Aalto University

Probabilistic machine learning | Intro (C)

Introduction to machine learning

Francesco Corona

Chemical and Metallurgical Engineering
School of Chemical Engineering

The Gaussian distribution

Probability theory | Intro (B)

The Gaussian distribution

A **Gaussian**, or **normal**, **distribution** is pervasive density model for continuous variables

For a single random variable X

$$p(x|\mu,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\Big(-\frac{1}{2}(x-\mu)\frac{1}{\sigma^2}(x-\mu)\Big)$$

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

For a D-dimensional variable X

$$p(x|\mu,\Sigma) = \frac{1}{(2\pi)^{D/2}}\frac{1}{|\Sigma|^{1/2}} \exp\Big(-\frac{1}{2}(x-\mu)^T\frac{1}{\Sigma}(x-\mu)\Big)$$

- μ is the D-dimensional mean vector
- Σ is the $D \times D$ covariance matrix
- $|\Sigma|$ is the determinant of Σ

The Gaussian distribution (cont.)

$$p(x|\mu,\Sigma) = \frac{1}{(2\pi)^{D/2}}\frac{1}{|\Sigma|^{1/2}}\exp\Big(-\frac{1}{2}(x-\mu)^T\frac{1}{\Sigma}(x-\mu)\Big)$$

Gaussians depends on x through the quadric form $\Delta^2 = (x-\mu)^T\Sigma^{-1}(x-\mu)$

- Quantity Δ is the **Mahalanobis distance** from μ to x
- It reduces to the Euclidean distance when $\Sigma = I_D$

Thence, the Gaussian density is constant on surfaces for which Δ^2 is constant

To be well-defined, all of the eigenvalues of Σ need be real and strictly positive

- Otherwise, the Gaussian cannot be properly normalised

Gaussian densities for which one or more eigenvalues of Σ are zero are singular

- They are confined to a subspace of lower dimensionality

The Gaussian distribution | Expectations

The expectation of $f(x) = x$ under the Gaussian distribution

$$\begin{aligned}\mathbb{E}[x] &= \frac{1}{(2\pi)^{D/2}}\frac{1}{|\Sigma|^{1/2}}\int \exp\Big(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\Big)\underbrace{x}\,dx \\ &= \mu\end{aligned}$$

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

---

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

$$\begin{aligned}\mathbb{E}[xx^T] &= \frac{1}{(2\pi)^{D/2}}\frac{1}{|\Sigma|^{1/2}}\int \exp\Big(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\Big)\underbrace{xx^T}\,dx \\ &= \mu\mu^T + \Sigma\end{aligned}$$

We refer to Σ as the **covariance matrix** of the Gaussian distribution

- Its inverse $\Lambda = \Sigma^{-1}$ is denoted as the **precision matrix**

Marginals, conditionals, and posteriors

The Gaussian distribution

The Gaussian distribution | Conditionals

If two vectors X_a and X_b are jointly (as X) distributed as a Gaussian, then the conditional distribution of one vector conditioned on the other one is a Gaussian distribution

$$X = \begin{pmatrix} X_a \\ X_b \end{pmatrix}$$

$$\mu = \begin{pmatrix} \mu_a \\ \mu_b \end{pmatrix}$$

$$p(x) = \underbrace{\frac{1}{(2\pi)^{D/2}} \frac{1}{|\Sigma|^{1/2}} \exp\Big(-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)\Big)}_{p(x_a, x_b)}$$

$$\Sigma = \begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{pmatrix}$$

$$\Lambda = \begin{pmatrix} \Lambda_{aa} & \Lambda_{ab} \\ \Lambda_{ba} & \Lambda_{bb} \end{pmatrix}$$

The conditional distribution $p(x_a|x_b)$

$$p(x_a|x_b) = \frac{1}{(2\pi)^{D_a/2}} \frac{1}{|\Sigma_{a|b}|^{1/2}} \exp\Big(-\frac{1}{2}(x_a-\mu_{a|b})^T \Sigma_{a|b}^{-1} (x_a-\mu_{a|b})\Big)$$

- $\mu_{a|b} = \mu_a + \Sigma_{ab}\Sigma_{bb}^{-1}(x_b - \mu_b)$
- $\Sigma_{a|b} = \Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba}$

The covariance matrix of the conditional distribution of $X_a|X_b$ is independent of x_b

