Probability Distributions

Chia-Ping Chen

Professor
National Sun Yat-sen University
Department of Computer Science and Engineering

Machine Learning

Topics

We introduce the distributions of a few random variables. Then we learn how to estimate such distributions from data.

- Binary Variables
- K-ary Variables
- Gaussian Variables
- Exponential Family
- Non-parametric Methods for Density Estimation

Binary Variables

Definition. Let X be a random variable. X is a **binary random variable** if it has exactly 2 possible values. In particular, X is a **Bernoulli random variable** if the possible values are 0 and 1. The distribution of a Bernoulli random variable is a **binary distribution function** given by

$$p(x) = \text{Bern}(x|\mu) = \mu^x(1-\mu)^{1-x}, \ x = 0, 1$$

It is clear that

$$p(1) = \mu, \ p(0) = 1 - \mu$$

and

$$\mathbb{E}(X) = \mu$$
$$\text{var}[X] = \mu(1-\mu)$$

Let X be a Bernoulli random variable with $p(x) = \text{Bern}(x|\mu)$. What is the **maximum-likelihood estimate** of μ based on $\mathcal{D} = \{x_1, \ldots, x_N\}$?

The likelihood of $\mathcal{D}$ as a function of μ is

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

$$\Rightarrow \quad \log p(\mathcal{D}|\mu) = \sum_{n=1}^{N} [x_n \log \mu + (1-x_n)\log(1-\mu)]$$

Setting the derivative with respect to μ to 0, we get

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

In **Bayesian learning**, we treat μ as a random variable.

1. We assume a prior distribution $p(\mu)$ of μ.
2. We update the distribution to $p(\mu|\mathcal{D})$ with data $\mathcal{D}$.
3. We obtain approximate $p(x)$ via $p(\mu|\mathcal{D})$ (point estimates or integration).

Benefit: Consider the density estimation of a Bernoulli random variable X based on $\mathcal{D} = \{x_1, \ldots, x_N\}$. When N is not large enough, the maximum likelihood estimation of μ is prone to over-fitting. The issue can be alleviated by Bayesian learning.

Definition. Consider Bayesian learning of a Bernoulli random variable X with $p(x) = \text{Bern}(x|\mu)$ based on $\mathcal{D}$. A **conjugate prior** $p(\mu)$ of μ makes the posterior $p(\mu|\mathcal{D})$ have the same functional form as $p(\mu)$.

According to the Bayes' rule

$$p(\mu|\mathcal{D}) \propto p(\mu)p(\mathcal{D}|\mu)$$

Hence, the problem we face here is to make $p(\mu)$ and $p(\mu|\mathcal{D})$ belong to the same family, given data likelihood $p(\mathcal{D}|\mu)$.

We can re-write the likelihood function as

$$\begin{aligned} p(\mathcal{D}|\mu) &= \prod_{n=1}^{N} \mu^{x_n}(1-\mu)^{1-x_n} \\ &= \mu^{\left(\sum_{n=1}^{N} x_n\right)}(1-\mu)^{\left(\sum_{n=1}^{N}(1-x_n)\right)} \\ &= \mu^{m}(1-\mu)^{N-m} \end{aligned}$$

where $m = \sum_{n=1}^{N} x_n$ is the number of 1s in $\mathcal{D}$.

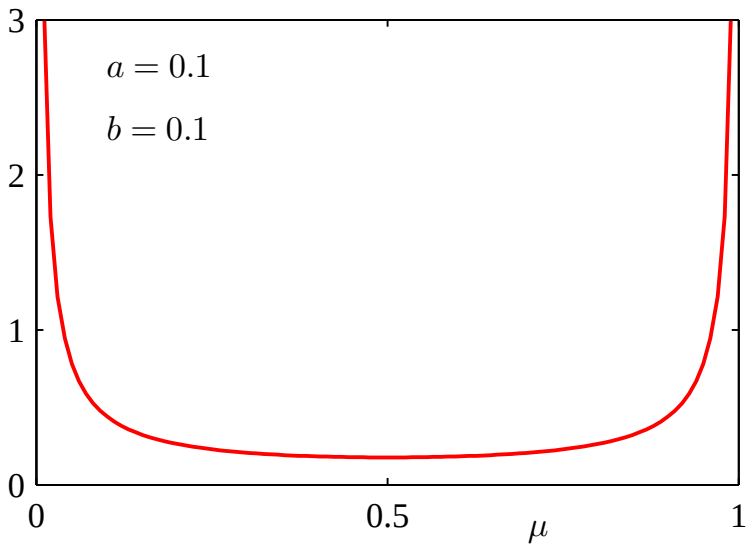
As a function of μ, the likelihood $p(\mathcal{D}|\mu)$ is proportional to powers of μ and $(1-\mu)$. Thus, if we let the prior $p(\mu)$ be proportional to powers of μ and $(1-\mu)$, then the posterior $p(\mu|\mathcal{D}) \propto p(\mu)p(\mathcal{D}|\mu)$ will also be proportional to powers of μ and $(1-\mu)$.

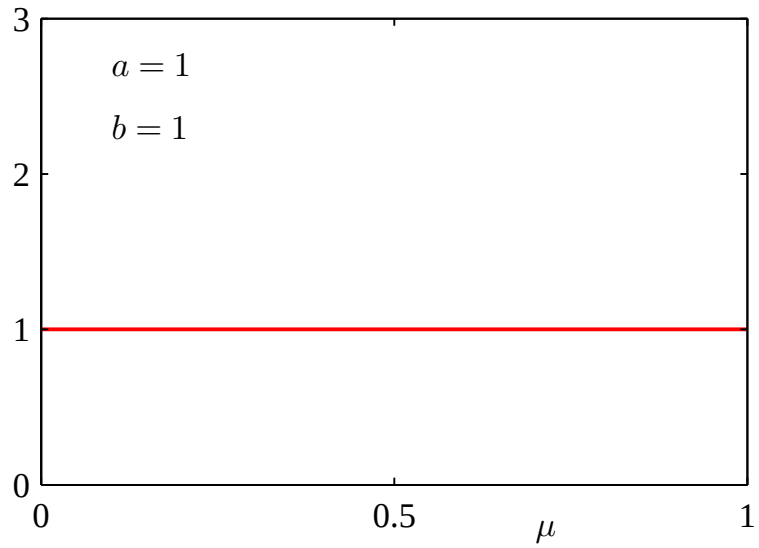
Beta Distribution

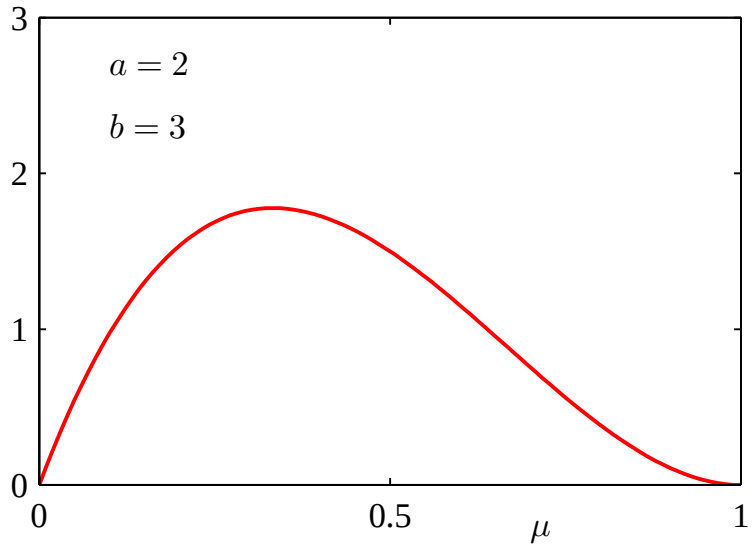
Definition. A **beta distribution** is

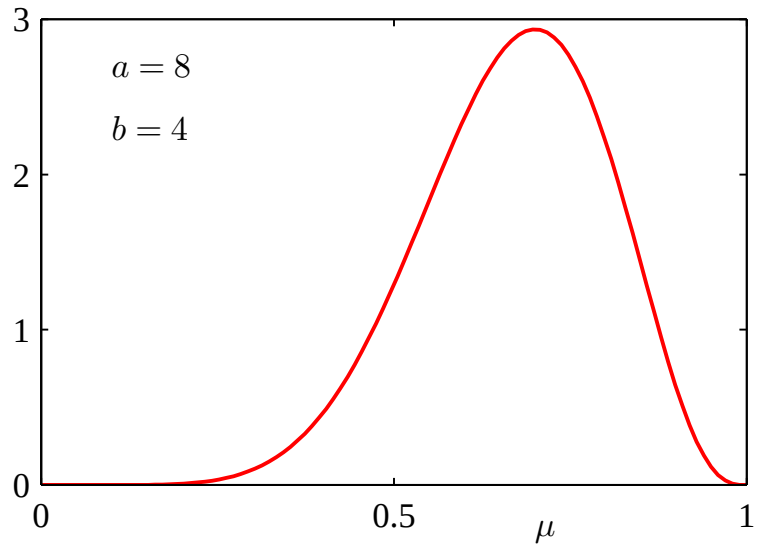
$$p(\mu) = \text{Beta}(\mu|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1},\ 0 \leq \mu \leq 1$$

The parameters in the distribution of a parameter, here a and b of $p(\mu)$, are **hyperparameters**.


$a = 0.1$
$b = 0.1$
3
2
1
0
0
0.5
1
μ


$a = 1$
$b = 1$
3
2
1
0
0
0.5
1
μ


$a = 2$
$b = 3$
3
2
1
0
0
0.5
1
μ

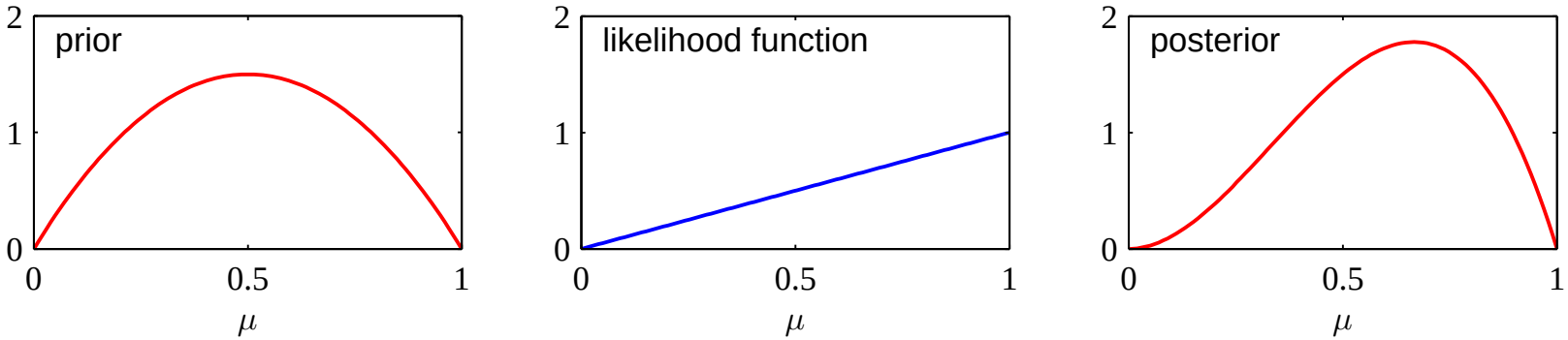

$a = 8$
$b = 4$
3
2
1
0
0
0.5
1
μ

Beta Prior and Posterior

Let X be a Bernoulli random variable with distribution $p(x) = \text{Bern}(x|\mu)$. Then a beta distribution $p(\mu) = \text{Beta}(\mu|a, b)$ is a conjugate prior for μ.

Let m be the number of 1s and $l = N - m$ be the number of 0s in the data set $\mathcal{D} = \{x_1, \ldots, x_N\}$. Then

$$\begin{aligned} p(\mu|\mathcal{D}) &\propto p(\mu)p(\mathcal{D}|\mu) \\ &= \text{Beta}(\mu|a, b)p(\mathcal{D}|\mu) \\ &\propto \mu^{a-1}(1-\mu)^{b-1}\mu^m(1-\mu)^l \\ &\propto \mu^{a+m-1}(1-\mu)^{b+l-1} \\ \Rightarrow \quad p(\mu|\mathcal{D}) &= \frac{\Gamma(a+m+b+l)}{\Gamma(a+m)\Gamma(b+l)}\mu^{a+m-1}(1-\mu)^{b+l-1} \\ &= \text{Beta}(\mu|a+m, b+l) \end{aligned}$$



One step of sequential Bayesian inference. $\mu \sim \text{Beta}(\mu|a, b)$ from $a = 2, b = 2$ to $a = 3, b = 2$ with a single observation of $x = 1$.

Let X be a Bernoulli random variable with $p(x) = \text{Bern}(x|\mu)$. We see

$$p(\mu) = \text{Beta}(\mu|a, b)$$
$$\downarrow$$
$$\downarrow \text{ Bayesian learning}$$
$$\downarrow$$
$$p(\mu|\mathcal{D}) = \text{Beta}(\mu|a + m, b + l)$$

The hyperparameters a and b in the prior distribution $p(\mu)$ can be interpreted as the **effective numbers of observations** for $x = 1$ and $x = 0$ prior to any observation.

Let X be a Bernoulli random variable with $p(x) = \text{Bern}(x|\mu)$. When μ is treated as a random variable, the distribution of X is the integration over the distribution of μ (the sum rule).

$$
\begin{aligned}
P(X = 1|\mathcal{D}) &= \int_0^1 P(X = 1, \mu|\mathcal{D}) d\mu \\
&= \int_0^1 P(X = 1|\mu) p(\mu|\mathcal{D}) d\mu \\
&= \int_0^1 \mu \, p(\mu|\mathcal{D}) d\mu \\
&= \mathbb{E}[\mu|\mathcal{D}] \\
&= \frac{a + m}{a + m + b + l}
\end{aligned}
$$

On average, data observation reduces parameter uncertainty.

