

STA 414/2104

Lecture 2, 15 January 2018

With thanks to Russ Salakhutdinov

The exponential family, maximum likelihood,
and optimization

Outline

- Seven distributions and their ML estimates
 - Bernoulli, binomial, and multinomial discrete
 - Beta and Dirichlet multivariate extension of beta
 - Normal and Student's t
- Mixture of Gaussians
- The Exponential Family and its ML estimates maximum likelihood



The 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 μ , not p as you may have encountered in earlier courses, so that:

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

- The probability distribution, known as the **Bernoulli distribution**, can be written as:

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

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

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

Parameter Estimation

- Suppose we observed a dataset $\mathcal{D} = \{x_1, \dots, x_N\}$


We can construct the likelihood function, which is a function of μ .

$$p(\mathcal{D}|\mu) = \prod_{n=1}^N p(x_n|\mu) = \prod_{n=1}^N \mu^{x_n} (1 - \mu)^{1-x_n}$$

- Equivalently, we can consider the log of the likelihood function:

$$\ln p(\mathcal{D}|\mu) = \sum_{n=1}^N \ln p(x_n|\mu) = \sum_{n=1}^N \{x_n \ln \mu + (1 - x_n) \ln(1 - \mu)\}$$

- Note that the likelihood function depends on the N observations x_n only

through the sum $\sum_n x_n$  Sufficient Statistic

order is not relevant

Parameter Estimation

- Suppose we observed a dataset $\mathcal{D} = \{x_1, \dots, x_N\}$

$$\ln p(\mathcal{D}|\mu) = \sum_{n=1}^N \ln p(x_n|\mu) = \sum_{n=1}^N \{x_n \ln \mu + (1 - x_n) \ln(1 - \mu)\}$$

Let's find the μ which maximizes the likelihood function. Setting the derivative of the log-likelihood function w.r.t μ to zero, we obtain:

$$\mu_{\text{ML}} = \frac{1}{N} \sum_{n=1}^N x_n = \frac{m}{N}$$

where m is the number of heads. If N is small, then problematic

This is a simplistic example of **parameter estimation**.

The Binomial Distribution

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

The probability of observing m heads given N coin flips and a parameter μ is given by the **binomial distribution**:

$$p(m \text{ heads} | N, \mu) =$$

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

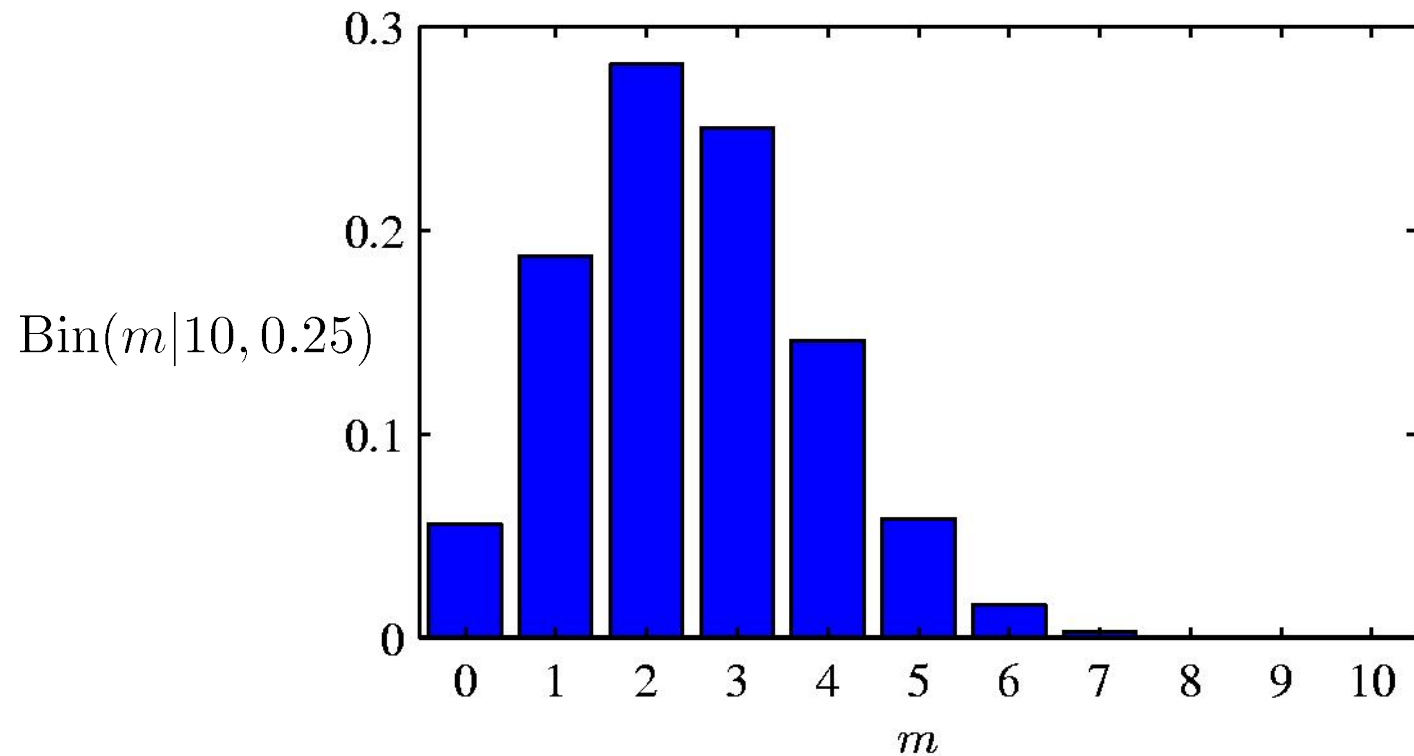
- Review exercise: show that the mean and variance are:

$$\mathbb{E}[m] \equiv \sum_{m=0}^N m \text{Bin}(m | N, \mu) = N\mu$$

$$\text{var}[m] \equiv \sum_{m=0}^N (m - \mathbb{E}[m])^2 \text{Bin}(m | N, \mu) = N\mu(1 - \mu)$$

Example

Below is a histogram plot of the binomial distribution as a function of m for $N=10$ and $\mu = 0.25$.



Beta Distribution

We can define a **distribution** over $\mu \in [0, 1]$ (e.g. it can be used as a prior over the parameter μ of the Bernoulli distribution).

$$\text{Beta}(\mu|a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \mu^{a-1} (1-\mu)^{b-1}$$

$$\mathbb{E}[\mu] = \frac{a}{a+b}$$

$$\text{var}[\mu] = \frac{ab}{(a+b)^2(a+b+1)}$$

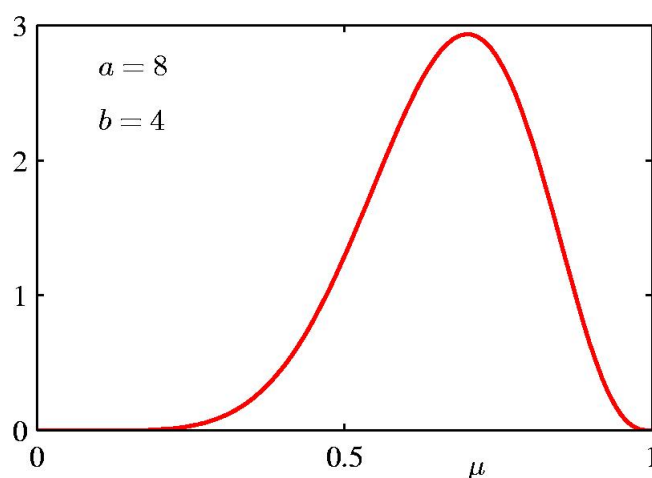
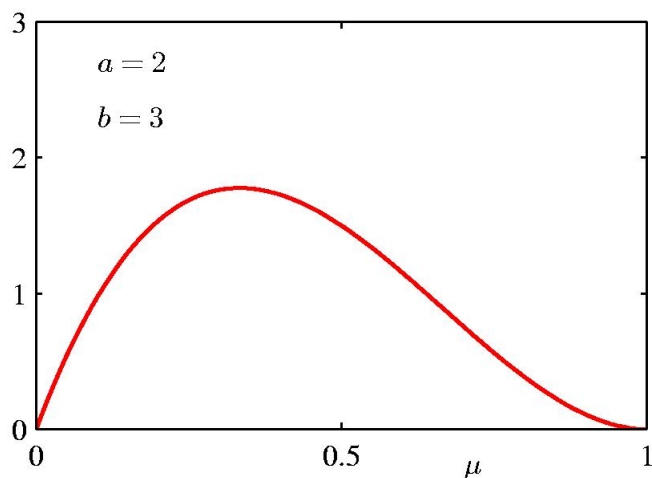
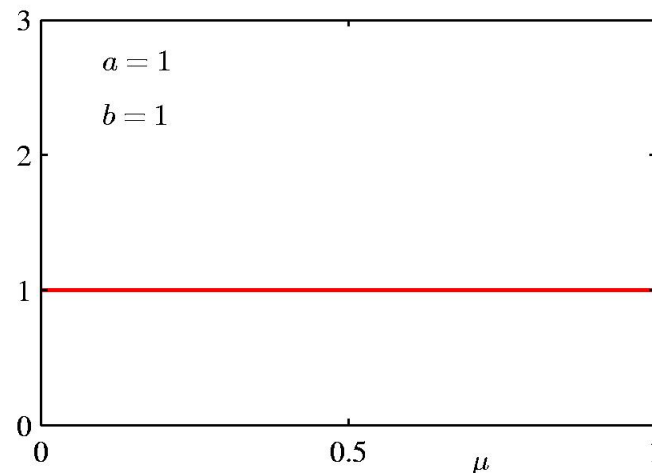
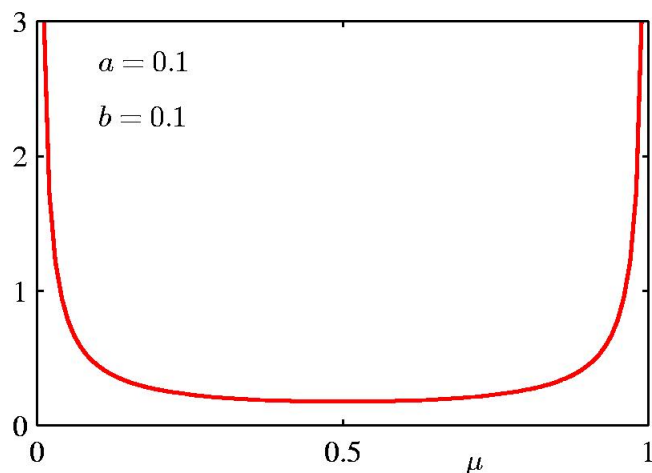
where the gamma function is defined as:

$$\Gamma(x) \equiv \int_0^{\infty} u^{x-1} e^{-u} du.$$

and ensures that the beta distribution is normalized.

Beta Distribution

if $a = b$, symmetry
if $a=1, b=1$, uniform distribution
if $a, b > 0$, hill shape



The beta distribution can be used to specify a posterior distribution over μ .
See for example John Rice's 3rd edition, Section 3.5 Example E; or Bishop 2.1.1

Multinomial Variables

- Consider a random variable that can take on one of K mutually exclusive states (e.g. rolling a die).
- We will use the so-called **1-of- K encoding** scheme. For example, 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:

1-of- K coding scheme: $\mathbf{x} = (0, 0, 1, 0, 0, 0)^T$

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

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

Multinomial Variables

- A multinomial variable can be viewed as a generalization of the Bernoulli trial 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 above probabilities sum to 1:

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

and that

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

Maximum Likelihood Estimation

- Suppose we observed a dataset $\mathcal{D} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$

We can construct the likelihood function, which is a function of μ .

m_k : number of times side k is rolled

$$p(\mathcal{D}|\mu) = \prod_{n=1}^N \prod_{k=1}^K \mu_k^{x_{nk}} = \prod_{k=1}^K \mu_k^{(\sum_n x_{nk})} = \prod_{k=1}^K \mu_k^{m_k}$$

- Note that the likelihood function depends on the N data points only though the following K quantities:

$$m_k = \sum_n x_{nk}, \quad k = 1, \dots, K.$$

where each represents the number of observations of $x_k=1$.

- These m_k are the sufficient statistics for this distribution.

Maximum Likelihood Estimation

$$p(\mathcal{D}|\boldsymbol{\mu}) = \prod_{n=1}^N \prod_{k=1}^K \mu_k^{x_{nk}} = \prod_{k=1}^K \mu_k^{(\sum_n x_{nk})} = \prod_{k=1}^K \mu_k^{m_k}$$

To find a maximum likelihood estimate for μ , we maximize the log-likelihood taking into account the constraint that $\sum_k \mu_k = 1$

- Create a Lagrangian function:

$$\sum_{k=1}^K m_k \ln \mu_k + \lambda \left(\sum_{k=1}^K \mu_k - 1 \right)$$

which leads to:

$$\mu_k = -m_k / \lambda \quad \underline{\mu_k^{\text{ML}} = \frac{m_k}{N}} \quad \lambda = -N$$

and thus we estimate μ_k as the fraction of observations for which $x_k=1$.
Is this simplistic?

The Multinomial Distribution

We can construct the joint distribution of the quantities $\{m_1, m_2, \dots, m_K\}$ given the parameters $\boldsymbol{\mu}$ and the total number N of observations:

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

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

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

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

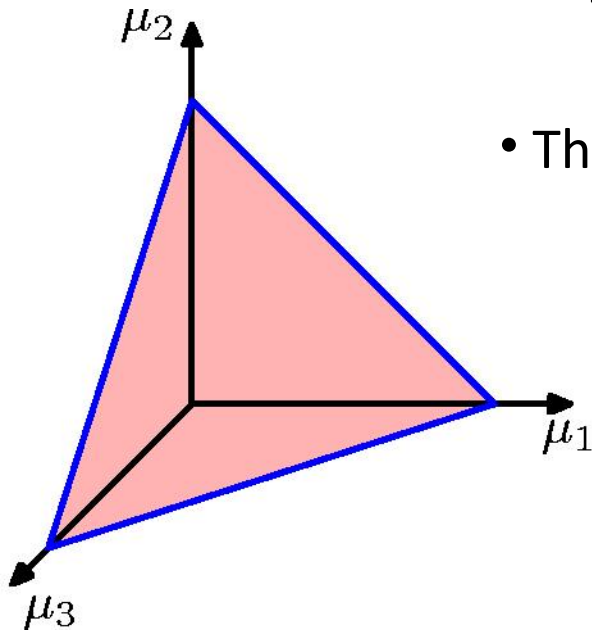
- The normalization coefficient is the number of ways of partitioning N objects into K groups with m_k objects in the k th group
- This is known as the **multinomial distribution**. Note that

$$\sum_k m_k = N.$$

The Dirichlet Distribution

Consider a distribution over μ_k – subject to constraints:

$$\forall k : \mu_k \geq 0 \quad \text{and} \quad \sum_{k=1}^K \mu_k = 1$$



- The **Dirichlet distribution** is defined as:

$$\text{Dir}(\mu|\alpha) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1) \cdots \Gamma(\alpha_K)} \prod_{k=1}^K \mu_k^{\alpha_k - 1}$$

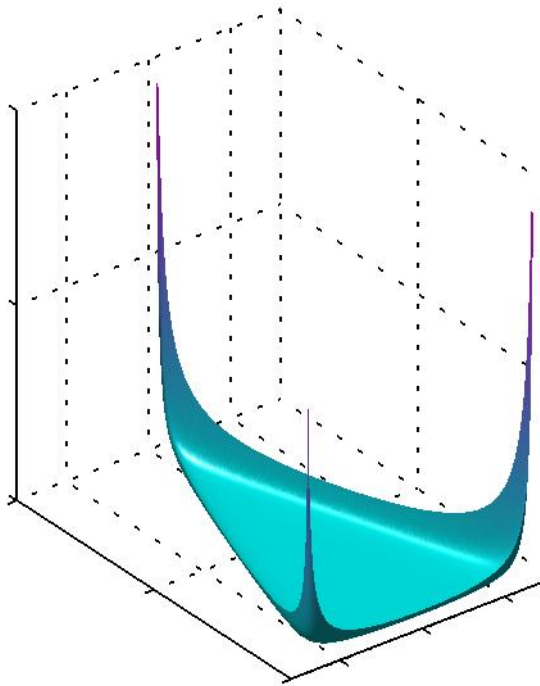
$$\alpha_0 = \sum_{k=1}^K \alpha_k$$

where $\alpha_1, \dots, \alpha_k$ are the parameters of the distribution, and $\Gamma(\cdot)$ is the gamma function.

- The Dirichlet distribution is confined to a **simplex** (the generalization of a triangle to arbitrary dimension) as a consequence of the constraints.

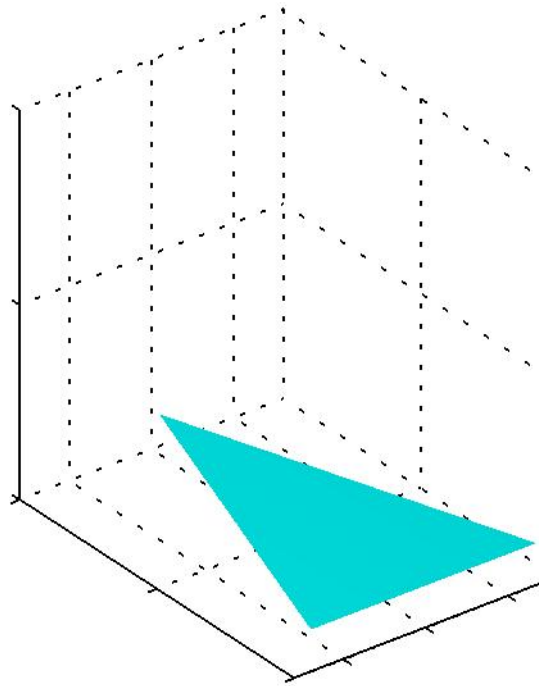
The Dirichlet Distribution

- Plots of the Dirichlet distribution over three variables.



