

**10707**

# **Deep Learning: Spring 2020**

Andrej Risteski

Machine Learning Department

## **Lecture 9:**

Common parametric  
distributions, Bayesian  
networks

# Unsupervised learning

Learning from data **without** labels.

What can we hope to do:

**Task A:** Fit a parametrized **structure** (e.g. clustering, low-dimensional subspace, manifold) to data to reveal something meaningful about data. (**Structure learning**)

**Task B:** Learn a (parametrized) **distribution** *close* to data generating distribution. (**Distribution learning**)

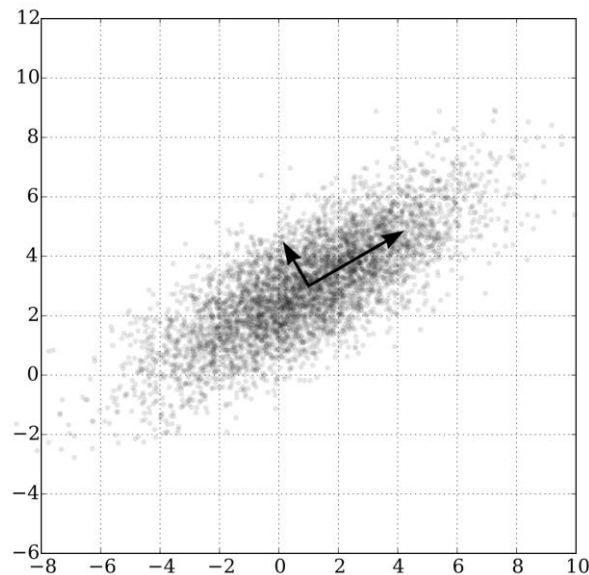
**Task C:** Learn a (parametrized) distribution that implicitly reveals an **“embedding”/“representation”** of data for downstream tasks. (**Representation/feature learning**)

*Entangled!* The “structure” and “distribution” often reveals an embedding.

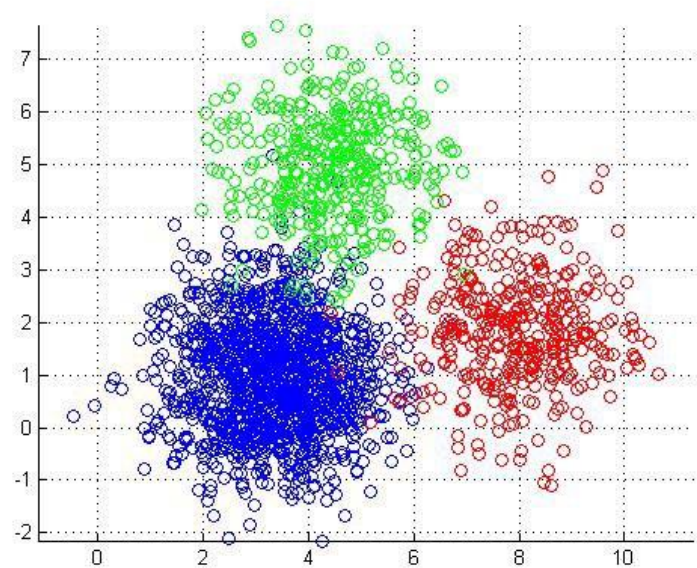
# Structure learning

Fit a parametrized **structure** (e.g. clustering, low-dimensional subspace) to data to reveal something meaningful about data.

**PCA**(principal component analysis),  
direction of highest variance



**Clustering**



# The classics: PCA

**Figure 1: Population structure within Europe.**

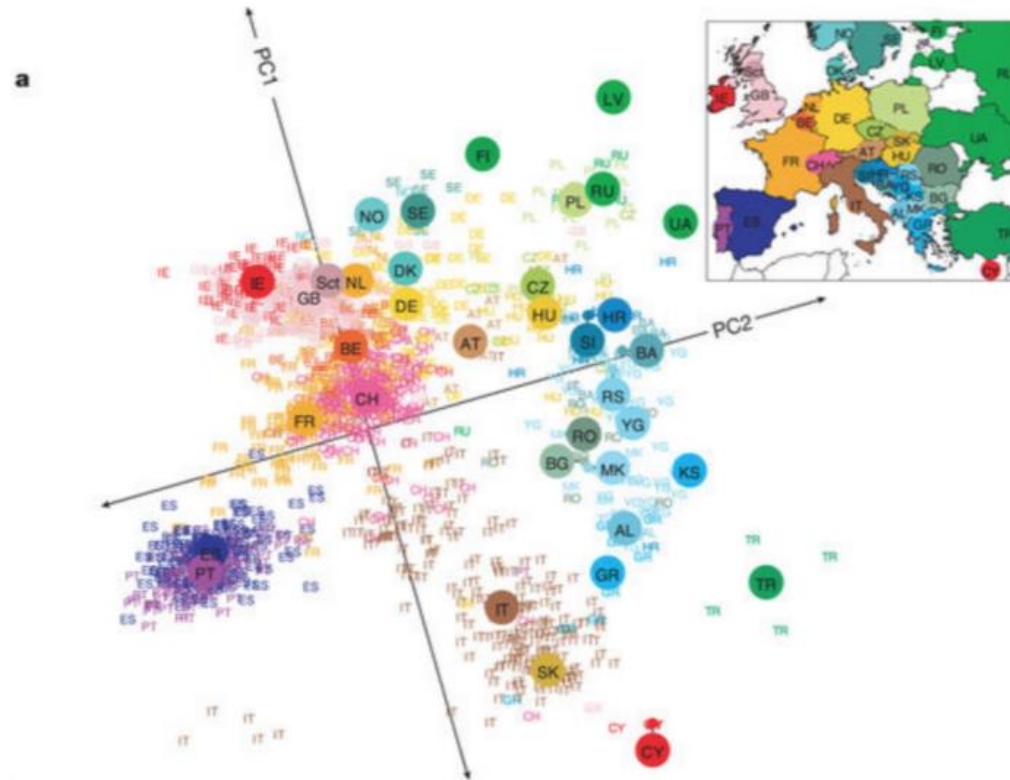


Figure 6: Plot from [1], depicting genomes for 1387 Europeans projected onto top 2 principal components. Colors/labels of datapoints correspond to geographic location of the individuals. Map of Europe (with same coloring) included in upper right for reference.

*Novembre et al '08*

# The classics: PCA

**Goal:** find a k-dimensional (linear) subspace explaining most of the variance in the data.

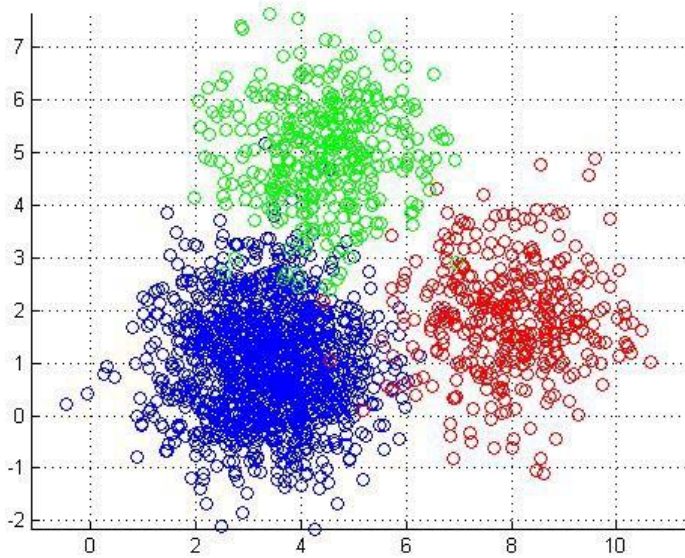
$$\max_{\{v_1, v_2, \dots, v_k\}} \frac{1}{m} \sum_{\text{samples } x_i} (\text{length of } x_i \text{ on } \text{span}(v_1, v_2, \dots, v_k))^2$$

$$= \max_{\text{orthonormal } \{v_1, v_2, \dots, v_k\}} \frac{1}{m} \sum_{\text{samples } x_i} \sum_{j=1}^k \langle x_i, v_j \rangle^2$$

Variance:  $\mathbb{E}[\sum_j \langle v_j, x \rangle^2]$



# The classics: clustering

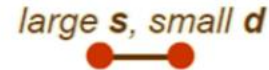
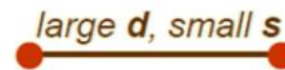


**Goal:** group the data into clusters of nearby points.

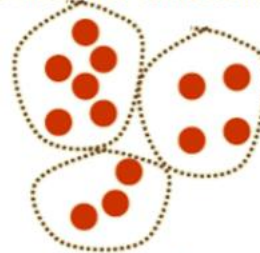
## What's needed for clustering?

