The Gaussian distribution

Probability distributions

Francesco Corona

Outline

The Gaussian distribution
Probability distributions

The Gaussian distribution

The **Gaussian** or **normal distribution**, is a classic model for the distribution of continuous variables

Definition

In the case of a single variable x, the Gaussian distribution can be written as

$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}}\exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) \tag{1}$$

- μ is the mean and σ^2 is the variance

The Gaussian distribution (cont.)

Definition

In the case of a D-dimensional variable $\mathbf{x}$, the Gaussian distribution is

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) \right) \tag{2}$$

- $\boldsymbol{\mu}$ is the D-dimensional mean vector
- $\boldsymbol{\Sigma}$ is the $D \times D$ covariance matrix
- $|\boldsymbol{\Sigma}|$ is the determinant of $\boldsymbol{\Sigma}$

The geometry
The Gaussian distribution

The geometry (cont.)

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) \right) \tag{3}$$

Let us start by considering the geometric form of the Gaussian distribution

- The Gaussian depends on $\mathbf{x}$ through the quadric form

$$\Delta^2 = (\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) \tag{4}$$

- Quantity Δ is the **Mahalanobis distance** from $\boldsymbol{\mu}$ to $\mathbf{x}$
- It reduces to the Euclidean distance when $\boldsymbol{\Sigma} = \mathbf{I}$

The Gaussian is constant on surfaces in $\mathbf{x}$-space for which Eq. 4 is constant

The geometry - Eigen-equation

Definition

For a square matrix $\mathbf{A}$ of size $M \times M$, the **eigenvector equation** is

$$\mathbf{A}\mathbf{u}_i = \lambda_i \mathbf{u}_i, \quad \text{with } i = 1, \ldots, M$$

- $\mathbf{u}_i$ is an **eigenvector** of $\mathbf{A}$ and λ_i is the associated **eigenvalue**

Remark

A vector is a linear function of another vector if changes in the first vector are directly, linearly, proportional to changes in the second

Proportionality between the two is expressed by a coefficient matrix **A**

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

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_r \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{r1} & a_{r2} & \cdots & a_{rn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

Each component of **y** may be a function of all components of **x**

$$\begin{aligned} y_1 &= a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n \\ y_2 &= a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n \\ \vdots &= \vdots \\ y_r &= a_{r1}x_1 + a_{r2}x_2 + \cdots + a_{rn}x_n \end{aligned}$$

Rows of **A** each multiply the column vector **x** to produce the column vector **y**

- The number of rows of **A** must equal the dimension of **y**
- The number of columns must equal the dimension of **x**

The geometry - Eigen-equation (cont.)

A square matrix $\mathbf{A}$ of size $M \times M$ can be used to transform one M-vector $\mathbf{x}$

- into another M-vector $\mathbf{y}$

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

If we can find a scalar variable λ_i and a particular value $\mathbf{x}$ $(= \mathbf{u}_i)$ such that

- identical transformations are given by

$$\mathbf{y}_i = \mathbf{A}\mathbf{u}_i = \lambda_i \mathbf{u}_i$$

we say that λ_i is an eigenvalue of $\mathbf{A}$ and $\mathbf{u}_i$ the corresponding eigenvector

The geometry - Eigen-equation (cont.)

Remark

Every matrix of size $M \times M$ has M eigenvalues and up to M eigenvectors

$$(\lambda_i \mathbf{I}_M - \mathbf{A})\mathbf{u}_i = \mathbf{0}, \quad i = 1, \ldots, M$$

If this equation is satisfied, then so is

$$(\lambda_i \mathbf{I}_M - \mathbf{A})\alpha\mathbf{u}_i = \mathbf{0}$$

for any scalar (arbitrary constant) α

The eigenvectors are specified only within an arbitrary multiplicative factor

The geometry - Eigen-equation (cont.)

The eigen-equation is a set of M simultaneous homogeneous linear equations

- The condition for a solution is the **characteristic equation**

$$|\mathbf{A} - \lambda \mathbf{I}_M| = 0$$

- An order M polynomial in λ so, M (non-distinct) solutions
- The **rank** of $\mathbf{A}$ equals the number of its nonzero eigenvalues
- Generally, the eigenvalues of a matrix are complex numbers[1]

If the eigenvalues are all distinct, then there are M independent eigenvectors

[1] Real roots are accompanied by real eigenvectors and complex roots (in conjugate pairs for real $\mathbf{A}$) are associated with complex-conjugate pairs of eigenvectors.

The geometry - Eigen-equation (cont.)

Rank($\mathbf{A}$) is defined as the number of linearly independent row/columns in $\mathbf{A}$

- Given a set of vectors $\{\mathbf{a}_1, \ldots, \mathbf{a}_K\}$, the set is said to be **linearly independent** if and only if $\sum_k \alpha_k \mathbf{a}_k = 0$ holds when all $\alpha_k = 0$
- No vector $\mathbf{a}_k$ can be expressed as linear combination of the others

The geometry - Eigen-equation (cont.)

A symmetric matrix $\mathbf{A}$, a covariance of the Gaussian, is such that $A_{ij} = A_{ji}$

$$\mathbf{A}^T = \mathbf{A}$$

The inverse of a symmetric matrix is also symmetric $(\mathbf{A}^{-1})^T = \mathbf{A}^{-1}$, with

$$\mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$$

The geometry - Eigen-equation (cont.)

Eigenvectors $\mathbf{u}_i$ of a real symmetric matrix can be chosen to be orthonormal

$$\mathbf{u}_i^T \mathbf{u}_j = I_{ij} \quad \text{with } I_{ij} \text{ elements of } \mathbf{I} \text{ such that } I_{ij} = \begin{cases} 1, & \text{if } i = j \\ 0, & \text{if } i \neq j \end{cases}$$

The eigenvalues λ_i of a symmetric matrix are real numbers

Evidence

Left multiply $\mathbf{A}\mathbf{u}_i = \lambda_i \mathbf{u}_i$ by $\mathbf{u}_j^T$ to get $\mathbf{u}_j^T \mathbf{A}\mathbf{u}_i = \lambda_i \mathbf{u}_j^T \mathbf{u}_i$ and then by exchange of indexes $\mathbf{u}_i^T \mathbf{A}\mathbf{u}_j = \lambda_j \mathbf{u}_i^T \mathbf{u}_j$. Subtract the two equations (after transposing the second one) and use the symmetry property of $\mathbf{A}$ to get $(\lambda_i - \lambda_j)\mathbf{u}_i^T \mathbf{u}_j = 0$

- $\mathbf{u}_i^T \mathbf{u}_j = 0$, for $\lambda_i \neq \lambda_j$

If $\lambda_i = \lambda_j$, then any linear combination $\alpha \mathbf{u}_i + \beta \mathbf{u}_j$ is also an eigenvector with the same eigenvalue and we can select arbitrarily one linear combination and pick the other one to be orthogonal to it

Since there are M eigenvalues, the corresponding M orthogonal eigenvectors form a complete set and any M-vector can be expressed as linear combination

The geometry - Eigen-equation (cont.)

Since the eigenvectors $\mathbf{u}_i$ can be chosen to be orthogonal and of unit length, we can take them to be the columns of an orthogonal $M \times M$ matrix $\mathbf{U}$

$$\mathbf{U}\mathbf{U}^T = \mathbf{I}$$

- Incidentally, note that also the rows of $\mathbf{U}$ are orthogonal, $\mathbf{U}^T\mathbf{U} = \mathbf{I}$

We can use $\mathbf{U}$ to transform a vector $\mathbf{x}$ into a new vector $\tilde{\mathbf{x}}$, that is $\tilde{\mathbf{x}} = \mathbf{U}\mathbf{x}$

- The length of the vector is preserved: $\tilde{\mathbf{x}}^T\tilde{\mathbf{x}} = \mathbf{x}^T\mathbf{U}^T\mathbf{U}\mathbf{x} = \mathbf{x}^T\mathbf{x}$
- The angle between two vectors is preserved: $\tilde{\mathbf{x}}^T\tilde{\mathbf{y}} = \mathbf{x}^T\mathbf{U}^T\mathbf{U}\mathbf{y} = \mathbf{x}^T\mathbf{y}$

Remark

Multiplication by $\mathbf{U}$ represents a rigid rotation of the coordinate system

The geometry - Eigen-equation (cont.)

We can write the eigenequation $\mathbf{A}\mathbf{u}_i = \lambda_i \mathbf{u}_i$, with $i = 1, \ldots, N$, in terms of $\mathbf{U}$

$$\mathbf{AU} = \mathbf{U}\boldsymbol{\Lambda}$$

- $\boldsymbol{\Lambda}$ is a $M \times M$ diagonal matrix, with diagonal $\text{diag}(\boldsymbol{\Lambda}) = (\lambda_1, \ldots, \lambda_M)^T$
- $\mathbf{U}^T\mathbf{AU} = \underbrace{\mathbf{U}^T\mathbf{U}}_{\mathbf{I}}\boldsymbol{\Lambda} \implies \mathbf{U}^T\mathbf{AU} = \boldsymbol{\Lambda}$, matrix $\mathbf{A}$ is diagonalised by $\mathbf{U}$
- $\underbrace{\mathbf{U}\mathbf{U}^T}_{\mathbf{I}}\mathbf{A}\underbrace{\mathbf{U}\mathbf{U}^T}_{\mathbf{I}} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^T \implies \mathbf{A} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^T \implies \mathbf{A} = \sum_{i=1}^{M} \lambda_i \mathbf{u}_i \mathbf{u}_i^T$
- $\mathbf{A}^{-1} = \mathbf{U}\boldsymbol{\Lambda}^{-1}\mathbf{U}^T \implies \mathbf{A}^{-1} = \sum_{i=1}^{M} \frac{1}{\lambda_i}\mathbf{u}_i \mathbf{u}_i^T$
- $|\mathbf{A}| = \prod_{i=1}^{M} \lambda_i$
- $\text{Trace}(\mathbf{A}) = \sum_{i=1}^{M} \lambda_i$

The geometry (cont.)

Consider the eigenvector equation for a real symmetric covariance matrix $\mathbf{\Sigma}$

$$\mathbf{\Sigma}\mathbf{u}_i = \lambda_i \mathbf{u}_i, \qquad \text{with } i = 1 \ldots, D \tag{5}$$

Real eigenvalues and its eigenvectors form an orthonormal set

$$\mathbf{u}_i^T \mathbf{u}_j = I_{ij} = \begin{cases} 1, & \text{if } i = j \\ 0, & \text{if } i \neq j \end{cases} \tag{6}$$

$\mathbf{\Sigma}$ can be expressed as an expansion of its eigenvectors ($\star$)

$$\mathbf{\Sigma} = \sum_{i=1}^{D} \lambda_i \mathbf{u}_i \mathbf{u}_i^T \tag{7}$$

The inverse covariance matrix $\mathbf{\Sigma}^{-1}$ can be expressed ($\star$) as

$$\mathbf{\Sigma}^{-1} = \sum_{i=1}^{D} \frac{1}{\lambda_i} \mathbf{u}_i \mathbf{u}_i^T \tag{8}$$

The geometry (cont.)

By substituting the inverse covariance matrix $\boldsymbol{\Sigma}^{-1}$ into the quadratic form Δ^2

$$\begin{aligned}
\Delta^2 &= (\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) \\
&= (\mathbf{x}-\boldsymbol{\mu})^T\Big(\sum_{i=1}^{D}\frac{1}{\lambda_i}\mathbf{u}_i\mathbf{u}_i^T\Big)(\mathbf{x}-\boldsymbol{\mu}) \\
&= \sum_{i=1}^{D}\frac{(\mathbf{x}-\boldsymbol{\mu})^T\mathbf{u}_i\mathbf{u}_i^T(\mathbf{x}-\boldsymbol{\mu})}{\lambda_i} \\
&= \sum_{i=1}^{D}\frac{y_i^2}{\lambda_i}, \quad \text{with } y_i = \mathbf{u}_i^T(\mathbf{x}-\boldsymbol{\mu}) \qquad (9)
\end{aligned}$$

The geometry (cont.)

We interpret $\{y_i\}$ as a new coordinate system defined by orthonormal vectors $\mathbf{u}_i$, that are shifted by $\boldsymbol{\mu}$ and rotated with respect to original coordinates $\{x_i\}$

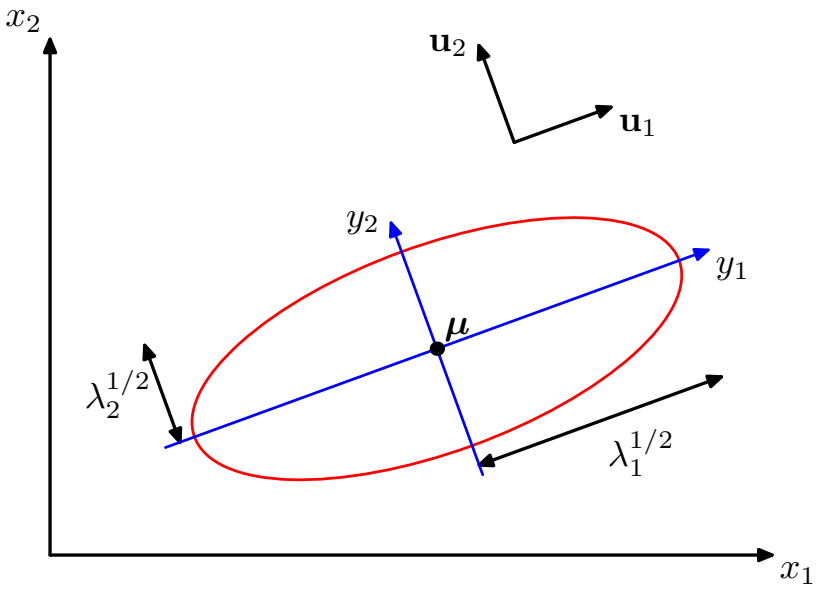
- Forming the vector $\mathbf{y} = (y_1, \ldots, y_D)^T$, we have

$$\mathbf{y} = \mathbf{U}(\mathbf{x} - \boldsymbol{\mu}) \tag{10}$$

$\mathbf{U}$ is an orthogonal matrix whose rows are $\mathbf{u}_i^T$ (i.e., $\mathbf{U}\mathbf{U}^T = \mathbf{I}$ and $\mathbf{U}^T\mathbf{U} = \mathbf{I}$)

The geometry (cont.)

The quadratic form and thus the Gaussian is constant on surfaces for which $\Delta^2 = \sum_{i=1}^{D} \frac{y_i^2}{\lambda_i}$ is constant



For $\lambda_i > 0$, the surfaces are ellipsoids

- Centred in $\boldsymbol{\mu}$ and axis along $\mathbf{u}_i$
- The scaling factors in the directions of the axes are $\lambda_i^{1/2}$

The red curve is the elliptical surface of a $2D$ Gaussian, in which the density is $\exp(-1/2)$ of its value at $\mathbf{x} = \boldsymbol{\mu}$

The geometry (cont.)

For the Gaussian to be well-defined, all of the eigenvalues λ_i of the covariance matrix need be strictly positive[2], otherwise it cannot be properly normalised

A Gaussian for which one or more eigenvalues are zero[3] is singular

- It is confined to a subspace of lower dimensionality

[2] A matrix whose eigenvalues are strictly positive is called **positive definite**

[3] A matrix in which all of the eigenvalues are nonnegative is called **positive semidefinite**

The geometry - Jacobian factor

Under a change of variable, a density does not transform like a regular function

For a change of variables $x = g(y)$, a function $f(x)$ becomes $\tilde{f}(y) = f(g(y))$

Definition

Consider a probability density $p_x(x)$ that corresponds to a density $p_y(y)$ wrt a new variable y and notice that $p_x(x)$ and $p_y(y)$ are different densities

- Observations falling in a range $(x, x + \delta x)$ have probability $p_x(x)\delta x$
- By transforming them, we make them fall in the range $(y, y + \delta y)$
- Observations falling in a range $(y, y + \delta y)$ have probability $p_y(y)\delta y$

$$p_x(x)\delta x \simeq p_y(y)\delta y$$

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

The geometry (cont.)

Consider now the Gaussian in the new coordinate system defined by $\{y_i\}$

In going from $\mathbf{x}$ to $\mathbf{y}$, we have a Jacobian matrix $\mathbf{J}$

$$J_{ij} = \frac{\partial x_i}{\partial y_j} = U_{ji} \tag{11}$$

Thus, the elements of the $\mathbf{J}$ are the elements of $\mathbf{U}^T$

Using the orthonormality property of $\mathbf{U}$, the square of the determinant of $\mathbf{J}$

$$|\mathbf{J}^2| = |\mathbf{U}^T|^2 = |\mathbf{U}^T||\mathbf{U}| = |\mathbf{U}^T\mathbf{U}| = |\mathbf{I}| = 1 \qquad \Longrightarrow |\mathbf{J}| = 1 \tag{12}$$

The geometry (cont.)

We also have that the determinant $|\mathbf{\Sigma}|$ of the covariance matrix can be written as the product of its eigenvalues, and thus we also have

$$|\mathbf{\Sigma}| = \prod_{j=1}^{D} \lambda_j \qquad \Longrightarrow \qquad |\mathbf{\Sigma}|^{1/2} = \prod_{j=1}^{D} \lambda_j^{1/2} \tag{13}$$

Hence, the Gaussian distribution in the coordinate system $\{y_i\}$ becomes

$$p(\mathbf{y}) = p(\mathbf{x})|\mathbf{J}| = \prod_{j=1}^{D} \frac{1}{(2\pi\lambda_j)^{1/2}} \exp\left(-\frac{y_j^2}{2\lambda_j}\right) \tag{14}$$

Here, it is the product of D independent univariate Gaussian distributions

- The eigenvectors define a new set of shifted and rotated coordinates
- Here, the joint distribution factorises into independent distributions

The geometry (cont.)

Using a result derived for the normalisation of the univariate Gaussian,

$$\int p(\mathbf{y})d\mathbf{y} = \prod_{j=1}^{D} \int_{-\infty}^{+\infty} \frac{1}{(2\pi\lambda_j)^{1/2}} \exp\left(-\frac{y_j^2}{2\lambda_j}\right) dy_j = 1 \tag{15}$$

which is the integral of the distribution in the $\mathbf{y}$ coordinate system

The geometry (cont.)

We look at the moments of the Gaussian distribution (back to $\mathbf{x}$-space)

The expectation of $\mathbf{x}$ under the Gaussian distribution is given by

$$\begin{aligned}\mathbb{E}[\mathbf{x}] &= \frac{1}{(2\pi)^{D/2}}\frac{1}{|\boldsymbol{\Sigma}|^{1/2}}\int \exp\Big(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\Big)\mathbf{x}d\mathbf{x} \\ &= \frac{1}{(2\pi)^{D/2}}\frac{1}{|\boldsymbol{\Sigma}|^{1/2}}\int \exp\Big(-\frac{1}{2}\mathbf{z}^T\boldsymbol{\Sigma}^{-1}\mathbf{z}\Big)(\mathbf{z}+\boldsymbol{\mu})d\mathbf{z} \qquad (16)\end{aligned}$$

- We have changed variables using $\mathbf{z} = \mathbf{x} + \boldsymbol{\mu}$

The exponent is an even function[4] of the components $\mathbf{z}$, which for integrals taken in $(-\infty, +\infty)$ will make the term in $\mathbf{z}$ in the factor $(\mathbf{z} + \boldsymbol{\mu})$ vanish[5]

$$\mathbb{E}[\mathbf{x}] = \boldsymbol{\mu} \qquad (17)$$

We refer to it as the **mean vector** of the Gaussian distribution

[4] A real-valued function is said to even if $f(x) = f(-x)$, or $f(x) - f(-x) = 0$

[5] Also, $\int x \exp(-cx^2)dx = -1/(2c)\exp(-cx^2)$

The geometry (cont.)

