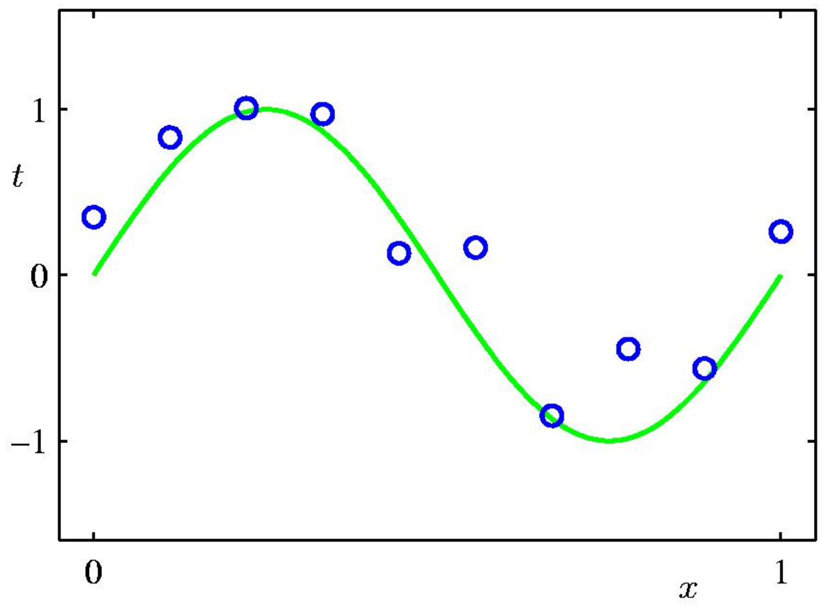
模式识别引论

An Introduction to Pattern Recognition

主讲: 李春光
www.pris.net.cn/teacher/lichunguang
模式识别与智能系统实验室
网络搜索教研中心 信息与通信工程学院 北京邮电大学

引例：多项式曲线拟合

- 给定N个训练数据：(x, t)
  - 数据其中绿色曲线为生成训练的真实曲线



- 多项式曲线拟合
  - 使用多项式模型去构造
  - 定义为

$$\boxed{y(x, \mathbf{w}) = w_0 + w_1 x + w_2 x^2 + \ldots + w_M x^M = \sum_{j=0}^{M} w_j x^j}$$

引例：多项式曲线拟合

- 学习任务:
  - 根据训练数据，估计模型中的参数

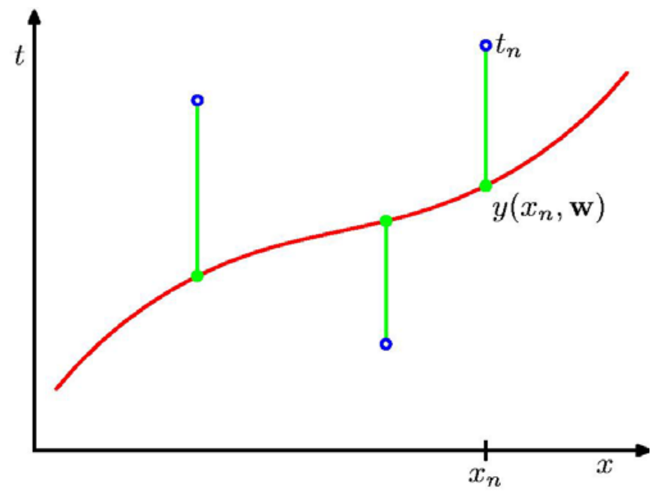
$$y(x, \mathbf{w}) = w_0 + w_1 x + w_2 x^2 + \ldots + w_M x^M = \sum_{j=0}^{M} w_j x^j$$

- 根据目标函数的不同，分为:
  - 最小二乘法

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

  - 正则化最小二乘法

$$\widetilde{E}(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^{N} \{y(x_n, \mathbf{w}) - t_n\}^2 + \frac{\lambda}{2} \|\mathbf{w}\|^2$$



最大似然估计法(MLE)

- 似然函数为

$$p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) = \prod_{n=1}^{N} \mathcal{N}\left(t_n | y(x_n, \mathbf{w}), \beta^{-1}\right)$$

- 对数似然函数为

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

- 通过最大化似然函数估计参数w_{ML}和β_{ML}

$$\boxed{\mathbf{w}_{\mathrm{ML}} = \left(\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{t}}$$

$$\boxed{\frac{1}{\beta_{\mathrm{ML}}} = \frac{1}{N}\sum_{n=1}^{N}\{y(x_n, \mathbf{w}_{\mathrm{ML}}) - t_n\}^2}$$

最大后验概率(MAP)估计法

- 先验分布密度

$$p(\mathbf{w}|\alpha) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I}) = \left(\frac{\alpha}{2\pi}\right)^{(M+1)/2} \exp\left\{-\frac{\alpha}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}\right\}$$

- 似然函数

$$p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) = \prod_{n=1}^{N} \mathcal{N}\left(t_n | y(x_n, \mathbf{w}), \beta^{-1}\right)$$

- 后验分布密度

$$p(\mathbf{w}|\mathbf{x}, \mathbf{t}, \alpha, \beta) \propto p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) p(\mathbf{w}|\alpha)$$

- 等价于正则化的最小二乘法

$$\boxed{\beta\widetilde{E}(\mathbf{w}) = \frac{\beta}{2}\sum_{n=1}^{N}\{y(x_n, \mathbf{w}) - t_n\}^2 + \frac{\alpha}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}}$$

贝叶斯估计法(Bayesian Estimation)

- 先验分布密度

$$p(\mathbf{w}|\alpha) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I}) = \left(\frac{\alpha}{2\pi}\right)^{(M+1)/2} \exp\left\{-\frac{\alpha}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}\right\}$$

- 似然函数

$$p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) = \prod_{n=1}^{N} \mathcal{N}\left(t_n|y(x_n, \mathbf{w}), \beta^{-1}\right)$$

- 后验分布密度

$$p(\mathbf{w}|\mathbf{x}, \mathbf{t}, \alpha, \beta) \propto p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta)p(\mathbf{w}|\alpha)$$

- 回归模型的预测性分布密度

$$\boxed{p(t|x, \mathbf{x}, \mathbf{t}) = \int p(t|x, \mathbf{w})p(\mathbf{w}|\mathbf{x}, \mathbf{t})\,\mathrm{d}\mathbf{w} = \mathcal{N}\left(t|m(x), s^2(x)\right)}$$

其中

$$m(x) = \beta\boldsymbol{\phi}(x)^{\mathrm{T}}\mathbf{S}\sum_{n=1}^{N}\boldsymbol{\phi}(x_n)t_n \qquad s^2(x) = \beta^{-1} + \boldsymbol{\phi}(x)^{\mathrm{T}}\mathbf{S}\boldsymbol{\phi}(x)$$

$$\mathbf{S}^{-1} = \alpha\mathbf{I} + \beta\sum_{n=1}^{N}\boldsymbol{\phi}(x_n)\boldsymbol{\phi}(x_n)^{\mathrm{T}} \qquad \boldsymbol{\phi}(x_n) = \left(x_n^0, \ldots, x_n^M\right)^{\mathrm{T}}$$

贝叶斯线性回归 内容提要

- 引子: 曲线拟合问题

- 常用的几种分布

- 贝叶斯线性回归
  - 两个例子
  - 等价核
  - 先验分布中的超参数的处理

几种常用的分布

- 伯努利(Bernoulli)分布
- 二项 (Binomial)分布
- Beta分布
- 多项分布
- 狄利克雷(Dirichlet)分布
- 高斯分布

二值(Binary)变量的伯努利分布

- **伯努利(Bernoulli)分布**
  - 比如，投一枚硬币观察结果为**Head or Tail, Head =1, Tail =0**

$$p(x=1|\mu)=\mu$$

$$\begin{aligned}\mathrm{Bern}(x|\mu) &= \mu^x(1-\mu)^{1-x} \\ \mathbb{E}[x] &= \mu \\ \mathrm{var}[x] &= \mu(1-\mu)\end{aligned}$$

伯努利分布参数的MLE

- Given

$$\mathcal{D} = \{x_1, \ldots, x_N\},\ m \text{ heads } (1),\ N - m \text{ tails } (0)$$

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

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

$$\Rightarrow \boxed{\mu_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N} x_n = \frac{m}{N}}$$