The mean vector of the conditional distribution of $X_a|X_b$ is a linear function of x_b

The Gaussian distribution | Marginals

If two vectors X_a and X_b are jointly (as X) distributed as a Gaussian, then each of the marginal distributions of the two components vectors area also Gaussian distributions

$$X = \begin{pmatrix} X_a \\ X_b \end{pmatrix} \qquad \mu = \begin{pmatrix} \mu_a \\ \mu_b \end{pmatrix}$$

$$\underbrace{p(x) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\Sigma|^{1/2}} \exp\Big(- \frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu) \Big)}_{p(x_a, x_b)} \qquad \Sigma = \begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{pmatrix}$$

$$\Lambda = \begin{pmatrix} \Lambda_{aa} & \Lambda_{ab} \\ \Lambda_{ba} & \Lambda_{bb} \end{pmatrix}$$

The marginal distribution $p(x_a)$

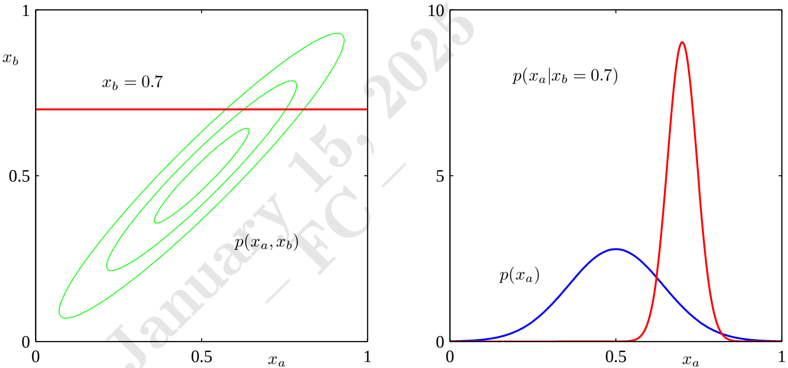
$$\begin{aligned} p(x_a) &= \int p(x_a, x_b)\, dx_b \\ &= \frac{1}{(2\pi)^{D_a/2}} \frac{1}{|\Sigma_{aa}|^{1/2}} \exp\Big(- \frac{1}{2}(x_a-\mu_a)^T \Sigma_{aa}^{-1} (x_a-\mu_a) \Big) \end{aligned}$$

The marginal distribution $p(x_b)$

$$\begin{aligned} p(x_b) &= \int p(x_a, x_b)\, dx_a \\ &= \frac{1}{(2\pi)^{D_b/2}} \frac{1}{|\Sigma_{bb}|^{1/2}} \exp\Big(- \frac{1}{2}(x_b-\mu_b)^T \Sigma_{bb}^{-1} (x_b-\mu_b) \Big) \end{aligned}$$

The Gaussian distribution | Conditionals and marginals

$$p(x) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\Sigma|^{1/2}} \exp\Big(- \frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu) \Big)$$



$$p(x_a|x_b) = \frac{1}{(2\pi)^{D_a/2}} \frac{1}{|\Sigma_{a|b}|^{1/2}} \exp\Big(- \frac{1}{2}(x_a-\mu_{a|b})^T \Sigma_{a|b}^{-1} (x_a-\mu_{a|b}) \Big)$$

$$p(x_a) = \frac{1}{(2\pi)^{D_a/2}} \frac{1}{|\Sigma_{aa}|^{1/2}} \exp\Big(- \frac{1}{2}(x_a-\mu_a)^T \Sigma_{aa}^{-1} (x_a-\mu_a) \Big)$$

The Gaussian distribution | Posteriors

Suppose we are only given some $p(x_a)$ and $p(x_b|x_a)$ and we are interested in $p(x_a|x_b)$

For the conditional $p(x_b|x_a)$, a likelihood function, we have

$$p(x_b|x_a) = \frac{1}{(2\pi)^{D_b/2}} \frac{1}{|\Sigma_{b|a}|^{1/2}} \exp\Big(- (x_b - \mu_{b|a})^T \frac{\Sigma_{b|a}^{-1}}{2} (x_b - \mu_{b|a}) \Big), \qquad \begin{matrix} \mu_{b|a} = A x_a + b \\ \Sigma_{b|a} \end{matrix}$$

