
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-information Prior
 - Non-parametric Methods
 - KNN
-

Parametric Distributions

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

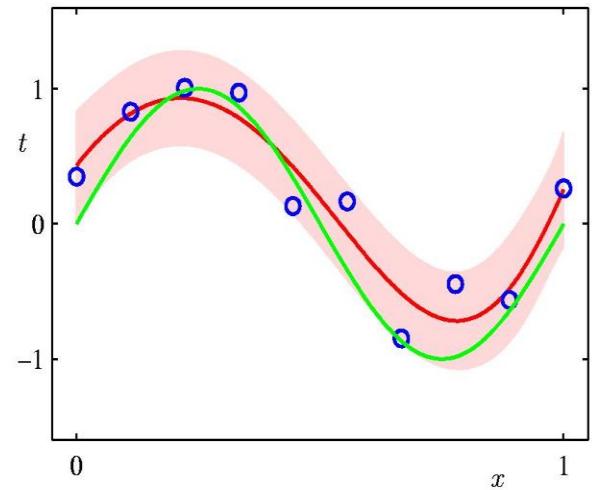
Need to determine θ given $\{\mathbf{x}_1, \dots, \mathbf{x}_N\}$

Representation: θ^* or $p(\theta)$?

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

Recall Curve Fitting

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



Binary Variables (1)

Coin flipping: heads=1, tails=0

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

Bernoulli Distribution

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

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

$$\text{var}[x] = \mu(1 - \mu)$$

Binary Variables (2)

N coin flips:

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

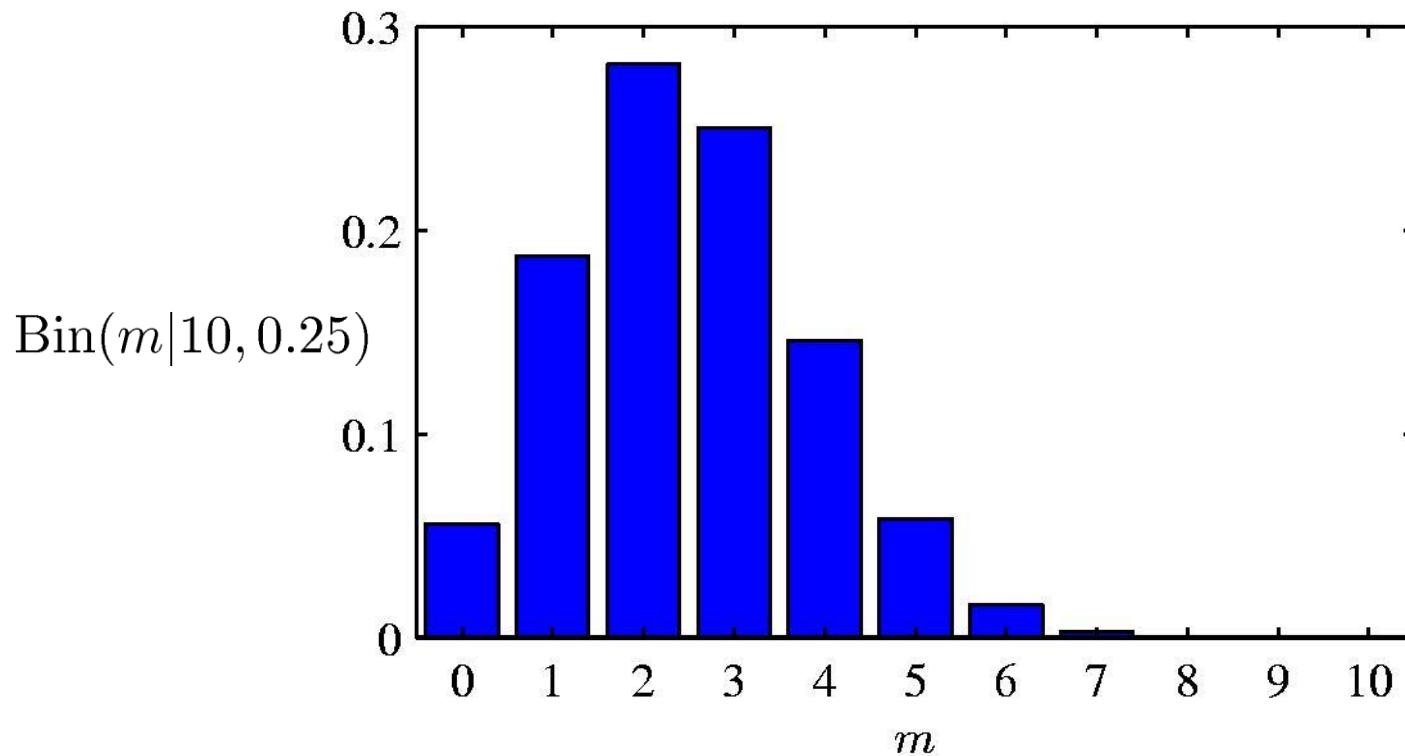
Binomial Distribution

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

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

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

Binomial Distribution



Parameter Estimation (1)

ML for Bernoulli

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

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

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

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

Parameter Estimation (2)

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

Prediction: *all* future tosses will land heads up

Overfitting to \mathcal{D}

Beta Distribution

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

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

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

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

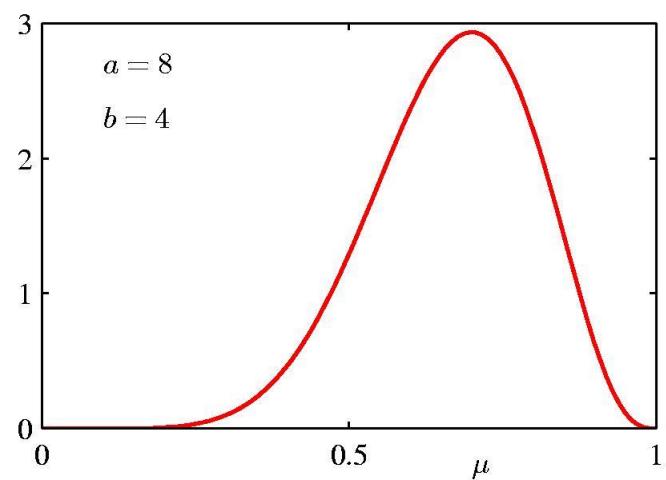
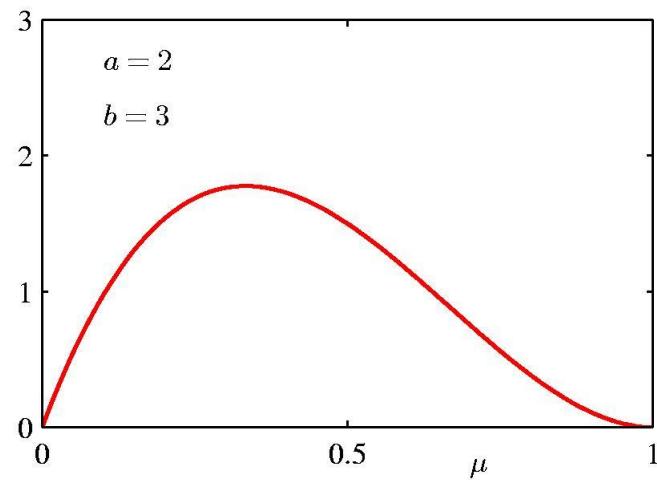
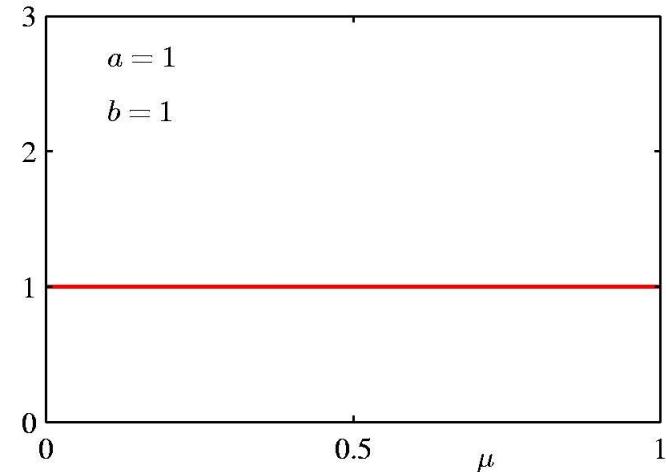
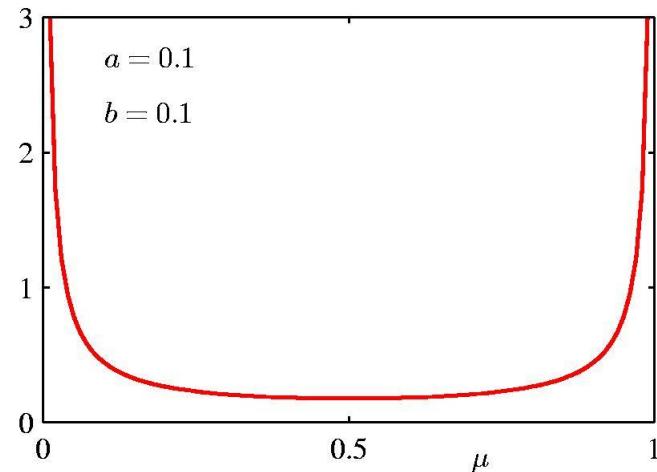
Bayesian Bernoulli