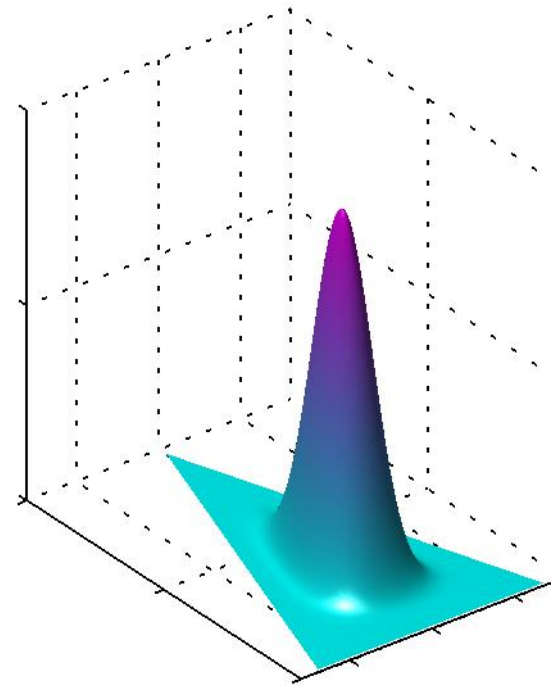
$$\alpha_k = 10^{-1}$$

similar to beta distribution plot



$$\alpha_k = 10^0$$

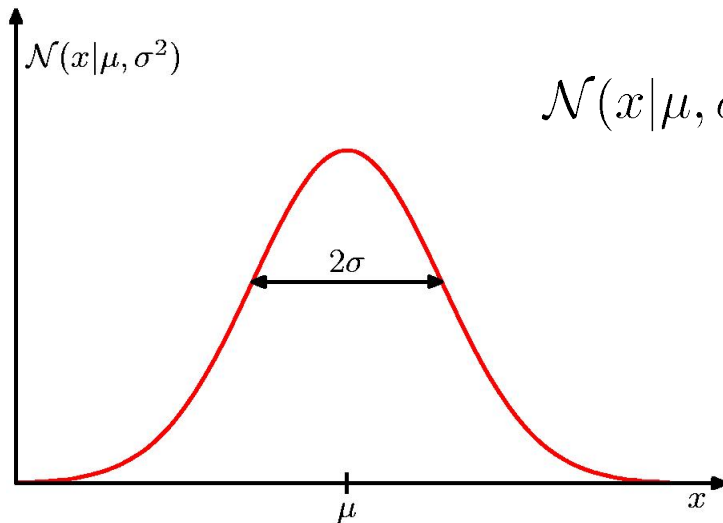
uniform distribution



$$\alpha_k = 10^1$$

Gaussian Univariate Distribution

- In the case of a single variable x , the Gaussian distribution takes the form



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

which is governed by two parameters:

- μ (mean)
- σ^2 (variance)

- The Gaussian distribution satisfies:

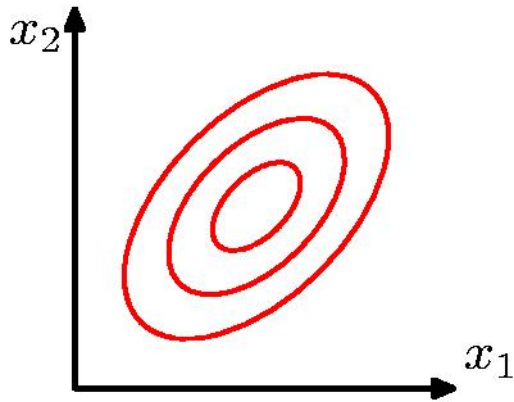
$$\mathcal{N}(x|\mu, \sigma^2) > 0$$

$$\int_{-\infty}^{\infty} \mathcal{N}(x|\mu, \sigma^2) dx = 1$$

Multivariate Gaussian Distribution

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

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



which is governed by two parameters:

- $\boldsymbol{\mu}$ is a D -dimensional mean vector
- $\boldsymbol{\Sigma}$ is a D -by- D covariance matrix

and $|\boldsymbol{\Sigma}|$ denotes the determinant of $\boldsymbol{\Sigma}$.

$$\mathbf{Z}^T \boldsymbol{\Sigma} \mathbf{Z} > 0 \text{ for } \mathbf{Z} \neq \mathbf{0}$$

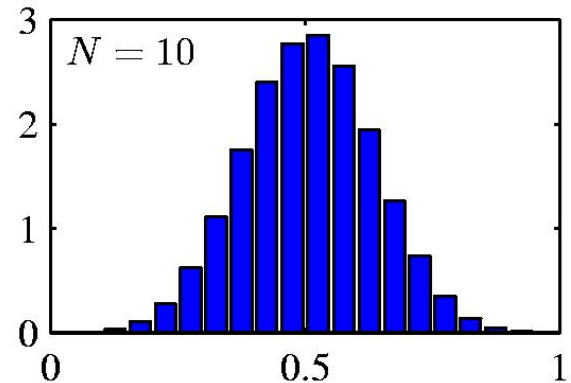
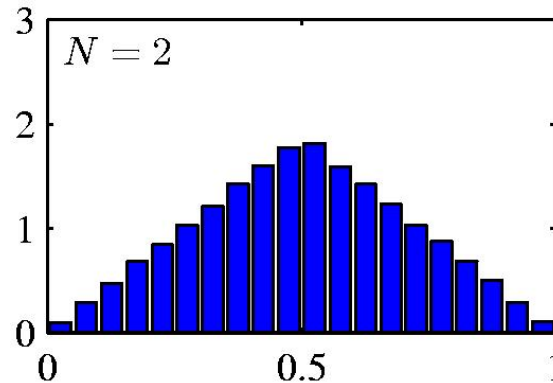
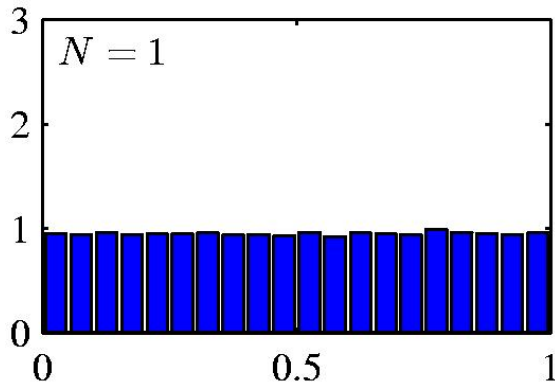
- The covariance matrix is a **positive definite matrix**. (Appendix C of Christopher Bishop's book gives a review of matrices.)

Central Limit Theorem

- The distribution of the sum of N i.i.d. random variables becomes increasingly Gaussian as N grows.
- 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.



Geometry of the Gaussian Distribution

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

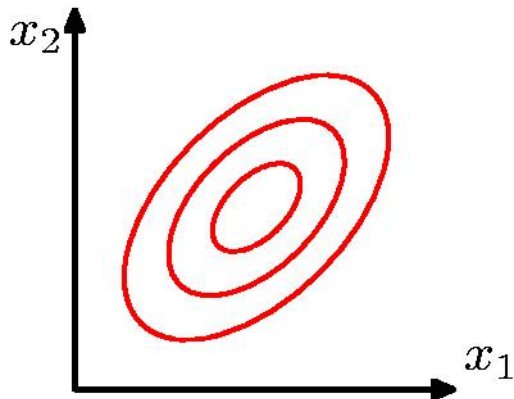
$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp \left\{ -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) \right\}$$

Let us analyse the functional dependence of the Gaussian on \mathbf{x} through the quadratic form:

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

a scalar

- Here Δ is known as the **Mahalanobis distance**. generalization of standard deviation



The Gaussian distribution will be constant on surfaces in \mathbf{x} -space for which Δ is constant.