There are D^2 second order moments $\mathbb{E}[x_i x_j]$ under the Gaussian distribution

$$\begin{aligned}\mathbb{E}[\mathbf{x}\mathbf{x}^T] &= \frac{1}{(2\pi)^{D/2}}\frac{1}{|\boldsymbol{\Sigma}|^{1/2}}\int \exp\Big(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\Big)\mathbf{x}\mathbf{x}^T d\mathbf{x} \\ &= \frac{1}{(2\pi)^{D/2}}\frac{1}{|\boldsymbol{\Sigma}|^{1/2}}\int \exp\Big(-\frac{1}{2}\mathbf{z}^T\boldsymbol{\Sigma}^{-1}\mathbf{z}\Big)(\mathbf{z}+\boldsymbol{\mu})(\mathbf{z}+\boldsymbol{\mu})^T d\mathbf{z} \quad (18)\end{aligned}$$

- The cross-terms involving $\boldsymbol{\mu}\mathbf{z}^T$ and $\mathbf{z}\boldsymbol{\mu}^T$ will vanish by symmetry
- The term $\boldsymbol{\mu}\boldsymbol{\mu}^T$ is constant and can be taken outside the integral

The geometry (cont.)

As for the term involving $\mathbf{z}\mathbf{z}^T$, using the eigenvector expansion of the covariance matrix together the completeness of the set of eigenvectors

$$\mathbf{z} = \sum_{j=1}^{D} y_j \mathbf{u}_j, \qquad \text{with } y_j = \mathbf{u}_j^T \mathbf{z} \tag{19}$$

$$\begin{aligned}
&\frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \int \exp\Big(-\frac{1}{2}\mathbf{z}^T \boldsymbol{\Sigma}^{-1}\mathbf{z}\Big)\mathbf{z}\mathbf{z}^T d\mathbf{z} \\
&\quad = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \sum_{i=1}^{D}\sum_{j=1}^{D} \mathbf{u}_i \mathbf{u}_j^T \int \exp\Big(-\sum_{k=1}^{D} \frac{y_k^2}{2\lambda_k}\Big) y_i y_j d\mathbf{y} \\
&\quad = \sum_{i=1}^{D} \mathbf{u}_i \mathbf{u}_i^T \lambda_i = \boldsymbol{\Sigma}
\end{aligned} \tag{20}$$

The geometry (cont.)

We used the eigenvector equation $\mathbf{\Sigma}\mathbf{u}_i = \lambda_i \mathbf{u}_i$ and the fact that the integral on the right-hand side of the middle line vanishes by symmetry unless $i = j$

In the final line we used $\mathbb{E}[x^2] = \int_{-\infty}^{+\infty} \mathcal{N}(x|\mu, \sigma^2) x^2 dx = \mu^2 + \sigma^2$ and $|\mathbf{\Sigma}|^{1/2} = \prod_{j=1}^{D} \lambda_j^{1/2}$, together with $\mathbf{\Sigma} = \sum_{i=1}^{D} \lambda_i \mathbf{u}_i \mathbf{u}_i^T$

As a result, we have

$$\mathbb{E}[\mathbf{x}\mathbf{x}^T] = \boldsymbol{\mu}\boldsymbol{\mu}^T + \mathbf{\Sigma} \tag{21}$$

The geometry (cont.)

For single random variables, we subtracted the mean before taking second moments and define a variance

Similarly, in the multivariate case it is again convenient to subtract off the mean, giving rise to the covariance of a random vector $\mathbf{x}$ defined by

$$\text{cov}[\mathbf{x}] = \text{cov}[\mathbf{x}, \mathbf{x}] = \mathbb{E}[(\mathbf{x} - \mathbb{E}[\mathbf{x}])(\mathbf{x}^T - \mathbb{E}[\mathbf{x}^T])] \tag{22}$$

For the specific case of a Gaussian distribution, we use of the first-order moment $\mathbb{E}[\mathbf{x}] = \boldsymbol{\mu}$ and the second-order moment $\mathbb{E}[\mathbf{x}\mathbf{x}^T] = \boldsymbol{\mu}\boldsymbol{\mu}^T + \boldsymbol{\Sigma}$

$$\text{cov}[\mathbf{x}] = \boldsymbol{\Sigma} \tag{23}$$

The parameter matrix $\boldsymbol{\Sigma}$ governs the covariance of $\mathbf{x}$ under the Gaussian distribution, hence it is called the **covariance matrix**

The geometry (cont.)

The Gaussian distribution is widely used as a density model

- Though, it suffers from some significant limitations

Remark

Consider the number of free parameters $D(D+3)/2$ in the distribution $(\star)$

- A general symmetric covariance matrix $\boldsymbol{\Sigma}$ has $D(D+1)/2$ independent parameters $(\star)$ and there are another D independent parameters in $\boldsymbol{\mu}$

For large D, this number grows quadratically with D, and the computational task of manipulating and inverting large matrices can become prohibitive

A way to *address* this problem is to use covariance matrices of restricted form

The geometry (cont.)

We can consider covariance matrices that are diagonal ($\mathbf{\Sigma} = \text{diag}(\sigma_i^2)$)

- A total of $2D$ independent parameters in the density model
- Axis-aligned ellipsoids of constant density

We can restrict covariance matrices to be proportional to the identity matrix ($\mathbf{\Sigma} = \sigma^2\mathbf{I}$, or isotropic covariance), we then get $D + 1$ independent parameters

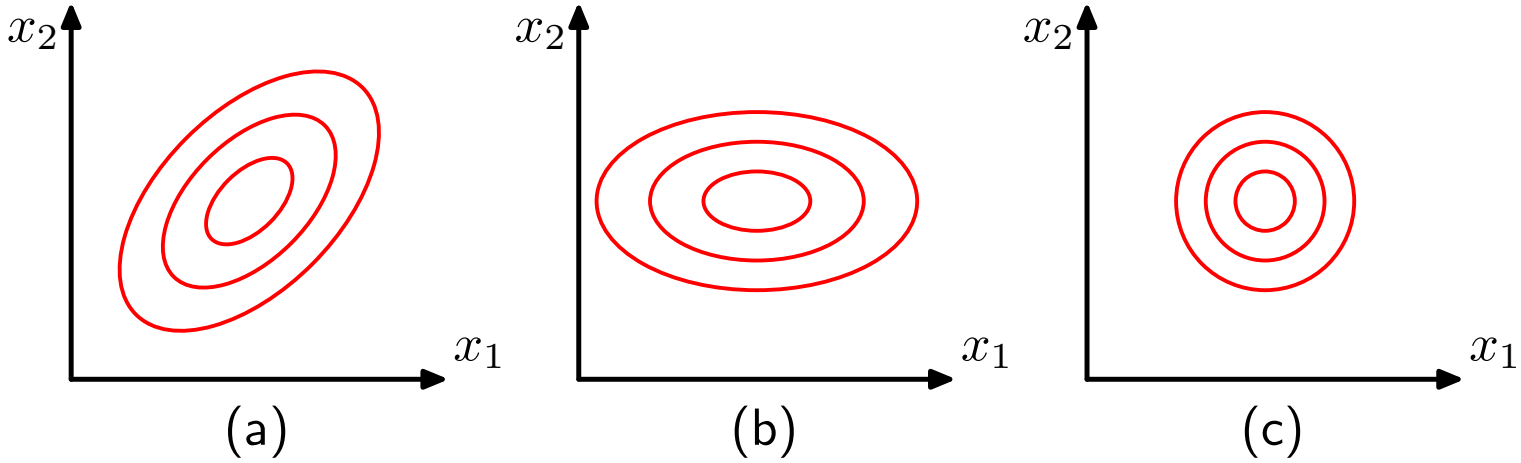
- Spherical surfaces of constant density

Such approaches limit the number of degrees of freedom in the distribution

- Inversion of the covariance matrix much faster

They also greatly restrict the form of the probability density and limit its ability to capture interesting correlations in the data

The geometry (cont.)



- (a): general $\mathbf{\Sigma}$
- (b): $\mathbf{\Sigma} = \text{diag}(\sigma_i^2)$
- (c): $\mathbf{\Sigma} = \sigma^2 \mathbf{I}$

Completing the square
The Gaussian distribution

Completing the square

Completing the square: A useful technique for manipulating Gaussians

Remark

$$\exp\left(-\frac{1}{2}\mathbf{x}^T\mathbf{A}\mathbf{x}+\mathbf{b}^T\mathbf{x}\right)$$

$$\frac{1}{2}\mathbf{x}^T\mathbf{A}\mathbf{x}-\mathbf{b}^T\mathbf{x}=\frac{1}{2}(\mathbf{x}^T-\mathbf{A}^{-1}\mathbf{b})^T\mathbf{A}^{-1}(\mathbf{x}^T-\mathbf{A}^{-1}\mathbf{b})-\frac{1}{2}\mathbf{b}^T\mathbf{A}^{-1}\mathbf{b}$$

$$\exp\left(-\frac{1}{2}\mathbf{x}^T\mathbf{A}^{-1}\mathbf{x}-\mathbf{b}^T\mathbf{x}\right)=\mathcal{N}(\mathbf{x}|\mathbf{A}^{-1}\mathbf{b},\mathbf{A}^{-1})|2\pi\mathbf{A}^{-1}|^{1/2}\exp\left(\frac{1}{2}\mathbf{b}^T\mathbf{A}^{-1}\mathbf{b}\right)$$

$$\int\exp\left(-\frac{1}{2}\mathbf{x}^T\mathbf{A}\mathbf{x}+\mathbf{b}^T\mathbf{x}\right)\mathrm{d}\mathbf{x}=|2\pi\mathbf{A}|^{1/2}\exp\left(\frac{1}{2}\mathbf{b}^T\mathbf{A}\mathbf{b}\right)$$

Conditional Gaussian distributions

The Gaussian distribution

Conditional Gaussian distributions

Property of the Gaussian: If two sets of variables are jointly Gaussian, then the conditional distribution of one set conditioned on the other is again Gaussian

Suppose $\mathbf{x}$ is a D-dimensional vector with Gaussian distribution $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$

- We partition $\mathbf{x}$ into two disjoint subsets $\mathbf{x}_a$ and $\mathbf{x}_b$

Without loss of generality:

- $\mathbf{x}_a$ comprises the first M components of $\mathbf{x}$
- $\mathbf{x}_b$ comprises the remaining $(D - M)$ ones
-

$$\mathbf{x} = \begin{pmatrix} \mathbf{x}_a \\ \mathbf{x}_b \end{pmatrix} \tag{24}$$

Conditional Gaussian distributions (cont.)

We also define corresponding partitions of

- the mean vector $\boldsymbol{\mu}$

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{pmatrix} \tag{25}$$

- the covariance matrix $\boldsymbol{\Sigma}$

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb} \end{pmatrix} \tag{26}$$

Because of the symmetry $\boldsymbol{\Sigma}^T = \boldsymbol{\Sigma}$, we have that $\boldsymbol{\Sigma}_{aa}$ and $\boldsymbol{\Sigma}_{bb}$ are also symmetric and $\boldsymbol{\Sigma}_{ba} = \boldsymbol{\Sigma}_{ab}^T$

Conditional Gaussian distributions (cont.)

In many situations, it is convenient to work with the **precision matrix**

$$\mathbf{\Lambda} = \mathbf{\Sigma}^{-1} \tag{27}$$

The corresponding partitioned precision matrix is given by the form

$$\mathbf{\Lambda} = \begin{pmatrix} \mathbf{\Lambda}_{aa} & \mathbf{\Lambda}_{ab} \\ \mathbf{\Lambda}_{ba} & \mathbf{\Lambda}_{bb} \end{pmatrix} \tag{28}$$

Because the inverse of a symmetric matrix is also symmetric, $\mathbf{\Lambda}_{aa}$ and $\mathbf{\Lambda}_{bb}$ are also symmetric and $\mathbf{\Lambda}_{ba} = \mathbf{\Lambda}_{ab}^T$ $(\star)$

Conditional Gaussian distributions (cont.)

We begin by finding an expression for the conditional distribution $p(\mathbf{x}_a|\mathbf{x}_b)$

From the product rule of probability, this conditional distribution can be evaluated from the joint distribution $p(\mathbf{x}) = p(\mathbf{x}_a, \mathbf{x}_b)$ by fixing $\mathbf{x}_b$ and normalising the result to obtain a valid probability distribution over $\mathbf{x}_a$

Instead of performing this normalisation explicitly, we can obtain the solution more efficiently by using the quadratic form in the exponent of the Gaussian

$$-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Lambda}(\mathbf{x}-\boldsymbol{\mu})$$

and reinstating the normalisation coefficient at the end of the manipulations

Conditional Gaussian distributions (cont.)

We use the partitioning:

$$\mathbf{x} = \begin{pmatrix} \mathbf{x}_a \\ \mathbf{x}_b \end{pmatrix} \quad \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{pmatrix} \quad \boldsymbol{\Lambda} = \begin{pmatrix} \boldsymbol{\Lambda}_{aa} & \boldsymbol{\Lambda}_{ab} \\ \boldsymbol{\Lambda}_{ba} & \boldsymbol{\Lambda}_{bb} \end{pmatrix}$$

$$\begin{aligned} -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Lambda}(\mathbf{x}-\boldsymbol{\mu}) = &-\frac{1}{2}(\mathbf{x}_a-\boldsymbol{\mu}_a)'\boldsymbol{\Lambda}_{aa}(\mathbf{x}_a-\boldsymbol{\mu}_a) - \frac{1}{2}(\mathbf{x}_a-\boldsymbol{\mu}_a)'\boldsymbol{\Lambda}_{ab}(\mathbf{x}_b-\boldsymbol{\mu}_b) \\ &- \frac{1}{2}(\mathbf{x}_b-\boldsymbol{\mu}_b)'\boldsymbol{\Lambda}_{ba}(\mathbf{x}_a-\boldsymbol{\mu}_a) - \frac{1}{2}(\mathbf{x}_b-\boldsymbol{\mu}_b)'\boldsymbol{\Lambda}_{bb}(\mathbf{x}_b-\boldsymbol{\mu}_b) \end{aligned}$$

We see that as a function of $\mathbf{x}_a$, this is again a quadratic form, and hence the corresponding conditional distribution $p(\mathbf{x}_a|\mathbf{x}_b)$ will be Gaussian

This distribution is completely characterised by its mean and its covariance, and our goal is to identify expressions for the mean and covariance of $p(\mathbf{x}_a|\mathbf{x}_b)$

Conditional Gaussian distributions (cont.)

We are given a quadratic form defining the exponent terms in a Gaussian

- We need to determine corresponding mean and covariance

The exponent in a general Gaussian distribution $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ can be written

$$-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) = -\frac{1}{2}\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x} + \mathbf{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \text{const} \qquad (29)$$

where 'const' denotes terms which are independent on $\mathbf{x}$ (i.e., $-\frac{1}{2}\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$)

If we take the general quadratic form and express it in the form given by the right-hand side of Equation 29, then we can equate the matrix of coefficients entering the second-order term in $\mathbf{x}$ to the inverse covariance matrix $\boldsymbol{\Sigma}^{-1}$ and the coefficient of the linear term in $\mathbf{x}$ to $\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$, from which we can obtain $\boldsymbol{\mu}$

Conditional Gaussian distributions (cont.)

We apply this procedure to the conditional Gaussian distribution $p(\mathbf{x}_a|\mathbf{x}_b)$:

- The quadratic form in the exponent

$$-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Lambda}(\mathbf{x}-\boldsymbol{\mu}) = -\frac{1}{2}(\mathbf{x}_a-\boldsymbol{\mu}_a)'\boldsymbol{\Lambda}_{aa}(\mathbf{x}_a-\boldsymbol{\mu}_a) - \frac{1}{2}(\mathbf{x}_a-\boldsymbol{\mu}_a)'\boldsymbol{\Lambda}_{ab}(\mathbf{x}_b-\boldsymbol{\mu}_b) - \frac{1}{2}(\mathbf{x}_b-\boldsymbol{\mu}_b)'\boldsymbol{\Lambda}_{ba}(\mathbf{x}_a-\boldsymbol{\mu}_a) - \frac{1}{2}(\mathbf{x}_b-\boldsymbol{\mu}_b)'\boldsymbol{\Lambda}_{bb}(\mathbf{x}_b-\boldsymbol{\mu}_b)$$

We will denote mean and covariance of this distribution by $\boldsymbol{\mu}_{a|b}$ and $\boldsymbol{\Sigma}_{a|b}$

Remark

Consider the functional dependence on $\mathbf{x}_a$ in which $\mathbf{x}_b$ is regarded as constant

Conditional Gaussian distributions (cont.)

$$-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Lambda}(\mathbf{x}-\boldsymbol{\mu}) = -\frac{1}{2}(\mathbf{x}_a-\boldsymbol{\mu}_a)'\boldsymbol{\Lambda}_{aa}(\mathbf{x}_a-\boldsymbol{\mu}_a) - \frac{1}{2}(\mathbf{x}_a-\boldsymbol{\mu}_a)'\boldsymbol{\Lambda}_{ab}(\mathbf{x}_b-\boldsymbol{\mu}_b) - \frac{1}{2}(\mathbf{x}_b-\boldsymbol{\mu}_b)'\boldsymbol{\Lambda}_{ba}(\mathbf{x}_a-\boldsymbol{\mu}_a) - \frac{1}{2}(\mathbf{x}_b-\boldsymbol{\mu}_b)'\boldsymbol{\Lambda}_{bb}(\mathbf{x}_b-\boldsymbol{\mu}_b)$$

- If we pick out all second-order terms in $\mathbf{x}_a$, we have

$$-\frac{1}{2}\mathbf{x}_a^T\boldsymbol{\Lambda}_{aa}\mathbf{x}_a \tag{30}$$

- From which, the covariance of $p(\mathbf{x}_a|\mathbf{x}_b)$ is given as

$$\boldsymbol{\Sigma}_{a|b} = \boldsymbol{\Lambda}_{aa}^{-1} \tag{31}$$

Conditional Gaussian distributions (cont.)

$$-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Lambda}(\mathbf{x}-\boldsymbol{\mu}) = -\frac{1}{2}(\mathbf{x}_a-\boldsymbol{\mu}_a)'\boldsymbol{\Lambda}_{aa}(\mathbf{x}_a-\boldsymbol{\mu}_a) - \frac{1}{2}(\mathbf{x}_a-\boldsymbol{\mu}_a)'\boldsymbol{\Lambda}_{ab}(\mathbf{x}_b-\boldsymbol{\mu}_b) - \frac{1}{2}(\mathbf{x}_b-\boldsymbol{\mu}_b)'\boldsymbol{\Lambda}_{ba}(\mathbf{x}_a-\boldsymbol{\mu}_a) - \frac{1}{2}(\mathbf{x}_b-\boldsymbol{\mu}_b)'\boldsymbol{\Lambda}_{bb}(\mathbf{x}_b-\boldsymbol{\mu}_b)$$

- Consider all terms that are linear in $\mathbf{x}_a$

$$\mathbf{x}_a^T\Big(\boldsymbol{\Lambda}_{aa}\boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{ab}(\mathbf{x}_b-\boldsymbol{\mu}_b)\Big) \tag{32}$$

It follows that the coefficient of $\mathbf{x}_a$ in this expression must equal $\boldsymbol{\Sigma}_{a|b}^{-1}\boldsymbol{\mu}_{a|b}$

$$\begin{aligned}\boldsymbol{\mu}_{a|b} &= \underbrace{\boldsymbol{\Sigma}_{a|b}}_{\boldsymbol{\Lambda}_{aa}^{-1}}\Big(\boldsymbol{\Lambda}_{aa}\boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{ab}(\mathbf{x}_b-\boldsymbol{\mu}_b)\Big)\\ &= \boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{aa}^{-1}\boldsymbol{\Lambda}_{ab}(\mathbf{x}_b-\boldsymbol{\mu}_b)\end{aligned} \tag{33}$$

Conditional Gaussian distributions (cont.)