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

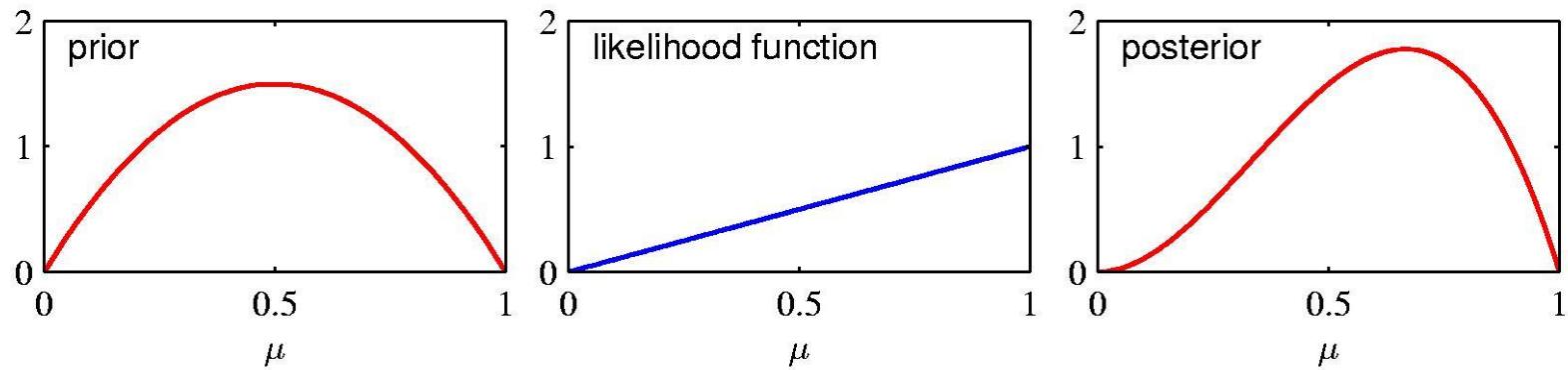
$$a_N = a_0 + m \quad b_N = b_0 + (N - m)$$

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

Beta Distribution



Prior \cdot Likelihood = Posterior



Properties of the Posterior

As the size of the data set, N , increase

$$a_N \rightarrow m$$

$$b_N \rightarrow N - m$$

$$\mathbb{E}[\mu] = \frac{a_N}{a_N + b_N} \rightarrow \frac{m}{N} = \mu_{\text{ML}}$$

$$\text{var}[\mu] = \frac{a_N b_N}{(a_N + b_N)^2 (a_N + b_N + 1)} \rightarrow 0$$

Prediction under the Posterior

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

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

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-information Priors
 - Non-parametric Methods
 - KNN
-

Multinomial Variables

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

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

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

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

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

ML Parameter estimation

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

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

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

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

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

The Multinomial Distribution

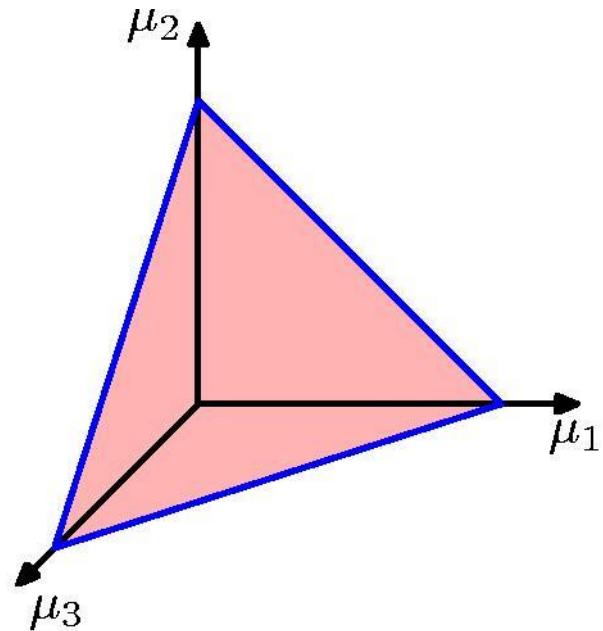
$$\begin{aligned}\text{Mult}(m_1, m_2, \dots, m_K | \boldsymbol{\mu}, N) &= \binom{N}{m_1 m_2 \dots m_K} \prod_{k=1}^K \mu_k^{m_k} \\ \mathbb{E}[m_k] &= N\mu_k \\ \text{var}[m_k] &= N\mu_k(1 - \mu_k) \\ \text{cov}[m_j m_k] &= -N\mu_j \mu_k\end{aligned}$$

The Dirichlet Distribution

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

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

Conjugate prior for the multinomial distribution.

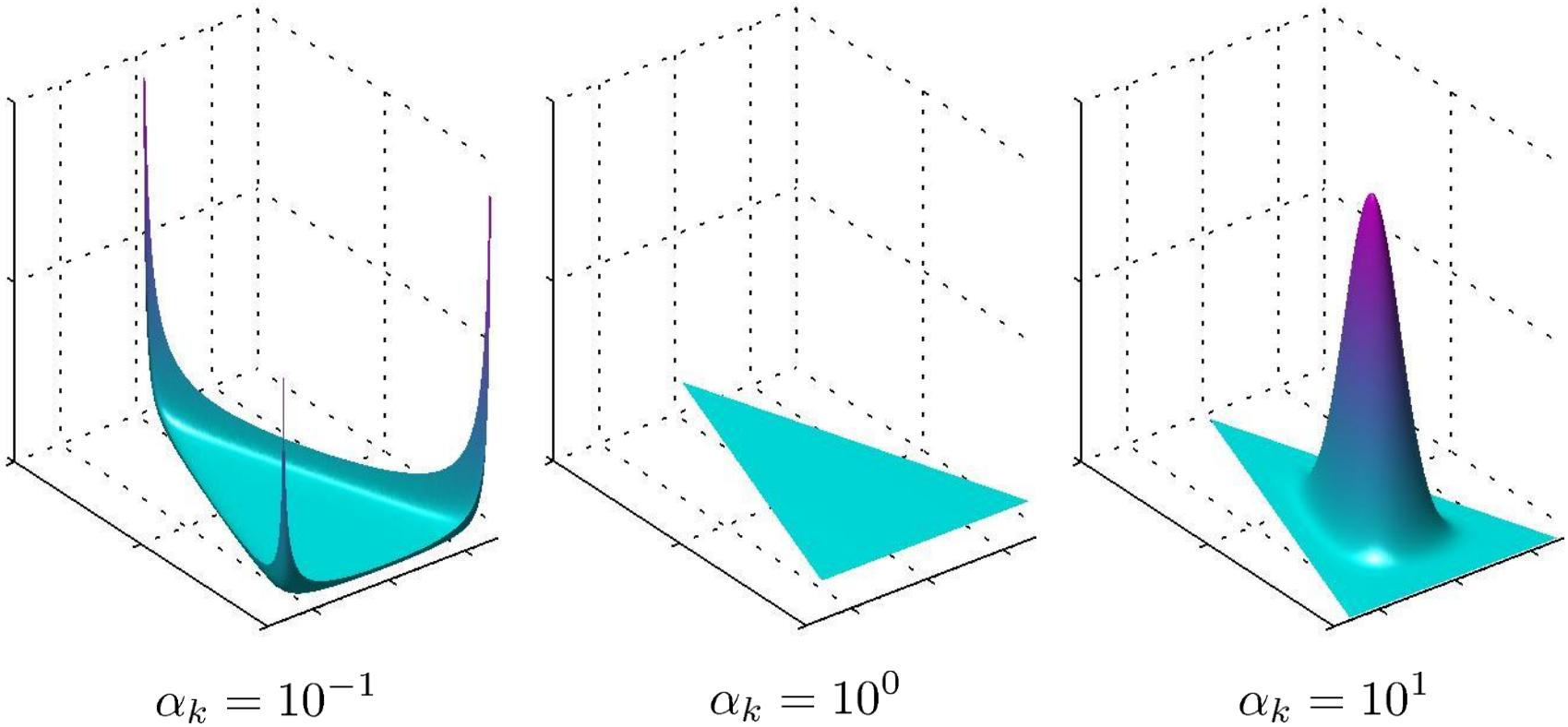


Bayesian Multinomial (1)