The mean $\mu_{b|a}$ is linear in x_a

For the marginal $p(x_a)$, a prior, we have

$$p(x_a) = \frac{1}{(2\pi)^{D_a/2}} \frac{1}{|\Sigma_{aa}|^{1/2}} \exp\Big(- \frac{1}{2} (x_a - \mu_a)^T \Sigma_{aa}^{-1} (x_a - \mu_a) \Big), \qquad \begin{matrix} \mu_a \\ \Sigma_{aa} \end{matrix}$$

From Bayes, we get a posterior $p(x_a|x_b)$

$$\begin{aligned} p(x_a|x_b) &= \frac{p(x_b|x_a) p(x_a)}{p(x_b)} \\ &= \frac{p(x_a, x_b)}{\int p(x_a, x_b) dx_a} \end{aligned}$$

The Gaussian distribution | Posteriors (cont.)

For the joint distribution $p(x_a, x_b) = p(x_b|x_a)p(x_a)$, we have

$$p(x_a, x_b) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\Sigma|^{1/2}} \exp\Big(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\Big)$$

The parameters μ and Σ can be determined

$$\mu = \begin{pmatrix} \mu_a \\ A\mu_a + b \end{pmatrix}$$

$$\Sigma = \begin{pmatrix} \Sigma_{aa} & \Sigma_{aa}A^T \\ A\Sigma_{aa} & \Sigma_{b|a} + A\Sigma_{aa}^{-1}A^T \end{pmatrix}$$

For the marginal $p(x_b)$, a marginal likelihood or model evidence

$$p(x_b) = \frac{1}{(2\pi)^{D_b/2}} \frac{1}{|\Sigma_b|^{1/2}} \exp\Big(-\frac{1}{2}(x_b-\mu_b)^T \Sigma_b^{-1}(x_b-\mu_b)\Big), \qquad \begin{aligned} \mu_b &= A\mu_a + b \\ \Sigma_{bb} &= \Sigma_{b|a} + A\Sigma_{aa}A^T \end{aligned}$$

The parameters μ_b and Σ_{bb} are from μ and Σ

The Gaussian distribution | Posteriors (cont.)

For the posterior $p(x_a|x_b)$, we have

$$p(x_a|x_b) = \frac{1}{(2\pi)^{D_a/2}} \frac{1}{|\Sigma_{a|b}|^{1/2}} \exp\Big(-\frac{1}{2}(x_a - \mu_{a|b})^T \Sigma_{a|b}^{-1} (x_a - \mu_{a|b}) \Big)$$

The mean $\mu_{a|b}$ and covariance $\Sigma_{a|b}$

$$\begin{aligned} \mu_{a|b} &= \Big(\Sigma_{aa}^{-1} + A^T \Sigma_{b|a}^{-1} A \Big)^{-1} \Big(A^T \Sigma_{b|a} (x_b - x_a) + \Sigma_{aa}^{-1} \mu_a \Big) \\ \Sigma_{a|b} &= \Big(\Sigma_{aa}^{-1} + A^T \Sigma_{b|a} A \Big)^{-1} \end{aligned}$$

We discussed an example of the **linear-Gaussian model**, the building block of many unsupervised techniques, like probabilistic principal component analysis and factor analysis, and many supervised techniques, like linear dynamical systems (Kalman filter)

Inference

The Gaussian distribution

The Gaussian | Inference

How to estimate the parameters of some multivariate Gaussian distribution from data?

We got data $\{x_1, \ldots, x_N\}$ which we assume to be independent draws from the Gaussian

$$p(x1, \ldots, x_N|\mu, \Sigma) = \prod_{n=1}^{N} p(x_n)$$
$$= \prod_{n=1}^{N} \frac{1}{(2\pi)^{D/2}} \frac{1}{|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x_n - \mu)^T \Sigma^{-1} (x_n - \mu)\right)$$

We can assume that μ and Σ are sure variables and maximise the likelihood function $p(x_1, \ldots, x_N|\mu, \Sigma)$ with respect to μ and Σ or by minimising the negative log-likelihood

$$\ln p(\{x_n\}|\mu, \Sigma) = \frac{1}{2}\sum_{n=1}^{N} (x_n - \mu)^T \Sigma^{-1} (x_n - \mu) + \frac{N}{2} \ln(|\Sigma|) + \frac{ND}{2} \ln(2\pi)$$

Optimisation yields the usual maximum likelihood estimates

$$\widehat{\mu}_{ML} = \frac{1}{N} \sum_{n=1}^{N} x_n$$

$$\widehat{\Sigma}_{ML} = \frac{1}{N(-1)} \sum_{n=1}^{N} (x_n - \widehat{\mu}_{ML})(x_n - \widehat{\mu}_{ML})^T$$

The Gaussian | Inference (cont.)

