MACHINE LEARNING

CHAPTER 2: PROBABILITY DISTRIBUTIONS

Learning Objectives

1、 What are binary, multinomial and Gaussian distributions and their conjugate prior distributions?

2、What are the common properties of Gaussian distributions?

3、What are exponential families and their properties?

4、How to choose non-informative prior*?

5、How to use non-parametric methods for learning?

6、What are KNN based methods?

Outlines

- Binary Distributions
- Multinomial Distributions
- Gaussian Distributions
- Exponential Families
- Non-informative Prior
- Non-parametric Methods
- KNN

Parametric Distributions

Basic building blocks: $p(\mathbf{x}|\boldsymbol{\theta})$

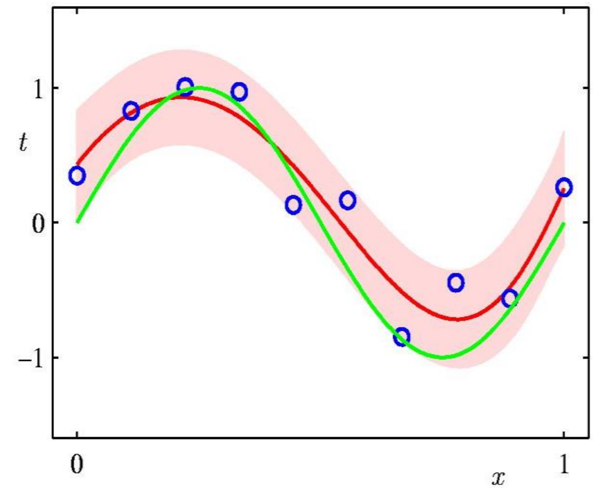
Need to determine $\boldsymbol{\theta}$ given $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$

Representation: $\boldsymbol{\theta}^\star$ or $p(\boldsymbol{\theta})$?

$$p(\theta|\mathbf{x}) \propto p(\mathbf{x}|\theta)\, p(\theta)$$

Recall Curve Fitting

$$p(t|x, \mathbf{x}, \mathbf{t}) = \int p(t|x, \mathbf{w}) p(\mathbf{w}|\mathbf{x}, \mathbf{t})\, \mathrm{d}\mathbf{w}$$



Binary Variables (1)

Coin flipping: heads=1, tails=0

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

Bernoulli Distribution

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

Binary Variables (2)

N coin flips:

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

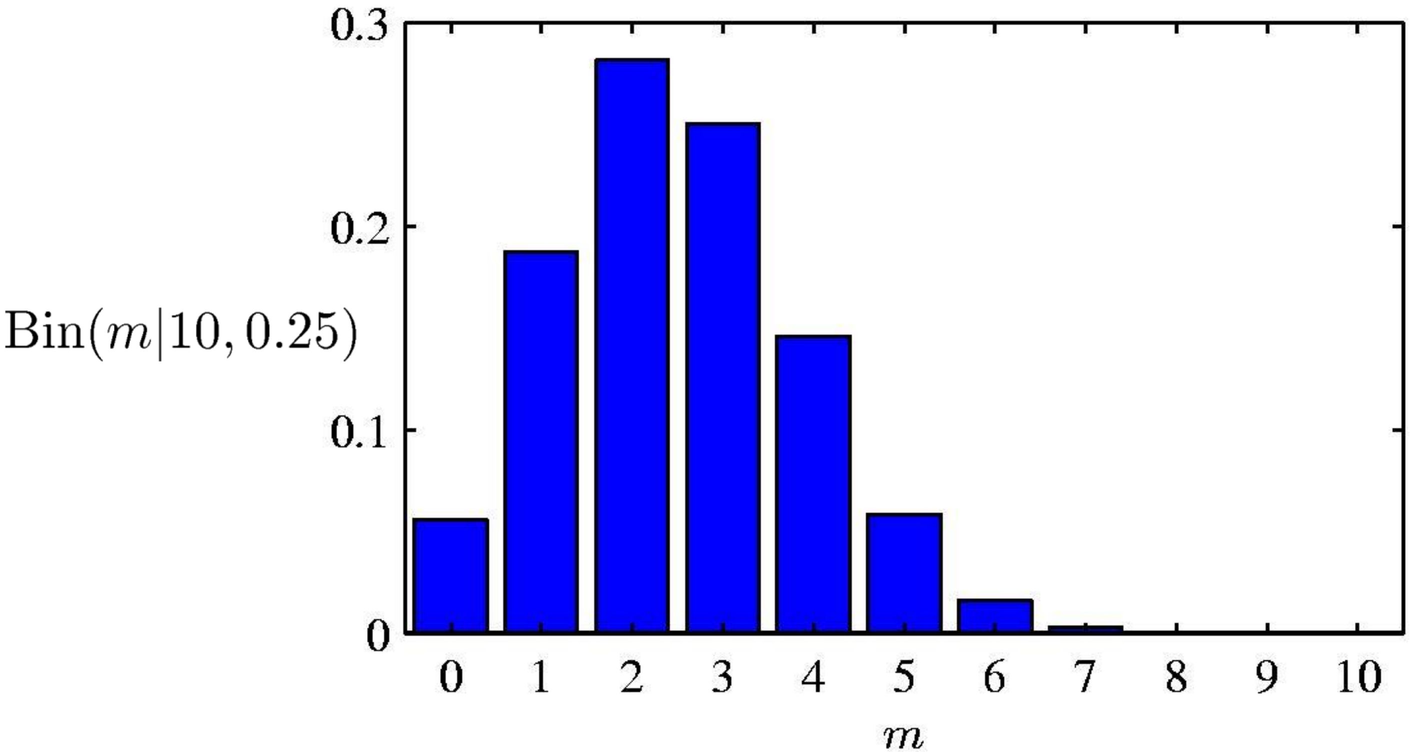
Binomial Distribution

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

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

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

Binomial Distribution



Parameter Estimation (1)

ML for Bernoulli

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

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

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

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

Parameter Estimation (2)

Example: $$\mathcal{D} = \{1, 1, 1\} \rightarrow \mu_{\mathrm{ML}} = \frac{3}{3} = 1$$

Prediction: *all* future tosses will land heads up

Overfitting to $\mathcal{D}$

Beta Distribution

Distribution over $\mu \in [0, 1]$.

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

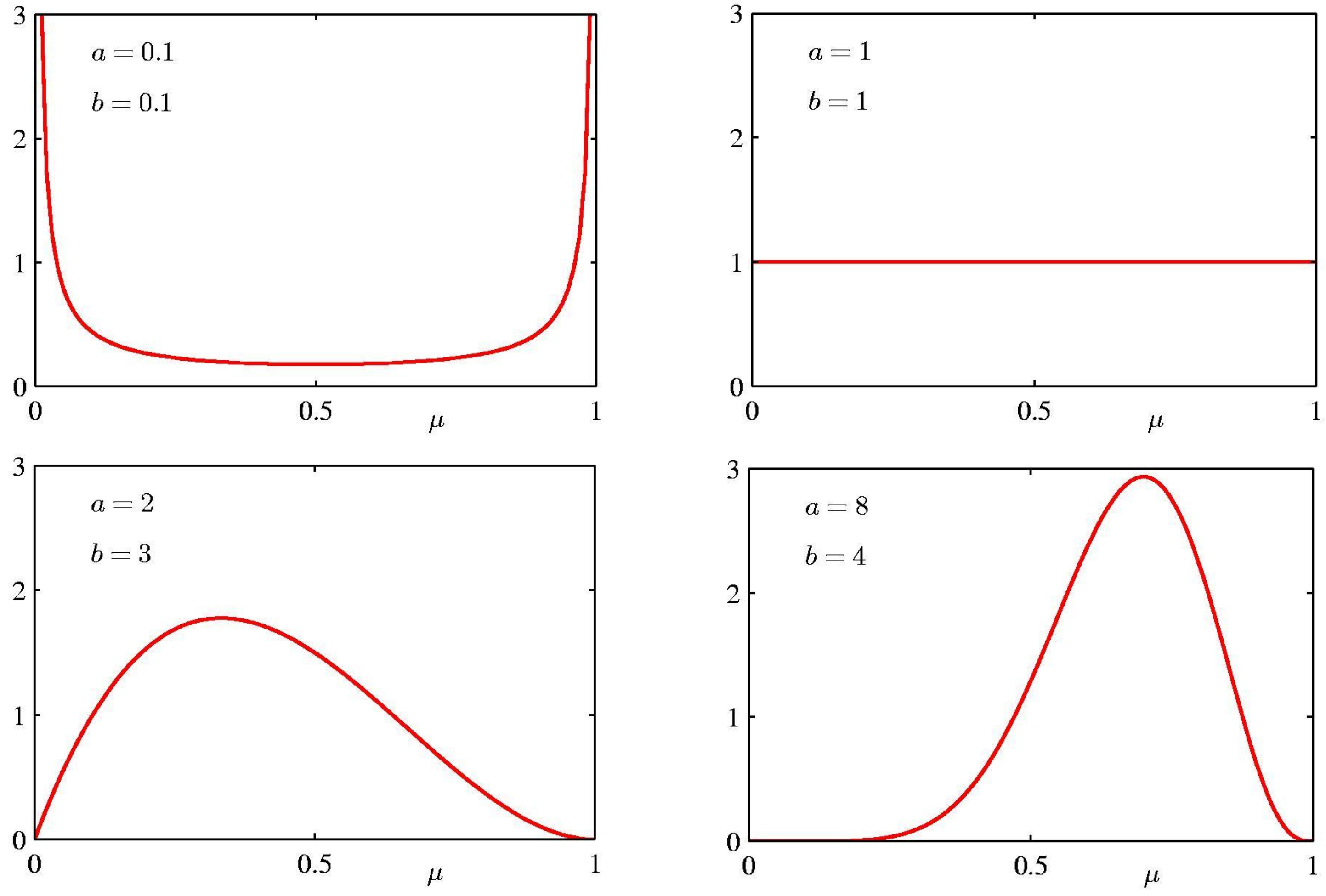
Bayesian Bernoulli

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

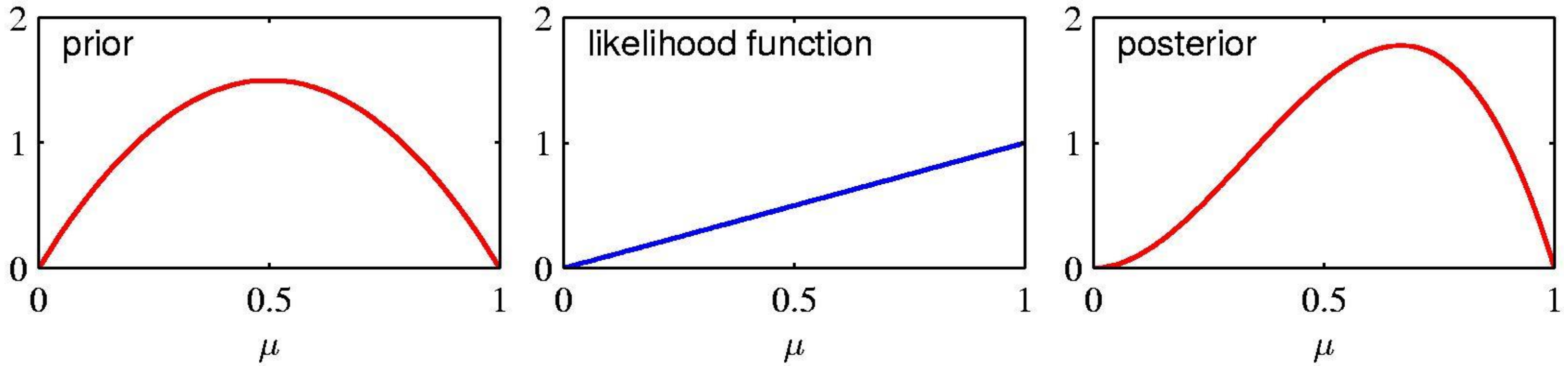
$$
a_N = a_0 + m \qquad b_N = b_0 + (N - m)
$$

The Beta distribution provides the *conjugate* prior for the Bernoulli distribution.

Beta Distribution



Prior · Likelihood = Posterior



Properties of the Posterior

As the size of the data set, N, increase

$$\begin{aligned} a_N &\rightarrow m \\ b_N &\rightarrow N - m \\ \mathbb{E}[\mu] &= \frac{a_N}{a_N + b_N} \rightarrow \frac{m}{N} = \mu_{\mathrm{ML}} \\ \mathrm{var}[\mu] &= \frac{a_N b_N}{(a_N + b_N)^2 (a_N + b_N + 1)} \rightarrow 0 \end{aligned}$$

Prediction under the Posterior

What is the probability that the next coin toss will land heads up?