最大似然估计(MLE)的过拟合现象

- 投一枚硬币，观察其落地后是正面或反面
- 假如N次观测全部为正面：

$$\mathcal{D} = \{1, 1, 1\} \Rightarrow \mu_{\mathrm{ML}} = \frac{3}{3} = 1$$

  - 则可以给出的预测是 后续所有观测全部为正面！
  - 原因: **MLE**估计存在对数据集的**Overfitting**

二项(Binomial)分布

- 二项(**Binomial**)分布
  - 比如 投一枚硬币**N**次，出现**m**次**Head**朝上

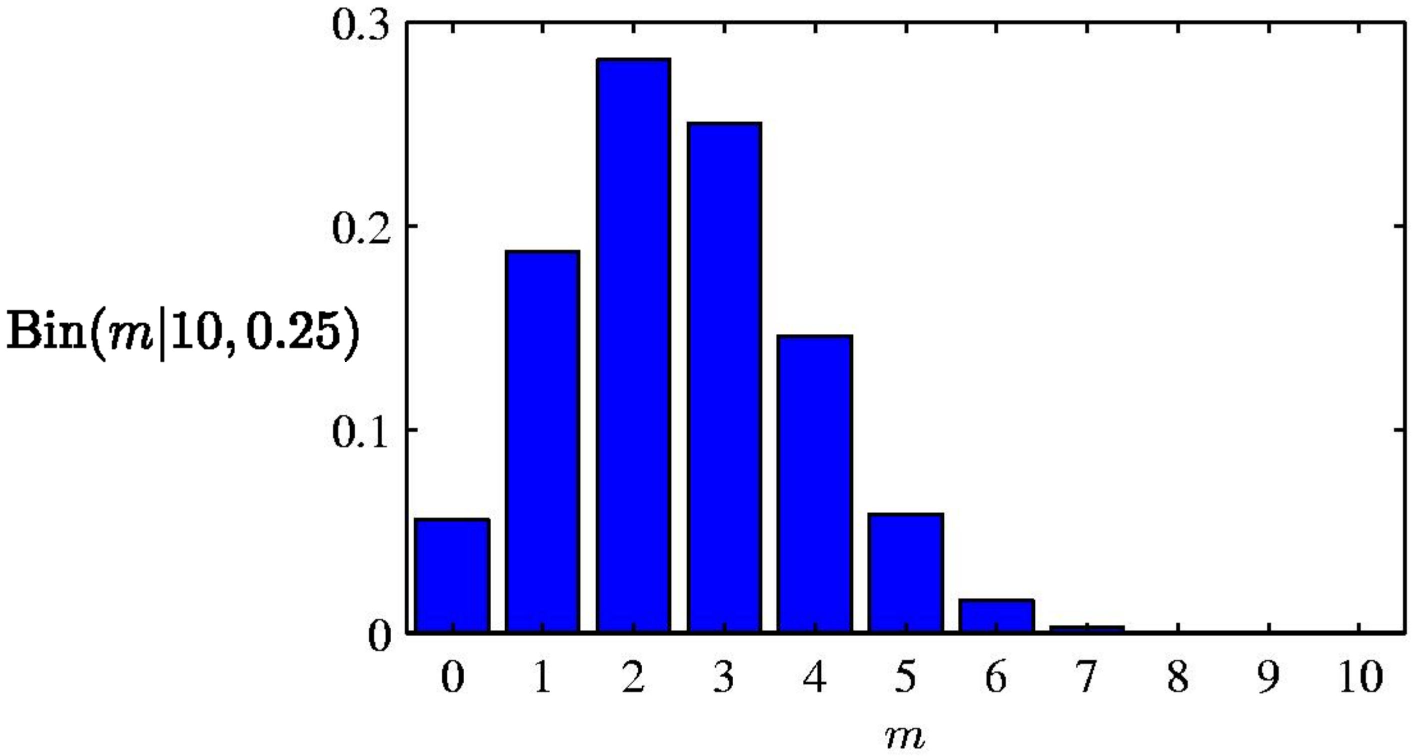
$$p(m \text{ heads}|N,\mu)$$

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

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

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

Binomial Distribution



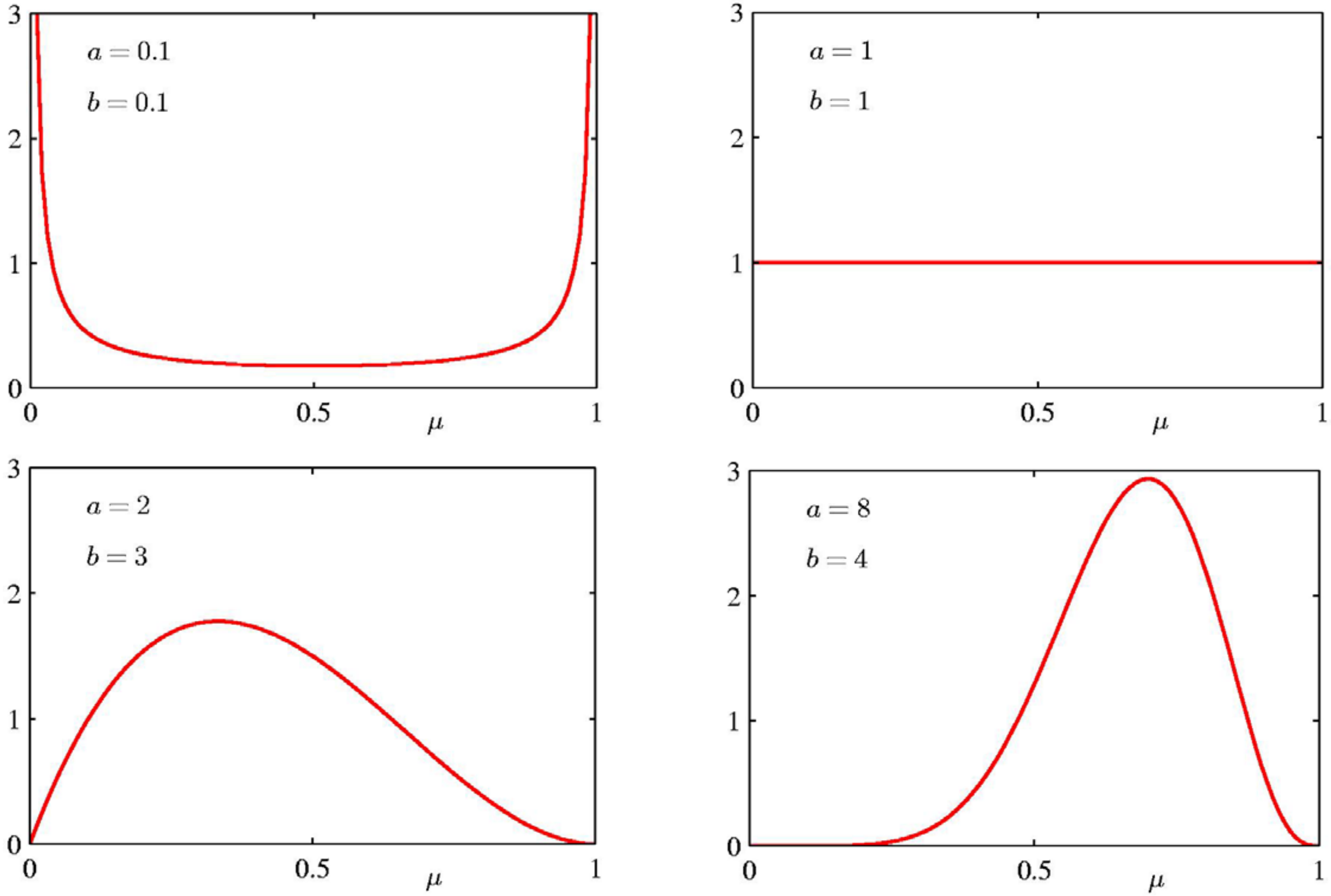
Beta分布

- Beta分布
  - 比如: 二项分布的参数$\mu \in [0,1]$可假设服从**Beta**分布

$$
\begin{aligned}
\mathrm{Beta}(\mu|a,b) &= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1} \\
\mathbb{E}[\mu] &= \frac{a}{a+b} \\
\mathrm{var}[\mu] &= \frac{ab}{(a+b)^2(a+b+1)}
\end{aligned}
$$

其中 $\Gamma(x) \equiv \int_0^\infty u^{x-1}e^{-u}\,\mathrm{d}u$

Beta Distribution



Beta分布的应用举例

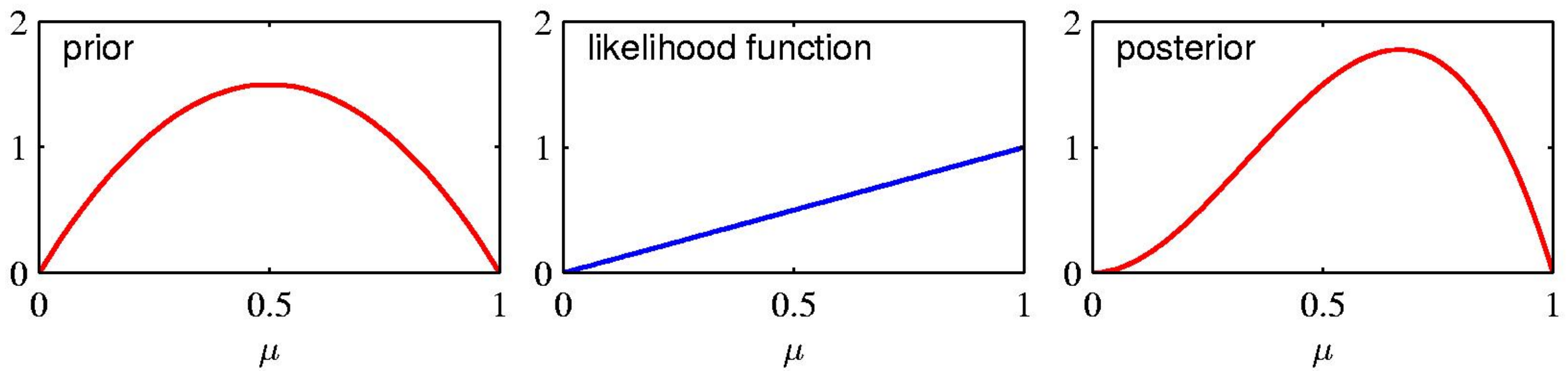
- Beta分布可以充当Binomial分布中参数的共轭先验分布 → 参数的后验分布仍为Beta分布

$$\boxed{\text{Posterior} = \text{Likelihood} \cdot \text{Prior}}$$

$$
\begin{aligned}
p(\mu|a_0, b_0, \mathcal{D}) &\propto p(\mathcal{D}|\mu)p(\mu|a_0, b_0) \\
&= \left(\prod_{n=1}^{N} \mu^{x_n}(1-\mu)^{1-x_n}\right) \mathrm{Beta}(\mu|a_0, b_0) \\
&\propto \mu^{m+a_0-1}(1-\mu)^{(N-m)+b_0-1} \\
&\propto \mathrm{Beta}(\mu|a_N, b_N)
\end{aligned}
$$

其中 $a_N = a_0 + m \qquad b_N = b_0 + (N-m)$

Likelihood · Prior = Posterior



后验分布的应用举例

- 考虑投硬币试验. 假设给定基于观测数据集D推理得到参数的后验分布，请问：在投硬币过程中，下一次正面朝上的概率是多少?

$$
\begin{aligned}
p(x=1|a_0,b_0,\mathcal{D}) &= \int_0^1 p(x=1|\mu)p(\mu|a_0,b_0,\mathcal{D})\,\mathrm{d}\mu \\
&= \int_0^1 \mu p(\mu|a_0,b_0,\mathcal{D})\,\mathrm{d}\mu \\
&= \mathbb{E}[\mu|a_0,b_0,\mathcal{D}] = \frac{a_N}{b_N}
\end{aligned}
$$

$$\Rightarrow \boxed{p(x=1|\mathcal{D}) = \frac{m+a}{m+a+l+b}}$$

其中，$a_N = m + a$，$b_N = l + b$

Bayesian Style…

多状态伯努利分布

- 具有K个互斥状态的随机变量的分布
  - 比如：分类/聚类/指派问题中的K维指示向量

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

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

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

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

多状态伯努利分布中参数的MLE

- 给定数据集 $\mathcal{D} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$

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

s.t. $\sum_k \mu_k = 1$

使用Lagrange法

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

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

多项分布

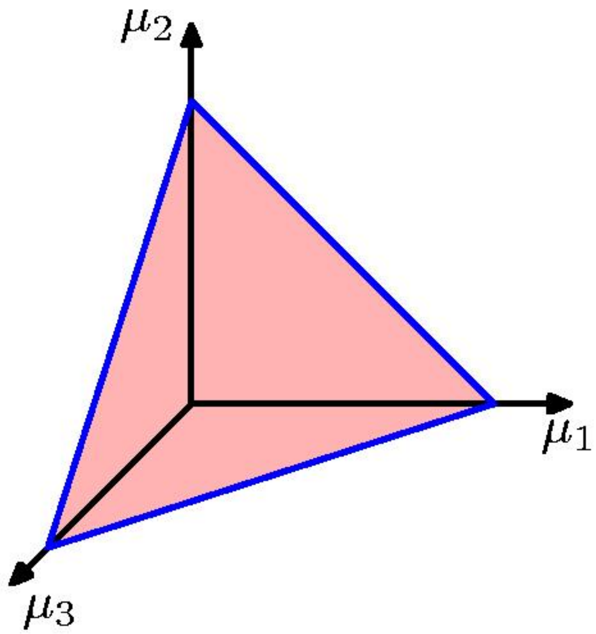
- 多项(**Multinomial**)分布
  - 比如 投一枚骰子**N**次，其中出现**k**点$\mathbf{m_k}$次，**k=1,…,K.**
- $$\begin{aligned}\mathrm{Mult}(m_1, m_2, \ldots, m_K|\boldsymbol{\mu}, N) &= \binom{N}{m_1 m_2 \ldots m_K} \prod_{k=1}^{K} \mu_k^{m_k} \\ \mathbb{E}[m_k] &= N\mu_k \\ \mathrm{var}[m_k] &= N\mu_k(1-\mu_k) \\ \mathrm{cov}[m_j m_k] &= -N\mu_j\mu_k\end{aligned}$$

狄利克雷(Dirichlet)分布

- 狄利克雷分布
  - 比如：多项分布中参数的先验分布可假设服从狄利克雷分布

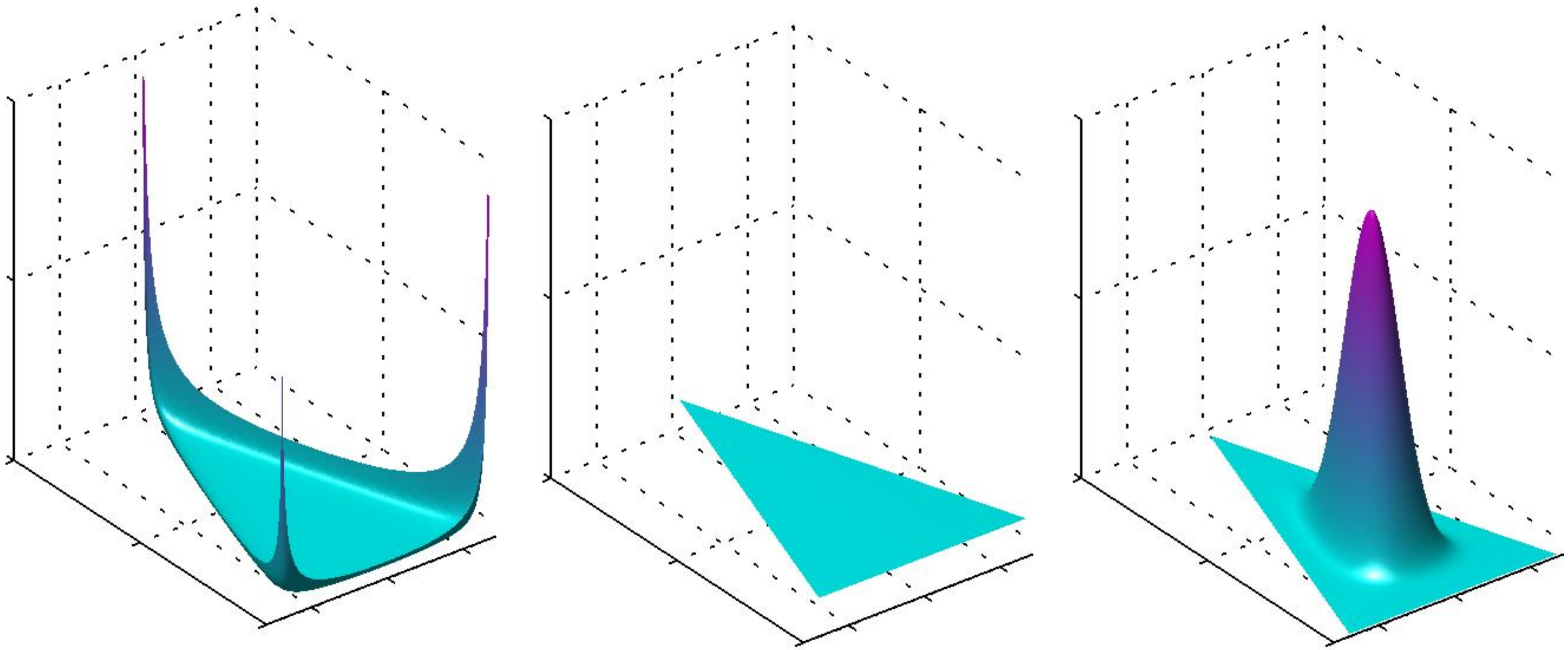
$$\mathrm{Dir}(\boldsymbol{\mu}|\boldsymbol{\alpha}) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_K)} \prod_{k=1}^{K} \mu_k^{\alpha_k - 1}$$