Geometry of the Gaussian Distribution

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

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

- Consider the eigenvalue equation for the covariance matrix:

$$\boldsymbol{\Sigma} \mathbf{u}_i = \lambda_i \mathbf{u}_i, \quad \text{where } i = 1, \dots, D.$$

- You should be able to show that the covariance can be expressed in terms of its eigenvectors:

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

- The inverse of the covariance is:

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

Geometry of the Gaussian Distribution

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

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

- Remember:

$$\Delta^2 = (\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) \quad \boldsymbol{\Sigma}^{-1} = \sum_{i=1}^D \frac{1}{\lambda_i} \mathbf{u}_i \mathbf{u}_i^T$$

- Hence:

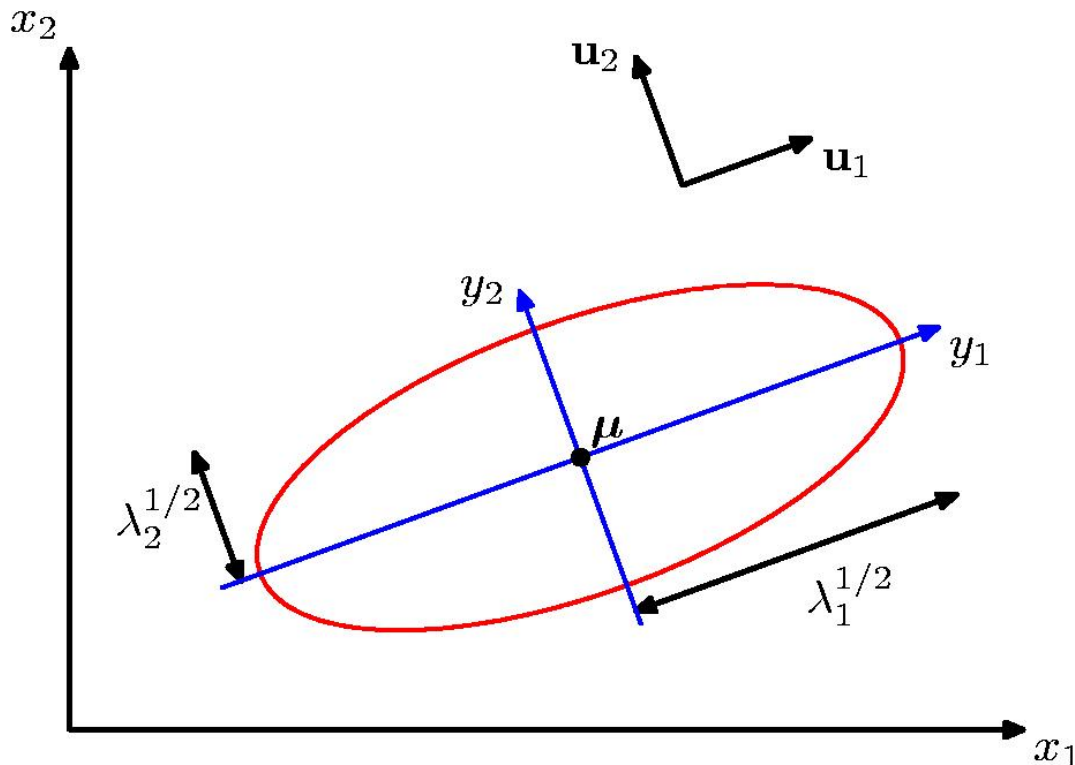
$$\Delta^2 = \sum_{i=1}^D \frac{y_i^2}{\lambda_i} \quad y_i = \mathbf{u}_i^T (\mathbf{x} - \boldsymbol{\mu})$$

- We can interpret $\{y_i\}$ as a new coordinate system defined by the orthonormal vectors \mathbf{u}_i that are shifted and rotated.

Geometry of the Gaussian Distribution

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

$$\Delta^2 = \sum_{i=1}^D \frac{y_i^2}{\lambda_i} \quad y_i = \mathbf{u}_i^T (\mathbf{x} - \boldsymbol{\mu})$$



- Red curve: surface of constant probability density
- The axes are defined by the eigenvectors \mathbf{u}_i of the covariance matrix with corresponding eigenvalues

Moments of the Gaussian Distribution

- The expectation of \mathbf{x} under the Gaussian distribution is:

$$\begin{aligned}\mathbb{E}[\mathbf{x}] &= \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \int \exp \left\{ -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) \right\} \mathbf{x} \, d\mathbf{x} \\ &= \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \int \underbrace{\exp \left\{ -\frac{1}{2}\mathbf{z}^T \boldsymbol{\Sigma}^{-1}\mathbf{z} \right\}}_{\text{cancels z out since odd/even function}} (\cancel{\mathbf{z}} + \boldsymbol{\mu}) \, d\mathbf{z}\end{aligned}$$

This is an even function of \mathbf{z} , so this multiplied by \mathbf{z} will vanish by symmetry.

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


Moments of the Gaussian Distribution

- Additional algebra leads to the second-order raw moments of the Gaussian distribution:

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

- The second-order central moments are given by the covariance matrix:

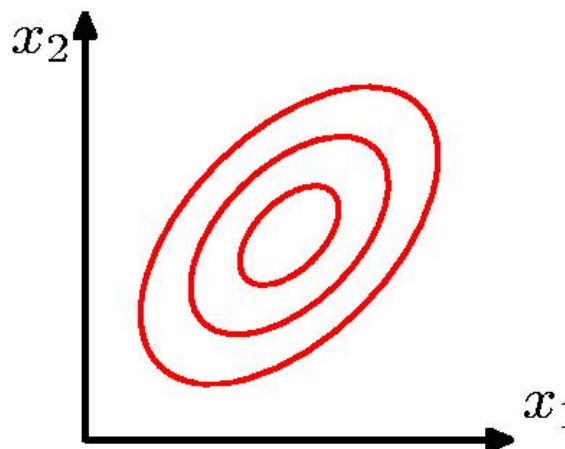
$$\text{cov}[\mathbf{x}] = \mathbb{E} [(\mathbf{x} - \mathbb{E}[\mathbf{x}])(\mathbf{x} - \mathbb{E}[\mathbf{x}])^T] = \boldsymbol{\Sigma}$$


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

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

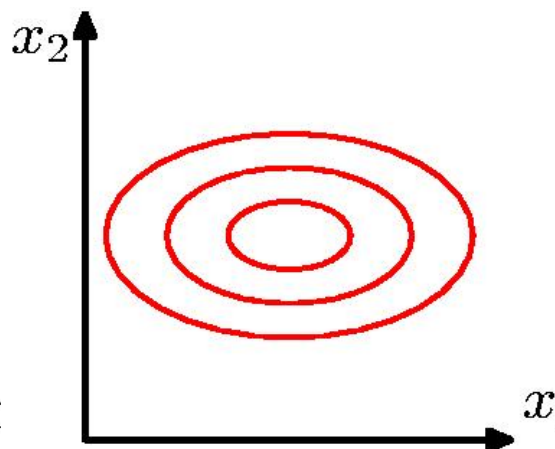
Moments of the Gaussian Distribution

- Contours of constant probability density:



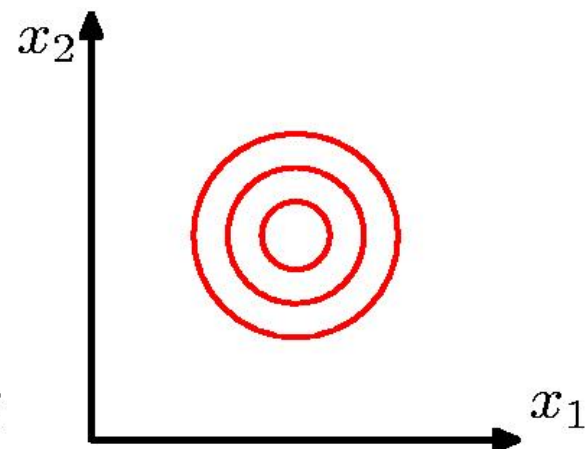
(a)

Covariance matrix is of general form.



(b)

Diagonal, axis-aligned covariance matrix



(c)

Spherical covariance matrix (proportional to identity matrix)

Partitioned Gaussian Distribution

- Consider a D -dimensional Gaussian distribution: $p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$
- Let us partition a datum \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 $\boldsymbol{\Sigma}_{aa}$. doesn't correspond

Conditional Distribution

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

$$p(\mathbf{x}_a | \mathbf{x}_b) = \mathcal{N}(\mathbf{x}_a | \boldsymbol{\mu}_{a|b}, \boldsymbol{\Sigma}_{a|b})$$

