

STA 414/2104: Machine Learning

Russ Salakhutdinov

Department of Computer Science

Department of Statistics

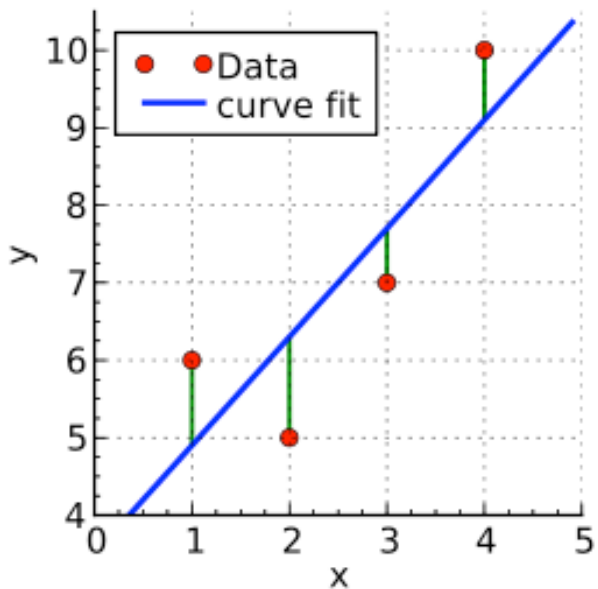
rsalakhu@cs.toronto.edu

<http://www.cs.toronto.edu/~rsalakhu/>

Lecture 2

Linear Least Squares

From last class: Minimize **the sum of the squares of the errors** between the predictions $y(\mathbf{x}_n, \mathbf{w})$ for each data point x_n and the corresponding real-valued targets t_n .



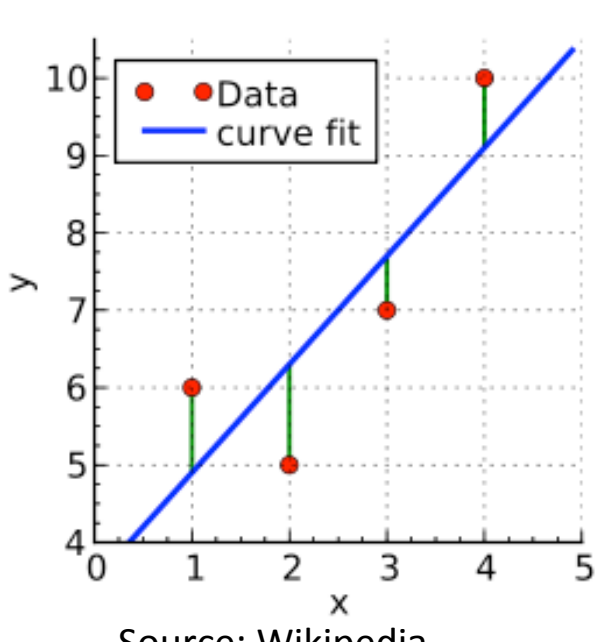
Source: Wikipedia

Loss function: sum-of-squared error function:

$$\begin{aligned} E(\mathbf{w}) &= \frac{1}{2} \sum_{n=1}^N (\mathbf{x}_n^T \mathbf{w} - t_n)^2 \\ &= \frac{1}{2} (\mathbf{X}\mathbf{w} - \mathbf{t})^T (\mathbf{X}\mathbf{w} - \mathbf{t}). \end{aligned}$$

Linear Least Squares

If $\mathbf{X}^T\mathbf{X}$ is nonsingular, then the unique solution is given by:



$$\mathbf{w}^* = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{t}$$

optimal weights

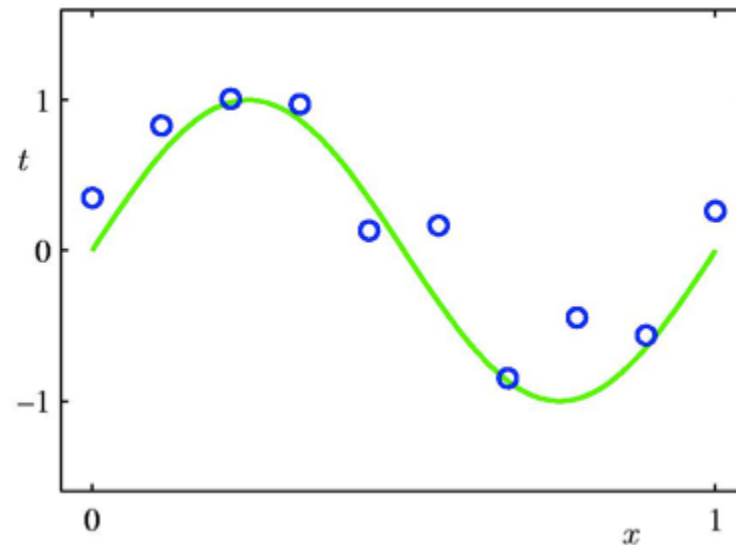
vector of target values

the design matrix has one input vector per row

- At an arbitrary input \mathbf{x}_0 , the prediction is $y(\mathbf{x}_0, \mathbf{w}) = \mathbf{x}_0^T \mathbf{w}^*$.
- The entire model is characterized by $d+1$ parameters \mathbf{w}^* .

Example: Polynomial Curve Fitting

Consider observing a **training set** consisting of N 1-dimensional observations:
 $\mathbf{x} = (x_1, x_2, \dots, x_N)^T$, together with corresponding real-valued targets:
 $\mathbf{t} = (t_1, t_2, \dots, t_N)^T$.



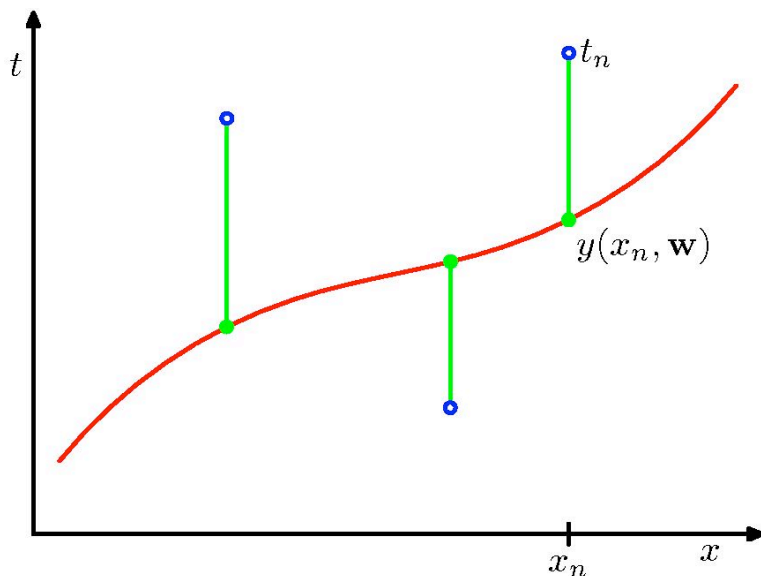
Goal: Fit the data using a polynomial function of the form:

$$y(x, \mathbf{w}) = w_0 + w_1x + w_2x^2 + \dots + w_Mx^M = \sum_{j=0}^M w_jx^j.$$

Note: the polynomial function is a nonlinear function of x , but it is a linear function of the coefficients \mathbf{w} → **Linear Models**.

Example: Polynomial Curve Fitting

- As for the least squares example: we can minimize the sum of the squares of the errors between the predictions $y(x_n, \mathbf{w})$ for each data point x_n and the corresponding target values t_n .