Mean and variance of conditional distribution $p(\mathbf{x}_a|\mathbf{x}_b)$ are given in terms of partitioned precision matrix of the original joint distribution $p(\mathbf{x}_a, \mathbf{x}_b)$

- We can express these results also in terms of partitioned covariance matrix

To do this $\star$, we make use of an identity for the inverse of a partitioned matrix

$$\begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix}^{-1} = \begin{pmatrix} \mathbf{M} & -\mathbf{M}\mathbf{B}\mathbf{D}^{-1} \\ -\mathbf{D}^{-1}\mathbf{C}\mathbf{M} & \mathbf{D}^{-1} + \mathbf{D}^{-1}\mathbf{C}\mathbf{M}\mathbf{B}\mathbf{D}^{-1} \end{pmatrix} \tag{34}$$

with $\mathbf{M} = (\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1}$

$\mathbf{M}^{-1}$ is the **Schur complement** of $\begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix}^{-1}$ with respect to sub-matrix $\mathbf{D}$

Conditional Gaussian distributions (cont.)

Using the definition

$$\begin{pmatrix} \boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb} \end{pmatrix}^{-1} = \begin{pmatrix} \boldsymbol{\Lambda}_{aa} & \boldsymbol{\Lambda}_{ab} \\ \boldsymbol{\Lambda}_{ba} & \boldsymbol{\Lambda}_{bb} \end{pmatrix} \tag{35}$$

and using

$$\begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix}^{-1} = \begin{pmatrix} \mathbf{M} & -\mathbf{M}\mathbf{B}\mathbf{D}^{-1} \\ -\mathbf{D}^{-1}\mathbf{C}\mathbf{M} & \mathbf{D}^{-1} + \mathbf{D}^{-1}\mathbf{C}\mathbf{M}\mathbf{B}\mathbf{D}^{-1} \end{pmatrix}, \quad \mathbf{M} = (\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1}$$

we have that

$$\boldsymbol{\Lambda}_{aa} = (\boldsymbol{\Sigma}_{aa} - \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}\boldsymbol{\Sigma}_{ba})^{-1} \tag{36}$$

$$\boldsymbol{\Lambda}_{ab} = -(\boldsymbol{\Sigma}_{aa} - \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}\boldsymbol{\Sigma}_{ba}^{-1})^{-1}\boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1} \tag{37}$$

Conditional Gaussian distributions (cont.)

We get expressions for the mean and variance of the conditional distribution $p(\mathbf{x}_a|\mathbf{x}_b)$ in terms of the partitioned covariance matrix

$$\boldsymbol{\mu}_{a|b} = \boldsymbol{\mu}_a + \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}(\mathbf{x}_b - \boldsymbol{\mu}_b) \quad (38)$$

$$\boldsymbol{\Sigma}_{a|b} = \boldsymbol{\Sigma}_{aa} - \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}\boldsymbol{\Sigma}_{ba} \quad (39)$$

Compare $\boldsymbol{\Sigma}_{a|b} = \boldsymbol{\Lambda}_{aa}^{-1}$ with $\boldsymbol{\Sigma}_{a|b} = \boldsymbol{\Sigma}_{aa} - \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}\boldsymbol{\Sigma}_{ba}$

- The conditional distribution $p(\mathbf{x}_a|\mathbf{x}_b)$ takes a simpler form when expressed in terms of the partitioned precision matrix than when it is expressed in terms of the partitioned covariance matrix

Conditional Gaussian distributions (cont.)

The mean $\boldsymbol{\mu}_{a|b} = \boldsymbol{\mu}_a + \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}(\mathbf{x}_b - \boldsymbol{\mu}_b)$ of conditional distribution $p(\mathbf{x}_a|\mathbf{x}_b)$ is a linear function of $\mathbf{x}_b$

The covariance $\boldsymbol{\Sigma}_{a|b} = \boldsymbol{\Sigma}_{aa} - \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}\boldsymbol{\Sigma}_{ba}$ of conditional distribution $p(\mathbf{x}_a|\mathbf{x}_b)$ is independent of $\mathbf{x}_b$

- This is an example of **Linear-Gaussian model**

Marginal Gaussian distributions

The Gaussian distribution

Marginal Gaussian distributions

If a joint distribution $p(\mathbf{x}_a, \mathbf{x}_b)$ is Gaussian, then, the conditional distribution $p(\mathbf{x}_a|\mathbf{x}_b)$ is Gaussian

We discuss the marginal distribution $p(\mathbf{x}_a)$

$$p(\mathbf{x}_a) = \int p(\mathbf{x}_a, \mathbf{x}_b) d\mathbf{x}_b \tag{40}$$

and we shall see that this is also Gaussian

Marginal Gaussian distributions (cont.)

The strategy for evaluating this distribution efficiently focuses on the quadratic form in the exponent of the joint distribution

- Goal: Identify mean and covariance of the marginal distribution $p(\mathbf{x}_a)$

Marginal Gaussian distributions (cont.)

$$-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Lambda}(\mathbf{x}-\boldsymbol{\mu}) = -\frac{1}{2}(\mathbf{x}_a-\boldsymbol{\mu}_a)'\boldsymbol{\Lambda}_{aa}(\mathbf{x}_a-\boldsymbol{\mu}_a) - \frac{1}{2}(\mathbf{x}_a-\boldsymbol{\mu}_a)'\boldsymbol{\Lambda}_{ab}(\mathbf{x}_b-\boldsymbol{\mu}_b) - \frac{1}{2}(\mathbf{x}_b-\boldsymbol{\mu}_b)'\boldsymbol{\Lambda}_{ba}(\mathbf{x}_a-\boldsymbol{\mu}_a) - \frac{1}{2}(\mathbf{x}_b-\boldsymbol{\mu}_b)'\boldsymbol{\Lambda}_{bb}(\mathbf{x}_b-\boldsymbol{\mu}_b)$$

We need to integrate out $\mathbf{x}_b$, because $p(\mathbf{x}_a) = \int p(\mathbf{x}_a, \mathbf{x}_b)d\mathbf{x}_b$

Picking out those terms that involve $\mathbf{x}_b$, we have

$$-\frac{1}{2}\mathbf{x}_b'\boldsymbol{\Lambda}_{bb}\mathbf{x}_b + \mathbf{x}_b'\mathbf{m} = -\frac{1}{2}(\mathbf{x}_b - \boldsymbol{\Lambda}_{bb}^{-1}\mathbf{m})'\boldsymbol{\Lambda}_{bb}(\mathbf{x}_b - \boldsymbol{\Lambda}_{bb}^{-1}\mathbf{m}) + \frac{1}{2}\mathbf{m}'\boldsymbol{\Lambda}_{bb}^{-1}\mathbf{m} \tag{41}$$

$$\text{with } \mathbf{m} = \boldsymbol{\Lambda}_{bb}\boldsymbol{\mu}_b - \boldsymbol{\Lambda}_{ba}(\mathbf{x}_a - \boldsymbol{\mu}_a)$$

The dependence on $\mathbf{x}_b$ has been cast into the standard quadratic form of a Gaussian distribution corresponding to the first term on the right-hand side above, plus a term that does not depend on $\mathbf{x}_b$ but that does depend on $\mathbf{x}_a$

Marginal Gaussian distributions (cont.)

When we take the exponential of this quadratic form, we see that the integration over $\mathbf{x}_b$ required by $p(\mathbf{x}_a) = \int p(\mathbf{x}_a, \mathbf{x}_b) d\mathbf{x}_b$ takes the form

$$\int \exp\left(-\frac{1}{2}(\mathbf{x}_b - \mathbf{\Lambda}_{bb}^{-1}\mathbf{m})^T \mathbf{\Lambda}_{bb}(\mathbf{x}_b - \mathbf{\Lambda}_{bb}^{-1}\mathbf{m})\right) d\mathbf{x}_b \tag{42}$$

It is the integral over an unnormalised Gaussian, and so the result will be the reciprocal of the normalisation coefficient

$$\frac{1}{(2\pi)^{D/2}} \frac{1}{|\mathbf{\Sigma}|^{1/2}}$$

which is independent of the mean, and it depends only on the determinant of the covariance matrix

Marginal Gaussian distributions (cont.)

By completing the square wrt $\mathbf{x}_b$, we can integrate out $\mathbf{x}_b$ and the only term remaining that depends on $\mathbf{x}_a$ from the contributions on the left-hand side of

$$-\frac{1}{2}\mathbf{x}_b^T\boldsymbol{\Lambda}_{bb}\mathbf{x}_b + \mathbf{x}_b\mathbf{m} = -\frac{1}{2}(\mathbf{x}_b - \boldsymbol{\Lambda}_{bb}^{-1}\mathbf{m})'\boldsymbol{\Lambda}_{bb}(\mathbf{x}_b - \boldsymbol{\Lambda}_{bb}^{-1}\mathbf{m}) + \frac{1}{2}\mathbf{m}'\boldsymbol{\Lambda}_{bb}^{-1}\mathbf{m}$$

is the last term on the right-hand side because $\mathbf{m} = \boldsymbol{\Lambda}_{bb}\boldsymbol{\mu}_b - \boldsymbol{\Lambda}_{ba}(\mathbf{x}_a - \boldsymbol{\mu}_a)$

We then combine this term with the remaining terms that depend on $\mathbf{x}_a$ from

$$\begin{aligned}-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})'\boldsymbol{\Lambda}(\mathbf{x} - \boldsymbol{\mu}) = &-\frac{1}{2}(\mathbf{x}_a - \boldsymbol{\mu}_a)'\boldsymbol{\Lambda}_{aa}(\mathbf{x}_a - \boldsymbol{\mu}_a) - \frac{1}{2}(\mathbf{x}_a - \boldsymbol{\mu}_a)'\boldsymbol{\Lambda}_{ab}(\mathbf{x}_b - \boldsymbol{\mu}_b) \\ &- \frac{1}{2}(\mathbf{x}_b - \boldsymbol{\mu}_b)'\boldsymbol{\Lambda}_{ba}(\mathbf{x}_a - \boldsymbol{\mu}_a) - \frac{1}{2}(\mathbf{x}_b - \boldsymbol{\mu}_b)'\boldsymbol{\Lambda}_{bb}(\mathbf{x}_b - \boldsymbol{\mu}_b)\end{aligned}$$

Marginal Gaussian distributions (cont.)

$$\begin{aligned}
&\frac{1}{2}\Big(\boldsymbol{\Lambda}_{bb}\boldsymbol{\mu}_b - \boldsymbol{\Lambda}_{ba}(\mathbf{x}_a - \boldsymbol{\mu}_a)\Big)^T \boldsymbol{\Lambda}_{bb}^{-1}\Big(\boldsymbol{\Lambda}_{bb}\boldsymbol{\mu}_b - \boldsymbol{\Lambda}_{ba}(\mathbf{x}_a - \boldsymbol{\mu}_a)\Big) \\
&\qquad - \frac{1}{2}\mathbf{x}_a^T\boldsymbol{\Lambda}_{aa}\mathbf{x}_a + \mathbf{x}_a^T(\boldsymbol{\Lambda}_{aa}\boldsymbol{\mu}_a + \boldsymbol{\Lambda}_{ab}\boldsymbol{\mu}_b) + \text{const} \\
&\qquad\qquad = -\frac{1}{2}\mathbf{x}_a^T(\boldsymbol{\Lambda}_{aa} - \boldsymbol{\Lambda}_{ab}\boldsymbol{\Lambda}_{bb}^{-1}\boldsymbol{\Lambda}_{ba})\mathbf{x}_a \\
&\qquad\qquad\qquad + \mathbf{x}_a^T(\boldsymbol{\Lambda}_{aa} - \boldsymbol{\Lambda}_{ab}\boldsymbol{\Lambda}_{bb}^{-1}\boldsymbol{\Lambda}_{ba})\boldsymbol{\mu}_a + \text{const} \qquad (43)
\end{aligned}$$

where 'const' denotes quantities independent of $\mathbf{x}_a$

Marginal Gaussian distributions (cont.)

Again, by comparison with the general exponent in a Gaussian

$$-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) = -\frac{1}{2}\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x} + \mathbf{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \text{const}$$

- The covariance of the marginal distribution $p(\mathbf{x}_a)$ is given by

$$\boldsymbol{\Sigma}_a = (\boldsymbol{\Lambda}_{aa} - \boldsymbol{\Lambda}_{ab}\boldsymbol{\Lambda}_{bb}^{-1}\boldsymbol{\Lambda}_{ba})^{-1} \tag{44}$$

- The mean of the marginal distribution $p(\mathbf{x}_a)$ is given by

$$\boldsymbol{\mu}_a = \boldsymbol{\Sigma}_a(\boldsymbol{\Lambda}_{aa} - \boldsymbol{\Lambda}_{ab}\boldsymbol{\Lambda}_{bb}^{-1}\boldsymbol{\Lambda}_{ba})^{-1}\boldsymbol{\mu}_a \tag{45}$$

Marginal Gaussian distributions (cont.)

Covariance in terms of partitioned precision matrix $\boldsymbol{\Lambda} = \begin{pmatrix} \boldsymbol{\Lambda}_{aa} & \boldsymbol{\Lambda}_{ab} \\ \boldsymbol{\Lambda}_{ba} & \boldsymbol{\Lambda}_{bb} \end{pmatrix}$

We can write this in terms of partitioned covariance matrix $\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb} \end{pmatrix}$

Marginal Gaussian distributions (cont.)

$$\begin{pmatrix} \boldsymbol{\Lambda}_{aa} & \boldsymbol{\Lambda}_{ab} \\ \boldsymbol{\Lambda}_{ba} & \boldsymbol{\Lambda}_{bb} \end{pmatrix}^{-1} = \begin{pmatrix} \boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb} \end{pmatrix} \tag{46}$$

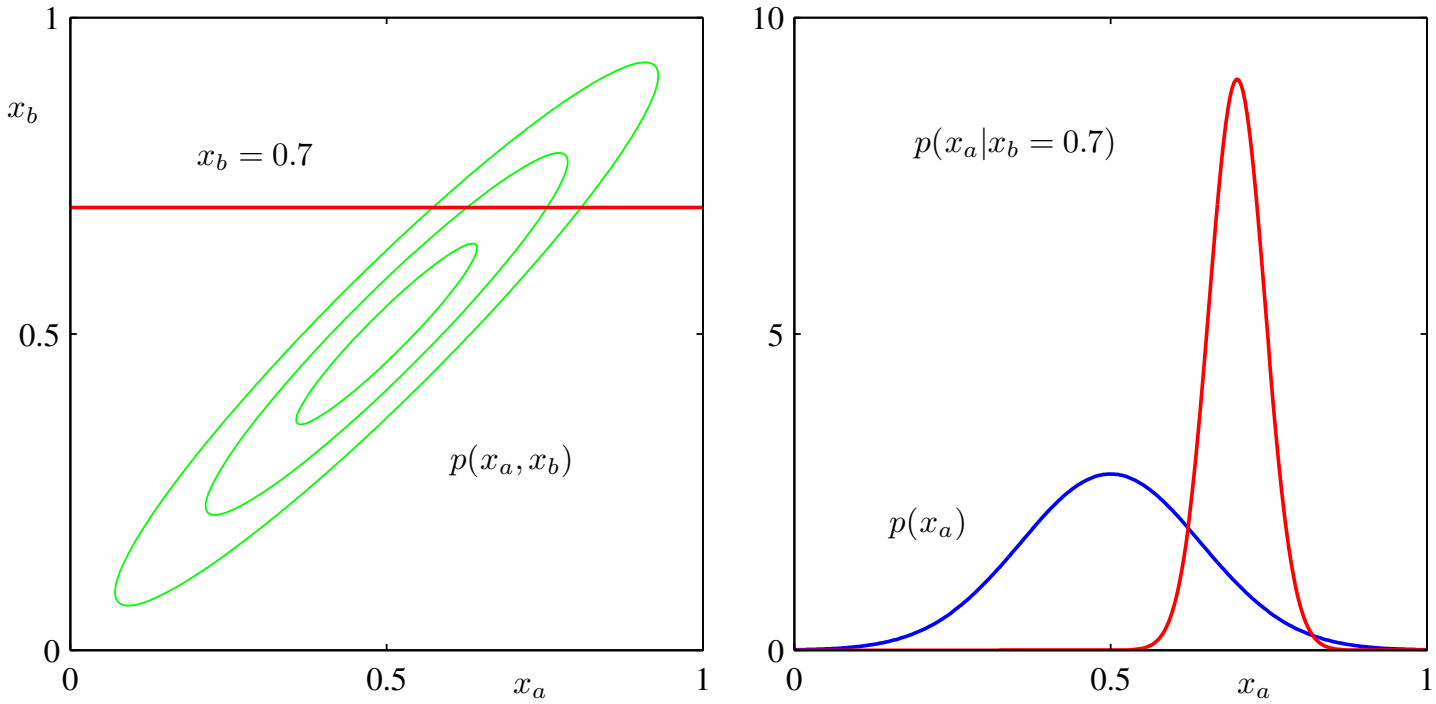
Using again $\begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix}^{-1} = \begin{pmatrix} \mathbf{M} & -\mathbf{MBD}^{-1} \\ -\mathbf{D}^{-1}\mathbf{CM} & \mathbf{D}^{-1} + \mathbf{D}^{-1}\mathbf{CMBD}^{-1} \end{pmatrix}$ we have:

$$\boldsymbol{\Sigma}_{aa} = (\boldsymbol{\Lambda}_{aa} - \boldsymbol{\Lambda}_{ab}\boldsymbol{\Lambda}_{bb}^{-1}\boldsymbol{\Lambda}_{ba}) \tag{47}$$

We obtain an intuitively satisfactory result that the marginal distribution $p(\mathbf{x}_a)$

$$\begin{aligned} \mathbb{E}[\mathbf{x}_a] &= \boldsymbol{\mu}_a \\ \text{cov}[\mathbf{x}_a] &= \boldsymbol{\Sigma}_{aa} \end{aligned} \tag{48}$$

Marginal Gaussian distributions (cont.)



Bayes's theorem for Gaussians

The Gaussian distribution

Bayes' theorem for Gaussian variables

We considered a Gaussian $p(\mathbf{x})$ in which we partitioned vector $\mathbf{x}$ into two subvectors $\mathbf{x} = (\mathbf{x}_a, \mathbf{x}_b)$ and then found expressions for

- the conditional distribution $p(\mathbf{x}_a|\mathbf{x}_b) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_{a|b}, \boldsymbol{\Lambda}_{aa}^{-1})$
- the marginal distribution $p(\mathbf{x}_a) = \mathcal{N}(\mathbf{x}_a|\boldsymbol{\mu}_a, \boldsymbol{\Sigma}_{aa})$

We also noted that the mean $\boldsymbol{\mu}_{a|b}$ of the conditional distribution $p(\mathbf{x}_a|\mathbf{x}_b)$ is a linear function of $\mathbf{x}_b$, $\boldsymbol{\mu}_{a|b} = \boldsymbol{\mu}_a + \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}(\mathbf{x}_b - \boldsymbol{\mu}_b)$

Bayes' theorem for Gaussian variables

A Gaussian marginal distribution $p(\mathbf{x})$, Gaussian conditional distribution $p(\mathbf{y}|\mathbf{x})$

- $p(\mathbf{y}|\mathbf{x})$ with mean a linear function of $\mathbf{x}$ and a covariance independent of $\mathbf{x}$

We wish to find:

- the marginal distribution $p(\mathbf{y})$
- the conditional distribution $p(\mathbf{x}|\mathbf{y})$

Bayes' theorem for Gaussian variables (cont.)

We take the marginal and conditional distributions to be

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}) \quad (49)$$

$$p(\mathbf{y}|\mathbf{x}) = \mathcal{N}(\mathbf{y}|\mathbf{A}\mathbf{x} + \mathbf{b}, \mathbf{L}^{-1}) \quad (50)$$

- $\boldsymbol{\mu}$, $\mathbf{A}$ and $\mathbf{b}$ are parameters governing the means
- $\boldsymbol{\Lambda}$ and $\mathbf{L}$ are precision matrices

If $\mathbf{x}$ is M-dimensional and $\mathbf{y}$ is D-simensional, then $\mathbf{A}$ is $D \times M$

Bayes' theorem for Gaussian variables (cont.)