By the total variance theorem of probability theory

$$\text{var}_{\boldsymbol{\theta}}(\boldsymbol{\theta}) = \mathbb{E}_{\mathcal{D}}[\text{var}_{\boldsymbol{\theta}}(\boldsymbol{\theta}|\mathcal{D})] + \text{var}_{\mathcal{D}}(\mathbb{E}_{\boldsymbol{\theta}}[\boldsymbol{\theta}|\mathcal{D}])$$

It follows from $\text{var}_{\mathcal{D}}(\mathbb{E}_{\boldsymbol{\theta}}[\boldsymbol{\theta}|\mathcal{D}]) \geq 0$ that

$$\text{var}_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \geq \mathbb{E}_{\mathcal{D}}[\text{var}_{\boldsymbol{\theta}}(\boldsymbol{\theta}|\mathcal{D})]$$

K-ary Variables

1-of-K Representation

Definition. Let X be a discrete random variable. X is a K-**ary random variable** if it has exactly K possible values.

A value (a.k.a. **state**) of X can be represented by a vector of size K, i.e. $\boldsymbol{x} = (x_1, \ldots, x_K)^T$, called 1-of-$K$ (a.k.a. 1-hot) representation: one component is 1 for state identity, and the other components are 0. The set of possible values are $\{\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(K)}\}$ with

$$x_j^{(k)} = \delta_{kj}$$

For example

$$\boldsymbol{x}^{(3)} = (0, 0, 1, 0, 0, 0)^T$$

We also denote a K-ary random variable by $\mathbf{x}$ since the values are represented by vectors.

K-ary Distribution

Definition. Let $\mathbf{x}$ be a K-ary random variable. The distribution of $\mathbf{x}$ is a K-**ary distribution**

$$P\left(\mathbf{x}=\boldsymbol{x}^{(j)}\right)=p\left(\boldsymbol{x}^{(j)}|\boldsymbol{\mu}\right)=\mu_j,\ j=1,\ldots,K$$

where $\boldsymbol{\mu}=(\mu_1,\ldots,\mu_K)^T$ is a vector of parameters.

The parameters $\boldsymbol{\mu}$ must satisfy

$$\mu_k\geq 0,\ \sum_{k=1}^{K}\mu_k=1$$

The expectation of $\mathbf{x}$ is

$$\mathbb{E}[\mathbf{x}|\boldsymbol{\mu}]=\sum_{k=1}^{K}\boldsymbol{x}^{(k)}\,p\left(\boldsymbol{x}^{(k)}|\boldsymbol{\mu}\right)=(\mu_1,\ldots,\mu_K)^T=\boldsymbol{\mu}$$

Let $\mathbf{x}$ be a K-ary random variable with K-ary distribution $p\left(\boldsymbol{x}^{(j)}|\boldsymbol{\mu}\right)$, and $\mathcal{D} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ be observations of $\mathbf{x}$. Using 1-of-K representation, the likelihood of a data point $\boldsymbol{x}_n$ can be written as

$$p(\boldsymbol{x}_n|\boldsymbol{\mu}) = \prod_{k=1}^{K} \mu_k^{x_{nk}}$$

So the likelihood of $\mathcal{D}$ is

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

where m_k is the number of points in $\mathcal{D}$ with $\boldsymbol{x}_n = \boldsymbol{x}^{(k)}$ or $x_{nk} = 1$.

The log data likelihood is

$$\log p(\mathcal{D}|\boldsymbol{\mu}) = \sum_{k=1}^{K} m_k \log \mu_k$$

Here the parameters $\mu_1, \ldots, \mu_K$ are not independent, so we need to maximize the **Lagrangian**

$$L(\boldsymbol{\mu}, \lambda) = \sum_{k=1}^{K} m_k \log \mu_k + \lambda \left(\sum_{k=1}^{K} \mu_k - 1 \right)$$

Setting the derivative of $L(\boldsymbol{\mu}, \lambda)$ with respect to $\boldsymbol{\mu}$ to 0, we get

$$\mu_k^{\mathsf{ML}} = \frac{m_k}{N}$$

In Bayesian learning, we start with a distribution over the parameters and update the distribution with data. Again, we use conjugate prior so the posterior is an update of the hyperparameters with data.

This dependency of the data likelihood function on $\boldsymbol{\mu}$ decides the conjugate prior of $\boldsymbol{\mu}$. In the case of a K-ary random variable, the data likelihood function depends on powers of μ_k

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

Definition. A **Dirichlet distribution** is

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

where

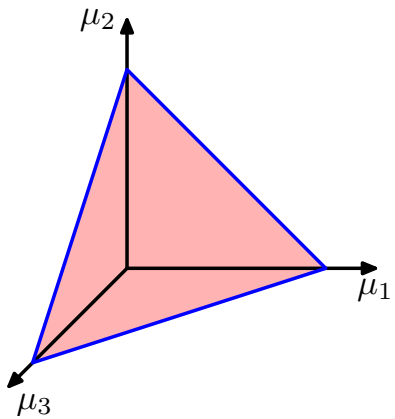
$$0 \leq \mu_k \leq 1, \ \sum_{k=1}^{K} \mu_k = 1$$

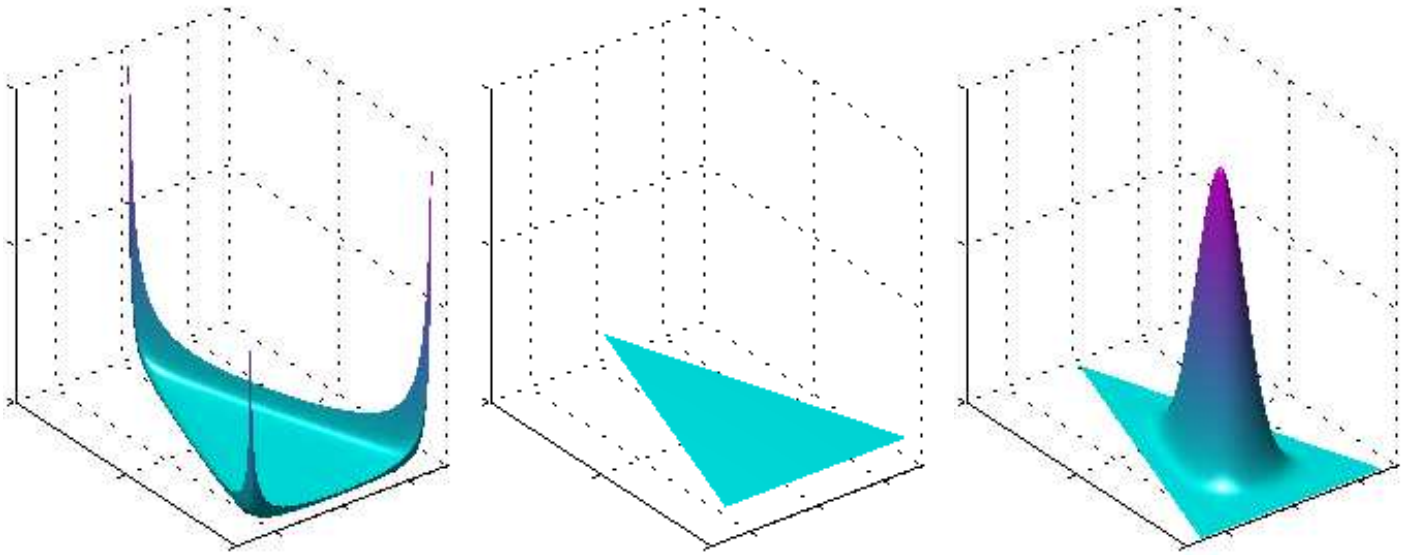
Here $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)^T$ are hyperparameters, and

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

Note $\text{Dir}(\boldsymbol{\mu}|\boldsymbol{\alpha})$ depends on powers of μ_k.

Dirichlet: Support and Density





Dirichlet distributions for $\alpha_k = 0.1, 1, 10$, respectively

Let $\mathbf{x}$ be a K-ary random variable with $p\left(\boldsymbol{x}^{(j)}|\boldsymbol{\mu}\right)$. Then a Dirichlet distribution $\mathrm{Dir}(\boldsymbol{\mu}|\boldsymbol{\alpha})$ is a conjugate prior for $\boldsymbol{\mu}$.

Based on $\mathcal{D}$, the posterior distribution of $\boldsymbol{\mu}$ is

$$p(\boldsymbol{\mu}|\mathcal{D}) \propto p(\boldsymbol{\mu})p(\mathcal{D}|\boldsymbol{\mu}) = \mathrm{Dir}(\boldsymbol{\mu}|\boldsymbol{\alpha})p(\mathcal{D}|\boldsymbol{\mu}) \propto \prod_{k=1}^{K} \mu_k^{\alpha_k+m_k-1}$$

By normalization

$$p(\boldsymbol{\mu}|\mathcal{D}) = \frac{\Gamma(\alpha_0+N)}{\Gamma(\alpha_1+m_1)\ldots\Gamma(\alpha_K+m_K)} \prod_{k=1}^{K} \mu_k^{\alpha_k+m_k-1} = \mathrm{Dir}(\boldsymbol{\mu}|\boldsymbol{\alpha}')$$

where $\alpha'_k = \alpha_k + m_k$.

Gaussian Variables

Definition. A **Gaussian distribution** or **Gaussian PDF** is

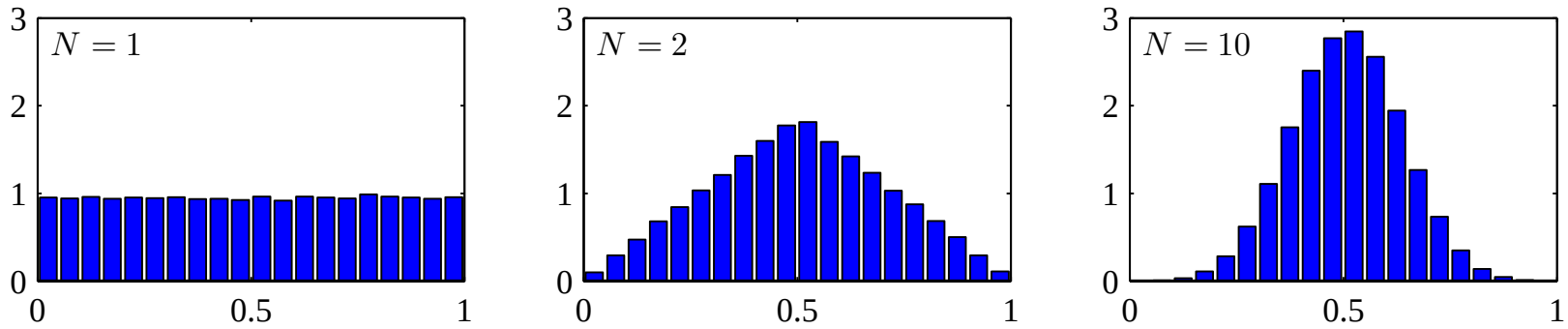
$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$

It is governed by parameters μ and σ^2.

A random variable with a Gaussian distribution is a Gaussian random variable. Let X be a Gaussian random variable with distribution $p(x) = \mathcal{N}(x|\mu,\sigma^2)$. Then it can be shown that

$$\mu = \mathbb{E}[X], \quad \sigma^2 = \text{var}(X)$$

The mean (or the sum) of a set of **i.i.d.** random variables has a distribution that becomes increasingly Gaussian as the number of terms in the sum increases.



Histogram plots of the mean of N uniform random variables in $[0, 1]$ for $N = 1, 2, 10$ respectively.

Definition. A **multi-variate Gaussian distribution** is

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

where D is the dimension of $\boldsymbol{x}$. It is governed by parameter vector $\boldsymbol{\mu}$ and parameter matrix $\boldsymbol{\Sigma}$.

A random vector with a multi-variate Gaussian distribution is a Gaussian random vector. Let $\mathbf{x}$ be a Gaussian random vector with distribution $p(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma})$. Then it can be shown that

$$\int p(\boldsymbol{x})d\boldsymbol{x} = 1,\ \mathbb{E}[\mathbf{x}] = \boldsymbol{\mu},\ \mathbb{E}\left[\mathbf{x}\mathbf{x}^T\right] = \boldsymbol{\mu}\boldsymbol{\mu}^T + \boldsymbol{\Sigma},\ \text{var}(\mathbf{x}) = \boldsymbol{\Sigma}$$

Definition. Let $\mathbf{x}$ be a Gaussian random vector of dimension D with $p(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$. The dependence of the PDF on $\boldsymbol{x}$ is through quadratic form

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

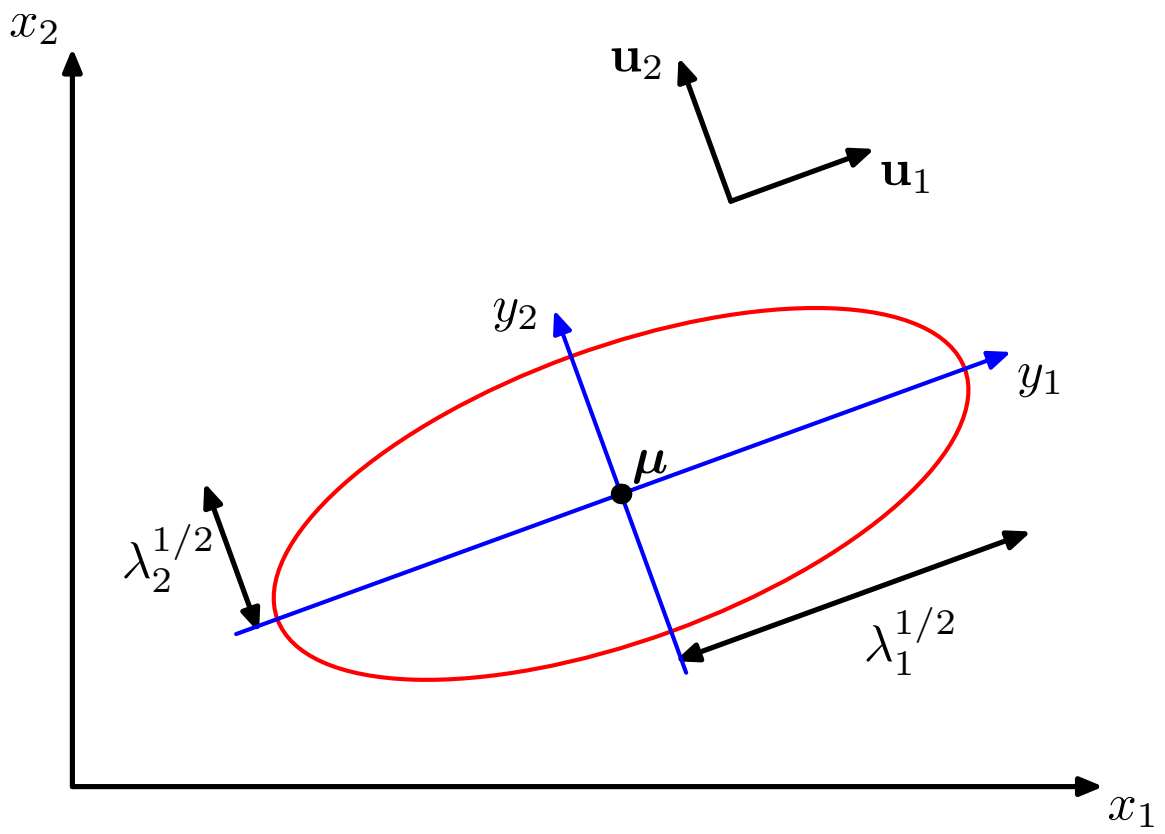
Δ is the **Mahalanobis distance** from $\boldsymbol{\mu}$ to $\boldsymbol{x}$.

