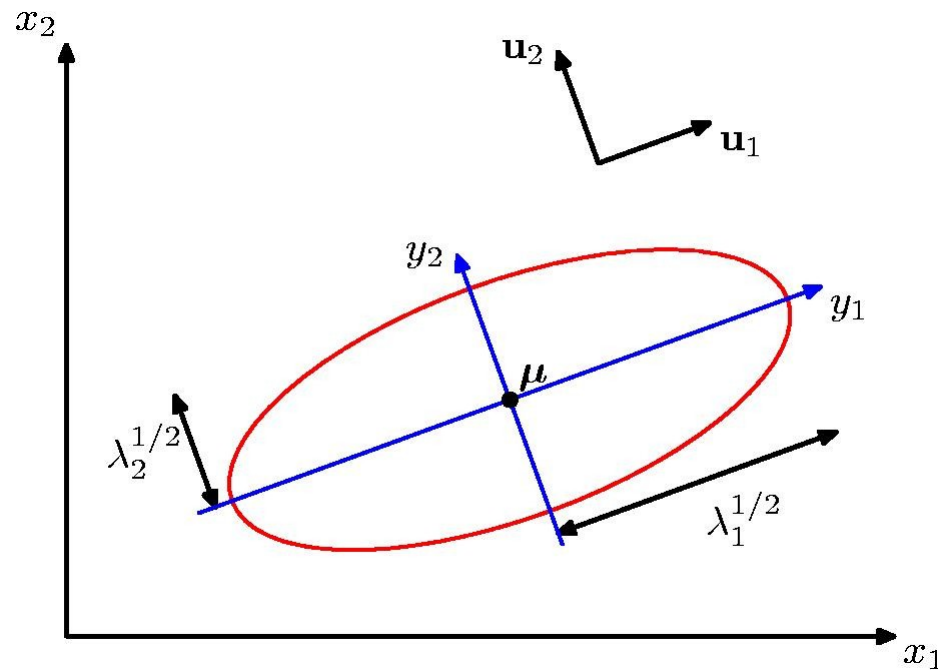


Crash course in probability theory and statistics – part 2



Machine Learning, Wed Apr 16, 2008

Motivation

All models are wrong, but some are useful.

This lecture introduces distributions that have proven useful in constructing models.

Densities, statistics and estimators

A **probability (density)** is any function $X \mapsto p(X)$ that satisfies the probability theory axioms.

A **statistic** is any function of observed data $x \mapsto f(x)$.

An **estimator** is a statistic used for estimating a parameter of the probability density $x \mapsto \mu$.

Estimators

Assume $\mathcal{D} = \{x_1, x_2, \dots, x_N\}$ are *independent, identically distributed (i.i.d)* outcomes of our experiments (observed data).

Desirable properties of an estimator are:

$$\hat{\mu} \rightarrow \mu \quad \text{for } N \rightarrow \infty$$

and

$$\mathbb{E}[\hat{\mu}] = \mu \quad \text{(unbiased)}$$

Estimators

A general way to get an estimator is using the *maximum likelihood (ML)* or *maximum a posterior (MAP)* approach:

ML:
$$\hat{\mu} = \underset{\mu}{\operatorname{maxarg}} p(x | \mu)$$

MAP:
$$\begin{aligned} \hat{\mu} &= \underset{\mu}{\operatorname{maxarg}} p(\mu | x) \\ &= \underset{\mu}{\operatorname{maxarg}} p(x | \mu)p(\mu) \end{aligned}$$

...possibly compensating for any bias in these.

Bayesian estimation

In a fully Bayesian approach we instead update our distribution based on observed data:

$$\hat{p}(y | x) = \int p(y | \mu)p(\mu | x) d\mu$$
$$\propto \int p(y | \mu)p(x | \mu)p(\mu) d\mu$$

Conjugate priors

If the prior distribution belongs to a certain class of functions, $p(\mu) \in C$, and the product of prior and likelihood belongs to the same class

$p(x|\mu) p(\mu) \in C$, then the prior is called a **conjugate prior**.

Conjugate priors

If the prior distribution belongs to a certain class of functions, $p(\mu) \in C$, and the product of prior and likelihood belongs to the same class

$p(x|\mu) p(\mu) \in C$, then the prior is called a **conjugate prior**.