### 1. Proximity measure, *either*

- similarity measure  $s(x_i, x_k)$ : large if  $x_i, x_k$  are similar
- dissimilarity (or distance) measure  $d(x_i, x_k)$ : small if  $x_i, x_k$  are similar



### 2. Criterion function to evaluate a clustering



*good clustering*



*bad clustering*

### 3. Algorithm to compute clustering

# K-means clustering

If the distance metric is the **Euclidean distance**, and the measure of cohesion is the **average distance from the centroid**: we get the **k-means objective**.

$$\operatorname{argmin}_{\{r_{nk}, \mu_k\}} \sum_{n=1}^N \sum_{k=1}^K r_{nk} \|\mathbf{x}_n - \mu_k\|^2$$

*Is point  $n$  in cluster  $k$ ?*

*Centroid of  $k$ -th cluster*

# K-means clustering

A natural iterative algorithm, which as we will see later is a variant of the **EM (expectation-maximization)** algorithm:

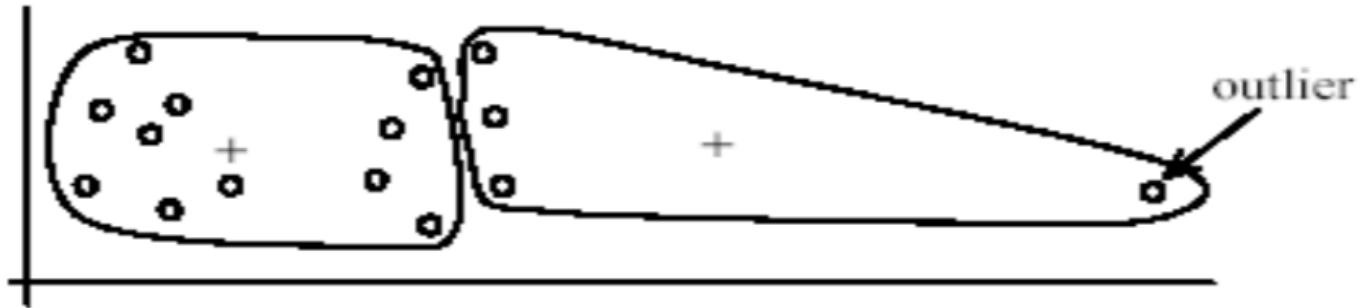
```
Input: Data set  $X = \{x^{(1)}, x^{(2)}, \dots, x^{(m)} | x^{(i)} \in \mathbb{R}^n\}$ 
Output: Cluster centroids  $\mu_{i=1, \dots, k} \in \mathbb{R}^n$ ; Cluster assignments  $c \in \mathbb{Z}$ 
1 Initialize  $k$  cluster centroids  $\mu_1, \dots, \mu_k \in \mathbb{R}^n$  randomly from  $X$ ;
2 repeat
3   for  $i = 1, \dots, m$  do // Update cluster assignments
4     | set  $c^{(i)} = \arg \min_j \|x^{(i)} - \mu_j\|^2$ ;
5   end
6   for  $j = 1, \dots, k$  do // Update cluster centroids
7     | set  $\mu_j = \frac{\sum_{i=1}^m 1_{\{c^{(i)}=j\}} x^{(i)}}{\sum_{i=1}^m 1_{\{c^{(i)}=j\}}}$ ;
8   end
9 until Convergence;
10 return  $\mu$  and  $c$ ;
```

Algorithm 1: Algorithm of batch-version for K-means

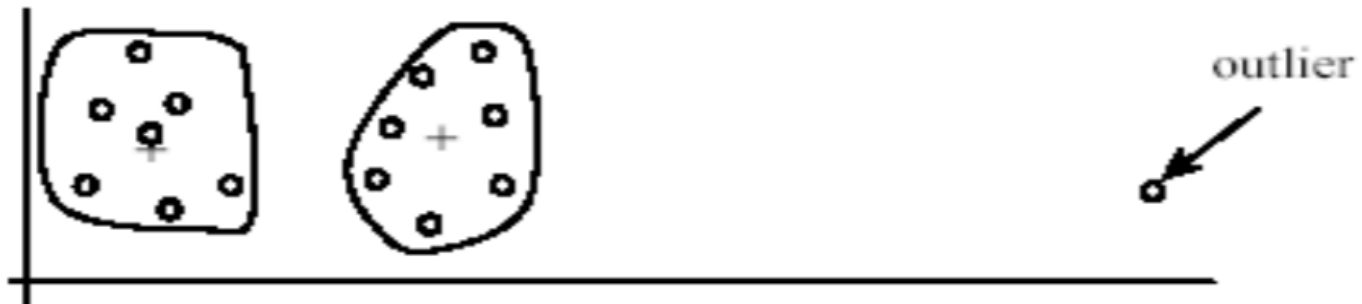


# Some weaknesses

Very sensitive to outliers:



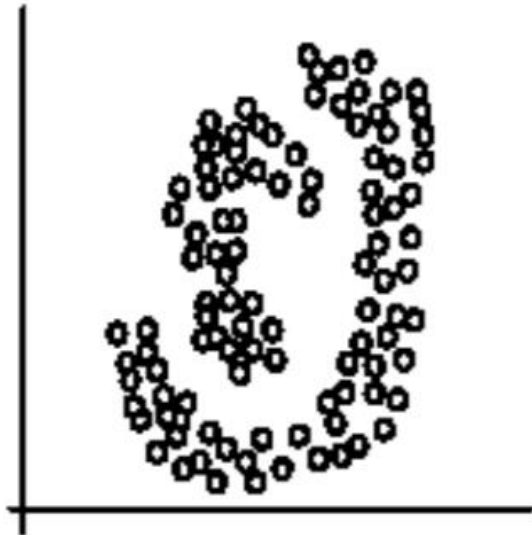
(A): Undesirable clusters



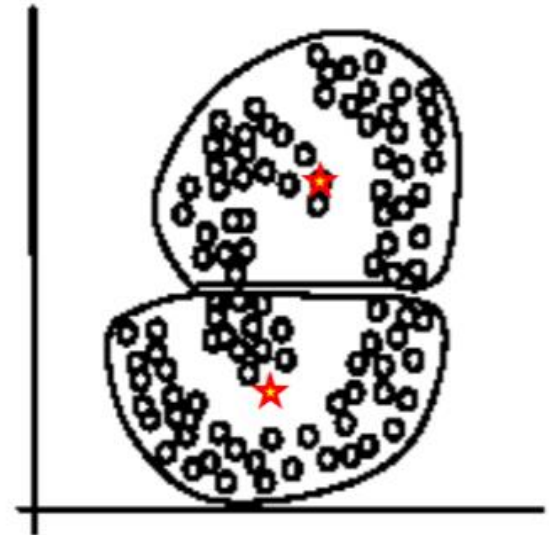
(B): Ideal clusters

# Some weaknesses

Not suitable for non-spherical clusters:



(A): Two natural clusters

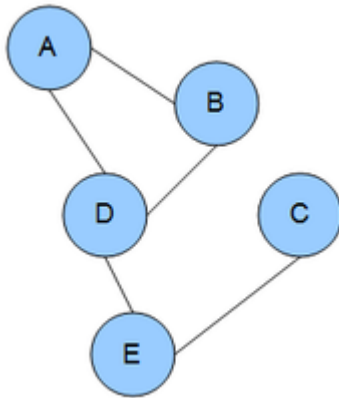


(B):  $k$ -means clusters

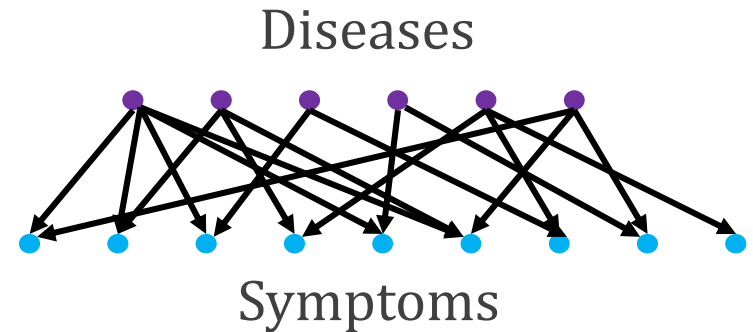
# Distribution learning

Some typical choices of parametrized distributions:

**Classical choices:** **fully-observed** graphical models (undirected and directed), **latent-variable** graphical models (mixture models, sparse coding, topic models).



**Markov Random Fields:**  
sparse independence  
structure: “A is independent of  
other vars, given B, D”



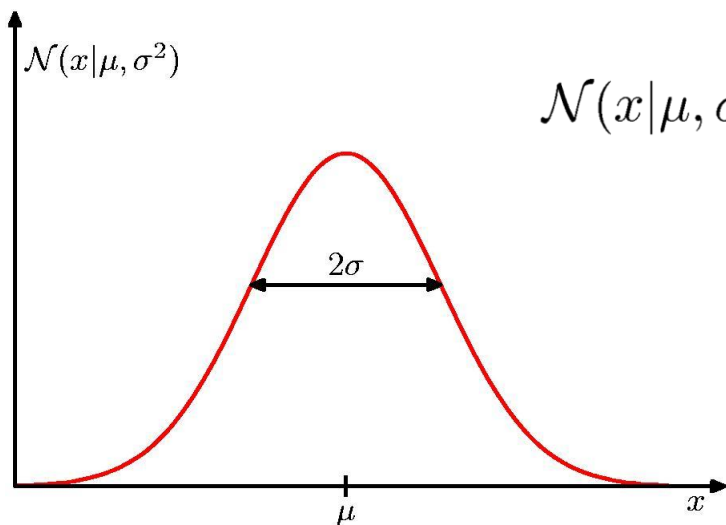
**Latent variable models:** data is  
“simple” conditioned on some  
unobserved (latent) variables

# Distribution learning

Before getting to graphical models, let's recap some basic distributions we will use as building blocks.

# Gaussian Univariate Distribution

In the case of a single variable  $x$ , the pdf of the **Gaussian distribution** is:



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

which is governed by two parameters:

$\mu$  (**mean**)  
 $\sigma$  (**variance**)

(In that, given the values of the mean/variance, there is a unique Gaussian with those mean/variance.)