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

An Example

	Prior	Data	Posterior
Total #	100	3	103
Head #	50	3	53
Tail #	50		50

The probability that the next coin toss will land heads up is 53/103。

Outlines

- Binary Distributions
- Multinomial Distributions
- Gaussian Distributions
- Exponential Families
- Non-informative Priors
- Non-parametric Methods
- KNN

Multinomial Variables

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

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

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

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

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

ML Parameter estimation

Given: $\mathcal{D} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$

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

Ensure $\sum_k \mu_k = 1$, use a Lagrange multiplier, λ.

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

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

The Multinomial Distribution

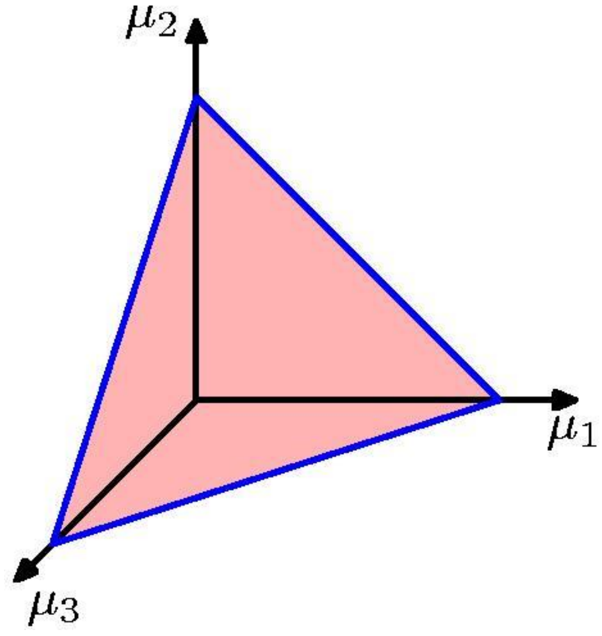
$$\begin{aligned}
\mathrm{Mult}(m_1, m_2, \ldots, m_K|\boldsymbol{\mu}, N) &= \binom{N}{m_1 m_2 \ldots m_K} \prod_{k=1}^{K} \mu_k^{m_k} \\
\mathbb{E}[m_k] &= N\mu_k \\
\mathrm{var}[m_k] &= N\mu_k(1-\mu_k) \\
\mathrm{cov}[m_j m_k] &= -N\mu_j\mu_k
\end{aligned}$$

The Dirichlet Distribution

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

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

Conjugate prior for the multinomial distribution.

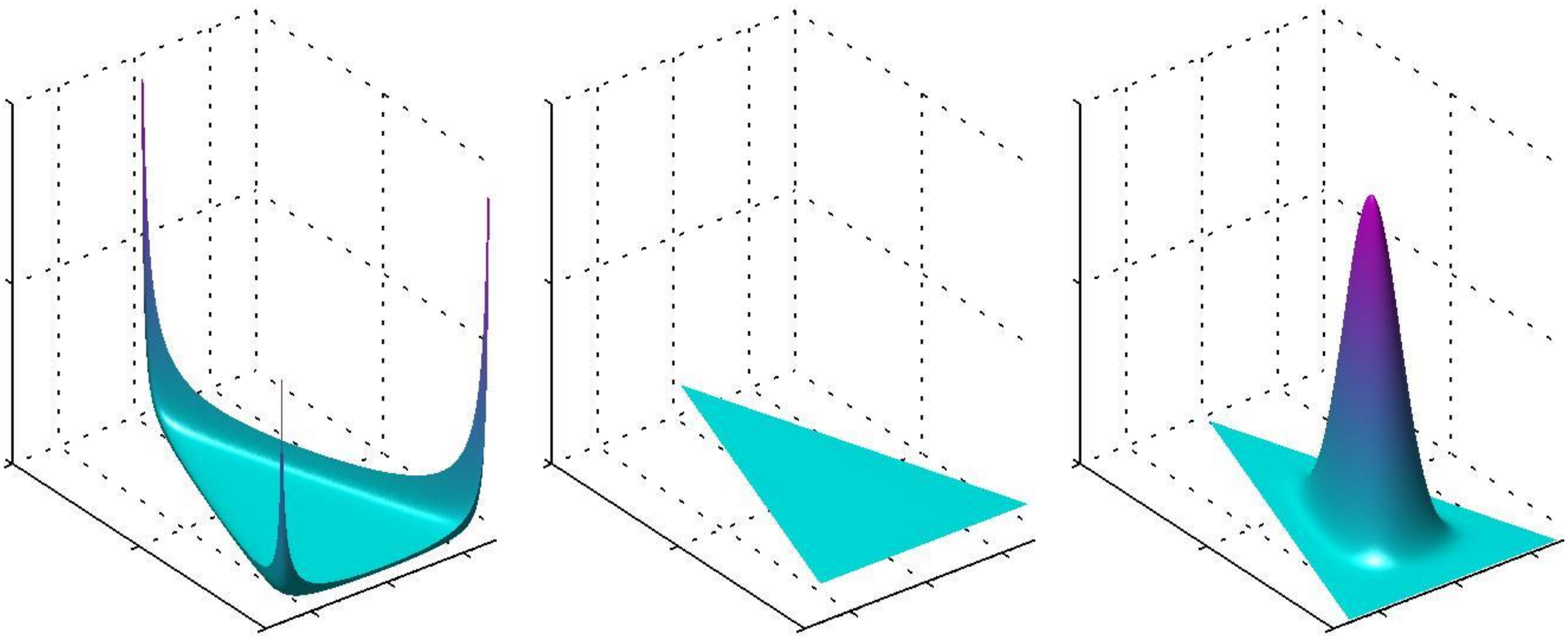


Bayesian Multinomial (1)

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

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

Bayesian Multinomial (2)

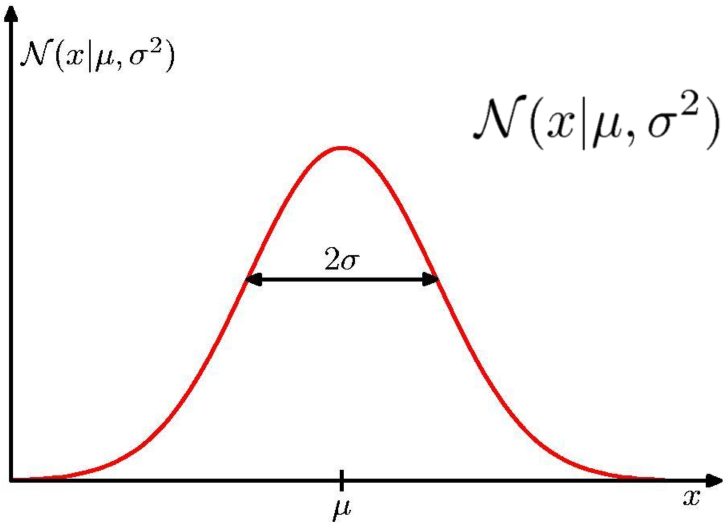


$\alpha_k = 10^{-1}$ $\alpha_k = 10^{0}$ $\alpha_k = 10^{1}$

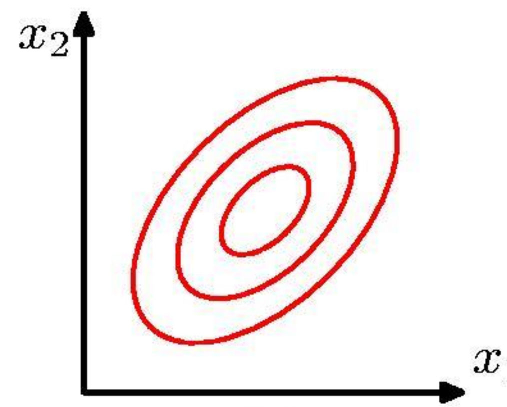
Outlines

- Binary Distributions
- Multinomial Distributions
- Gaussian Distributions
- Exponential Families
- Non-informative Priors
- Non-parametric Methods
- KNN

The Gaussian Distribution



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

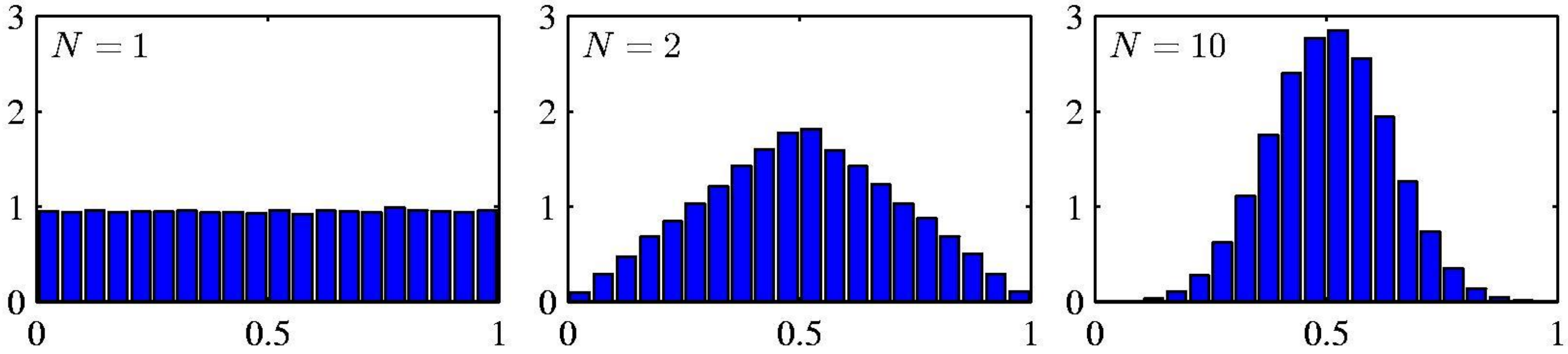


$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$

Central Limit Theorem

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

Example: N uniform $[0,1]$ random variables.



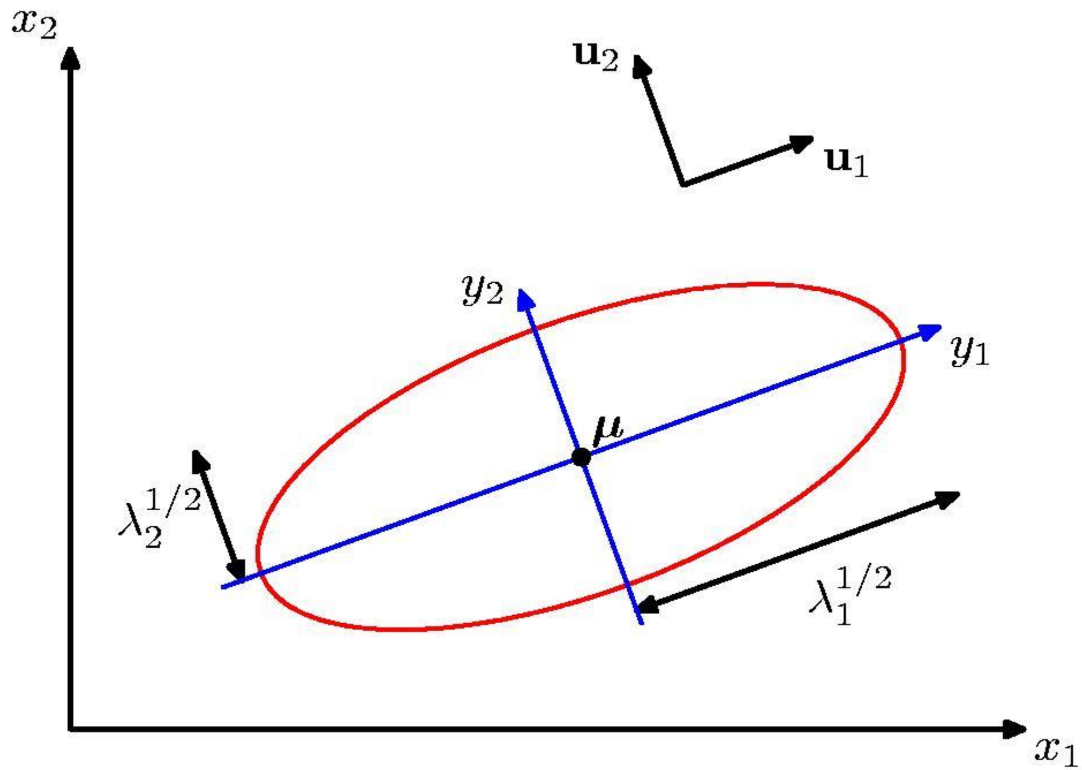
Geometry of the Multivariate Gaussian

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

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

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

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



Moments of the Multivariate Gaussian (1)