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



Dirichlet 分布

- 3个变量的Dirichlet分布



$\alpha_k = 10^{-1}$ $\alpha_k = 10^{0}$ $\alpha_k = 10^{1}$

Dirichlet分布的应用举例

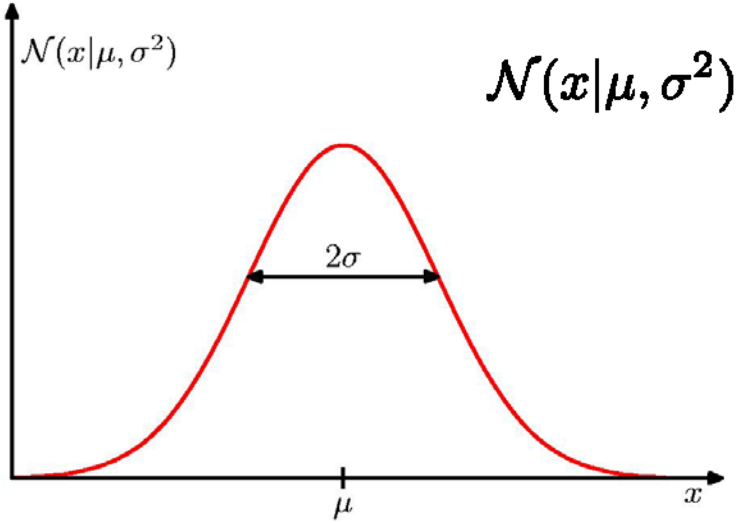
- **Dirichlet**分布可以充当多项分布中参数的共轭先验分布
  - 参数的后验分布仍为**Multinomial**分布

$$\boxed{\text{Posterior} = \text{Likelihood} \cdot \text{Prior}}$$

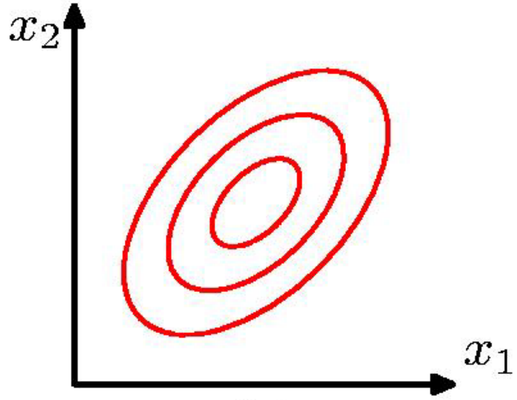
$$p(\boldsymbol{\mu}|\mathcal{D},\boldsymbol{\alpha}) \propto p(\mathcal{D}|\boldsymbol{\mu})p(\boldsymbol{\mu}|\boldsymbol{\alpha}) \propto \prod_{k=1}^{K} \mu_k^{\alpha_k+m_k-1}$$

$$\begin{aligned} p(\boldsymbol{\mu}|\mathcal{D},\boldsymbol{\alpha}) &= \mathrm{Dir}(\boldsymbol{\mu}|\boldsymbol{\alpha}+\mathbf{m}) \\ &= \frac{\Gamma(\alpha_0+N)}{\Gamma(\alpha_1+m_1)\cdots\Gamma(\alpha_K+m_K)} \prod_{k=1}^{K} \mu_k^{\alpha_k+m_k-1} \end{aligned}$$

高斯(Gaussian)分布

- 

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



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

多变量/多元高斯分布的几何

- 多元高斯分布中的去相关

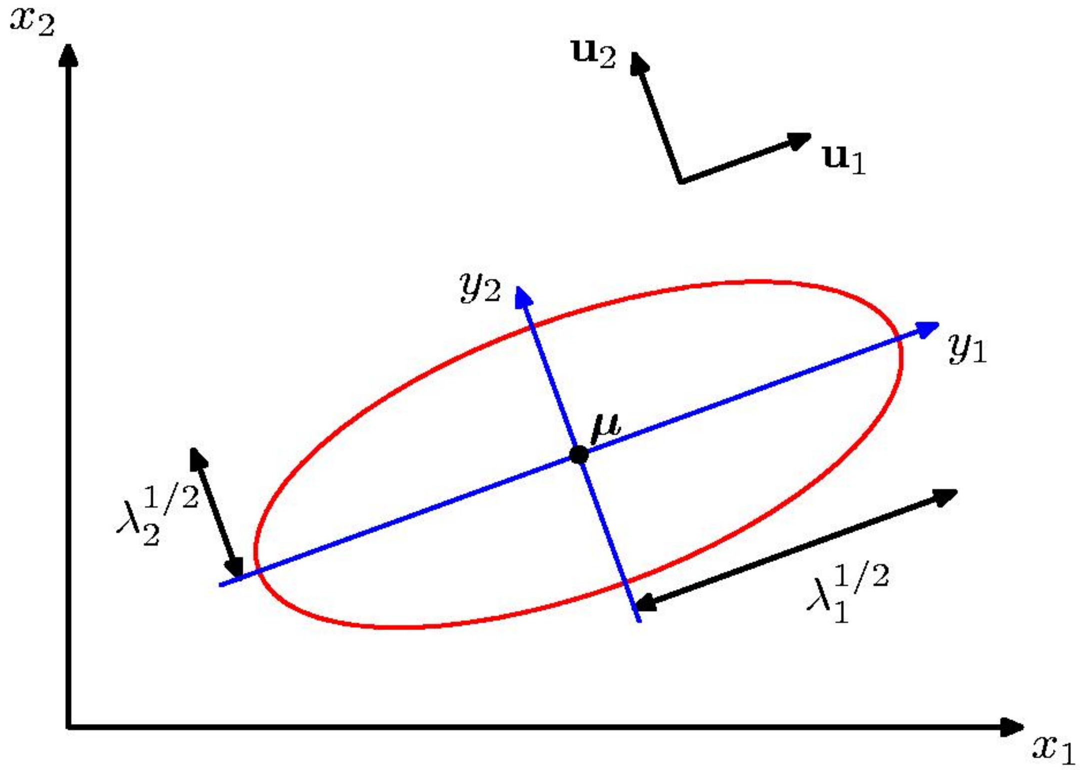
$$\boldsymbol{\Delta}^2 = (\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})$$

  - 马氏距离(Mahalanobis距离)

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

$$\boldsymbol{\Delta}^2 = \sum_{i=1}^{D}\frac{y_i^2}{\lambda_i}$$

$$y_i = \mathbf{u}_i^{\mathrm{T}}(\mathbf{x}-\boldsymbol{\mu})$$



高斯分布中的参数估计

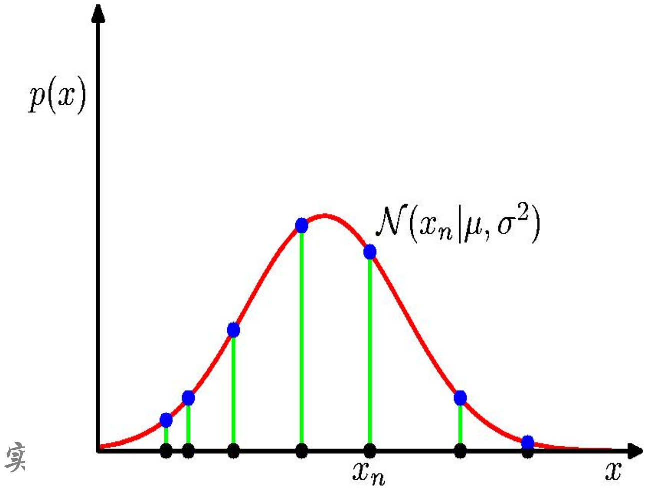
- 最大似然估计法(Maximal Likelihood Estimation)
  - 给定**N**个数据点的数据集**X**，假设**i.i.d., e.g.** 高斯分布
    - independent and identically distributed: i.i.d.
  - 数据集由特定参数下的高斯分布生成的概率为

$$p(\mathbf{x}|\mu,\sigma^2)=\prod_{n=1}^{N}\mathcal{N}\left(x_n|\mu,\sigma^2\right)$$

    - 称为似然函数(Likelihood function)
  - 最大似然估计原则：在给定数据的情况下，寻找最大化似然函数的参数

$$\max_{\mu,\sigma^2} L\left(\mu,\sigma^2 \mid X\right)=p\left(X \mid \mu,\sigma^2\right)$$

$$\max_{\mu,\sigma^2} l\left(\mu,\sigma^2 \mid X\right)=\log p\left(X \mid \mu,\sigma^2\right)$$



高斯分布参数的最大似然估计

- 目标函数:

$$\ln p\left(\mathbf{x}|\mu,\sigma^2\right) = -\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2 - \frac{N}{2}\ln\sigma^2 - \frac{N}{2}\ln(2\pi)$$

- 均值的估计

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

- 方差的估计

$$\sigma^2_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N}(x_n-\mu_{\mathrm{ML}})^2$$

高斯分布中均值的贝叶斯估计

- 假设方差已知, 给定 i. i. d. 数据 $\mathbf{x} = \{x_1, \ldots, x_N\}$ 则，似然函数为:

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

  - 假设均值服从高斯分布 $p(\mu) = \mathcal{N}\left(\mu|\mu_0, \sigma_0^2\right).$
  - 后验分布为: $\boxed{\text{Posterior} = \text{Likelihood} \cdot \text{Prior}}$

$$p(\mu|\mathbf{x}) \propto p(\mathbf{x}|\mu)p(\mu).$$

$$\Rightarrow \quad p(\mu|\mathbf{x}) = \mathcal{N}\left(\mu|\mu_N, \sigma_N^2\right)$$

高斯分布中均值的贝叶斯估计

- 后验分布

$$p(\mu|\mathbf{x}) = \mathcal{N}\left(\mu|\mu_N, \sigma_N^2\right)$$

  - 其中

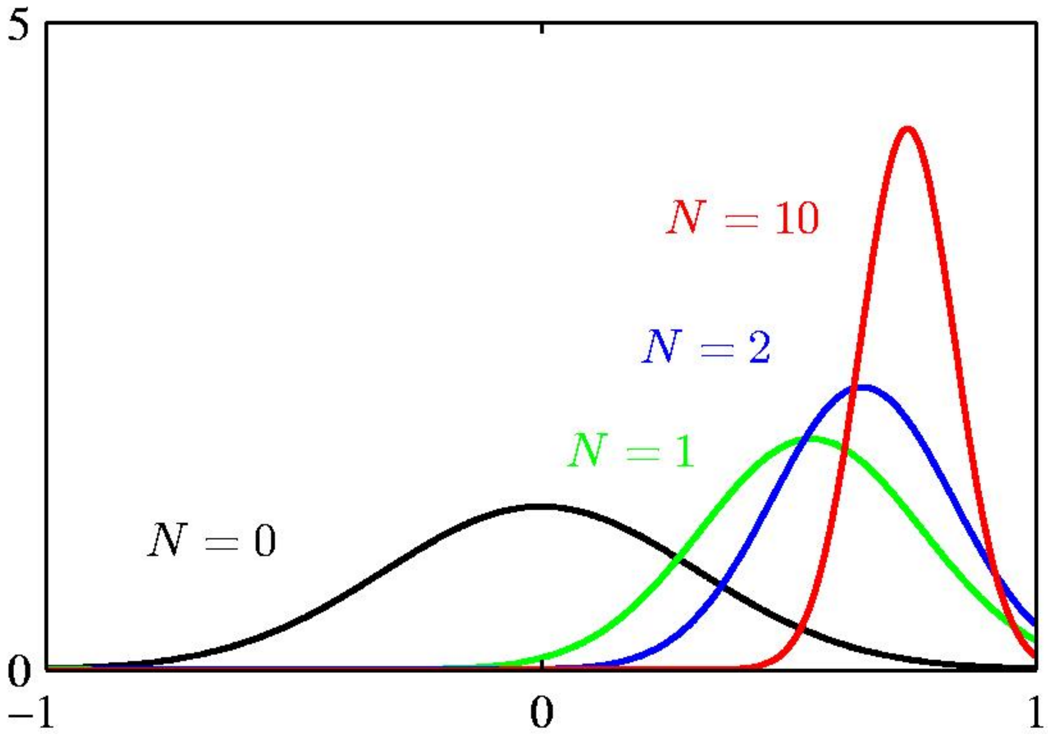
$$\begin{aligned} \mu_N &= \frac{\sigma^2}{N\sigma_0^2+\sigma^2}\mu_0 + \frac{N\sigma_0^2}{N\sigma_0^2+\sigma^2}\mu_{\mathrm{ML}}, \\ \frac{1}{\sigma_N^2} &= \frac{1}{\sigma_0^2} + \frac{N}{\sigma^2}. \end{aligned} \qquad \mu_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N} x_n$$