Loss function: sum-of-squared error function:

$$E(\mathbf{w}) = \frac{1}{2} \sum_{i=1}^N (y(x_n, \mathbf{w}) - t_n)^2.$$

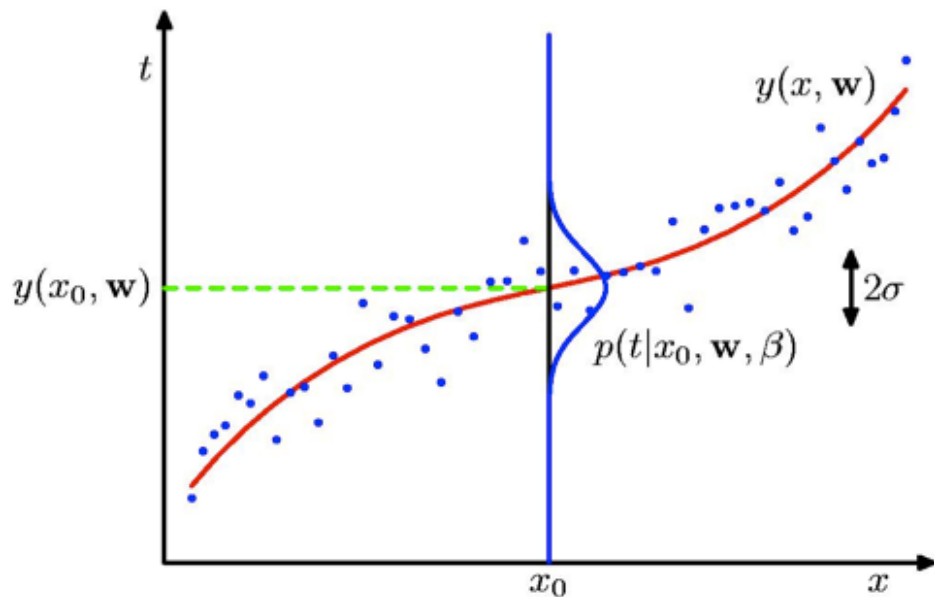
- Similar to the linear least squares: Minimizing sum-of-squared error function has a unique solution \mathbf{w}^* .

Probabilistic Perspective

- So far we saw that polynomial curve fitting can be expressed in terms of error minimization. We now view it from probabilistic perspective.
- Suppose that our model arose from a statistical model:

$$t = y(\mathbf{x}, \mathbf{w}) + \epsilon,$$

where ϵ is a random error having Gaussian distribution with zero mean, and is independent of \mathbf{x} .



Thus we have:

$$p(t|\mathbf{x}, \mathbf{w}, \beta) = \mathcal{N}(t|y(\mathbf{x}, \mathbf{w}), \beta^{-1}),$$

where β is a precision parameter, corresponding to the inverse variance.

I will use probability distribution and probability density interchangeably. It should be obvious from the context.

Maximum Likelihood

If the data are assumed to be independently and identically distributed (*i.i.d assumption*), the likelihood function takes form:

$$p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) = \prod_{i=1}^N \mathcal{N}(t_n | y(\mathbf{x}_n, \mathbf{w}), \beta^{-1}).$$

It is often convenient to maximize the log of the likelihood function:

$$\ln p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) = -\frac{\beta}{2} \underbrace{\sum_{n=1}^N (y(\mathbf{x}_n, \mathbf{w}) - t_n)^2}_{\beta E(\mathbf{w})} + \frac{N}{2} \ln \beta - \frac{N}{2} \ln(2\pi).$$

- Maximizing log-likelihood with respect to \mathbf{w} (under the assumption of a Gaussian noise) is equivalent to minimizing the *sum-of-squared error* function.

- Determine \mathbf{w}_{ML} by maximizing log-likelihood. Then maximizing

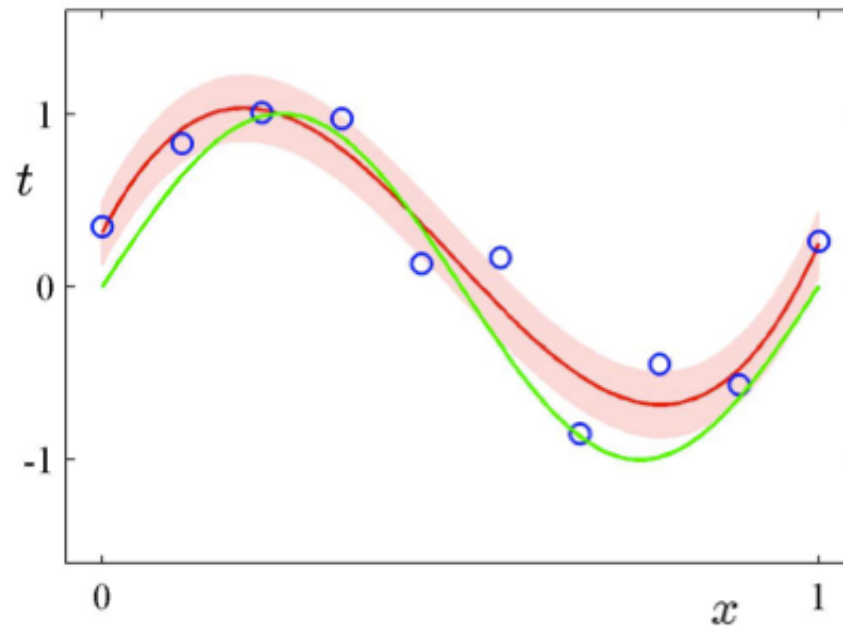
w.r.t. β :

$$\frac{1}{\beta_{ML}} = \frac{1}{N} \sum_n (y(\mathbf{x}_n, \mathbf{w}_{ML}) - t_n)^2.$$

Predictive Distribution

Once we determined the parameters \mathbf{w} and β , we can make prediction for new values of \mathbf{x} :

$$p(t|\mathbf{x}, \mathbf{w}_{ML}, \beta_{ML}) = \mathcal{N}(t|y(\mathbf{x}, \mathbf{w}_{ML}), \beta_{ML}^{-1}).$$



Later we will consider Bayesian linear regression.

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 , so that:

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

- The probability distribution, known as 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 maximize 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

Parameter Estimation

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

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

- 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.

Binomial Distribution

- We can also 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:

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

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

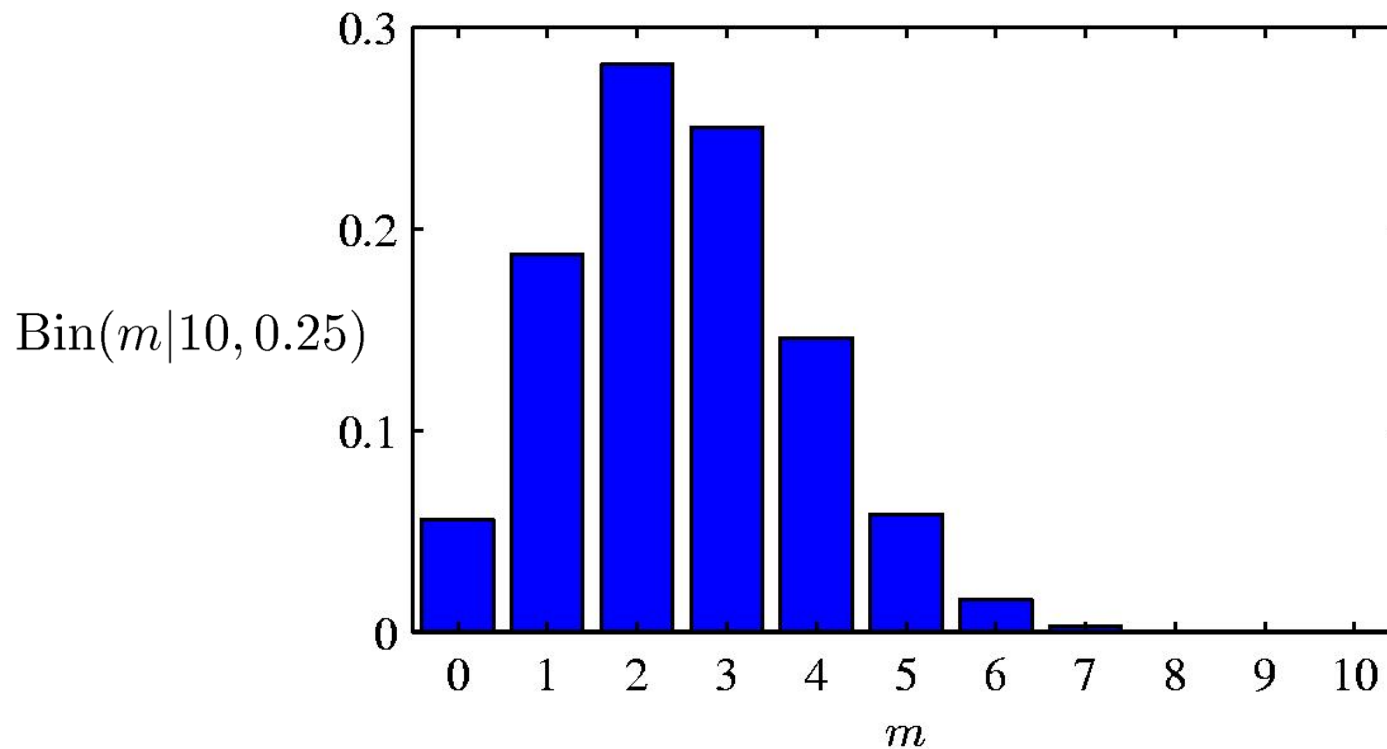
- The mean and variance can be easily derived as:

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

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

Example

- 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 a prior over the parameter μ of the Bernoulli distribution).

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

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

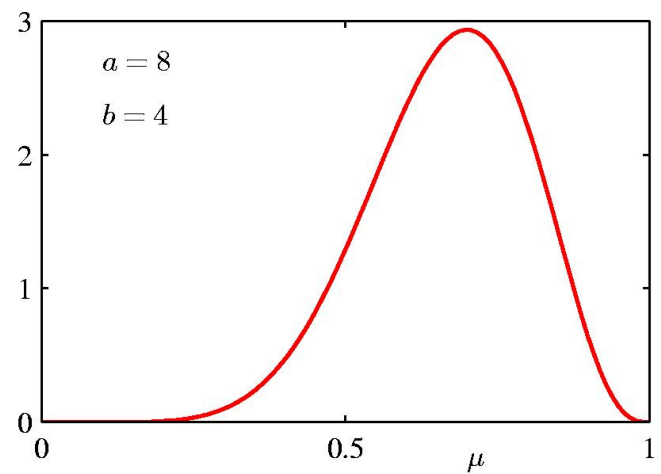
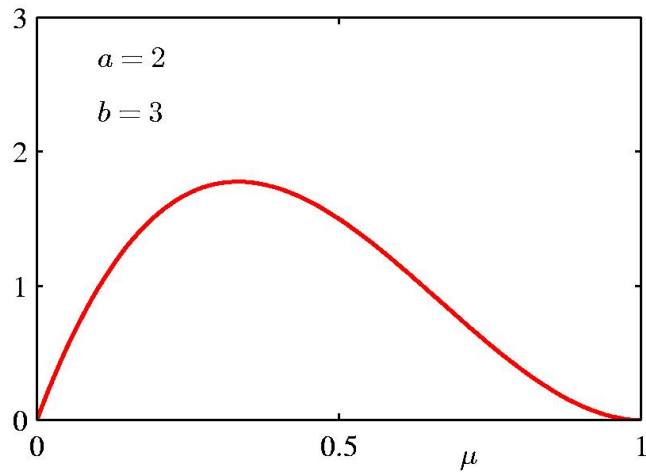
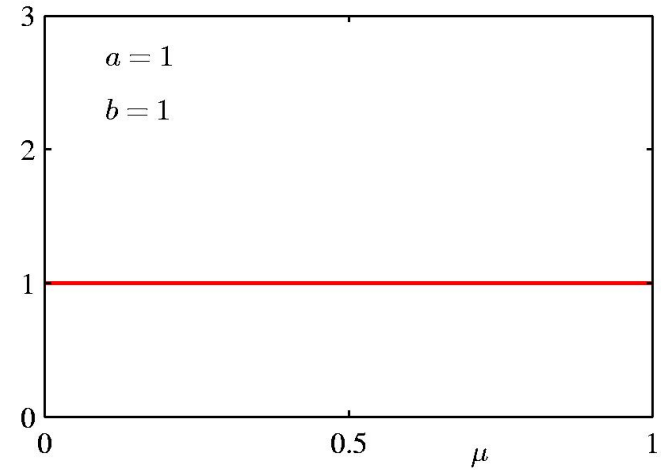
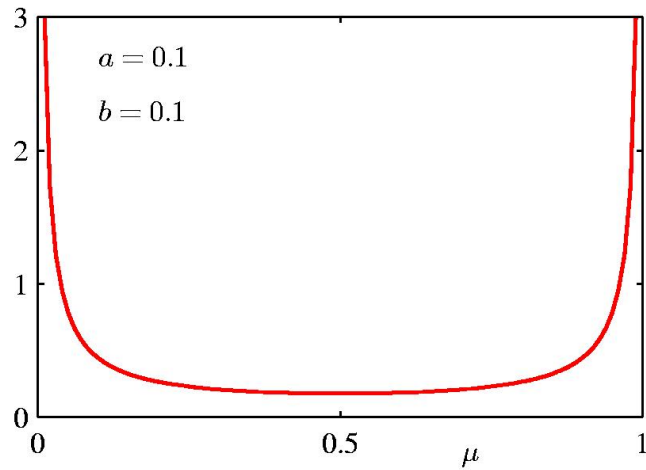
$$\text{var}[\mu] = \frac{ab}{(a+b)^2(a+b+1)}$$

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



Multinomial Variables

- 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 resented 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

- Multinomial 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

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

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 .

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

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

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

which represents the number of observations of $x_k=1$.

- These are called the sufficient statistics for this distribution.

Maximum Likelihood Estimation

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

- To find a maximum likelihood solution for $\boldsymbol{\mu}$, we need to maximize the log-likelihood taking into account the constraint that $\sum_k \mu_k = 1$
- Forming the Lagrangian:

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

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

which is the fraction of observations for which $x_k=1$.

Multinomial Distribution

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

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

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

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

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

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

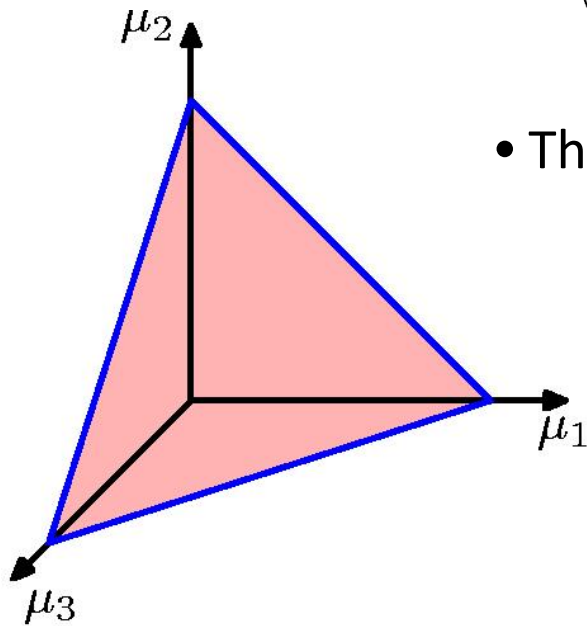
- Note that

$$\sum_k m_k = N.$$

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:

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

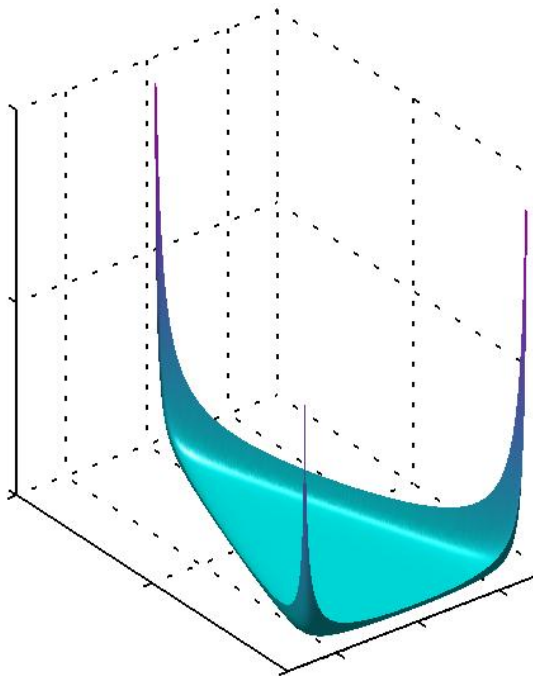
$$\alpha_0 = \sum_{k=1}^K \alpha_k$$

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

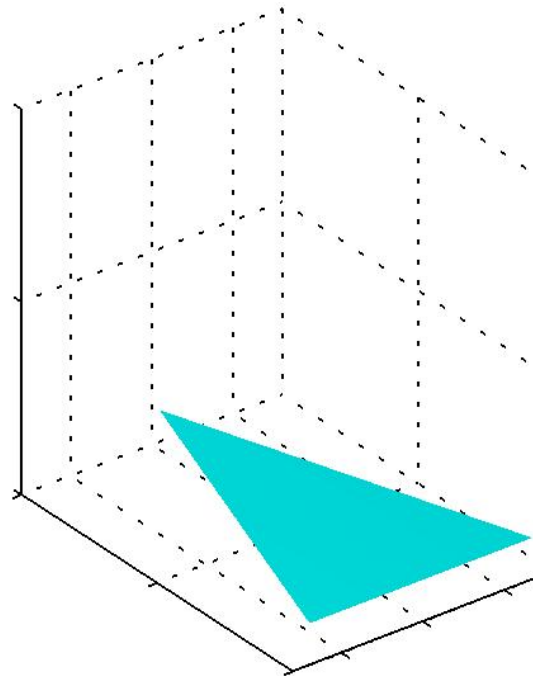
- The Dirichlet distribution is confined to a simplex as a consequence of the constraints.

Dirichlet Distribution

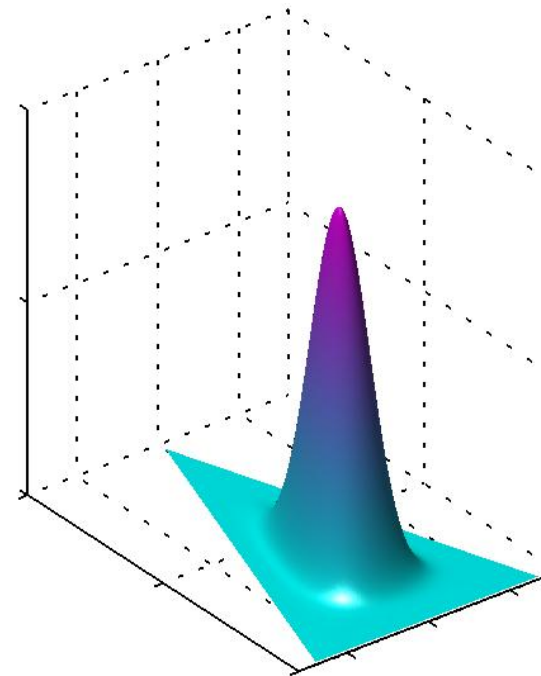
- Plots of the Dirichlet distribution over three variables.



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



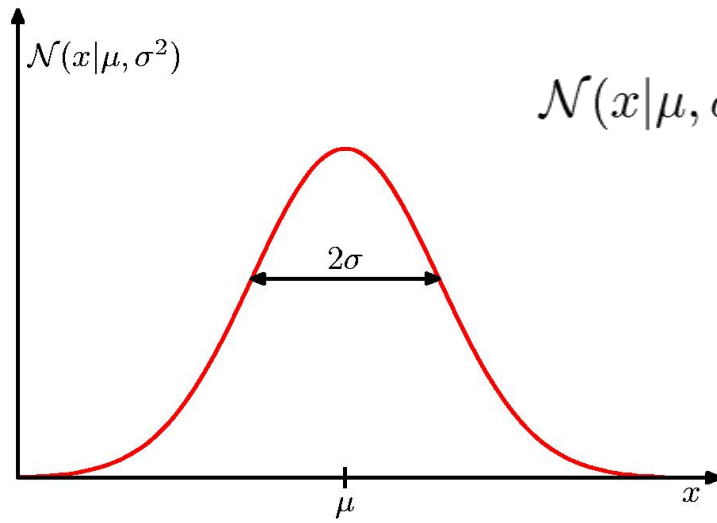
$$\alpha_k = 10^0$$



$$\alpha_k = 10^1$$

Gaussian Univariate Distribution

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

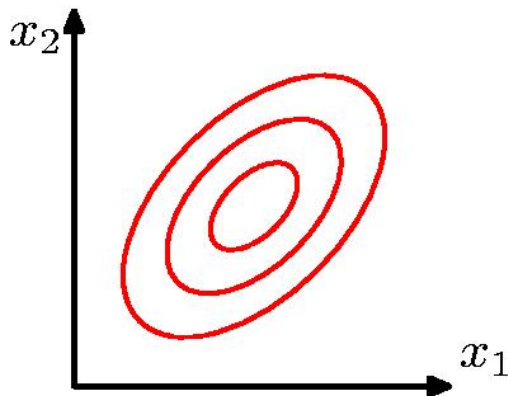
$$\mathcal{N}(x|\mu, \sigma^2) > 0$$

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

Multivariate Gaussian Distribution

- For a D-dimensional vector \mathbf{x} , the Gaussian distribution takes 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 $\boldsymbol{\Sigma}$.

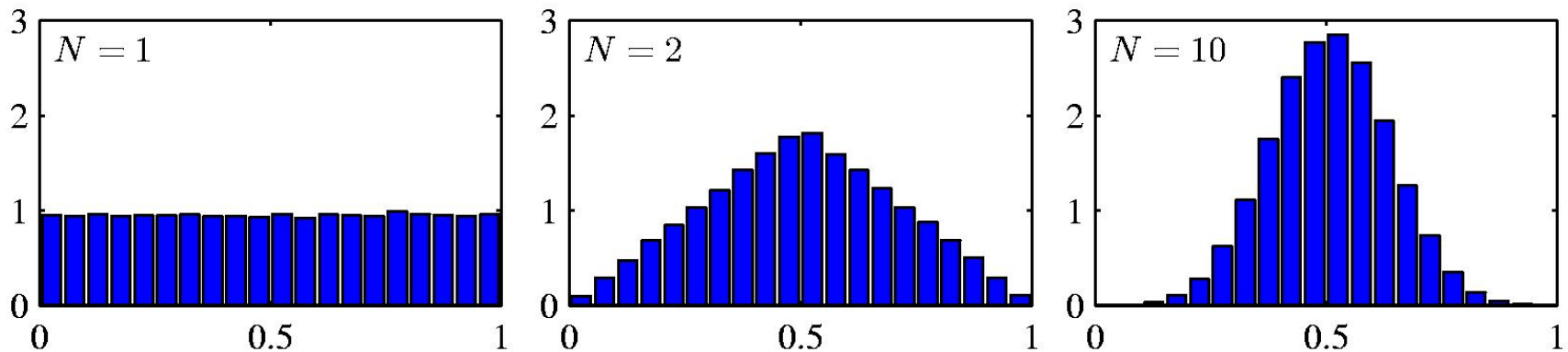
- Note that the covariance matrix is a symmetric positive definite matrix.