First we find an expression for the joint distribution over $\mathbf{x}$ and $\mathbf{y}$, by defining

$$\mathbf{z} = \begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix} \tag{51}$$

and considering the log of the joint distribution $p(\mathbf{z}) = p(\mathbf{x}, \mathbf{y}) = p(\mathbf{y}|\mathbf{x})p(\mathbf{x})$

$$\begin{aligned} \ln p(\mathbf{z}) &= \ln p(\mathbf{x}) + \ln p(\mathbf{y}|\mathbf{x}) \\ &= -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Lambda}(\mathbf{x} - \boldsymbol{\mu}) \\ &\quad -\frac{1}{2}(\mathbf{y} - \mathbf{A}\mathbf{x} - \mathbf{b})^T \mathbf{L}(\mathbf{y} - \mathbf{A}\mathbf{x} - \mathbf{b}) \\ &\quad + \text{const} \end{aligned} \tag{52}$$

with 'const' denoting terms independent of $\mathbf{x}$ and $\mathbf{y}$

- It is again a quadratic function of the components of $\mathbf{z}$
- Thus $p(\mathbf{z})$ is again a Gaussian distribution

Bayes' theorem for Gaussian variables (cont.)

To find the precision of the Gaussian $p(\mathbf{z})$, consider the second order terms of

$$\begin{aligned}\ln p(\mathbf{z}) &= \ln p(\mathbf{x}) + \ln p(\mathbf{y}|\mathbf{x}) \\ &= -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Lambda}(\mathbf{x}-\boldsymbol{\mu}) - \frac{1}{2}(\mathbf{y}-\mathbf{A}\mathbf{x}-\mathbf{b})^T\mathbf{L}(\mathbf{y}-\mathbf{A}\mathbf{x}-\mathbf{b}) + \text{const}\end{aligned}$$

which can be written as

$$-\frac{1}{2}\mathbf{x}^T(\boldsymbol{\Lambda}+\mathbf{A}^T\mathbf{L}\mathbf{A})\mathbf{x} - \frac{1}{2}\mathbf{y}^T\mathbf{A}\mathbf{y} + \frac{1}{2}\mathbf{y}^T\mathbf{L}\mathbf{A}\mathbf{x} + \frac{1}{2}\mathbf{x}^T\mathbf{A}^T\mathbf{L}\mathbf{y}$$

$$= \frac{1}{2}\begin{pmatrix}\mathbf{x}\\ \mathbf{y}\end{pmatrix}^T\begin{pmatrix}\boldsymbol{\Lambda}+\mathbf{A}^T\mathbf{L}\mathbf{A} & -\mathbf{A}^T\mathbf{L}\\ -\mathbf{L}\mathbf{A} & \mathbf{L}\end{pmatrix}\begin{pmatrix}\mathbf{x}\\ \mathbf{y}\end{pmatrix} = -\frac{1}{2}\mathbf{z}^T\mathbf{R}\mathbf{z}$$

So the Gaussian distribution over $\mathbf{z}$ has precision (inverse covariance) matrix $\mathbf{R}$

Bayes' theorem for Gaussian variables (cont.)

The covariance matrix is found by taking the inverse of the precision matrix

$$\text{cov}[\mathbf{z}] = \mathbf{R}^{-1} = \begin{pmatrix} \mathbf{\Lambda}^{-1} & \mathbf{\Lambda}^{-1}\mathbf{A}^T \\ \mathbf{A}\mathbf{\Lambda}^{-1} & \mathbf{L}^{-1} + \mathbf{A}\mathbf{\Lambda}^{-1}\mathbf{A}^T \end{pmatrix} \tag{53}$$

- The matrix inversion formula seen earlier is used ($\star$)

Bayes' theorem for Gaussian variables (cont.)

To find the mean of the Gaussian $p(\mathbf{z})$, consider the linear terms of

$$\begin{aligned} \ln p(\mathbf{z}) &= \ln p(\mathbf{x}) + \ln p(\mathbf{y}|\mathbf{x}) \\ &= -\frac{1}{2}(\mathbf{x}-\mu)^T \mathbf{\Lambda}(\mathbf{x}-\mu) - \frac{1}{2}(\mathbf{y}-\mathbf{A}\mathbf{x}-\mathbf{b})^T \mathbf{L}(\mathbf{y}-\mathbf{A}\mathbf{x}-\mathbf{b}) + \text{const} \end{aligned}$$

which can be written as

$$\mathbf{x}^T \mathbf{\Lambda}\mu - \mathbf{x}^T \mathbf{A}^T \mathbf{L}\mathbf{b} + \mathbf{y}^T \mathbf{L}\mathbf{b} = \begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix}^T \begin{pmatrix} \mathbf{\Lambda}\mu - \mathbf{A}^T \mathbf{L}\mathbf{b} \\ \mathbf{L}\mathbf{b} \end{pmatrix} \tag{54}$$

Bayes' theorem for Gaussian variables (cont.)

Using our earlier result in Equation 29[6], we find that the mean of $\mathbf{z}$ is

$$\mathbb{E}[\mathbf{z}] = \mathbf{R}^{-1} \begin{pmatrix} \boldsymbol{\Lambda}\boldsymbol{\mu} - \mathbf{A}^T\mathbf{L}\mathbf{b} \\ \mathbf{L}\mathbf{b} \end{pmatrix} \tag{55}$$

We can now make use of Equation 53[7] for the covariance of $\mathbf{z}$ to get

$$\mathbb{E}[\mathbf{z}] = \begin{pmatrix} \boldsymbol{\mu} \\ \mathbf{A}\boldsymbol{\mu} + \mathbf{b} \end{pmatrix} \tag{56}$$

---

[6] $-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) = -\frac{1}{2}\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x} + \mathbf{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \text{const}$

[7] $\text{cov}[\mathbf{z}] = \mathbf{R}^{-1} = \begin{pmatrix} \boldsymbol{\Lambda}^{-1} & \boldsymbol{\Lambda}^{-1}\mathbf{A}^T \\ -\mathbf{A}\boldsymbol{\Lambda}^{-1} & \mathbf{L}^{-1} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^T \end{pmatrix}$

Bayes' theorem for Gaussian variables (cont.)

We turn now our attention to the marginal distribution $p(\mathbf{y})$

- which we have marginalised over $\mathbf{x}$

The marginal distribution over a subset of components of the random vector takes a simple form when expressed in terms of partitioned covariance matrix

$$\mathbb{E}[\mathbf{x}_a] = \boldsymbol{\mu}_a \qquad \text{and} \qquad \text{cov}[\mathbf{x}_a] = \boldsymbol{\Sigma}_{aa}$$

Use $\text{cov}[\mathbf{z}] = \mathbf{R}^{-1} = \begin{pmatrix} \boldsymbol{\Lambda}^{-1} & \boldsymbol{\Lambda}^{-1}\mathbf{L} \\ \mathbf{A}\boldsymbol{\Lambda}^{-1} & \mathbf{L}^{-1} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^T \end{pmatrix}$ and $\mathbb{E}[\mathbf{z}] = \begin{pmatrix} \boldsymbol{\mu} \\ \mathbf{A}\boldsymbol{\mu} + \mathbf{b} \end{pmatrix}$ then

$$\mathbb{E}[\mathbf{y}] = \mathbf{A}\boldsymbol{\mu} + \mathbf{b} \tag{57}$$

$$\text{cov}[\mathbf{y}] = \mathbf{L}^{-1}\mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^T \tag{58}$$

are the searched mean and covariance of the marginal distribution $p(\mathbf{y})$

Bayes' theorem for Gaussian variables (cont.)

Finally, we seek an expression for the conditional $p(\mathbf{x}|\mathbf{y})$

The conditional distribution is easier in terms of partitioned precision matrix

$$\boldsymbol{\Sigma}_{a|b} = \boldsymbol{\Lambda}_{aa}^{-1} \qquad \text{and} \qquad \boldsymbol{\mu}_{a|b} = \boldsymbol{\mu}_a - \boldsymbol{\Sigma}_{aa}^{-1}\boldsymbol{\Sigma}_{ab}(\mathbf{x}_b - \boldsymbol{\mu}_b)$$

Use $\text{cov}[\mathbf{z}] = \mathbf{R}^{-1} = \begin{pmatrix} \boldsymbol{\Lambda}^{-1} & \boldsymbol{\Lambda}^{-1}\mathbf{L} \\ \mathbf{A}\boldsymbol{\Lambda}^{-1} & \mathbf{L}^{-1} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^T \end{pmatrix}$ and $\mathbb{E}[\mathbf{z}] = \begin{pmatrix} \boldsymbol{\mu} \\ \mathbf{A}\boldsymbol{\mu} + \mathbf{b} \end{pmatrix}$ then

$$\mathbb{E}[\mathbf{x}|\mathbf{y}] = (\boldsymbol{\Lambda} + \mathbf{A}^T\mathbf{L}\mathbf{A}^{-1})\Big(\mathbf{A}^T\mathbf{L}(\mathbf{y} - \mathbf{b}) + \boldsymbol{\Lambda}\boldsymbol{\mu}\Big) \tag{59}$$

$$\text{cov}[\mathbf{x}|\mathbf{y}] = (\boldsymbol{\Lambda} + \mathbf{A}^T\mathbf{L}\mathbf{A})^{-1} \tag{60}$$

are the searched mean and covariance of the conditional distribution $p(\mathbf{x}|\mathbf{y})$

Bayes' theorem for Gaussian variables (cont.)

So what about the Bayes' theorem?

The evaluation of the conditional $p(\mathbf{x}|\mathbf{y})$ is an example of Bayes' theorem

- We can interpret the distribution $p(\mathbf{x})$ as a prior over $\mathbf{x}$

The conditional distribution $p(\mathbf{x}|\mathbf{y})$ is the corresponding posterior over $\mathbf{x}$

- If $\mathbf{y}$ is observed

Having found the marginal and conditional distributions, we effectively expressed the joint distribution $p(\mathbf{z}) = p(\mathbf{x})p(\mathbf{y}|\mathbf{x})$ in the form $p(\mathbf{x}|\mathbf{y})p(\mathbf{y})$

Maximum likelihood
The Gaussian distribution

Maximum likelihood for the Gaussian

Given data $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_N)^T$ in which the observations $\{\mathbf{x}_n\}$ are assumed to be drawn independently from a multivariate Gaussian distribution, we can estimate the parameters of the distribution by maximum likelihood

- The log likelihood function is

$$\ln p(\mathbf{X}|\boldsymbol{\mu}), \boldsymbol{\Sigma} = \frac{ND}{2}\ln(2\pi) - \frac{N}{2}\ln|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}_n - \boldsymbol{\mu}) \tag{61}$$

The likelihood function depends on the data set on thru terms

$$\sum_{n=1}^{N}\mathbf{x}_n \qquad \text{and} \qquad \sum_{n=1}^{N}\mathbf{x}_n\mathbf{x}_n^T \tag{62}$$

These are the **sufficient statistics** for the Gaussian distribution

Maximum likelihood for the Gaussian (cont.)

The derivative of log likelihood with respect to $\boldsymbol{\mu}$ can be written as [8]

$$\frac{\partial}{\partial \boldsymbol{\mu}} \ln p(\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \boldsymbol{\Sigma}^{-1}(\mathbf{x}_n - \boldsymbol{\mu}) \tag{63}$$

Set to zero, it gets us the maximum likelihood estimate of the mean

$$\boldsymbol{\mu}_{ML} = \frac{1}{N} \sum_{n=1}^{N} \mathbf{x}_n \tag{64}$$

which is, as expected, the mean of the observed set of data points

[8] We used $\frac{\partial}{\partial \mathbf{x}}(\mathbf{x}^T \mathbf{a}) = \frac{\partial}{\partial \mathbf{x}}(\mathbf{a}^T \mathbf{x}) = \mathbf{a}$

Maximum likelihood for the Gaussian (cont.)

Maximisation of the log likelihood function $\ln p(\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ wrt to $\boldsymbol{\Sigma}$ is tough $(\star)$
As expected, the result takes the form

$$\boldsymbol{\Sigma}_{ML} = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu}_{ML})(\mathbf{x}_n - \boldsymbol{\mu}_{ML})^T \tag{65}$$

It involves $\boldsymbol{\mu}_{ML}$ as a result of the joint maximisation wrt $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, whereas the solution for $\boldsymbol{\mu}_{ML}$ does not depend on $\boldsymbol{\Sigma}_{ML}$ (first solve for $\boldsymbol{\mu}_{ML}$ and then $\boldsymbol{\Sigma}_{ML}$)

Maximum likelihood for the Gaussian (cont.)

If we evaluate the expectations of the maximum likelihood solutions under the true distribution, we obtain the following results

$$\mathbb{E}[\boldsymbol{\mu}_{ML}] = \boldsymbol{\mu} \tag{66}$$

$$\mathbb{E}[\boldsymbol{\Sigma}_{ML}] = \frac{N-1}{N}\boldsymbol{\Sigma} \tag{67}$$

- The expectation of the maximum likelihood estimate for the mean is equal to the true mean, unbiased
- The maximum likelihood estimate for the covariance underestimates the true covariance, it is biased

We correct the bias, by defining the estimator $\tilde{\boldsymbol{\Sigma}}$

$$\tilde{\boldsymbol{\Sigma}} = \frac{1}{N-1}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu}_{ML})(\mathbf{x}_n - \boldsymbol{\mu}_{ML})^T \tag{68}$$

which has an expectation that equals the true $\boldsymbol{\Sigma}$

Bayesian inference for the Gaussian
The Gaussian distribution

Bayesian inference for the Gaussian

The maximum likelihood framework gives point estimates for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$

Now we develop a Bayesian treatment by introducing prior distributions

- over these parameters

We start simple, with a single Gaussian random variable x

- We will suppose that the variance σ^2 is known
- We consider the task of inferring the mean μ

We are also given a set of N observations $\mathbf{x} = \{x_1, \ldots, x_n\}$

Bayesian inference for the Gaussian (cont.)

The likelihood function can be viewed as a function of μ and takes the form

$$p(\mathbf{x}|\mu) = \prod_{n=1}^{N} p(x_n|\mu) = \frac{1}{(2\pi\sigma^2)} \exp\left(-\frac{1}{2\sigma^2} \sum_{n=1}^{N} (x_n - \mu)^2 \right) \tag{69}$$

It takes the form of the exponential of a quadratic form in μ and if we choose a Gaussian prior $p(\mu)$, it is a conjugate distribution for the likelihood function

- The posterior is a product of two exponentials of quadratic functions of μ
- Thus, it is also a Gaussian

Bayesian inference for the Gaussian (cont.)

We take our prior distribution to be

$$p(\mu) = \mathcal{N}(\mu|\mu_0, \sigma_0^2) \tag{70}$$

The posterior distribution is given by

$$p(\mu|\mathbf{x}) \propto p(\mathbf{x}|\mu)p(\mu) \tag{71}$$

Some manipulations ($\star$) involving completing the square in the exponent allow to show that the posterior distribution is given by

$$p(\mu|\mathbf{x}) = \mathcal{N}(\mu|\mu_N, \sigma_N^2) \tag{72}$$

$$\mu_N = \frac{\sigma^2}{N\sigma_0^2 + \sigma_0^2}\mu_0 + \frac{N\sigma^2}{N\sigma^2 + \sigma^2}\mu_{ML} \tag{73}$$

$$\frac{1}{\sigma_N^2} = \frac{1}{\sigma_0^2} + \frac{N}{\sigma^2} \tag{74}$$

$\mu_{ML} = \frac{1}{N}\sum_{n=1}^{N} x_n$ is the maximum likelihood estimate for μ, the sample mean

Bayesian inference for the Gaussian (cont.)

Note that the mean of the posterior distribution $p(\mu|\mathbf{x})$ is a compromise between the prior mean μ_0 and the maximum likelihood solution μ_{ML}

$$\mu_N = \frac{\sigma^2}{N\sigma_0^2 + \sigma_0^2}\mu_0 + \frac{N\sigma^2}{N\sigma^2 + \sigma^2}\mu_{ML}$$

- It reduces to the prior mean, if the number of observed data points $N = 0$
- For $N \to \infty$, the posterior mean equals the maximum likelihood solution

Bayesian inference for the Gaussian (cont.)

Consider the result for the variance of the posterior distribution $p(\mu|\mathbf{x})$

$$\frac{1}{\sigma_N^2} = \frac{1}{\sigma_0^2} + \frac{N}{\sigma^2}$$

It is most naturally expressed in terms of inverse variance, or precision

Precisions are additive, so the precision of the posterior is given by

- the precision of the prior, plus one contribution of the data precision from each of the observed data poins

Bayesian inference for the Gaussian (cont.)

$$\frac{1}{\sigma_N^2} = \frac{1}{\sigma_0^2} + \frac{N}{\sigma^2}$$

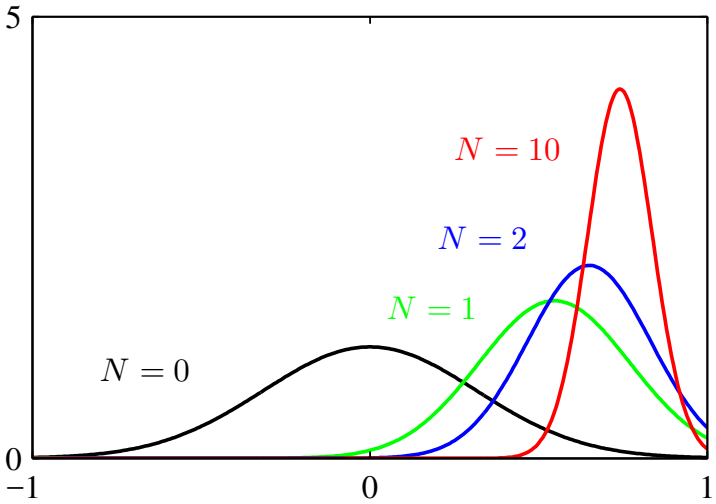
As we increase the number of observed data points, the precision steadily increases, giving a posterior distribution with steadily decreasing variance

- With no observed data points, we have the prior variance
- If $N \to \infty$, variance $\sigma_N^2 \to 0$ and the posterior distribution becomes infinitely peaked around the ML solution μ_{ML}

Bayesian inference for the Gaussian (cont.)

Bayesian inference for the mean μ of a Gaussian with known variance σ

- Data from a Gaussian of mean 0.8 and variance 0.1



The prior is chosen to have mean 0

In both prior and likelihood function, the variance is set to the true value

- the prior distribution over μ, the curve $N = 0$, itself Gaussian
- the posterior distribution for increasing numbers N of points

We see that the maximum likelihood result of a point estimate for μ is recovered precisely from the Bayesian formalism in the limit $N \to \infty$

For finite N, in the limit $\sigma_0^2 \to \infty$ in which the prior has infinite variance then the posterior mean is the ML result, while the posterior variance is $\sigma_N^2 = \sigma^2/N$

Bayesian inference for the Gaussian (cont.)

The analysis of Bayesian inference for the mean of a D-dimensional Gaussian variable $\mathbf{x}$ with known covariance and unknown mean is straightforward ($\star$)

Bayesian inference for the Gaussian (cont.)