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

thanks to anti-symmetry of $\mathbf{z}$

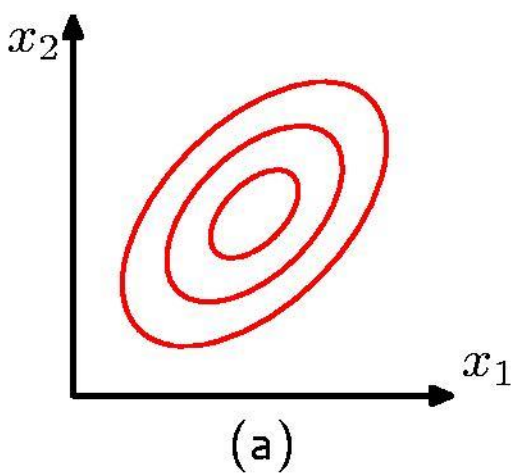
$$\mathbb{E}[\mathbf{x}] = \boldsymbol{\mu}$$

Moments of the Multivariate Gaussian (2)

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

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

$$\operatorname{cov}[A\mathbf{x}] = A\boldsymbol{\Sigma}A^{T}$$



Properties of Gaussians

$$\left.\begin{array}{l} X \sim N(\mu, \sigma^2) \\ Y = aX + b \end{array}\right\} \Rightarrow \quad Y \sim N(a\mu + b, a^2\sigma^2)$$

$$\left.\begin{array}{l} X_1 \sim N(\mu_1, \sigma_1{}^2) \\ X_2 \sim N(\mu_2, \sigma_2{}^2) \end{array}\right\} \Rightarrow p(X_1) \cdot p(X_2) \sim N\left(\frac{\sigma_2{}^2}{\sigma_1{}^2 + \sigma_2{}^2}\mu_1 + \frac{\sigma_1{}^2}{\sigma_1{}^2 + \sigma_2{}^2}\mu_2, \quad \frac{1}{\sigma_1{}^{-2} + \sigma_2{}^{-2}} \right)$$

$p(X) \sim N(\mu, \sigma^2)$

Precision

$$\begin{cases} \sigma^{-2} = \sigma_1{}^{-2} + \sigma_2{}^{-2} \\ \sigma^{-2}\mu = \sigma_1{}^{-2}\mu_1 + \sigma_2{}^{-2}\mu_2 \end{cases}$$

Properties of Gaussians

$$p_{X1}(x)p_{X2}(x) \propto e^{-\frac{(x-\mu_1)^2}{2\sigma_1{}^2}} e^{-\frac{(x-\mu_2)^2}{2\sigma_2{}^2}}$$

$$\Updownarrow$$

$$p_X(x) \propto e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Quadratic terms of x (x^2) are equal

1st order terms of *x* are also equal

$$\begin{cases} \sigma^{-2} = \sigma_1{}^{-2} + \sigma_2{}^{-2} \\ \sigma^{-2}\mu = \sigma_1{}^{-2}\mu_1 + \sigma_2{}^{-2}\mu_2 \end{cases}$$

Multivariate Gaussians

$$\left.\begin{array}{l} X \sim N(\mu,\Sigma) \\ Y = AX + B \end{array}\right\} \Rightarrow \quad Y \sim N(A\mu + B, A\Sigma A^T)$$

$$\left.\begin{array}{l} X_1 \sim N(\mu_1,\Sigma_1) \\ X_2 \sim N(\mu_2,\Sigma_2) \end{array}\right\} \Rightarrow p(X_1)\cdot p(X_2) \sim N\left(\frac{\Sigma_2}{\Sigma_1+\Sigma_2}\mu_1 + \frac{\Sigma_1}{\Sigma_1+\Sigma_2}\mu_2, \quad \frac{1}{\Sigma_1^{-1}+\Sigma_2^{-1}}\right)$$

(where division "–" denotes matrix inversion)

- We **stay Gaussian** as long as we start with Gaussians and perform only **linear transformations**

Multivariate Gaussians

$$p_X(x) \sim N(\mu,\ \Sigma)$$

Precision

$$\left[\begin{array}{lcccc} \Sigma^{-1} & = & {\Sigma_1}^{-1} & + & {\Sigma_2}^{-1} \\ \Sigma^{-1}\mu & = & {\Sigma_1}^{-1}\mu_1 & + & {\Sigma_2}^{-1}\mu_2 \end{array}\right.$$

Mean

Bayes' Theorem for Gaussian Variables

Given $$y = Ax + v$$

$$p(x) = \mathcal{N}(x|\mu, \Sigma) \qquad p(v) = \mathcal{N}(v|0, Q)$$

we have

$$p(y|x) = \mathcal{N}(y|Ax, Q)$$

$$p(y) = \mathcal{N}(y|A\mu, A\Sigma A^T + Q)$$

Then what is $p(x|y)$?

Bayes' Theorem for Gaussian Variables

Given $$x = Hy + u$$

$$p(x|y) = \mathcal{N}(x|Hy, L) \qquad p(u) = \mathcal{N}(u|0, L)$$

we have

$$p(x|y) \propto p(y|x)p(x) = \mathcal{N}(y|Ax, Q)\mathcal{N}(x|\mu, \Sigma)$$

$$-\frac{1}{2}(x - Hy)^{\boldsymbol{T}}\ L^{-1}(x - Hy) \quad \propto \quad -\frac{1}{2}(y - Ax)^{\boldsymbol{T}}\ Q^{-1}(y - Ax) -\frac{1}{2}(x - \mu)^{\boldsymbol{T}}\ \Sigma^{-1}(x - \mu)$$

Bayes' Theorem for Gaussian Variables

$$-\frac{1}{2}(x-Hy)^{\boldsymbol{T}}\ L^{-1}(x-Hy) \quad \propto \quad -\frac{1}{2}(y-Ax)^{\boldsymbol{T}}\ Q^{-1}(y-Ax)$$
$$-\frac{1}{2}(x-\mu)^{\boldsymbol{T}}\ \Sigma^{-1}(x-\mu)$$

Quadratic terms of x (x^T**x) are equal

$$\begin{cases} L^{-1} = A^T Q^{-1} A + \Sigma^{-1} \\ L^{-1} H y = A^T Q^{-1} y + \Sigma^{-1}\mu \end{cases}$$

1st order terms of x (x^T **) are also equal

Bayes' Theorem for Gaussian Variables

$$p(x|y) = \mathcal{N}(x|Hy, L)$$

where

$$\left[\begin{array}{lcl} L^{-1} & = & A^T Q^{-1} A \quad + \quad \Sigma^{-1} \\ Hy & = & L\{A^T Q^{-1} y + \Sigma^{-1} \mu\} \end{array}\right.$$

Matrix Inversion Lemma

If A, C, BCD are non-sigular square matrix (the inverse exists) then

$$[A+BCD]^{-1} = A^{-1} - A^{-1}B[C^{-1}+DA^{-1}B]^{-1}DA^{-1}$$

Matrix Inversion Lemma Proof

$$[A+BCD]\left[A^{-1}-A^{-1}B[C^{-1}+DA^{-1}B]^{-1}DA^{-1}\right]$$

$$=I+BCDA^{-1}-B[C^{-1}+DA^{-1}B]^{-1}DA^{-1}$$

$$-BCDA^{-1}B[C^{-1}+DA^{-1}B]^{-1}DA^{-1}$$

$$=I+BCDA^{-1}-B\left\{I+CDA^{-1}B\right\}[C^{-1}+DA^{-1}B]^{-1}DA^{-1}$$

$$=I+BCDA^{-1}-BC\left\{C^{-1}+DA^{-1}B\right\}[C^{-1}+DA^{-1}B]^{-1}DA^{-1}$$

$$=I$$

Bayes' Theorem for Gaussian Variables

Then

$$L = \Sigma - \Sigma A^T (A^T \Sigma A + Q)^{-1} A \Sigma$$

$$\left[\begin{array}{ccc} L & = & (I - KA)\Sigma \\ Hy & = & \mu + K(y - A\mu) \end{array}\right.$$

Kalman Gain $\longrightarrow$ $K = \Sigma A^T (A^T \Sigma A + Q)^{-1}$

$$\boxed{p(x|y) = \mathcal{N}(x|\mu + K(y - A\mu), (I - KA)\Sigma)}$$

Partitioned Gaussian Distributions

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \qquad \boldsymbol{\Lambda} \equiv \boldsymbol{\Sigma}^{-1} \qquad \boldsymbol{\Lambda} = \begin{pmatrix} \boldsymbol{\Lambda}_{aa} & \boldsymbol{\Lambda}_{ab} \\ \boldsymbol{\Lambda}_{ba} & \boldsymbol{\Lambda}_{bb} \end{pmatrix}$$

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

$$\mathbf{x}_a = \boldsymbol{A}\mathbf{x}_b + \boldsymbol{w} \quad \boldsymbol{\Sigma}_{a|b} = \boldsymbol{\Sigma}_w$$

$$\mathbf{x}_a - \mu_a = \boldsymbol{A}(\mathbf{x}_b - \mu_b) + \boldsymbol{w} \;\Rightarrow\; \mu_{a|b} - \mu_a = \boldsymbol{A}(\mathbf{x}_b - \mu_b)$$

$$\boldsymbol{\Sigma}_{ab} = A\boldsymbol{\Sigma}_{bb} \qquad \Rightarrow \qquad A = \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}$$

$$\boxed{\boldsymbol{\Sigma}_{aa} = A\boldsymbol{\Sigma}_{bb}A^T + \boldsymbol{\Sigma}_w = A\boldsymbol{\Sigma}_{bb}A^T + \boldsymbol{\Sigma}_{a|b} = \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}\boldsymbol{\Sigma}_{ba} + \boldsymbol{\Sigma}_{a|b}}$$

Partitioned Gaussian Distributions

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

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

$$\boxed{\mu_{a|b} = \mu_a + \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}(\mathbf{x}_b - \mu_b)}$$

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

Inverse Covariance Matrix*

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

$$\boxed{-\frac{1}{2}(\mathbf{x}_a-\boldsymbol{\mu}_a)^{\mathrm{T}}\boldsymbol{\Lambda}_{aa}(\mathbf{x}_a-\boldsymbol{\mu}_a)} - \frac{1}{2}(\mathbf{x}_a-\boldsymbol{\mu}_a)^{\mathrm{T}}\boldsymbol{\Lambda}_{ab}(\mathbf{x}_b-\boldsymbol{\mu}_b)$$

$$-\frac{1}{2}(\mathbf{x}_b-\boldsymbol{\mu}_b)^{\mathrm{T}}\boldsymbol{\Lambda}_{ba}(\mathbf{x}_a-\boldsymbol{\mu}_a) - \frac{1}{2}(\mathbf{x}_b-\boldsymbol{\mu}_b)^{\mathrm{T}}\boldsymbol{\Lambda}_{bb}(\mathbf{x}_b-\boldsymbol{\mu}_b).$$

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

$$-\frac{1}{2}\left(\mathbf{x}_a-\boldsymbol{\mu}_{a|b}\right)^{T}\boldsymbol{\Sigma}_{\boldsymbol{a|b}}^{-1}\left(\mathbf{x}_a-\boldsymbol{\mu}_{a|b}\right) = \boxed{-\frac{1}{2}\mathbf{x}_a{}^{T}\,\boldsymbol{\Sigma}_{\boldsymbol{a|b}}^{-1}\mathbf{x}_a} + \mathbf{x}_a{}^{T}\,\boldsymbol{\Sigma}_{\boldsymbol{a|b}}^{-1}\boldsymbol{\mu}_{a|b} + \text{const.}$$

$$\Rightarrow \quad \underbrace{\boldsymbol{\Sigma}_{a|b} = \boldsymbol{\Lambda}_{aa}^{-1}}_{-\frac{1}{2}\mathbf{x}_a{}^{T} * \mathbf{x}_a} \qquad \underbrace{\boldsymbol{\Lambda}_{aa}\boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{ab}(\mathbf{x}_b-\boldsymbol{\mu}_b) = \boldsymbol{\Sigma}_{a|b}^{-1}\boldsymbol{\mu}_{a|b}}_{\mathbf{x}_a{}^{T} *}$$