The maximum likelihood framework gave us point estimates for μ and Σ, in a Bayesian treatment one models the parameters as random variables and introduces their priors

We define a probabilistic model for all the random variables

$$p(\{x_n\}, \lambda, \Sigma) = p(\{x_n\}|\mu, \Sigma)p(\mu, \Sigma)$$

Then, we can use the Bayes rule to determine the posterior

$$p(\mu, \Sigma|\{x_n\}) = \frac{p(\{x_n\}|\mu, \Sigma)p(\mu, \Sigma)}{p(\{x_n\})}$$

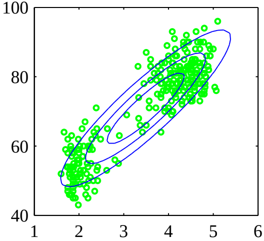
The conjugate prior for μ, assuming a known Σ^{-1} is a Gaussian distribution

The conjugate prior for Σ^{-1}, assuming a known μ is a **Wishart distribution**

The conjugate joint prior for μ and Σ^{-1} is a **Gaussian-Wishart distribution**

The Gaussian | Inference | Mixtures

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



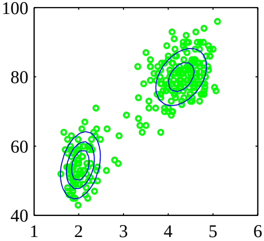
A Gaussian distribution inferred by maximum likelihood

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

A linear combination of two Gaussians distributions

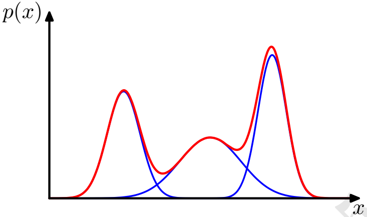
- Inference by maximum likelihood

Linear superpositions improve data representation



Superpositions obtained taking linear combinations of basic distributions, such as Gaussians, can be formulated as probabilistic data models, or **Gaussian mixture models**

The Gaussian | Inference | Mixtures (cont.)

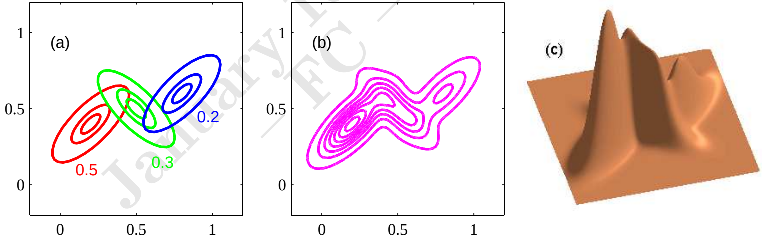


The one-dimensional mixture of Gaussians

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

We can get very complex densities

With sufficient number of Gaussians we can approximate almost any continuous density



- The contours at constant density for each of the Gaussian components
- The contours at constant density of the mixture distribution $p(x)$
- The surface plot of the mixture distribution $p(x)$
- The numbers (first plot) are the weights

The Gaussian | Inference | Mixtures (cont.)

We consider a linear superposition of K Gaussians

$$p(x) = \sum_{k=1}^{K} \pi_k p(x|\mu_k, \Sigma_k)$$

with

$$\begin{cases} \pi_k \in [0,1], k = 1, \dots, K \\ \sum_{k=1}^{K} \pi_k = 1 \end{cases}$$

Such a density model is a **mixture of Gaussians** in which each Gaussian density $p(x|\mu_k, \Sigma_k)$ is a component of the mixture, with its own mean vector μ_k and covariance matrix Σ_k

Mixing coefficients $\{\pi_k\}$, means $\{\mu_k\}$, and covariances $\{\Sigma_k\}$ are the model parameters

A way to infer the model parameters from data $\{x_n\}$ is to maximise the (log) likelihood

$$\ln p(\{x_n\}|\{\pi_k\}, \{\mu_k\}, \{\Sigma_k\}) = \sum_{n=1}^{N} \ln \Big(\sum_{k=1}^{K} \pi_k p(x_n|\mu_k, \Sigma_k) \Big)$$