Let $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_D$ be orthonormal eigenvectors of $\boldsymbol{\Sigma}$ with eigenvalues $\lambda_1 \leq \cdots \leq \lambda_D$. Then $\boldsymbol{\Sigma} = \sum_{i=1}^{D} \lambda_i \boldsymbol{u}_i \boldsymbol{u}_i^T$, $\boldsymbol{\Sigma}^{-1} = \sum_{i=1}^{D} \lambda_i^{-1} \boldsymbol{u}_i \boldsymbol{u}_i^T$, and

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

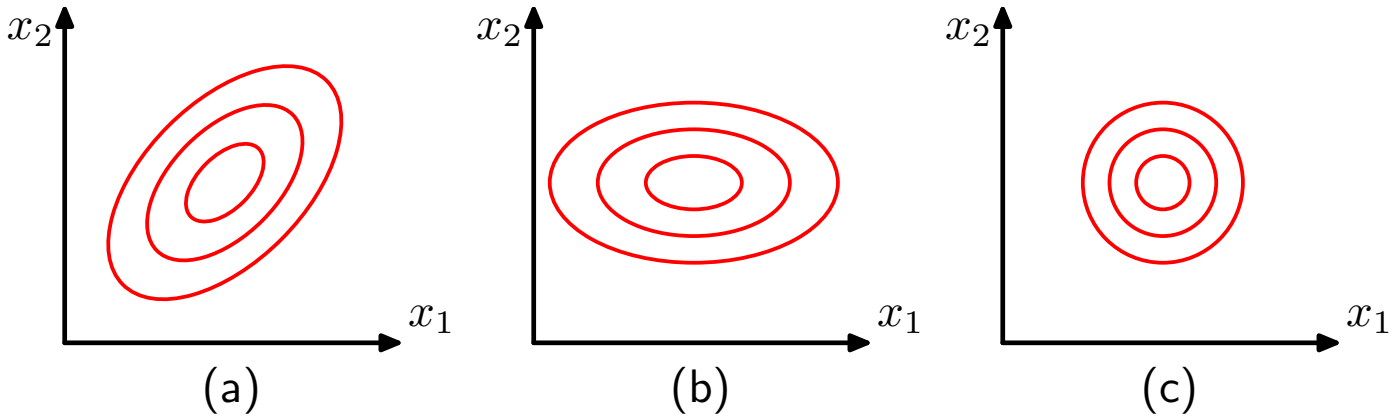
where $y_i = \boldsymbol{u}_i^T (\boldsymbol{x} - \boldsymbol{\mu})$ and $(\boldsymbol{x} - \boldsymbol{\mu}) = \sum_{i=1}^{D} y_i \boldsymbol{u}_i$.

Contour of a Gaussian PDF



The contour of a 2-D Gaussian PDF, on which $\Delta^2 = 1$ so the density is $e^{-1/2}$ of the value at $\boldsymbol{x} = \boldsymbol{\mu}$.

Examples



Examples 2-D Gaussian PDF with general, diagonal, and isotropic covariance matrix, respectively.

Let $\mathbf{x}$ be a Gaussian random vector with $p(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Then the expectation of $\mathbf{x}$ is $\boldsymbol{\mu}$.

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

where $\boldsymbol{z} = \boldsymbol{x} - \boldsymbol{\mu}$.

Let $\mathbf{x}$ be a Gaussian random vector with $p(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Then the expectation of $\mathbf{x}\mathbf{x}^{\mathbf{T}}$ is $\boldsymbol{\mu}\boldsymbol{\mu}^T + \boldsymbol{\Sigma}$.

$$
\begin{aligned}
\mathbb{E}\left[\mathbf{x}\mathbf{x}^T\right] &= \frac{1}{(2\pi)^{\frac{D}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}} \int \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\}\boldsymbol{x}\boldsymbol{x}^T d\boldsymbol{x} \\
&= \frac{1}{(2\pi)^{\frac{D}{2}}}\frac{1}{|\boldsymbol{\Sigma}|^{\frac{1}{2}}} \int \exp\left\{-\frac{1}{2}\boldsymbol{z}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{z}\right\}(\boldsymbol{z}+\boldsymbol{\mu})(\boldsymbol{z}+\boldsymbol{\mu})^T d\boldsymbol{z} \\
&= \frac{1}{(2\pi)^{\frac{D}{2}}}\frac{1}{|\boldsymbol{\Sigma}|^{\frac{1}{2}}} \int \exp\left\{-\frac{1}{2}\boldsymbol{z}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{z}\right\}(\boldsymbol{z}\boldsymbol{z}^T+\boldsymbol{\mu}\boldsymbol{\mu}^T) d\boldsymbol{z} \\
&= \frac{1}{(2\pi)^{\frac{D}{2}}}\frac{1}{|\boldsymbol{\Sigma}|^{\frac{1}{2}}} \int \exp\left\{-\frac{1}{2}\boldsymbol{z}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{z}\right\}\boldsymbol{z}\boldsymbol{z}^T d\boldsymbol{z} + \boldsymbol{\mu}\boldsymbol{\mu}^T \\
&= \sum_{i=1}^{D} \boldsymbol{u}_i\boldsymbol{u}_i^T\lambda_i + \boldsymbol{\mu}\boldsymbol{\mu}^T = \boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^T
\end{aligned}
$$

Recall that $\boldsymbol{z} = \boldsymbol{x} - \boldsymbol{\mu} = \sum y_i \boldsymbol{u}_i$.

$$\begin{aligned}
&\frac{1}{(2\pi)^{\frac{D}{2}}} \frac{1}{|\boldsymbol{\Sigma}|^{\frac{1}{2}}} \int \exp\left\{-\frac{1}{2}\boldsymbol{z}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{z}\right\} \boldsymbol{z}\boldsymbol{z}^T d\boldsymbol{z} \\
= &\sum_{i=1}^{D}\sum_{j=1}^{D} \int \frac{1}{(2\pi)^{\frac{D}{2}}} \frac{1}{|\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}\sum_{k=1}^{D}\frac{y_k^2}{\lambda_k}\right\} y_i y_j \boldsymbol{u}_i \boldsymbol{u}_j^T d\boldsymbol{y} \\
= &\sum_{i=1}^{D}\sum_{j=1}^{D} \boldsymbol{u}_i \boldsymbol{u}_j^T \int \frac{1}{(2\pi)^{\frac{D}{2}}} \frac{1}{|\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}\sum_{k=1}^{D}\frac{y_k^2}{\lambda_k}\right\} y_i y_j d\boldsymbol{y} \\
= &\sum_{i=1}^{D} \lambda_i \boldsymbol{u}_i \boldsymbol{u}_i^T = \boldsymbol{\Sigma}
\end{aligned}$$

where

$$\int \frac{1}{(2\pi)^{\frac{D}{2}}} \frac{1}{|\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}\sum_{k=1}^{D}\frac{y_k^2}{\lambda_k}\right\} y_i y_j d\boldsymbol{y} = \begin{cases} 0, & j \neq i \\ \lambda_i, & j = i \end{cases}$$

Covariance Matrix

Let $\mathbf{x}$ be a Gaussian random vector with $p(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Then the covariance matrix of $\mathbf{x}$ is $\boldsymbol{\Sigma}$.

$$\begin{aligned}
\text{cov}[\mathbf{x}] &= \mathbb{E}\left[(\mathbf{x} - \mathbb{E}[\mathbf{x}])(\mathbf{x} - \mathbb{E}[\mathbf{x}])^T\right] \\
&= \mathbb{E}\left[\mathbf{x}\mathbf{x}^T\right] - \mathbb{E}\left[\mathbf{x}\right]\mathbb{E}\left[\mathbf{x}^T\right] \\
&= \boldsymbol{\Sigma}
\end{aligned}$$

Let random vectors $\mathbf{x}_a$ and $\mathbf{x}_b$ be joint Gaussian. Then $\mathbf{x}_a$ is conditional Gaussian given $\mathbf{x}_b$.

Let the PDF of $\mathbf{x} = (\mathbf{x}_a, \mathbf{x}_b)^T$ be $p(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Partition $\mathbf{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma}$ and the precision matrix $\boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1}$ as follows

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

It can be shown that the conditional distribution of $\mathbf{x}_a$ given $\mathbf{x}_b$ is $p(\boldsymbol{x}_a|\boldsymbol{x}_b) = \mathcal{N}(\boldsymbol{x}_a|\boldsymbol{\mu}_{a|b}, \boldsymbol{\Sigma}_{a|b})$ where

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

Note $\boldsymbol{\Lambda}_{a|b} = \boldsymbol{\Sigma}_{a|b}^{-1} = \boldsymbol{\Lambda}_{aa}$.

Let random vectors $\mathbf{x}_a$ and $\mathbf{x}_b$ be joint Gaussian. Then $\mathbf{x}_a$ (and $\mathbf{x}_b$) is Gaussian.

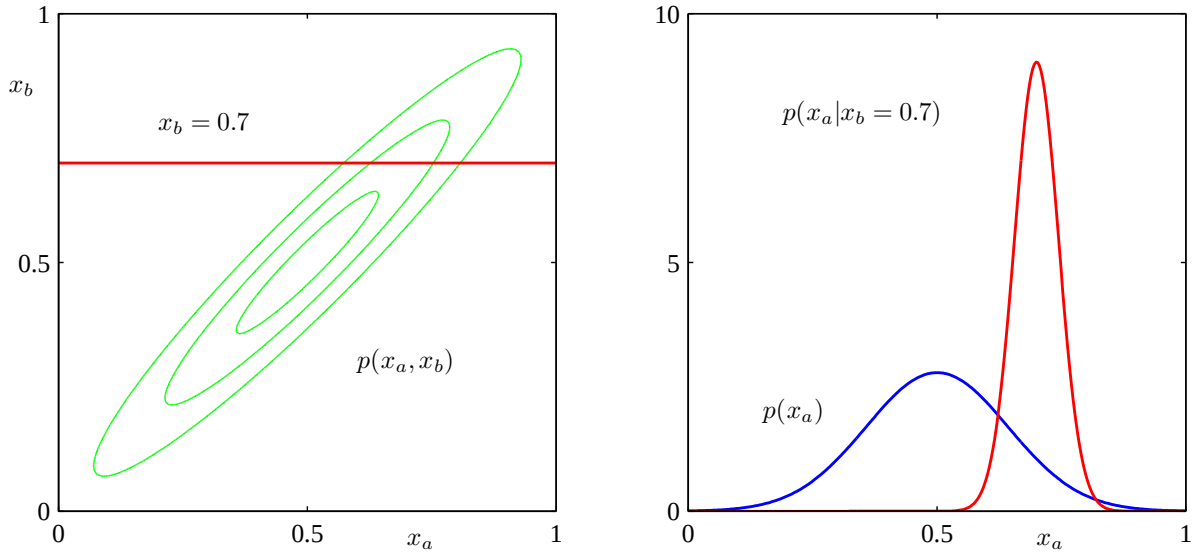
Let the PDF of $\mathbf{x} = (\mathbf{x}_a, \mathbf{x}_b)^T$ be $p(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Partition $\mathbf{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma}$ and the precision matrix $\boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1}$ as follows

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

It can be shown that the distribution of $\mathbf{x}_a$ is $p(\boldsymbol{x}_a) = \mathcal{N}(\boldsymbol{x}_a|\boldsymbol{\mu}_{a|\emptyset}, \boldsymbol{\Sigma}_a)$ where

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

Conditional Gaussian



Left: Contours of joint Gaussian PDF.
Right: A marginal and a conditional Gaussian PDF.

Let $\mathbf{y}$ be Gaussian and $\mathbf{z}$ be conditional Gaussian given $\mathbf{y}$. Then $\mathbf{y}$ and $\mathbf{z}$ are joint Gaussian.

Let the PDF of $\mathbf{y}$ be $p(\boldsymbol{y}) = \mathcal{N}(\boldsymbol{y}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1})$ and the conditional PDF of $\mathbf{z}$ given $\mathbf{y}$ be $p(\boldsymbol{z}|\boldsymbol{y}) = \mathcal{N}(\boldsymbol{z}|\boldsymbol{A}\boldsymbol{y} + \boldsymbol{b}, \boldsymbol{L}^{-1})$. Then the joint PDF of $\mathbf{x} = (\mathbf{y}, \mathbf{z})^T$ is

$$
\begin{aligned}
p(\boldsymbol{x}) &= p(\boldsymbol{y})p(\boldsymbol{z}|\boldsymbol{y}) \\
&= \mathcal{N}\left(\begin{pmatrix}\boldsymbol{y}\\ \boldsymbol{z}\end{pmatrix} \middle| \begin{pmatrix}\boldsymbol{\mu}\\ \boldsymbol{A}\boldsymbol{\mu}+\boldsymbol{b}\end{pmatrix}, \begin{pmatrix}\boldsymbol{\Lambda}+\boldsymbol{A}^T\boldsymbol{L}\boldsymbol{A} & -\boldsymbol{A}^T\boldsymbol{L}\\ -\boldsymbol{L}\boldsymbol{A} & \boldsymbol{L}\end{pmatrix}^{-1}\right)
\end{aligned}
$$

Let $\mathbf{y}$ be Gaussian and $\mathbf{z}$ be conditional Gaussian given $\mathbf{y}$. Since $\mathbf{y}$ and $\mathbf{z}$ are joint Gaussian, we have

- $\mathbf{z}$ is Gaussian with PDF
- $\mathbf{y}$ is conditional Gaussian given $\mathbf{z}$

The PDF of $\mathbf{z}$ is

$$p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}|\boldsymbol{A}\boldsymbol{\mu} + \boldsymbol{b}, \boldsymbol{L}^{-1} + \boldsymbol{A}\boldsymbol{\Lambda}^{-1}\boldsymbol{A}^T)$$

The conditional PDF of $\mathbf{y}$ given $\mathbf{z}$ is

$$p(\boldsymbol{y}|\boldsymbol{z}) = \mathcal{N}(\boldsymbol{y}|\boldsymbol{\Sigma}\{\boldsymbol{A}^T\boldsymbol{L}(\boldsymbol{z} - \boldsymbol{b}) + \boldsymbol{\Lambda}\boldsymbol{\mu}\}, \boldsymbol{\Sigma})$$

where $\boldsymbol{\Sigma} = (\boldsymbol{\Lambda} + \boldsymbol{A}^T\boldsymbol{L}\boldsymbol{A})^{-1}$.