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

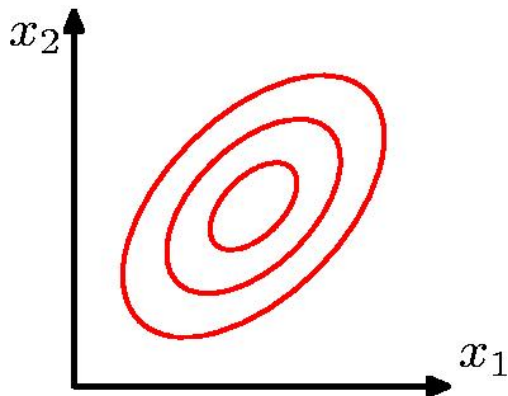
- For a D-dimensional vector \mathbf{x} , the Gaussian distribution takes form:

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

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

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

- Here Δ is known as Mahalanobis distance.



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

Geometry of the Gaussian Distribution

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

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

- Consider the eigenvalue equation for the covariance matrix:

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

- 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:

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

Geometry of the Gaussian Distribution

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

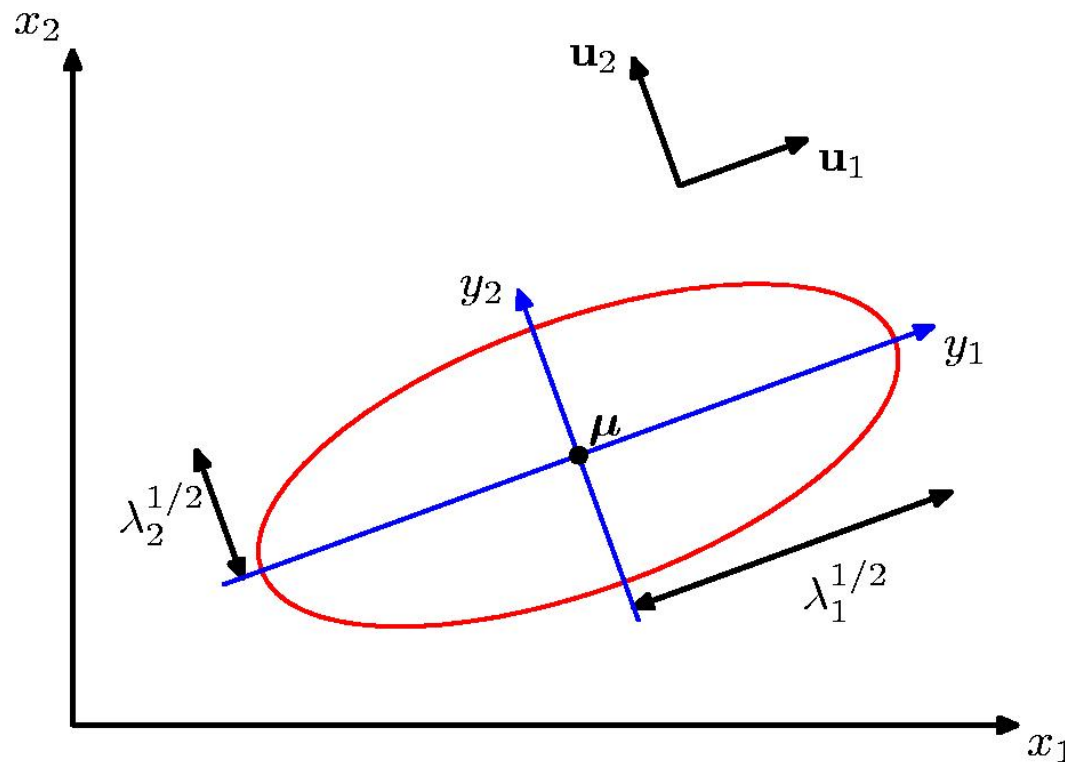
$$\Delta^2 = \sum_{i=1}^D \frac{y_i^2}{\lambda_i} \quad y_i = \mathbf{u}_i^T (\mathbf{x} - \boldsymbol{\mu})$$

- 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 axis 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:

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

The term in \mathbf{z} in the factor $(\mathbf{z} + \boldsymbol{\mu})$ will vanish by symmetry.

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

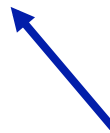
Moments of the Gaussian Distribution

- The second order moments of the Gaussian distribution:

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

- The covariance is given by:

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

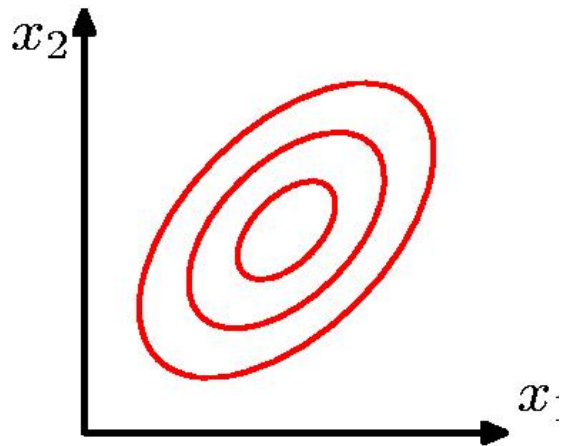


$$\mathbb{E}[\mathbf{x}] = \boldsymbol{\mu}$$

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

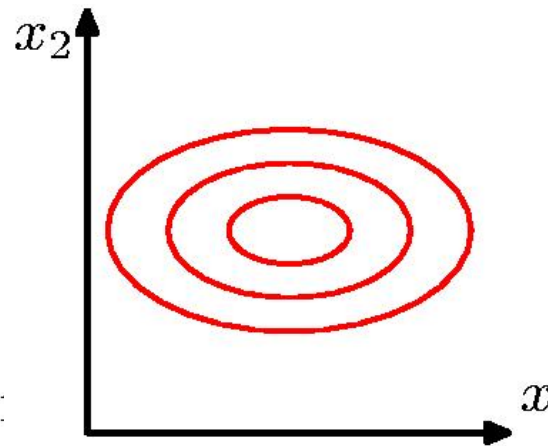
Moments of the Gaussian Distribution

- Contours of constant probability density:



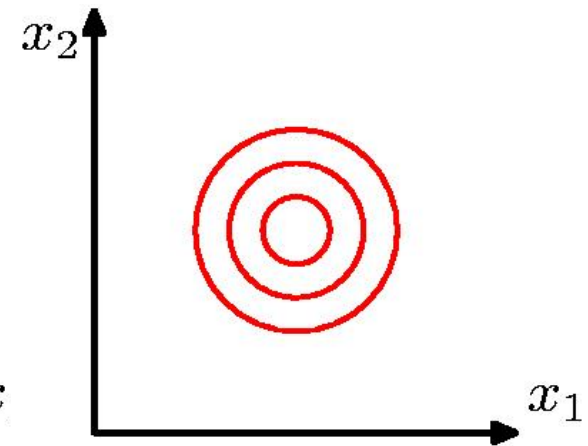
(a)

Covariance matrix is of general form.



(b)

Diagonal, axis-aligned covariance matrix.



(c)

Spherical (proportional to identity) covariance matrix.