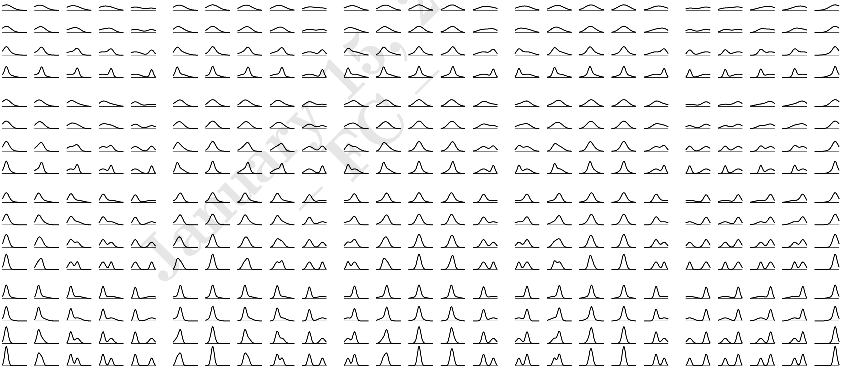
This solution treats parameters as sure variables and their point-estimates are inferred

- Standard non-linear program
- Expectation-maximisation

The Gaussian | Inference | Mixtures (cont.)

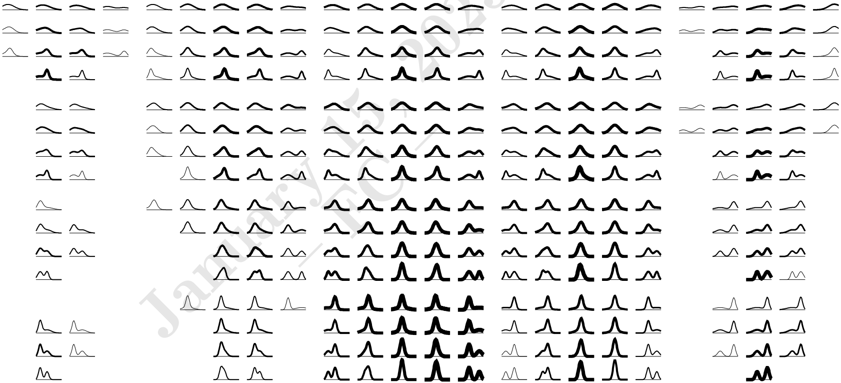
A discretised parameter space for a two-Gaussian mixture with $\pi_1 = 0.6$ and $\pi_2 = 0.4$

- The standard deviations of $\{\Sigma_1 = \sigma_1^2 I, \Sigma_2 = \sigma_2^2 I\}$ vary vertically
- The mean vectors $\{\mu_1, \mu_2\}$ vary horizontally



The Gaussian | Inference | Mixtures (cont.)

$$p(x) = \underbrace{0.6}_{\pi_1}\, p(x|\mu_1, \Sigma_1) + \underbrace{0,4}_{\pi_2}\, p(x|\mu_2, \Sigma_2)$$



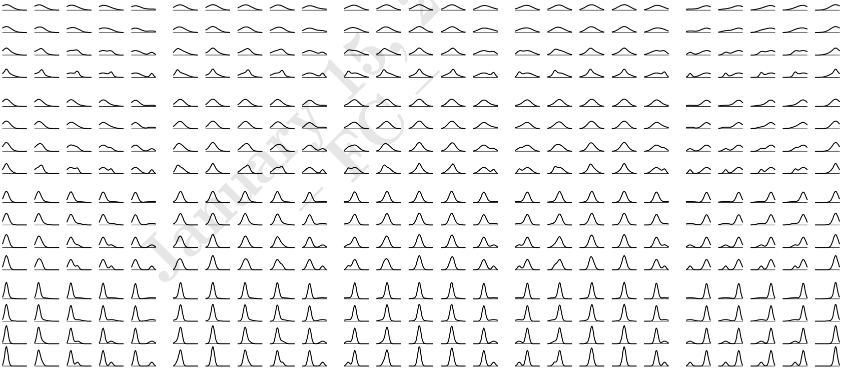
The likelihood function $p(\{x_n\}|\{\pi_k\}, \{\mu_k\}, \{\Sigma_k\})$ for some $\{x_n\}$ shown as line thickness

- Sub-hypothesis for which the maximum likelihood is smaller than e^{-8} blanked

The Gaussian | Inference | Mixtures (cont.)

