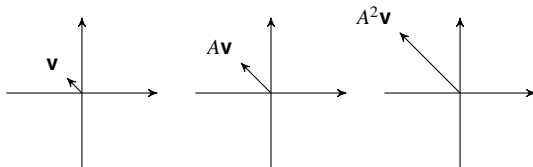
ILLUSTRATION

Suppose we *repeatedly* apply a symmetric matrix A to a vector $\mathbf{v} \in \mathbb{R}^d$: We compute

$$A\mathbf{v}, \quad A(A\mathbf{v}) = A^2\mathbf{v}, \quad A(A(A\mathbf{v})) = A^3\mathbf{v}, \quad \dots$$

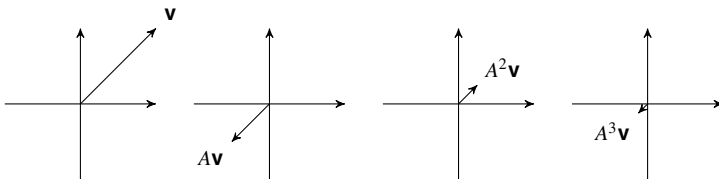
How does $\mathbf{v}$ change?

Example 1: $\mathbf{v}$ is an eigenvector with eigenvalue 2



The direction of $\mathbf{v}$ does not change, but its length doubles with each application of A .

Example 2: $\mathbf{v}$ is an eigenvector with eigenvalue $-\frac{1}{2}$



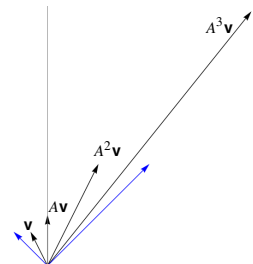
EFFECT OF A SYMMETRIC MATRIX

If $\mathbf{v}$ is an arbitrary vector, we can represent it as a linear combination $\mathbf{v} = \sum_j v_j^A \xi_j$ of eigenvectors of A , and obtain

$$A^n \mathbf{v} = \sum_{j=1}^d v_j^A \lambda_j^n \xi_j$$

The weight λ_j^n grows most rapidly for eigenvalue with largest absolute value.

The direction of $A^n \mathbf{v}$ converges to the direction of the eigenvector with largest absolute eigenvalue as n grows large.



$$A = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} \quad \mathbf{v} = \begin{pmatrix} -\frac{1}{2} \\ 1 \end{pmatrix}$$

eigenvectors shown in blue
length proportional to eigenvalue

QUADRATIC FORMS

In applications, symmetric matrices often occur in quadratic forms.

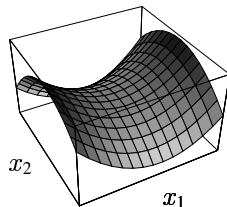
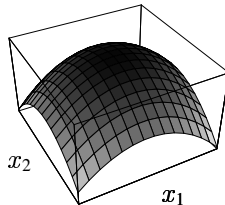
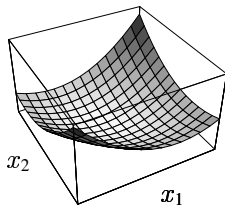
Definition

The **quadratic form** defined by a symmetric matrix A is the function

$$\begin{aligned} q_A : \mathbb{R}^d &\rightarrow \mathbb{R} \\ \mathbf{x} &\mapsto \langle \mathbf{x}, A\mathbf{x} \rangle \end{aligned}$$

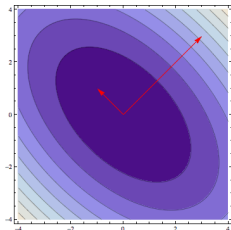
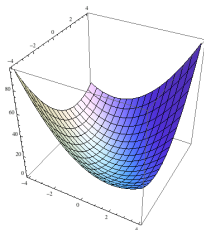
Intuition

A quadratic form is the d -dimensional analogue of a quadratic function ax^2 , with a vector substituted for the scalar x and the matrix A substituted for the scalar $a \in \mathbb{R}$.