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

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

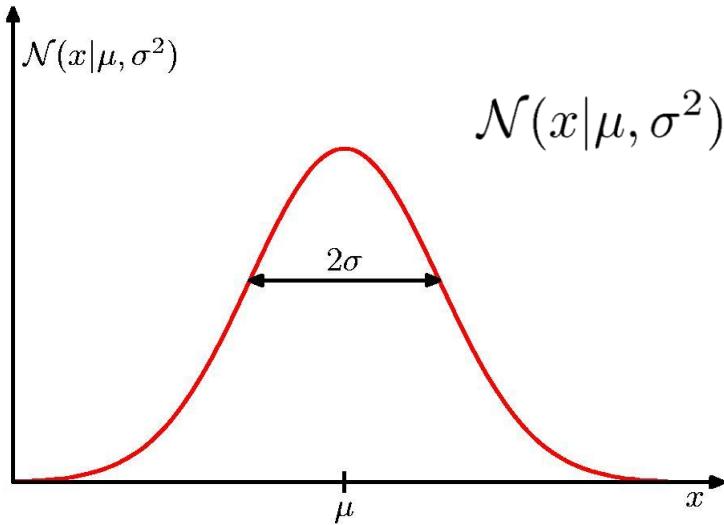
Bayesian Multinomial (2)



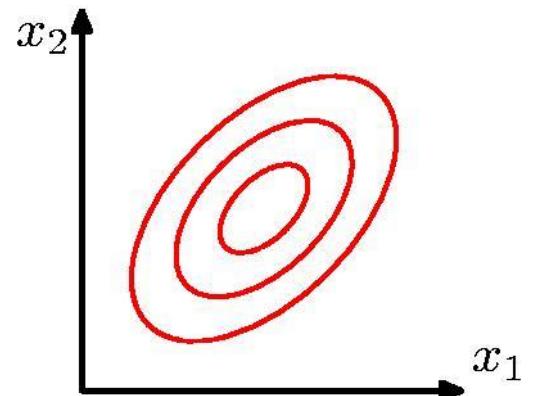
Outlines

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

The Gaussian Distribution



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

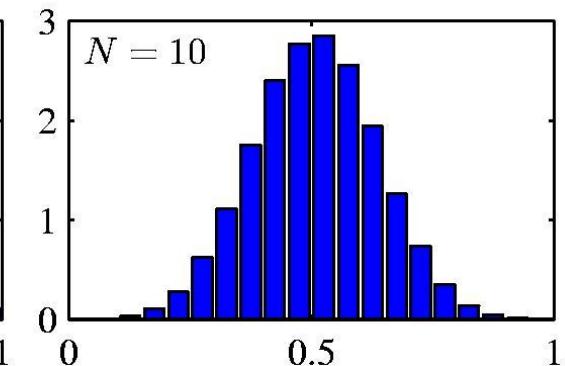
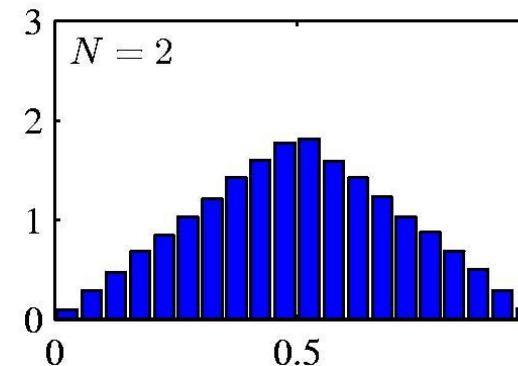
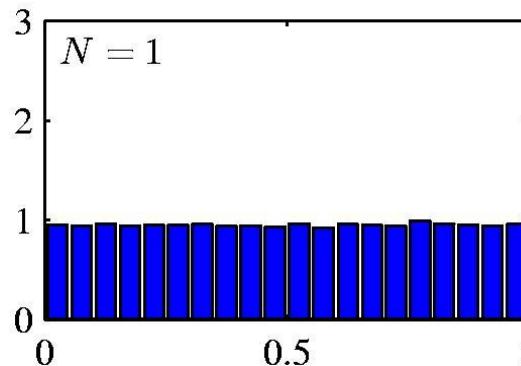


$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp \left\{ -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right\}$$

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.



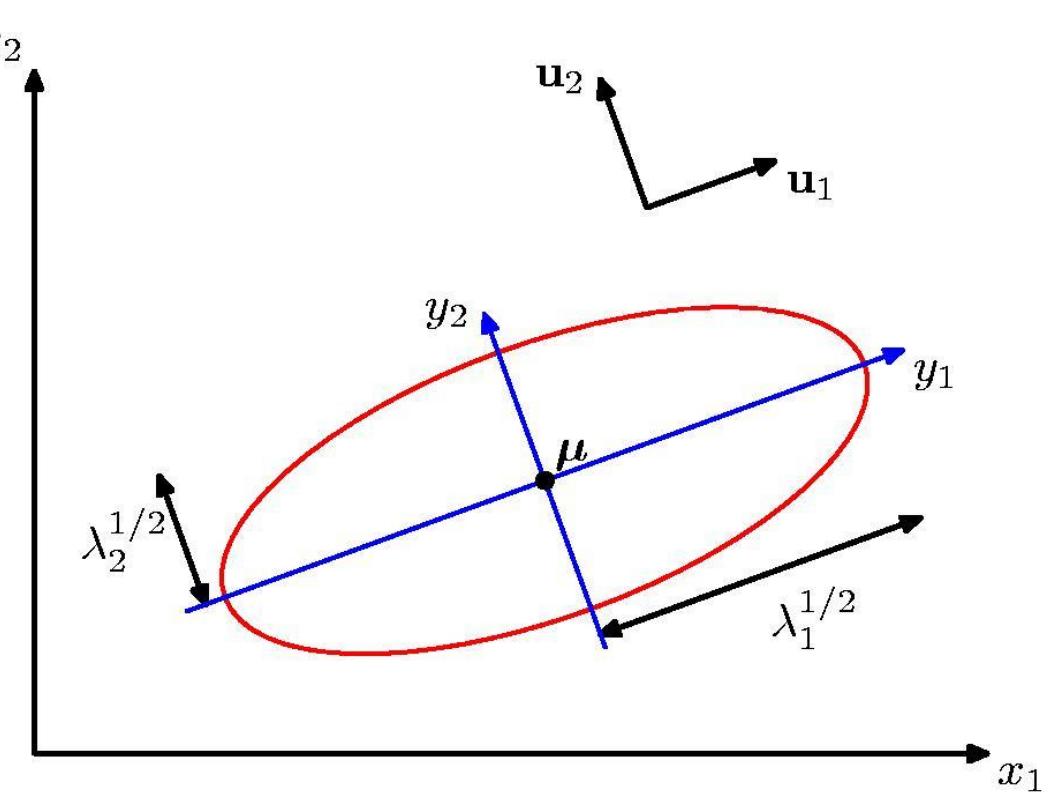
Geometry of the Multivariate Gaussian

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

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

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

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



Moments of the Multivariate Gaussian (1)