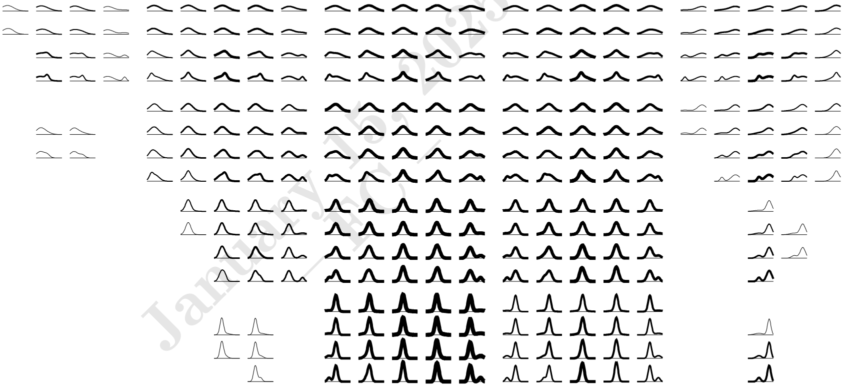
A discretised parameter space for a two-Gaussian mixture with $\pi_1 = 0.8$ and $\pi_2 = 0.2$

- The standard deviations of $\{\Sigma_1 = \sigma_1 I, \Sigma_2 = \sigma_2 I\}$ vary vertically
- The mean vectors $\{\mu_1, \mu_2\}$ vary horizontally



The Gaussian | Inference | Mixtures (cont.)

$$p(x) = \underbrace{0.8}_{\pi_1}\, p(x|\mu_1, \Sigma_1) + \underbrace{0,2}_{\pi_2}\, p(x|\mu_2, \Sigma_2)$$



The likelihood function $p(\{x_n\}|\{\pi_k\}, \{\mu_k\}, \{\Sigma_k\})$ for some $\{x_n\}$ shown as line thickness

- Sub-hypothesis for which the maximum likelihood is smaller than e^{-8} blanked

The Gaussian | Inference | Mixtures (cont.)

A mixture of K Gaussians in terms of latent variables

$$p(x) = \sum_{k=1}^{K} \pi_k p(x|\mu_k, \Sigma_k)$$

with

$$\begin{cases} \pi_k \in [0,1], k = 1, \dots, K \\ \sum_{k=1}^{K} \pi_k = 1 \end{cases}$$

A mixing coefficient π_k can be understood as the prior probability $p(k)$ of selecting the k-th Gaussian and $p(x|\mu_k, \Sigma_k)$ as the conditional probability density at x, given the k

$$p(x) = \sum_{k=1}^{K} \underbrace{p(x|k)p(k)}_{p(x,k)}$$

Using the Bayes' rule, we obtain the posterior probability of selecting the k-th Gaussian

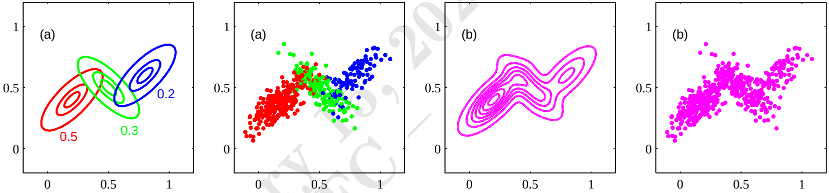
$$p(k|x) = \frac{p(k)p(x|k)}{\sum_l p(l)p(x|l)}$$

This solution treats means and covariance matrixes as sure variables, weights random

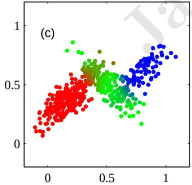
The Gaussian | Inference | Mixtures (cont.)

We can draw samples from the joint distribution $p(x, k))$, using ancestral sampling, and show them at the corresponding of x after colouring them according to the k value

Samples from the marginal $p(x)$ are obtained by sampling the joint and ignoring the k



For each sample, we can depict the posterior probability for each Gaussian component



A point x for which $p(k = 1(2, 3)) = 1|x)$ is red (blue, green)

Other points x have weighting doses of red, blue, and green