	$N = 0$	$N \rightarrow \infty$
μ_N	μ_0	μ_{ML}
σ_N^2	σ_0^2	0

高斯分布中均值的贝叶斯估计

- 不同训练样本数目对均值估计的影响

$$p(\mu|\mathbf{x}) = \mathcal{N}\left(\mu|\mu_N, \sigma_N^2\right)$$



高斯分布中均值的贝叶斯估计

- 序贯估计(Sequential Estimation)

$$
\begin{aligned}
p(\mu|\mathbf{x}) &\propto p(\mu)p(\mathbf{x}|\mu) \\
&= \left[p(\mu)\prod_{n=1}^{N-1} p(x_n|\mu)\right] p(x_N|\mu) \\
&\propto \mathcal{N}\left(\mu|\mu_{N-1},\sigma_{N-1}^2\right) p(x_N|\mu)
\end{aligned}
$$

– 当第**N+1**个样本到来时，旧后验概率**(**基于**N**个样本的估计**)**变成“先验”

高斯分布中方差的贝叶斯估计

- 假设均值已知, 给定 i. i. d. 数据 $\mathbf{x} = \{x_1, \ldots, x_N\}$ 则，似然函数为:

$$p(\mathbf{x}|\lambda) = \prod_{n=1}^{N} \mathcal{N}(x_n|\mu, \lambda^{-1}) \propto \lambda^{N/2} \exp\left\{-\frac{\lambda}{2}\sum_{n=1}^{N}(x_n - \mu)^2\right\}.$$

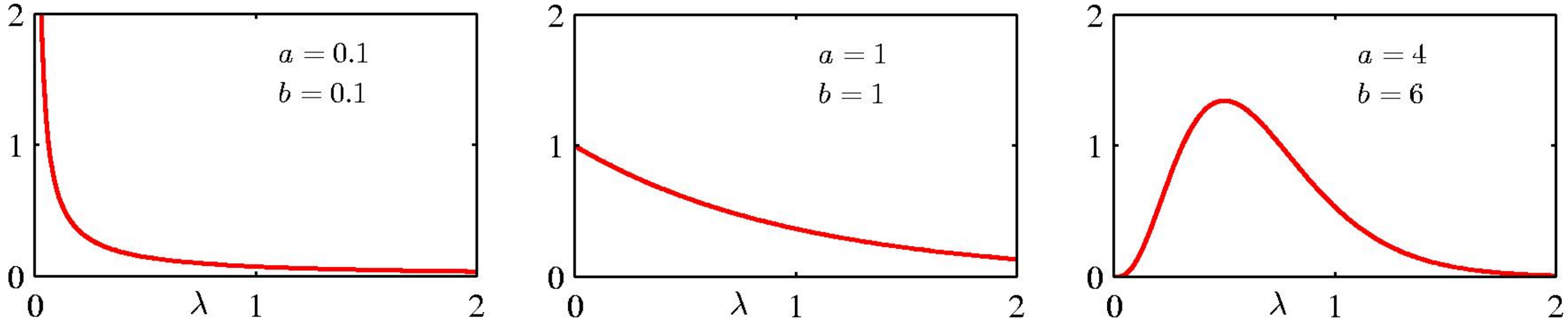
  - 注意: 似然函数具有**Gamma**分布的形式
  - 代替直接估计方差，这里估计精度**(Precision)**参数

Gamma分布

- Gamma分布具有下述形式:

$$\mathrm{Gam}(\lambda|a,b) = \frac{1}{\Gamma(a)} b^a \lambda^{a-1} \exp(-b\lambda)$$

$$\mathbb{E}[\lambda] = \frac{a}{b} \qquad \mathrm{var}[\lambda] = \frac{a}{b^2}$$



高斯分布中方差的贝叶斯估计

- 假设精度参数服从Gamma分布

$$\mathrm{Gam}(\lambda|a_0, b_0)$$

  – 则后验分布为

$$p(\lambda|\mathbf{x}) \propto \lambda^{a_0-1}\lambda^{N/2}\exp\left\{-b_0\lambda - \frac{\lambda}{2}\sum_{n=1}^{N}(x_n-\mu)^2\right\}$$

$$\Rightarrow \quad \mathrm{Gam}(\lambda|a_N, b_N)$$

$$\begin{aligned} a_N &= a_0 + \frac{N}{2} \\ b_N &= b_0 + \frac{1}{2}\sum_{n=1}^{N}(x_n-\mu)^2 = b_0 + \frac{N}{2}\sigma^2_{\mathrm{ML}}. \end{aligned}$$

高斯分布中均值-方差的贝叶斯估计

- 假设均值和方差均未知, 给定 i. i. d. 数据

  $\mathbf{x} = \{x_1, \ldots, x_N\}$

  则联合似然函数为:

$$p(\mathbf{x}|\mu, \lambda) = \prod_{n=1}^{N} \left(\frac{\lambda}{2\pi}\right)^{1/2} \exp\left\{-\frac{\lambda}{2}(x_n - \mu)^2\right\}$$

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

  如何选择先验?

高斯分布中均值-方差的贝叶斯估计

- 高斯-伽玛(Gaussian-Gamma)先验分布

$$p(\mu, \lambda) = \mathcal{N}(\mu|\mu_0, (\beta\lambda)^{-1})\mathrm{Gam}(\lambda|a, b)$$

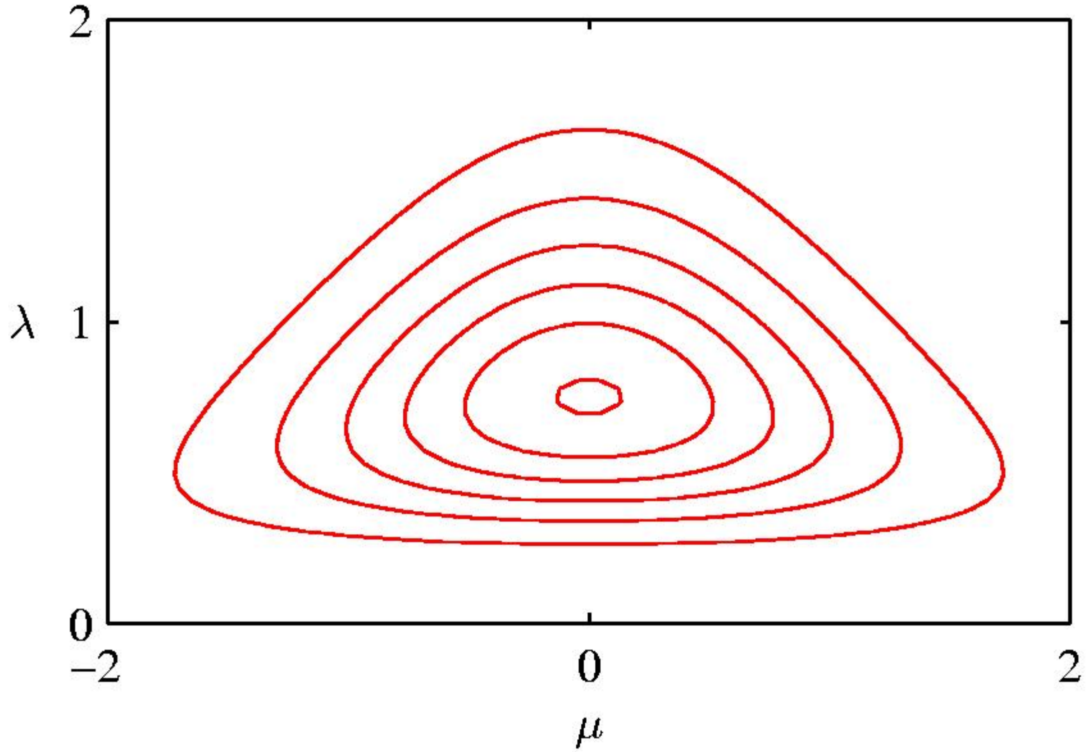
$$\propto \quad \underbrace{\exp\left\{-\frac{\beta\lambda}{2}(\mu - \mu_0)^2\right\}}\, \underbrace{\lambda^{a-1}\exp\{-b\lambda\}}$$

- Quadratic in μ.
- Linear in λ.

- Gamma distribution over λ .
- Independent of μ.

Gaussian-Gamma分布

- $$p(\mu, \lambda) = \mathcal{N}(\mu|\mu_0, (\beta\lambda)^{-1})\mathrm{Gam}(\lambda|a, b)$$
$$\propto \quad \exp\left\{-\frac{\beta\lambda}{2}(\mu-\mu_0)^2\right\}\lambda^{a-1}\exp\{-b\lambda\}$$



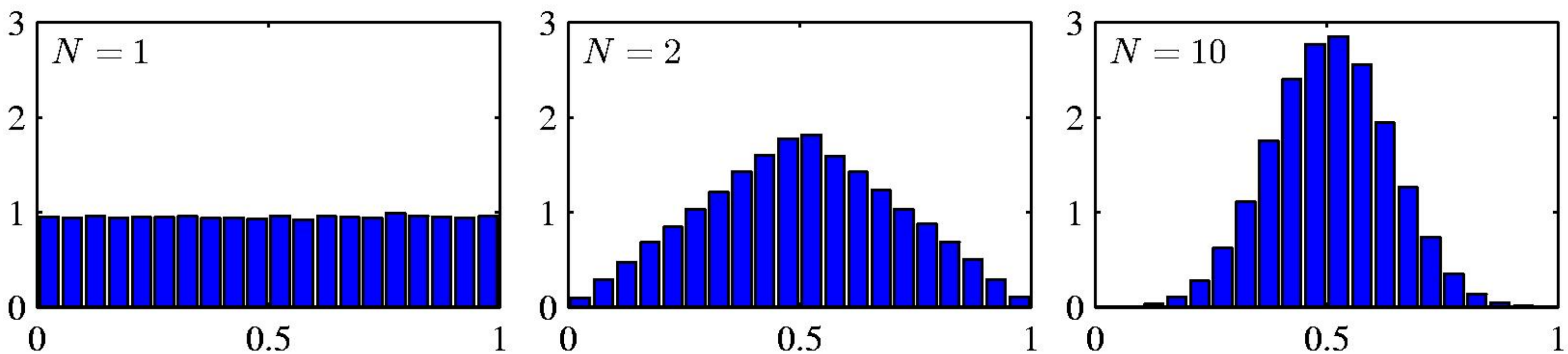
多元高斯分布的共轭先验

- 多变量高斯分布的均值/协方差矩阵的共轭先验(**Conjugate priors**)
  – 假设协方差矩阵已知，则均值向量的共轭先验为高斯分布
  – 假设均值向量已知，则协方差矩阵的共轭先验是威沙特**(**Wishart**)**分布
  $$\mathcal{W}(\mathbf{\Lambda}|\mathbf{W},\nu) = B|\mathbf{\Lambda}|^{(\nu-D-1)/2}\exp\left(-\frac{1}{2}\mathrm{Tr}(\mathbf{W}^{-1}\mathbf{\Lambda})\right).$$
  – 假设均值向量未知、协方差矩阵未知，则两者的联合共轭先验分布为高斯**-**威沙特分布

$$p(\boldsymbol{\mu},\mathbf{\Lambda}|\boldsymbol{\mu}_0,\beta,\mathbf{W},\nu) = \mathcal{N}(\boldsymbol{\mu}|\boldsymbol{\mu}_0,(\beta\mathbf{\Lambda})^{-1})\,\mathcal{W}(\mathbf{\Lambda}|\mathbf{W},\nu)$$

中心极限(Central Limit)定理

- N 个i.i.d. (独立同分布) 随机变量的和的分布，当N增大时，趋于高斯分布
  - 举例: **N个在[0,1]区间内均匀(uniform)分布的随机变量的和的分布**

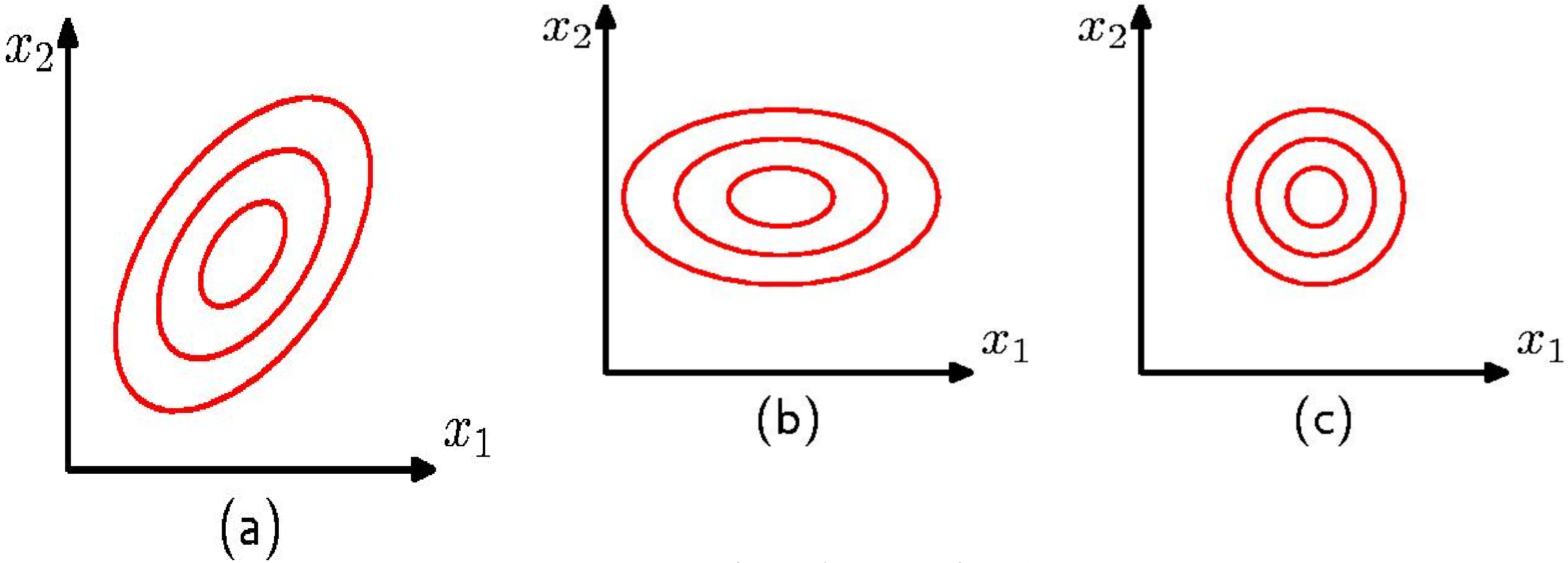


多元高斯分布的矩

- 二阶矩和协方差矩阵

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

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



多元高斯随机变量的分块

- 假设x服从多元高斯分布

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

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

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

- 条件分布和边缘分布仍然为高斯分布.