Density Estimation for Gaussians

Let $\mathbf{x}$ be Gaussian with PDF $p(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$. In the **density estimation** of $\mathbf{x}$, we use a data set to estimate $p(\boldsymbol{x})$.

- Maximum likelihood estimate
- Bayesian learning

Let $\mathcal{D} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ be a data set of a Gaussian random vector $\mathbf{x}$. For data likelihood, we have

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

$$\Rightarrow \log p(\boldsymbol{x}_n|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = -\frac{D}{2}\log 2\pi - \frac{1}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}(\boldsymbol{x}_n - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_n - \boldsymbol{\mu})$$

$$\Rightarrow \log p(\mathcal{D}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \log p(\boldsymbol{x}_i|\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

$$= -\frac{ND}{2}\log\, 2\pi - \frac{N}{2}\log|\boldsymbol{\Sigma}|$$

$$- \frac{1}{2}\sum_{n=1}^{N}(\boldsymbol{x}_n - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_n - \boldsymbol{\mu})$$

Maximum likelihood estimate maximizes data likelihood.

- For $\boldsymbol{\mu}$

$$\nabla_{\boldsymbol{\mu}} \log p(\mathcal{D}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = 0 \;\Rightarrow\; \sum_{n=1}^{N} \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_n - \boldsymbol{\mu}) = 0$$

$$\Rightarrow\; \boldsymbol{\mu}_{\mathsf{ML}} = \frac{1}{N} \sum_{n=1}^{N} \boldsymbol{x}_n$$

- For $\boldsymbol{\Sigma}$, it can be shown that

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

Let x be a Gaussian random variable with PDF $p(x) = \mathcal{N}(x|\mu, \sigma^2)$, and $\mathcal{D} = \{x_1, \ldots, x_N\}$ be a data set. We consider the following scenarios of Bayesian learning of Gaussian distribution.

- Given σ^2 and a conjugate prior of μ
- Given μ and a conjugate prior of precision $\lambda = 1/\sigma^2$
- Given a conjugate prior of μ and λ

The likelihood of $\mathcal{D}$ as a function of μ is

$$p(\mathcal{D}|\mu) = \prod_{n=1}^{N} p(x_n|\mu)$$
$$= \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2\right\}$$

It is a log quadratic function of μ, so a conjugate prior of μ is log quadratic, i.e. Gaussian

$$p(\mu) = \mathcal{N}(\mu|\mu_0, \sigma_0^2)$$

According to the Bayes' rule

$$p(\mu|\mathcal{D}) \propto p(\mu)p(\mathcal{D}|\mu)$$

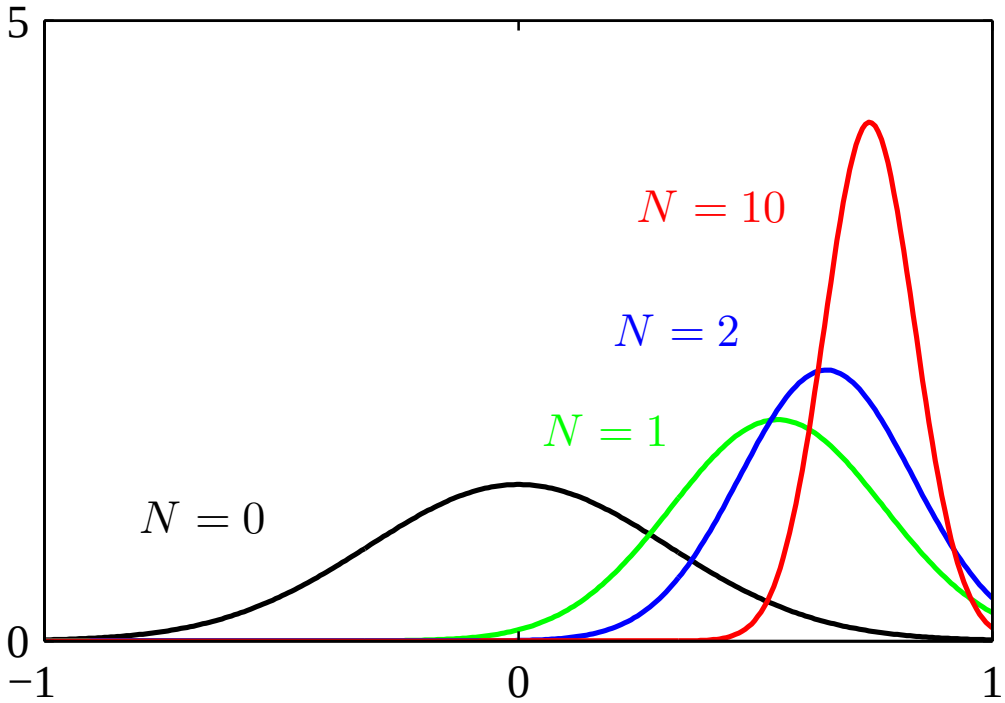
we have

$$p(\mu|\mathcal{D}) = \mathcal{N}(\mu|\mu_N, \sigma_N^2)$$

where

$$\mu_N = \frac{\sigma^2}{N\sigma_0^2 + \sigma^2}\mu_0 + \frac{N\sigma_0^2}{N\sigma_0^2 + \sigma^2}\mu_{\mathsf{ML}}$$

$$\frac{1}{\sigma_N^2} = \frac{1}{\sigma_0^2} + \frac{N}{\sigma^2}$$



Plots of $p(\mu|\mathcal{D})$ assuming Gaussian $p(\mu)$ and $p(x|\mu) = \mathcal{N}(x|\mu, \sigma^2)$.
The true $p(x)$ is Gaussian with mean 0.8 and variance 0.1.

The likelihood of $\mathcal{D}$ as a function of λ is

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

$$\propto \lambda^{N/2} \exp\left\{-\frac{\lambda}{2}\sum_{n=1}^{N}(x_n-\mu)^2\right\}$$

So a conjugate prior of λ is **gamma distribution**

$$p(\lambda) = \mathsf{Gam}(\lambda|a,b) = \frac{1}{\Gamma(a)} b^a \lambda^{a-1} e^{-b\lambda},\ \lambda \geq 0$$

According to the Bayes' rule

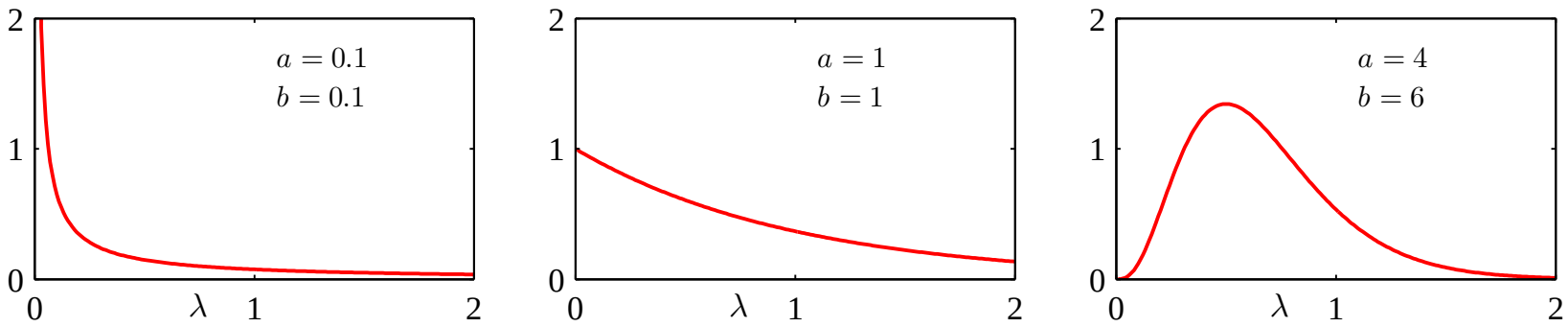
$$p(\lambda|\mathcal{D}) \propto p(\lambda)p(\mathcal{D}|\lambda)$$

we have

$$p(\lambda|\mathcal{D}) = \mathsf{Gam}\left(\lambda|a_N, b_N\right)$$

where

$$a_N = a + \frac{N}{2},\ b_N = b + \frac{N}{2}\sigma^2_{\mathsf{ML}}$$



The marginal distribution of x is

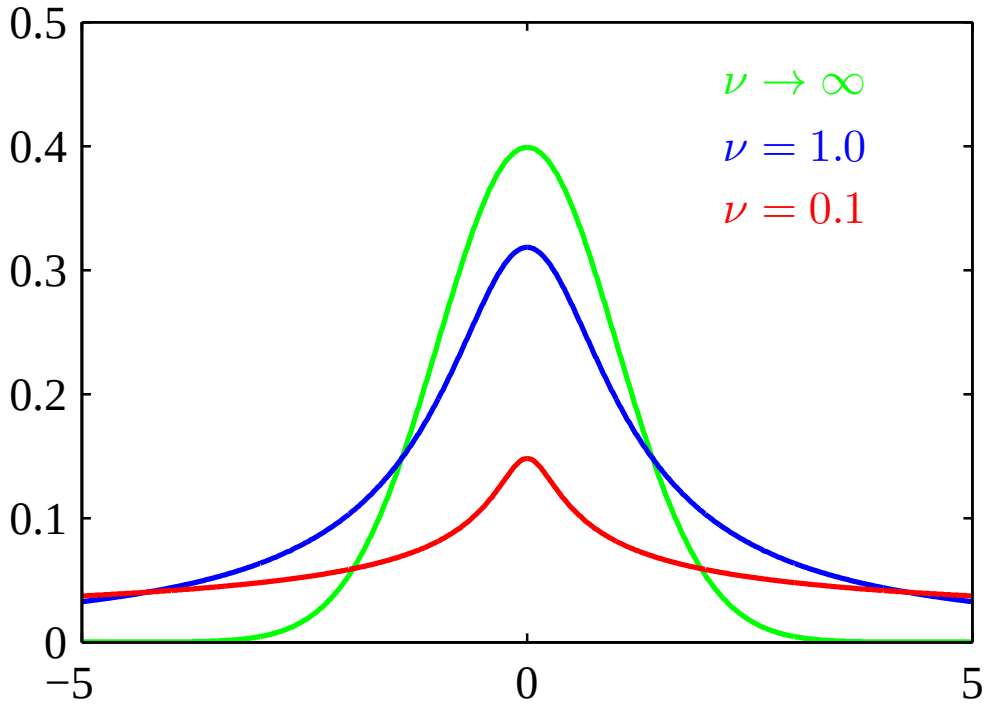
$$p(x|\mu, a, b) = \int_0^\infty \mathcal{N}(x|\mu, \lambda^{-1})\text{Gam}(\lambda|a, b)d\lambda$$
$$= \frac{b^a}{\Gamma(a)}\left(\frac{1}{2\pi}\right)^{1/2}\left[b + \frac{(x-\mu)^2}{2}\right]^{-a-1/2}\Gamma(a+1/2)$$

This is a **Student's t-distribution**

$$\text{St}(x|\mu, \lambda, \nu) = \frac{\Gamma(\nu/2+1/2)}{\Gamma(\nu/2)}\left(\frac{\lambda}{\pi\nu}\right)^{1/2}\left[1 + \frac{\lambda(x-\mu)^2}{\nu}\right]^{-\nu/2-1/2}$$

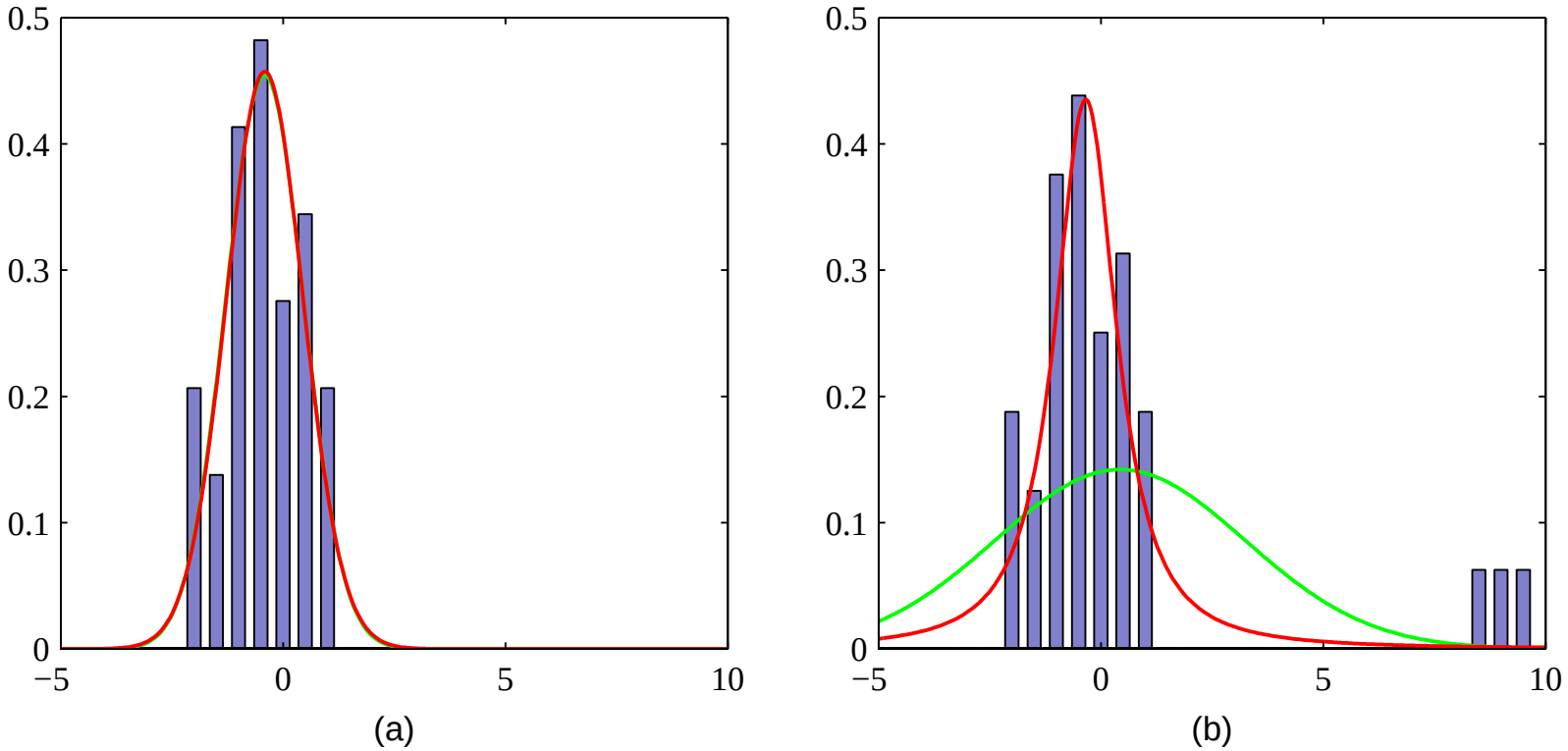
with parameters $\nu = 2a$ and $\lambda = a/b$.