QUADRATIC FORMS

Here is the quadratic form for the matrix $A = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$:

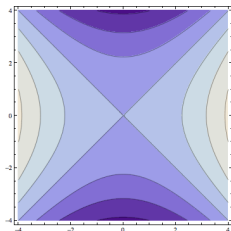
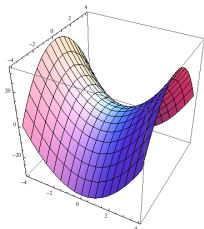


- Left: The function value q_A is graphed on the vertical axis.
- Right: Contours. Each line in $\mathbb{R}^2$ corresponds to a constant function value of q_A . Dark colour = small values.
- The red lines are eigenvector directions of A . Their lengths represent the (absolute) values of the eigenvalues.
- In this case, both eigenvalues are positive. If all eigenvalues are positive, the contours are ellipses. So:

positive/negative definite matrices $\leftrightarrow$ elliptic quadratic forms

QUADRATIC FORMS

In this plot, the eigenvectors are axis-parallel, and one eigenvalue is negative:



The matrix here is $A = \begin{pmatrix} 2 & 0 \\ 0 & -2 \end{pmatrix}$.

Intuition

- If we change the sign of one of the eigenvalue, the quadratic function along the corresponding eigen-axis flips.
- There is a point which is a minimum of the function along one axis direction, and a maximum along the other. Such a point is called a *saddle point*.

APPLICATION: COVARIANCE MATRIX

Reminder

If $\mathbf{X} = (X_1, \dots, X_d)$ is a random vector with values in $\mathbb{R}^d$, the covariance matrix is

$$\text{Cov}[\mathbf{X}] := (\text{Cov}[X_i, X_j])_{i,j \leq d} = \begin{pmatrix} \text{Cov}[X_1, X_1] & \cdots & \text{Cov}[X_1, X_d] \\ \vdots & & \vdots \\ \text{Cov}[X_d, X_1] & \cdots & \text{Cov}[X_d, X_d] \end{pmatrix}$$

where

$$\text{Cov}[X_1, X_2] = \mathbb{E}[(X_1 - \mathbb{E}[X_1])(X_2 - \mathbb{E}[X_2])] .$$

Covariance matrix of a Gaussian

If a random vector $\mathbf{X} \in \mathbb{R}^d$ has Gaussian distribution with density $p(\mathbf{x}; \mu, \Sigma)$, its covariance matrix is $\text{Cov}[\mathbf{X}] = \Sigma$. In other words, a Gaussian is parameterized by its covariance.

Observation

Since $\text{Cov}[X_i, X_j] = \text{Cov}[X_j, X_i]$, the covariance matrix is symmetric.

What is the eigenstructure of Σ ?

- We know: Σ symmetric $\Rightarrow$ there is an eigenvector ONB
- Call the eigenvectors in this ONB $\xi_1, \dots, \xi_d$ and their eigenvalues $\lambda_1, \dots, \lambda_d$
- We can rotate the coordinate system to $\xi_1, \dots, \xi_d$. In the new coordinate system, Σ has the form

$$\Sigma_{[\xi_1, \dots, \xi_d]} = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_d \end{pmatrix} = \text{diag}(\lambda_1, \dots, \lambda_d)$$

EXAMPLE

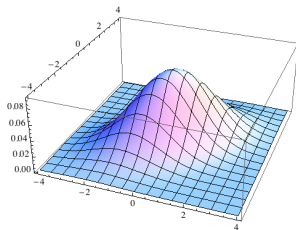
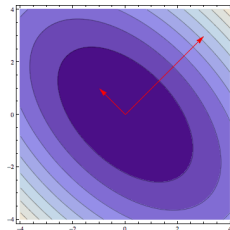
Quadratic form

$$\langle \mathbf{x}, \Sigma \mathbf{x} \rangle \quad \text{with} \quad \Sigma = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$$

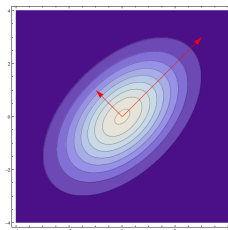
The eigenvectors are $(1, 1)$ and $(-1, 1)$ with eigenvalues 3 and 1.

Gaussian density

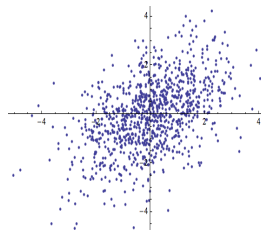
$p(\mathbf{x}; \mu, \Sigma)$ with $\mu = (0, 0)$.



density



density contour



1000 sample points

The coordinates X'_i as random variables

Write $\mathbf{e}_1, \dots, \mathbf{e}_d$ for the ONB of axis vectors. We can represent each ξ_i as

$$\xi_i = \sum_{j=1}^d \alpha_{ij} \mathbf{e}_j$$

Then $O = (\alpha_{ij})$ is the orthogonal transformation matrix between the two bases.