Inverse Matrix Lemma*

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

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

Inverse Covariance Matrix*

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

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

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

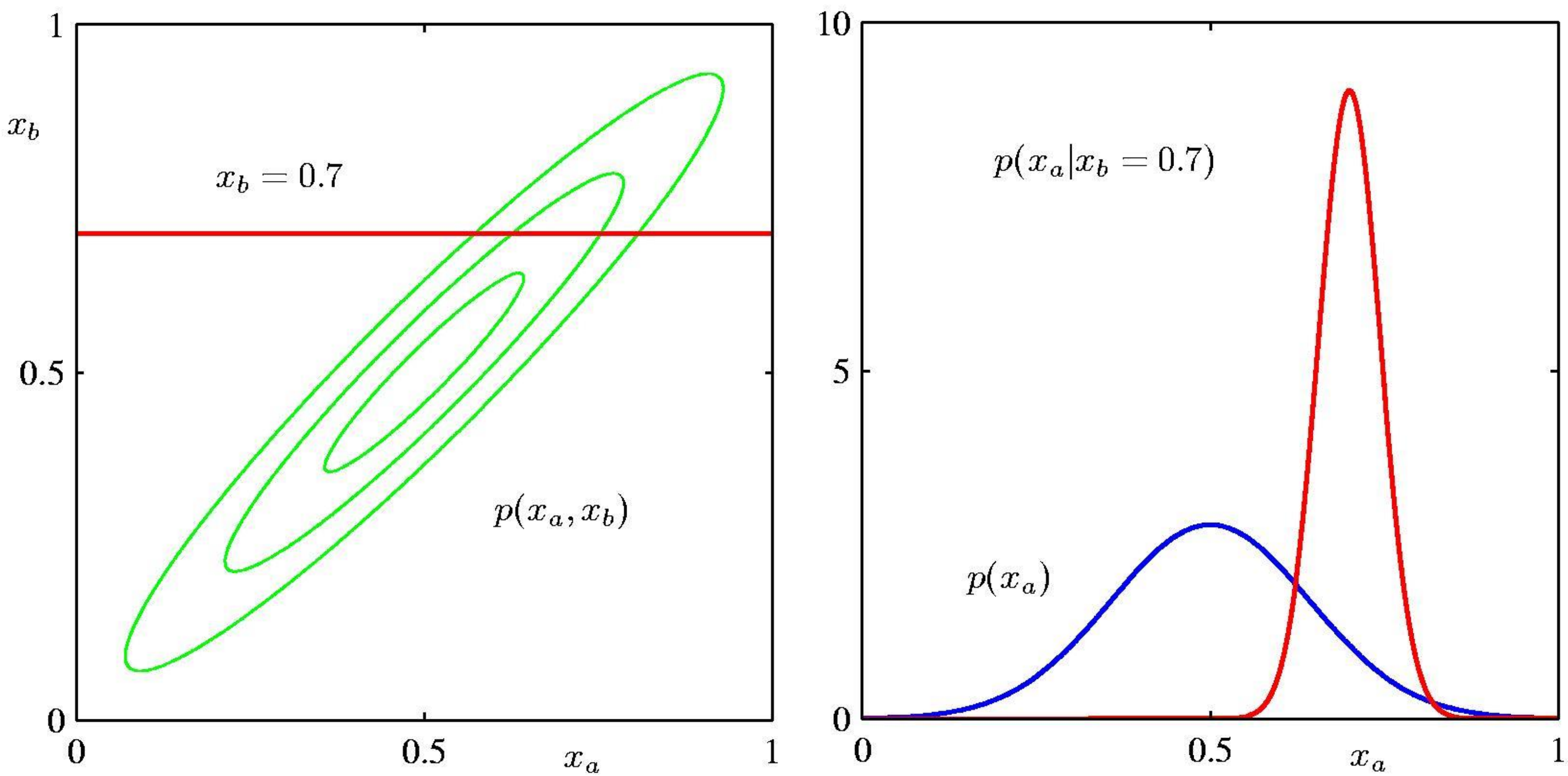
Partitioned Conditionals and Marginals*

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

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

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

Partitioned Conditionals and Marginals



Bayes' Theorem for Gaussian Variables*

Given

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

we have

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

where

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

Maximum Likelihood for the Gaussian (1)

Given i.i.d. data $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_N)^{\mathrm{T}}$, the log likelihood function is given by

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

Sufficient statistics

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

Maximum Likelihood for the Gaussian (2)

Set the derivative of the log likelihood function to zero,

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

and solve to obtain

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

Similarly

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

Maximum Likelihood for the Gaussian (3)

Under the true distribution

$$
\begin{aligned}
\mathbb{E}[\boldsymbol{\mu}_{\mathrm{ML}}] &= \boldsymbol{\mu} \\
\mathbb{E}[\boldsymbol{\Sigma}_{\mathrm{ML}}] &= \frac{N-1}{N}\boldsymbol{\Sigma}.
\end{aligned}
$$

Hence define

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

Sequential Estimation

Contribution of the N^{th} data point, $\mathbf{x}_N$

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

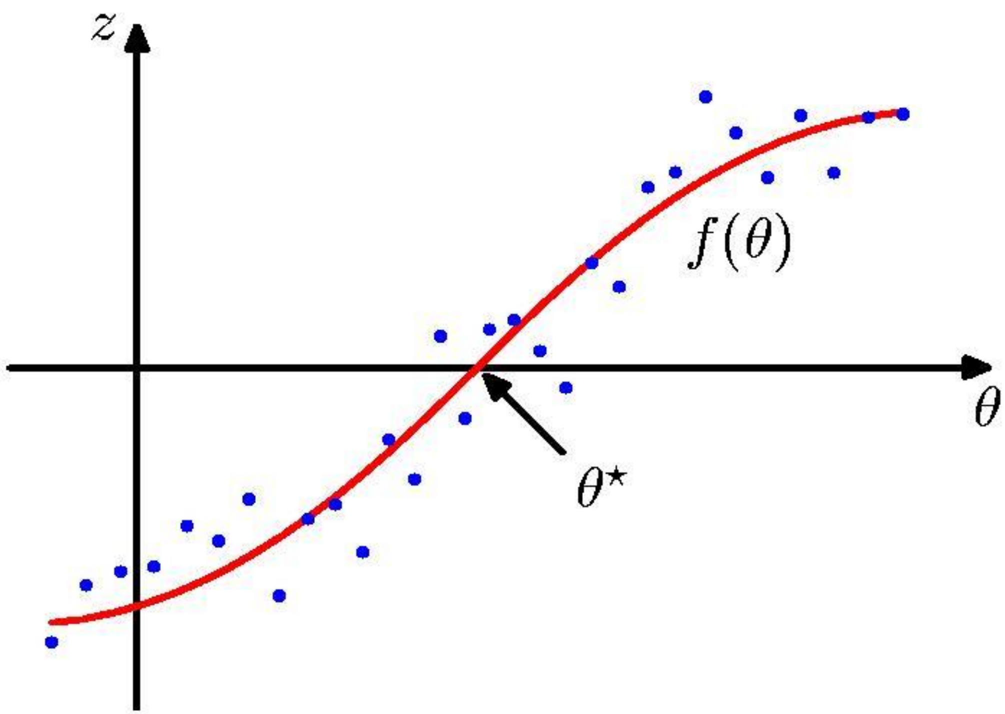
The Robbins-Monro Algorithm (1)*

Consider θ and z governed by $p(z,\theta)$ and define the *regression function*

$$f(\theta) \equiv \mathbb{E}[z|\theta] = \int z p(z|\theta)\,\mathrm{d}z$$

Seek $\theta^{\star}$ such that $f(\theta^{\star}) = 0$.

The Robbins-Monro Algorithm (2)*



Assume we are given samples from $p(z,\theta)$, one at the time.

The Robbins-Monro Algorithm (3)*

Successive estimates of $\theta^\star$ are then given by

$$\theta^{(N)} = \theta^{(N-1)} - a_{N-1} z(\theta^{(N-1)}).$$

Conditions on a_N for convergence :

$$\lim_{N \to \infty} a_N = 0 \qquad \sum_{N=1}^{\infty} a_N = \infty \qquad \sum_{N=1}^{\infty} a_N^2 < \infty$$

Robbins-Monro for Maximum Likelihood (1)*

Regarding

$$-\lim_{N\to\infty}\frac{1}{N}\sum_{n=1}^{N}\frac{\partial}{\partial\theta}\ln p(x_n|\theta)=\mathbb{E}_x\left[-\frac{\partial}{\partial\theta}\ln p(x|\theta)\right]$$

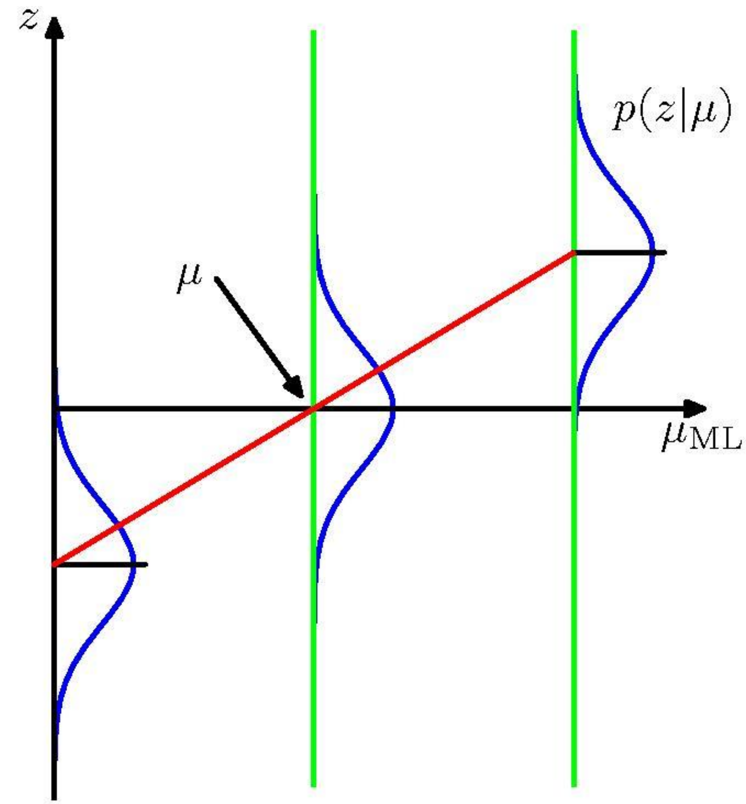
as a regression function, finding its root is equivalent to finding the maximum likelihood solution θ_{ML}. Thus

$$\theta^{(N)}=\theta^{(N-1)}-a_{N-1}\frac{\partial}{\partial\theta^{(N-1)}}\left[-\ln p(x_N|\theta^{(N-1)})\right].$$

Robbins-Monro for Maximum Likelihood (2)*

Example: estimate the mean of a Gaussian.

$$\begin{aligned} z &= \frac{\partial}{\partial \mu_{\mathrm{ML}}}\left[-\ln p(x|\mu_{\mathrm{ML}}, \sigma^2)\right] \\ &= -\frac{1}{\sigma^2}(x - \mu_{\mathrm{ML}}) \end{aligned}$$



The distribution of z is Gaussian with mean $\mu - \mu_{\mathrm{ML}}$.

For the Robbins-Monro update equation, $a_N = \sigma^2/N$.

⇒ SEQUENTIAL estimation

Bayesian Inference for the Gaussian (1)

Assume σ^2 is known. Given i.i.d. data $\mathbf{x} = \{x_1, \ldots, x_N\}$, the likelihood function for μ is given by

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

This has a Gaussian shape as a function of μ (but it is *not* a distribution over μ).

Bayesian Inference for the Gaussian (2)

Combined with a Gaussian prior over μ,

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

this gives the posterior

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

Completing the square over μ, we see that

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

Bayesian Inference for the Gaussian (3)

… where

$$\mu_N = \frac{\sigma^2}{N\sigma_0^2 + \sigma^2}\mu_0 + \frac{N\sigma_0^2}{N\sigma_0^2 + \sigma^2}\mu_{\mathrm{ML}}, \qquad \mu_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N} x_n$$

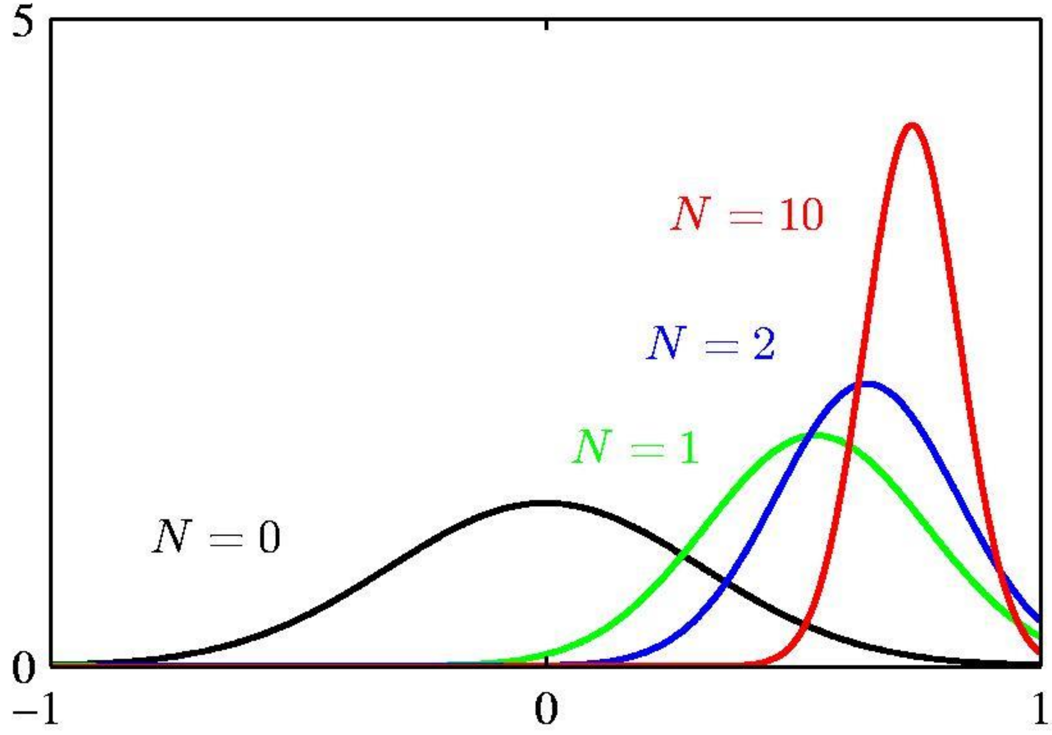
$$\frac{1}{\sigma_N^2} = \frac{1}{\sigma_0^2} + \frac{N}{\sigma^2}.$$

Note:

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

Bayesian Inference for the Gaussian (4)

Example: $p(\mu|\mathbf{x}) = \mathcal{N}\left(\mu|\mu_N, \sigma_N^2\right)$ for $N = 0,\ 1,\ 2$ and $10.$



Bayesian Inference for the Gaussian (5)

Sequential Estimation

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

The posterior obtained after observing $N-1$ data points becomes the prior when we observe the N^{th} data point.