Long Tail of Student's t-Distribution



Plots of Student's t-distribution $\mathrm{St}(x|\mu, \lambda, \nu)$ with 3 ν's, $\mu = 0$, and $\lambda = 1$.

Robustness to Outliers



The likelihood of $\mathcal{D}$ as a function of μ and λ is

$$
\begin{aligned}
p(\mathcal{D}|\mu,\lambda) &= \prod_{n=1}^{N} p(x_n|\mu,\lambda) \\
&= \prod_{n=1}^{N} \left(\frac{\lambda}{2\pi}\right)^{1/2} \exp\left\{-\frac{\lambda}{2}(x_n-\mu)^2\right\} \\
&\propto \left[\lambda^{1/2}\exp\left(-\frac{\lambda\mu^2}{2}\right)\right]^N \exp\left\{\lambda\mu\sum_{n=1}^{N}x_n - \frac{\lambda}{2}\sum_{n=1}^{N}x_n^2\right\}
\end{aligned}
$$

So a conjugate prior of μ and λ is

$$
p(\mu,\lambda) \propto \left[\lambda^{1/2}\exp\left(-\frac{\lambda\mu^2}{2}\right)\right]^{\beta_0} \exp\left\{c\lambda\mu - d\lambda\right\}
$$

The conjugate prior can be further reduced

$$
\begin{aligned}
p(\mu, \lambda) &\propto \left[\lambda^{1/2} \exp\left(-\frac{\lambda\mu^2}{2}\right)\right]^{\beta_0} \exp\{c\lambda\mu - d\lambda\} \\
&= \exp\left\{c\lambda\mu - \frac{\beta}{2}\lambda\mu^2\right\} \lambda^{\beta_0/2} \exp\{-d\lambda\} \\
&= \exp\left\{-\frac{\beta_0\lambda}{2}(\mu - c/\beta_0)^2\right\} \lambda^{\beta_0/2} \exp\left\{-\left(d - \frac{c^2}{2\beta_0}\right)\lambda\right\} \\
&= \mathcal{N}(\mu|\alpha_0, (\beta_0\lambda)^{-1})\mathsf{Gam}(\lambda|a_0, b_0)
\end{aligned}
$$

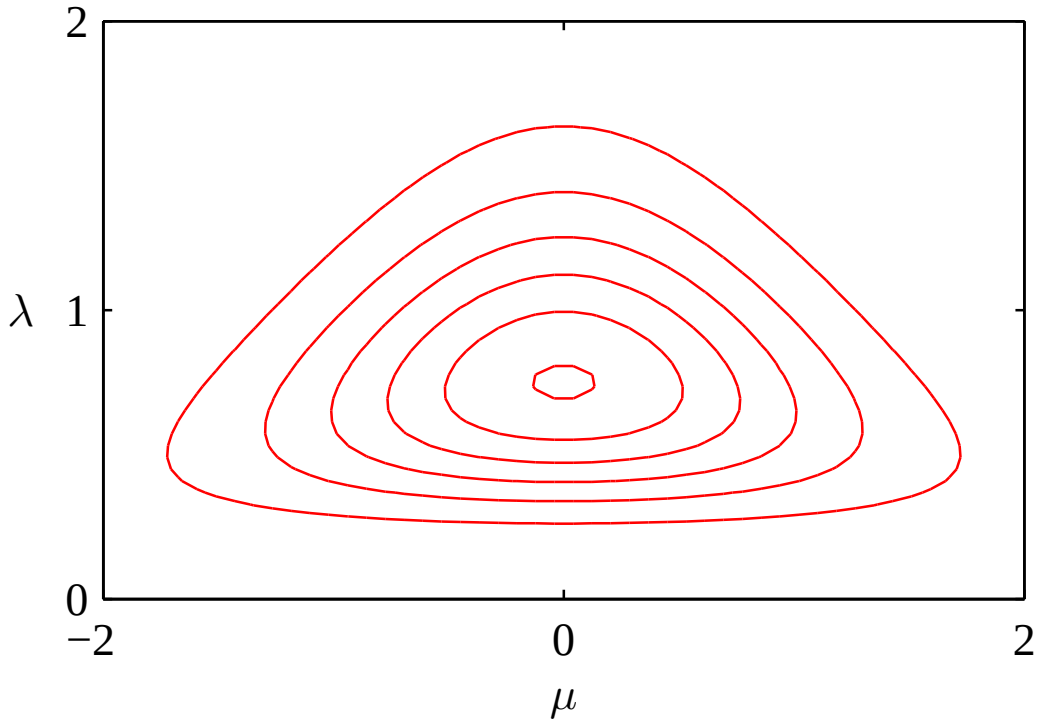
with

$$
\alpha_0 = \frac{c}{\beta_0}, \ a_0 = 1 + \frac{\beta_0}{2}, \ b_0 = d - \frac{c^2}{2\beta_0}
$$

This is a **normal-gamma distribution** defined by

$$
\text{Nor-Gam}(\mu, \lambda|\alpha, \beta, a, b) = \mathcal{N}(\mu|\alpha, (\beta\lambda)^{-1})\mathsf{Gam}(\lambda|a, b)
$$

Example



Plot of Nor-Gam$(\mu, \lambda|\alpha, \beta, a, b)$ with $\alpha = 0,\ \beta = 2,\ a = 5,\ b = 6.$

According to the Bayes' rule

$$p(\mu, \lambda|\mathcal{D}) \propto p(\mu, \lambda)p(\mathcal{D}|\mu, \lambda)$$

The posterior distribution of μ and λ is

$$p(\mu, \lambda|\mathcal{D}) = \text{Nor-Gam}(\mu, \lambda|\alpha_N, \beta_N, a_N, b_N)$$

with parameters

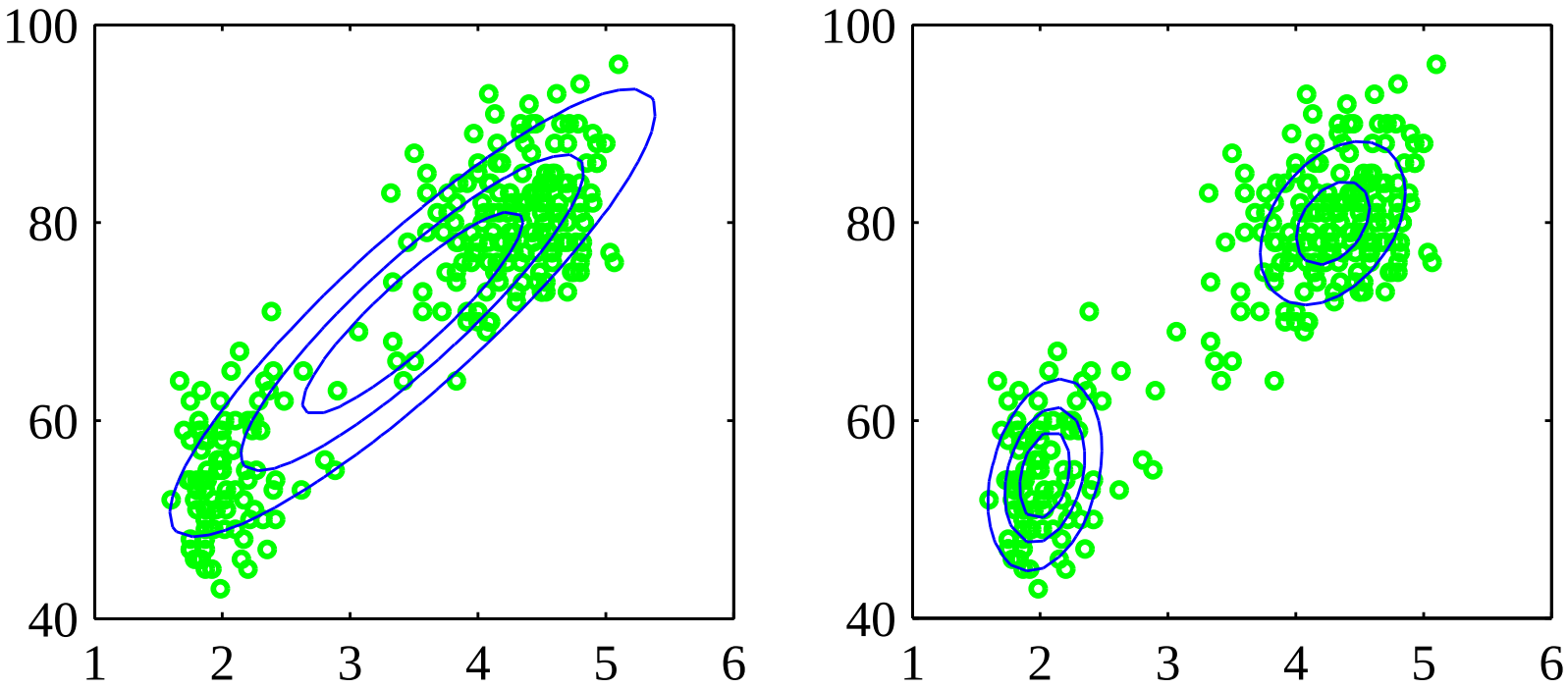
$$\beta_N = \beta_0 + N,\ \alpha_N = \frac{c_N}{\beta_N},\ a_N = 1 + \frac{\beta_N}{2},\ b_N = d_N - \frac{c_N^2}{2\beta_N}$$

where

$$c_N = c + \sum_{n=1}^{N} x_n,\ d_N = d + \frac{1}{2}\sum_{n=1}^{N} x_n^2$$

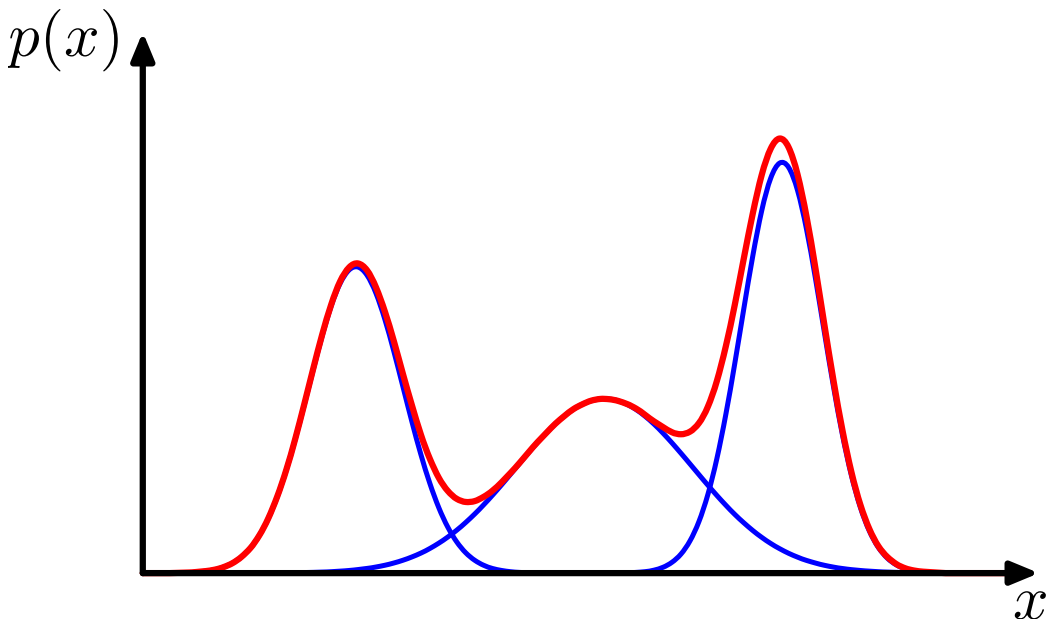
An Issue with Gaussian PDF

A Gaussian distribution is not good for data with multiple clusters.



Fitting the old faithful data set with a Gaussian (left) and a mixture of Gaussians (right).

With sufficient components, a linear combination of Gaussian PDFs can approximate any distribution.