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

thanks to anti-symmetry of \mathbf{z}

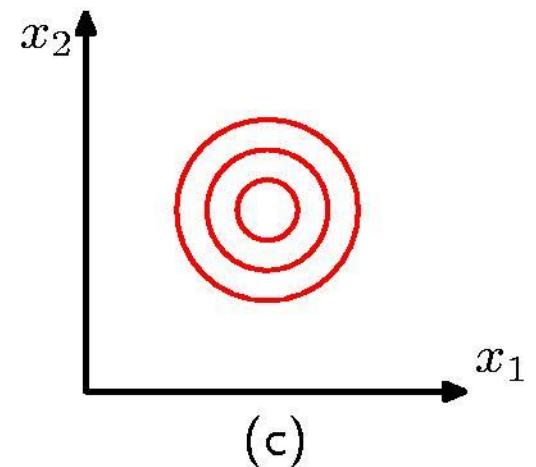
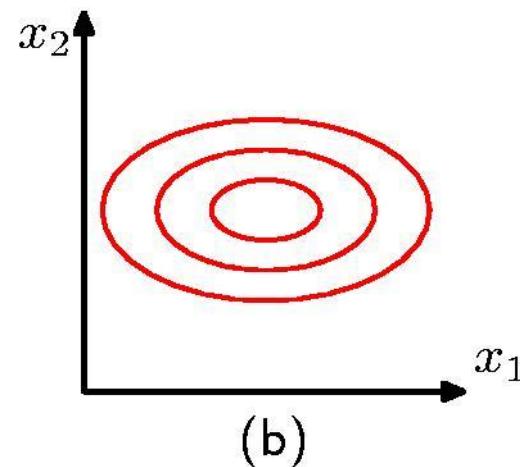
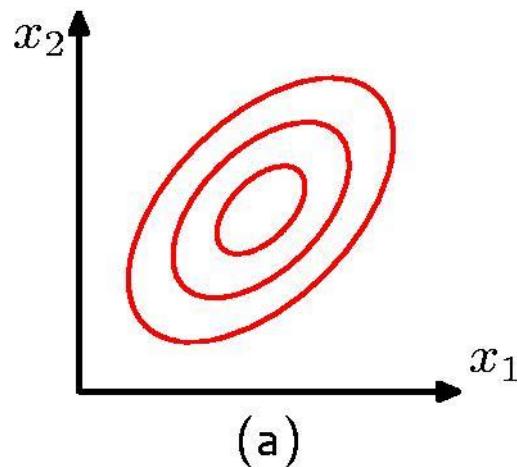
$$\mathbb{E}[\mathbf{x}] = \boldsymbol{\mu}$$

Moments of the Multivariate Gaussian (2)

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

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

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



Properties of Gaussians

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

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



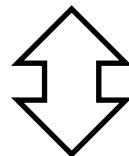
Precision

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

$$\left[\begin{array}{rcl} \sigma^{-2} & = & \sigma_1^{-2} + \sigma_2^{-2} \\ \sigma^{-2}\mu & = & \sigma_1^{-2}\mu_1 + \sigma_2^{-2}\mu_2 \end{array} \right]$$

Properties of Gaussians

$$p_{X_1}(x)p_{X_2}(x) \propto e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}} e^{-\frac{(x-\mu_2)^2}{2\sigma_2^2}}$$



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

Quadratic terms of x (x^2) are equal

1st order terms of x are also equal

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

Multivariate Gaussians

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

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

(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

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

Mean

Bayes' Theorem for Gaussian Variables

Given

$$y = Ax + v$$

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

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

$$x = Hy + u$$

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

we have

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

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

Bayes' Theorem for Gaussian Variables

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

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

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

Matrix Inversion Lemma

If A, C, BCD are non-singular 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

$$\begin{aligned} & [A + BCD] \left[A^{-1} - A^{-1}B[C^{-1} + DA^{-1}B]^{-1}DA^{-1} \right] \\ &= I + BCDA^{-1} - B[C^{-1} + DA^{-1}B]^{-1}DA^{-1} \\ &\quad - BCDA^{-1}B[C^{-1} + DA^{-1}B]^{-1}DA^{-1} \\ &= I + BCDA^{-1} - B\{I + CDA^{-1}B\}[C^{-1} + DA^{-1}B]^{-1}DA^{-1} \\ &= I + BCDA^{-1} - BC\{C^{-1} + DA^{-1}B\}[C^{-1} + DA^{-1}B]^{-1}DA^{-1} \\ &= I \end{aligned}$$

Bayes' Theorem for Gaussian Variables

Then

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

$$\begin{cases} L &= (I - KA)\Sigma \\ Hy &= \mu + K(y - A\mu) \end{cases}$$

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

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

Partitioned Gaussian Distributions

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

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

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

$$\mathbf{x}_a - \boldsymbol{\mu}_a = \mathbf{A}(\mathbf{x}_b - \boldsymbol{\mu}_b) + \mathbf{w} \Rightarrow \boldsymbol{\mu}_{a|b} - \boldsymbol{\mu}_a = \mathbf{A}(\mathbf{x}_b - \boldsymbol{\mu}_b)$$

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

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

Partitioned Gaussian Distributions

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

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

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

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

Inverse Covariance Matrix

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

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

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

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

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

Inverse Matrix Lemma

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

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

Inverse Covariance Matrix

$$\begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{pmatrix}^{-1} = \begin{pmatrix} \Lambda_{aa} & \Lambda_{ab} \\ \Lambda_{ba} & \Lambda_{bb} \end{pmatrix}$$

$$\Lambda_{aa} = (\Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba})^{-1}$$

$$\Lambda_{ab} = -(\Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba})^{-1}\Sigma_{ab}\Sigma_{bb}^{-1}$$

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

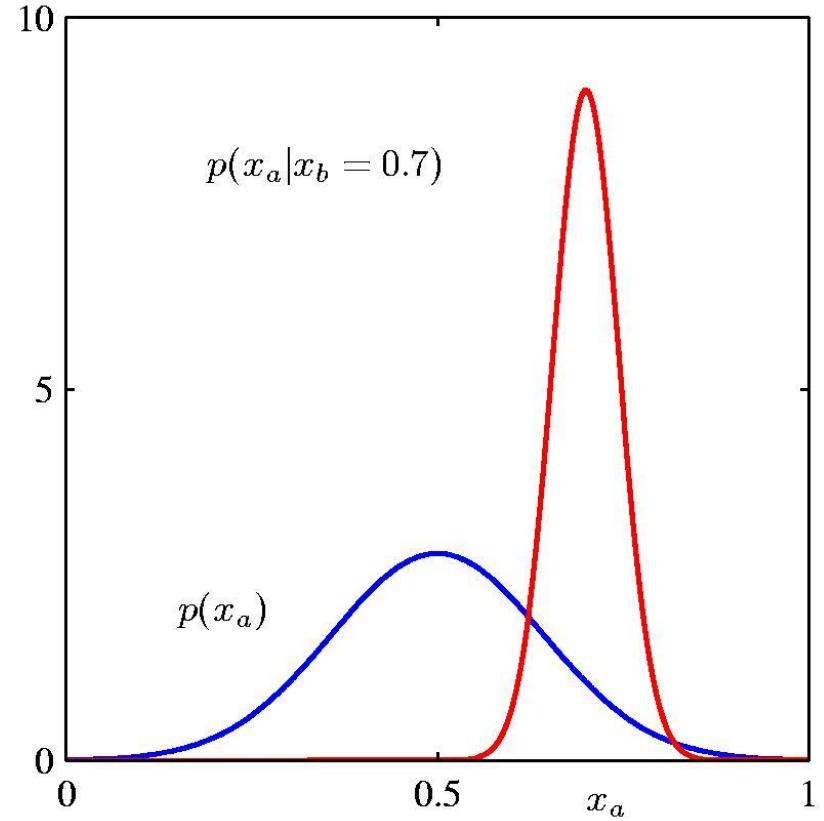
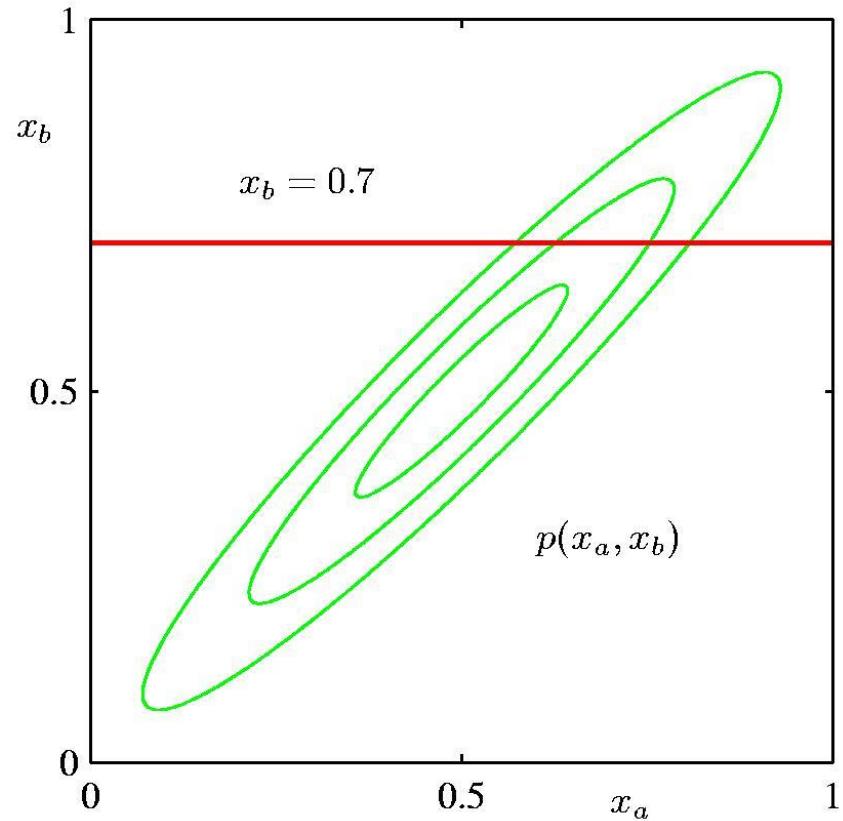
Partitioned Conditionals and Marginals