Typical situation:

$$p(\mu | \theta) \propto f(\mu, \theta)$$

$$p(x | \mu) f(\mu, \theta) \propto f(\mu, \theta_x)$$

where f is a well known function with known normalizing constant

$$C(\theta)^{-1} = \int f(\mu, \theta) d\mu$$

Conjugate priors

$$\begin{aligned} p(y | x, \theta) &= \int p(y | \mu) p(\mu | x, \theta) d\mu \\ &\propto \int p(y | \mu) p(x | \mu) p(\mu | \theta) d\mu \\ &\propto \int p(y | \mu) p(x | \mu) f(\mu, \theta) d\mu \\ &\propto \int p(y | \mu) f(\mu, \theta_x) d\mu \\ &\propto \int f(\mu, \theta_{x,y}) d\mu = \frac{1}{C(\theta_{x,y})} \end{aligned}$$

$$p(y | x, \theta) = \frac{C(\theta_{x,y})^{-1}}{\int C(\theta_{x,z})^{-1} dz}$$

Bernoulli and binomial distribution

Bernoulli distribution: single event with binary outcome.

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

Binomial: sum of N Bernoulli outcomes.

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

Mean for Bernoulli:

$$\mathbb{E}[x] = 1 \cdot \mu + 0 \cdot (1 - \mu) = \mu$$

Mean for Binomial:

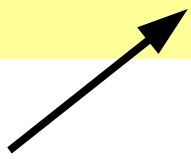
$$\mathbb{E}[m] = \sum_{m=1}^N \mathbb{E}[x] = N\mu$$

Bernoulli and binomial distribution

Bernoulli maximum likelihood for $\mathcal{D} = \{x_1, \dots, x_N\}$

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

$$\begin{aligned} \log p(\mathcal{D} \mid \mu) &= \sum_{n=1}^N \{x_n \log \mu + (1 - x_n) \log(1 - \mu)\} \\ &= m \log \mu + (N - m) \log(1 - \mu) \end{aligned}$$

$$m = \sum_{n=1}^N x_n$$


Sufficient statistic

Bernoulli and binomial distribution

Bernoulli maximum likelihood for $\mathcal{D} = \{x_1, \dots, x_N\}$

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

$$\begin{aligned} \log p(\mathcal{D} \mid \mu) &= \sum_{n=1}^N \{x_n \log \mu + (1 - x_n) \log(1 - \mu)\} \\ &= m \log \mu + (N - m) \log(1 - \mu) \end{aligned}$$

ML estimate: $\hat{\mu} = \frac{m}{N}$

Bernoulli and binomial distribution

Bernoulli maximum likelihood for $\mathcal{D} = \{x_1, \dots, x_N\}$

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

$$\begin{aligned} \log p(\mathcal{D} \mid \mu) &= \sum_{n=1}^N \{x_n \log \mu + (1 - x_n) \log(1 - \mu)\} \\ &= m \log \mu + (N - m) \log(1 - \mu) \end{aligned}$$

ML estimate: $\hat{\mu} = \frac{m}{N}$

Average, so $\hat{\mu} \rightarrow \mathbb{E}[x] = \mu$

$$\mathbb{E}[\hat{\mu}] = \frac{1}{N} \sum_{n=1}^N \mathbb{E}[x_n] = \frac{1}{N} \sum_{n=1}^N \mu = \mu$$

Bernoulli and binomial distribution

Similar for binomial distribution:

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

$$\begin{aligned} \log p(\mathcal{D} \mid \mu) &\propto \sum_{n=1}^N \{x_n \log \mu + (1 - x_n) \log(1 - \mu)\} \\ &= m \log \mu + (N - m) \log(1 - \mu) \end{aligned}$$

ML estimate:

$$\hat{\mu} = \frac{m}{N}$$

Beta distribution

Beta distribution:

$$\text{Beta}(\mu \mid \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \mu^{\alpha-1} (1 - \mu)^{\beta-1}$$