Definition. A **mixture of Gaussians** is a superposition of Gaussians

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

where $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ is a **component** of the mixture, with a **mixing coefficient** π_k. The mixing coefficients of a mixture of Gaussians must satisfy

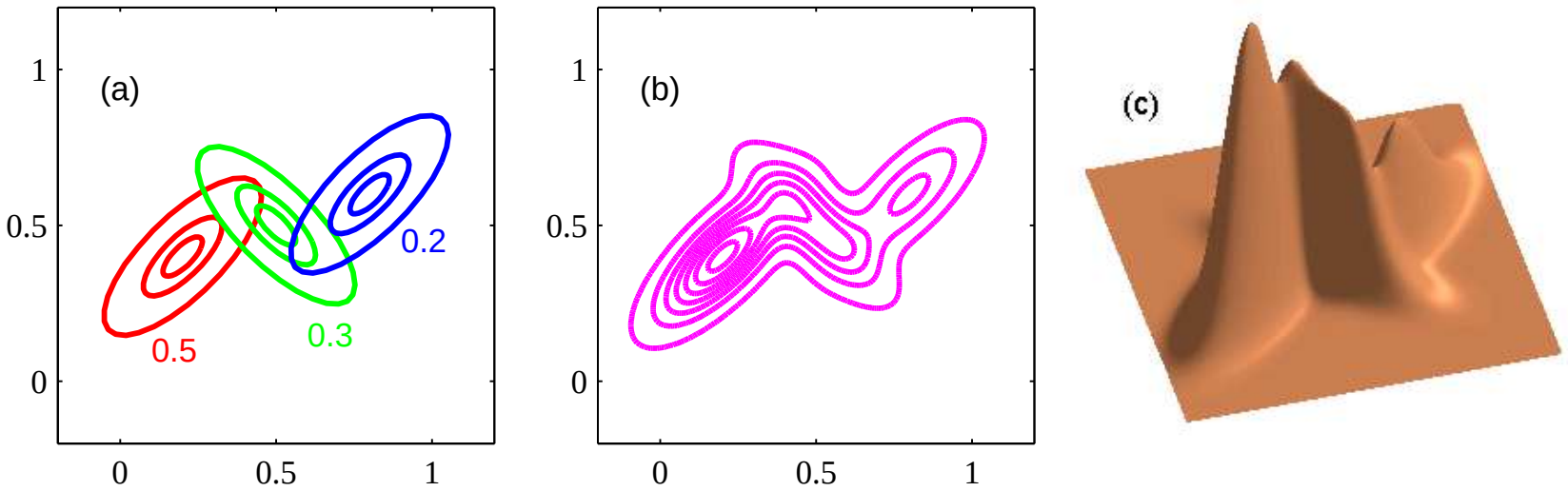
$$\pi_k \geq 0, \ \sum_k \pi_k = 1$$

The parameters of a mixture of Gaussians are

$$\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k \text{ for } k = 1, \ldots, K$$

2-D Example

A mixture of bi-variate Gaussians.



Left: Contours of the PDFs of 3 Gaussian components
Middle: Contours of the PDF of the mixture
Right: 3-D plot of the PDF of the mixture

Exponential Family*

Definition. Let $\mathbf{x}$ be a random vector with a parametric distribution $p(\boldsymbol{x}|\boldsymbol{\eta})$. The distribution $p(\boldsymbol{x}|\boldsymbol{\eta})$ is in the **exponential family** if it has the following form

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

- $\boldsymbol{\eta}$ are the natural parameters.
- $g(\boldsymbol{\eta})$ ensures normalization of $p(\boldsymbol{x}|\boldsymbol{\eta})$, i.e.

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

A Bernoulli distribution is in the exponential family.

$$\begin{aligned}\text{Bern}(x|\mu) &= \mu^x(1-\mu)^{1-x} \\ &= \exp\{x\log\mu + (1-x)\log(1-\mu)\} \\ &= (1-\mu)\exp\left\{x\log\left(\frac{\mu}{1-\mu}\right)\right\} \\ &= h(x)g(\boldsymbol{\eta})\exp\{\boldsymbol{\eta}^T\boldsymbol{u}(x)\}\end{aligned}$$

with

$$\eta = \log\left(\frac{\mu}{1-\mu}\right)$$

$$u(x) = x$$

$$g(\eta) = 1-\mu = \frac{1}{1+\exp(\eta)}$$

$$h(x) = 1$$

A K-ary distribution is in the exponential family.

$$
\begin{aligned}
p(\boldsymbol{x}|\boldsymbol{\mu}) &= \prod_{k=1}^{K} \mu_k^{x_k} = \exp\left\{\sum_{k=1}^{K} x_k \log \mu_k\right\} \\
&= \exp\left\{\sum_{k=1}^{K-1} x_k \log \mu_k + \left(1 - \sum_{k=1}^{K-1} x_k\right) \log \mu_K\right\} \\
&= \mu_K \exp\left\{\sum_{k=1}^{K-1} x_k \log\left(\frac{\mu_k}{\mu_K}\right)\right\} \\
&= h(\boldsymbol{x}) g(\boldsymbol{\eta}) \exp\{\boldsymbol{\eta}^T \boldsymbol{u}(\boldsymbol{x})\}
\end{aligned}
$$

with

$$
\eta_k = \log\left(\frac{\mu_k}{\mu_K}\right), \; g(\boldsymbol{\eta}) = \mu_K = 1 - \sum_{k=1}^{K-1} \mu_k, \; h(\boldsymbol{x}) = 1, \; \boldsymbol{u}(\boldsymbol{x}) = \boldsymbol{x}
$$

A Gaussian PDF is in the exponential family.

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

with

$$\boldsymbol{\eta} = \begin{pmatrix}\frac{\mu}{\sigma^2}\\ -\frac{1}{2\sigma^2}\end{pmatrix},\ \boldsymbol{u}(x) = \begin{pmatrix}x\\ x^2\end{pmatrix}$$

$$h(x) = (2\pi)^{-\frac{1}{2}},\ g(\boldsymbol{\eta}) = (-2\eta_2)^{\frac{1}{2}}\exp\left(\frac{\eta_1^2}{4\eta_2}\right)$$

Let $\mathbf{x}$ be a random vector with

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

Consider the density estimation of $\mathbf{x}$ with $\mathcal{D} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$.

$$\begin{aligned} p(\mathcal{D}|\boldsymbol{\eta}) &= \prod_{n=1}^{N} p(\boldsymbol{x}_n|\boldsymbol{\eta}) \\ &= \prod_{n=1}^{N} \left(h(\boldsymbol{x}_n)g(\boldsymbol{\eta})\exp\left\{\boldsymbol{\eta}^T\boldsymbol{u}(\boldsymbol{x}_n)\right\}\right) \\ \Rightarrow \quad \log p(\mathcal{D}|\boldsymbol{\eta}) &= \sum_{n=1}^{N} \log h(\boldsymbol{x}_n) + N\log g(\boldsymbol{\eta}) + \boldsymbol{\eta}^T \sum_{n=1}^{N} \boldsymbol{u}(\boldsymbol{x}_n) \\ \Rightarrow \quad \nabla \log p(\mathcal{D}|\boldsymbol{\eta}) &= N\nabla \log g(\boldsymbol{\eta}) + \sum_{n=1}^{N} \boldsymbol{u}(\boldsymbol{x}_n) \end{aligned}$$

Maximum likelihood estimate is a stationary point of $\log p(\mathcal{D}|\boldsymbol{\eta})$

$$\nabla \log p(\mathcal{D}|\boldsymbol{\eta})\Big|_{\boldsymbol{\eta}=\boldsymbol{\eta}_{\mathsf{ML}}} = 0$$

With $p(\boldsymbol{x}|\boldsymbol{\eta})$ in the exponential family, we have

$$-\nabla \log g(\boldsymbol{\eta}_{\mathsf{ML}}) = \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{u}(\boldsymbol{x}_n)$$

Note $\boldsymbol{\eta}_{\mathsf{ML}}$ depends on the data set only through $\sum_{n=1}^{N} \boldsymbol{u}(\boldsymbol{x}_n)$. This is an example of **sufficient statistics**.

Conjugate Prior

Let $\mathbf{x}$ be a random vector with $p(\boldsymbol{x}|\boldsymbol{\eta})$ in the exponential family

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

For a conjugate prior of $\boldsymbol{\eta}$, we match the dependency of the data likelihood function on $\boldsymbol{\eta}$, i.e. a power of $g(\boldsymbol{\eta})$ and an exponent linear in $\boldsymbol{\eta}$

$$p(\boldsymbol{\eta}|\boldsymbol{\chi},\nu) = f(\boldsymbol{\chi},\nu)g(\boldsymbol{\eta})^\nu\exp\left\{\nu\boldsymbol{\eta}^T\boldsymbol{\chi}\right\}$$

According to the Bayes' rule

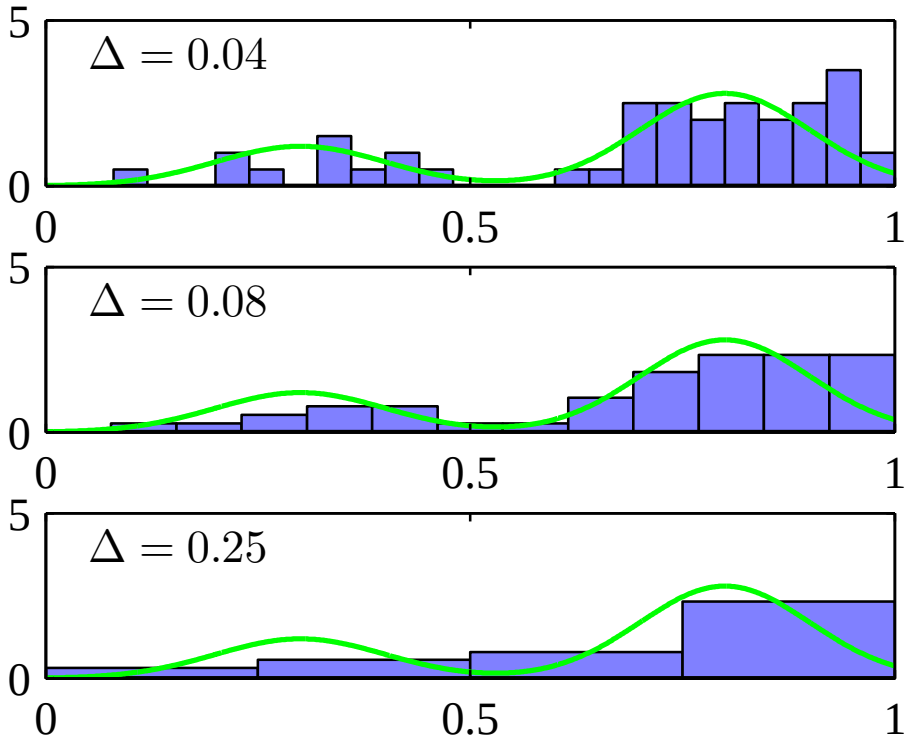
$$
\begin{aligned}
p(\boldsymbol{\eta}|\mathcal{D}, \boldsymbol{\chi}, \nu) &\propto p(\boldsymbol{\eta}|\boldsymbol{\chi}, \nu) p(\mathcal{D}|\boldsymbol{\eta}) \\
&= g(\boldsymbol{\eta})^{\nu+N} \exp\left\{\boldsymbol{\eta}^T \left(\sum_{n=1}^{N} \boldsymbol{u}(\boldsymbol{x}_n) + \nu\boldsymbol{\chi}\right)\right\} \\
&= g(\boldsymbol{\eta})^{\nu'} \exp\left\{\nu' \boldsymbol{\eta}^T \boldsymbol{\chi}'\right\} \\
&\propto p(\boldsymbol{\eta}|\boldsymbol{\chi}', \nu')
\end{aligned}
$$

where

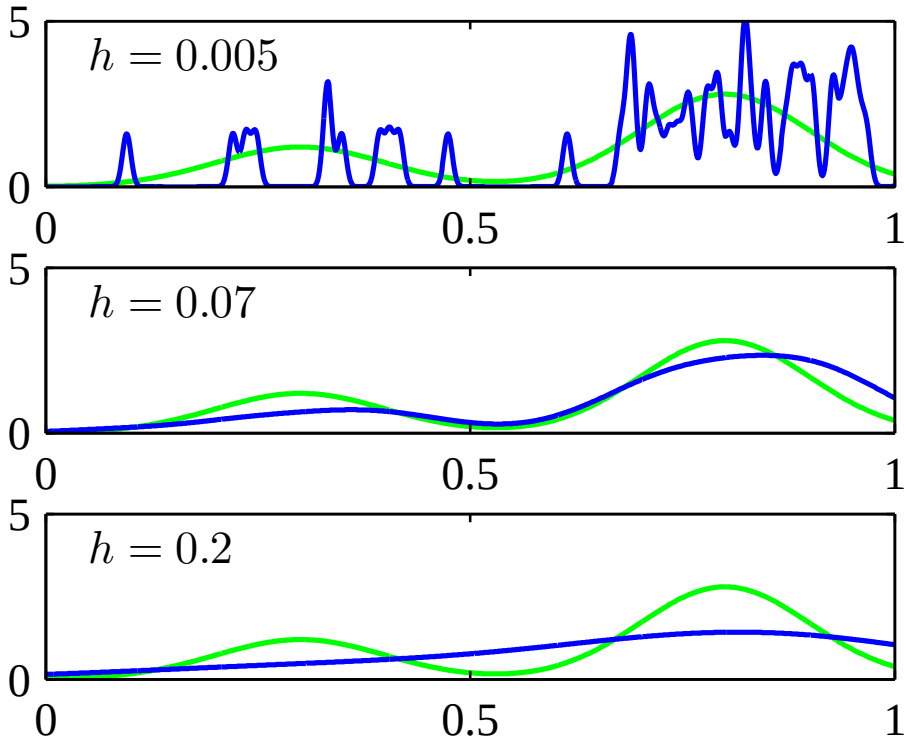
$$
\nu' = \nu + N, \ \boldsymbol{\chi}' = \frac{1}{\nu'}\left(\nu\boldsymbol{\chi} + \sum_{n=1}^{N} \boldsymbol{u}(\boldsymbol{x}_n)\right)
$$

Nonparametric Methods

Histogram Density Estimation



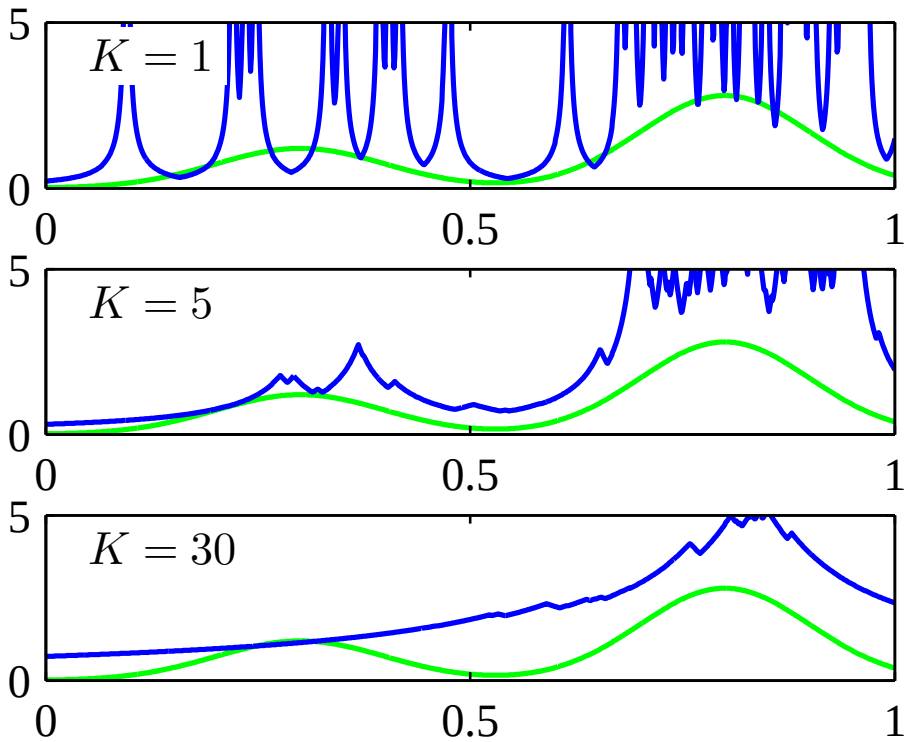
3 cases of histogram density estimation with 50 data points generated from the distribution shown by the green curve.