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

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

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

Partitioned Conditionals and Marginals



Bayes' Theorem for Gaussian Variables

Given

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

we have

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

where

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

Maximum Likelihood for the Gaussian (1)

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

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

Sufficient statistics

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

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

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

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

Similarly

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

Maximum Likelihood for the Gaussian (3)

Under the true distribution

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

Hence define

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

Sequential Estimation

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

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

correction given \mathbf{x}_N
correction weight
old estimate

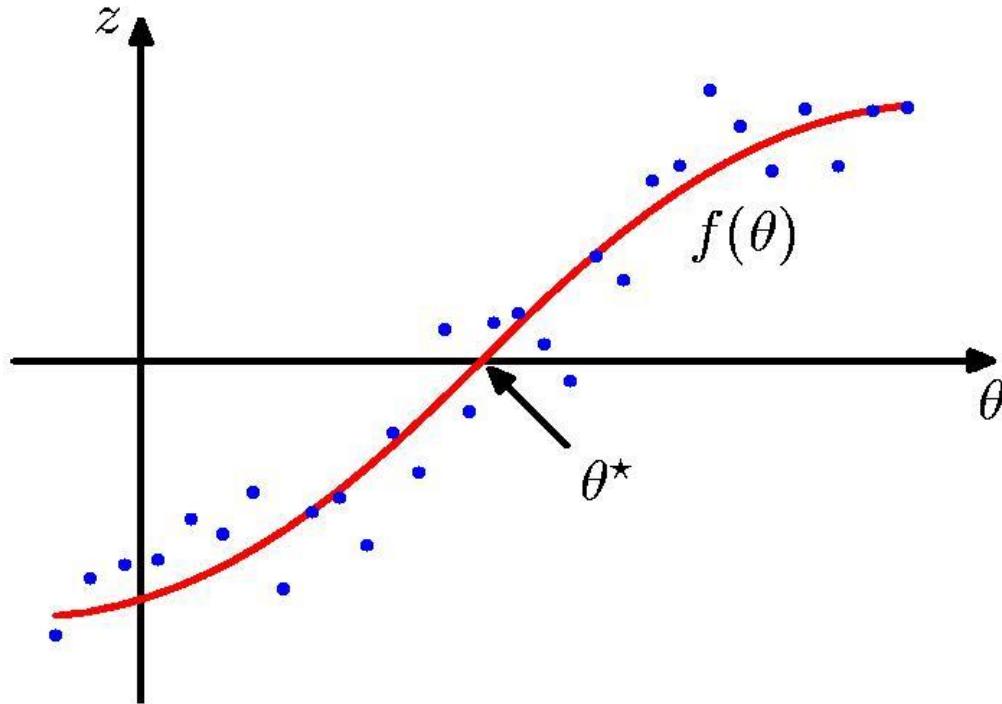
The Robbins-Monro Algorithm (1)

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

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

Seek θ^* such that $f(\theta^*) = 0$.

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 θ^* are then given by

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

Conditions on a_N for convergence :

$$\lim_{N \rightarrow \infty} a_N = 0 \quad \sum_{N=1}^{\infty} a_N = \infty \quad \sum_{N=1}^{\infty} a_N^2 < \infty$$

Robbins-Monro for Maximum Likelihood (1)

Regarding

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

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

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

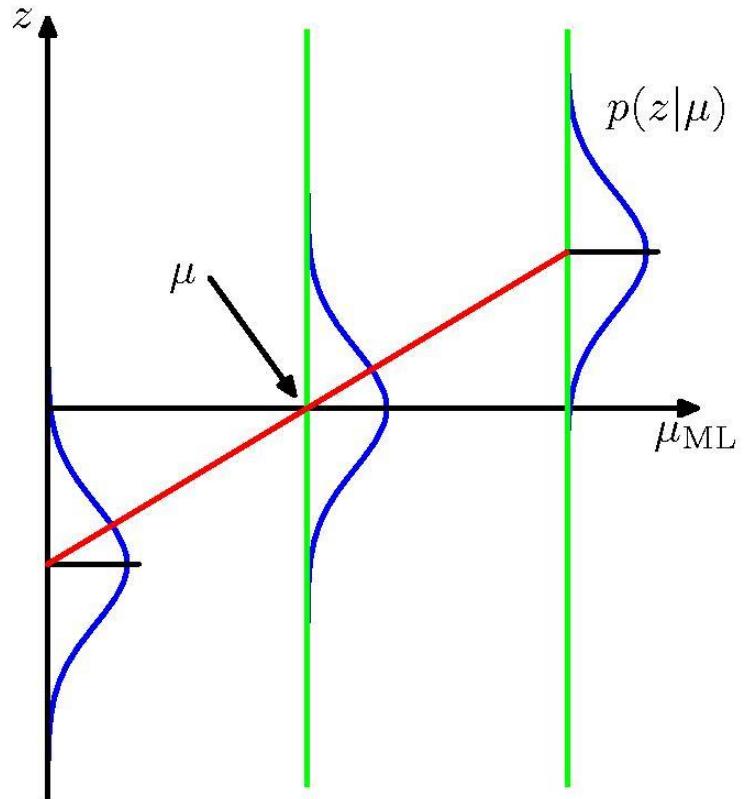
Robbins-Monro for Maximum Likelihood (2)

Example: estimate the mean of a Gaussian.