# Why are mean and variance as claimed?

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

We wish to show that  $\mathbb{E}[x] = \mu$ ,  $\mathbb{E}[(x - \mu)^2] = \sigma^2$

By change of variables, suffices to show this for  $\mu = 0, \sigma = 1$

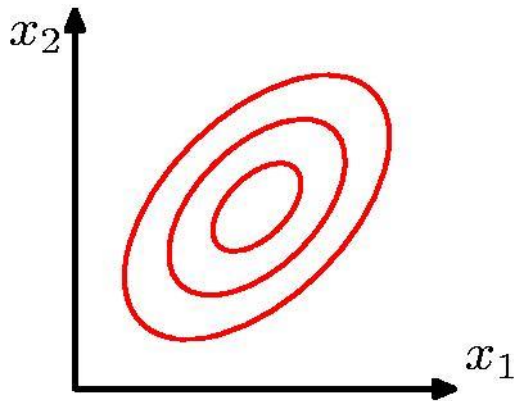
$$\mathbb{E}[x] = \int_{-\infty}^{\infty} x \frac{1}{\sqrt{2\pi}} \exp(-x^2/2) dx = \frac{1}{2} \int_{-\infty}^{\infty} (x + (-x)) \frac{1}{\sqrt{2\pi}} \exp(-x^2/2) dx = 0$$

$$\begin{aligned} \mathbb{E}[(x - \mu)^2] &= \int_{-\infty}^{\infty} x^2 \frac{1}{\sqrt{2\pi}} \exp(-x^2/2) dx = 2 \int_0^{\infty} x^2 \frac{1}{\sqrt{2\pi}} \exp(-x^2/2) dx \\ &= \frac{2}{\sqrt{2\pi}} \int_0^{\infty} x \underbrace{(x \exp(-x^2/2))}_{:= dv, v = -\left(\exp\left(-\frac{x^2}{2}\right)\right)} dx = \frac{2}{\sqrt{2\pi}} \int_0^{\infty} x dv = \frac{2}{\sqrt{2\pi}} \left( xv \Big|_0^{\infty} - \int_0^{\infty} v dx \right) \\ &= \frac{2}{\sqrt{2\pi}} \left( \int_0^{\infty} \exp\left(-\frac{x^2}{2}\right) dx \right) = \frac{1}{\sqrt{2\pi}} \left( \int_{-\infty}^{\infty} \exp\left(-\frac{x^2}{2}\right) dx \right) = 1 \end{aligned}$$

# Multivariate Gaussian Distribution

For a D-dimensional vector  $\mathbf{x}$ , the Gaussian distribution takes form:

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



which is governed by two parameters:

$\boldsymbol{\mu}$  (**mean**) is a D-dimensional mean vector.

$\boldsymbol{\Sigma}$  (**covariance**) is a D by D positive definite matrix,  $|\boldsymbol{\Sigma}|$  denotes the determinant of  $\boldsymbol{\Sigma}$ .

(In that, given the values of the mean/variance, there is a unique Gaussian with those mean/variance.)

# Why are mean and covariance as claimed?

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

We wish to show that  $\mathbb{E}[\mathbf{x}] = \boldsymbol{\mu}$ ,  $\mathbb{E}[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^T] = \boldsymbol{\Sigma}$

By change of variables, suffices to show this for  $\boldsymbol{\mu} = 0$

Expectation is easy:

$$\mathbb{E}[\mathbf{x}] = \int \mathbf{x} \frac{1}{\sqrt{\frac{D}{2\pi^2}|\boldsymbol{\Sigma}|}} \exp\left(-\frac{1}{2}\mathbf{x}^T \boldsymbol{\Sigma}^{-1} \mathbf{x}\right) d\mathbf{x} = \frac{1}{2} \int (\mathbf{x} + (-\mathbf{x})) \frac{1}{\sqrt{\frac{D}{2\pi^2}|\boldsymbol{\Sigma}|}} \exp\left(-\frac{1}{2}\mathbf{x}^T \boldsymbol{\Sigma}^{-1} \mathbf{x}\right) d\mathbf{x} = 0$$

Covariance can be reduced to  $\boldsymbol{\Sigma}=\mathbf{I}$ :

One can produce a sample from a Gaussian w/ covariance  $\boldsymbol{\Sigma}$  as

$\mathbf{x} = \boldsymbol{\Sigma}^{1/2} \mathbf{y}$ ,  $\mathbf{y}$  is sampled from a Gaussian w/ covariance  $\mathbf{I}$ .

(Very useful, check it!)

Since the covariance of  $\mathbf{x}$  is  $\mathbb{E}[\mathbf{x}\mathbf{x}^T] = \boldsymbol{\Sigma}^{\frac{1}{2}} \mathbb{E}[\mathbf{y}\mathbf{y}^T] \boldsymbol{\Sigma}^{\frac{1}{2}}$ , it suffices to show the covariance of  $\mathbf{y}$  is  $\mathbf{I}$ .



# Why are mean and covariance as claimed?

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

We wish to show that  $\mathbb{E}[\mathbf{x}] = \boldsymbol{\mu}$ ,  $\mathbb{E}[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^T] = \boldsymbol{\Sigma}$

By change of variables, suffices to show this for  $\boldsymbol{\mu} = \mathbf{0}$

Expectation is easy:

$$\mathbb{E}[\mathbf{x}] = \int \mathbf{x} \frac{1}{\sqrt{\frac{D}{2\pi^2}|\boldsymbol{\Sigma}|}} \exp\left(-\frac{1}{2}\mathbf{x}^T \boldsymbol{\Sigma}^{-1} \mathbf{x}\right) d\mathbf{x} = \frac{1}{2} \int (\mathbf{x} + (-\mathbf{x})) \frac{1}{\sqrt{\frac{D}{2\pi^2}|\boldsymbol{\Sigma}|}} \exp\left(-\frac{1}{2}\mathbf{x}^T \boldsymbol{\Sigma}^{-1} \mathbf{x}\right) d\mathbf{x} = \mathbf{0}$$

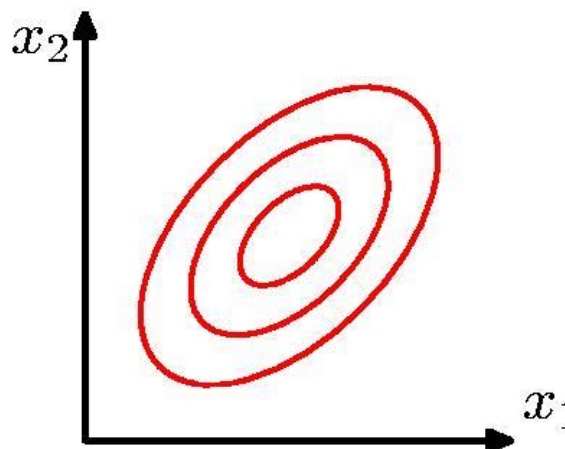
Covariance can be reduced to  $\boldsymbol{\Sigma}=\mathbf{I}$ :

A Gaussian with  $\boldsymbol{\Sigma}=\mathbf{I}$  is just a product distribution of standard Gaussians (check this!)

Hence, the covariance matrix has diagonals 1 (we showed this already), and off-diagonals 0 (since the off-diagonals are  $\mathbb{E}[\mathbf{x}_i \mathbf{x}_j] = \mathbb{E}[\mathbf{x}_i] \mathbb{E}[\mathbf{x}_j] = \mathbf{0}$ )

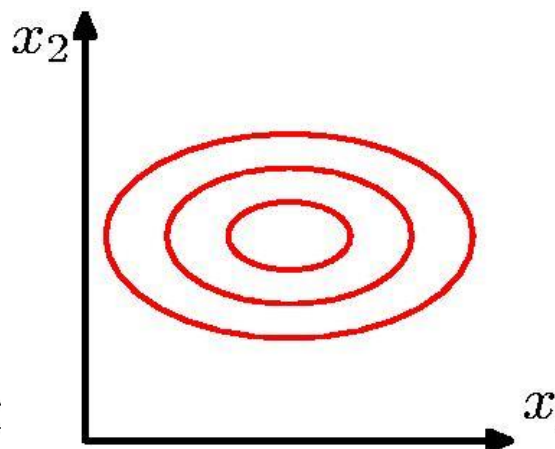
# Multivariate Gaussian Distribution

Contours of constant probability density:



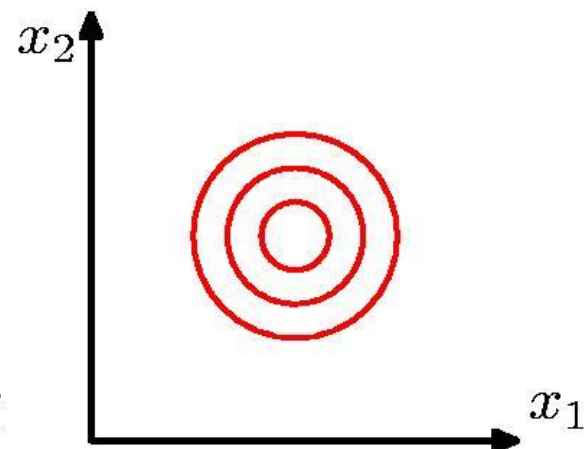
(a)

Covariance matrix is of general form.



(b)

Diagonal, axis-aligned covariance matrix.



(c)

Spherical (proportional to identity) covariance matrix.