多元高斯随机变量的分块

- 条件分布仍为高斯分布，且

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

  - 凑全平方项: completing the square

$$-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) = -\frac{1}{2}\mathbf{x}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{x} + \mathbf{x}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \text{const}$$

$$\begin{aligned}
-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) = \\
-\frac{1}{2}(\mathbf{x}_a-\boldsymbol{\mu}_a)^{\mathrm{T}}\boldsymbol{\Lambda}_{aa}(\mathbf{x}_a-\boldsymbol{\mu}_a) - \frac{1}{2}(\mathbf{x}_a-\boldsymbol{\mu}_a)^{\mathrm{T}}\boldsymbol{\Lambda}_{ab}(\mathbf{x}_b-\boldsymbol{\mu}_b) \\
-\frac{1}{2}(\mathbf{x}_b-\boldsymbol{\mu}_b)^{\mathrm{T}}\boldsymbol{\Lambda}_{ba}(\mathbf{x}_a-\boldsymbol{\mu}_a) - \frac{1}{2}(\mathbf{x}_b-\boldsymbol{\mu}_b)^{\mathrm{T}}\boldsymbol{\Lambda}_{bb}(\mathbf{x}_b-\boldsymbol{\mu}_b).
\end{aligned}$$

分块矩阵求逆运算的恒等式

- 使用到下列矩阵恒等式

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

  – 其中 $\mathbf{M} = (\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1}$

  – 推导: 基于分块矩阵方程的块消除

分块的多元高斯随机变量

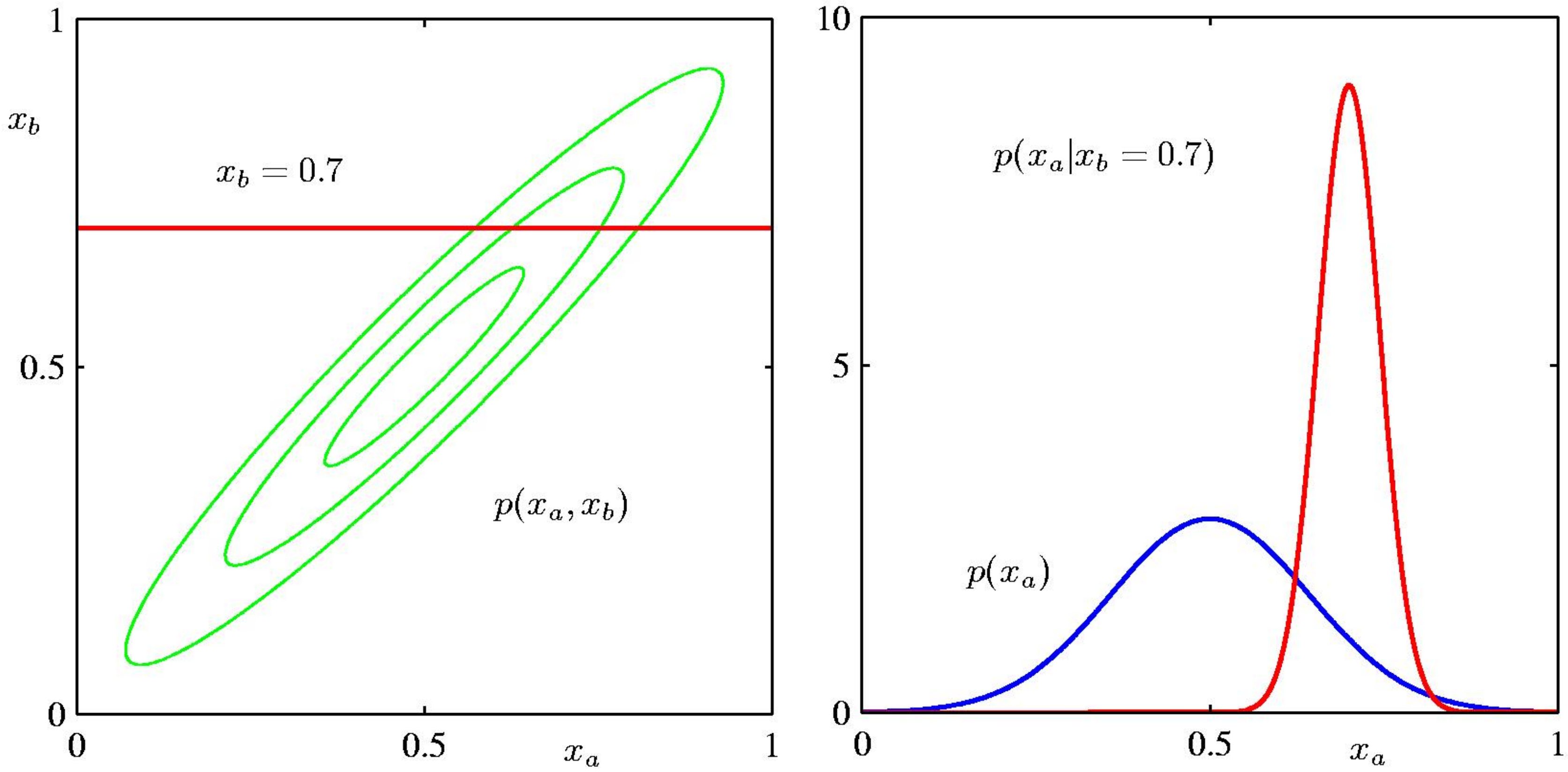
- $$p(\mathbf{x}_a|\mathbf{x}_b) = \mathcal{N}(\mathbf{x}_a|\boldsymbol{\mu}_{a|b}, \boldsymbol{\Sigma}_{a|b})$$

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

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

分块的条件分布和边缘分布

-



Bayes定理用于高斯随机变量

- 边缘分布密度和条件概率密度

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

$$
\begin{aligned}
p(\mathbf{y}) &= \mathcal{N}(\mathbf{y}|\mathbf{A}\boldsymbol{\mu}+\mathbf{b}, \mathbf{L}^{-1}+\mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^{\mathrm{T}}) \\
p(\mathbf{x}|\mathbf{y}) &= \mathcal{N}(\mathbf{x}|\boldsymbol{\Sigma}\{\mathbf{A}^{\mathrm{T}}\mathbf{L}(\mathbf{y}-\mathbf{b})+\boldsymbol{\Lambda}\boldsymbol{\mu}\}, \boldsymbol{\Sigma})
\end{aligned}
$$

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

贝叶斯线性回归 内容提要

- 引子: 曲线拟合问题
- 常用的几种分布
- 贝叶斯线性回归
  - 两个例子
  - 等价核
  - 先验分布中的超参数的处理

贝叶斯线性回归

- 假设观测数据来采样于一个确定性(**deterministic**) 函数，训练数据中存在i.i.d.加性高斯噪声(**additive Gaussian noise**)

$$\boxed{t = y(\mathbf{x}, \mathbf{w}) + \epsilon \quad \text{where} \quad p(\epsilon|\beta) = \mathcal{N}(\epsilon|0, \beta^{-1})}$$

  – 由于噪声的存在，**t** 的取值具有不确定性，因此我们使用下述高斯分布:

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

THE NEXT:

- 确定先验分布
- 根据训练数据推理后验分布
- 计算回归模型的预测性分布

先验分布

- 共轭先验(Conjugate prior):
  - 一种特定形式的先验分布，它能使后验分布与先验分布具有相同的泛函形式，从而简化后续的贝叶斯分析

$$p(\mathbf{w}|\mathbf{x},\mathbf{t},\alpha,\beta) \propto p(\mathbf{t}|\mathbf{x},\mathbf{w},\beta)p(\mathbf{w}|\alpha)$$

- 举例:
  - **Bernoulli**分布的共轭先验是**Beta**分布
  - 多项分布的共轭先验是狄利克雷**(**Dirichlet**)**分布
  - 高斯分布的共轭先验是高斯分布

贝叶斯线性回归

- 先验分布
  - 假设观测数据中存在**i.i.d.**加性高斯噪声 **(**additive Gaussian noise**)**
  - 选择共轭先验：$p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_0, \mathbf{S}_0).$
    - 一种特定形式的先验分布，使得后验分布与先验分布具有相同的泛函形式，从而简化后续的贝叶斯分析
- 后验分布

$$\boxed{\text{Posterior} = \text{Likelihood} \cdot \text{Prior}}$$

$$\Rightarrow p(\mathbf{w}|\mathbf{t}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N)$$

其中

$$\begin{aligned} \mathbf{m}_N &= \mathbf{S}_N \left(\mathbf{S}_0^{-1}\mathbf{m}_0 + \beta\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{t}\right) \\ \mathbf{S}_N^{-1} &= \mathbf{S}_0^{-1} + \beta\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}. \end{aligned}$$

贝叶斯线性回归

- 通常选择先验分布为

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I})$$

- 则后验分布为:

$$p(\mathbf{w}|\mathbf{t}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N)$$

  - 其中 $$\begin{aligned} \mathbf{m}_N &= \beta \mathbf{S}_N \boldsymbol{\Phi}^{\mathrm{T}} \mathbf{t} \\ \mathbf{S}_N^{-1} &= \alpha \mathbf{I} + \beta \boldsymbol{\Phi}^{\mathrm{T}} \boldsymbol{\Phi}. \end{aligned}$$

Bayesian Estimation:
得到的不是参数w本身,
而是w的后验分布

  - 后验分布的对数为:

$$\ln p(\mathbf{w}|\mathbf{t}) = -\frac{\beta}{2} \sum_{n=1}^{N} \{t_n - \mathbf{w}^{\mathrm{T}} \phi(\mathbf{x}_n)\}^2 - \frac{\alpha}{2} \mathbf{w}^{\mathrm{T}} \mathbf{w} + \text{const}$$

Maximum A Posterior: 使用w的后验分布的最大值所对应的w

贝叶斯线性回归 内容提要

- 引子: 曲线拟合问题
- 常用的几种分布
- 贝叶斯线性回归
  – 两个例子
  – 等价核
  – 先验分布中的超参数的处理

考虑一个例子:

- 数据生成:

$$t = f\left(x, \mathbf{a}\right) + \varepsilon$$

$$\text{其中 } x \sim U\left[-1,1\right], \quad f\left(x, \mathbf{a}\right) = a_0 + a_1 x = -0.3 + 0.5x$$

$$\varepsilon \sim N\left(0, \sigma^2\right)$$

- 线性回归模型:

$$t = y\left(x, \mathbf{w}\right) + \varepsilon = \mathbf{w}^T x + \varepsilon$$

– 任务: 基于给定数据，估计线性回归模型
  - 如果使用MLE: 即确定参数w
  - 如果使用MAP估计: 即确定参数w

考虑一个例子:

- 数据生成:

$$t = f\left(x, \mathbf{a}\right) + \varepsilon$$

其中 $x \sim U\left[-1,1\right], \quad f\left(x, \mathbf{a}\right) = a_0 + a_1 x = -0.3 + 0.5x$

$$\varepsilon \sim N\left(0, \sigma^2\right)$$

- 线性回归模型:

$$t = y\left(x, \mathbf{w}\right) + \varepsilon = \mathbf{w}^T x + \varepsilon \qquad t \sim N\left(t \mid \mathbf{w}^T x, \sigma^2\right)$$