We can represent random vector $\mathbf{X} \in \mathbb{R}^d$ sampled from the Gaussian in the eigen-ONB as

$$\mathbf{X}_{[\xi_1, \dots, \xi_d]} = (X'_1, \dots, X'_d) \quad \text{with} \quad X'_i = \sum_{j=1}^d \alpha_{ij} X_j$$

Since the X_j are random variables (and the α_{ij} are fixed), each X'_i is a scalar random variable.

INTERPRETATION

Meaning of the random variables X'_i

For any Gaussian $p(\mathbf{x}; \mu, \Sigma)$, we can

1. shift the origin of the coordinate system into μ
2. rotate the coordinate system to the eigen-ONB of Σ .

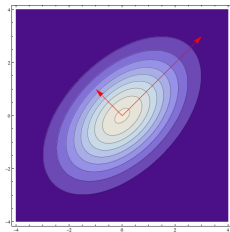
In this new coordinate system, the Gaussian has covariance matrix

$$\Sigma_{[\xi_1, \dots, \xi_d]} = \text{diag}(\lambda_1, \dots, \lambda_d)$$

where λ_i are the eigenvalues of Σ .

Gaussian in the new coordinates

A Gaussian vector $\mathbf{X}_{[\xi_1, \dots, \xi_d]}$ represented in the new coordinates consists of d independent 1D Gaussian variables X'_i . Each X'_i has mean 0 and variance λ_i .



A multidimensional Gaussian consists of independent, orthogonal scalar normal variables in *some* coordinate system.

MLE FOR GAUSSIANS

Consider data $\mathcal{D} = (\mathbf{x}_1, \dots, \mathbf{x}_n)$ in $\mathbb{R}^d$, and the Gaussian model

$$\mathcal{M} = \{g(\cdot | \mu, \Sigma) \mid \mu \in \mathbb{R}^d, \Sigma \in \mathbb{R}^{d \times d} \text{ positive definite} \}$$

where g denotes the Gaussian density on $\mathbb{R}^d$.

Ansatz

ML estimation treats the two parameters separately. I will write θ to denote either $\theta = \mu$ (for fixed Σ), or $\theta = \Sigma$ (for fixed μ), depending on which parameter we estimate.

To find the maximum likelihood estimator, we have to solve

$$0 = \nabla_{\theta} \log \prod_{i=1}^n g(\mathbf{x}_i | \theta) = \nabla_{\theta} \sum_{i=1}^n \log \left(\frac{1}{\sqrt{(2\pi)^d |\Sigma|}} \right) - \frac{1}{2} (\mathbf{x}_i - \mu)^{\top} \Sigma^{-1} (\mathbf{x}_i - \mu)$$

Maximum likelihood estimators

The solutions are

$$\hat{\mu}_{\text{ML}} := \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i \quad \text{and} \quad \hat{\Sigma}_{\text{ML}} := \frac{1}{n} \sum_{i=1}^n (\mathbf{x}_i - \mu)(\mathbf{x}_i - \mu)^{\top}.$$

Note $\hat{\mu}$ does not depend on Σ . We can estimate $\hat{\mu}$ first, then plug in $\hat{\mu}$ for μ in $\hat{\Sigma}_{\text{ML}}$.

REFRESHER – MATRIX DERIVATIVES OF SCALAR FORMS

We will use the following facts:

$$\mathbf{x}^\top \mathbf{A} \mathbf{y} = \mathbf{y}^\top \mathbf{A}^\top \mathbf{x} = \text{Tr} [\mathbf{x}^\top \mathbf{A} \mathbf{y}] \quad (\text{scalars equal their own transpose and trace})$$

$$\text{Tr} [A] = \text{Tr} [A^\top] \quad \text{Tr} [ABC] = \text{Tr} [CAB] = \text{Tr} [BCA]$$

$$\begin{aligned} \frac{\partial}{\partial A_{ij}} \text{Tr} [A^\top B] &= \frac{\partial}{\partial A_{ij}} \sum_n [A^\top B]_{nn} = \frac{\partial}{\partial A_{ij}} \sum_n \sum_m A_{nm}^\top B_{mn} = \frac{\partial}{\partial A_{ij}} \sum_{mn} A_{mn} B_{mn} = B_{ij} \\ \Rightarrow \frac{\partial}{\partial A} \text{Tr} [A^\top B] &= B \end{aligned}$$