# Central Limit Theorem

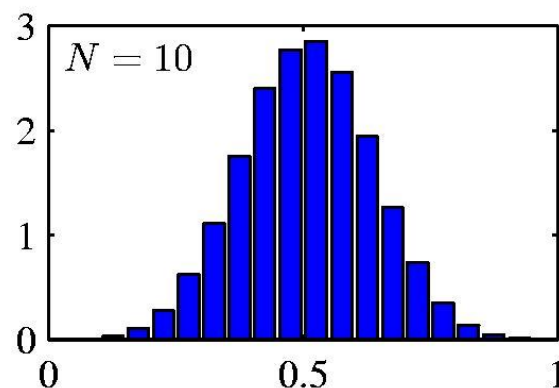
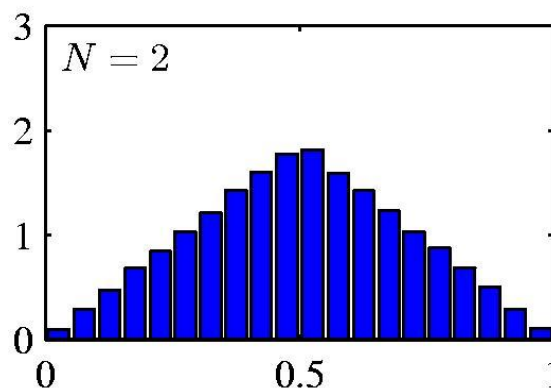
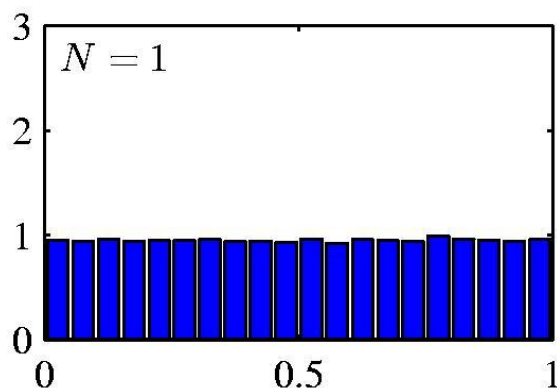
The distribution of the sum of  $N$  i.i.d. random variables becomes “increasingly Gaussian” as  $N$  grows. (Can be formalized.)

Consider  $N$  variables, each of which has a uniform distribution over the interval  $[0,1]$ .

Let us look at the distribution over the mean:

$$\frac{x_1 + x_2 + \dots + x_N}{N}.$$

As  $N$  increases, the distribution tends towards a Gaussian distribution. (Can be made quantitative.)



# Marginals and conditional of Gaussians

Consider a D-dimensional Gaussian distribution:  $p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$

Let us partition  $\mathbf{x}$  into two disjoint subsets  $\mathbf{x}_a$  and  $\mathbf{x}_b$ :

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

In many situations, it will be more convenient to work with the *precision matrix* (inverse of the covariance matrix):

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

(Note that  $\boldsymbol{\Lambda}_{aa}$  is not given by the inverse of  $\boldsymbol{\Sigma}_{aa}$ )

# Helpful result: inverting block matrices

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

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

## Block matrix inversion lemma:

$$\begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix}^{-1} = \begin{bmatrix} (\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1} & -(\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1}\mathbf{B}\mathbf{D}^{-1} \\ -\mathbf{D}^{-1}\mathbf{C}(\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1} & \mathbf{D}^{-1} + \mathbf{D}^{-1}\mathbf{C}(\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1}\mathbf{B}\mathbf{D}^{-1} \end{bmatrix}.$$

Here,  $\mathbf{D}$  and  $\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C}$  must be invertible.

So, for instance,  $\boldsymbol{\Lambda}_{aa}^{-1} = \boldsymbol{\Sigma}_{aa} - \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}\boldsymbol{\Sigma}_{ba}$  etc.

# Conditionals of Gaussian

It turns out that the conditional distribution is also a Gaussian distribution:

$$p(\mathbf{x}_B | \mathbf{x}_A) = \mathcal{N}(\mathbf{x}_B | \boldsymbol{\mu}_{B|A}, \boldsymbol{\Sigma}_{B|A})$$

Covariance does  
not depend on  $\mathbf{x}_A$ .

$$\begin{aligned}\boldsymbol{\Sigma}_{B|A} &= \boldsymbol{\Lambda}_{BB}^{-1} = \boldsymbol{\Sigma}_{BB} - \boldsymbol{\Sigma}_{BA} \boldsymbol{\Sigma}_{AA}^{-1} \boldsymbol{\Sigma}_{AB} \\ \boldsymbol{\mu}_{B|A} &= \boldsymbol{\mu}_B - \boldsymbol{\Lambda}_{BB}^{-1} \boldsymbol{\Lambda}_{BA} (\mathbf{x}_A - \boldsymbol{\mu}_A) \\ &= \boldsymbol{\mu}_B - \boldsymbol{\Sigma}_{BA} \boldsymbol{\Sigma}_{AA}^{-1} (\mathbf{x}_A - \boldsymbol{\mu}_A)\end{aligned}$$

Linear function  
of  $\mathbf{x}_A$ .

# Why is this true?

First, write out the conditional explicitly:

$$\begin{aligned} p(x_B \mid x_A) &= \frac{p(x_A, x_B; \mu, \Sigma)}{\int_{x_B \in \mathbf{R}^m} p(x_A, x_B; \mu, \Sigma) dx_B} \\ &= \frac{1}{Z'} \exp \left( -\frac{1}{2} \begin{bmatrix} x_A - \mu_A \\ x_B - \mu_B \end{bmatrix}^T \begin{bmatrix} \Sigma_{AA} & \Sigma_{AB} \\ \Sigma_{BA} & \Sigma_{BB} \end{bmatrix}^{-1} \begin{bmatrix} x_A - \mu_A \\ x_B - \mu_B \end{bmatrix} \right) \end{aligned}$$

where  $Z'$  is a normalizing constant.

The expression inside the exponential is a quadratic in  $x_A$  : so the conditional is a Gaussian!

We will massage the expression to get it into a nicer form to extract the covariance and mean. (We will do the matrix analogue of “**completing the square**”.)

# Why is this true?

First, write out the conditional explicitly:

$$\begin{aligned} p(x_B \mid x_A) &= \frac{p(x_A, x_B; \mu, \Sigma)}{\int_{x_B \in \mathbf{R}^m} p(x_A, x_B; \mu, \Sigma) dx_A} \\ &= \frac{1}{Z'} \exp \left( -\frac{1}{2} \begin{bmatrix} x_A - \mu_A \\ x_B - \mu_B \end{bmatrix}^T \begin{bmatrix} \Sigma_{AA} & \Sigma_{AB} \\ \Sigma_{BA} & \Sigma_{BB} \end{bmatrix}^{-1} \begin{bmatrix} x_A - \mu_A \\ x_B - \mu_B \end{bmatrix} \right) \end{aligned}$$

where  $Z'$  is a normalizing constant.

Using the notation we introduced, we have:

$$\begin{aligned} p(x_B \mid x_A) &= \frac{1}{Z'} \exp \left( -\frac{1}{2} \begin{bmatrix} x_A - \mu_A \\ x_B - \mu_B \end{bmatrix}^T \begin{bmatrix} \Lambda_{AA} & \Lambda_{AB} \\ \Lambda_{BA} & \Lambda_{BB} \end{bmatrix} \begin{bmatrix} x_A - \mu_A \\ x_B - \mu_B \end{bmatrix} \right) \\ &= \frac{1}{Z'} \exp \left( - \left[ \begin{aligned} &\frac{1}{2} (x_A - \mu_A)^T \Lambda_{AA} (x_A - \mu_A) + \frac{1}{2} (x_A - \mu_A)^T \Lambda_{AB} (x_B - \mu_B) \\ &\frac{1}{2} (x_B - \mu_B)^T \Lambda_{BA} (x_A - \mu_A) + \frac{1}{2} (x_B - \mu_B)^T \Lambda_{BB} (x_B - \mu_B) \end{aligned} \right] \right). \end{aligned}$$

Expanding block form



# Why is this true?

Using the notation we introduced, we have:

$$p(x_B | x_A) = \frac{1}{Z'} \exp \left( - \left[ \underbrace{\frac{1}{2} (x_A - \mu_A)^T \Lambda_{AA} (x_A - \mu_A)}_c + \underbrace{\frac{1}{2} (x_A - \mu_A)^T \Lambda_{AB} (x_B - \mu_B)}_{\frac{1}{2} b^T z} + \underbrace{\frac{1}{2} (x_B - \mu_B)^T \Lambda_{BA} (x_A - \mu_A)}_{\frac{1}{2} b^T z} + \underbrace{\frac{1}{2} (x_B - \mu_B)^T \Lambda_{BB} (x_B - \mu_B)}_{z^T A z} \right] \right).$$

The “completion of squares” trick:

$$\frac{1}{2} z^T A z + b^T z + c = \frac{1}{2} (z + A^{-1} b)^T A (z + A^{-1} b) + c - \frac{1}{2} b^T A^{-1} b$$

Apply above:  $z = x_B - \mu_B, A = \Lambda_{BB}, b = \Lambda_{BA}(x_A - \mu_A), c = \frac{1}{2} (x_A - \mu_A)^T \Lambda_{AA} (x_A - \mu_A)$

$$p(x_B | x_A) = \frac{1}{Z'} \exp \left( - \left[ \frac{1}{2} \left( x_B - \mu_B + \Lambda_{BB}^{-1} \Lambda_{BA} (x_A - \mu_A) \right)^T \Lambda_{BB} \left( x_B - \mu_B + \Lambda_{BB}^{-1} \Lambda_{BA} (x_A - \mu_A) \right) + \frac{1}{2} (x_A - \mu_A)^T \Lambda_{AA} (x_A - \mu_A) - \frac{1}{2} (x_A - \mu_A)^T \Lambda_{AA} \Lambda_{BB}^{-1} \Lambda_{BA} (x_A - \mu_A) \right] \right)$$

# Marginal Distribution

It turns out that the marginal distribution is also a Gaussian distribution:

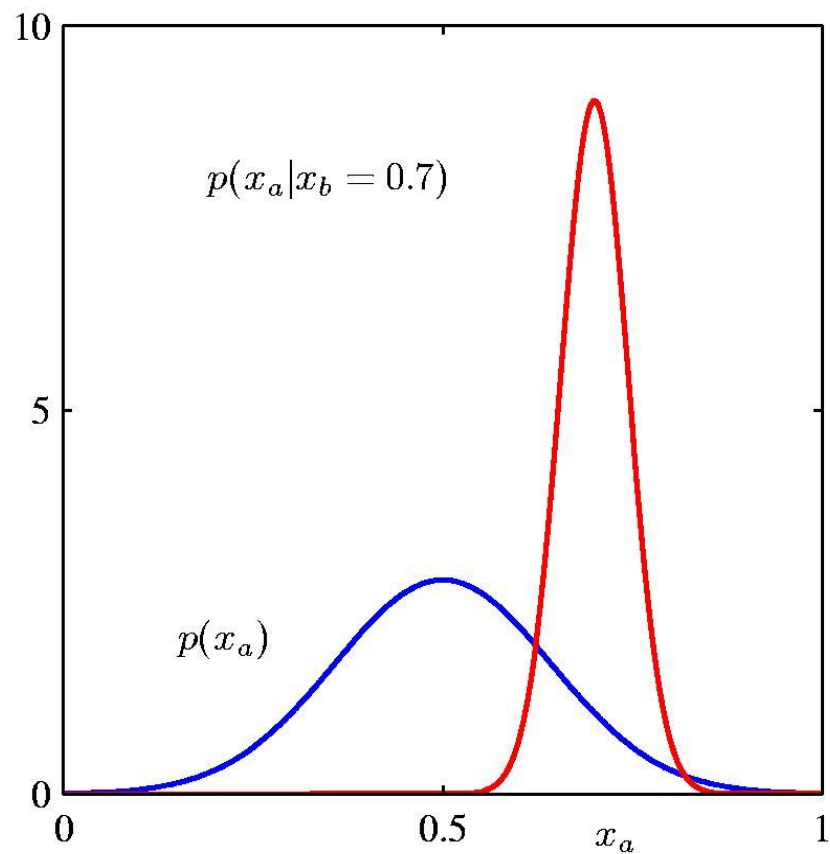
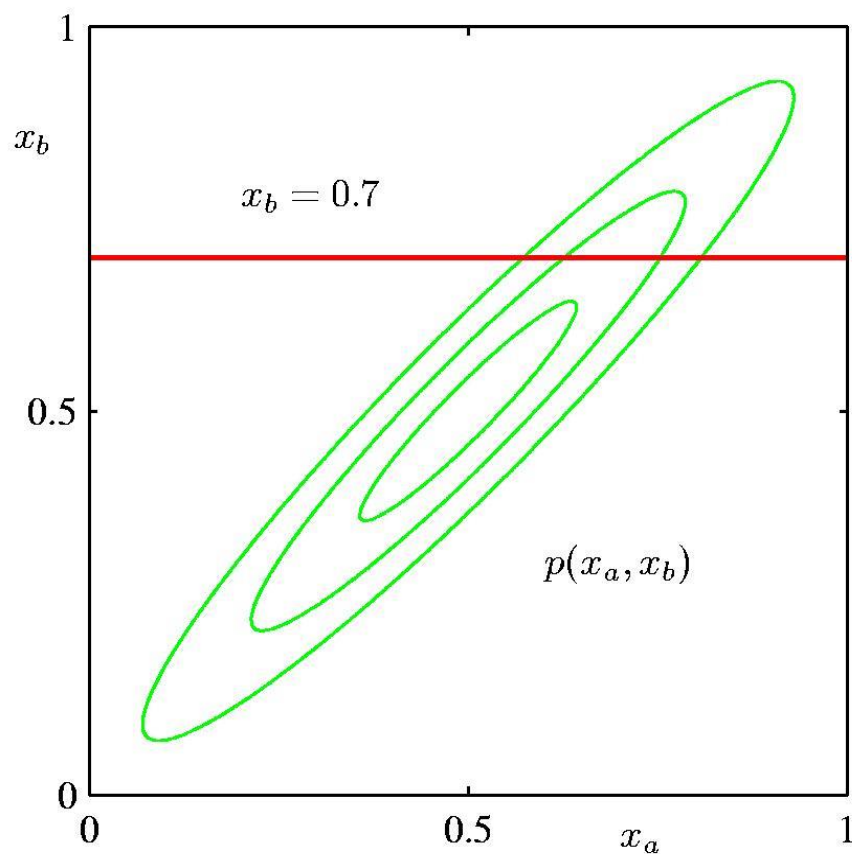
$$\begin{aligned} p(\mathbf{x}_a) &= \int p(\mathbf{x}_a, \mathbf{x}_b) d\mathbf{x}_b \\ &= \mathcal{N}(\mathbf{x}_a | \boldsymbol{\mu}_a, \boldsymbol{\Sigma}_{aa}) \end{aligned}$$

Similarly as before, one can show that the corresponding covariance

and mean are (surprisingly!!):  $\boldsymbol{\Sigma}_{AA}, \boldsymbol{\mu}_A$

(i.e. we just drop the corresponding rows/columns)

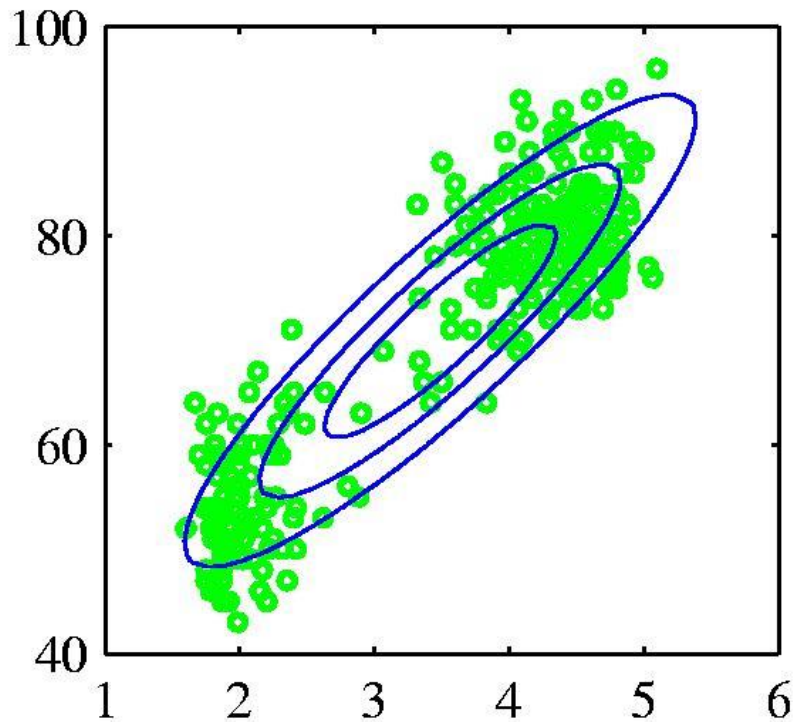
# Conditional and Marginal Distributions



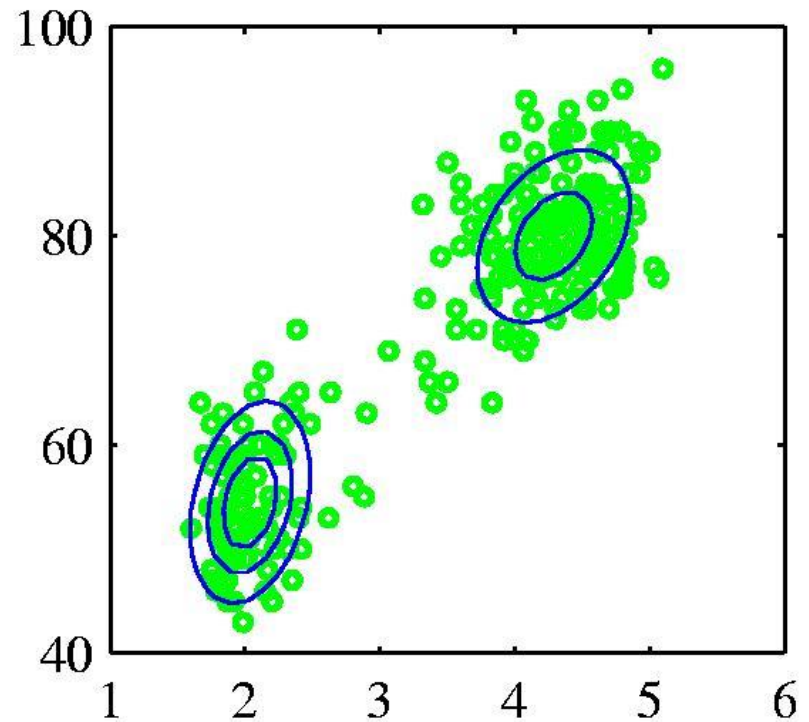
# Mixture of Gaussians

When modeling real-world data, Gaussian assumption may not be appropriate.

Consider the following example: Old Faithful Dataset



Single Gaussian



Mixture of two  
Gaussians

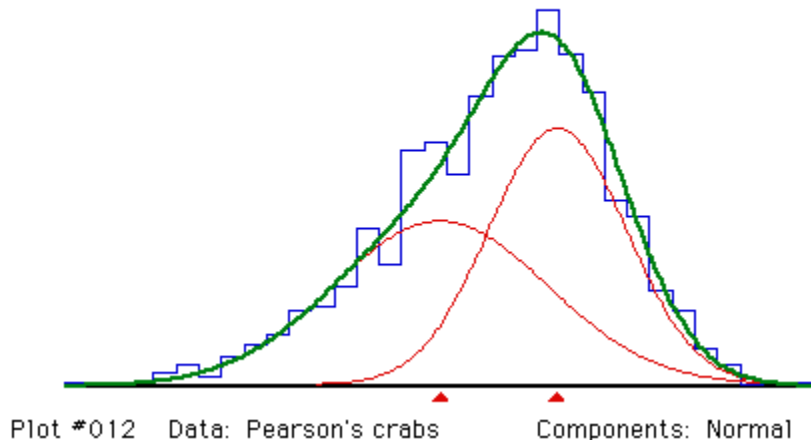
# Mixture of Gaussians

When modeling real-world data, Gaussian assumption may not be appropriate.