Bayesian Inference for the Gaussian (6)

Now assume μ is known. The likelihood function for $\lambda = 1/\sigma^2$ is given by

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

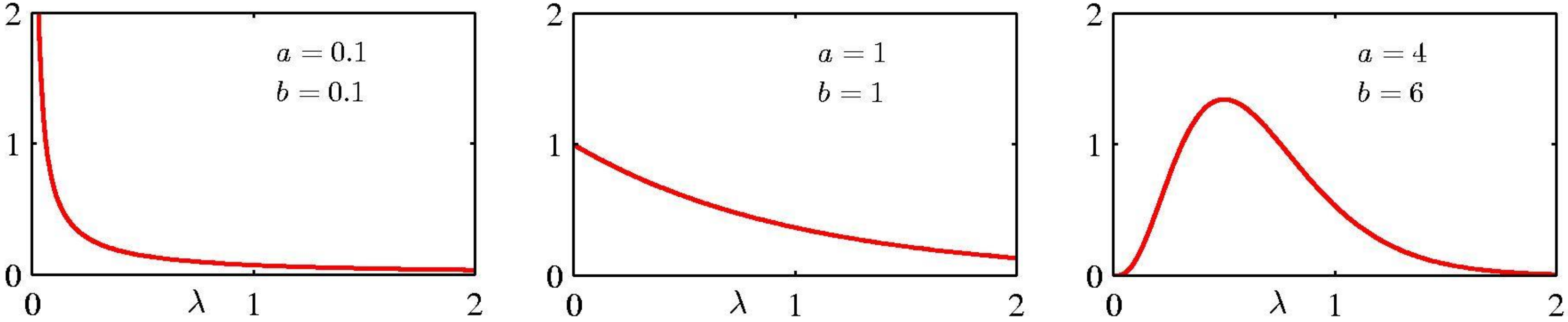
This has a Gamma shape as a function of λ.

Bayesian Inference for the Gaussian (7)

The Gamma distribution

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

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



Bayesian Inference for the Gaussian (8)

Now we combine a Gamma prior, $\mathrm{Gam}(\lambda|a_0, b_0)$, with the likelihood function for λ to obtain

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

which we recognize as $\mathrm{Gam}(\lambda|a_N, b_N)$ with

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

Bayesian Inference for the Gaussian (9)

If both μ and λ are unknown, the joint likelihood function is given by

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

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

We need a prior with the same functional dependence on μ and λ.

Bayesian Inference for the Gaussian (10)

The Gaussian-gamma distribution prior

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

Then the posterior is given by

$$\beta_N = \beta + N \qquad c_N = c + \sum_{n=1}^{N} x_N \qquad d_N = d + \frac{1}{2}\sum_{n=1}^{N} x_N^2$$

Bayesian Inference for the Gaussian (11)

The Gaussian-gamma distribution

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

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

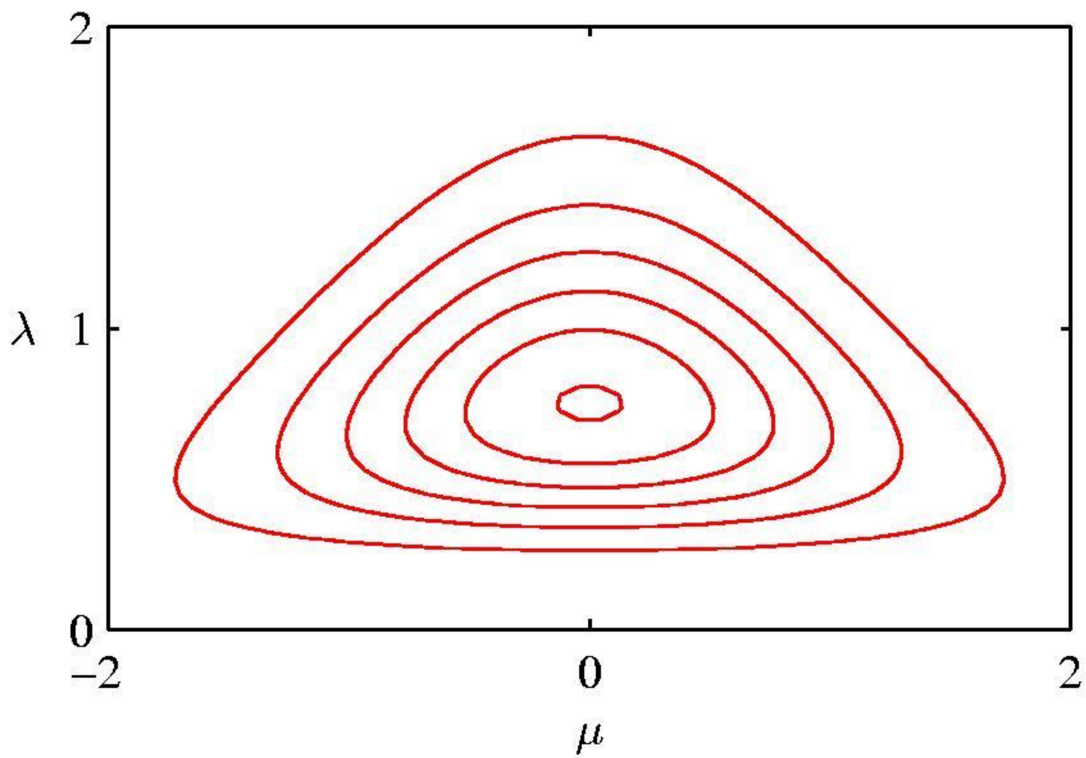
- Quadratic in μ.
- Linear in λ.

- Gamma distribution over λ.
- Independent of μ.

$$\begin{aligned} \mu_0 &= c/\beta \\ a &= 1 + \beta/2 \\ b &= d - c^2/2\beta \end{aligned}$$

Bayesian Inference for the Gaussian (12)

The Gaussian-gamma distribution



Bayesian Inference for the Gaussian (13)*

Multivariate conjugate priors

- $\boldsymbol{\mu}$ unknown, $\mathbf{\Lambda}$ known: $p(\boldsymbol{\mu})$ Gaussian.
- $\mathbf{\Lambda}$ unknown, $\boldsymbol{\mu}$ known: $p(\mathbf{\Lambda})$ Wishart,

$$\mathcal{W}(\mathbf{\Lambda}|\mathbf{W},\nu) = B|\mathbf{\Lambda}|^{(\nu-D-1)/2}\exp\left(-\frac{1}{2}\mathrm{Tr}(\mathbf{W}^{-1}\mathbf{\Lambda})\right).$$

- $\mathbf{\Lambda}$ and $\boldsymbol{\mu}$ unknown: $p(\boldsymbol{\mu},\mathbf{\Lambda})$ Gaussian-Wishart, $p(\boldsymbol{\mu},\mathbf{\Lambda}|\boldsymbol{\mu}_0,\beta,\mathbf{W},\nu) =$

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

Student’s t-Distribution*

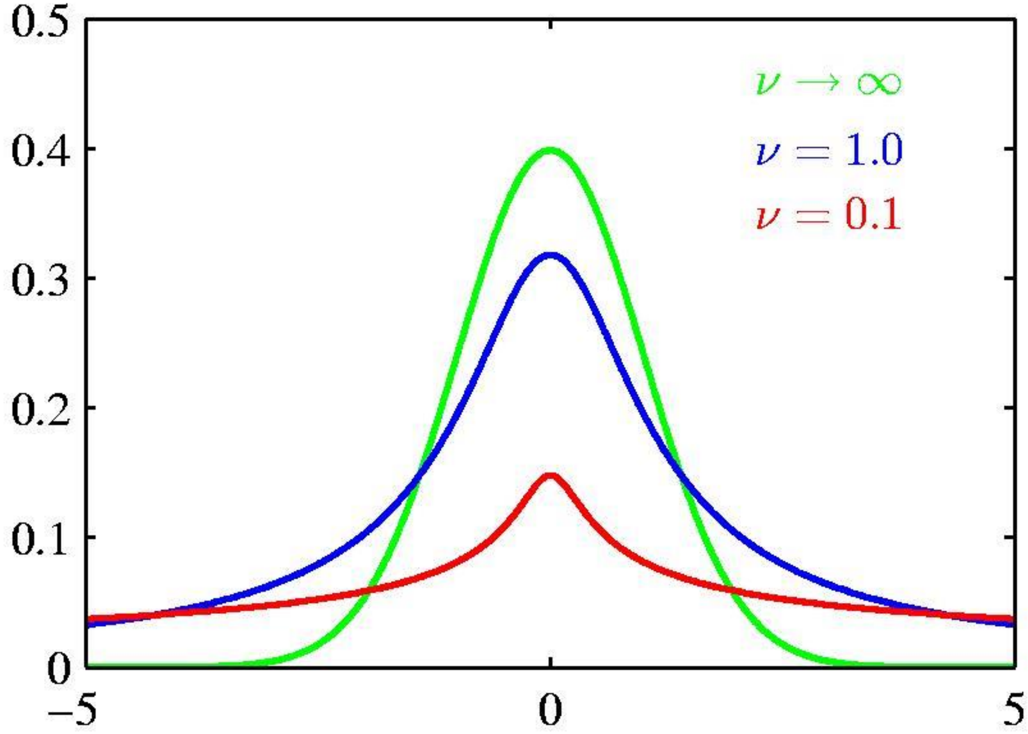
$$
\begin{aligned}
p(x|\mu,a,b) &= \int_0^\infty \mathcal{N}(x|\mu,\tau^{-1})\mathrm{Gam}(\tau|a,b)\,\mathrm{d}\tau \\
&= \int_0^\infty \mathcal{N}\left(x|\mu,(\eta\lambda)^{-1}\right)\mathrm{Gam}(\eta|\nu/2,\nu/2)\,\mathrm{d}\eta \\
&= \frac{\Gamma(\nu/2+1/2)}{\Gamma(\nu/2)}\left(\frac{\lambda}{\pi\nu}\right)^{1/2}\left[1+\frac{\lambda(x-\mu)^2}{\nu}\right]^{-\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

where

$$
\lambda = a/b \qquad \eta = \tau b/a \qquad \nu = 2a.
$$

Infinite mixture of Gaussians.