Covariance does not depend on \mathbf{x}_b



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



linear gaussian

Linear function of \mathbf{x}_b

Marginal Distribution

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

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

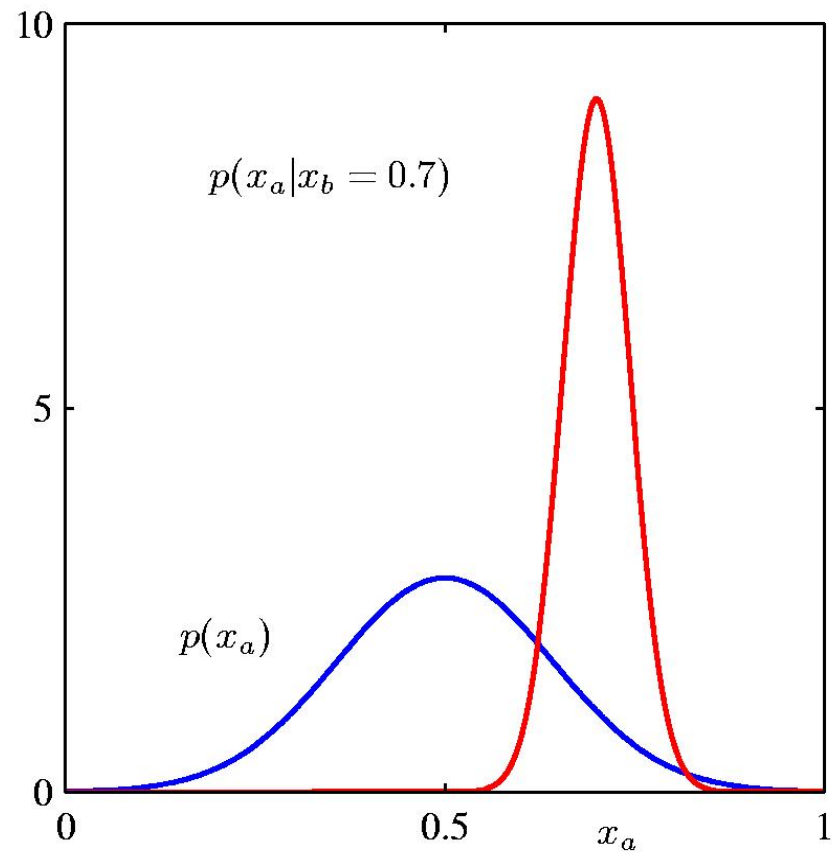
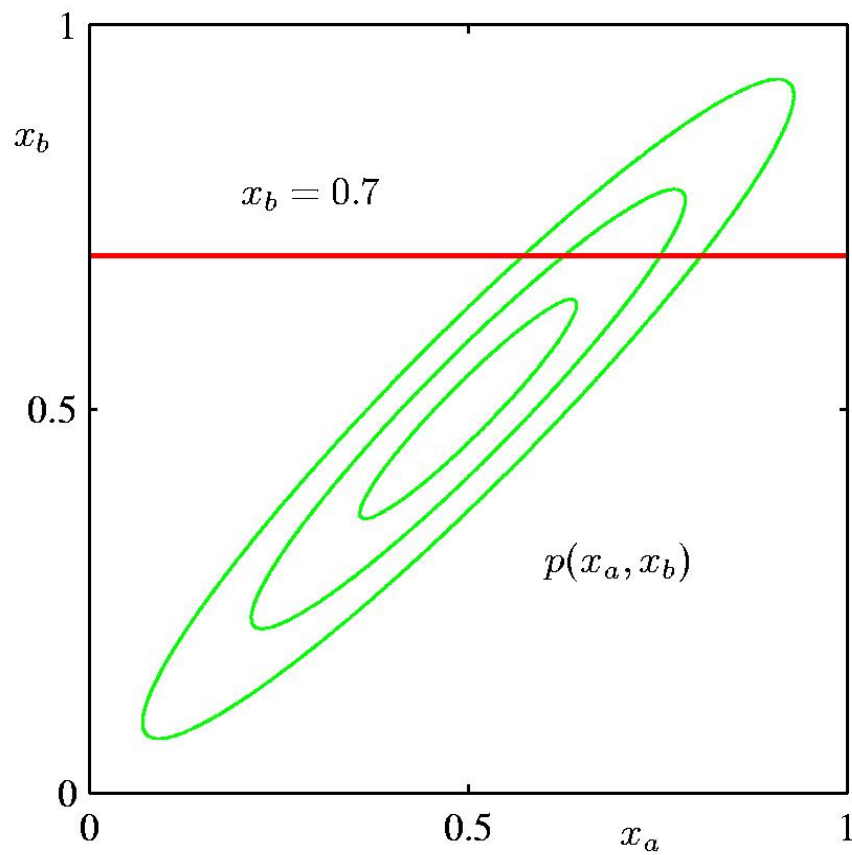
- For a marginal distribution, the mean and covariance are most simply expressed in terms of partitioned covariance matrix.

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

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

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

Conditional and Marginal Distributions



Maximum Likelihood Estimation

- Suppose we observed i.i.d data $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$.

We can construct the log-likelihood function, which is a function of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$:

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

- Note that the likelihood function depends on the N data points only through the following sums:

$$\sum_{n=1}^N \mathbf{x}_n \qquad \sum_{n=1}^N \mathbf{x}_n \mathbf{x}_n^T$$

Sufficient Statistics

Maximum Likelihood Estimation

- To find a maximum likelihood (ML) estimate of the mean, we set the derivative of the log-likelihood function to zero:

$$\frac{\partial}{\partial \boldsymbol{\mu}} \ln p(\mathbf{X} | \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^N \boldsymbol{\Sigma}^{-1} (\mathbf{x}_n - \boldsymbol{\mu}) = 0$$

and solve to obtain:

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

More difficult to find is the ML estimate of $\boldsymbol{\Sigma}$ which is:

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

Maximum Likelihood Estimation

- Evaluating the expectation of the ML estimates under the true distribution, we obtain:

$$\begin{aligned}\mathbb{E}[\boldsymbol{\mu}_{\text{ML}}] &= \boldsymbol{\mu} && \swarrow \text{Unbiased estimate} \\ \mathbb{E}[\boldsymbol{\Sigma}_{\text{ML}}] &= \frac{N-1}{N} \boldsymbol{\Sigma}. && \swarrow \text{Biased estimate}\end{aligned}$$

- Note that the maximum likelihood estimate of $\boldsymbol{\Sigma}$ is biased.
- We can correct the bias by defining a different estimator:

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

Sequential Estimation

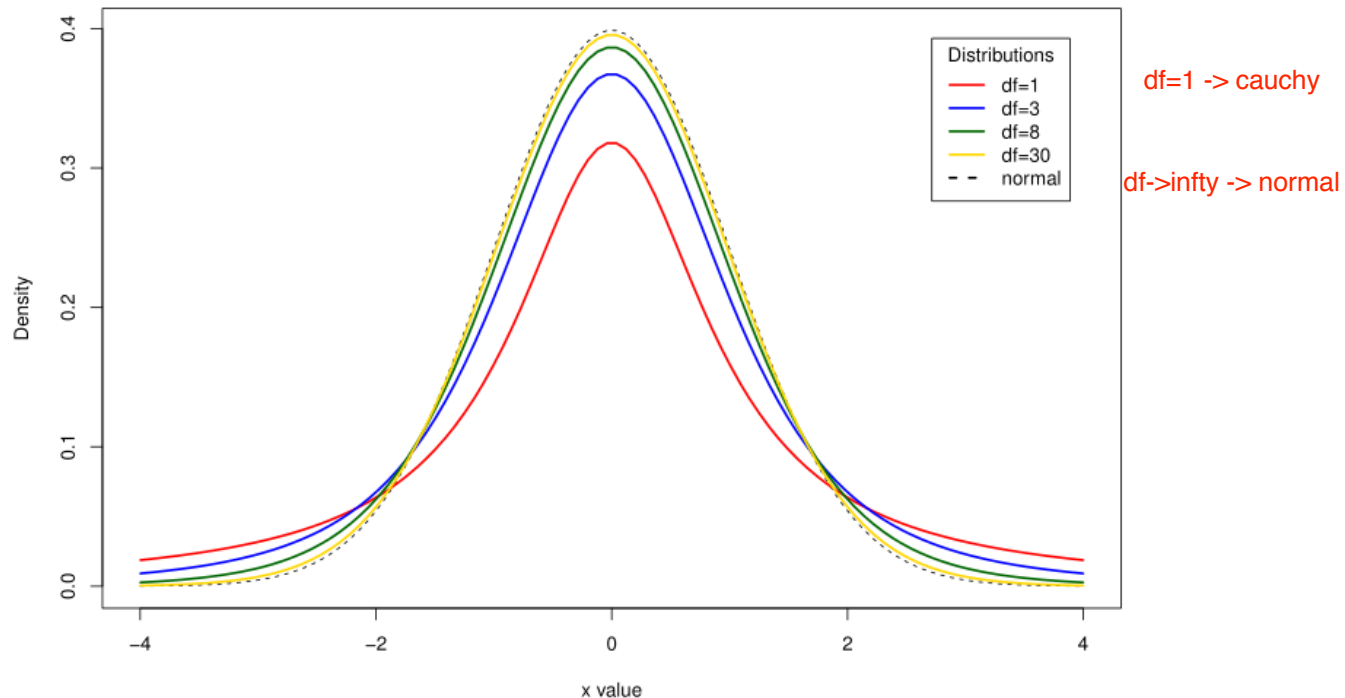
- Sequential estimation allows data points to be processed one at a time and then discarded. This is important for *online algorithms*.
- Let's consider the contribution of the N^{th} data point \mathbf{x}_n :

$$\begin{aligned}\mu_{\text{ML}}^{(N)} &= \frac{1}{N} \sum_{n=1}^N \mathbf{x}_n \\ &= \frac{1}{N} \mathbf{x}_N + \frac{1}{N} \sum_{n=1}^{N-1} \mathbf{x}_n \\ &= \frac{1}{N} \mathbf{x}_N + \frac{N-1}{N} \mu_{\text{ML}}^{(N-1)} \\ &= \underbrace{\mu_{\text{ML}}^{(N-1)}}_{\text{old estimate}} + \underbrace{\frac{1}{N} (\mathbf{x}_N - \mu_{\text{ML}}^{(N-1)})}_{\text{correction given } \mathbf{x}_N \text{ with correction weight } \frac{1}{N}}\end{aligned}$$

The diagram illustrates the sequential estimation process. It shows the equation for the maximum likelihood estimate $\mu_{\text{ML}}^{(N)}$ after N data points. The final line of the equation is broken down into two parts: the old estimate $\mu_{\text{ML}}^{(N-1)}$ and a correction term $\frac{1}{N} (\mathbf{x}_N - \mu_{\text{ML}}^{(N-1)})$. Red arrows point from these terms to labels on the right: 'old estimate' for the first term, and 'correction given \mathbf{x}_N with correction weight $\frac{1}{N}$ ' for the second term.