Historical aside: Karl Pearson's crabs (1894)

Pearson had access to ~1000 collected width/length crab measurements from an island on Malta.

He (ahead of his time) argued that there is a mixture that matches the measurements much better than a Gaussian (fitted the mixture by hand a novel procedure!).



Thus, he argued that there are two separate species of crabs.

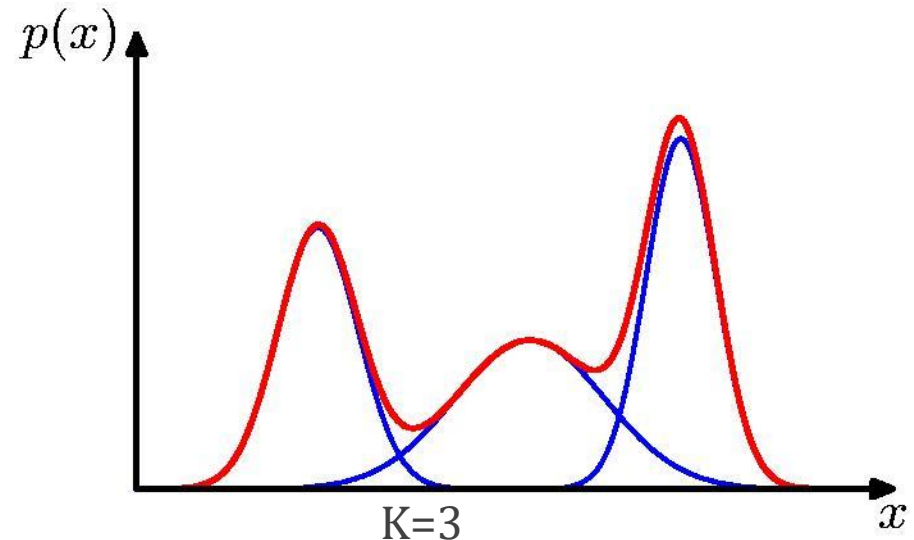
# Mixture of Gaussians

We can combine simple models into a complex model by defining a superposition of  $K$  Gaussian densities of the form:

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

Mixing coefficient

$$\forall k : \pi_k \geq 0 \quad \sum_{k=1}^K \pi_k = 1$$

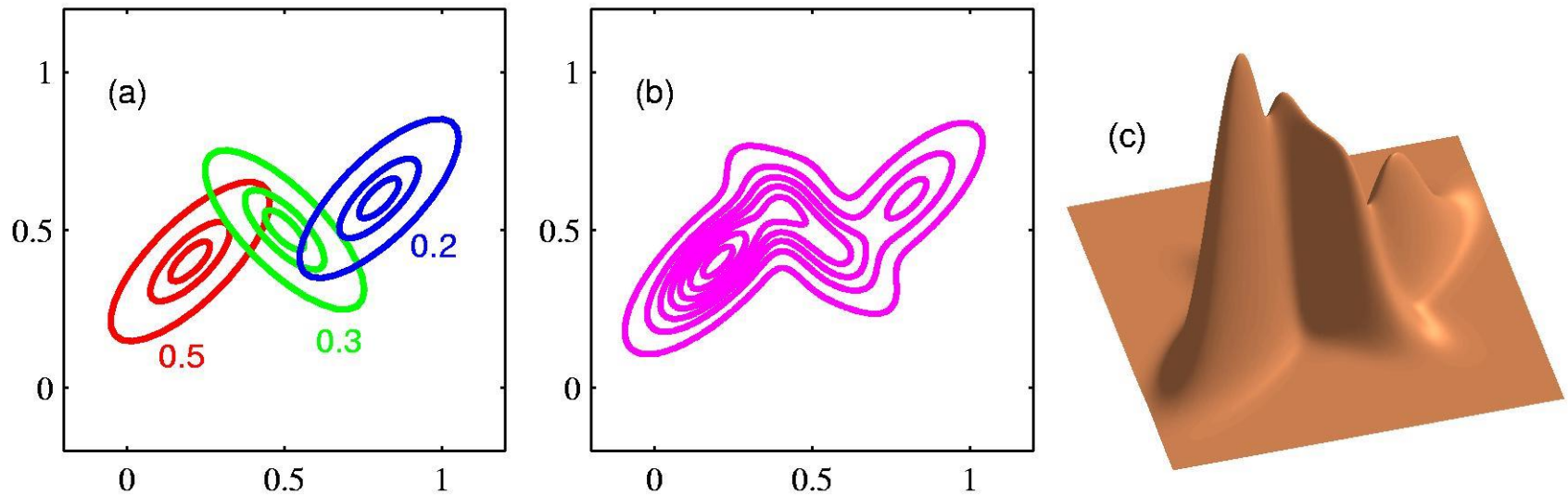


Note that each Gaussian component has its own mean  $\boldsymbol{\mu}_k$  and covariance  $\boldsymbol{\Sigma}_k$ . The parameters  $\boldsymbol{\pi}_k$  are called mixing coefficients.

Mote generally, mixture models can comprise linear combinations of other distributions.

# Mixture of Gaussians

Illustration of a mixture of 3 Gaussians in a 2-dimensional space:



(a) Contours of constant density of each of the mixture components, along with the mixing coefficients

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

(c) A surface plot of the distribution  $p(\mathbf{x})$ .

# Bernoulli Distribution

Consider a single *binary* random variable  $x \in \{0, 1\}$ . For example,  $x$  can describe the outcome of flipping a coin:

Coin flipping: heads = 1, tails = 0.

The probability of  $x=1$  will be denoted by the **parameter**  $\mu$ , so that:

$$p(x = 1|\mu) = \mu \quad 0 \leq \mu \leq 1.$$

The resulting **Bernoulli distribution** can be written as:

$$\text{Bern}(x|\mu) = \mu^x (1 - \mu)^{1-x}$$

$$\mathbb{E}[x] = \mu$$

$$\text{var}[x] = \mu(1 - \mu)$$



# Binomial Distribution

We can also work out the distribution of the number  $m$  of observations of  $x=1$  (e.g. the number of heads), i.e.

$\sum_{i=1}^N X_i$ ,  $X_i$  is Bernoulli with parameter  $\mu$

The probability of observing  $m$  heads given  $N$  coin flips and a parameter  $\mu$  is given by:

$$p(m \text{ heads} | N, \mu) = \text{Bin}(m | N, \mu) = \binom{N}{m} \mu^m (1 - \mu)^{N-m}$$

The **mean and variance** can be easily derived as:

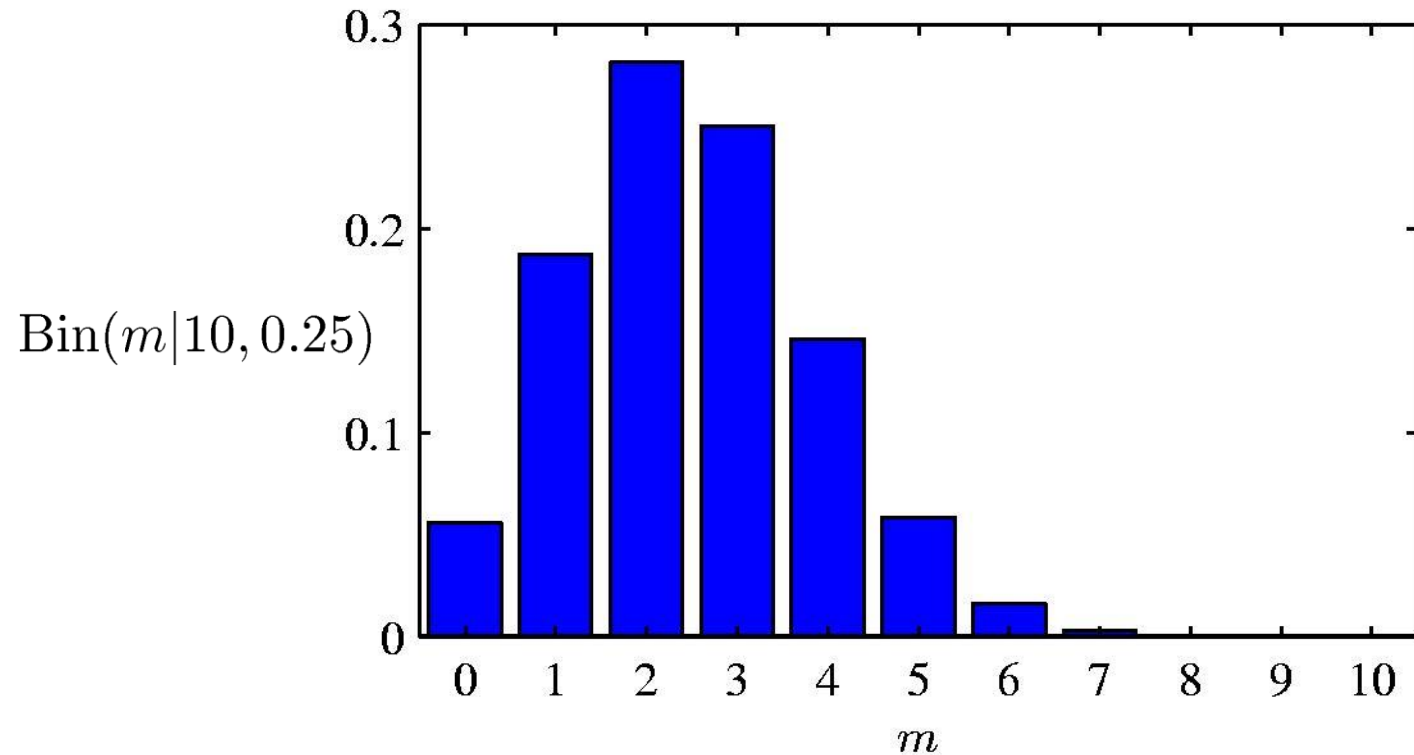
$$\mathbb{E}[m] = N \mu$$

$$\text{var}[m] = N \mu (1 - \mu)$$

By law of large number, for large  $N$ , close to  $\mathcal{N}(N \mu, N \mu (1 - \mu))$

# Example

Histogram plot of the Binomial distribution as a function of  $m$  for  $N=10$  and  $\mu = 0.25$ .