$$\begin{aligned} \frac{\partial}{\partial A} \text{Tr} [A^\top BAC] &= \frac{\partial}{\partial A} \text{Tr} [F_1(A)^\top B F_2(A) C] \quad \text{with } F_1 \text{ and } F_2 \text{ both identity maps} \\ &= \frac{\partial}{\partial F_1} \text{Tr} [F_1^\top B F_2 C] \frac{\partial F_1}{\partial A} + \frac{\partial}{\partial F_2} \text{Tr} [F_1^\top B F_2 C] \frac{\partial F_2}{\partial A} \\ &= \frac{\partial}{\partial F_1} \text{Tr} [F_1^\top B F_2 C] \frac{\partial F_1}{\partial A} + \frac{\partial}{\partial F_2} \text{Tr} [C F_1^\top B F_2] \frac{\partial F_2}{\partial A} \\ &= \frac{\partial}{\partial F_1} \text{Tr} [F_1^\top B F_2 C] \frac{\partial F_1}{\partial A} + \frac{\partial}{\partial F_2} \text{Tr} [F_2^\top B^\top F_1 C^\top] \frac{\partial F_2}{\partial A} \\ &= B F_2 C + B^\top F_1 C^\top = BAC + B^\top A C^\top \end{aligned}$$

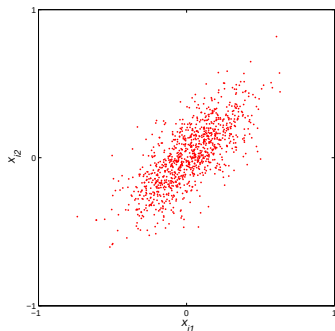
$$\begin{aligned} \frac{\partial}{\partial A_{ij}} \log |A| &= \frac{1}{|A|} \frac{\partial}{\partial A_{ij}} |A| = \frac{1}{|A|} \frac{\partial}{\partial A_{ij}} \sum_k (-1)^{i+k} A_{ik} |A|_{ik} = \frac{1}{|A|} (-1)^{i+j} |A|_{ij} \\ \Rightarrow \frac{\partial}{\partial A} \log |A| &= (A^{-1})^\top \end{aligned}$$

$$\begin{aligned}\frac{\partial(-\ell)}{\partial\mu} &= \frac{\partial}{\partial\mu} \left[\frac{n}{2} \log |2\pi\Sigma| + \frac{1}{2} \sum_i (\mathbf{x}_i - \mu)^\top \Sigma^{-1} (\mathbf{x}_i - \mu) \right] \\&= \frac{1}{2} \sum_i \frac{\partial}{\partial\mu} \left[(\mathbf{x}_i - \mu)^\top \Sigma^{-1} (\mathbf{x}_i - \mu) \right] \\&= \frac{1}{2} \sum_i \frac{\partial}{\partial\mu} \left[\mathbf{x}_i^\top \Sigma^{-1} \mathbf{x}_i + \mu^\top \Sigma^{-1} \mu - 2\mu^\top \Sigma^{-1} \mathbf{x}_i \right] \\&= \frac{1}{2} \sum_i \left(\frac{\partial}{\partial\mu} \left[\mu^\top \Sigma^{-1} \mu \right] - 2 \frac{\partial}{\partial\mu} \left[\mu^\top \Sigma^{-1} \mathbf{x}_i \right] \right) \\&= \frac{1}{2} \sum_i \left(2\Sigma^{-1} \mu - 2\Sigma^{-1} \mathbf{x}_i \right) \\&= n\Sigma^{-1} \mu - \Sigma^{-1} \sum_i \mathbf{x}_i \\&= 0 \Rightarrow \hat{\mu} = \frac{1}{n} \sum_i \mathbf{x}_i\end{aligned}$$