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

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

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



Bayesian Inference for the Gaussian (1)

Assume σ^2 is known. Given i.i.d. data

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

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

this gives the posterior

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

Completing the square over μ , we see that

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

Bayesian Inference for the Gaussian (3)

... where

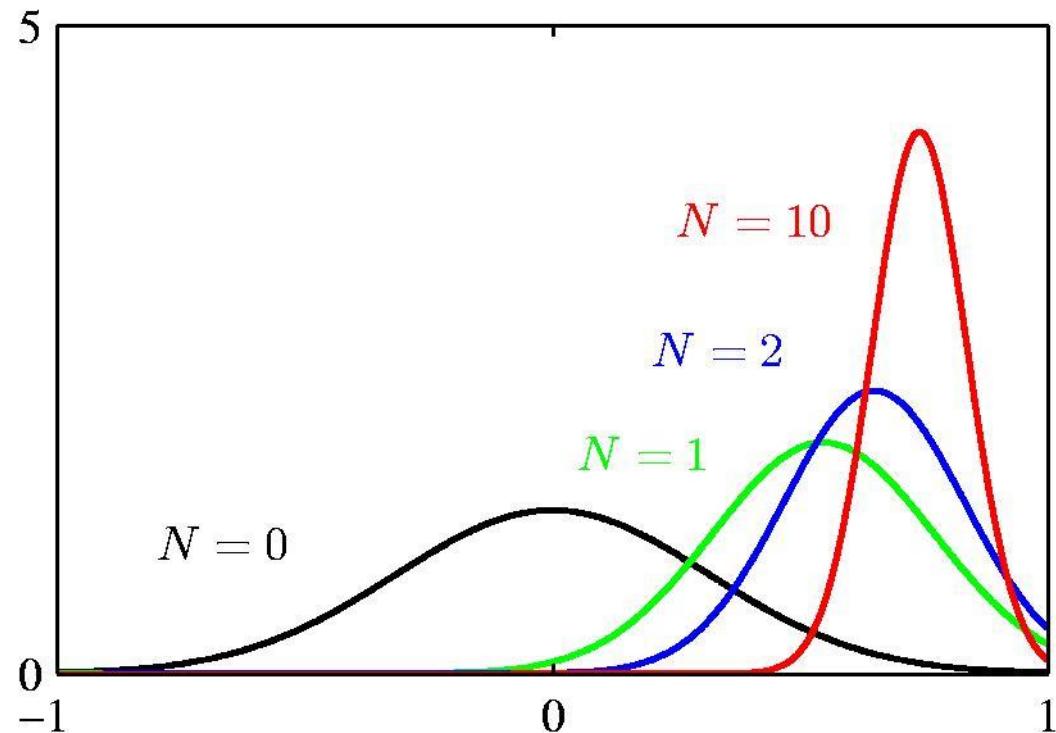
$$\begin{aligned}\mu_N &= \frac{\sigma^2}{N\sigma_0^2 + \sigma^2} \mu_0 + \frac{N\sigma_0^2}{N\sigma_0^2 + \sigma^2} \mu_{\text{ML}}, & \mu_{\text{ML}} &= \frac{1}{N} \sum_{n=1}^N x_n \\ \frac{1}{\sigma_N^2} &= \frac{1}{\sigma_0^2} + \frac{N}{\sigma^2}.\end{aligned}$$

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



Bayesian Inference for the Gaussian (5)

Sequential Estimation

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

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 .

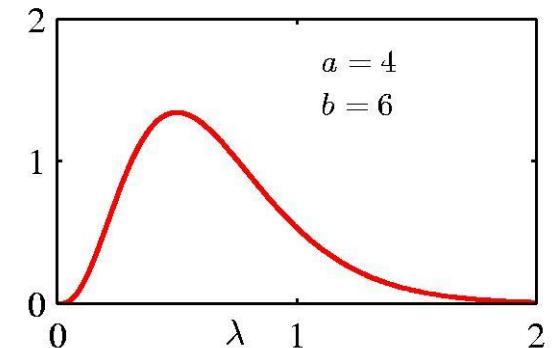
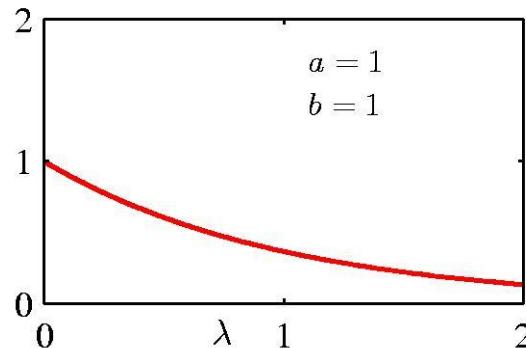
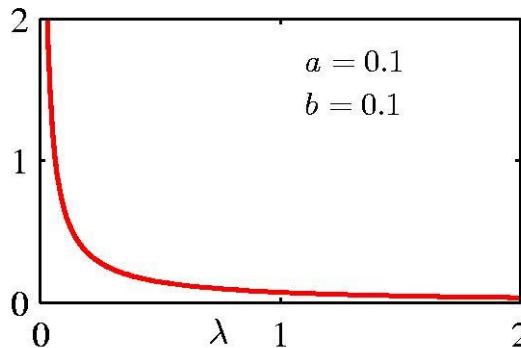
Bayesian Inference for the Gaussian (7)

The Gamma distribution

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

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

$$\text{var}[\lambda] = \frac{a}{b^2}$$



Bayesian Inference for the Gaussian (8)

Now we combine a Gamma prior, $\text{Gam}(\lambda|a_0, b_0)$, 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 $\text{Gam}(\lambda|a_N, b_N)$ with

$$a_N = a_0 + \frac{N}{2}$$

$$b_N = b_0 + \frac{1}{2} \sum_{n=1}^N (x_n - \mu)^2 = b_0 + \frac{N}{2} \sigma_{\text{ML}}^2.$$

Bayesian Inference for the Gaussian (9)

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

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

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

Bayesian Inference for the Gaussian (10)

The Gaussian-gamma distribution

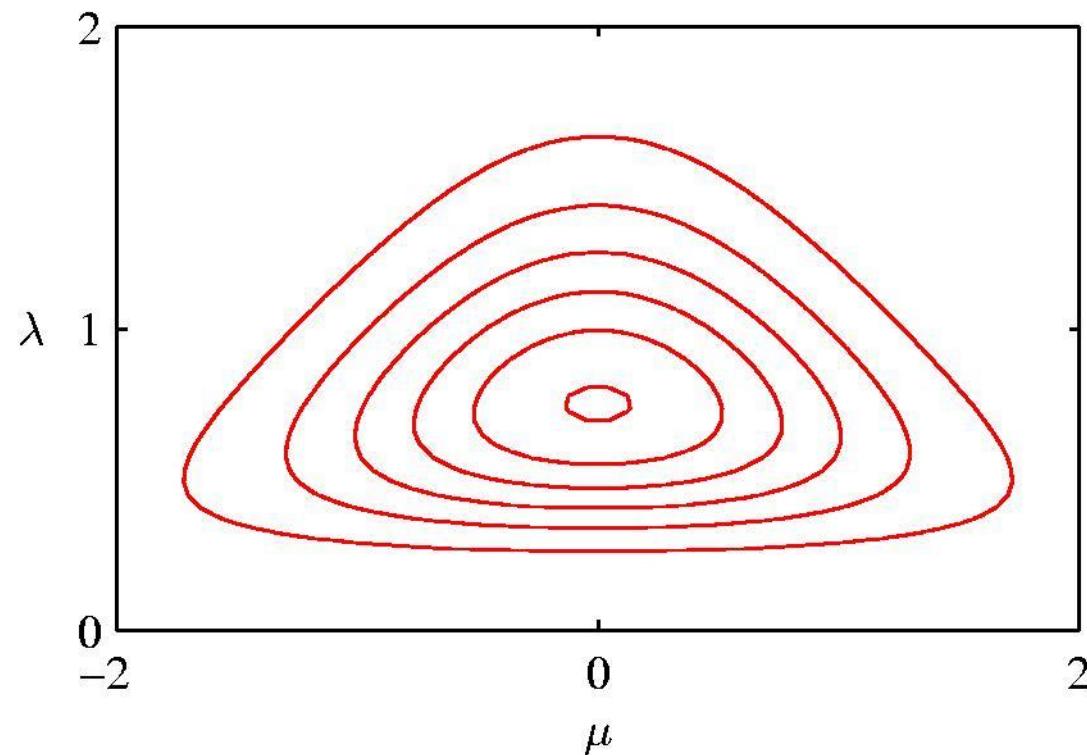
$$p(\mu, \lambda) = \mathcal{N}(\mu | \mu_0, (\beta\lambda)^{-1}) \text{Gam}(\lambda | a, b)$$

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


- Quadratic in μ .
- Linear in λ .
- Gamma distribution over λ .
- Independent of μ .