# Categorical distribution

Consider a random variable that can take on one of  $K$  possible mutually exclusive states (e.g. roll of a dice).

We will use so-called 1-of- $K$  encoding scheme.

If a random variable can take on  $K=6$  states, and a particular observation of the variable corresponds to the state  $x_3=1$ , then  $\mathbf{x}$  will be represented as:

$$\mathbf{x} = (0, 0, 1, 0, 0, 0)^T$$

If we denote the probability of  $x_k=1$  by the parameter  $\mu_k$ , then the categorical distribution over  $\mathbf{x}$  is defined as:

$$p(\mathbf{x}|\boldsymbol{\mu}) = \prod_{k=1}^K \mu_k^{x_k} \quad \forall k : \mu_k \geq 0 \quad \text{and} \quad \sum_{k=1}^K \mu_k = 1$$

# Categorical distribution

Categorical distribution can be viewed as a generalization of Bernoulli distribution to more than two outcomes.

$$p(\mathbf{x}|\boldsymbol{\mu}) = \prod_{k=1}^K \mu_k^{x_k}$$

It is easy to see that the distribution is normalized:

$$\sum_{\mathbf{x}} p(\mathbf{x}|\boldsymbol{\mu}) = \sum_{k=1}^K \mu_k = 1$$

and

$$\mathbb{E}[\mathbf{x}|\boldsymbol{\mu}] = \sum_{\mathbf{x}} p(\mathbf{x}|\boldsymbol{\mu})\mathbf{x} = (\mu_1, \dots, \mu_K)^T = \boldsymbol{\mu}$$

# Multinomial Distribution

We can construct the joint distribution of the quantities  $\{m_i\}$  given the parameters  $\{\mu_i\}$  and the total number  $N$  of observations:

$$\text{Mult}(m_1, m_2, \dots, m_K | \boldsymbol{\mu}, N) = \binom{N}{m_1 m_2 \dots m_K} \prod_{k=1}^K \mu_k^{m_k}$$

$$\mathbb{E}[m_k] = N\mu_k$$

$$\text{var}[m_k] = N\mu_k(1 - \mu_k)$$

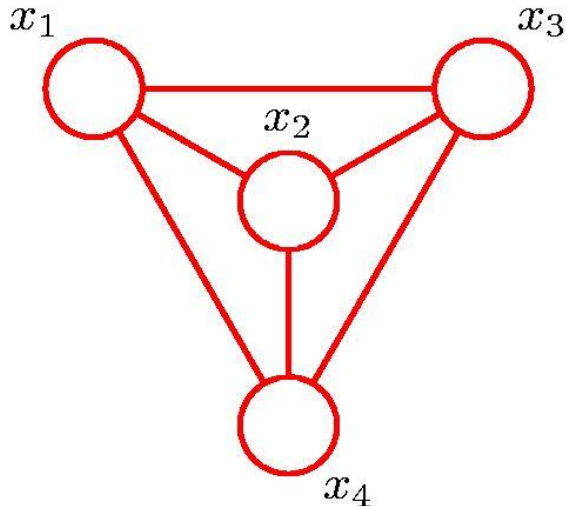
$$\text{cov}[m_j, m_k] = -N\mu_j\mu_k$$

The normalization coefficient is the number of ways of partitioning  $N$  objects into  $K$  groups of size  $m_1, m_2, \dots, m_K$ .

Note that  $\sum_k m_k = N$ .

# Graphical Models

Recall: **graph** contains a set of nodes connected by edges.



In a **probabilistic graphical model**, each node represents a random variable, links represent “probabilistic dependencies” between random variables.

Graph specifies how joint distribution over all random variables **decomposes** into a **product** of factors, each factor depending on a subset of the variables.

Two types of graphical models:

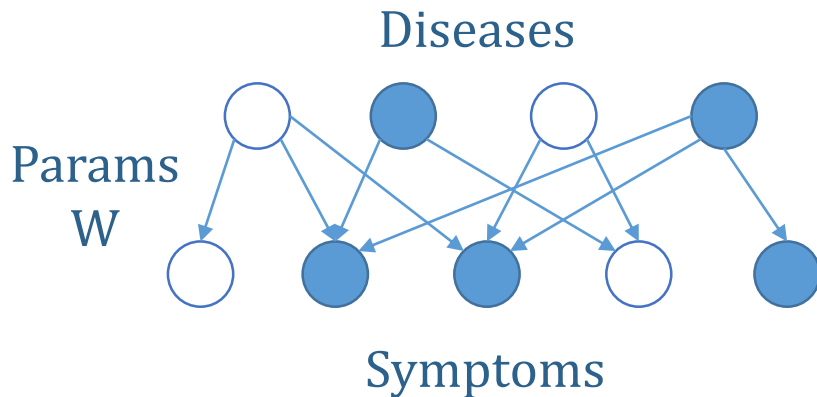
- **Bayesian networks**, also known as **Directed Graphical Models** (the links have a particular directionality indicated by the arrows)
- **Markov Random Fields**, also known as **Undirected Graphical Models** (the links do not carry arrows and have no directional significance).

# Bayesian Networks

**Directed Graphs** are useful for expressing causal relationships between random variables.

Your **symptoms**: fever + red spots.

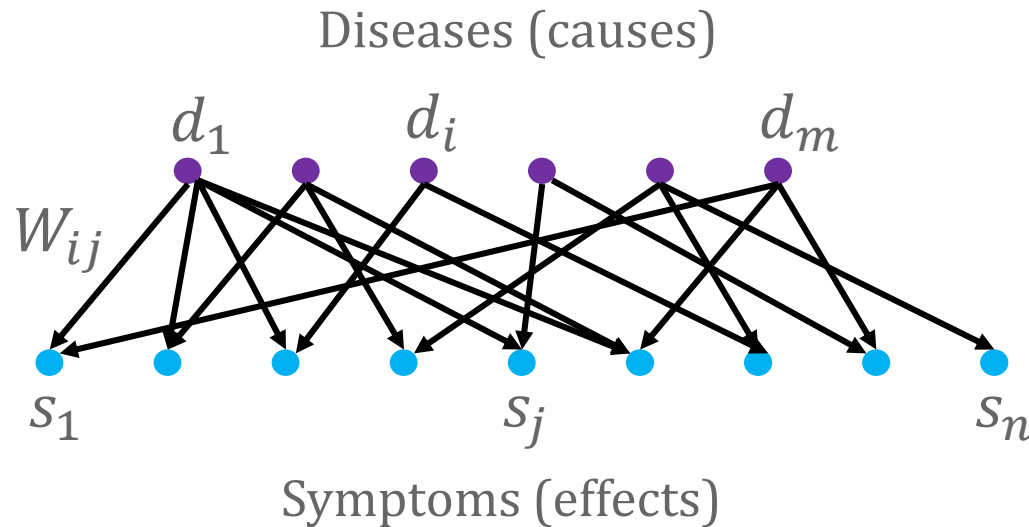
**Probability** that you have measles?



**Bayesian network** succinctly describes  $\Pr[\text{symptom} | \text{diseases}]$

# Noisy-OR networks

$$d_i, s_j \in \{0,1\}$$
$$W_{ij} \geq 0$$



- Each  $d_i$  is on **independently** with prob.  $\rho$
- When  $d_i$  is on, it **activates**  $s_j$  with probability  $1 - \exp(-W_{ij})$ .
- $s_j$  is **on** if one of  $d_i$ 's **activates**  $s_j$



# Bayesian Networks

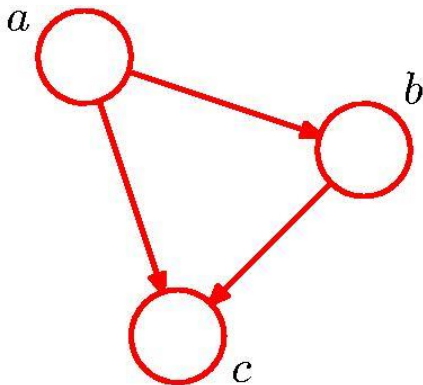
**Directed Graphs** are useful for expressing causal relationships between random variables.

“Deriving” Bayesian Networks as restrictions of arbitrary distributions:

An **arbitrary** joint distribution  $p(a, b, c)$  over three random variables  $a, b$ , and  $c$  can be written as

$$p(a, b, c) = p(c|a, b)p(a, b) = p(c|a, b)p(b|a)p(a)$$

Associate a **graph** with the decomposition:



- Node for each of the random variables.
- Add **directed** links to the graph from the nodes corresponding to the vars on which the distribution is conditioned.