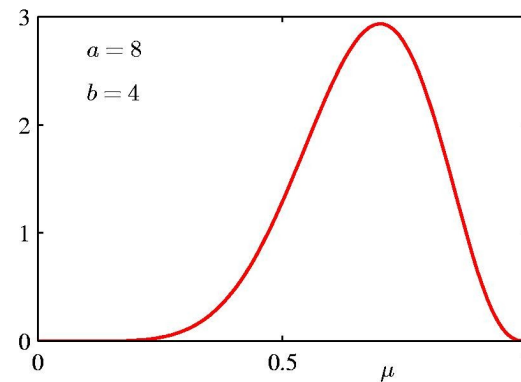
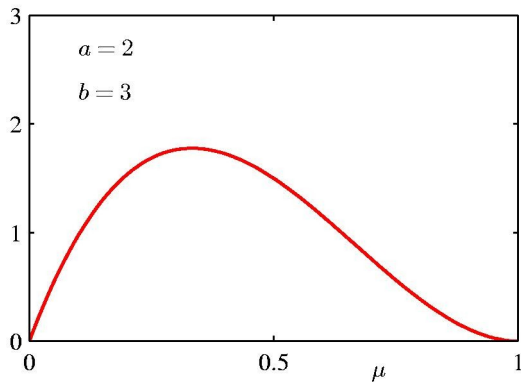
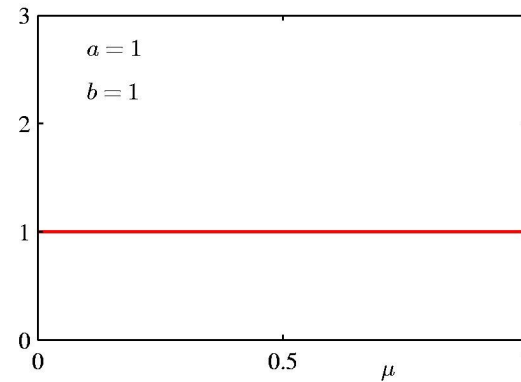
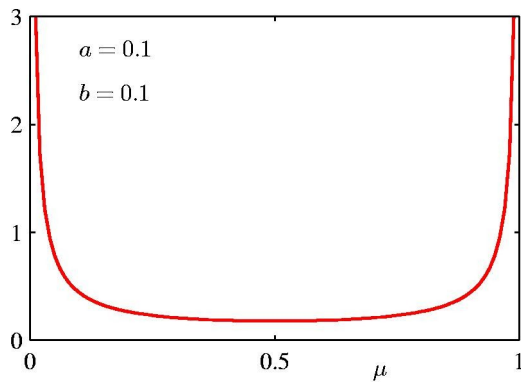
Gamma function



Beta distribution

Beta distribution:

$$\text{Beta}(\mu \mid \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \mu^{\alpha-1} (1 - \mu)^{\beta-1}$$



Beta distribution

Beta distribution:

$$\text{Beta}(\mu \mid \alpha, \beta) = \frac{1}{B(\alpha, \beta)} \mu^{\alpha-1} (1 - \mu)^{\beta-1}$$

Beta function

$$B(\alpha, \beta) = \int_0^1 \mu^{\alpha-1} (1 - \mu)^{\beta-1} d\mu$$

Beta distribution

Beta distribution:

$$\text{Beta}(\mu \mid \alpha, \beta) = \frac{1}{B(\alpha, \beta)} \underbrace{\mu^{\alpha-1} (1 - \mu)^{\beta-1}}_{f(\mu, \alpha, \beta)}$$

Normalizing constant

$f(\mu, \alpha, \beta)$

$$B(\alpha, \beta) = \int_0^1 \mu^{\alpha-1} (1 - \mu)^{\beta-1} d\mu$$

Beta distribution

Beta distribution:

$$\text{Beta}(\mu \mid \alpha, \beta) = \frac{1}{B(\alpha, \beta)} \mu^{\alpha-1} (1 - \mu)^{\beta-1}$$

Conjugate to Bernoulli/Binomial:

$$p(\mu \mid \alpha, \beta) \propto \mu^{\alpha-1} (1 - \mu)^{\beta-1}$$

$$\begin{aligned} p(x \mid \mu)p(\mu \mid \alpha, \beta) &\propto \mu^x (1 - \mu)^{1-x} \mu^{\alpha-1} (1 - \mu)^{\beta-1} \\ &= \mu^{\alpha-1+x} (1 - \mu)^{\beta-1+(1-x)} \end{aligned}$$

Posterior distribution: $\text{Beta}(\mu \mid \alpha + x, \beta + 1 - x)$

Beta distribution

Beta distribution:

$$\text{Beta}(\mu \mid \alpha, \beta) = \frac{1}{B(\alpha, \beta)} \mu^{\alpha-1} (1 - \mu)^{\beta-1}$$

$$\begin{aligned} \hat{p}(y \mid x, \alpha, \beta) &\propto \int \mu^{\alpha-1+x+y} (1 - \mu)^{\beta-1+(1-x)+(1-y)} d\mu \\ &= B(\alpha + x + y, \beta + (1 - x) + (1 - y)) \end{aligned}$$

$$\hat{p}(y \mid x, \alpha, \beta) = \frac{B(\alpha + x + y, \beta + 2 - x - y)}{B(\alpha + x + 1, \beta + 1 - x) + B(\alpha + x, \beta + 2 - x)}$$

Posterior distribution:

$$\text{Beta}(\mu \mid \alpha + x, \beta + 1 - x)$$

Multinomial distribution

One out of K classes: x bitvector with

$$\sum_{k=1}^K x_k = 1$$

Distribution:

$$p(\mathbf{x} \mid \mu) = \prod_{k=1}^K \mu_k^{x_k} \quad \mathbb{E}[x] = \mu$$

Likelihood:

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

$$m_k = \sum_{n=1}^N x_{k,n}$$

Sufficient
statistic



Multinomial distribution

Maximum likelihood estimate:

$$\hat{\mu}_k = \frac{m_k}{N}$$

$$\hat{\mu}_k \rightarrow \mu_k$$

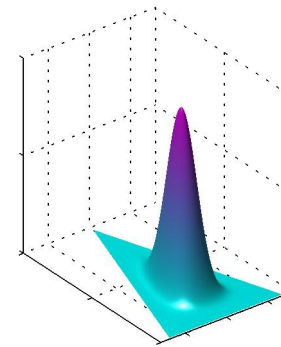
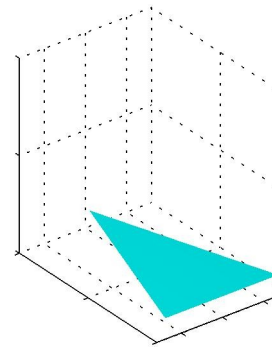
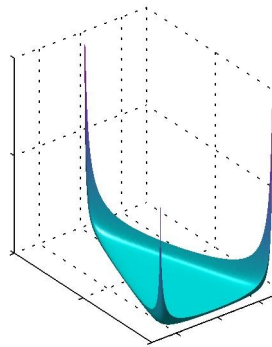
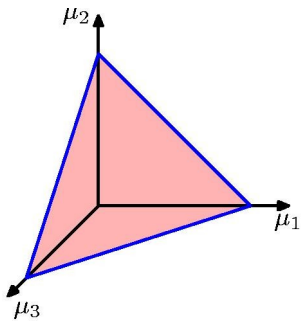
$$\mathbb{E}[\hat{\mu}_k] = \mathbb{E} \left[\frac{1}{N} \sum_{n=1}^N x_{n,k} \right] = \frac{1}{N} \sum_{n=1}^N \mathbb{E}[x_{n,k}]$$

$$= \frac{1}{N} N \cdot \mathbb{E}[x_{n,k}] = \mu_k$$

Dirichlet distribution

Dirichlet distribution:

$$p(\mu \mid \alpha, \beta) = \frac{\Gamma\left(\sum_{k=1}^K \alpha_k\right)}{\prod_{k=1}^K \Gamma(\alpha_k)} \prod_{k=1}^K \mu_k^{\alpha_k - 1}$$



Dirichlet distribution

Dirichlet distribution:

$$p(\mu \mid \alpha, \beta) = \frac{\Gamma\left(\sum_{k=1}^K \alpha_k\right)}{\prod_{k=1}^K \Gamma(\alpha_k)} \prod_{k=1}^K \mu_k^{\alpha_k - 1}$$

Normalizing constant

$f(\mu, \alpha, \beta)$

Dirichlet distribution

Dirichlet distribution:

$$p(\mu \mid \alpha, \beta) = \frac{\Gamma\left(\sum_{k=1}^K \alpha_k\right)}{\prod_{k=1}^K \Gamma(\alpha_k)} \prod_{k=1}^K \mu_k^{\alpha_k - 1}$$

Conjugate to Multinomial:

$$p(\mu \mid \alpha) \propto \prod_{k=1}^K \mu_k^{\alpha_k - 1}$$

$$p(x \mid \mu)p(\mu \mid \alpha) \propto \left(\prod_{k=1}^K \mu_k^{x_k}\right) \left(\prod_{k=1}^K \mu_k^{\alpha_k - 1}\right) = \prod_{k=1}^K \mu_k^{\alpha_k - 1 + x_k}$$

Gaussian/Normal distribution

Scalar variable:

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

Vector variable:

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



Symmetric, real, $D \times D$ matrix

Geometry of a Gaussian

Sufficient statistic:

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

There's a linear transformation U so $\Sigma^{-1} = U^T \Lambda^{-1} U$ where Λ is a diagonal matrix. Then

$$\begin{aligned} \Delta^2 &= (\mathbf{x} - \mu)^T \Sigma^{-1} (\mathbf{x} - \mu) = (\mathbf{x} - \mu)^T U^T \Lambda^{-1} U (\mathbf{x} - \mu) \\ &= \mathbf{y}^T \Lambda^{-1} \mathbf{y} = \sum_{i=1}^D \frac{y_i^2}{\lambda_i} \end{aligned}$$