Bayesian Inference for the Gaussian (11)

The Gaussian-gamma distribution



Bayesian Inference for the Gaussian (12)

Multivariate conjugate priors

- μ unknown, Λ known: $p(\mu)$ Gaussian.
- Λ unknown, μ known: $p(\Lambda)$ Wishart,

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

- Λ and μ unknown: $p(\mu, \Lambda)$ Gaussian-Wishart, $p(\mu, \Lambda | \mu_0, \beta, \mathbf{W}, \nu) = \mathcal{N}(\mu | \mu_0, (\beta \Lambda)^{-1}) \mathcal{W}(\Lambda | \mathbf{W}, \nu)$
-

Student's t-Distribution

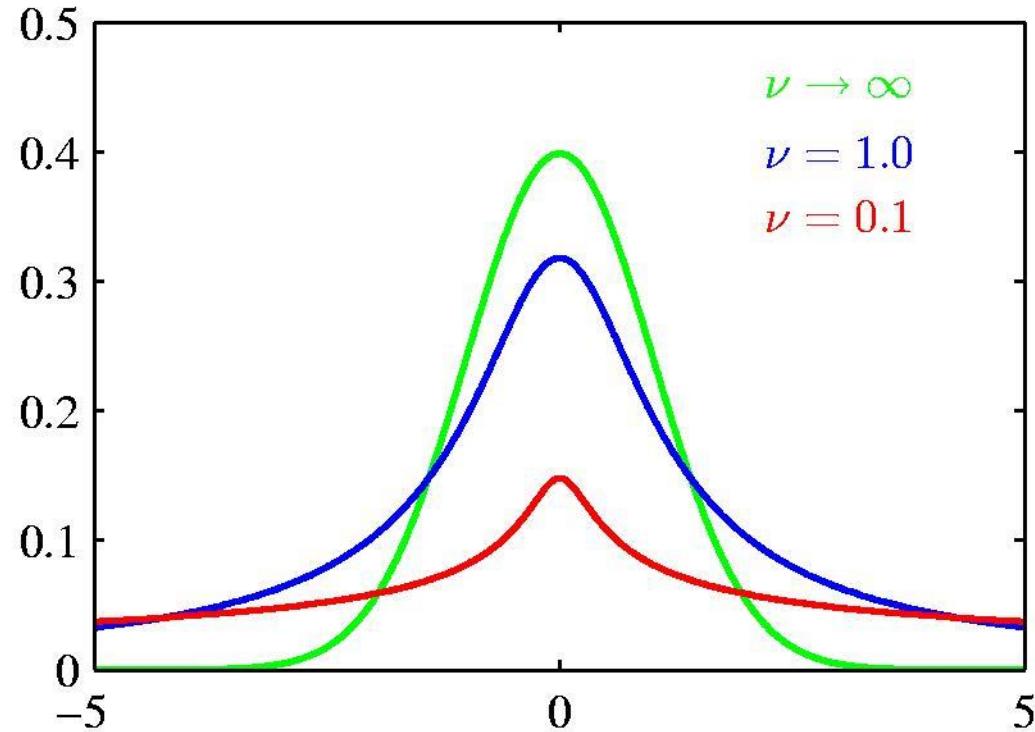
$$\begin{aligned} p(x|\mu, a, b) &= \int_0^\infty \mathcal{N}(x|\mu, \tau^{-1}) \text{Gam}(\tau|a, b) d\tau \\ &= \int_0^\infty \mathcal{N}(x|\mu, (\eta\lambda)^{-1}) \text{Gam}(\eta|\nu/2, \nu/2) d\eta \quad \leftarrow \text{---} \\ &= \frac{\Gamma(\nu/2 + 1/2)}{\Gamma(\nu/2)} \left(\frac{\lambda}{\pi\nu}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

where

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

Infinite mixture of Gaussians. -----

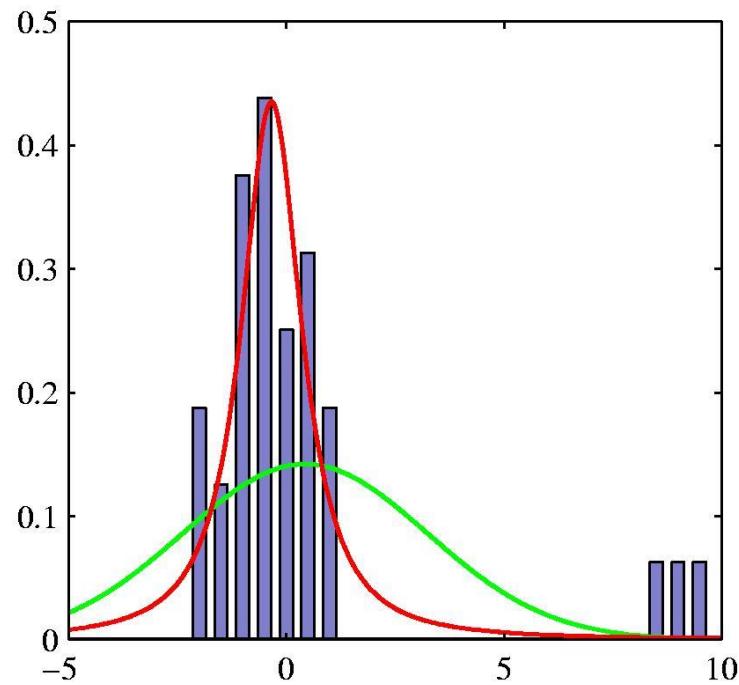
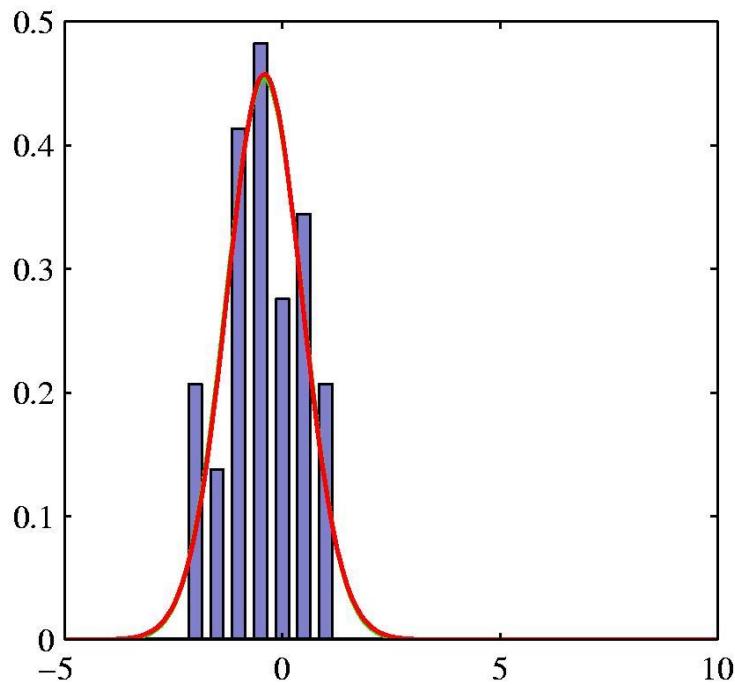
Student's t-Distribution



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

Student's t-Distribution

Robustness to outliers: Gaussian vs t-distribution.