Partitioned Gaussian Distribution

- 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 $\boldsymbol{\Sigma}_{aa}$.

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 function of \mathbf{x}_b .



Marginal Distribution

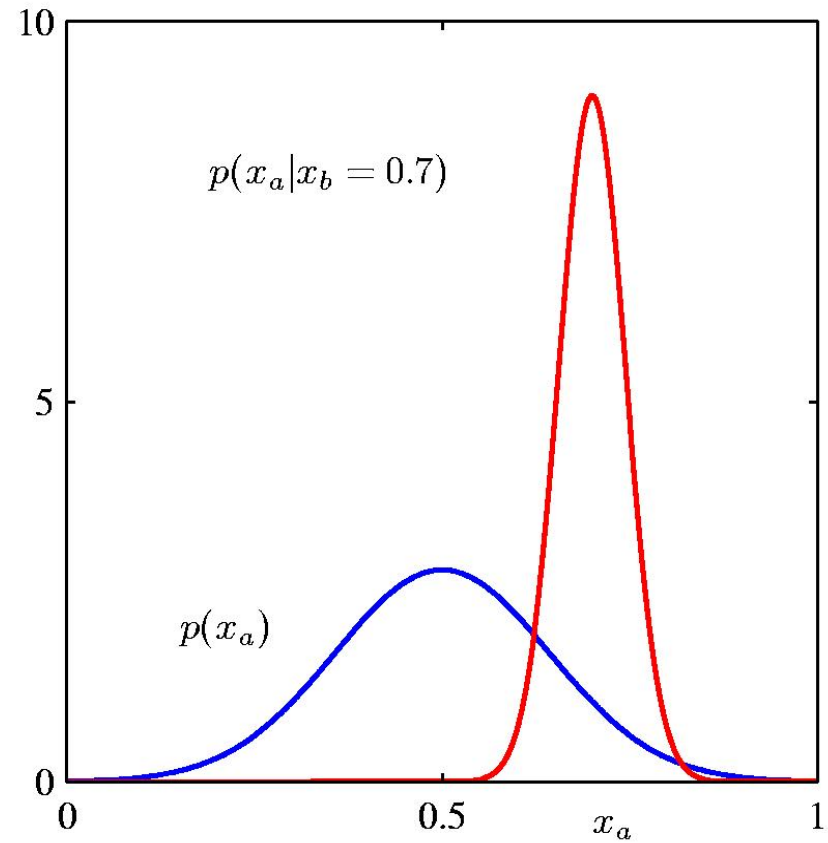
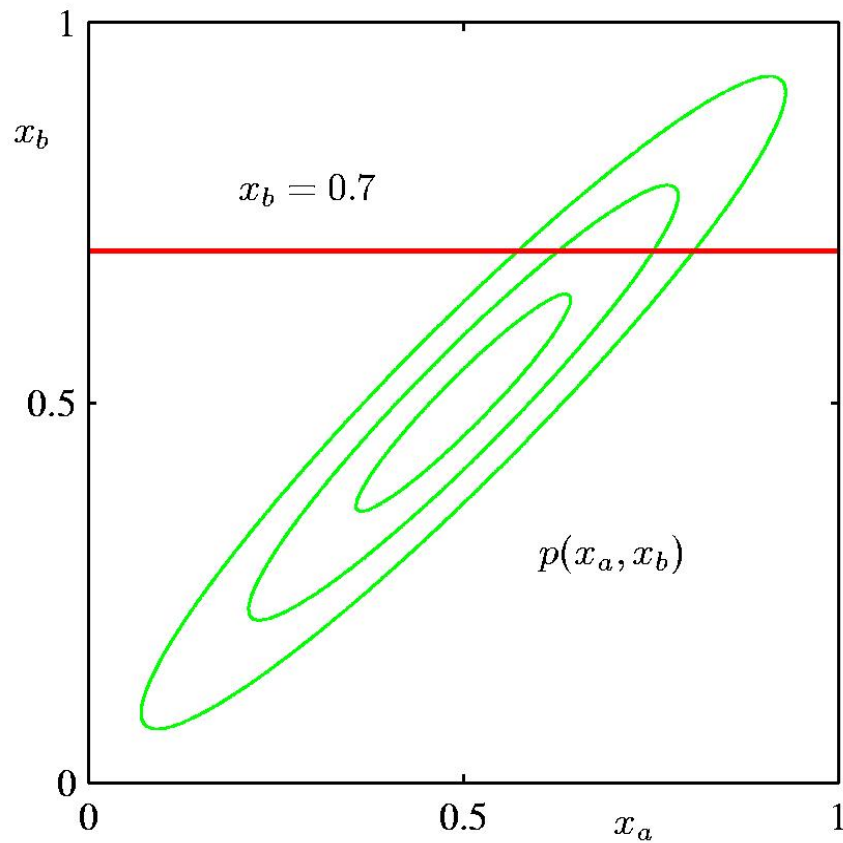
- It turns out that the marginal distribution is also a Gaussian distribution:

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

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

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

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 μ and Σ :

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

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

Sufficient Statistics

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

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

Maximum Likelihood Estimation

- To find a maximum likelihood 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:

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

- Similarly, we can find the ML estimate of $\boldsymbol{\Sigma}$:

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

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. Important for on-line applications.
- Let us consider the contribution of the N^{th} data point \mathbf{x}_n :

$$\begin{aligned}\boldsymbol{\mu}_{\text{ML}}^{(N)} &= \frac{1}{N} \sum_{n=1}^N \mathbf{x}_n \\ &= \frac{1}{N} \mathbf{x}_N + \frac{1}{N} \sum_{n=1}^{N-1} \mathbf{x}_n \\ &= \frac{1}{N} \mathbf{x}_N + \frac{N-1}{N} \boldsymbol{\mu}_{\text{ML}}^{(N-1)} \\ &= \boldsymbol{\mu}_{\text{ML}}^{(N-1)} + \frac{1}{N} (\mathbf{x}_N - \boldsymbol{\mu}_{\text{ML}}^{(N-1)})\end{aligned}$$

The diagram consists of three red arrows pointing from the final equation to text labels on the right. The first arrow starts from the term $\boldsymbol{\mu}_{\text{ML}}^{(N-1)}$ and points to the label "old estimate". The second arrow starts from the fraction $\frac{1}{N}$ and points to the label "correction weight". The third arrow starts from the term $(\mathbf{x}_N - \boldsymbol{\mu}_{\text{ML}}^{(N-1)})$ and points to the label "correction given \mathbf{x}_N ".