So far, we have assumed that the variance of the Gaussian distribution over the data is known and our goal is to infer the mean

- Let us suppose that the mean is known and we wish to infer the variance

Calculations are simpler, if we choose a conjugate form for the prior

- The likelihood function for $\lambda = 1/\sigma^2$ is

$$p(\mathbf{x}|\lambda) = \prod_{n=1}^{N} \mathcal{N}(x_n|\mu, \lambda^{-1}) \propto \lambda^{N/2} \exp\left(-\frac{\lambda}{2} \sum_{n=1}^{N} (x_n - \mu)^2 \right) \tag{75}$$

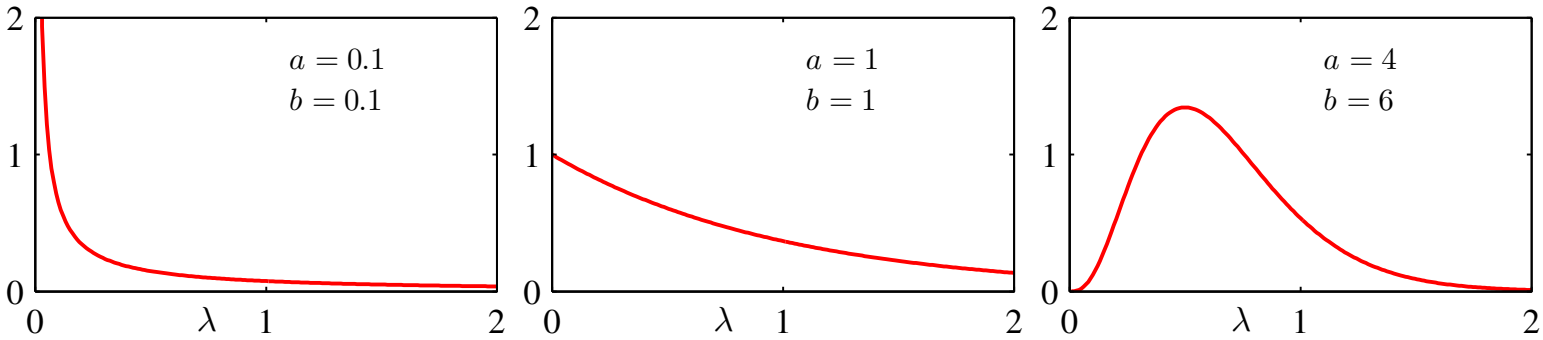
The corresponding conjugate prior should therefore be proportional to the product of a power of λ and the exponential of a linear function of λ

Bayesian inference for the Gaussian (cont.)

This corresponds to the **gamma distribution**, which is defined as

$$\text{Gam}(\lambda|a,b) = \frac{1}{\Gamma(a)} b^a \lambda^{a-1} \exp(-b\lambda) \tag{76}$$

The gamma function $\Gamma(\cdot)$ assures a correct normalisation



The integral of the gamma distribution is finite if $a > 0$
The distribution itself is finite for $a \geq 1$

The mean and variance of the gamma distribution are $\begin{cases} \mathbb{E}[\lambda] = a/b \\ \text{var}[\lambda] = a/b^2 \end{cases}$

Bayesian inference for the Gaussian (cont.)

Consider a prior distribution $\text{Gam}(\lambda|a_0, b_0)$, multiply by the likelihood function

- Obtain a posterior distribution

$$p(\lambda|\mathbf{x}) \propto \lambda^{a_0-1}\lambda^{N/2}\exp\left(-b_0\lambda - \frac{\lambda}{2}\sum_{n=1}^{N}(x_n-\mu)^2\right) \tag{77}$$

- Recognise the Gamma distribution $\text{Gam}(\lambda|a_N, b_N)$, with

$$a_N = a_0 + \frac{N}{2} \tag{78}$$

$$b_N = b_0 + \frac{1}{2}\sum_{n=1}^{N}(x_n-\mu)^2 = b_0 + \frac{N}{2}\sigma^2_{ML} \tag{79}$$

where σ^2_{ML} is the maximum likelihood estimator of the variance

Bayesian inference for the Gaussian (cont.)

Note that there is no need to keep track of the normalisation constants in the prior and in the likelihood function in

$$p(\lambda|\mathbf{x}) \propto \lambda^{a_0-1}\lambda^{N/2}\exp\left(-b_0\lambda - \frac{\lambda}{2}\sum_{n=1}^{N}(x_n-\mu)^2\right)$$

Because, if required, the correct coefficient can be found at the end using the normalised form of the gamma distro

$$\mathrm{Gam}(\lambda|a,b) = \frac{1}{\Gamma(a)}b^a\lambda^{a-1}\exp(-b\lambda)$$

Bayesian inference for the Gaussian (cont.)

The effect of observing N points increases the value of coefficient a by $N/2$

$$a_N = a_0 + \frac{N}{2}$$

We can interpret a_0 in the prior in terms of $2a_0$ 'effective' prior observations

Similarly, N points contribute $N\sigma^2_{ML}/2$ to parameter b, σ^2_{ML} is the variance

$$b_N = b_0 + \frac{1}{2}\sum_{n=1}^{N}(x_n - \mu)^2 = b_0 + \frac{N}{2}\sigma^2_{ML}$$

We interpret b_0 in the prior as arising from the $2a_0$
'effective' prior observations having variance $2b_0/(2a_0) = b_0/a_0$

Bayesian inference for the Gaussian (cont.)

Suppose now that both mean and variance are unknown

To find a conjugate prior, consider the dependence of the likelihood on μ and λ

$$p(\mathbf{x}|\mu,\lambda) = \prod_{n=1}^{N}\left(\frac{\lambda}{2\pi}\right)\exp\left(-\frac{\lambda}{2}(x_n-\mu)^2\right)$$

$$\propto \left(\lambda^{1/2}\exp\left(-\frac{\lambda\mu^2}{2}\right)\right)^N \exp\left(\lambda\mu\sum_{n=1}^{N}x_n - \frac{\lambda}{2}\sum_{n=1}^{N}x_n^2\right) \quad (80)$$

We wish to identify the prior distribution $p(\mu,\lambda)$ with the same functional dependence on μ and λ as the likelihood function above

Bayesian inference for the Gaussian (cont.)

$$\begin{aligned} p(\mu, \lambda) &\propto \left(\lambda^{1/2} \exp\left(-\frac{\lambda\mu^2}{2}\right)\right)^{\beta} \exp\left(c\lambda\mu - d\lambda\right) \\ &= \exp\left(-\frac{\beta\lambda}{2}(\mu - c/\beta)^2\right)\lambda^{\beta/2} \exp\left(-\left(d - \frac{c^2}{2\beta}\right)\lambda\right) \end{aligned} \quad (81)$$

where c, d and β are constant

Because $p(\mu, \lambda) = p(\mu|\lambda)p(\lambda)$, we can find $p(\mu|\lambda)$ and $p(\lambda)$ by inspection

- $p(\mu|\lambda)$ is a Gaussian whose precision is a linear function of λ
- $p(\lambda)$ is a gamma distribution

Bayesian inference for the Gaussian (cont.)

The normalised prior takes the form

$$p(\mu, \lambda) = \mathcal{N}\left(\mu|\mu_o, (\beta\lambda)^{-1}\right)\text{Gam}(\lambda|a, b) \tag{82}$$

where we defined the new constants

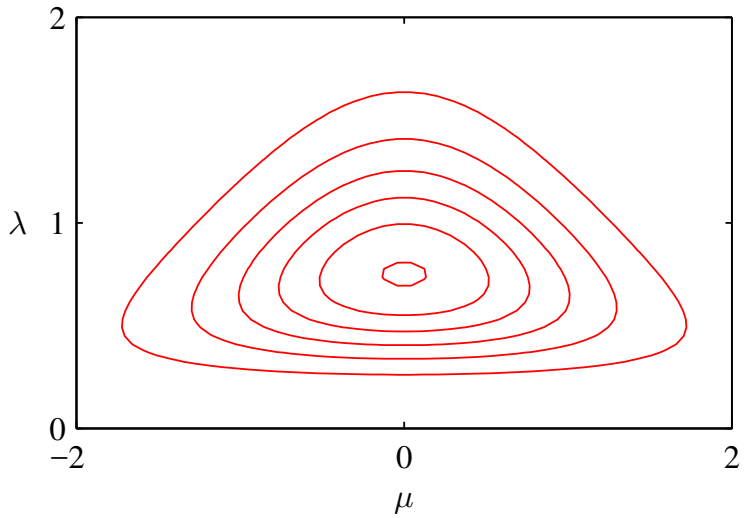
- $\mu_0 = c/\beta$
- $a = (1 + \beta)/2$
- $b = d - c^2/2\beta$

This distribution is called **normal-gamma** or **Gaussian-gamma distribution**

- It is not simply the product of an independent Gaussian prior over μ and gamma prior over λ, because the precision of μ is a linear function of λ
- Even if we choose a prior in which μ and λ are independent, the posterior distribution would exhibit coupling between precision of μ and value of λ

Bayesian inference for the Gaussian (cont.)

Contour plot of the normal-gamma distribution



Parameter values $\mu_0 = 0$, $\beta = 2$, $a = 5$ and $b = 6$

Bayesian inference for the Gaussian (cont.)

In the case of a D-variate Gaussian distribution $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1})$ over a variable $\mathbf{x}$

- the conjugate prior distribution for the mean $\boldsymbol{\mu}$ assuming a known precision is again a Gaussian distribution
- the conjugate prior distribution for the precision matrix $\boldsymbol{\Lambda}$ assuming a known mean is a **Wishart distribution**

$$\mathbf{W}(\boldsymbol{\Lambda}|\mathbf{W}, \nu) = B|\boldsymbol{\Lambda}|^{(\nu-D-1)/2} \exp\left(-\frac{1}{2}\mathrm{Tr}(\mathbf{W}^{-1}\boldsymbol{\Lambda})\right) \tag{83}$$

where ν is the **number of degrees of freedom** of the distribution, $\mathbf{W}$ is a $D \times D$ scale matrix and $\mathrm{Tr}(\cdot)$ denotes the trace of a matrix

The normalisation constant B is given by

$$B(\mathbf{W}, \nu) = |\mathbf{W}|^{-\nu/2}\left(2^{\nu D/2}\pi^{D(D-1)/4}\prod_{i=1}^{D}\Gamma\left(\frac{\nu+1-1}{2}\right)\right)^{-1} \tag{84}$$

Bayesian inference for the Gaussian (cont.)

- the conjugate prior distribution assuming both mean $\boldsymbol{\mu}$ and precision $\boldsymbol{\Lambda}$ unknown is a **normal-Wishart** or **Gaussian-Wishart distribution**

$$p(\boldsymbol{\mu}, \boldsymbol{\Lambda}|\boldsymbol{\mu}_0, \beta, \mathbf{W}, \nu) = \mathcal{N}\Big(\boldsymbol{\mu}|\boldsymbol{\mu}_0, (\beta\boldsymbol{\Lambda}^{-1})\Big)\mathcal{W}(\boldsymbol{\Lambda}|\mathbf{W}, \nu) \tag{85}$$

Student's t-distribution
The Gaussian distribution

Student's t-distribution

The conjugate prior for the precision of a Gaussian is a gamma distribution

If we have a univariate Gaussian $\mathcal{N}(x|\mu, \tau^{-1})$ together with a gamma prior $\text{Gam}(\tau|a, b)$ and we integrate out precision, we get the marginal of x

$$\begin{aligned} p(x|\mu, a, b) &= \int_0^{+\infty} \mathcal{N}(x|\mu, \tau^{-1})\text{Gam}(\tau|a, b)d\tau \\ &= \int_0^{+\infty} \frac{b^a e^{-b\tau}\tau^{a-1}}{\Gamma(a)}\left(\frac{\tau}{2\pi}\right)^{1/2} \exp\left(-\frac{\tau}{2}(x-\mu)^2\right)d\tau \\ &= \frac{b^a}{\Gamma(a)}\left(\frac{1}{2\pi}\right)^{1/2}\left(b + \frac{(x-\mu)^2}{2}\right)^{-a-1/2}\Gamma(a+1/2) \end{aligned} \quad (86)$$

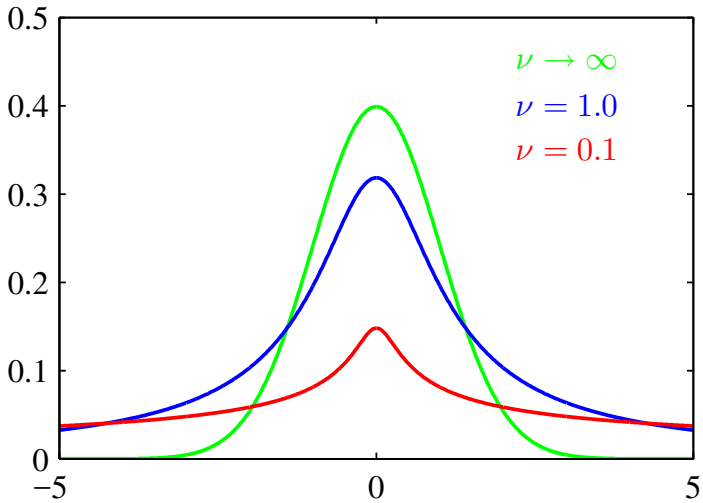
where we can make the change of variables $z = \tau\left(b + (x-\mu)^2/2\right)$

Student's t-distribution (cont.)

By convention, we define new parameters $\nu = 2a$ and $\lambda = a/b$ to get

$$p(x|\mu, a, b) = \text{Stu}(x|\mu, \lambda, \nu) = \frac{\Gamma(\frac{\nu}{2} + \frac{1}{2})}{\Gamma(\frac{\nu}{2})} \left(\frac{\lambda}{\pi\nu}\right)^{1/2} \left(1 + \frac{\lambda(x-\mu)^2}{\nu}\right)^{-\nu/2-1/2} \tag{87}$$

which is known as **Student's t-distribution** with parameter λ being the **precision** and parameter ν being called the **degree of freedom**



Student's t-distribution for $\mu = 0$, $\lambda = 1$ and for various values of ν

The limit $\nu \to \infty$ equals a Gaussian with mean μ and precision λ

For $\nu = 1$, the t-distribution reduces to the Cauchy distribution

Student's t-distribution (cont.)

$$\int_0^{+\infty} \mathcal{N}(x|\mu, \tau^{-1})\text{Gam}(\tau|a, b)d\tau$$

The Student's t-distribution is obtained by adding up an infinite number of Gaussian distributions having the same mean but different precisions

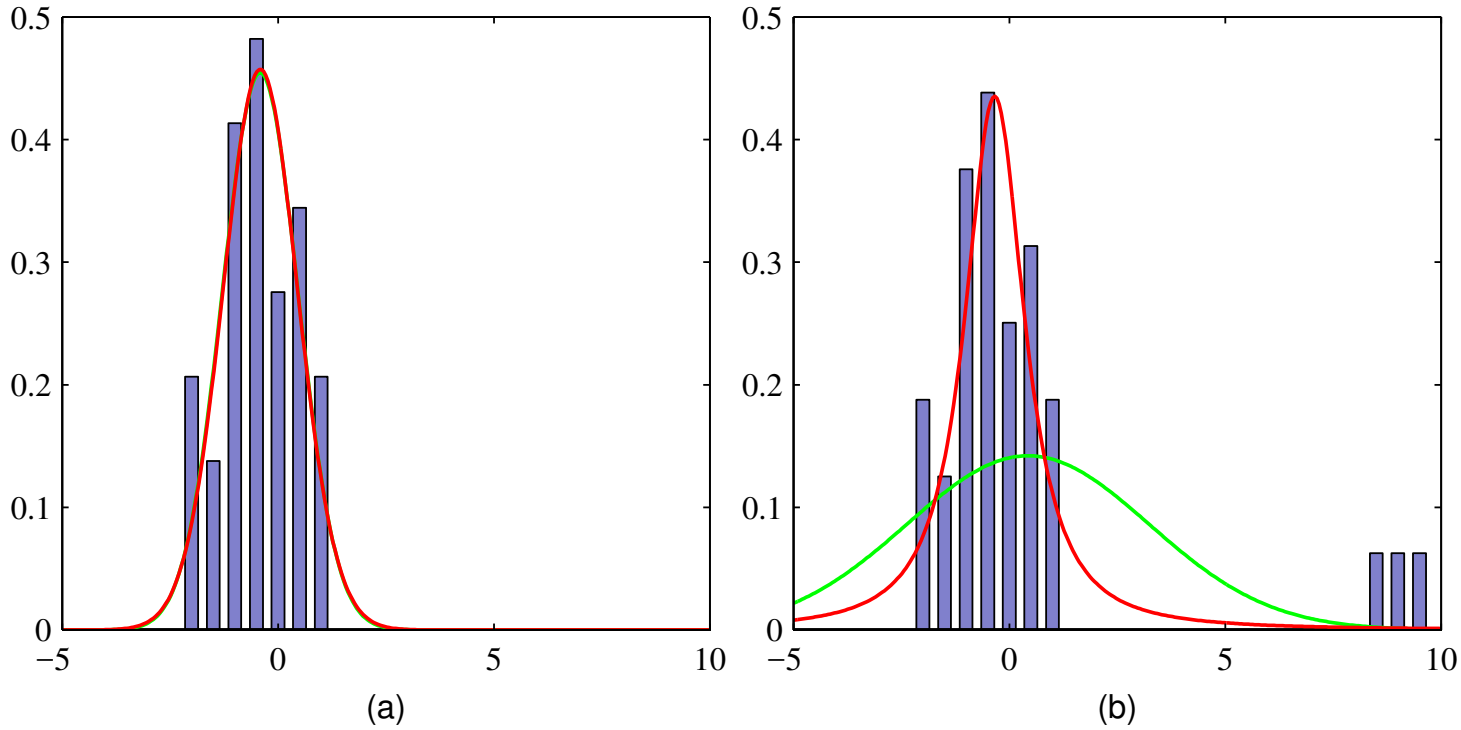
- This can be interpreted as an infinite mixture of Gaussians

The result is a distribution that in general has longer 'tails' than a Gaussian

- The t-distribution has an important property called robustness
- It is less sensitive than the Gaussian to the presence of outliers

Student's t-distribution (cont.)

Maximum likelihood fits for a Gaussian (green) and a t-distribution (red)



The Gaussian is strongly distorted by the presence of outlying points

Student's t-distribution (cont.)