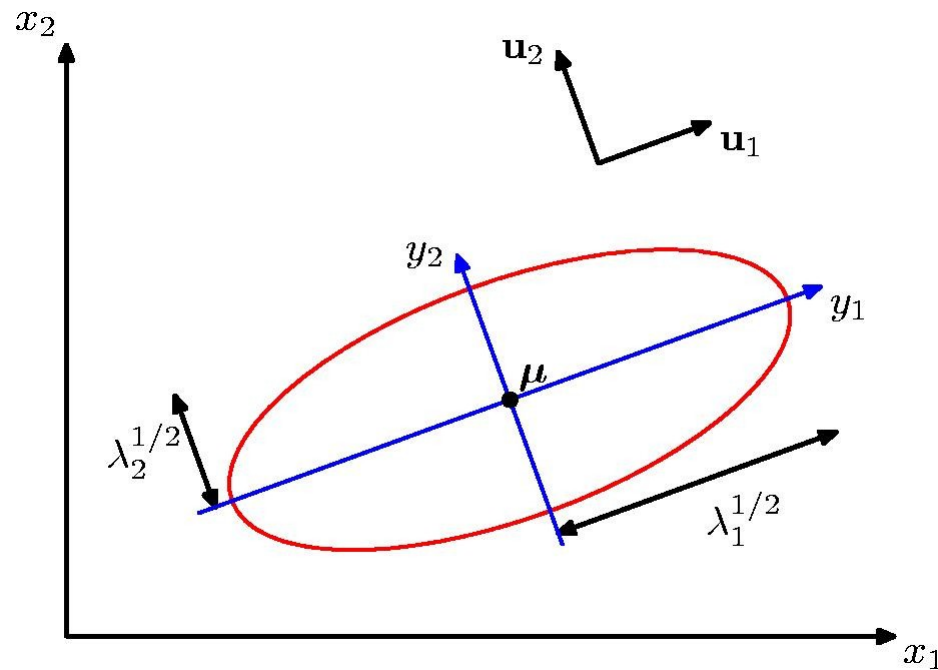
where $\mathbf{y} := U(\mathbf{x} - \mu)$

Geometry of a Gaussian

Gaussian constant when

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

is constant – an ellipsis in the U coordinate system:

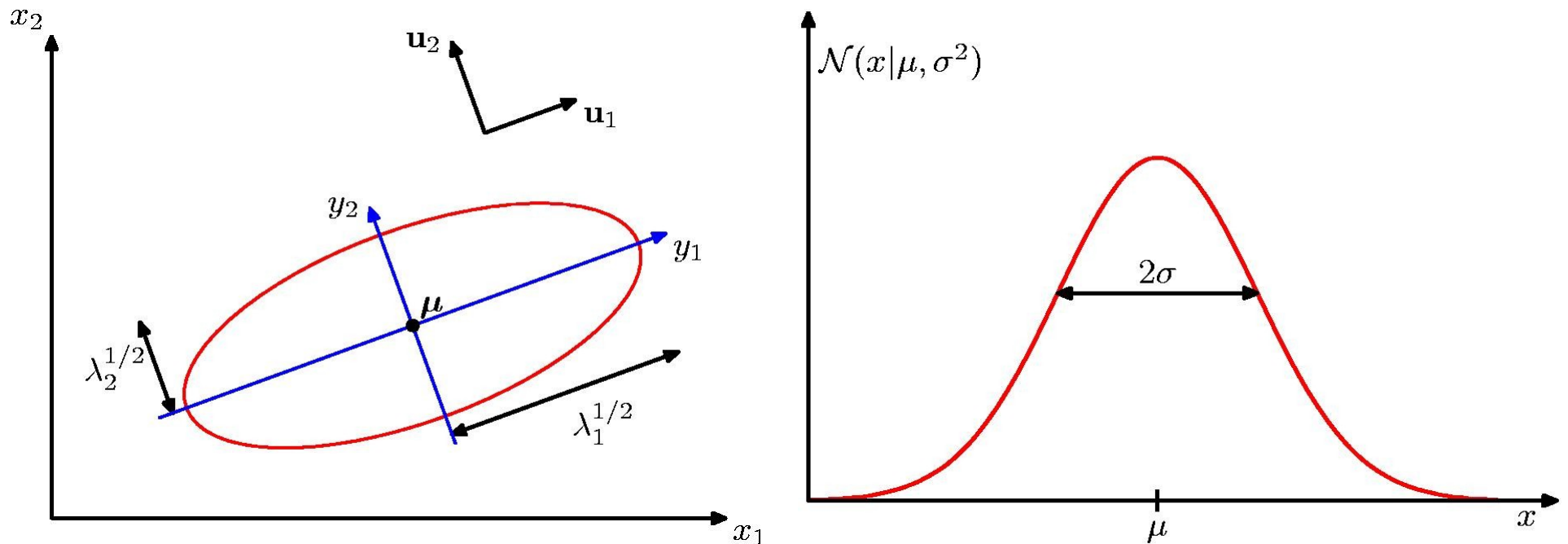


Geometry of a Gaussian

Gaussian constant when

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

is constant – an ellipsis in the U coordinate system:



Parameters of a Gaussian

Parameters are mean and variance/covariance:

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

$$\text{var}[x] = \sigma^2$$

$$\text{cov}[\mathbf{x}, \mathbf{x}] = \Sigma$$

ML estimates are:

$$\hat{\mu} = \frac{1}{N} \sum_{n=1}^N x_n$$

$$\hat{\sigma}^2 = \frac{1}{N} \sum_{n=1}^N (x_n - \hat{\mu})^2 \quad \hat{\Sigma} = \frac{1}{N} \sum_{n=1}^N (\mathbf{x}_n - \hat{\mu})(\mathbf{x}_n - \hat{\mu})^T$$

Parameters of a Gaussian

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

Unbiased

$$\mathbb{E}[\hat{\Sigma}] = \frac{N-1}{N} \Sigma$$

Biased

ML estimates are:

$$\hat{\mu} = \frac{1}{N} \sum_{n=1}^N x_n$$

$$\hat{\sigma}^2 = \frac{1}{N} \sum_{n=1}^N (x_n - \hat{\mu})^2 \quad \hat{\Sigma} = \frac{1}{N} \sum_{n=1}^N (\mathbf{x}_n - \hat{\mu})(\mathbf{x}_n - \hat{\mu})^T$$

Parameters of a Gaussian

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

Unbiased

$$\mathbb{E}[\hat{\Sigma}] = \frac{N-1}{N} \Sigma$$

Biased

Intuition: The variance estimator is based on the mean estimator – which is fitted to data and fits better than the real mean, thus the variance is under estimated.

Correction:

$$\tilde{\Sigma} = \frac{N}{N-1} \mathbb{E}[\hat{\Sigma}] = \frac{1}{N-1} \sum_{n=1}^N (\mathbf{x}_n - \hat{\mu})(\mathbf{x}_n - \hat{\mu})^T$$

Gaussian is its own conjugate

For *fixed* variance:

$$\underbrace{\frac{1}{\sqrt{2\pi\sigma^2}}}_{\text{Normalizing constant}} \underbrace{\exp\left\{-\frac{1}{2\sigma^2}(x - \mu)^2\right\}}_{f(\mu, \sigma^2)}$$

Normalizing constant

$f(\mu, \sigma^2)$

Gaussian is its own conjugate

For *fixed* variance:

$$\frac{1}{\sqrt{2\pi\sigma^2}} \exp \left\{ -\frac{1}{2\sigma^2} (x - \mu)^2 \right\}$$

$$p(\mu | \mu_0, \sigma_0^2) \propto \exp \left\{ -\frac{1}{2\sigma_0^2} (\mu - \mu_0)^2 \right\}$$

$$\begin{aligned} p(x | \mu, \sigma^2) p(\mu | \mu_0, \sigma_0^2) &\propto \exp \left\{ -\frac{1}{2\sigma^2} (x - \mu)^2 \right\} \exp \left\{ -\frac{1}{2\sigma_0^2} (\mu - \mu_0)^2 \right\} \\ &= \exp \left\{ -\frac{1}{2\sigma_x^2} (\mu - \mu_x)^2 \right\} \end{aligned}$$

with:

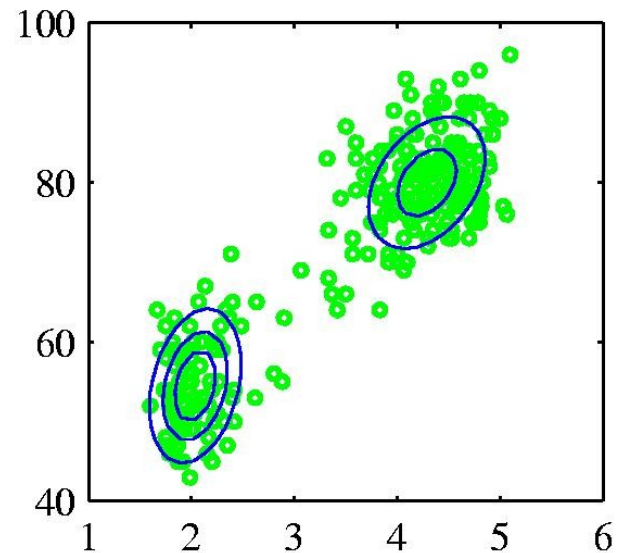
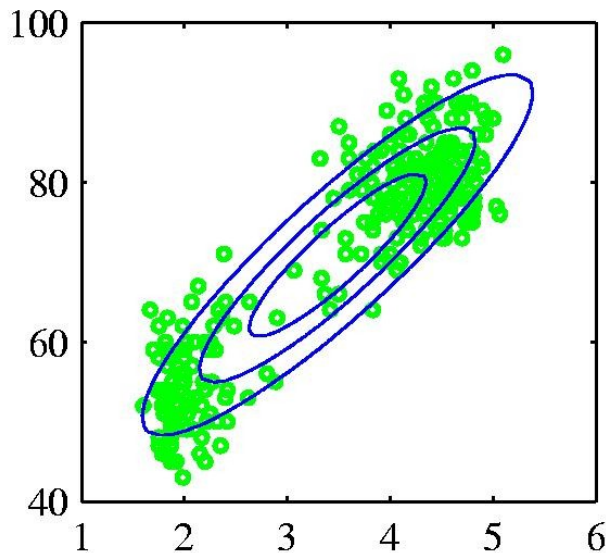
$$\mu_x = \left(\frac{\mu_0}{\sigma_0^2} + \frac{x}{\sigma^2} \right) / \left(\frac{1}{\sigma_0^2} + \frac{1}{\sigma^2} \right) \quad \sigma_x^2 = \left(\frac{1}{\sigma_0^2} + \frac{1}{\sigma^2} \right)^{-1}$$