Student's t -distribution

- You're familiar with the t -distribution (aka Students' t -distribution) as the quotient of a standard normal and a χ^2 distribution



- In Bayesian machine learning, we often generalize the t -distribution. A common parameterization is to consider a Gaussian distribution with known mean (not necessarily zero!) and unknown variance such that the variance has a Gamma prior distribution

Student's t -distribution

- Therefore, Student's t -distribution is

$$\begin{aligned} p(x|\mu, a, b) &= \int_0^\infty \mathcal{N}(x|\mu, \tau^{-1}) \text{Gam}(\tau|a, b) d\tau \\ &= \int_0^\infty \mathcal{N}(x|\mu, (\eta\lambda)^{-1}) \text{Gam}(\eta|\nu/2, \nu/2) d\eta \\ &= \frac{\Gamma(\nu/2 + 1/2)}{\Gamma(\nu/2)} \left(\frac{\lambda}{\pi\nu} \right)^{1/2} \left[1 + \frac{\lambda(x - \mu)^2}{\nu} \right]^{-\nu/2 - 1/2} \\ &= \text{St}(x|\mu, \lambda, \nu) \end{aligned}$$

assume tau's distribution is gamma

tau⁻¹ is called precision

Infinite mixture of Gaussians

where

$$\lambda = a/b$$

$$\eta = \tau b/a$$

$$\nu = 2a.$$



Sometimes called the precision parameter



Degrees of freedom

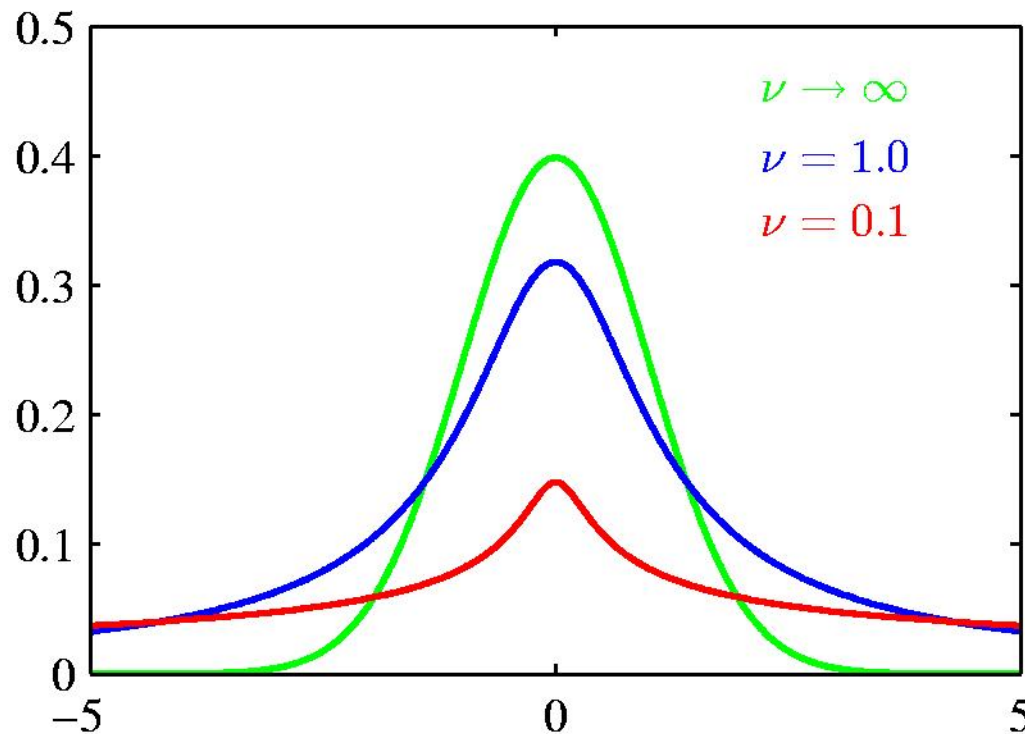
Student's t -distribution

- Setting $\nu = 1$ recovers the Cauchy distribution

The limit $\nu \rightarrow \infty$ corresponds to a Gaussian distribution

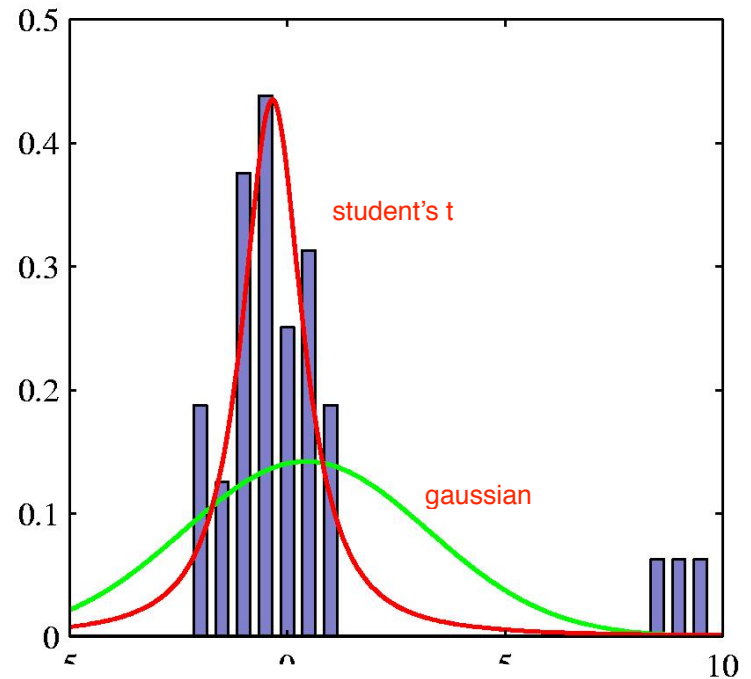
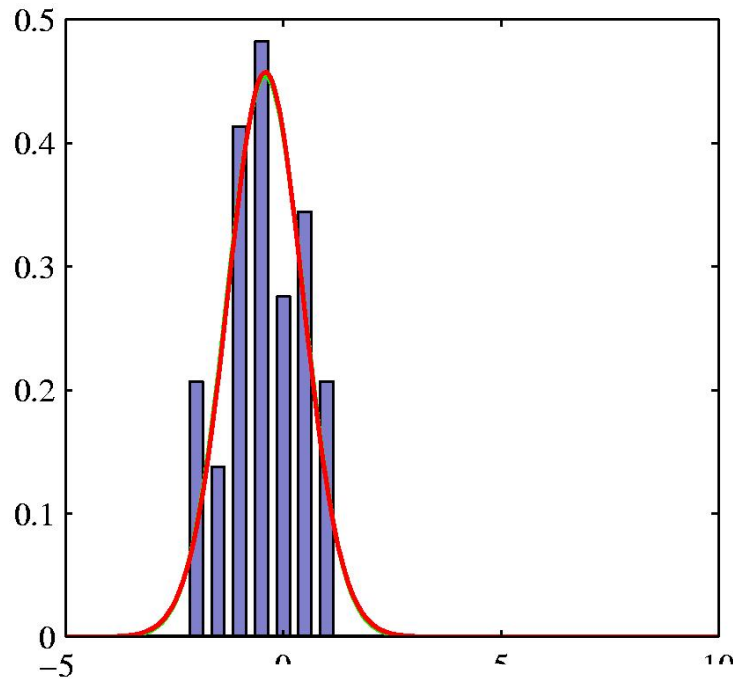
	$\nu = 1$	$\nu \rightarrow \infty$
$\text{St}(x \mu, \lambda, \nu)$	Cauchy	$\mathcal{N}(x \mu, \lambda^{-1})$

cauchy: more relaxed about the outliers



Student's t -distribution

- Robustness to outliers: Gaussian vs. t -distribution.