Substituting parameters $\nu = 2a$, $\lambda = 1/b$ and $\eta = \tau b/a$, the t-distribution is

$$\begin{aligned} \mathrm{St}(x|\mu,\lambda,\nu) &= \int_0^{+\infty} \mathcal{N}\left(x|\mu,(\eta\lambda)^{-1}\right)\mathrm{Gam}(\frac{\eta/\nu}{2},\nu/2)d\eta \quad (88)\\ &= \frac{\Gamma(\frac{\nu}{2}+\frac{1}{2})}{\Gamma(\frac{\nu}{2})}\left(\frac{\lambda}{\pi\nu}\right)^{1/2}\left(1+\frac{\lambda(x-\mu)^2}{\nu}\right)^{-\nu/2-1/2} \quad (89)\end{aligned}$$

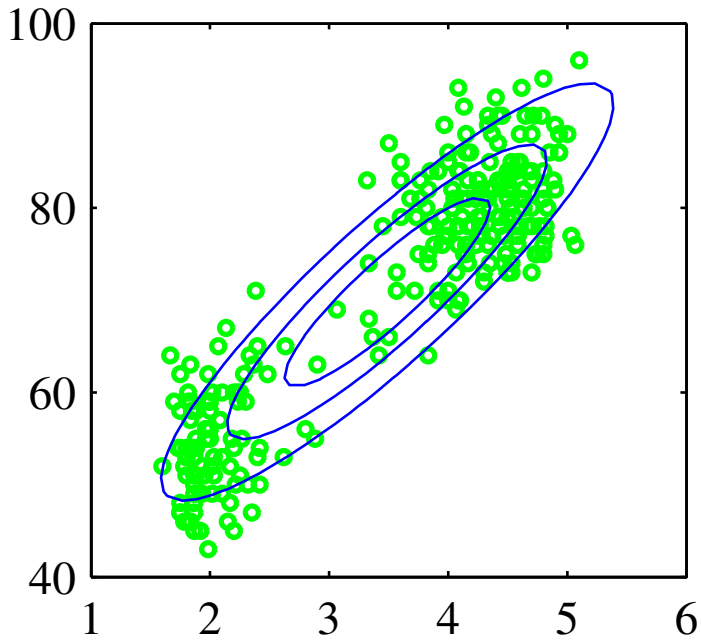
We can also generalise this to a D-dimensional Gaussian $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Lambda})$ to get

$$\begin{aligned} \mathrm{St}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Lambda},\nu) &= \int_0^{+\infty} \mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu},(\eta\boldsymbol{\Lambda})^{-1}\right)\mathrm{Gam}(\frac{\eta/\nu}{2},\nu/2)d\eta \quad (90)\\ &= \frac{\Gamma(\frac{D}{2}+\frac{\nu}{2})}{\Gamma(\frac{\nu}{2})}\frac{|\boldsymbol{\Lambda}|^{1/2}}{(\pi\nu)^{D/2}}\left(1+\frac{(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Lambda}(\mathbf{x}-\boldsymbol{\mu})}{\nu}\right)^{-D/2-\nu/2} \quad (91)\end{aligned}$$

Mixtures of Gaussians
The Gaussian distribution

Mixtures of Gaussians

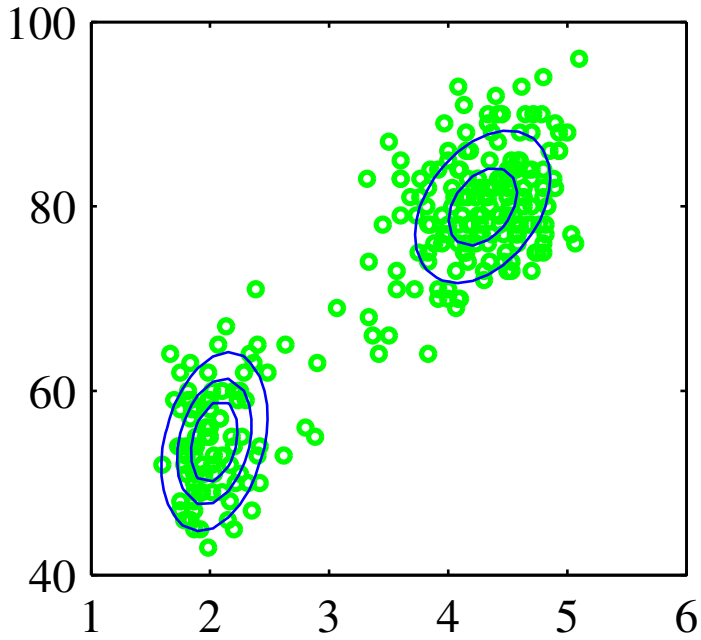
While the Gaussian distribution has some important analytical properties, it suffers from significant limitations when it comes to modelling real data sets



A Gaussian distribution fitted to data using maximum likelihood

This distribution fails to capture the two clumps in the data and places much of its probability mass in the central region between the clumps where the data are relatively sparse

Mixtures of Gaussians (cont.)



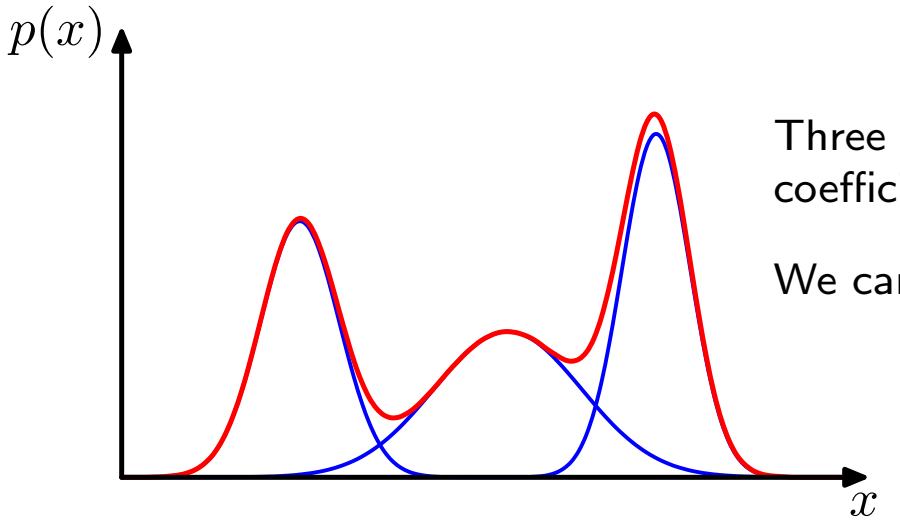
A linear combination of two Gaussians fitted using maximum likelihood

A linear superposition gives a better characterisation of the data set

Superpositions, by taking linear combinations of basic distributions such as Gaussians, can be formulated as probabilistic models (**mixture distributions**)

Mixtures of Gaussians (cont.)

A Gaussian mixture distribution in one dimension



Three Gaussians (blue, scaled by a coefficient) and their sum (red)

We can get very complex densities

By using a sufficient number of Gaussians, and by adjusting their means and covariances as well as the coefficients in the linear combination, almost any continuous density can be approximated to arbitrary accuracy

Mixtures of Gaussians (cont.)

Definition

We consider a linear superposition of K Gaussian densities of the form

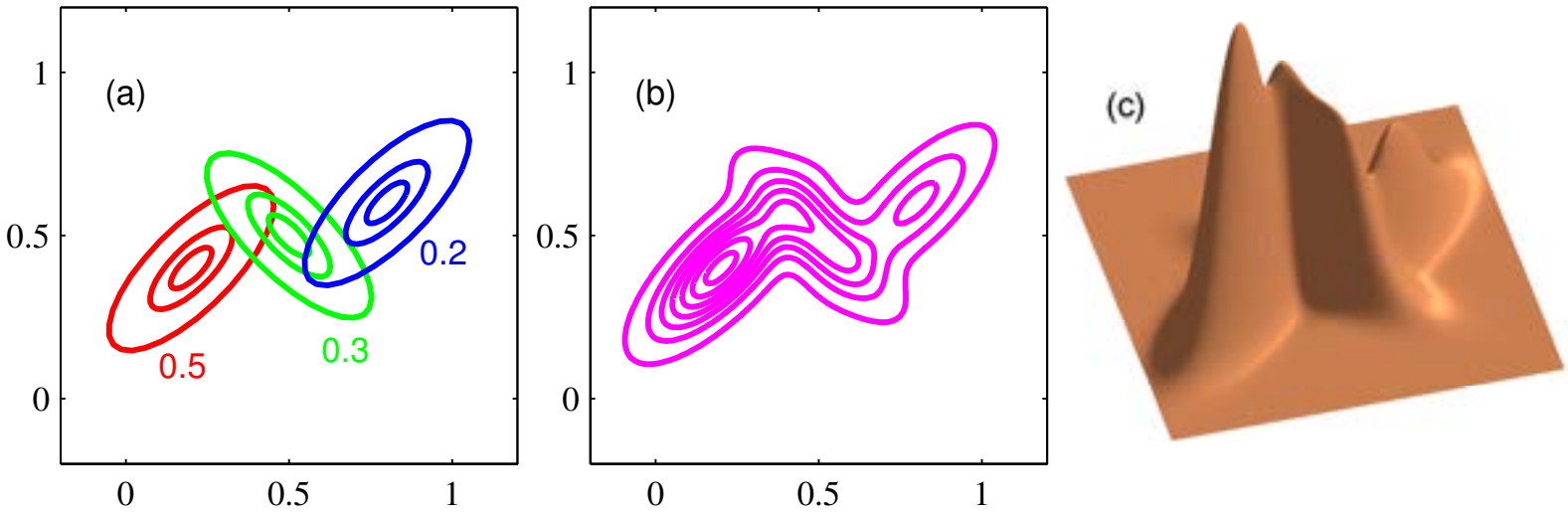
$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{92}$$

Such a model is a **mixture of Gaussians**, each Gaussian density $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ is a **component** of the mixture, with its own mean $\boldsymbol{\mu}_k$ and covariance $\boldsymbol{\Sigma}_k$

The parameters π_k in $p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ are **mixing coefficients**

Mixtures of Gaussians (cont.)

A Gaussian mixture distribution in two dimensions



- The contours at constant density for each of the mixture components
- The contours at constant density of the mixture distribution $p(\mathbf{x})$
- The surface plot of the mixture distribution $p(\mathbf{x})$

The numbers in the first plot are the mixing coefficients

Mixtures of Gaussians (cont.)

If we integrate both sides of $p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ with respect to $\mathbf{x}$ and note that both $p(\mathbf{x})$ and the Gaussian components are normalised, we get

$$\sum_{k=1}^{K} \pi_k = 1 \tag{93}$$

The requirements that $p(\mathbf{x}) \geq 0$ and $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \geq 0$ imply that $0 \leq \pi_k \leq 1$

Mixtures of Gaussians (cont.)

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

$\pi_k \in [0, 1]$ and $\sum \pi_k = 1$: The mixing coefficients can be seen as probabilities

From sum and probabilities rules, the marginal density is given by

$$p(\mathbf{x}) = \sum_{k=1}^{K} p(k) p(\mathbf{x}|k) \tag{94}$$

By $p(k) = \pi_k$, we view coefficients as prior probabilities (of picking the k-th component) and $p(\mathbf{x}|k) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ the probability of $\mathbf{x}$ conditioned on k

Mixtures of Gaussians (cont.)

A role is played by posterior probabilities (or **responsibilities**) $p(k|\mathbf{x})$

By the Bayes' theorem, the posterior probabilities can be written as

$$\begin{aligned} \gamma_k(\mathbf{x}) &\equiv p(k|\mathbf{x}) \\ &= \frac{p(k)p(\mathbf{x}|k)}{\sum_l p(l)p(\mathbf{x}|l)} \\ &= \frac{\pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_\mathbf{k})}{\sum_l \pi_l \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_\mathbf{l})} \end{aligned} \tag{95}$$

Mixtures of Gaussians (cont.)

The Gaussian mixture distribution is governed by the parameters $\boldsymbol{\pi}$, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$

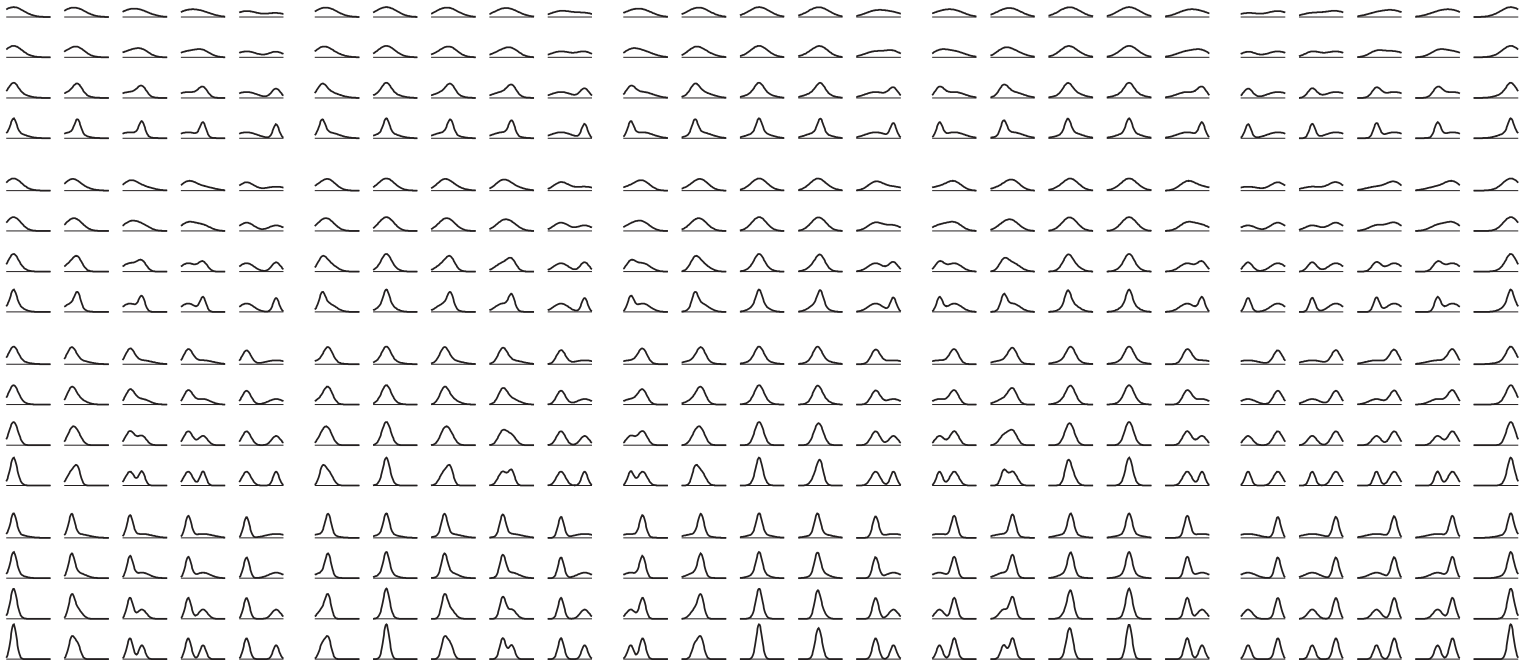
$$\begin{aligned} \boldsymbol{\pi} &\equiv \{\pi_1, \ldots, \pi_K\} \\ \boldsymbol{\mu} &\equiv \{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K\} \\ \boldsymbol{\Sigma} &\equiv \{\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_K\} \end{aligned}$$

One way to set the values of the parameters is to use (log) maximum likelihood

$$\ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \ln \left(\sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right) \tag{96}$$

Maximum likelihood (cont.)

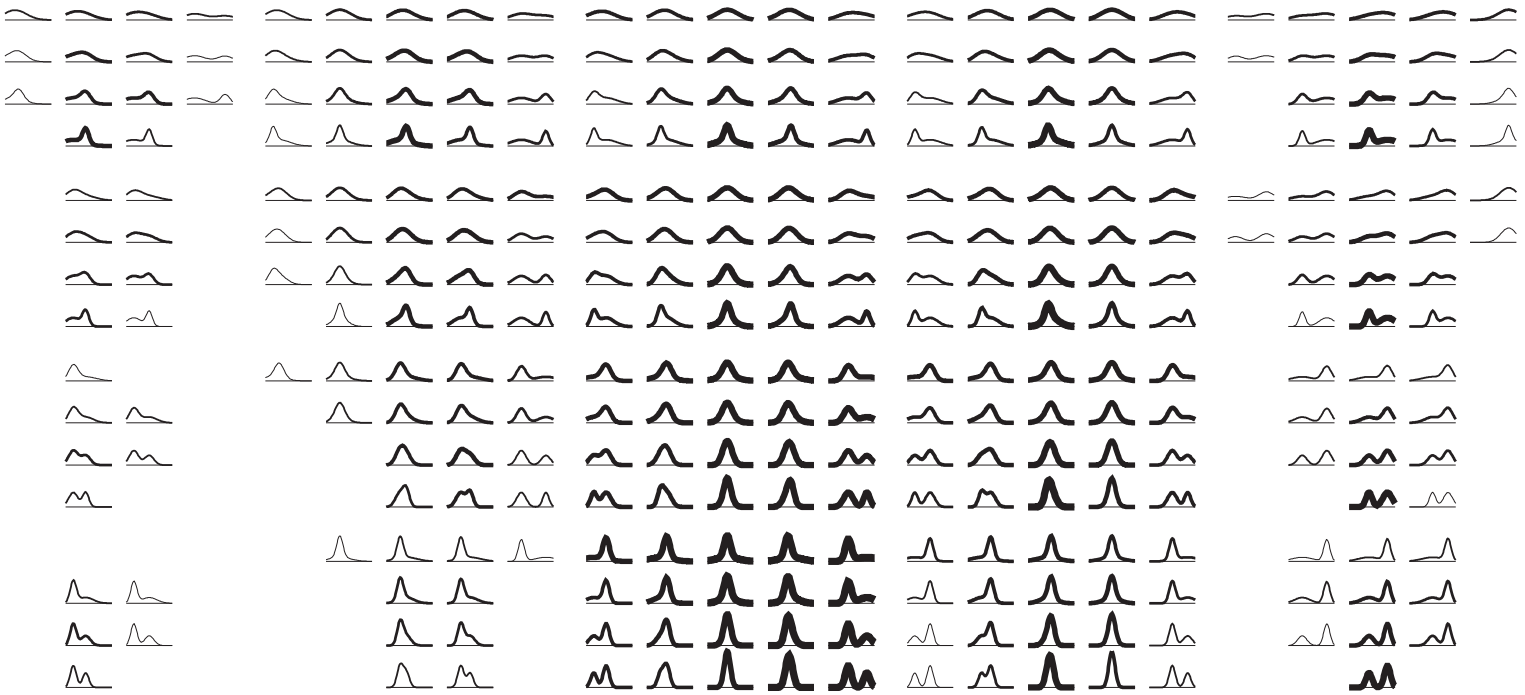
(Discretized) hypothesis space for two-Gaussian mix ($\pi_1 = 0.6$ and $\pi_2 = 0.4$)



- Means vary horizontally
- Standard deviations vary vertically

Maximum likelihood (cont.)

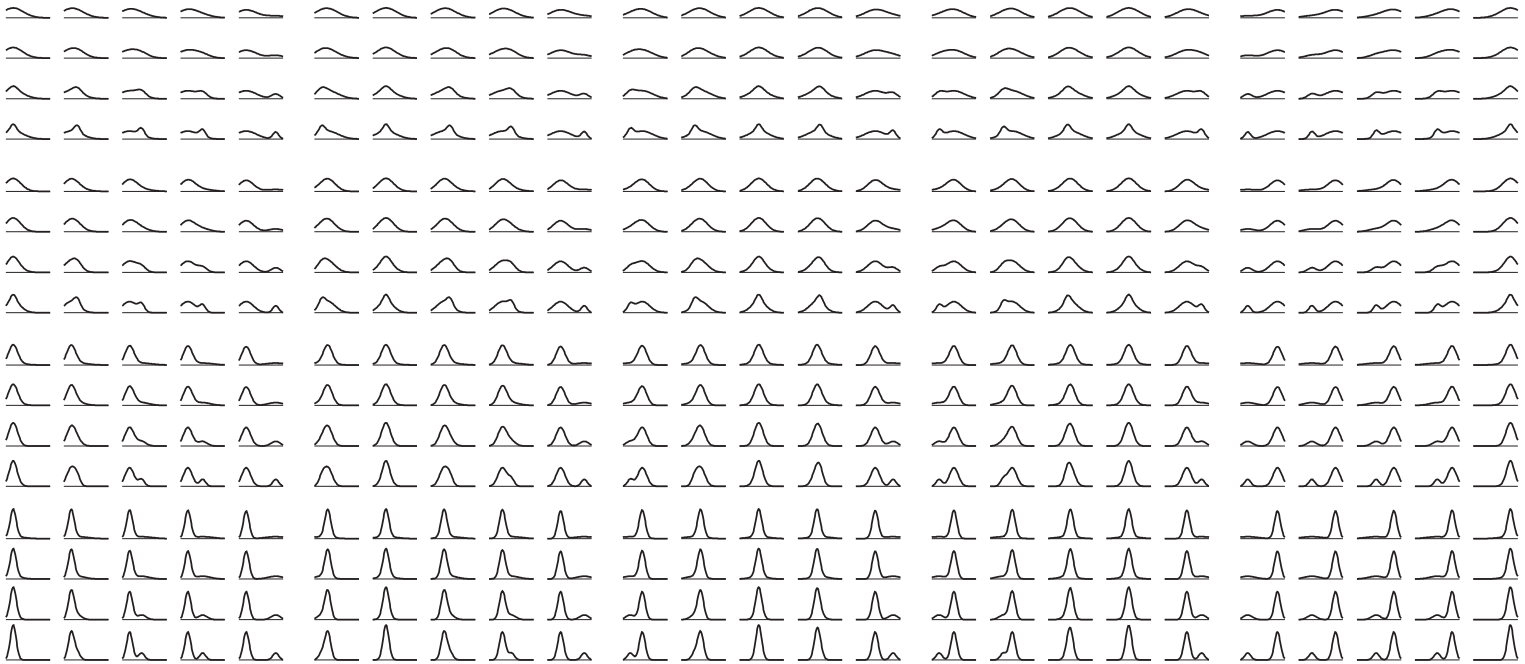
Likelihood function, given the data, as line thickness



Sub-hypothesis smaller than e^{-8} the maximum likelihood are blanked

Maximum likelihood (cont.)