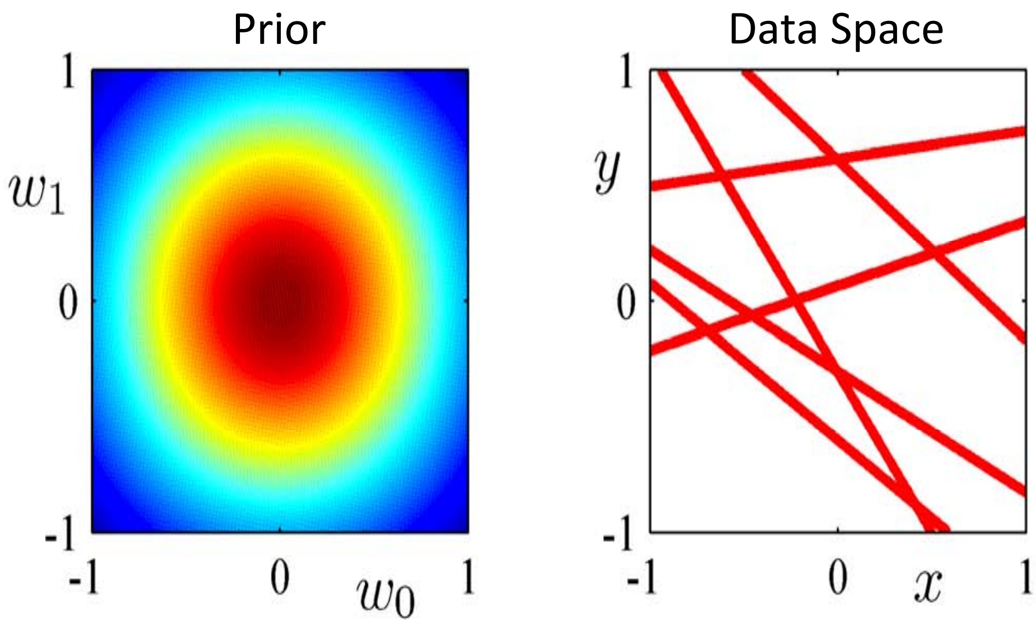
  – 任务: 基于给定数据，估计线性回归模型
    - 在贝叶斯估计中, 我们基于给定训练数据去推理参数w的后验分布

$$\boxed{p\left(\mathbf{w} \mid \mathbf{x}, \mathbf{t}\right)}$$

      – 从后验分布中可以采样很多w

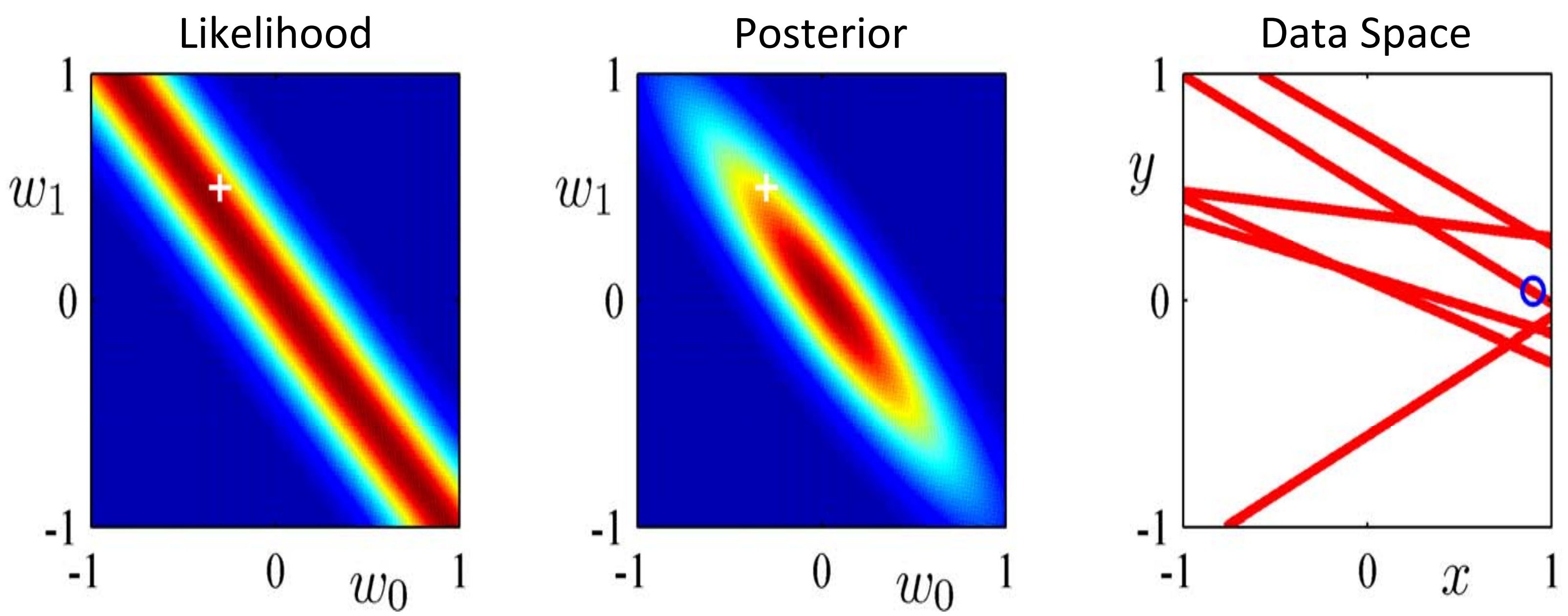
Bayesian Linear Regression (1)

0 data points observed



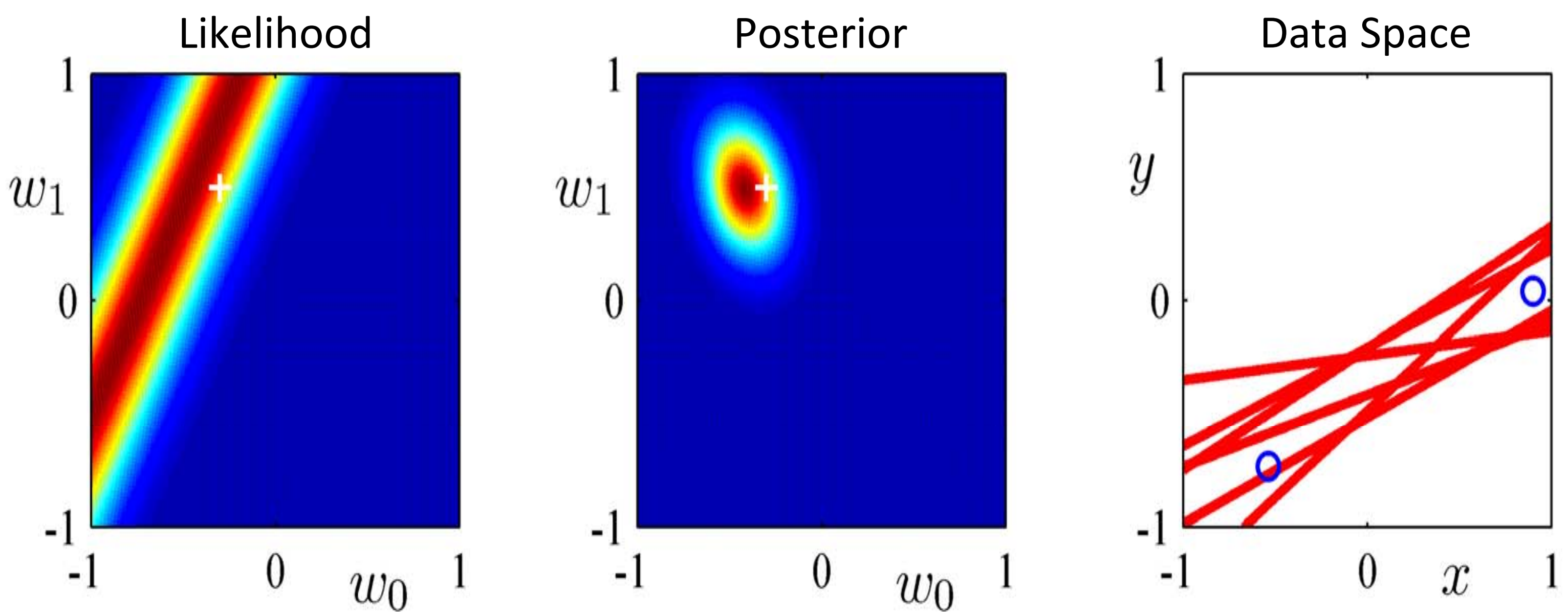
Bayesian Linear Regression (2)

1 data point observed



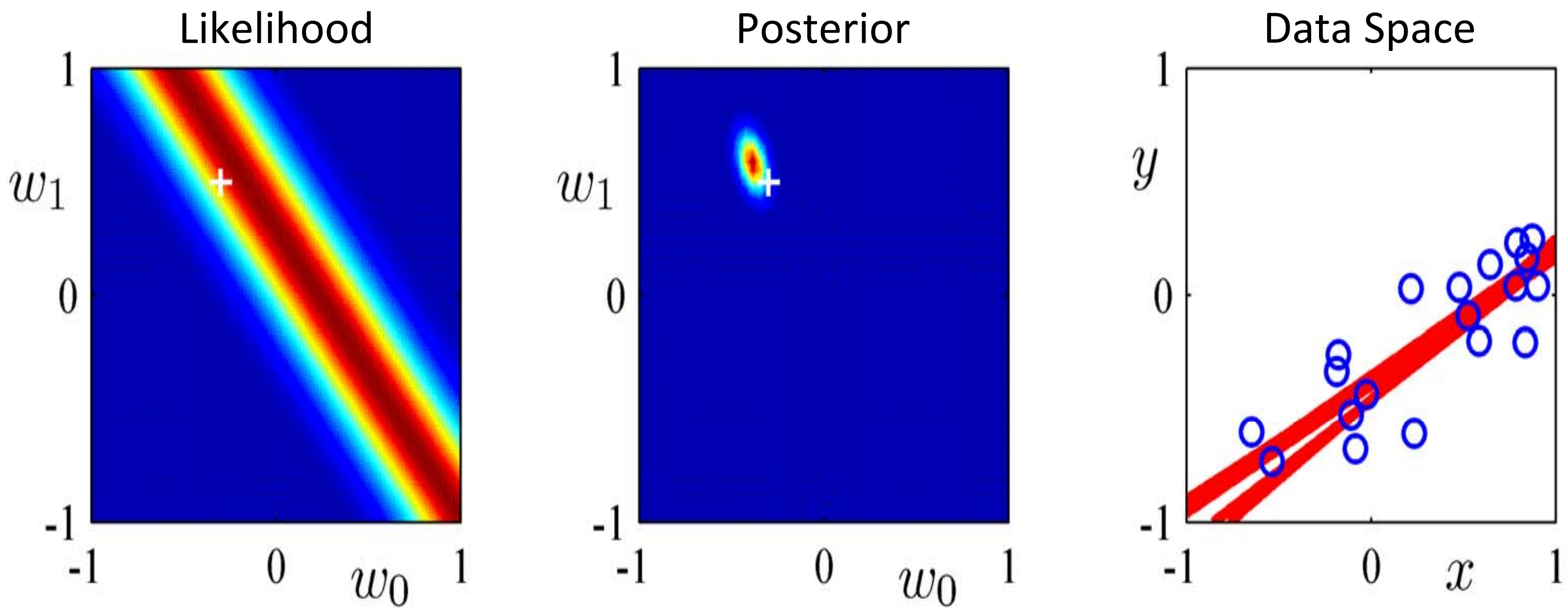
Bayesian Linear Regression (3)

2 data points observed



Bayesian Linear Regression (4)

20 data points observed



预测性分布(Predictive distribution)

- 给定一个新的x, 预测其输出t:
  - “使用所有**w**进行平均”

$$
\begin{aligned}
p(t|\mathbf{t},\alpha,\beta) &= \int p(t|\mathbf{w},\beta)p(\mathbf{w}|\mathbf{t},\alpha,\beta)\,\mathrm{d}\mathbf{w} \\
&= \mathcal{N}(t|\mathbf{m}_N^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}),\sigma_N^2(\mathbf{x}))
\end{aligned}
$$

  - 其中

$$
\sigma_N^2(\mathbf{x}) = \frac{1}{\beta} + \boldsymbol{\phi}(\mathbf{x})^{\mathrm{T}}\mathbf{S}_N\boldsymbol{\phi}(\mathbf{x}).
$$

例：使用高斯基函数的回归问题

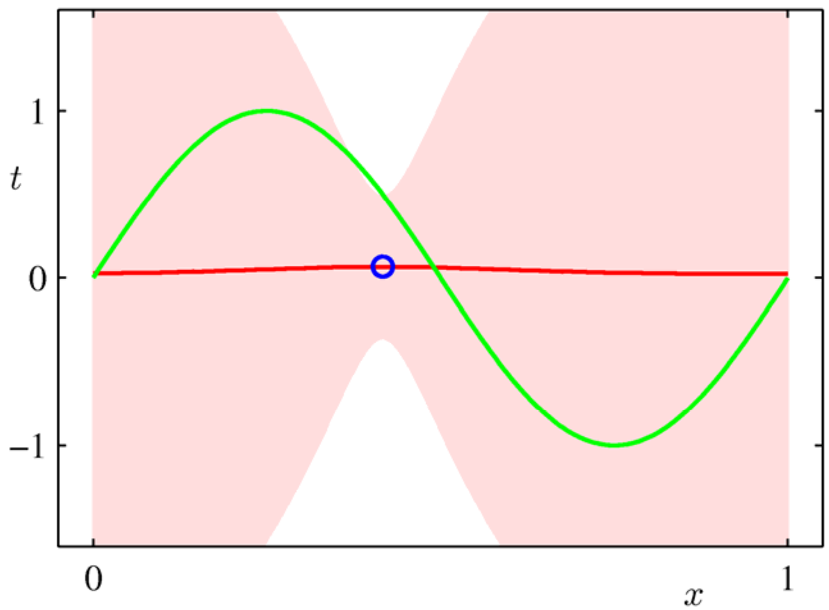
- 假设训练数据采样于一个正弦曲线
  - 我们使用9个高斯函数作为基函数
  - 使用不同数目的训练数据，可以获得不同的预测性分布