3 cases of kernel density estimation with the same data set

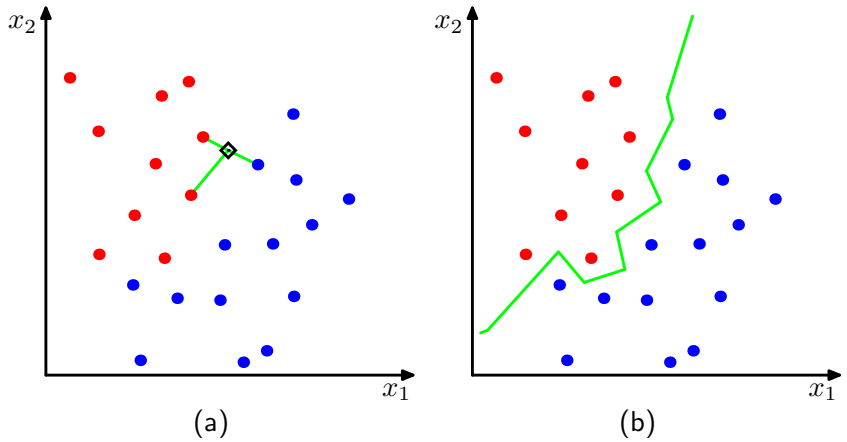
$$p(x) = \frac{1}{N}\sum_n k(x, x_n), \text{ where } k(x, x') = \frac{1}{(2\pi h^2)^{1/2}} \exp\left\{-\frac{(x-x')^2}{2h^2}\right\}$$

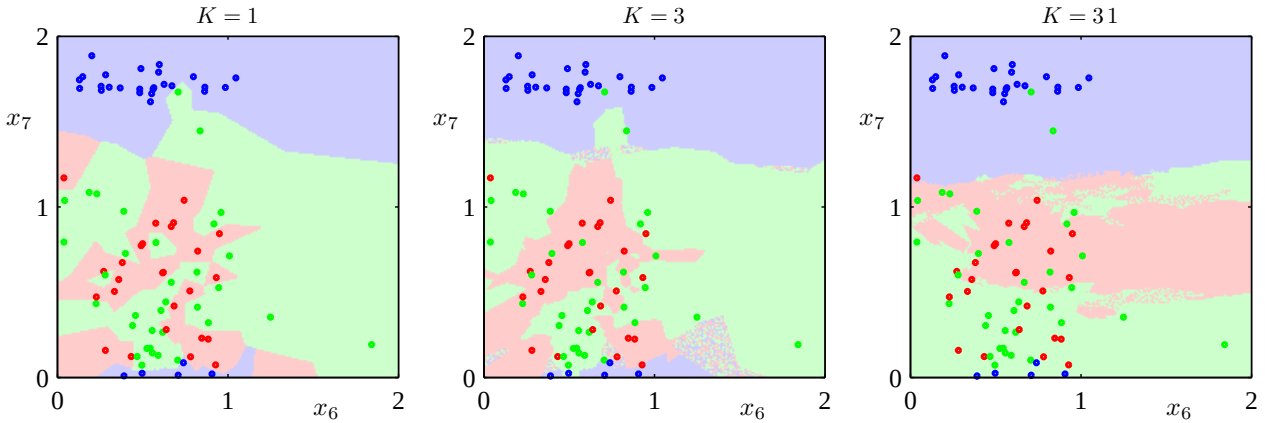
K-Nearest-Neighbor Density Estimation



3 cases of KNN density estimation with the same data set. Here $\hat{p}(x) = \frac{K}{NV(x)}$ where $V(x)$ is the volume of a sphere centered on x and containing K data points.

K-Nearest-Neighbor Classifiers





* **Joint, Marginal, and Conditional Gaussians**

Log of Gaussian Distribution

For a Gaussian random vector $\mathbf{x}$, the log distribution

- is quadratic in $\boldsymbol{x}$
- the second-order term depends on precision/covariance
- the first-order term depends on precision and mean

$$\begin{aligned}\log \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}) &= -\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Lambda}(\boldsymbol{x}-\boldsymbol{\mu}) + \text{const} \\ &= -\frac{1}{2}\boldsymbol{x}^T \boldsymbol{\Lambda}\boldsymbol{x} + \boldsymbol{x}^T \boldsymbol{\Lambda}\boldsymbol{\mu} + \text{const}\end{aligned}$$

Let $\mathbf{x} \sim \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1})$ be a Gaussian random vector. Partition $\mathbf{x}$ and $\boldsymbol{\mu}$ into sub-vectors

$$\mathbf{x} = \begin{pmatrix} \mathbf{x}_a \\ \mathbf{x}_b \end{pmatrix}, \quad \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{pmatrix}$$

Partition the covariance and precision into sub-matrices

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

The conditional distribution of $\mathbf{x}_a$ given $\mathbf{x}_b$ is Gaussian.

The conditional distribution of $\mathbf{x}_a$ given $\mathbf{x}_b = \boldsymbol{x}_b$ is Gaussian

$$\mathcal{N}(\boldsymbol{x}_a|\boldsymbol{\mu}_{a|b}, \boldsymbol{\Lambda}_{a|b}^{-1})$$

where the conditional mean is

$$\boldsymbol{\mu}_{a|b} = \boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{aa}^{-1}\boldsymbol{\Lambda}_{ab}(\boldsymbol{x}_b - \boldsymbol{\mu}_b)$$

and the conditional precision is

$$\boldsymbol{\Lambda}_{a|b} = \boldsymbol{\Lambda}_{aa}$$

Proof: Gaussian

The conditional distribution of $\mathbf{x}_a$ given $\mathbf{x}_b = \boldsymbol{x}_b$ is

$$p(\boldsymbol{x}_a|\boldsymbol{x}_b) = \frac{p(\boldsymbol{x}_a, \boldsymbol{x}_b)}{p(\boldsymbol{x}_b)} \propto p(\boldsymbol{x}_a, \boldsymbol{x}_b)$$

The logarithm of $p(\boldsymbol{x}_a|\boldsymbol{x}_b)$, apart from a constant,

$$-\tfrac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Lambda} (\boldsymbol{x} - \boldsymbol{\mu})$$

$$= -\tfrac{1}{2}[(\boldsymbol{x}_a - \boldsymbol{\mu}_a)^T \boldsymbol{\Lambda}_{aa} (\boldsymbol{x}_a - \boldsymbol{\mu}_a) + (\boldsymbol{x}_a - \boldsymbol{\mu}_a)^T \boldsymbol{\Lambda}_{ab} (\boldsymbol{x}_b - \boldsymbol{\mu}_b)$$

$$+ (\boldsymbol{x}_b - \boldsymbol{\mu}_b)^T \boldsymbol{\Lambda}_{ba} (\boldsymbol{x}_a - \boldsymbol{\mu}_a) + (\boldsymbol{x}_b - \boldsymbol{\mu}_b)^T \boldsymbol{\Lambda}_{bb} (\boldsymbol{x}_b - \boldsymbol{\mu}_b)]$$

is **quadratic** in $\boldsymbol{x}_a$. Hence $\mathbf{x}_a$ given $\mathbf{x}_b = \boldsymbol{x}_b$ is Gaussian.

Proof: Conditional Precision

The **second-order** term in $\log p(\boldsymbol{x}_a|\boldsymbol{x}_b)$ is

$$-\frac{1}{2}\boldsymbol{x}_a^T\boldsymbol{\Lambda}_{aa}\boldsymbol{x}_a$$

Since $\mathbf{x}_a$ given $\mathbf{x}_b = \boldsymbol{x}_b$ is Gaussian, this term must be

$$-\frac{1}{2}\boldsymbol{x}_a^T\boldsymbol{\Lambda}_{a|b}\boldsymbol{x}_a$$

Hence

$$\boldsymbol{\Lambda}_{a|b} = \boldsymbol{\Lambda}_{aa}$$

Note this conditional precision is independent of $\boldsymbol{x}_b$.

Proof: Conditional Mean

The **first-order** term in $\log p(\boldsymbol{x}_a|\boldsymbol{x}_b)$ is

$$\boldsymbol{x}_a^T \{\boldsymbol{\Lambda}_{aa}\boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{ab}(\boldsymbol{x}_b - \boldsymbol{\mu}_b)\}$$

Since $\mathbf{x}_a$ given $\mathbf{x}_b = \boldsymbol{x}_b$ is Gaussian, this term must be

$$\boldsymbol{x}_a^T \boldsymbol{\Lambda}_{a|b}\boldsymbol{\mu}_{a|b}$$

So $\boldsymbol{\Lambda}_{aa}\boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{ab}(\boldsymbol{x}_b - \boldsymbol{\mu}_b) = \boldsymbol{\Lambda}_{a|b}\boldsymbol{\mu}_{a|b}$. Hence

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

Note this conditional mean is **linear** in $\boldsymbol{x}_b$.

Inverse of a Partitioned Matrix

Consider $\begin{pmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{C} & \boldsymbol{D} \end{pmatrix}$ where $\boldsymbol{A}$ and $\boldsymbol{D}$ are invertible. Then

$$\begin{pmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{C} & \boldsymbol{D} \end{pmatrix}^{-1} = \begin{pmatrix} \boldsymbol{M} & -\boldsymbol{M}\boldsymbol{B}\boldsymbol{D}^{-1} \\ -\boldsymbol{D}^{-1}\boldsymbol{C}\boldsymbol{M} & \boldsymbol{D}^{-1} + \boldsymbol{D}^{-1}\boldsymbol{C}\boldsymbol{M}\boldsymbol{B}\boldsymbol{D}^{-1} \end{pmatrix}$$

where $\boldsymbol{M} = \left(\boldsymbol{A} - \boldsymbol{B}\boldsymbol{D}^{-1}\boldsymbol{C}\right)^{-1}$.

Note that $\boldsymbol{M}^{-1} = \boldsymbol{A} - \boldsymbol{B}\boldsymbol{D}^{-1}\boldsymbol{C}$ is also known as the **Schur complement** of $\boldsymbol{D}$.

Inverse Covariance Matrix

For a covariance matrix

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

$$= \begin{pmatrix} \left(\mathbf{\Sigma}_{aa} - \mathbf{\Sigma}_{ab}\mathbf{\Sigma}_{bb}^{-1}\mathbf{\Sigma}_{ba}\right)^{-1} & -\left(\mathbf{\Sigma}_{aa} - \mathbf{\Sigma}_{ab}\mathbf{\Sigma}_{bb}^{-1}\mathbf{\Sigma}_{ba}\right)^{-1}\mathbf{\Sigma}_{ab}\mathbf{\Sigma}_{bb}^{-1} \\ * & * \end{pmatrix}$$

Hence

$$\mathbf{\Lambda}_{aa} = \left(\mathbf{\Sigma}_{aa} - \mathbf{\Sigma}_{ab}\mathbf{\Sigma}_{bb}^{-1}\mathbf{\Sigma}_{ba}\right)^{-1}$$

$$\mathbf{\Lambda}_{ab} = -\left(\mathbf{\Sigma}_{aa} - \mathbf{\Sigma}_{ab}\mathbf{\Sigma}_{bb}^{-1}\mathbf{\Sigma}_{ba}\right)^{-1}\mathbf{\Sigma}_{ab}\mathbf{\Sigma}_{bb}^{-1}$$

Similarly

$$\mathbf{\Sigma}_{aa} = \left(\mathbf{\Lambda}_{aa} - \mathbf{\Lambda}_{ab}\mathbf{\Lambda}_{bb}^{-1}\mathbf{\Lambda}_{ba}\right)^{-1}$$

$$\mathbf{\Sigma}_{ab} = -\left(\mathbf{\Lambda}_{aa} - \mathbf{\Lambda}_{ab}\mathbf{\Lambda}_{bb}^{-1}\mathbf{\Lambda}_{ba}\right)^{-1}\mathbf{\Lambda}_{ab}\mathbf{\Lambda}_{bb}^{-1}$$

We can use covariance instead of precision.