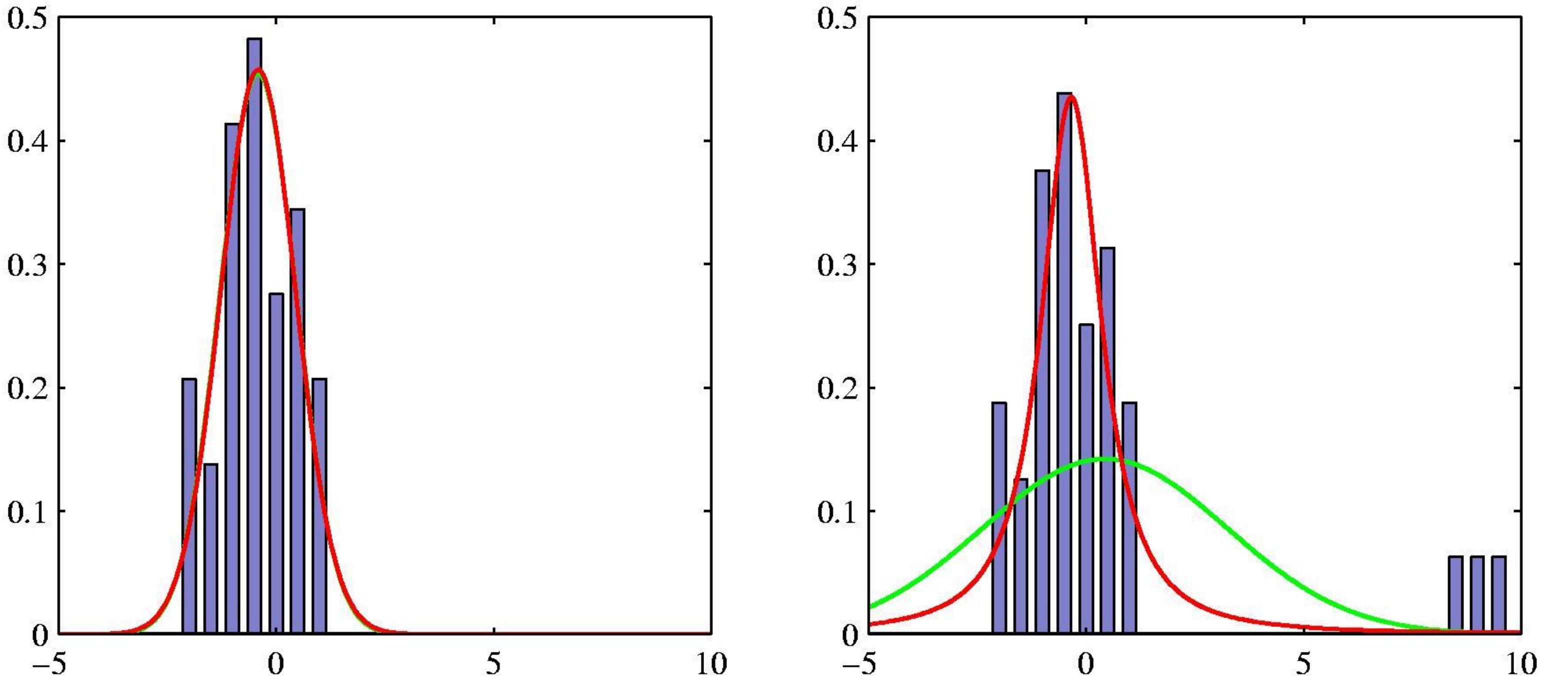
Student's t-Distribution*



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

Student’s t-Distribution*

Robustness to outliers: Gaussian vs t-distribution.



Student's t-Distribution*

The D-variate case:

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

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

Properties:

$$\begin{aligned}\mathbb{E}[\mathbf{x}] &= \boldsymbol{\mu}, && \text{if } \nu > 1 \\ \mathrm{cov}[\mathbf{x}] &= \frac{\nu}{(\nu-2)}\boldsymbol{\Lambda}^{-1}, && \text{if } \nu > 2 \\ \mathrm{mode}[\mathbf{x}] &= \boldsymbol{\mu}\end{aligned}$$

Periodic variables*

- Examples: calendar time, direction, …
- We require

$$\begin{aligned} p(\theta) &\geqslant 0 \\ \int_0^{2\pi} p(\theta)\,\mathrm{d}\theta &= 1 \\ p(\theta + 2\pi) &= p(\theta). \end{aligned}$$

von Mises Distribution (1)*

This requirement is satisfied by

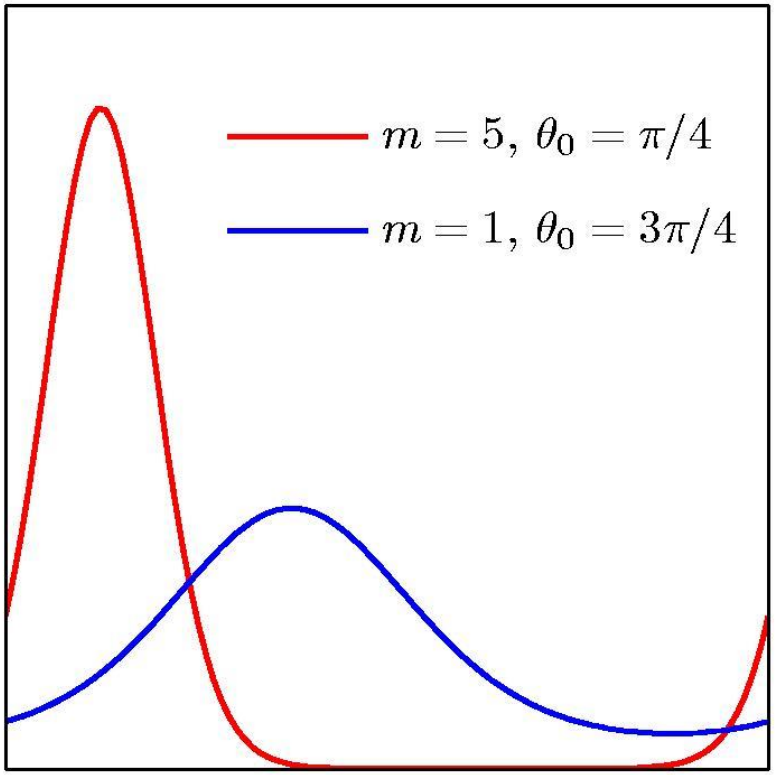
$$p(\theta|\theta_0, m) = \frac{1}{2\pi I_0(m)} \exp\left\{m\cos(\theta - \theta_0)\right\}$$

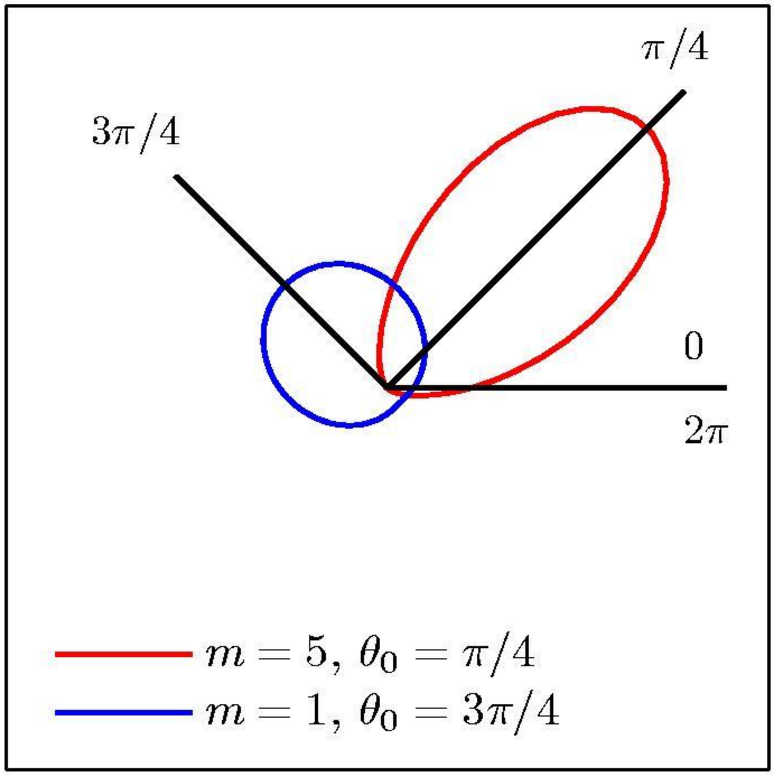
where

$$I_0(m) = \frac{1}{2\pi}\int_0^{2\pi} \exp\left\{m\cos\theta\right\}\,\mathrm{d}\theta$$

is the 0^{th} order modified Bessel function of the 1^{st} kind.

von Mises Distribution (2)*





Maximum Likelihood for von Mises*

Given a data set, $\mathcal{D} = \{\theta_1, \ldots, \theta_N\}$, the log likelihood function is given by

$$\ln p(\mathcal{D}|\theta_0, m) = -N \ln(2\pi) - N \ln I_0(m) + m \sum_{n=1}^{N} \cos(\theta_n - \theta_0).$$

Maximizing with respect to θ_0 we directly obtain

$$\theta_0^{\mathrm{ML}} = \tan^{-1}\left\{\frac{\sum_n \sin\theta_n}{\sum_n \cos\theta_n}\right\}.$$

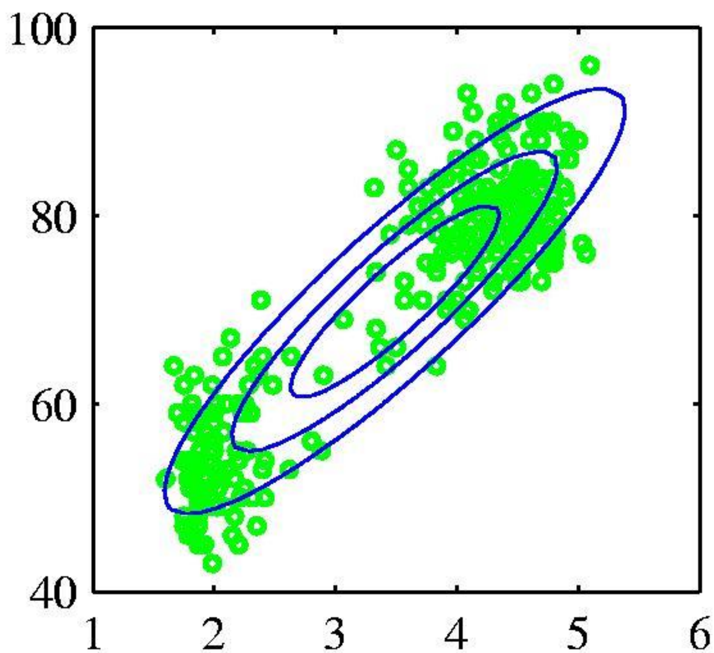
Similarly, maximizing with respect to m we get

$$\frac{I_1(m_{\mathrm{ML}})}{I_0(m_{\mathrm{ML}})} = \frac{1}{N}\sum_{n=1}^{N}\cos(\theta_n - \theta_0^{\mathrm{ML}})$$

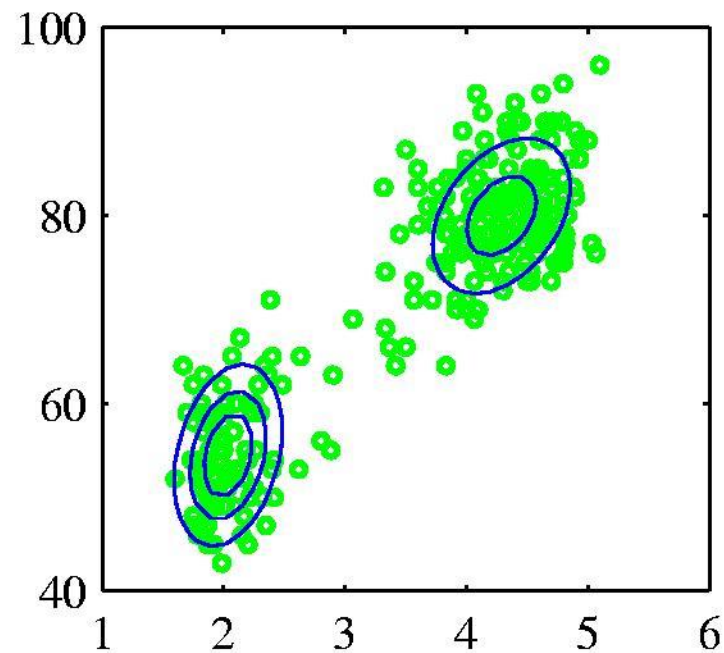
which can be solved numerically for m_{ML}.