$$\begin{aligned}\frac{\partial(-\ell)}{\partial \Sigma^{-1}} &= \frac{\partial}{\partial \Sigma^{-1}} \left[\frac{n}{2} \log |2\pi \Sigma| + \frac{1}{2} \sum_i (\mathbf{x}_i - \mu)^\top \Sigma^{-1} (\mathbf{x}_i - \mu) \right] \\&= \frac{\partial}{\partial \Sigma^{-1}} \left[\frac{n}{2} \log |2\pi I| \right] - \frac{\partial}{\partial \Sigma^{-1}} \left[\frac{n}{2} \log |\Sigma^{-1}| \right] \\&\quad + \frac{1}{2} \sum_i \frac{\partial}{\partial \Sigma^{-1}} \left[(\mathbf{x}_i - \mu)^\top \Sigma^{-1} (\mathbf{x}_i - \mu) \right] \\&= -\frac{n}{2} \Sigma^\top + \frac{1}{2} \sum_i (\mathbf{x}_i - \mu)(\mathbf{x}_i - \mu)^\top \\&= 0 \Rightarrow \hat{\Sigma} = \frac{1}{n} \sum_i (\mathbf{x}_i - \mu)(\mathbf{x}_i - \mu)^\top\end{aligned}$$

MULTIVARIATE LINEAR REGRESSION

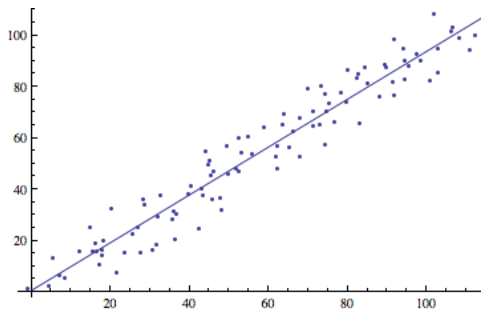
The relationship between variables can also be modelled as a **conditional** distribution.



- data $\mathcal{D} = \{(\mathbf{x}_1, \mathbf{y}_1) \dots, (\mathbf{x}_n, \mathbf{y}_n)\}$
- each $\mathbf{x}_i$ ($\mathbf{y}_i$) is a vector of d_x (d_y) features,
- $\mathbf{y}_i$ is conditionally independent of all else, given $\mathbf{x}_i$.

A simple form of supervised (predictive) learning: model $\mathbf{y}$ as a **linear** function of $\mathbf{x}$, with **Gaussian** noise.

$$p(\mathbf{y}|\mathbf{x}, \mathbf{W}, \Sigma_y) = |2\pi\Sigma_y|^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2}(\mathbf{y} - \mathbf{W}\mathbf{x})^\top \Sigma_y^{-1}(\mathbf{y} - \mathbf{W}\mathbf{x}) \right\}$$



MULTIVARIATE LINEAR REGRESSION – ML ESTIMATE

ML estimates are obtained by maximizing the (conditional) likelihood, as before:

$$\begin{aligned}\ell &= \sum_i \log p(\mathbf{y}_i | \mathbf{x}_i, \mathbf{W}, \Sigma_y) \\&= -\frac{n}{2} \log |2\pi \Sigma_y| - \frac{1}{2} \sum_i (\mathbf{y}_i - \mathbf{W} \mathbf{x}_i)^\top \Sigma_y^{-1} (\mathbf{y}_i - \mathbf{W} \mathbf{x}_i) \\ \frac{\partial(-\ell)}{\partial \mathbf{W}} &= \frac{\partial}{\partial \mathbf{W}} \left[\frac{n}{2} \log |2\pi \Sigma_y| + \frac{1}{2} \sum_i (\mathbf{y}_i - \mathbf{W} \mathbf{x}_i)^\top \Sigma_y^{-1} (\mathbf{y}_i - \mathbf{W} \mathbf{x}_i) \right] \\&= \frac{1}{2} \sum_i \frac{\partial}{\partial \mathbf{W}} \left[(\mathbf{y}_i - \mathbf{W} \mathbf{x}_i)^\top \Sigma_y^{-1} (\mathbf{y}_i - \mathbf{W} \mathbf{x}_i) \right] \\&= \frac{1}{2} \sum_i \frac{\partial}{\partial \mathbf{W}} \left[\mathbf{y}_i^\top \Sigma_y^{-1} \mathbf{y}_i + \mathbf{x}_i^\top \mathbf{W}^\top \Sigma_y^{-1} \mathbf{W} \mathbf{x}_i - 2 \mathbf{x}_i^\top \mathbf{W}^\top \Sigma_y^{-1} \mathbf{y}_i \right] \\&= \frac{1}{2} \sum_i \left[\frac{\partial}{\partial \mathbf{W}} \text{Tr} \left[\mathbf{W}^\top \Sigma_y^{-1} \mathbf{W} \mathbf{x}_i \mathbf{x}_i^\top \right] - 2 \frac{\partial}{\partial \mathbf{W}} \text{Tr} \left[\mathbf{W}^\top \Sigma_y^{-1} \mathbf{y}_i \mathbf{x}_i^\top \right] \right] \\&= \frac{1}{2} \sum_i \left[2 \Sigma_y^{-1} \mathbf{W} \mathbf{x}_i \mathbf{x}_i^\top - 2 \Sigma_y^{-1} \mathbf{y}_i \mathbf{x}_i^\top \right] \\&= 0 \quad \Rightarrow \quad \hat{\mathbf{W}} = \sum_i \mathbf{y}_i \mathbf{x}_i^\top \left(\sum_i \mathbf{x}_i \mathbf{x}_i^\top \right)^{-1}\end{aligned}$$