Student's t-Distribution

- Consider Student's t-Distribution

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

Infinite mixture
of Gaussians

where

$$\lambda = a/b$$

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

$$\nu = 2a.$$

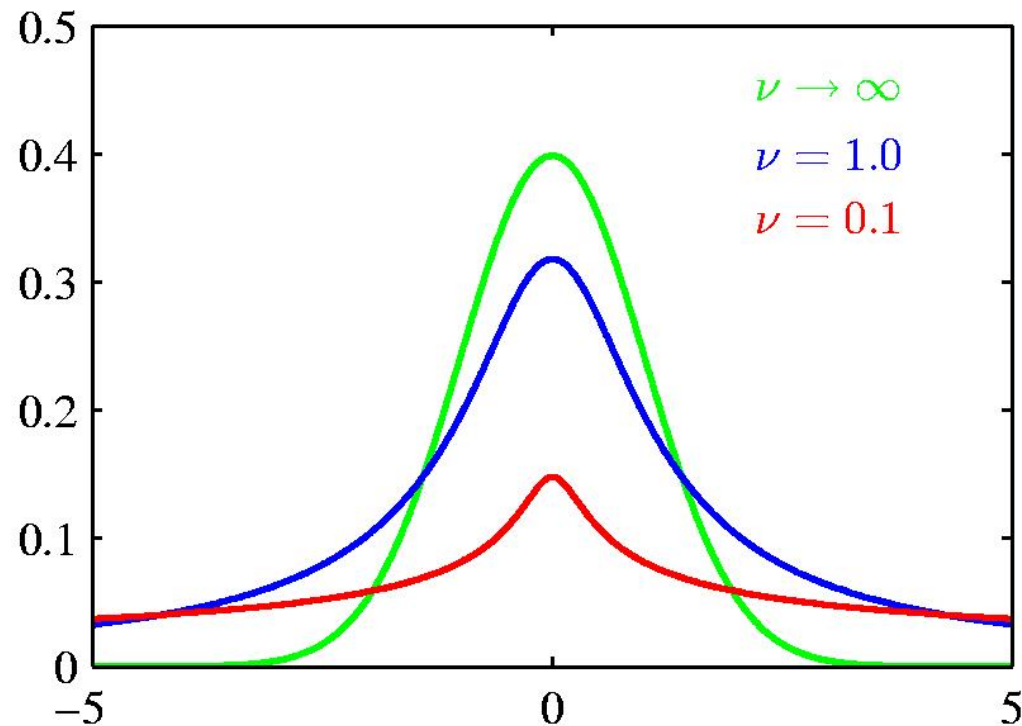
Sometimes called
the precision
parameter.

Degrees of freedom

Student's t-Distribution

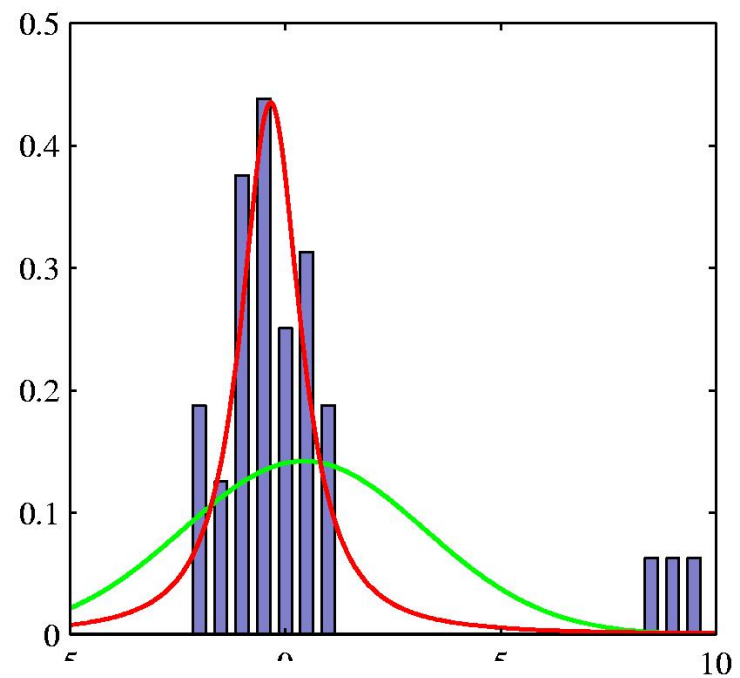
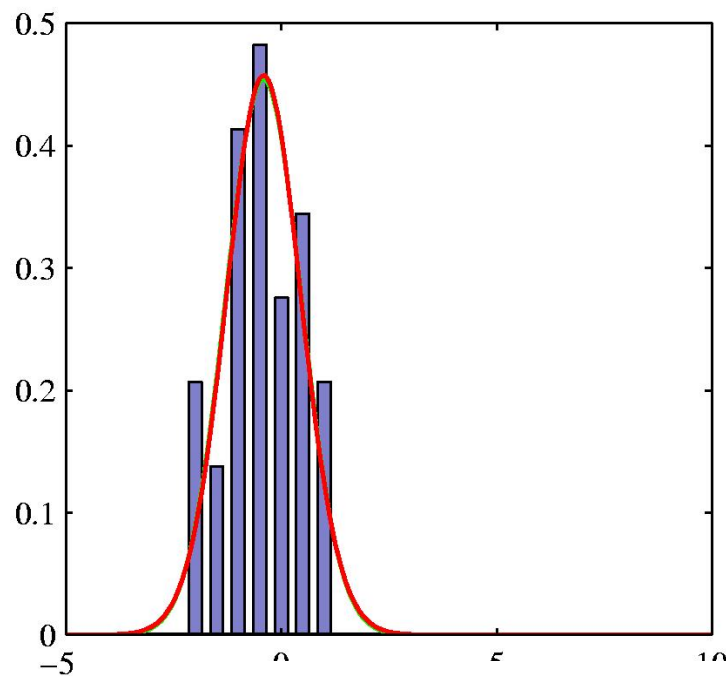
- Setting $\nu = 1$ recovers Cauchy distribution
- The limit $\nu \rightarrow \infty$ corresponds to a Gaussian distribution.

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



Student's t-Distribution

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



Student's t-Distribution

- The multivariate extension of the t-Distribution:

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

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

- Properties:

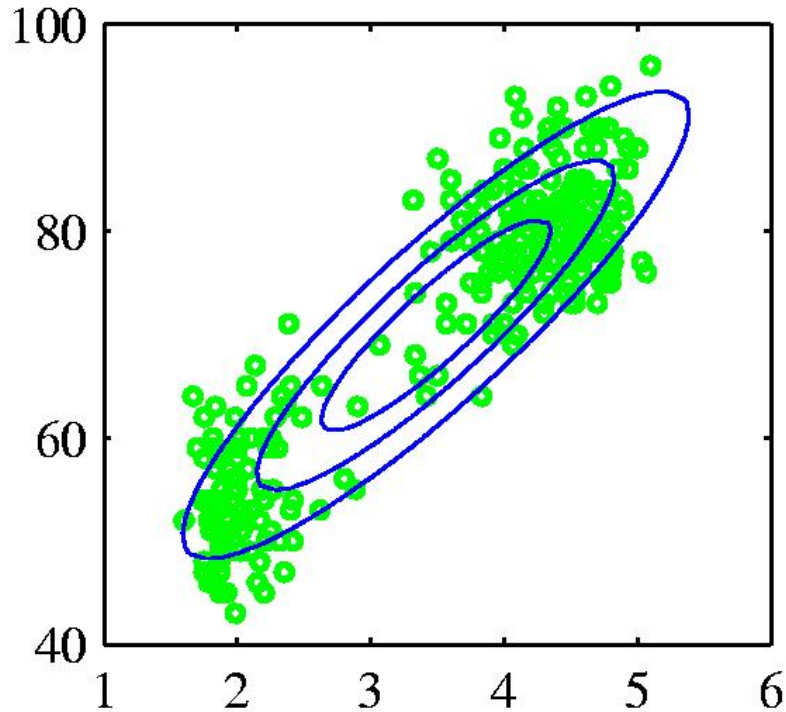
$$\mathbb{E}[\mathbf{x}] = \boldsymbol{\mu}, \quad \text{if } \nu > 1$$