Mixtures of Gaussians

A Gaussian has a single mode (peak) and cannot model multi-modal distributions.

Instead, we can use *mixtures*:

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



Mixtures of Gaussians

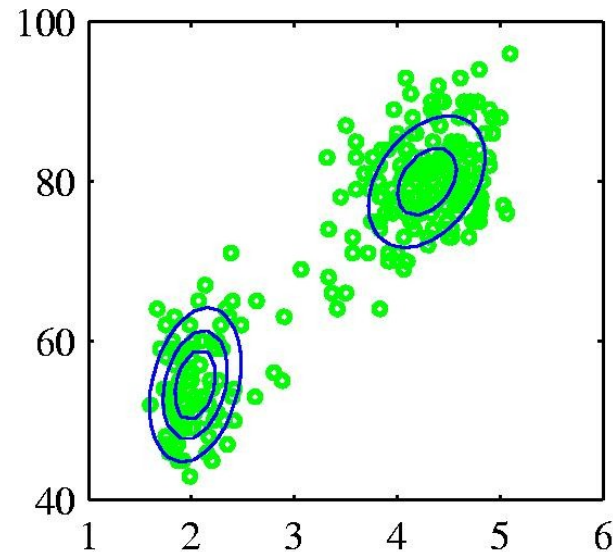
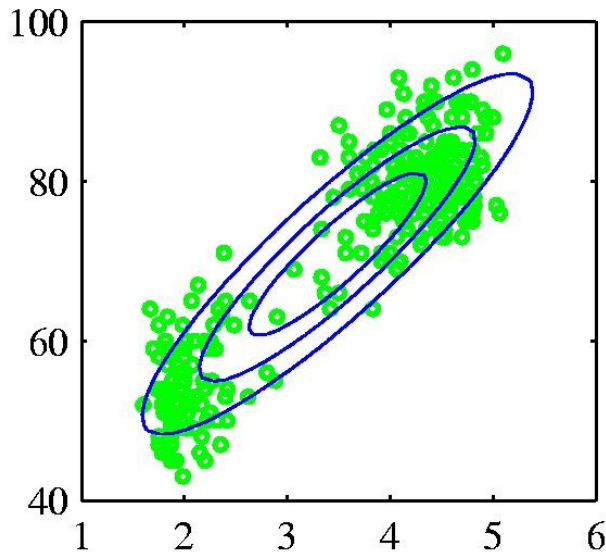
A Gaussian has a single mode (peak) and cannot model multi-modal distributions.

Instead, we can use *mixtures*:

$$p(\mathbf{x}) = \sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{x} \mid \mu_k, \Sigma_k)$$