MULTIVARIATE LINEAR REGRESSION – POSTERIOR

Let y_i be scalar (so that $\mathbf{W}$ is a row vector) and write $\mathbf{w}$ for the column vector of weights.

A conjugate prior for $\mathbf{w}$ is

$$P(\mathbf{w}|\mathbf{A}) = \mathcal{N}(\mathbf{0}, \mathbf{A}^{-1})$$

Then the **log** posterior on $\mathbf{w}$ is

$$\begin{aligned}\log P(\mathbf{w}|\mathcal{D}, \mathbf{A}, \sigma_y) &= \log P(\mathcal{D}|\mathbf{w}, \mathbf{A}, \sigma_y) + \log P(\mathbf{w}|\mathbf{A}, \sigma_y) - \log P(\mathcal{D}|\mathbf{A}, \sigma_y) \\&= -\frac{1}{2} \mathbf{w}^\top \mathbf{A} \mathbf{w} - \frac{1}{2} \sum_i (y_i - \mathbf{w}^\top \mathbf{x}_i)^2 \sigma_y^{-2} + \text{const} \\&= -\frac{1}{2} \mathbf{w}^\top \underbrace{(\mathbf{A} + \sigma_y^{-2} \sum_i \mathbf{x}_i \mathbf{x}_i^\top)}_{\Sigma_w^{-1}} \mathbf{w} + \mathbf{w}^\top \sum_i (y_i \mathbf{x}_i) \sigma_y^{-2} + \text{const} \\&= -\frac{1}{2} \mathbf{w}^\top \Sigma_w^{-1} \mathbf{w} + \mathbf{w}^\top \underbrace{\Sigma_w^{-1} \Sigma_w}_{\mu_w} \sum_i (y_i \mathbf{x}_i) \sigma_y^{-2} + \text{const} \\&= \log \mathcal{N}(\Sigma_w \sum_i (y_i \mathbf{x}_i) \sigma_y^{-2}, \Sigma_w)\end{aligned}$$

MAP AND ML FOR LINEAR REGRESSION

As the posterior is Gaussian, the MAP and posterior mean weights are the same:

$$\mathbf{w}^{\text{MAP}} = \underbrace{\left(A + \frac{\sum_i \mathbf{x}_i \mathbf{x}_i^T}{\sigma_y^2} \right)^{-1}}_{\Sigma_w} \frac{\sum_i y_i \mathbf{x}_i}{\sigma_y^2} = \left(A \sigma_y^2 + \sum_i \mathbf{x}_i \mathbf{x}_i^T \right)^{-1} \sum_i y_i \mathbf{x}_i$$

Compare this to the (transposed) ML weight vector for scalar outputs:

$$\mathbf{w}^{\text{ML}} = \hat{\mathbf{W}}^T = \left(\sum_i \mathbf{x}_i \mathbf{x}_i^T \right)^{-1} \sum_i y_i \mathbf{x}_i$$

- The prior acts to “inflate” the apparent covariance of inputs.
- As A is positive definite, [shrinks](#) the weights towards the prior mean (here $\mathbf{0}$).
- If $A = \alpha I$ this is known as the [ridge regression](#) estimator.
- The MAP/shrinkage/ridge weight estimate often has lower squared error (despite bias) and makes more accurate predictions on test inputs than the ML estimate.
- An example of prior-based [regularization](#) of estimates.

Remarks

- Models the *conditional* $P(\mathbf{y}|\mathbf{x})$.
- If we also model $P(\mathbf{x})$, then learning is indistinguishable from unsupervised. In particular if $P(\mathbf{x})$ is Gaussian, and $P(\mathbf{y}|\mathbf{x})$ is linear-Gaussian, then $\mathbf{x}, \mathbf{y}$ are *jointly Gaussian*.