# Bayesian Networks

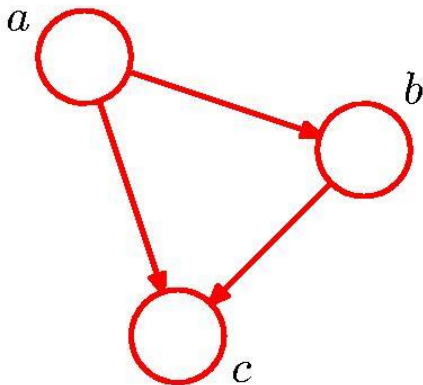
**Directed Graphs** are useful for expressing causal relationships between random variables.

“Deriving” Bayesian Networks as restrictions of arbitrary distributions:

An **arbitrary** joint distribution  $p(a, b, c)$  over three random variables  $a, b$ , and  $c$  can be written as

$$p(a, b, c) = p(c|a, b)p(a, b) = p(c|a, b)p(b|a)p(a)$$

Associate a **graph** with the decomposition:



Different ordering => different graphical representation.

Joint distribution over  $K$  variables factorizes:

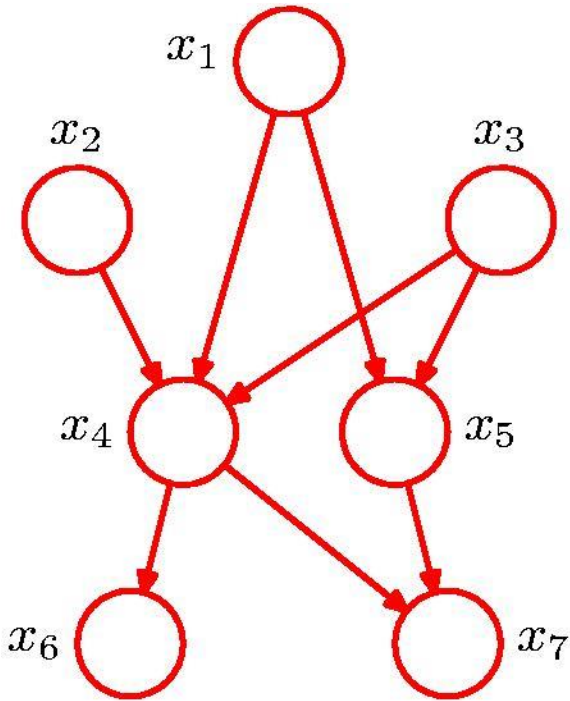
$$p(x_1, \dots, x_K) = p(x_K|x_1, \dots, x_{K-1}) \dots p(x_2|x_1)p(x_1)$$

Corresponding undirected graph is fully connected:

(as each lower-numbered node points to each higher-numbered node)

# Bayesian Networks

A graph that is **not** fully connected conveys information about the conditional **independence** structure of the distribution it encodes.



E.g. consider the graph on the left.

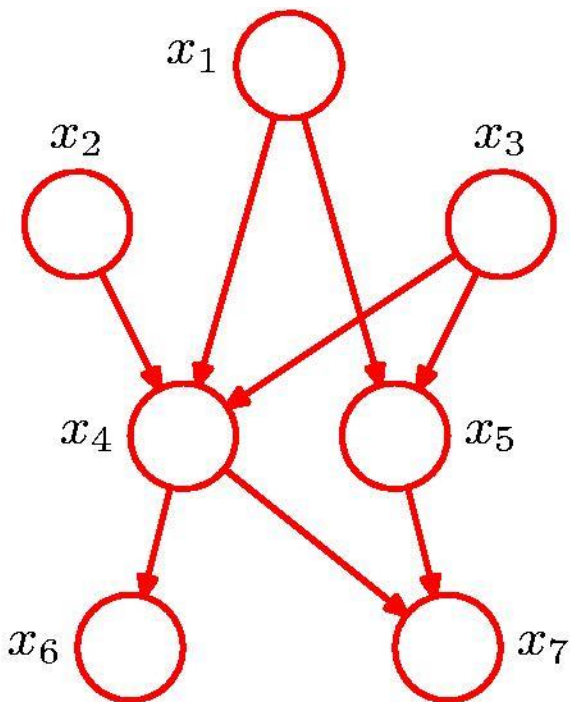
It encodes distributions over  $x_1, \dots, x_7$  that can be written as the product:

$$p(x_1, \dots, x_7) = p(x_1)p(x_2)p(x_3)p(x_4|x_1, x_2, x_3) \\ p(x_5|x_1, x_3)p(x_6|x_4)p(x_7|x_4, x_5)$$

Note the change from the previous slide: e.g.  $x_5$  is **not** conditioned on all of  $x_1, x_2, x_3, x_4$  but only on  $x_1, x_3$ .

# The general case: factorization

The joint distribution defined by the graph is given by the product of a conditional distribution for each node **conditioned on its parents**:



$$p(\mathbf{x}) = \prod_{k=1}^K p(x_k | \text{pa}_k)$$

where  $\text{pa}_k$  denotes a set of parents for the node  $x_k$ .

Each of the conditional distributions will typically have some parametric form. (e.g. product of Bernoullis in the noisy-OR case)

Important restriction: There must be **no directed cycles!** (i.e. graph is a DAG)

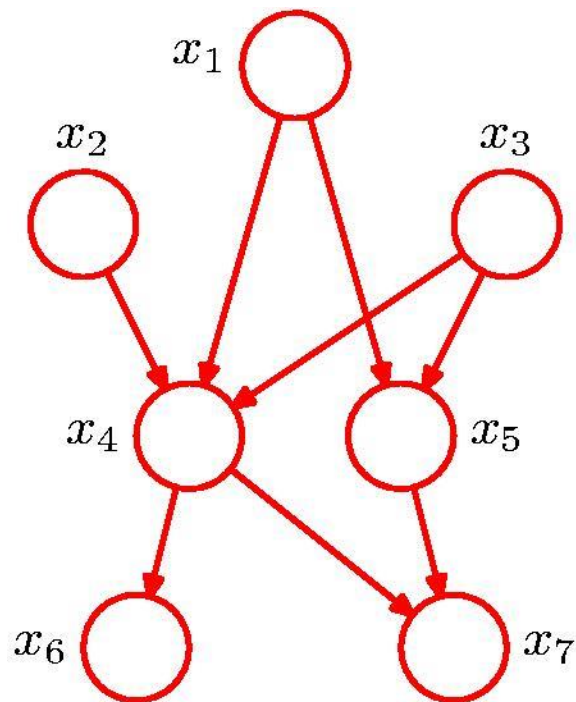
# Crucial property: easy sampling

Consider a joint distribution over  $K$  random variables  $p(x_1, x_2, \dots, x_K)$  that factorizes as:

$$p(\mathbf{x}) = \prod_{k=1}^K p(x_k | \text{pa}_k)$$

Suppose each of the conditional distributions are easy to sample from. How do we sample from the joint?

Start at the top and sample in order.



$$\hat{x}_1 \sim p(x_1)$$

$$\hat{x}_2 \sim p(x_2)$$

$$\hat{x}_3 \sim p(x_3)$$

$$\hat{x}_4 \sim p(x_4 | \hat{x}_1, \hat{x}_2, \hat{x}_3)$$

$$\hat{x}_5 \sim p(x_5 | \hat{x}_1, \hat{x}_3)$$

The parent variables are set to their sampled values



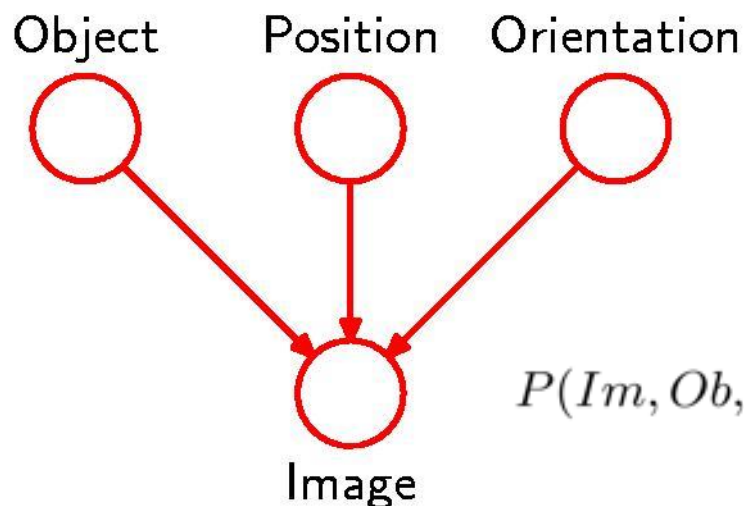
To obtain a sample from the marginal distribution, e.g.  $p(x_2, x_5)$ , sample from the full joint distribution, retain  $\hat{x}_2, \hat{x}_5$ , discard the remaining values.

# Typical deep learning application

Higher-up nodes will typically represent **latent** (hidden) random variables.

The role of latent variables is to allow modeling a **complicated** distribution over observed variables **constructed** from **simpler** conditional distributions.

*Latent-variable model of image*



Object identity, position, and orientation have independent *prior probabilities*.

Image has probability distr that depends on object identity, position, and orientation (*conditional distribution/likelihood*).

$$P(Im, Ob, Po, Or) = \underbrace{P(Im|Ob, Po, Or)}_{\text{Likelihood}} \underbrace{P(Ob)P(Po)P(Or)}_{\text{Prior}}$$

Likelihood and prior are modeled by parametric distribution whose parameters are fitted throughout training.

# Why restrict connectivity?

Why would we not want fully connected graphs?

Restricts the richness of the class!!

Consider discrete joint distribution over  $n$  variables, where each variable takes one of  $k$  values.

To **fully** specify it in general, we need the values of probabilities of every possibly outcome – so we need to specify  $k^n$  values.

To specify the conditionals  $p(x_1|x_2, x_3, \dots x_d)$  we need to specify  $k^d$  values.

Hence, in a graph of in-degree at most  $d$ , we need to specify at most  $n k^d \ll k^n$  values!!