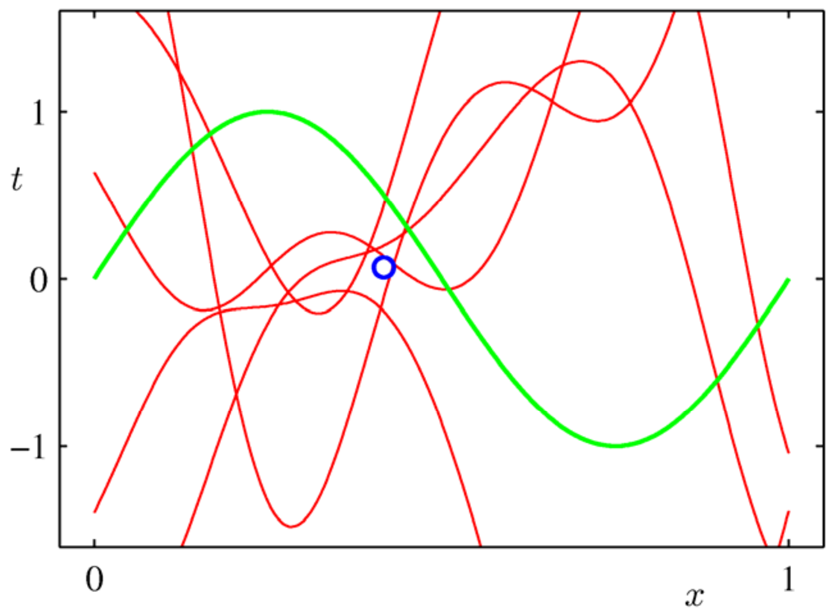
$$
\begin{aligned}
p(t|\mathbf{t},\alpha,\beta) &= \int p(t|\mathbf{w},\beta)p(\mathbf{w}|\mathbf{t},\alpha,\beta)\,\mathrm{d}\mathbf{w} \\
&= \mathcal{N}(t|\mathbf{m}_N^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}),\sigma_N^2(\mathbf{x}))
\end{aligned}
$$

例：使用高斯基函数的回归问题

- 使用1个训练样本获得的预测性分布

$$
\begin{aligned}
p(t|\mathbf{t},\alpha,\beta) &= \int p(t|\mathbf{w},\beta)p(\mathbf{w}|\mathbf{t},\alpha,\beta)\,\mathrm{d}\mathbf{w} \\
&= \mathcal{N}(t|\mathbf{m}_N^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}),\sigma_N^2(\mathbf{x}))
\end{aligned}
$$

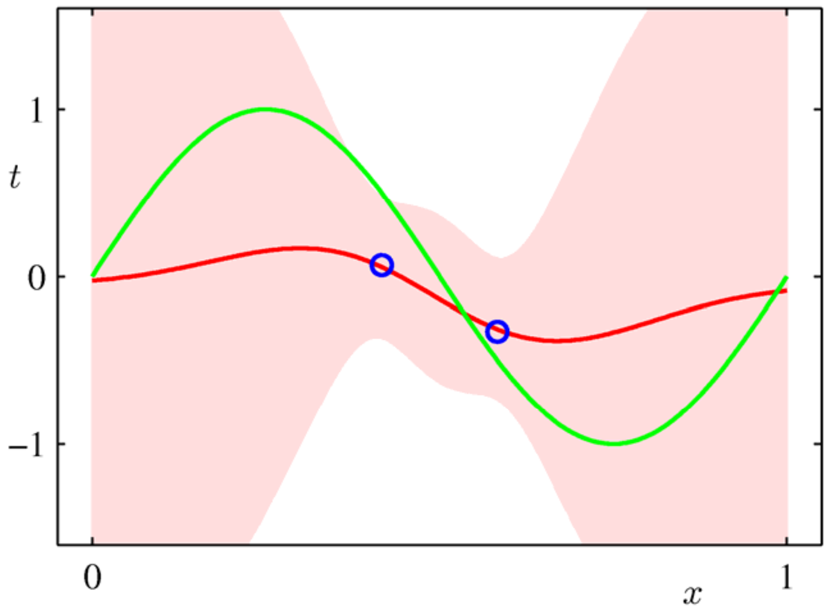


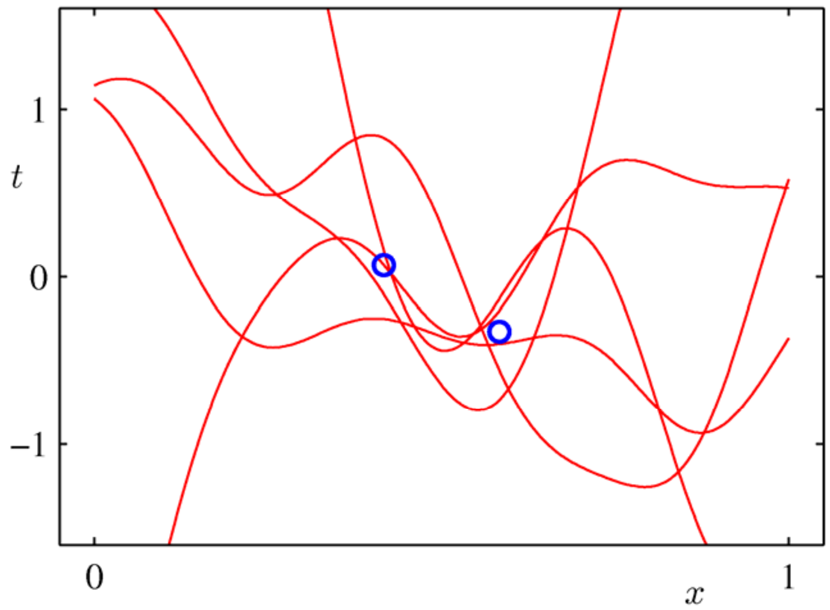


例：使用高斯基函数的回归问题

- 使用2个训练样本获得的预测性分布

$$
\begin{aligned}
p(t|\mathbf{t},\alpha,\beta) &= \int p(t|\mathbf{w},\beta)p(\mathbf{w}|\mathbf{t},\alpha,\beta)\,\mathrm{d}\mathbf{w} \\
&= \mathcal{N}(t|\mathbf{m}_N^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}),\sigma_N^2(\mathbf{x}))
\end{aligned}
$$

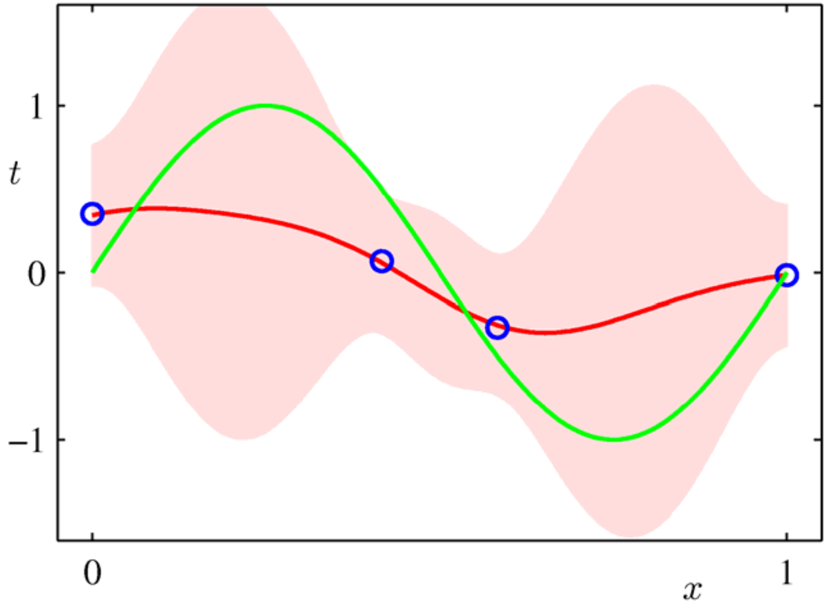


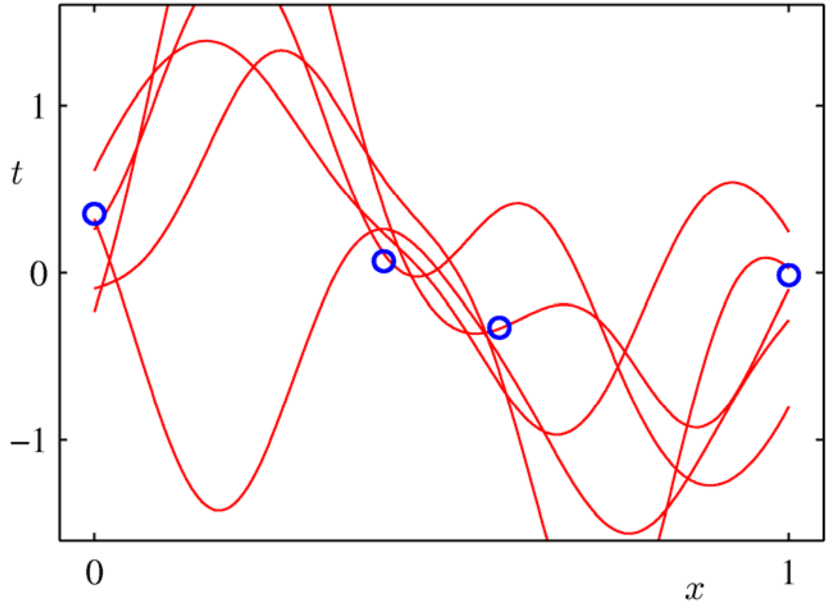


例：使用高斯基函数的回归问题

- 使用4个训练样本获得的预测性分布

$$
\begin{aligned}
p(t|\mathbf{t}, \alpha, \beta) &= \int p(t|\mathbf{w}, \beta) p(\mathbf{w}|\mathbf{t}, \alpha, \beta) \, \mathrm{d}\mathbf{w} \\
&= \mathcal{N}(t|\mathbf{m}_N^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{x}), \sigma_N^2(\mathbf{x}))
\end{aligned}
$$

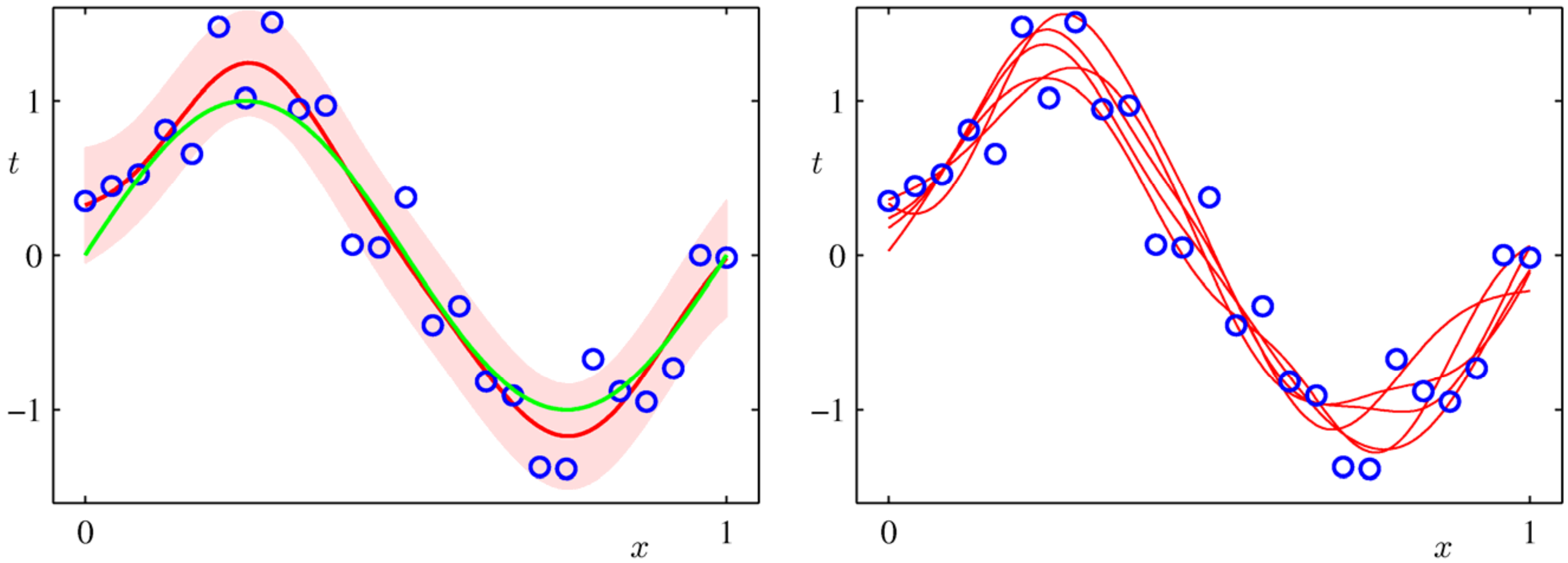




例：使用高斯基函数的回归问题

- 使用25个训练样本获得的预测性分布

$$
\begin{aligned}
p(t|\mathbf{t},\alpha,\beta) &= \int p(t|\mathbf{w},\beta)p(\mathbf{w}|\mathbf{t},\alpha,\beta)\,\mathrm{d}\mathbf{w} \\
&= \mathcal{N}(t|\mathbf{m}_N^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}),\sigma_N^2(\mathbf{x}))
\end{aligned}
$$