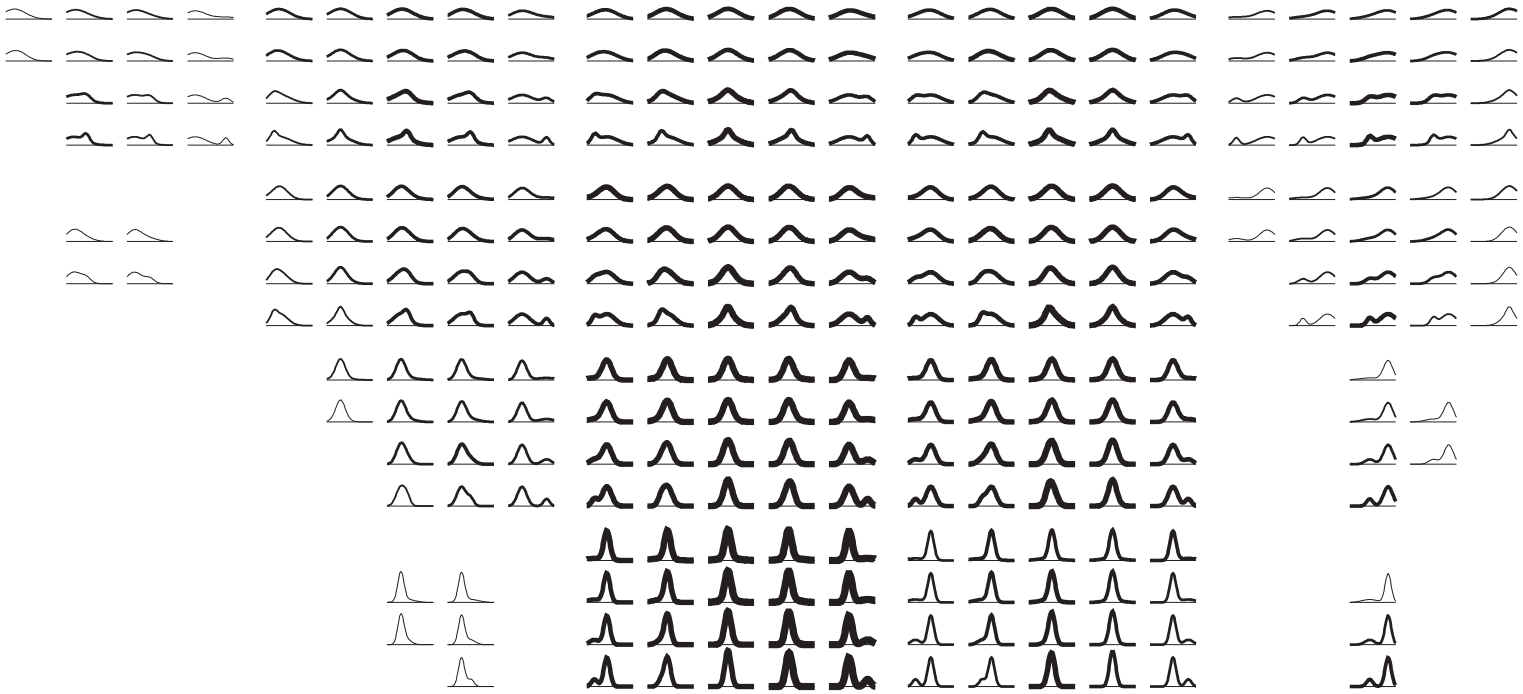
(Discretized) hypothesis space for two-Gaussian mix ($\pi_1 = 0.8$ and $\pi_2 = 0.2$)



- Means vary horizontally
- Standard deviations vary vertically

Maximum likelihood (cont.)

Likelihood function, given the data, as line thickness



Sub-hypothesis smaller than e^{-8} the maximum likelihood are blanked

Mixtures of Gaussians (cont.)

$$\ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \ln \left(\sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right) \tag{97}$$

It is bad, very bad, because of the summation over k inside the logarithm

- No longer closed-form solution
- Iterative numerical optimisation

- Expectation-maximisation

Mixtures of Gaussians (cont.)

If we define a joint distribution over observed and latent variables, the distribution of the observed variables alone is obtained by marginalisation

Complex marginal distributions over observed variables can be expressed in terms of joint distributions over the space of observed and latent variables

- The introduction of latent variables allows complicated distributions to be formed from simpler components

Remark

Mixture distributions can be interpreted in terms of discrete latent variables

- It is the case with the Gaussian mixture

Mixture models can be used in the problem of finding clusters in a point set

Mixtures of Gaussians (cont.)

We motivated the Gaussian mixture model as a linear superposition of Gaussians to provide a richer class of density models than the Gaussian

We formulate now Gaussian mixtures in terms of **discrete latent variables**

- This provides a deeper insight into this important distribution
- It also serves to motivate the expectation-maximisation algorithms

Mixtures of Gaussians (cont.)

A Gaussian mix distribution is a linear superposition of Gaussian components

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{98}$$

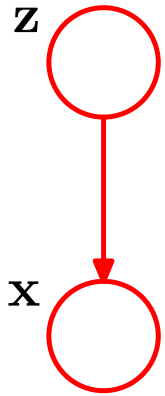
Let us introduce a K-dim binary random variable $\mathbf{z}$ with 1-of-K representation in which an element z_k is equal to 1 and all other elements are equal to 0

- Values of z_k satisfy $z_k \in \{0, 1\}$ and $\sum_k z_k = 1$
- There are K possible states for the vector $\mathbf{z}$

Mixtures of Gaussians (cont.)

We define a joint distribution $p(\mathbf{x}, \mathbf{z}) = p(\mathbf{z})p(\mathbf{x}|\mathbf{z})$, that is in terms of

- the marginal distribution $p(\mathbf{z})$
- the conditional distribution $p(\mathbf{x}|\mathbf{z})$



The marginal distribution over $\mathbf{z}$ is specified in terms of mixing coefficients π_k

$$p(z_k = 1) = \pi_k \tag{99}$$

where parameters $\{\pi_k\}$ must satisfy

$$0 \leq \pi_k \leq 1 \quad \text{and} \quad \sum_{k=1}^{K} \pi_k = 1 \tag{100}$$

Mixtures of Gaussians (cont.)

Because $\mathbf{z}$ uses a 1-of-K representation, we write this distribution in the form

$$p(\mathbf{z}) = \prod_{k=1}^{K} \pi_k^{z_k} \tag{101}$$

The conditional distribution of $\mathbf{x}$ given a particular value for $\mathbf{z}$ is Gaussian

$$p(\mathbf{x}|z_k = 1) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \quad \text{or } p(\mathbf{x}|\mathbf{z}) = \prod_{k=1}^{K} \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)^{z_k} \tag{102}$$

The joint distribution is given by $p(\mathbf{z})p(\mathbf{x}|\mathbf{z})$, and the marginal distribution of $\mathbf{x}$ is then obtained by summing the joint distribution over all possible states of $\mathbf{z}$

$$p(\mathbf{x}) = \sum_{\mathbf{z}} p(\mathbf{z})p(\mathbf{x}|\mathbf{z}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{103}$$

Mixtures of Gaussians (cont.)

If we have several observations $\mathbf{x}_1, \ldots, \mathbf{x}_N$, then, since we have represented the marginal distribution in the form $p(\mathbf{x}) = \sum_{\mathbf{z}} p(\mathbf{x}, \mathbf{z})$, it follows that

- for every observed point $\mathbf{x}_n$ there is a corresponding latent variable $\mathbf{z}_n$

Definition

An equivalent formulation of Gaussian mixtures with an explicit latent variable

- We are now able to work with the joint distribution $p(\mathbf{x}, \mathbf{z})$ instead of the marginal distribution $p(\mathbf{x})$, which leads to significant simplifications

Mixtures of Gaussians (cont.)

Another important quantity is the conditional probability $\mathbf{z}$ given $\mathbf{x}$

Let $\gamma(z_k)$ denote $p(z_k = 1|\mathbf{x})$, using Bayes' theorem

$$\begin{aligned} \gamma(z_k) \equiv p(z_k = 1|\mathbf{x}) &= \frac{p(z_k = 1)p(\mathbf{x}|z_k = 1)}{\sum_{j=1}^{K} p(z_j = 1)p(\mathbf{x}|z_j = 1)} \\ &= \frac{\pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)} \end{aligned} \tag{104}$$

where π_k is viewed as the prior probability of $z_k = 1$ and the quantity $\gamma(z_k)$ as corresponding posterior probability, once we have observed $\mathbf{x}$

- Again, $\gamma(z_k)$ can be viewed as the **responsability** that component k takes for 'explaining' the observation $\mathbf{x}$

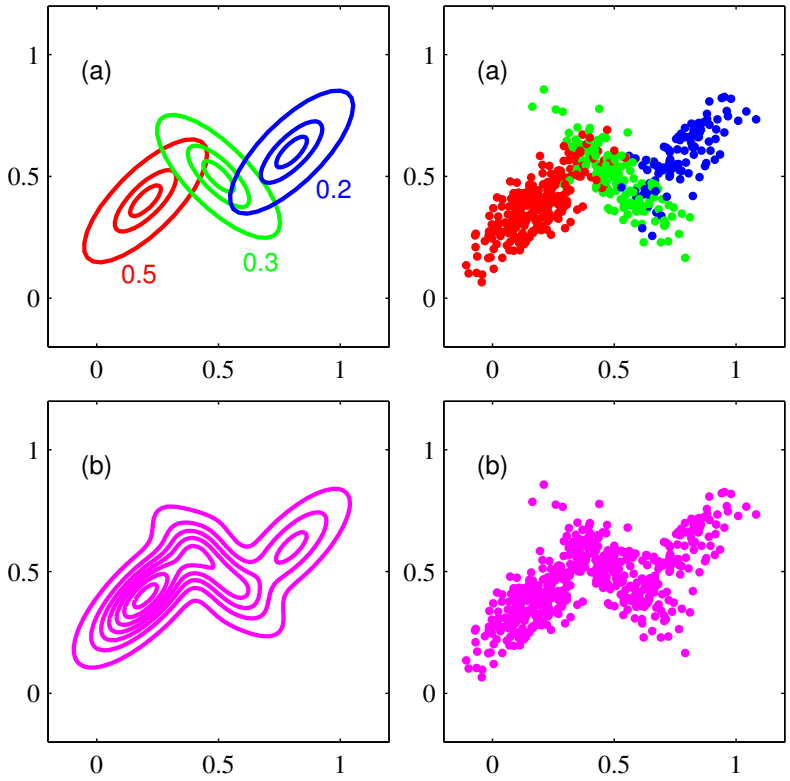
Mixtures of Gaussians (cont.)

We can depict samples from the joint distribution $p(\mathbf{x}, \mathbf{z})$ by plotting points at the corresponding values of $\mathbf{x}$ and colouring them according to the value of $\mathbf{z}$

- The Gaussian component was responsible for generating them

Similarly, samples from the marginal distribution $p(\mathbf{x})$ are obtained by taking the samples from the joint distribution and ignoring the values of $\mathbf{z}$

Mixtures of Gaussians (cont.)



An example of 500 points as drawn from some mixture of 3 Gaussians

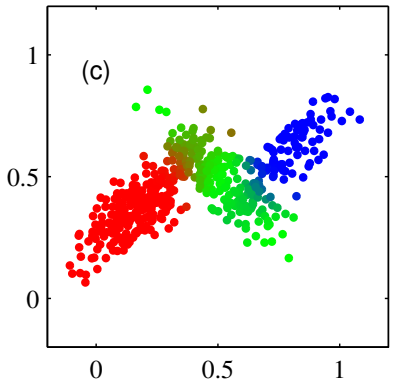
a Samples from the joint distribution $p(\mathbf{z})p(\mathbf{x}|\mathbf{z})$ in which the 3 states of $\mathbf{z}$, corresponding to 3 components of the mixture, are in red, green and blue

b The corresponding samples from the marginal distribution $p(\mathbf{x})$, which is obtained by simply ignoring the values of $\mathbf{z}$ and just plotting the $\mathbf{x}$ values

The data set in (a) is said to be complete, whereas that in (b) is incomplete

Mixtures of Gaussians (cont.)

We can illustrate 'responsibilities' by evaluating, for every point, the posterior probability for each component in the GMM distribution from which data arose



A point for which $\gamma(z_{n1(2,3)}) = 1$ is red (blue, green)

A point for which $\gamma(z_{n2}) = \gamma(z_{n3}) = 0.5$
uses equal portions of blue and green

Responsibilities $\gamma(z_{nk})$ associated with point $\mathbf{x}_n$ can be represented by dyeing the corresponding point with proportions of R, B, and G ink given by $\gamma(z_{nk})$

Maximum likelihood
Mixtures of Gaussians

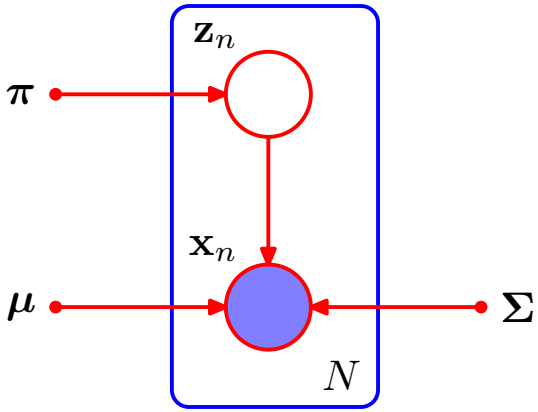
Maximum likelihood

Suppose we have a data set of observations $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$

We wish to model this data using a mixture of Gaussians

- We can represent this data set as an $N \times D$ matrix $\mathbf{X}$, in which the n-th row is given by $\mathbf{x}_n^T$
- The corresponding latent variables can be represented by a $N \times K$ matrix $\mathbf{Z}$ with rows $\mathbf{z}_n^T$

If we assume that the points are drawn independently from the distribution, the Gaussian mixture model can be expressed for the iid set using a PGM



The GMM distribution for N independent and identically distributed data points $\{\mathbf{x}_n\}$

- with corresponding latent points $\{\mathbf{z}_n\}$

Maximum likelihood

From $p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$, the log likelihood function is given by

$$\ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \ln \left(\sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right) \tag{105}$$

The Gaussian mixture has components whose covariance matrices $\boldsymbol{\Sigma}_k = \sigma_k^2 \mathbf{I}$ [9]

Suppose that one of the components, the j-th, has its mean $\boldsymbol{\mu}_j$ exactly equal to one of the data points, the n-th, so that $\boldsymbol{\mu}_j = \mathbf{x}_n$

- The contribution of point $\mathbf{x}_n$ to the likelihood is

$$\mathcal{N}(\mathbf{x}_n|\mathbf{x}_n, \sigma_j^2 \mathbf{I}) = \frac{1}{(2\pi)^{1/2}} \frac{1}{\sigma_j} \tag{106}$$

- This term goes to infinity in the limit $\sigma_j \to 0$
- The whole log likelihood goes to infinity with it

[9] It is a computational simplification due to the presence of singularities when maximising the likelihood, though the conclusions will hold for general covariance matrices $\boldsymbol{\Sigma}_k$

Maximum likelihood

This is not a great situation, because the maximisation of the log likelihood is always not well posed when one of the Gaussians collapses onto a data point

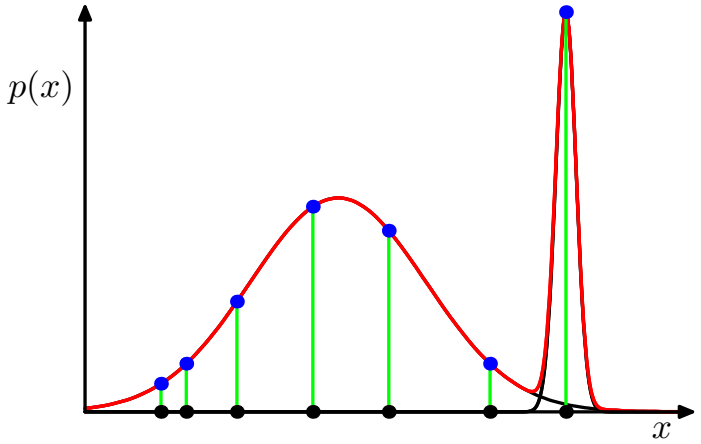
- It does not happen with a single Gaussian

Remark

If a single Gaussian collapses onto a point, it contributes multiplicative factors to the likelihood function arising from other points and these factors go to zero exponentially fast, giving an overall likelihood that goes to 0 rather than ∞

If we have (at least) two components in the mixture, one of the components can have a finite variance and therefore assign finite probability to all of the points while the other component can shrink onto one specific point and thereby contribute an ever increasing additive value to the log likelihood

Maximum likelihood (cont.)



These singularities provide an example of over-fitting, as it occurs with MLE

In applying MLE to the GMM we must take steps to avoid finding pathological solutions and seek good local maxima

- We can hope to avoid the singularities by using suitable heuristics, as detecting when a Gaussian component is collapsing and resetting its mean to a randomly chosen value while also resetting its covariance to some large value, and then continuing with the optimisation

Maximum likelihood (cont.)

A further issue in MLE solutions arises from the fact that for any given ML solution, a K-component mixture have a total of $K!$ equivalent solutions

- The $K!$ ways of assigning K sets of parameters to K components

For any given (nondegenerate) point in the space of parameters there are a further $K! - 1$ additional points all of which give rise to the same distribution

- The problem is known as **identifiability**

Maximum likelihood (cont.)

Maximising the log likelihood function for a Gaussian mixture model

$$\ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \ln \left(\sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right)$$

is a more complex problem than for the case of a single Gaussian

The difficulty arises from the summation over k inside the log

- the log function no longer acts directly on the Gaussian

Set derivatives of log likelihood to zero: Does not give a closed form solution

- One approach is to apply gradient-based optimisation techniques, feasible
- An alternative approach known as the EM algorithm, broad applicability

EM for mixtures of Gaussians
Mixtures of Gaussians

EM for mixtures of Gaussians

The **expectation-maximisation**, or **EM algorithm** is an elegant and powerful method for finding maximum likelihood solutions for latent variables models

We motivate the EM algorithm in the context of the Gaussian mixture model

EM for mixtures of Gaussians (cont.)

First, conditions that must be satisfied at a maximum of likelihood functions

Setting the derivatives of $\ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \ln \left(\sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right)$ with respect to the means $\boldsymbol{\mu}_k$ of the Gaussian components to zero

$$0 = -\sum_{n=1}^{N} \underbrace{\frac{\pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_i \pi_j \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}}_{\gamma(z_{nk})} \boldsymbol{\Sigma}_k (\mathbf{x}_n - \boldsymbol{\mu}_k) \tag{107}$$

where for the Gaussian distribution we made use of the form

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp \left(-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right)$$

Posterior probabilities, or responsibilities $\gamma(z_{nk})$ appear on the RHS

EM for mixtures of Gaussians (cont.)

$$0 = -\sum_{n=1}^{N} \underbrace{\frac{\pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_i \pi_j \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}}_{\gamma(z_{nk})} \boldsymbol{\Sigma}_k(\mathbf{x}_n - \boldsymbol{\mu}_k)$$

Multiplying by $\boldsymbol{\Sigma}_k^{-1}$, assuming it is non-singular, and reordering

$$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \mathbf{x_n} \tag{108}$$

where together with $\gamma(z_{nk}) = \frac{\pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_i \pi_j \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}$, we defined

$$N_k = \sum_{n=1}^{N} \gamma_{nk} \tag{109}$$

N_k is understood as effective number of points assigned to cluster k

EM for mixtures of Gaussians (cont.)

$$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \mathbf{x_n}$$

The mean $\boldsymbol{\mu}_k$ for the k-th Gaussian component is obtained by taking a weighted mean of all of the points in the data set

- The weighting factor for data point $\mathbf{x}_n$ is the posterior probability $\gamma(z_{nk})$ that component k was responsible for generating $\mathbf{x}_n$

EM for mixtures of Gaussians (cont.)