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

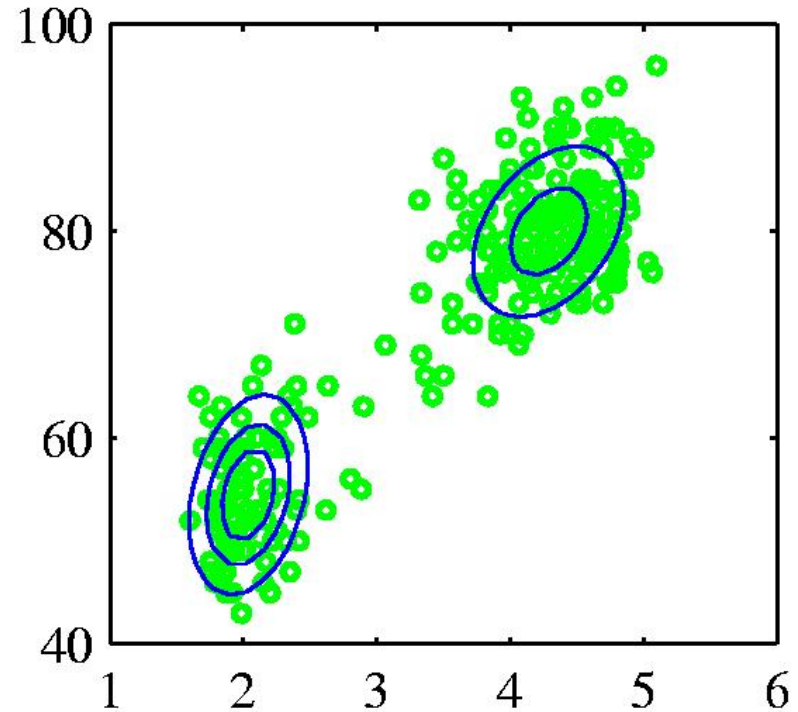
$$\text{mode}[\mathbf{x}] = \boldsymbol{\mu}$$

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

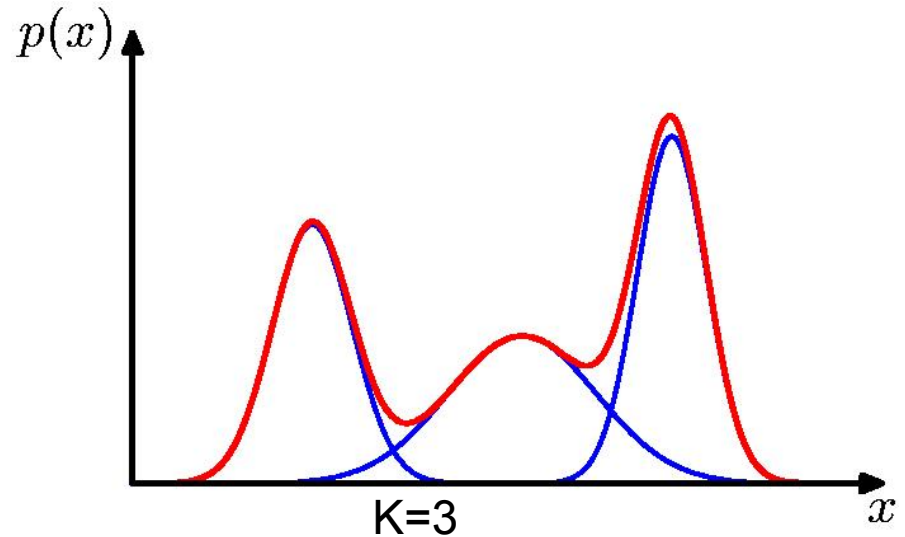
- We can combine simple models into a complex model by defining a superposition of K Gaussian densities of the form:

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

↓
Mixing coefficient

Component

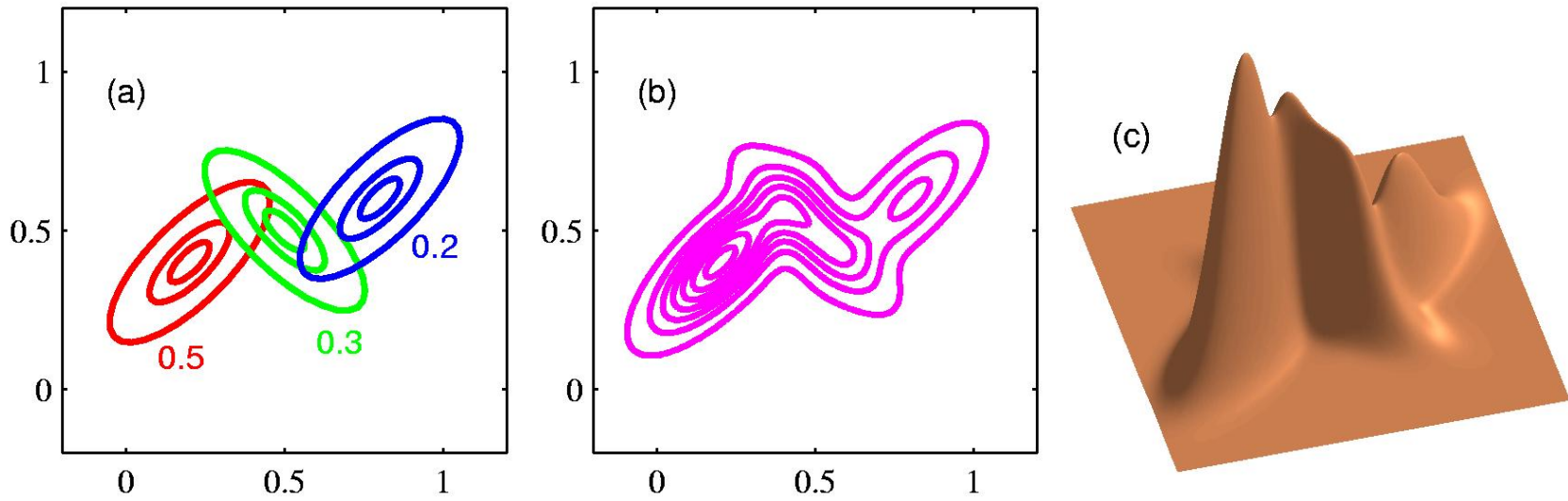
$$\forall k : \pi_k \geq 0 \quad \sum_{k=1}^K \pi_k = 1$$



- Note that each Gaussian component has its own mean μ_k and covariance Σ_k . The parameters π_k are called mixing coefficients.
- Note 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 $p(\mathbf{x})$.

Maximum Likelihood Estimation

- Given a dataset D , we can determine model parameters μ_k, Σ_k, π_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

- **Solution:** use standard, iterative, numeric optimization methods or the Expectation Maximization algorithm.

The Exponential Family

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

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

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 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:

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

- 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 μ_k must satisfy

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

- In some cases it will be convenient to remove the constraint by expressing the distribution over the M-1 parameters.

Multinomial Distribution

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

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

- Let $\mu_M = 1 - \sum_{k=1}^{M-1} \mu_k$

- This leads to:

$$\eta_k = \ln \left(\frac{\mu_k}{1 - \sum_{j=1}^{M-1} \mu_j} \right) \quad \text{and} \quad \mu_k = \frac{\exp(\eta_k)}{\underbrace{1 + \sum_{j=1}^{M-1} \exp(\eta_j)}_{\text{Softmax function}}}.$$

- Here the parameters η_k are independent.

- Note that:

$$0 \leq \mu_k \leq 1 \quad \text{and} \quad \sum_{k=1}^{M-1} \mu_k \leq 1.$$

Multinomial Distribution

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

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

- The Multinomial distribution can therefore be written as:

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

where

$$\begin{aligned} \boldsymbol{\eta} &= (\eta_1, \dots, \eta_{M-1}, 0)^T \\ \mathbf{u}(\mathbf{x}) &= \mathbf{x} \\ h(\mathbf{x}) &= 1 \\ g(\boldsymbol{\eta}) &= \left(1 + \sum_{k=1}^{M-1} \exp(\eta_k) \right)^{-1}. \end{aligned}$$

Gaussian Distribution

- The Gaussian distribution can be written as:

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

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 $\boldsymbol{\eta}$:

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

- Thus

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

ML for the Exponential Family

- Remember the Exponential Family:

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

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

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

- Thus

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

- Note that 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 η .

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

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

- Therefore we have

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

Sufficient Statistic