贝叶斯线性回归 内容提要

- 引子: 曲线拟合问题
- 常用的几种分布
- 贝叶斯线性回归
  - 两个例子
  - 等价核
  - 先验分布中的超参数的处理

等价核 (Equivalent Kernel)

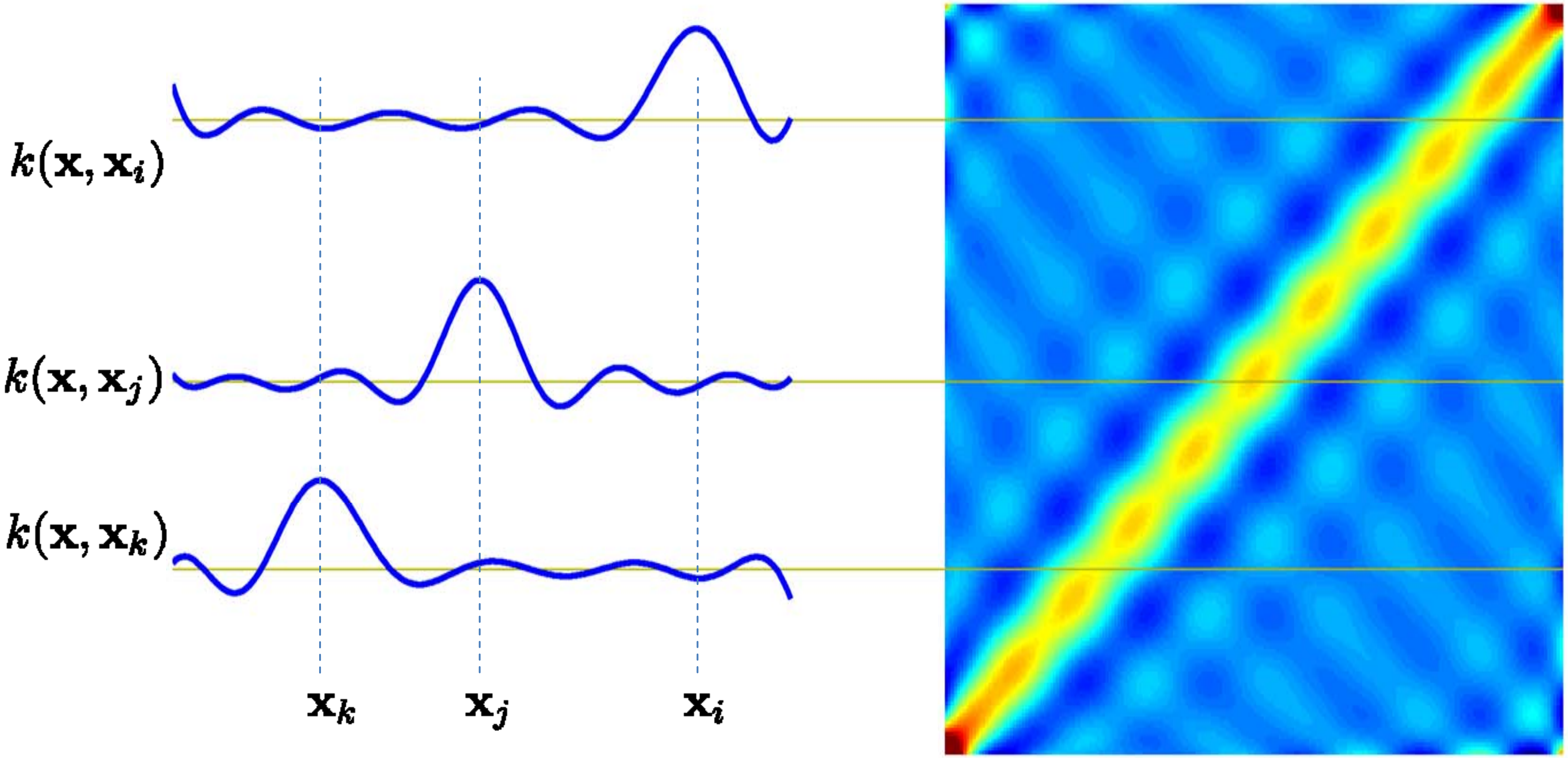
The predictive mean can be written

$$
\begin{aligned}
y(\mathbf{x}, \mathbf{m}_N) &= \mathbf{m}_N^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{x}) = \beta \boldsymbol{\phi}(\mathbf{x})^{\mathrm{T}} \mathbf{S}_N \boldsymbol{\Phi}^{\mathrm{T}} \mathbf{t} \\
&= \sum_{n=1}^{N} \beta \boldsymbol{\phi}(\mathbf{x})^{\mathrm{T}} \mathbf{S}_N \boldsymbol{\phi}(\mathbf{x}_n) t_n \\
&= \sum_{n=1}^{N} k(\mathbf{x}, \mathbf{x}_n) t_n .
\end{aligned}
$$

Equivalent kernel or *smoother matrix.*

This is a weighted sum of the training data target values, t_n.

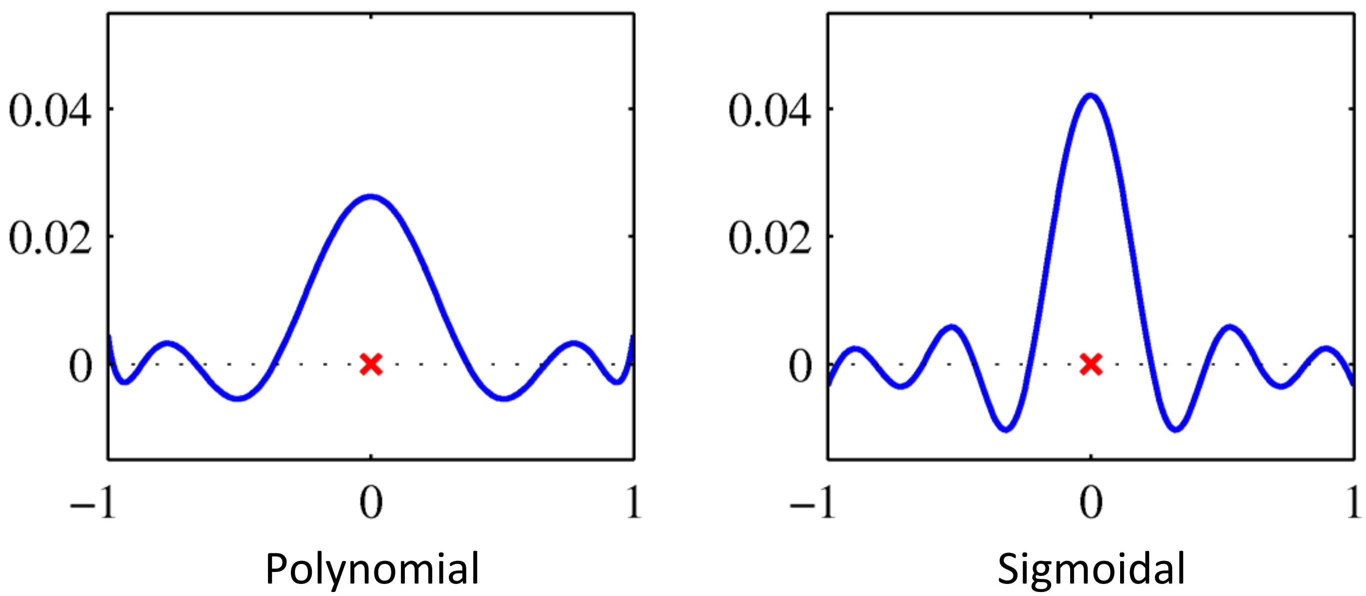
等价核 (Equivalent Kernel)



- t_n对应的权值依赖于x到x_n的距离，距 x_n越近权值越大

等价核 (Equivalent Kernel)

- 非局部基(Nonlocal basis)也对应局部等价核



等价核 (Equivalent Kernel)

- 等价核作为协方差矩阵函数(**covariance function**)

$$
\begin{aligned}
\operatorname{cov}[y(\mathbf{x}), y(\mathbf{x}')] &= \operatorname{cov}[\boldsymbol{\phi}(\mathbf{x})^{\mathrm{T}}\mathbf{w}, \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}')] \\
&= \boldsymbol{\phi}(\mathbf{x})^{\mathrm{T}}\mathbf{S}_N\boldsymbol{\phi}(\mathbf{x}') = \beta^{-1}k(\mathbf{x}, \mathbf{x}').
\end{aligned}
$$

  – 若不引入基函数，而直接定义核函数，则得出高斯过程(**Gaussian Processes**)

等价核 (Equivalent Kernel)

- 基本性质：
  - 归一化
    $$\sum_{n=1}^{N} k(\mathbf{x}, \mathbf{x}_n) = 1$$
    - 注意到等价核的值可正可负
  - 等价核也可以写成内积形式
    $$k(\mathbf{x}, \mathbf{z}) = \boldsymbol{\psi}(\mathbf{x})^{\mathrm{T}} \boldsymbol{\psi}(\mathbf{z})$$
    - 其中
    $$\boldsymbol{\psi}(\mathbf{x}) = \beta^{1/2} \mathbf{S}_N^{1/2} \boldsymbol{\phi}(\mathbf{x})$$

贝叶斯线性回归 内容提要

- 引子: 曲线拟合问题
- 常用的几种分布
- 贝叶斯线性回归
  - 两个例子
  - 等价核
  - 先验分布中的超参数的处理

先验分布中的超参数的处理方法

- 完整的贝叶斯预测性分布为:

$$p(t|\mathbf{t}) = \iiint p(t|\mathbf{w}, \beta) p(\mathbf{w}|\mathbf{t}, \alpha, \beta) p(\alpha, \beta|\mathbf{t}) \, \mathrm{d}\mathbf{w} \, \mathrm{d}\alpha \, \mathrm{d}\beta$$

  – 但是，由于积分无法解析地进行，我们使用近似方法

$$p(t|\mathbf{t}) \simeq p\left(t|\mathbf{t}, \widehat{\alpha}, \widehat{\beta}\right) = \int p\left(t|\mathbf{w}, \widehat{\beta}\right) p\left(\mathbf{w}|\mathbf{t}, \widehat{\alpha}, \widehat{\beta}\right) \mathrm{d}\mathbf{w}$$

  – 其中 $\left(\widehat{\alpha}, \widehat{\beta}\right)$ 是超参数联合分布 $p(\alpha, \beta|\mathbf{t})$ 的模 **(mode)**，假设超参数的分布**sharply peaked**
  – 这种方法被称为:
    - **Empirical Bayes**
    - Type II or generalized maximum likelihood
    - **Evidence approximation**

如何估计超参数 (1)

- 由贝叶斯定理，得出

$$p(\alpha, \beta|\mathbf{t}) \propto p(\mathbf{t}|\alpha, \beta)p(\alpha, \beta)$$

- 如果假设超参数的联合分布是均匀分布，则可以得出

$$\begin{aligned} p(\alpha, \beta|\mathbf{t}) &\propto p(\mathbf{t}|\alpha, \beta) \\ &= \int p(\mathbf{t}|\mathbf{w}, \beta)p(\mathbf{w}|\alpha)\,\mathrm{d}\mathbf{w}. \end{aligned}$$

  – 若积分中的两项来自高斯分布，则有

$$\ln p(\mathbf{t}|\alpha, \beta) = \frac{M}{2}\ln\alpha + \frac{N}{2}\ln\beta - E(\mathbf{m}_N) + \frac{1}{2}\ln|\mathbf{S}_N| - \frac{N}{2}\ln(2\pi).$$

如何估计超参数 (2)

- 寻找超参数，使得 $\ln p(\mathbf{t}|\alpha, \beta)$ 最大化

- 求微分，令其为0，则得到

$$\alpha = \frac{\gamma}{\mathbf{m}_N^{\mathrm{T}} \mathbf{m}_N}$$

$$\frac{1}{\beta} = \frac{1}{N-\gamma} \sum_{n=1}^{N} \left\{ t_n - \mathbf{m}_N^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{x}_n) \right\}^2$$

– 其中 $\left(\beta \boldsymbol{\Phi}^{\mathrm{T}} \boldsymbol{\Phi}\right) \mathbf{u}_i = \lambda_i \mathbf{u}_i, \qquad \gamma = \sum_i \frac{\lambda_i}{\alpha + \lambda_i}.$

Q / A

- Any Questions...