The conditional mean and covariance of $\mathbf{x}_a$ given $\mathbf{x}_b = \boldsymbol{x}_b$ can be expressed by covariance as

$$\boldsymbol{\mu}_{a|b} = \boldsymbol{\mu}_a + \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}(\boldsymbol{x}_b - \boldsymbol{\mu}_b)$$

$$\boldsymbol{\Sigma}_{a|b} = \boldsymbol{\Sigma}_{aa} - \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}\boldsymbol{\Sigma}_{ba}$$

This follow from

$$\begin{aligned}\boldsymbol{\Lambda}_{aa}^{-1}\boldsymbol{\Lambda}_{ab} &= \left(\boldsymbol{\Sigma}_{aa} - \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}\boldsymbol{\Sigma}_{ba}\right)\left(-\left(\boldsymbol{\Sigma}_{aa} - \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}\boldsymbol{\Sigma}_{ba}\right)^{-1}\boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}\right) \\ &= -\boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}\end{aligned}$$

$$\boldsymbol{\Sigma}_{a|b} = \boldsymbol{\Lambda}_{a|b}^{-1} = \boldsymbol{\Lambda}_{aa}^{-1} = \boldsymbol{\Sigma}_{aa} - \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}\boldsymbol{\Sigma}_{ba}$$

$\mathbf{x}_a$ is Gaussian with mean $\boldsymbol{\mu}_a$ and covariance $\boldsymbol{\Sigma}_{aa}$.

Proof. By marginalization

$$p(\boldsymbol{x}_a) = \int p(\boldsymbol{x}_a, \boldsymbol{x}_b) d\boldsymbol{x}_b$$

The integration function $p(\boldsymbol{x}_a, \boldsymbol{x}_b)$ is exponential with exponent

$$\begin{aligned} -\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Lambda}(\boldsymbol{x}-\boldsymbol{\mu}) &= -\frac{1}{2}\boldsymbol{x}^T\boldsymbol{\Lambda}\boldsymbol{x} + \boldsymbol{x}^T\boldsymbol{\Lambda}\boldsymbol{\mu} + \text{const} \\ &= -\frac{1}{2}\boldsymbol{x}_a^T\boldsymbol{\Lambda}_{aa}\boldsymbol{x}_a - \frac{1}{2}\boldsymbol{x}_b^T\boldsymbol{\Lambda}_{bb}\boldsymbol{x}_b - \boldsymbol{x}_b^T\boldsymbol{\Lambda}_{ba}\boldsymbol{x}_a \\ &\quad + \boldsymbol{x}_a^T\left(\boldsymbol{\Lambda}_{aa}\boldsymbol{\mu}_a + \boldsymbol{\Lambda}_{ab}\boldsymbol{\mu}_b\right) + \boldsymbol{x}_b^T\left(\boldsymbol{\Lambda}_{ba}\boldsymbol{\mu}_a + \boldsymbol{\Lambda}_{bb}\boldsymbol{\mu}_b\right) \\ &\quad + \text{const} \end{aligned}$$

The terms involved in the integration over $\boldsymbol{x}_b$ is

$$
\begin{aligned}
&-\frac{1}{2}\boldsymbol{x}_b^T\boldsymbol{\Lambda}_{bb}\boldsymbol{x}_b - \boldsymbol{x}_b^T\boldsymbol{\Lambda}_{ba}\boldsymbol{x}_a + \boldsymbol{x}_b^T\left(\boldsymbol{\Lambda}_{ba}\boldsymbol{\mu}_a + \boldsymbol{\Lambda}_{bb}\boldsymbol{\mu}_b\right) \\
&= -\frac{1}{2}\boldsymbol{x}_b^T\boldsymbol{\Lambda}_{bb}\boldsymbol{x}_b + \boldsymbol{x}_b^T\boldsymbol{m} \\
&= -\frac{1}{2}(\boldsymbol{x}_b - \boldsymbol{\Lambda}_{bb}^{-1}\boldsymbol{m})^T\boldsymbol{\Lambda}_{bb}(\boldsymbol{x}_b - \boldsymbol{\Lambda}_{bb}^{-1}\boldsymbol{m}) + \frac{1}{2}\boldsymbol{m}^T\boldsymbol{\Lambda}_{bb}^{-1}\boldsymbol{m}
\end{aligned}
$$

where $\boldsymbol{m} = \boldsymbol{\Lambda}_{bb}\boldsymbol{\mu}_b + \boldsymbol{\Lambda}_{ba}\boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{ba}\boldsymbol{x}_a = \boldsymbol{\Lambda}_{bb}\boldsymbol{\mu}_b - \boldsymbol{\Lambda}_{ba}(\boldsymbol{x}_a - \boldsymbol{\mu}_a)$.
After the integration over $\boldsymbol{x}_b$, the remaining exponent is

$$
\begin{aligned}
&-\frac{1}{2}\boldsymbol{x}_a^T\boldsymbol{\Lambda}_{aa}\boldsymbol{x}_a + \boldsymbol{x}_a^T\left(\boldsymbol{\Lambda}_{aa}\boldsymbol{\mu}_a + \boldsymbol{\Lambda}_{ab}\boldsymbol{\mu}_b\right) + \frac{1}{2}\boldsymbol{m}^T\boldsymbol{\Lambda}_{bb}^{-1}\boldsymbol{m} \\
&= -\frac{1}{2}\boldsymbol{x}_a^T(\boldsymbol{\Lambda}_{aa} - \boldsymbol{\Lambda}_{ab}\boldsymbol{\Lambda}_{bb}^{-1}\boldsymbol{\Lambda}_{ba})\boldsymbol{x}_a + \boldsymbol{x}_a^T(\boldsymbol{\Lambda}_{aa} - \boldsymbol{\Lambda}_{ab}\boldsymbol{\Lambda}_{bb}^{-1}\boldsymbol{\Lambda}_{ba})\boldsymbol{\mu}_a + \text{const}
\end{aligned}
$$

Thus, the covariance of $\boldsymbol{x}_a$ is

$$\boldsymbol{\Sigma}_{a|\emptyset} = (\boldsymbol{\Lambda}_{aa} - \boldsymbol{\Lambda}_{ab}\boldsymbol{\Lambda}_{bb}^{-1}\boldsymbol{\Lambda}_{ba})^{-1} = \boldsymbol{\Sigma}_{aa}$$

The mean of $\boldsymbol{x}_a$ is

$$\boldsymbol{\mu}_{a|\emptyset} = \boldsymbol{\mu}_a$$

Suppose $\mathbf{x} \sim \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and

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

The conditional distribution of $\mathbf{x}_a$ given $\mathbf{x}_b = \boldsymbol{x}_b$ is Gaussian

$$\begin{aligned} p(\boldsymbol{x}_a|\boldsymbol{x}_b) &= \mathcal{N}(\boldsymbol{x}_a|\boldsymbol{\mu}_{a|b}, \boldsymbol{\Sigma}_{a|b}) \\ \boldsymbol{\mu}_{a|b} &= \boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{aa}^{-1}\boldsymbol{\Lambda}_{ab}(\boldsymbol{x}_b - \boldsymbol{\mu}_b) \\ \boldsymbol{\Sigma}_{a|b} &= \boldsymbol{\Lambda}_{aa}^{-1} \end{aligned}$$

The marginal distribution of $\mathbf{x}_a$ is Gaussian

$$\begin{aligned} p(\boldsymbol{x}_a) &= \mathcal{N}(\boldsymbol{x}_a|\boldsymbol{\mu}_{a|\emptyset}, \boldsymbol{\Sigma}_{a|\emptyset}) \\ \boldsymbol{\mu}_{a|\emptyset} &= \boldsymbol{\mu}_a \\ \boldsymbol{\Sigma}_{a|\emptyset} &= \boldsymbol{\Sigma}_{aa} \end{aligned}$$

A linear Gaussian model for $\mathbf{y}$ and $\mathbf{z}$ assumes

- $\mathbf{y}$ is Gaussian
- $\mathbf{z}$ is conditional Gaussian given $\mathbf{y} = \boldsymbol{y}$
- The conditional mean of $\mathbf{z}$ given $\mathbf{y} = \boldsymbol{y}$ is linear in $\boldsymbol{y}$
- The conditional covariance of $\mathbf{z}$ given $\mathbf{y} = \boldsymbol{y}$ does not depend on $\boldsymbol{y}$

That is

$$p(\boldsymbol{y}) = \mathcal{N}(\boldsymbol{y}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1})$$

and

$$p(\boldsymbol{z}|\boldsymbol{y}) = \mathcal{N}(\boldsymbol{z}|\boldsymbol{A}\boldsymbol{y} + \boldsymbol{b}, \boldsymbol{L}^{-1})$$

The joint distribution of a linear Gaussian model is Gaussian.

By product rule

$$p(\boldsymbol{y}, \boldsymbol{z}) = p(\boldsymbol{y})p(\boldsymbol{z}|\boldsymbol{y})$$

So

$$\begin{aligned}\log p(\boldsymbol{y}, \boldsymbol{z}) &= \log p(\boldsymbol{y}) + \log p(\boldsymbol{z}|\boldsymbol{y}) \\ &= -\frac{1}{2}(\boldsymbol{y} - \boldsymbol{\mu})^T \boldsymbol{\Lambda}(\boldsymbol{y} - \boldsymbol{\mu}) \\ &\quad - \frac{1}{2}(\boldsymbol{z} - \boldsymbol{A}\boldsymbol{y} - \boldsymbol{b})^T \boldsymbol{L}(\boldsymbol{z} - \boldsymbol{A}\boldsymbol{y} - \boldsymbol{b}) + \text{const}\end{aligned}$$

which is quadratic in $\begin{pmatrix}\boldsymbol{y} \\ \boldsymbol{z}\end{pmatrix} = \boldsymbol{x} = \begin{pmatrix}\boldsymbol{x}_a \\ \boldsymbol{x}_b\end{pmatrix}$.

The second-order term in $\log p(\boldsymbol{x})$ is

$$-\frac{1}{2}\boldsymbol{y}^T(\boldsymbol{\Lambda}+\boldsymbol{A}^T\boldsymbol{L}\boldsymbol{A})\boldsymbol{y}-\frac{1}{2}\boldsymbol{z}^T\boldsymbol{L}\boldsymbol{z}+\frac{1}{2}\boldsymbol{z}^T\boldsymbol{L}\boldsymbol{A}\boldsymbol{y}+\frac{1}{2}\boldsymbol{y}^T\boldsymbol{A}^T\boldsymbol{L}\boldsymbol{z}$$

$$=-\frac{1}{2}\begin{pmatrix}\boldsymbol{y}\\ \boldsymbol{z}\end{pmatrix}^T\begin{pmatrix}\boldsymbol{\Lambda}+\boldsymbol{A}^T\boldsymbol{L}\boldsymbol{A} & -\boldsymbol{A}^T\boldsymbol{L}\\ -\boldsymbol{L}\boldsymbol{A} & \boldsymbol{L}\end{pmatrix}\begin{pmatrix}\boldsymbol{y}\\ \boldsymbol{z}\end{pmatrix}=-\frac{1}{2}\boldsymbol{x}^T\boldsymbol{\Lambda}_{\mathbf{x}}\boldsymbol{x}$$

Hence the precision and covariance of $\mathbf{x}$ are

$$\boldsymbol{\Lambda}_{\mathbf{x}}=\begin{pmatrix}\boldsymbol{\Lambda}+\boldsymbol{A}^T\boldsymbol{L}\boldsymbol{A} & -\boldsymbol{A}^T\boldsymbol{L}\\ -\boldsymbol{L}\boldsymbol{A} & \boldsymbol{L}\end{pmatrix}$$

$$\boldsymbol{\Sigma}_{\mathbf{x}}=\boldsymbol{\Lambda}_{\mathbf{x}}^{-1}=\begin{pmatrix}\boldsymbol{\Lambda}+\boldsymbol{A}^T\boldsymbol{L}\boldsymbol{A} & -\boldsymbol{A}^T\boldsymbol{L}\\ -\boldsymbol{L}\boldsymbol{A} & \boldsymbol{L}\end{pmatrix}^{-1}$$

$$=\begin{pmatrix}\boldsymbol{\Lambda}^{-1} & \boldsymbol{\Lambda}^{-1}\boldsymbol{A}^T\\ \boldsymbol{A}\boldsymbol{\Lambda}^{-1} & \boldsymbol{L}^{-1}+\boldsymbol{A}\boldsymbol{\Lambda}^{-1}\boldsymbol{A}^T\end{pmatrix}$$

The first-order term in $\log p(\boldsymbol{x})$ is

$$\begin{aligned}\boldsymbol{x}^T\boldsymbol{\Sigma}_{\mathbf{x}}^{-1}\boldsymbol{\mu}_{\mathbf{x}} &= \boldsymbol{y}^T\boldsymbol{\Lambda}\boldsymbol{\mu} - \boldsymbol{y}^T\boldsymbol{A}^T\boldsymbol{L}\boldsymbol{b} + \boldsymbol{z}^T\boldsymbol{L}\boldsymbol{b} \\ &= \begin{pmatrix}\boldsymbol{y}\\ \boldsymbol{z}\end{pmatrix}^T \begin{pmatrix}\boldsymbol{\Lambda}\boldsymbol{\mu} - \boldsymbol{A}^T\boldsymbol{L}\boldsymbol{b}\\ \boldsymbol{L}\boldsymbol{b}\end{pmatrix}\end{aligned}$$

Hence the mean of $\mathbf{x}$ is

$$\begin{aligned}\boldsymbol{\mu}_{\mathbf{x}} &= \boldsymbol{\Sigma}_{\mathbf{x}} \begin{pmatrix}\boldsymbol{\Lambda}\boldsymbol{\mu} - \boldsymbol{A}^T\boldsymbol{L}\boldsymbol{b}\\ \boldsymbol{L}\boldsymbol{b}\end{pmatrix} \\ &= \begin{pmatrix}\boldsymbol{\mu}\\ \boldsymbol{A}\boldsymbol{\mu} + \boldsymbol{b}\end{pmatrix} \\ &= \begin{pmatrix}\boldsymbol{\mu}_a\\ \boldsymbol{\mu}_b\end{pmatrix}\end{aligned}$$

Consider a linear Gaussian model for $\mathbf{y}$ and $\mathbf{z}$ where $\mathbf{y}$ is Gaussian and $\mathbf{z}$ is conditional Gaussian given $\mathbf{y} = \boldsymbol{y}$. Then

- $\mathbf{z}$ is Gaussian.
- $\mathbf{y}$ is conditional Gaussian given $\mathbf{z} = \boldsymbol{z}$.