Setting derivatives of $\ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \ln \left(\sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right)$ wrt the covariance matrices $\boldsymbol{\Sigma}_k$ of the Gaussian components to zero

$$\boldsymbol{\Sigma}_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk})(\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^T \tag{110}$$

which has the same form as the result for a single Gaussian fitted to data

- Again, each point is weighted by the corresponding posterior probability

Denominator is the effective number of points associated with the component

EM for mixtures of Gaussians (cont.)

We maximise $\ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \ln \left(\sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right)$ wrt the mixing coefficients π_k, taking into account of the constraint $\sum_{k=1}^{K} \pi_k = 1$

Using Lagrange multipliers and maximising

$$\ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) + \lambda \left(\sum_{k=1}^{K} \pi_k - 1 \right) \tag{111}$$

gives us an expression with responsabilities

$$0 = \sum_{n=1}^{K} \frac{\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_j \pi_k \mathcal{N}(\pi_k|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_k)} + \lambda \tag{112}$$

EM for mixtures of Gaussians (cont.)

Multiplying both sides by π_k and summing over k using constraint $\sum_{k=1}^{K} \pi_k = 1$, we find that $\lambda = N$ which can be used to eliminate λ

$$\pi_k = \frac{N_k}{N} \tag{113}$$

The mixing coefficient for the k-th component is the average responsibility taken by that component for explaining data

EM for mixtures of Gaussians (cont.)

$$\boldsymbol{\mu}_k = \frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})\mathbf{x_n}$$

$$\boldsymbol{\Sigma}_k = \frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})(\mathbf{x}_n-\boldsymbol{\mu}_k)(\mathbf{x}_n-\boldsymbol{\mu}_k)^T$$

$$\pi_k = \frac{N_k}{N}$$

Remark

These do not make for a closed-form solution for the mixture parameters

- It suggest an iterative scheme for maximising the likelihood

EM for mixtures of Gaussians (cont.)

We choose initial values for the means, covariances, and mixing coefficients

Then we alternate between two updates, until convergence

Pseudocode 2.1

- **E-step** or **Expectation-step**, the values of the parameters are used to evaluate posterior probabilities

$$\gamma(z_k) \equiv p(z_k = 1|\mathbf{x}) = \frac{p(z_k = 1)p(\mathbf{x}|z_k = 1)}{\sum_{j=1}^{K} p(z_j = 1)p(\mathbf{x}|z_j = 1)} = \frac{\pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}$$

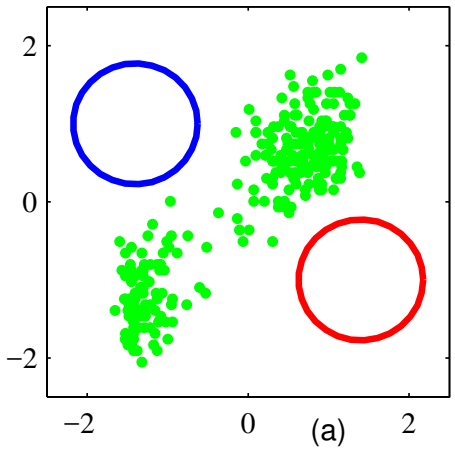
- **M-step** or **Maximisation-step**, probabilities are used to re-estimate means covariances and mixing coefficients

$$\begin{aligned}
\boldsymbol{\mu}_k &= \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \mathbf{x_n} \\
\boldsymbol{\Sigma}_k &= \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk})(\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^T \\
\pi_k &= \frac{N_k}{N}
\end{aligned}$$

EM for mixtures of Gaussians (cont.)

At each update, the parameters resulting from an E-step followed by an M-step are guaranteed to increase the log likelihood function

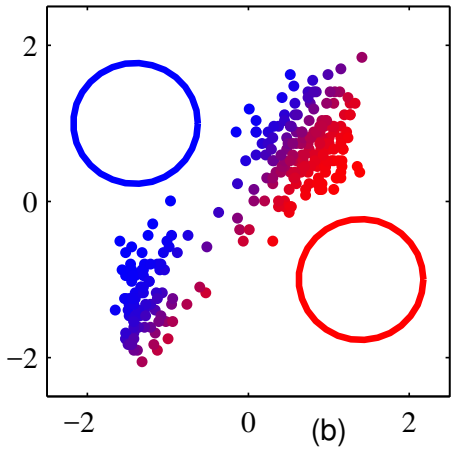
A mixture of two Gaussians, with centres initialised as in the $k-$means example, and precision initialised to be proportional to the unit matrix



Data points as green dots

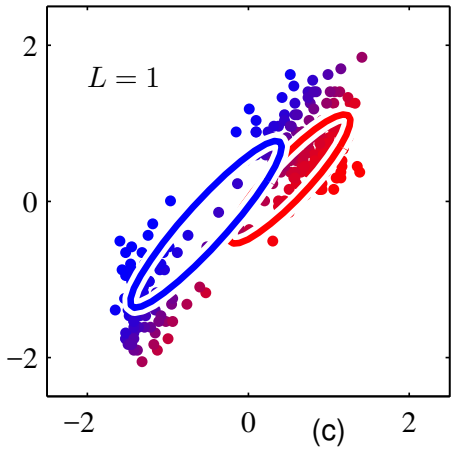
Contours at one standard deviation

EM for mixtures of Gaussians (cont.)



Result of the **first E-step**

Points dyed with a proportion of ink equal to the posterior probability of having been generated by the component Gaussians

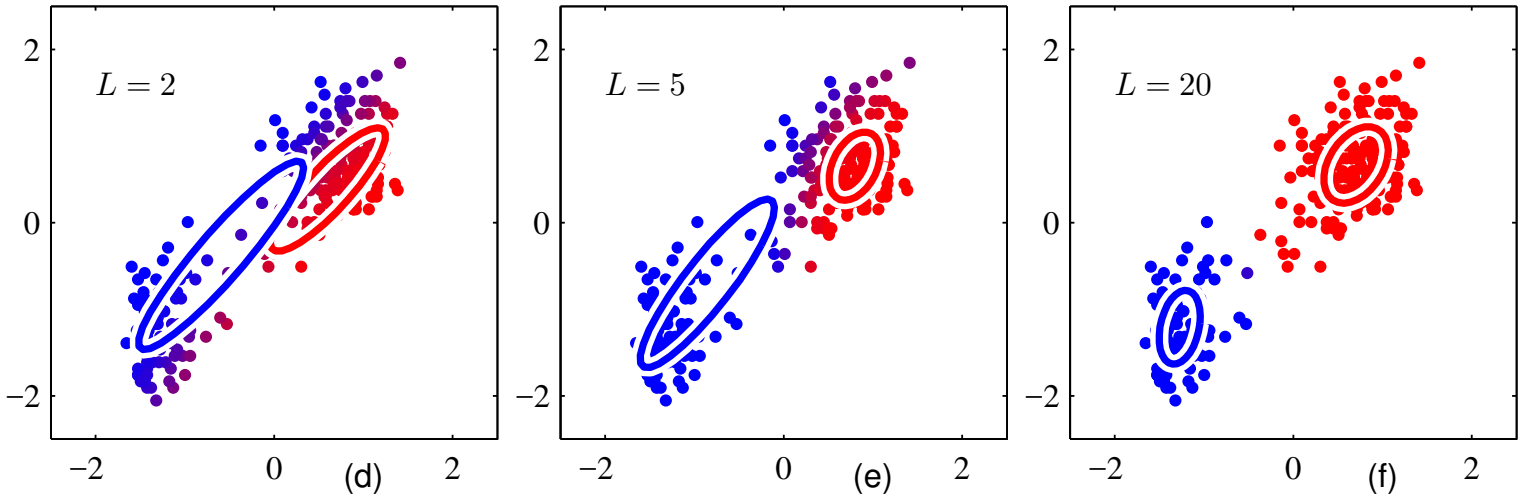


Result of the **first M-step**

The mean of the blue Gaussian moves to the mean of the dataset weighted by the probabilities of the blue points

The covariance of the blue Gaussian equals that of the blue points

EM for mixtures of Gaussians (cont.)



The EM algorithm takes many more iterations to reach convergence compared with *k*-means, and that each cycle requires significantly more computation

Remark

It is common to run ***k*-means** to find a suitable initialisation for a Gaussian mixture model that is subsequently adapted using EM

- Covariances can be initialised to sample covariances of the clusters
- Mixing coefficients can be set to fractions of points in retrieved clusters

k-means clustering
Mixtures of Gaussians

k-means clustering

We consider the problem of identifying groups, or clusters, of data points

We have data $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ consisting of N observations

- $\mathbf{x}$ is a random D-dimensional Euclidean variable

Goal: Partition the data into some number K of clusters

- We suppose for the moment that the value of K is given

Definition

A **cluster** can be seen as a group of points whose inter-point distances are small, compared with the distances to points outside of the cluster

To formalise this notion, introduce a set of D-dimensional vectors $\{\boldsymbol{\mu}_k\}_{k=1}^{K}$

- $\boldsymbol{\mu}_k$ is a **prototype** associated with the k-th cluster

We can think of the various $\boldsymbol{\mu}_k$ as representing the centres of the clusters

k-means clustering (cont.)

Given data $\{\mathbf{x}_n\}_{n=1}^N$ and having defined cluster prototypes $\{\boldsymbol{\mu}_k\}_{k=1}^K$

- Goal: Find an assignment of data points to clusters
- Goal: Find an optimal set of prototype vectors $\{\boldsymbol{\mu}_k\}_{k=1}^K$

The set of vectors $\{\boldsymbol{\mu}_k\}_{k=1}^K$ is such that the sum of the squares of the distances of each point to its closest vector $\boldsymbol{\mu}_k$, is a minimum

k-means clustering (cont.)

For each point $\mathbf{x}_n$, we introduce a set of **binary indicator variables** $r_{nk} \in \{0, 1\}$ with $k = 1, \ldots, K$ telling which of the K clusters the point $\mathbf{x}_n$ is assigned to

- If $\mathbf{x}_n$ is assigned to cluster k, then $r_{nk} = 1$ and $r_{nj} = 0$ for $j \neq k$
- This scheme is known as **1-of-K coding**

Definition

We define an objective function, sometimes called a **distortion measure**

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \|\mathbf{x}_n - \boldsymbol{\mu}_k\|^2 \tag{114}$$

A sum of squares of distances between each point and assigned prototype

- Find $\{r_{nk}\}_{k=1}^{K}$ and $\{\boldsymbol{\mu}_k\}_{k=1}^{K}$ such that J is minimised

k-means clustering (cont.)

We can do this through an iterative procedure in which each iteration involves two successive steps corresponding to successive optimisations wrt r_{nk} and $\boldsymbol{\mu}_k$

Pseudocode

The procedure starts by setting some initial values for $\boldsymbol{\mu}_k$

- **1-st phase**: Minimise J with respect to the r_{nk}, keeping $\boldsymbol{\mu}_k$ fixed
- **2-nd phase**: Minimise J with respect to the $\boldsymbol{\mu}_k$, keeping r_{nk} fixed

The two-stage optimisation is repeated until convergence

These two stages of updating r_{nk} and updating $\boldsymbol{\mu}_k$ correspond respectively to the E (expectation) and M (maximisation) steps of the EM algorithm

- We use E-step and M-step also in the context of the k-means algorithm

k-means clustering (cont.)

Consider first the determination of the r_{nk}, with the $\boldsymbol{\mu}_k$ fixed

The objective $J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \|\mathbf{x}_n - \boldsymbol{\mu}_k\|^2$ is a linear function of r_{nk}

- This optimisation can be performed to give a closed form solution

The terms involving different n are independent and we can optimise for each n separately by setting r_{nk} to be 1 for whichever k gives minimum $\|\mathbf{x}_n - \boldsymbol{\mu}_k\|^2$

- We assign the n-th point $\mathbf{x}_n$ to the closest cluster centre $\boldsymbol{\mu}_k$

$$r_{nk} = \begin{cases} 1, & \text{if } k = \text{argmin}_j \|\mathbf{x}_n - \boldsymbol{\mu}_k\|^2 \\ 0, & \text{otherwise} \end{cases} \tag{115}$$

k-means clustering (cont.)

Now consider the determination of the $\boldsymbol{\mu}_\mathbf{k}$, with the r_{nk} fixed

The objective $J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \|\mathbf{x}_n - \boldsymbol{\mu}_k\|^2$ is a quadratic function of $\boldsymbol{\mu}_k$, and it can be minimised by setting its derivative with respect to $\boldsymbol{\mu}_k$ to zero

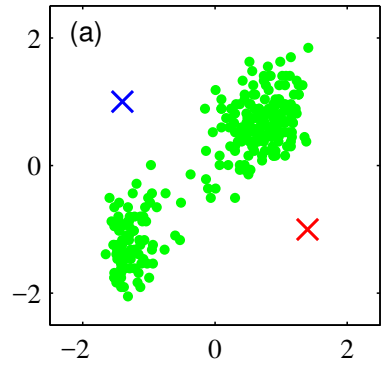
$$2 \sum_{n=1}^{N} r_{nk} (\mathbf{x}_n - \boldsymbol{\mu}_k) = 0 \tag{116}$$

Solving for $\boldsymbol{\mu}_k$

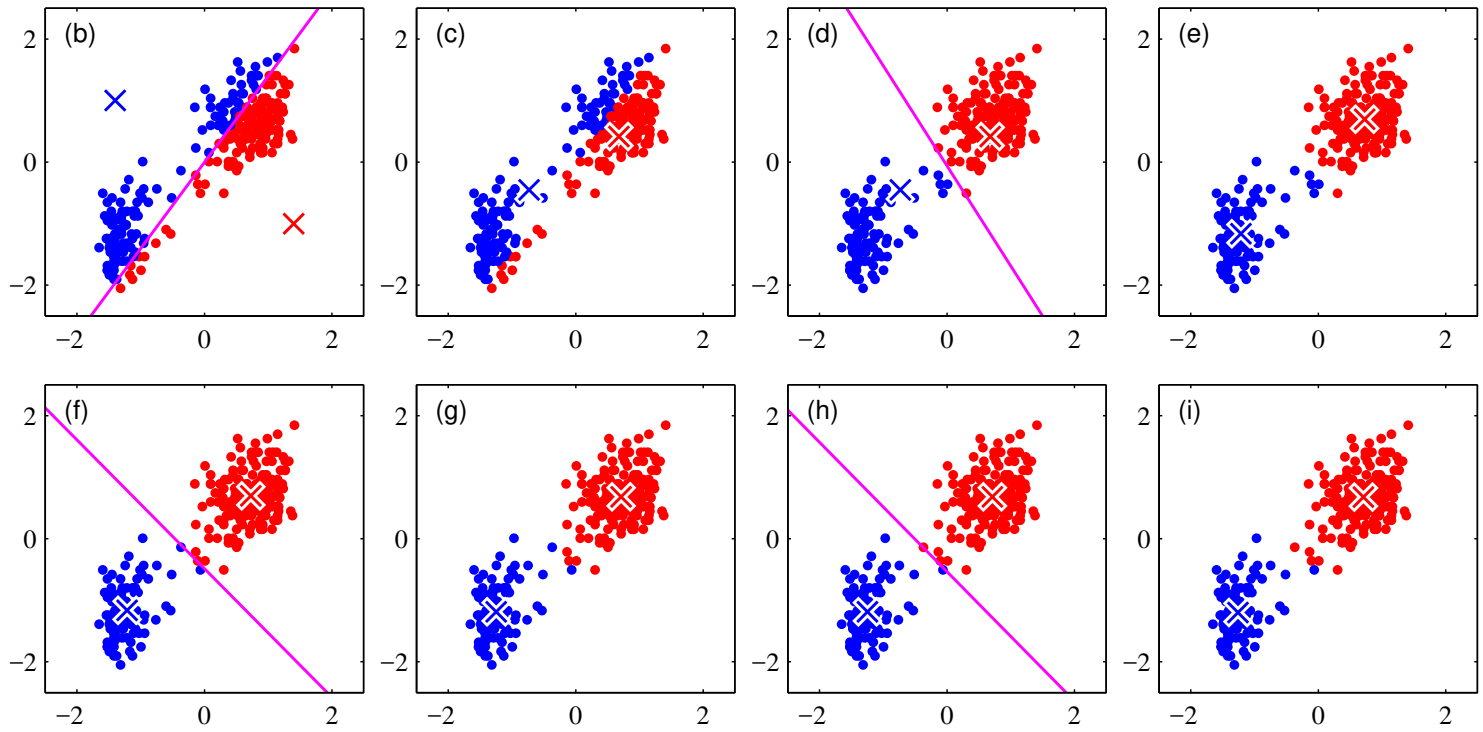
$$\boldsymbol{\mu}_k = \frac{\sum_{n=1}^{N} r_{nk} \mathbf{x}_n}{\sum_{n=1}^{N} r_{nk}} \tag{117}$$

The denominator equals the number of points assigned to cluster k, and overall we set $\boldsymbol{\mu}_k$ equal to the mean of all points $\mathbf{x}_n$ assigned to cluster k

Green points are data $\{\mathbf{x}_n\}$ in a 2D Euclidean space and the initial choices for centres $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ are the red and blue crosses

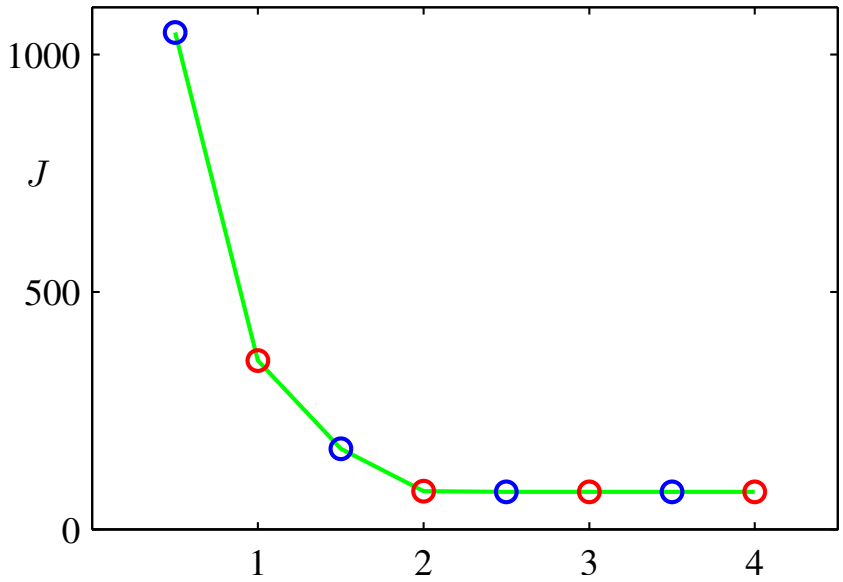


- In the initial E step, each point is assigned either to the red cluster or to the blue cluster, according to which cluster centre is nearer
- In the subsequent M step, each cluster centre is re-computed to be the mean of the points assigned to the corresponding cluster



k-means clustering (cont.)

Cost J after each E step (blue points) and M step (red points) of the *k*-means



The algorithm has converged after the third M step, and the final EM cycle produces no changes in either the assignments or the prototype vectors