Student's t -distribution

- The multivariate extension of the t -distribution:

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

where $\Delta^2 = (\mathbf{x} - \boldsymbol{\mu})^\mathrm{T} \boldsymbol{\Lambda} (\mathbf{x} - \boldsymbol{\mu})$

- Properties:

$$\mathbb{E}[\mathbf{x}] = \boldsymbol{\mu}, \quad \text{if } \nu > 1$$

$$\text{cov}[\mathbf{x}] = \frac{\nu}{(\nu - 2)} \boldsymbol{\Lambda}^{-1}, \quad \text{if } \nu > 2$$

$$\text{mode}[\mathbf{x}] = \boldsymbol{\mu}$$

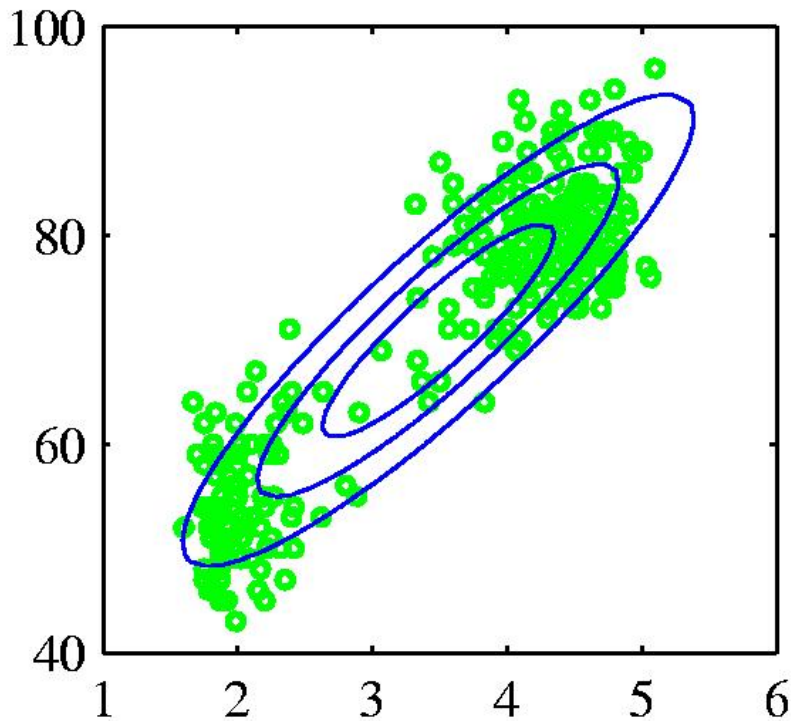
Outline

- Seven distributions and their ML estimates
 - Bernoulli, binomial, and multinomial
 - Beta and Dirichlet
 - Normal and Student's t
- **Mixture of Gaussians**
- The Exponential Family and its ML estimates

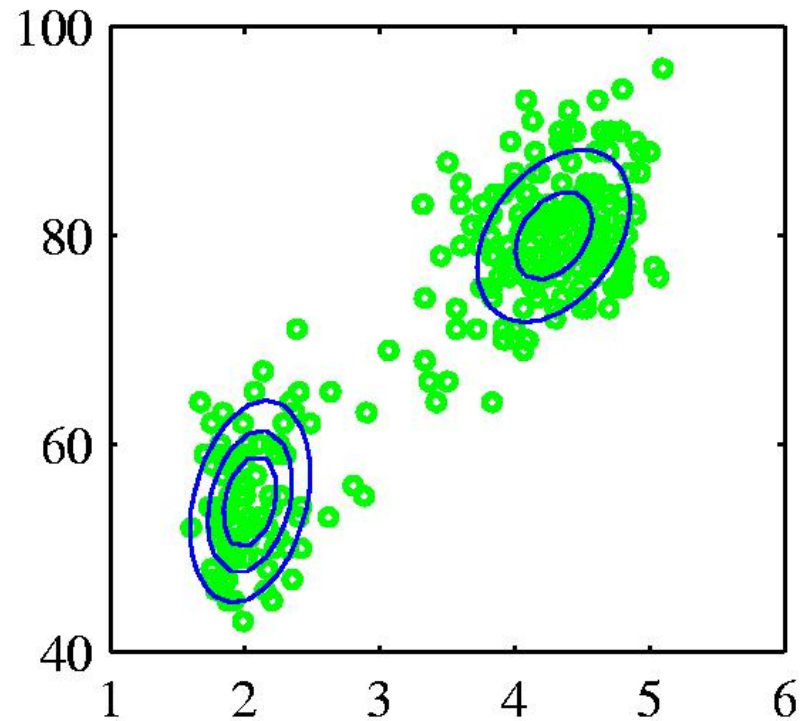


Mixture of Gaussians

- When modelling real-world data, assuming a normal distribution may not be appropriate
- Consider the following example: the Old Faithful dataset



Single Gaussian



Mixture of two Gaussians

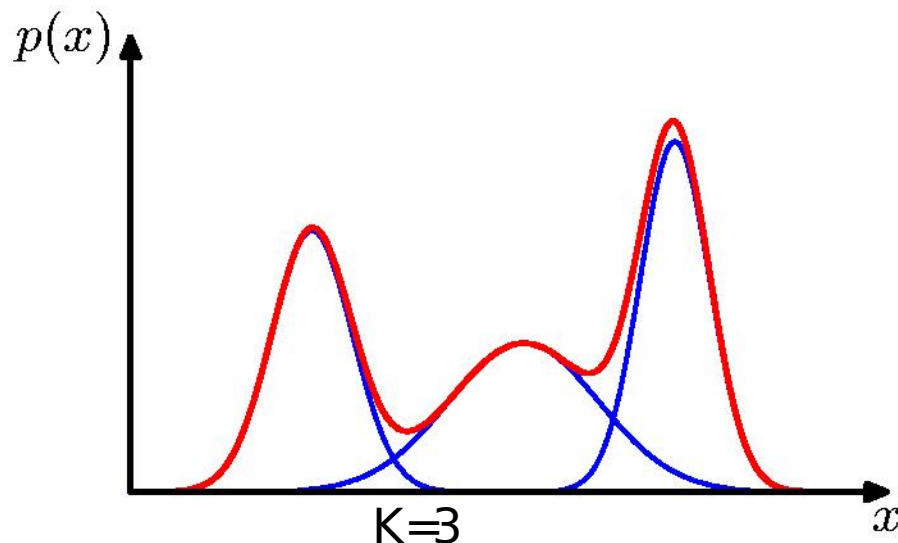
Mixture of Gaussians

- 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

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

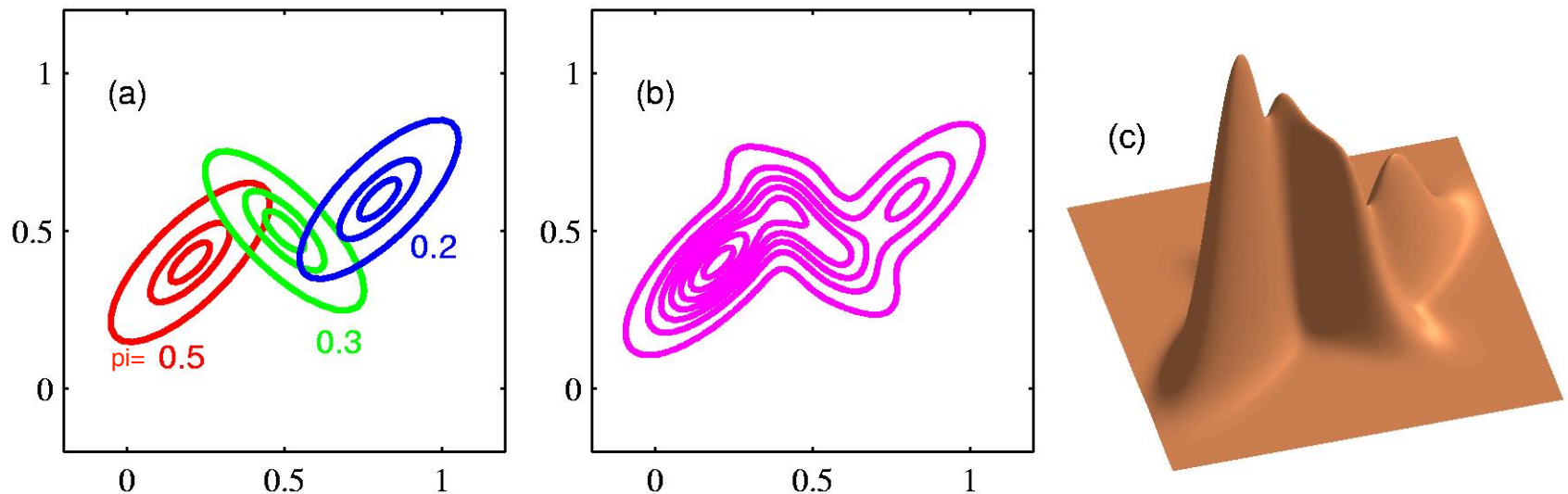


Note that each Gaussian component has its own mean $\boldsymbol{\mu}_k$ and covariance $\boldsymbol{\Sigma}_k$. The parameters π_k are called mixing coefficients.