Prob. for selecting given Gaussian

Conditional distribution



Mixtures of Gaussians

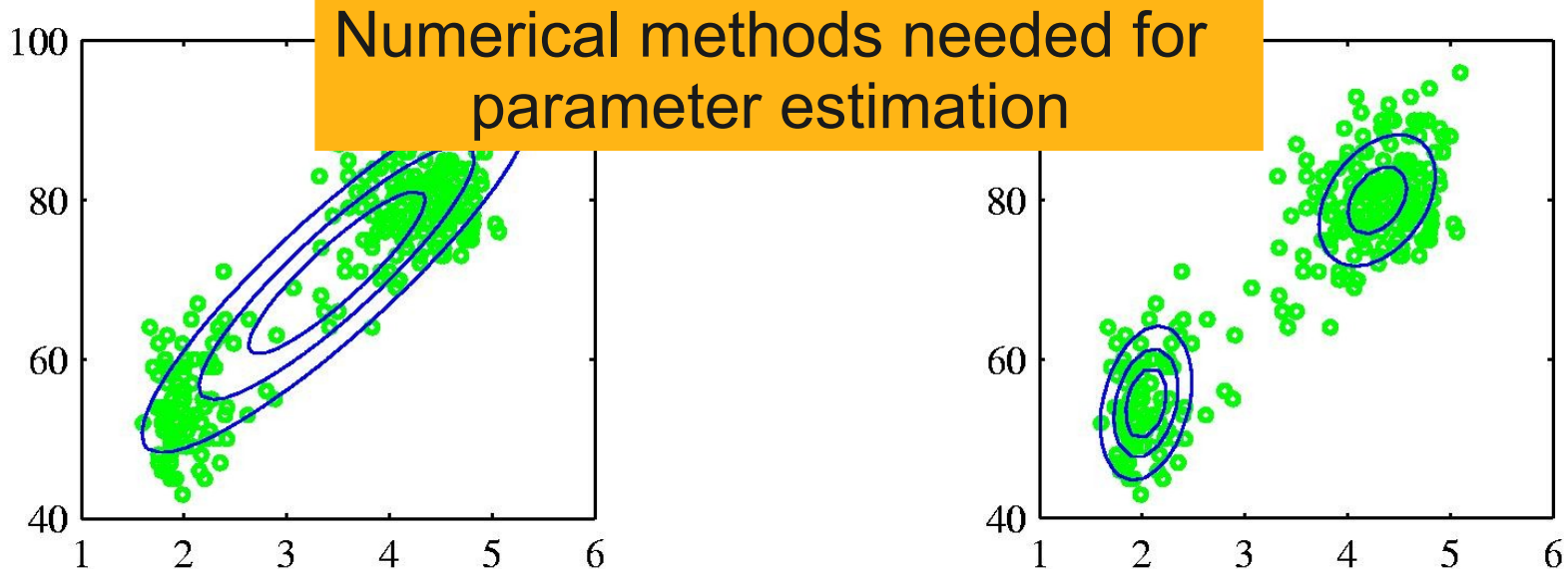
A Gaussian has a single mode (peak) and cannot model multi-modal distributions.

Instead, we can use *mixtures*:

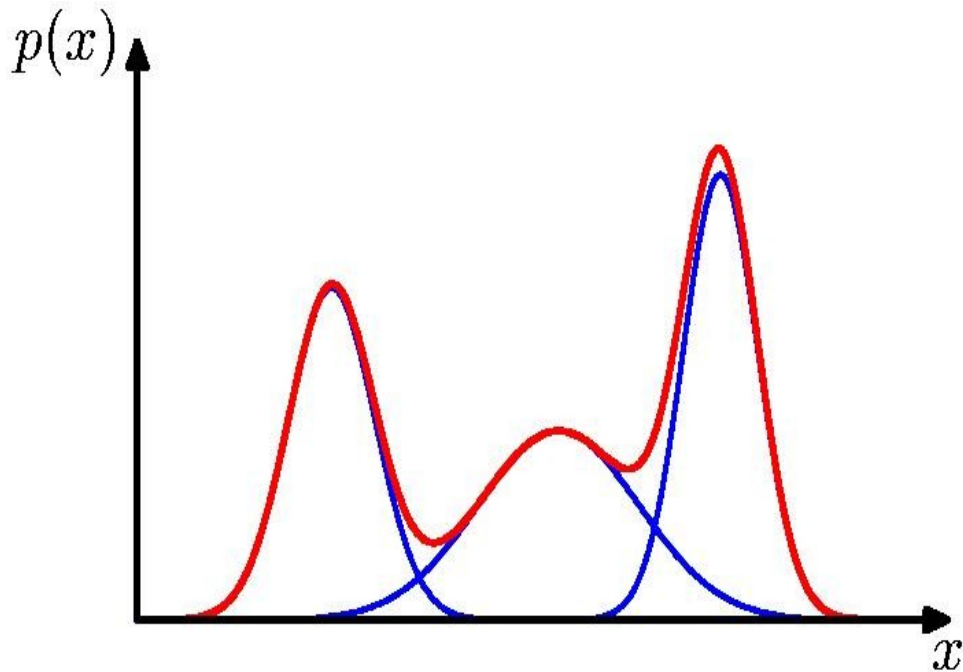
$$p(\mathbf{x}) = \sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{x} \mid \mu_k, \Sigma_k)$$

Prob. for selecting given Gaussian

Conditional distribution



Summary



- Bernoulli and Binomial, with Beta prior
- Multinomial with Dirichlet prior
- Gaussian with Gaussian prior
- Mixtures of Gaussians