Mixtures of Gaussians (1)

Old Faithful data set



Single Gaussian

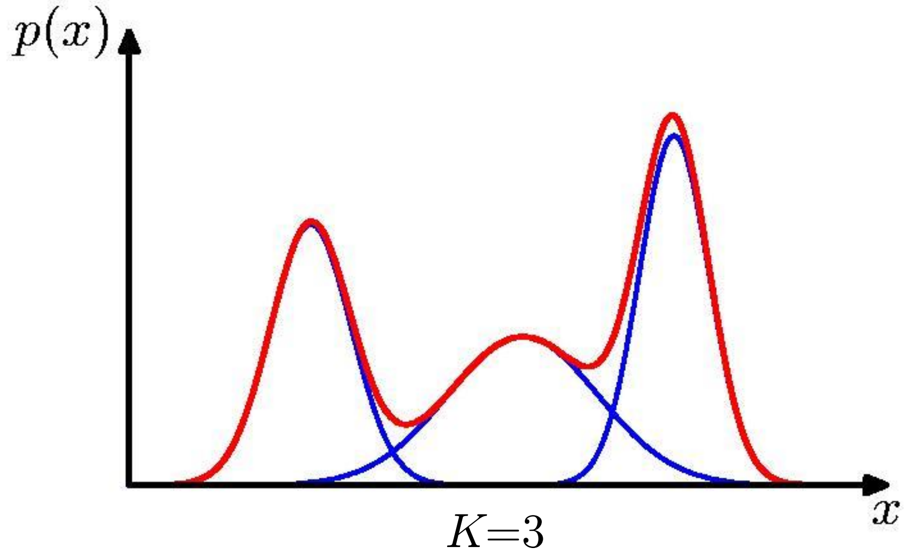


Mixture of two Gaussians

Mixtures of Gaussians (2)

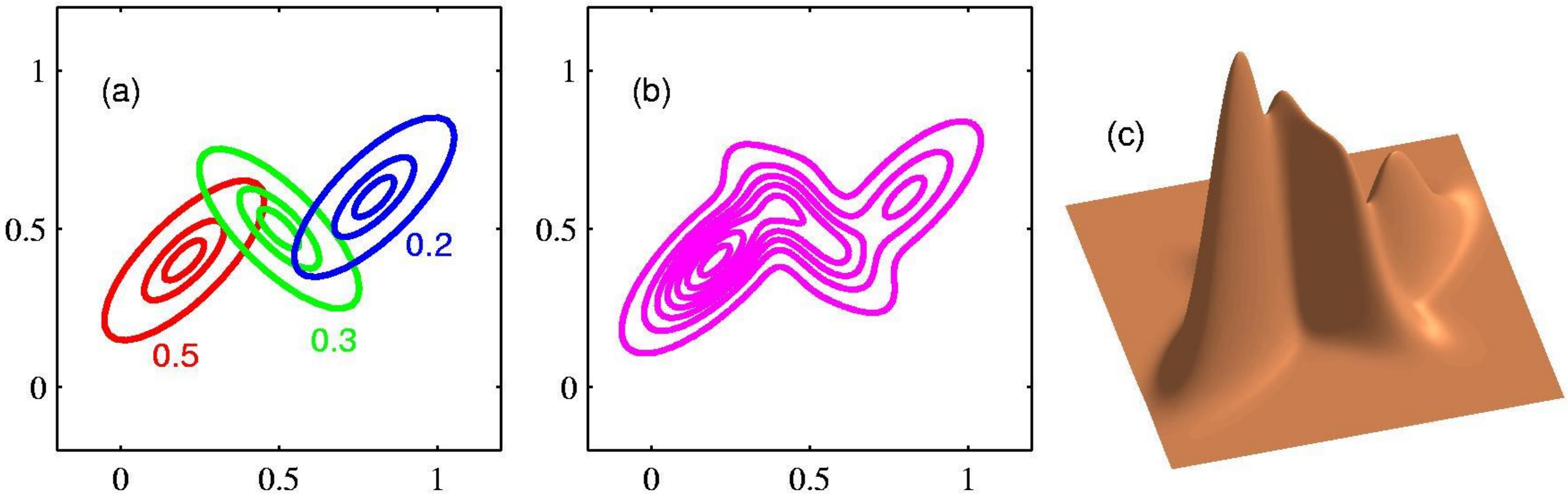
Combine simple models into a complex model:

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



$$\forall k : \pi_k \geqslant 0 \qquad \sum_{k=1}^{K} \pi_k = 1$$

Mixtures of Gaussians (3)



Mixtures of Gaussians (4)

Determining parameters $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$, and $\boldsymbol{\pi}$ using maximum log likelihood

$$\ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \underbrace{\ln \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right\}}_{\text{Log of a sum; no closed form maximum.}}$$

Solution: use standard, iterative, numeric optimization methods or the *expectation maximization* algorithm (Chapter 9).

Mixtures of Gaussians (5)

The posterior probability of each data point being responsible for each cluster

$$\begin{aligned}\gamma_k(\mathbf{x}) &\equiv p(k|\mathbf{x}) \\ &= \frac{p(k)p(\mathbf{x}|k)}{\sum_l p(l)p(\mathbf{x}|l)} \\ &= \frac{\pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_l \pi_l \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)}\end{aligned}$$

Outlines

- Binary Distributions
- Multinomial Distributions
- Gaussian Distributions
- Exponential Families
- Non-informative Priors
- Non-parametric Methods
- KNN

The Exponential Family (1)

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

where $\boldsymbol{\eta}$ is the *natural parameter* and

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

so $\mathrm{g}(\boldsymbol{\eta})$ can be interpreted as a normalization coefficient.

$\mathbf{u}(\mathbf{x})$: statistics of $\mathbf{x}$

The Exponential Family (2.1)

The Bernoulli Distribution

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

Comparing with the general form we see that

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

The Exponential Family (2.2)

The Bernoulli distribution can hence be written as

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

where

$$\begin{aligned} u(x) &= x \\ h(x) &= 1 \\ g(\eta) &= 1 - \sigma(\eta) = \sigma(-\eta). \end{aligned}$$

The Exponential Family (3.1)

The Multinomial Distribution

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

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

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

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

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

The Exponential Family (3.2)

Let $\mu_M = 1 - \sum_{k=1}^{M-1} \mu_k$. This leads to

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

Here the η_k parameters are independent. Note that

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

The Exponential Family (3.3)

The Multinomial distribution can then be written as

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

where

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

The Exponential Family (4)

The Gaussian Distribution

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

where

$$
\boldsymbol{\eta} = \begin{pmatrix} \mu/\sigma^2 \\ -1/2\sigma^2 \end{pmatrix} \qquad h(\mathbf{x}) = (2\pi)^{-1/2}
$$

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

ML for the Exponential Family (1)*

From the definition of $g(\boldsymbol{\eta})$ we get

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

Thus

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

ML for the Exponential Family (2)*

Give a data set, $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$, the likelihood function is given by

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

Thus we have

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

Conjugate priors

For any member of the exponential family, there exists a prior

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

Combining with the likelihood function, we get

$$p(\boldsymbol{\eta}|\mathbf{X}, \boldsymbol{\chi}, \nu) \propto g(\boldsymbol{\eta})^{\nu+N} \exp\left\{\boldsymbol{\eta}^{\mathrm{T}} \left(\sum_{n=1}^{N} \mathbf{u}(\mathbf{x}_n) + \nu \boldsymbol{\chi}\right)\right\}.$$

Prior corresponds to ν pseudo-observations with value $\boldsymbol{\chi}$.

Outlines

- Binary Distributions
- Multinomial Distributions
- Gaussian Distributions
- Exponential Families
- Non-informative Priors
- Non-parametric Methods
- KNN

Non-informative Priors (1)*

With little or no information available a-priori, we might choose a non-informative prior.

- λ discrete, K-nomial : $p(\lambda) = 1/K$.
- $\lambda \in [a,b]$ real and bounded: $p(\lambda) = 1/b - a$.
- λ real and unbounded: improper!

A constant prior may no longer be constant after a change of variable; consider $p(\lambda)$ constant and $\lambda = \eta^2$:

$$p_\eta(\eta) = p_\lambda(\lambda) \left| \frac{\mathrm{d}\lambda}{\mathrm{d}\eta} \right| = p_\lambda(\eta^2) 2\eta \propto \eta$$

Non-informative Priors (2)*

Translation invariant priors. Consider

$$p(x|\mu) = f(x-\mu) = f((x+c)-(\mu+c)) = f(\widehat{x}-\widehat{\mu}) = p(\widehat{x}|\widehat{\mu}).$$

For a corresponding prior over μ, we have

$$\int_A^B p(\mu)\,\mathrm{d}\mu = \int_{A-c}^{B-c} p(\mu)\,\mathrm{d}\mu = \int_A^B p(\mu - c)\,\mathrm{d}\mu$$

for any A and B. Thus $p(\mu) = p(\mu - c)$ and $p(\mu)$ must be constant.

Non-informative Priors (3)*

Example: The mean of a Gaussian, μ; the conjugate prior is also a Gaussian,

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

As $\sigma_0^2 \rightarrow \infty$, this will become constant over μ.

Non-informative Priors (4)*

Scale invariant priors. Consider $p(x|\sigma) = (1/\sigma)f(x/\sigma)$ and make the change of variable $\widehat{x} = cx$

$$p_{\widehat{x}}(\widehat{x}) = p_x(x)\left|\frac{\mathrm{d}x}{\mathrm{d}\widehat{x}}\right| = p_x\left(\frac{\widehat{x}}{c}\right)\frac{1}{c} = \frac{1}{c\sigma}f\left(\frac{\widehat{x}}{c\sigma}\right) = p_x(\widehat{x}|\widehat{\sigma}).$$

For a corresponding prior over σ, we have

$$\int_A^B p(\sigma)\,\mathrm{d}\sigma = \int_{A/c}^{B/c} p(\sigma)\,\mathrm{d}\sigma = \int_A^B p\left(\frac{1}{c}\sigma\right)\frac{1}{c}\,\mathrm{d}\sigma$$

for any A and B. Thus $p(\sigma) \;/\; 1/\sigma$ and so this prior is improper too. Note that this corresponds to $p(\ln\sigma)$ being constant.

Non-informative Priors (5)*

Example: For the variance of a Gaussian, σ^2, we have

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

If $\lambda = 1/\sigma^2$ and $p(\sigma) \;/\; 1/\sigma$, then $p(\lambda) \;/\; 1/\lambda$.

- We know that the conjugate distribution for λ is the Gamma distribution,

$$\mathrm{Gam}(\lambda|a_0, b_0) \propto \lambda^{a_0 - 1} \exp(-b_0 \lambda).$$

- A non-informative prior is obtained when $a_0 = 0$ and $b_0 = 0$.

Outlines

- Binary Distributions
- Multinomial Distributions
- Gaussian Distributions
- Exponential Families
- Non-information Priors
- Non-parametric Methods
- KNN

Non-parametric Methods (1)

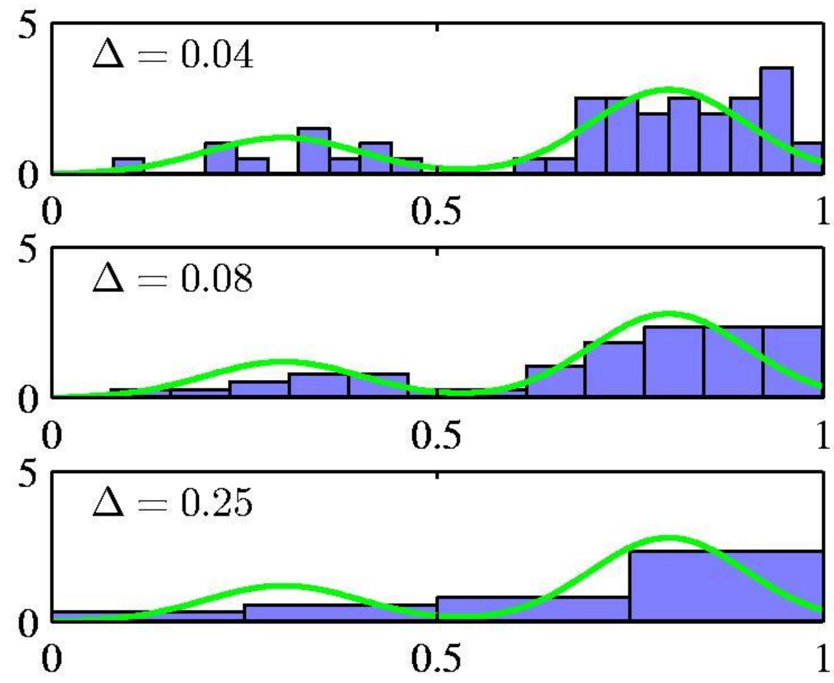
- Parametric distribution models are restricted to specific forms, which may not always be suitable; for example, consider modelling a multimodal distribution with a single, unimodal model.
- Non-parametric approaches make few assumptions about the overall shape of the distribution being modelled.

Non-parametric Methods (2)

Histogram methods partition the data space into distinct bins with widths Δ_i and count the number of observations, n_i, in each bin.

$$p_i = \frac{n_i}{N\Delta_i}$$

- Often, the same width is used for all bins, $\Delta_i = \Delta$.
- Δ acts as a smoothing parameter.



- In a D-dimensional space, using M bins in each dimension will require M^D bins!

Non-parametric Methods (3)

- Assume observations drawn from a density $p(\mathrm{x})$ and consider a small region R containing $\mathbf{x}$ such that

$$P = \int_{\mathcal{R}} p(\mathbf{x})\,\mathrm{d}\mathbf{x}.$$

- The probability that K out of N observations lie inside R is $\mathrm{Bin}(K|N,P)$ and if N is large

$$K \simeq NP.$$

- If the volume of R, V, is sufficiently small, $p(\mathbf{x})$ is approximately constant over R and

$$P \simeq p(\mathbf{x})V$$

Thus

$$\boxed{p(\mathbf{x}) = \frac{K}{NV}.}$$

V small, yet $K>0$, therefore N large?