- More generally, **mixture models** can comprise linear combinations of other distributions.

Mixture of Gaussians

- Illustration of a mixture of three 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 $p(\mathbf{x})$.

Maximum Likelihood Estimation

Given a dataset \mathbf{X} , we can determine model parameters $\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \pi_k$ by maximizing the log-likelihood function:

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


Log of a sum: no closed-form solution

Solutions:

- Use standard, iterative, numerical optimization methods (e.g. conjugate gradients), or
- Use the Expectation Maximization algorithm (to come around March)

Outline

- Seven distributions and their ML estimates
 - Bernoulli, binomial, and multinomial
 - Beta and Dirichlet
 - Normal and Student's t
- Mixture of Gaussians
- **The Exponential Family and its ML estimates**



The Exponential Family

- The exponential family of distributions over \mathbf{x} is defined to be a set of distributions of the form:

$$p(\mathbf{x}|\boldsymbol{\eta}) = h(\mathbf{x})g(\boldsymbol{\eta}) \exp \{ \boldsymbol{\eta}^T \mathbf{u}(\mathbf{x}) \}$$

can separate \mathbf{x} and $\boldsymbol{\eta}$

where

- $\boldsymbol{\eta}$ is the vector of natural parameters
- $\mathbf{u}(\mathbf{x})$ is the vector of sufficient statistics

The function $g(\boldsymbol{\eta})$ can be interpreted as the coefficient that ensures that the distribution $p(\mathbf{x}|\boldsymbol{\eta})$ is normalized:

$$g(\boldsymbol{\eta}) \int h(\mathbf{x}) \exp \{ \boldsymbol{\eta}^T \mathbf{u}(\mathbf{x}) \} d\mathbf{x} = 1$$

Bernoulli Distribution

- The Bernoulli distribution is a member of the exponential family:

$$\begin{aligned} p(x|\mu) &= \text{Bern}(x|\mu) = \mu^x (1 - \mu)^{1-x} \\ &= \exp \{x \ln \mu + (1 - x) \ln(1 - \mu)\} \\ &= (1 - \mu) \exp \left\{ \ln \left(\frac{\mu}{1 - \mu} \right) x \right\} \quad \text{x is sufficient statistics} \end{aligned}$$

- Comparing with the general form of the exponential family:

$$p(\mathbf{x}|\boldsymbol{\eta}) = h(\mathbf{x})g(\boldsymbol{\eta}) \exp \{ \boldsymbol{\eta}^T \mathbf{u}(\mathbf{x}) \}$$

we see that

$$\eta = \ln \left(\frac{\mu}{1 - \mu} \right) \quad \text{and so} \quad \mu = \underbrace{\sigma(\eta)}_{\text{Logistic sigmoid}} = \frac{1}{1 + \exp(-\eta)}.$$

Bernoulli Distribution

- The Bernoulli distribution is a member of the exponential family:

$$\begin{aligned} p(x|\mu) &= \text{Bern}(x|\mu) = \mu^x (1 - \mu)^{1-x} \\ &= \exp \{ x \ln \mu + (1 - x) \ln(1 - \mu) \} \\ &= (1 - \mu) \exp \left\{ \ln \left(\frac{\mu}{1 - \mu} \right) x \right\} \end{aligned}$$

$$p(\mathbf{x}|\boldsymbol{\eta}) = h(\mathbf{x})g(\boldsymbol{\eta}) \exp \{ \boldsymbol{\eta}^T \mathbf{u}(\mathbf{x}) \}$$

- The Bernoulli distribution can therefore be written as:

$$p(x|\eta) = \sigma(-\eta) \exp(\eta x)$$

where

$$u(x) = x$$

$$h(x) = 1$$

$$g(\eta) = 1 - \sigma(\eta) = \sigma(-\eta).$$

Multinomial Distribution

- The multinomial distribution is a member of the exponential family:

$$p(\mathbf{x}|\boldsymbol{\mu}) = \prod_{k=1}^M \mu_k^{x_k} = \exp \left\{ \sum_{k=1}^M x_k \ln \mu_k \right\} = h(\mathbf{x})g(\boldsymbol{\eta}) \exp (\boldsymbol{\eta}^T \mathbf{u}(\mathbf{x}))$$

where $\mathbf{x} = (x_1, \dots, x_M)^T$ $\boldsymbol{\eta} = (\eta_1, \dots, \eta_M)^T$

and

$$\begin{aligned}\eta_k &= \ln \mu_k \\ \mathbf{u}(\mathbf{x}) &= \mathbf{x} \\ h(\mathbf{x}) &= 1 \\ g(\boldsymbol{\eta}) &= 1.\end{aligned}$$

NOTE: The parameters η_k are not independent since the corresponding η_k must

satisfy
$$\sum_{k=1}^M \mu_k = 1.$$

- Sometimes it's convenient to remove the constraint by expressing the distribution over the $M - 1$ parameters; Bishop makes a start in this direction

Gaussian Distribution

- The Gaussian distribution is a member of the exponential family:

$$\begin{aligned} p(x|\mu, \sigma^2) &= \frac{1}{(2\pi\sigma^2)^{1/2}} \exp \left\{ -\frac{1}{2\sigma^2} (x - \mu)^2 \right\} \\ &= \frac{1}{(2\pi\sigma^2)^{1/2}} \exp \left\{ -\frac{1}{2\sigma^2} x^2 + \frac{\mu}{\sigma^2} x - \frac{1}{2\sigma^2} \mu^2 \right\} \\ &= h(x)g(\boldsymbol{\eta}) \exp \{ \boldsymbol{\eta}^T \mathbf{u}(x) \} \end{aligned}$$

where

$$\begin{aligned} \boldsymbol{\eta} &= \begin{pmatrix} \mu/\sigma^2 \\ -1/2\sigma^2 \end{pmatrix} & h(\mathbf{x}) &= (2\pi)^{-1/2} \\ \mathbf{u}(x) &= \begin{pmatrix} x \\ x^2 \end{pmatrix} & g(\boldsymbol{\eta}) &= (-2\eta_2)^{1/2} \exp \left(\frac{\eta_1^2}{4\eta_2} \right). \end{aligned}$$

ML for the Exponential Family

Remember the exponential family: $p(\mathbf{x}|\boldsymbol{\eta}) = h(\mathbf{x})g(\boldsymbol{\eta}) \exp \{ \boldsymbol{\eta}^T \mathbf{u}(\mathbf{x}) \}$

From the definition of the normalizer $g(\boldsymbol{\eta})$:

$$g(\boldsymbol{\eta}) \int h(\mathbf{x}) \exp \{ \boldsymbol{\eta}^T \mathbf{u}(\mathbf{x}) \} d\mathbf{x} = 1$$

We can take a derivative w.r.t $\boldsymbol{\eta}$:

$$\underbrace{\nabla g(\boldsymbol{\eta}) \int h(\mathbf{x}) \exp \{ \boldsymbol{\eta}^T \mathbf{u}(\mathbf{x}) \} d\mathbf{x}}_{1/g(\boldsymbol{\eta})} + g(\boldsymbol{\eta}) \underbrace{\int h(\mathbf{x}) \exp \{ \boldsymbol{\eta}^T \mathbf{u}(\mathbf{x}) \} \mathbf{u}(\mathbf{x}) d\mathbf{x}}_{\mathbb{E}[\mathbf{u}(\mathbf{x})]} = 0$$

• Thus

$$-\nabla \ln g(\boldsymbol{\eta}) = \mathbb{E}[\mathbf{u}(\mathbf{x})]$$

The covariance of $\mathbf{u}(\mathbf{x})$ can be expressed in terms of the second derivative of $g(\boldsymbol{\eta})$, and similarly for the higher moments.

ML for the Exponential Family

- Suppose we observed i.i.d data $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$.

We can construct the log-likelihood function, which is a function of the natural parameter $\boldsymbol{\eta}$.

$$p(\mathbf{x}|\boldsymbol{\eta}) = h(\mathbf{x})g(\boldsymbol{\eta}) \exp \{ \boldsymbol{\eta}^T \mathbf{u}(\mathbf{x}) \}$$

$$p(\mathbf{X}|\boldsymbol{\eta}) = \left(\prod_{n=1}^N h(\mathbf{x}_n) \right) g(\boldsymbol{\eta})^N \exp \left\{ \boldsymbol{\eta}^T \sum_{n=1}^N \mathbf{u}(\mathbf{x}_n) \right\}.$$

- Exercise: differentiate the log w.r.t. $\boldsymbol{\eta}$ to find:

$$-\nabla \ln g(\boldsymbol{\eta}_{\text{ML}}) = \underbrace{\frac{1}{N} \sum_{n=1}^N \mathbf{u}(\mathbf{x}_n)}_{\text{Sufficient Statistic}}$$