Non-parametric Methods (4)

Kernel Density Estimation: fix V, estimate K from the data. Let R be a hypercube centred on x and define the kernel function (Parzen window)

$$k((\mathbf{x}-\mathbf{x}_n)/h) = \begin{cases} 1, & |(x_i - x_{ni})/h| \leqslant 1/2, \qquad i = 1, \ldots, D, \\ 0, & \text{otherwise.} \end{cases}$$

It follows that

$$K = \sum_{n=1}^{N} k\left(\frac{\mathbf{x}-\mathbf{x}_n}{h}\right)$$ and hence $$p(\mathbf{x}) = \frac{1}{N}\sum_{n=1}^{N} \frac{1}{h^D} k\left(\frac{\mathbf{x}-\mathbf{x}_n}{h}\right).$$

Non-parametric Methods (5)

To avoid discontinuities in p(x), use a smooth kernel, e.g. a Gaussian

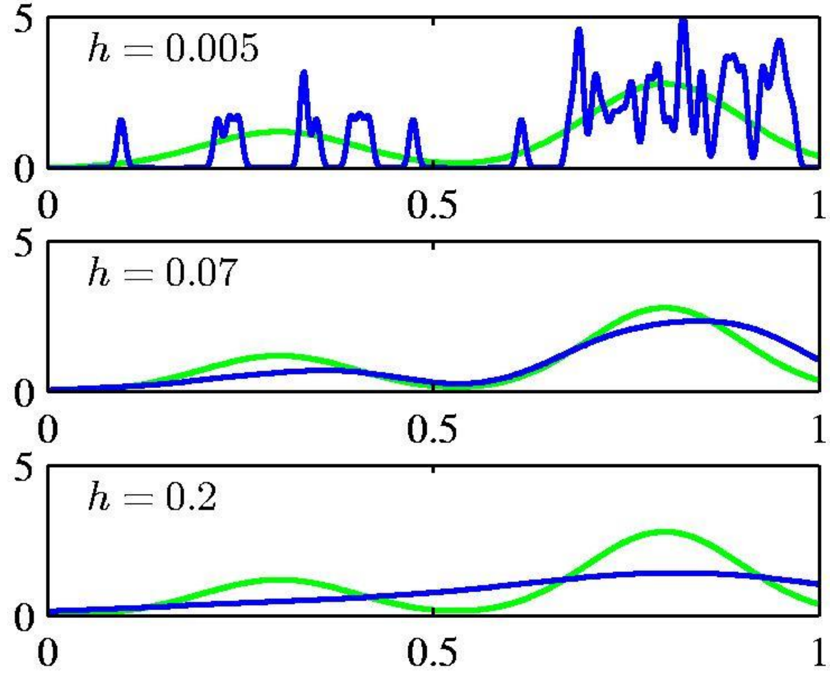
$$p(\mathbf{x}) = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{(2\pi h^2)^{D/2}}$$

$$\exp\left\{-\frac{\|\mathbf{x}-\mathbf{x}_n\|^2}{2h^2}\right\}$$

Any kernel such that

$$\begin{aligned} k(\mathbf{u}) &\geqslant 0, \\ \int k(\mathbf{u})\,\mathrm{d}\mathbf{u} &= 1 \end{aligned}$$

will work.

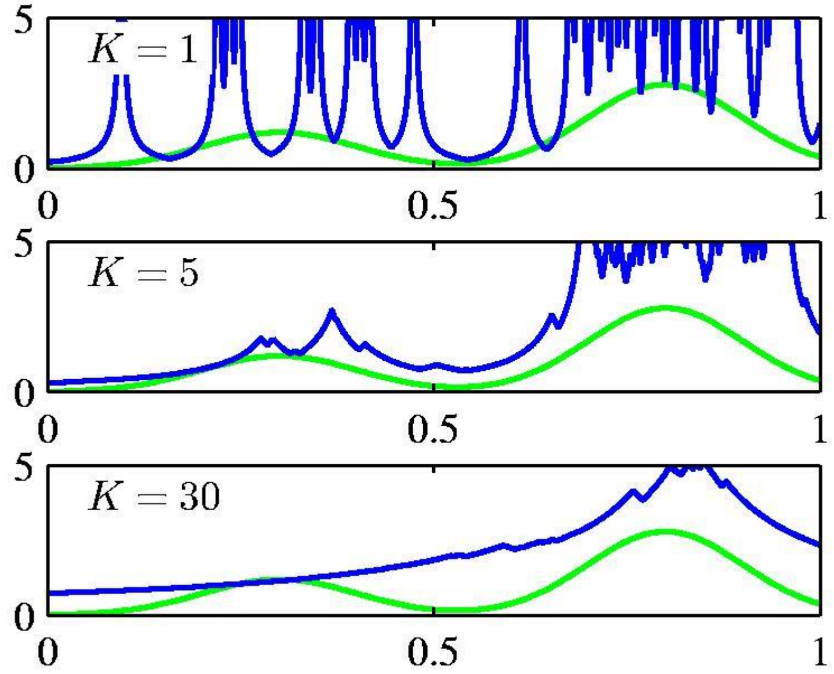


h acts as a smoother.

Non-parametric Methods (6)

Nearest Neighbour Density Estimation: fix K, estimate V from the data. Consider a hypersphere centred on x and let it grow to a volume, V^*, that includes K of the given N data points. Then

$$p(\mathbf{x}) \simeq \frac{K}{NV^\star}.$$



K acts as a smoother.

Non-parametric Methods (7)

- Nonparametric models (not histograms) requires storing and computing with the entire data set.
- Parametric models, once fitted, are much more efficient in terms of storage and computation.

Outlines

- Binary Distributions
- Multinomial Distributions
- Gaussian Distributions
- Exponential Families
- Non-informative Priors
- Non-parametric Methods
- KNN

K-Nearest-Neighbours for Classification (1)

- Given a data set with N_k data points from class $\mathrm{C_k}$, we have $\sum_k N_k = N$

number of data in a region

$$p(\mathbf{x}) = \frac{K}{NV}$$

number of total data

volume of the region

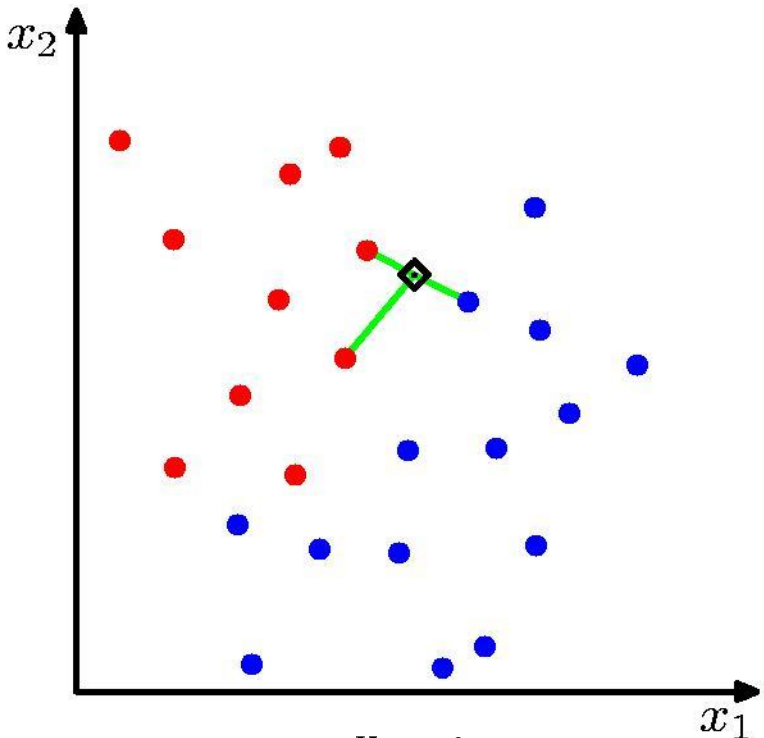
and correspondingly

$$p(\mathbf{x}|\mathcal{C}_k) = \frac{K_k}{N_k V}.$$

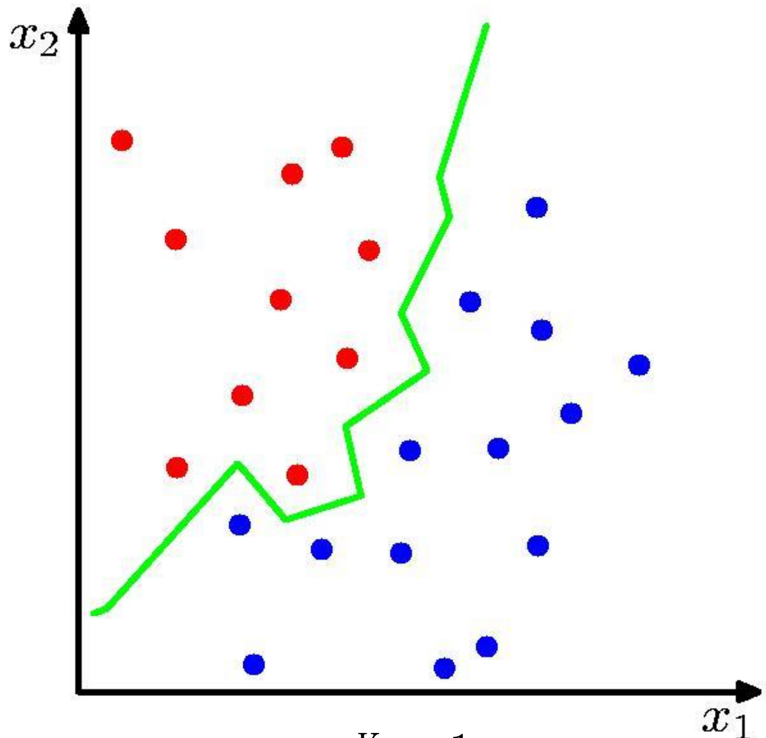
- Since $p(\mathcal{C}_k) = N_k/N$, Bayes' theorem gives

$$p(\mathcal{C}_k|\mathbf{x}) = \frac{p(\mathbf{x}|\mathcal{C}_k)p(\mathcal{C}_k)}{p(\mathbf{x})} = \frac{K_k}{K}.$$

K-Nearest-Neighbours for Classification (2)

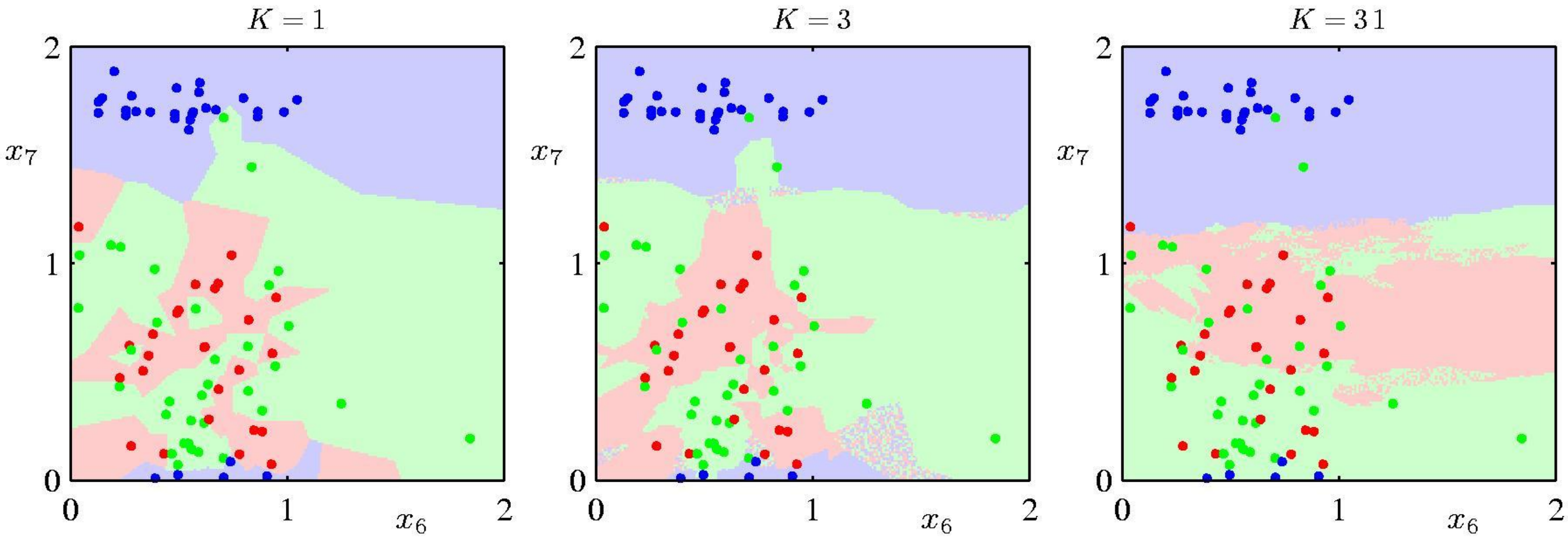


K = 3



K = 1

K-Nearest-Neighbours for Classification (3)



- K acts as a smother
- For $N \to \infty$, the error rate of the 1-nearest-neighbour classifier is never more than twice the optimal error (obtained from the true conditional class distributions).

Summary

- Binary Distributions
- Multinomial Distributions
- Gaussian Distributions
- Exponential Families
- Non-information Priors
- Non-parametric Methods
- KNN