Student's t-Distribution

The D -variate case:

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

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

Properties:

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

Periodic variables

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

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

von Mises Distribution (1)

This requirement is satisfied by

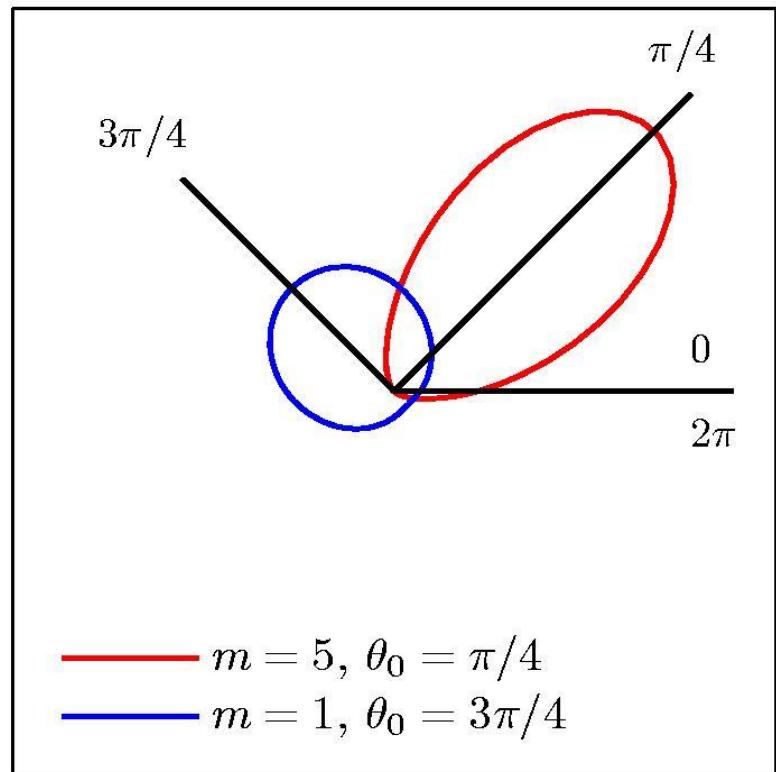
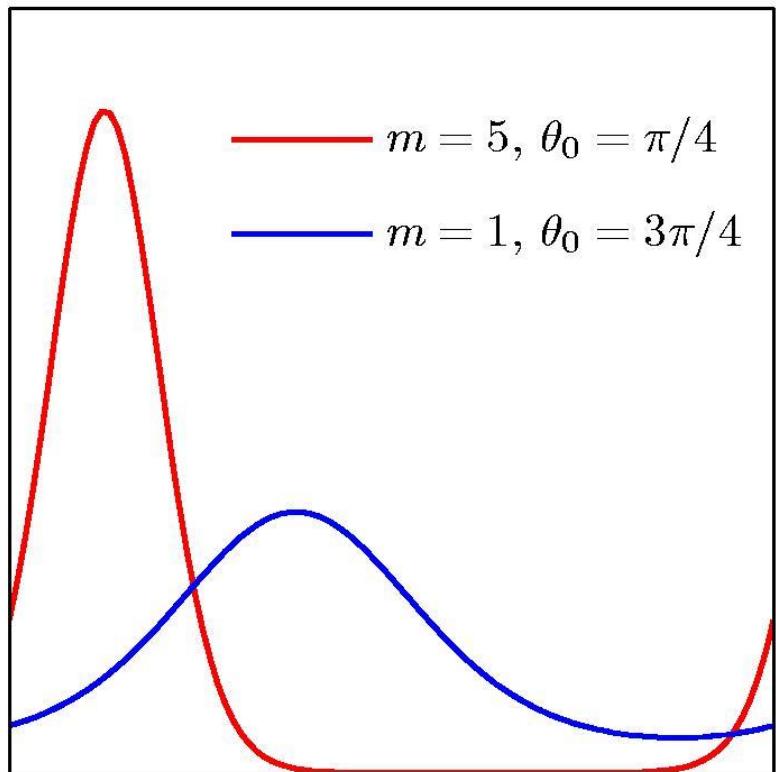
$$p(\theta|\theta_0, m) = \frac{1}{2\pi I_0(m)} \exp \{m \cos(\theta - \theta_0)\}$$

where

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

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

von Mises Distribution (2)



Maximum Likelihood for von Mises

Given a data set, $\mathcal{D} = \{\theta_1, \dots, \theta_N\}$, 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

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

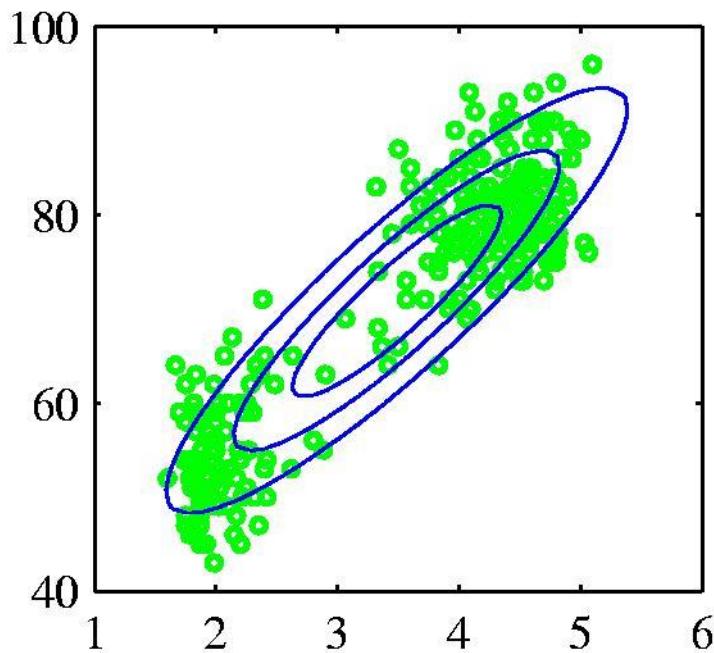
Similarly, maximizing with respect to m we get

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

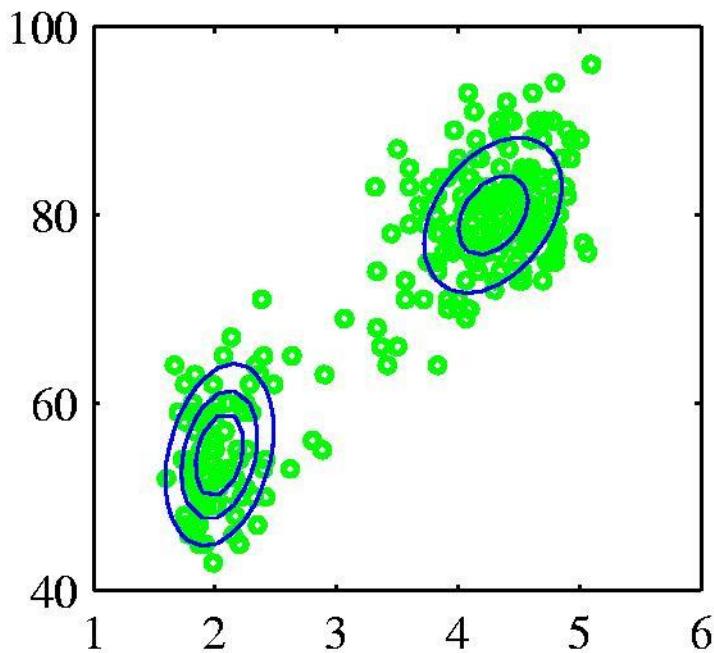
which can be solved numerically for m_{ML} .

Mixtures of Gaussians (1)

Old Faithful data set



Single Gaussian



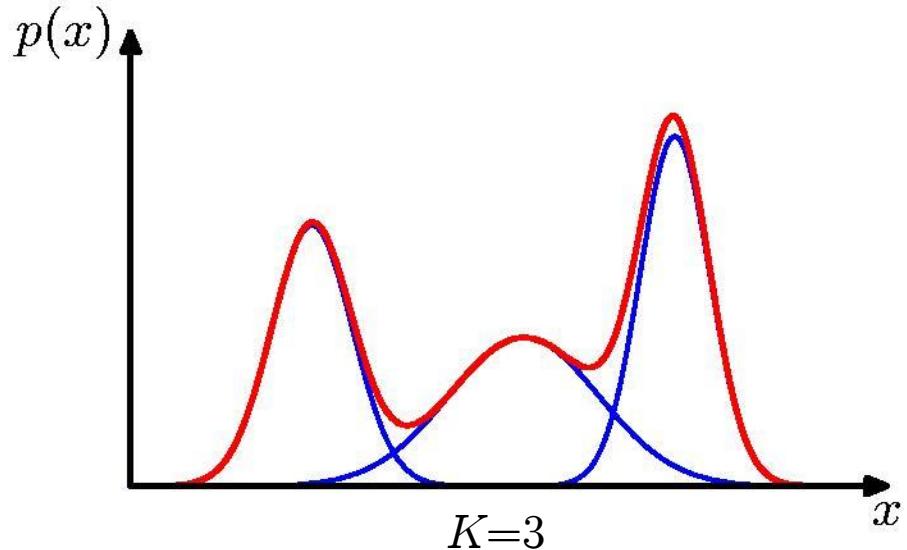
Mixture of two Gaussians

Mixtures of Gaussians (2)

Combine simple models
into a complex model:

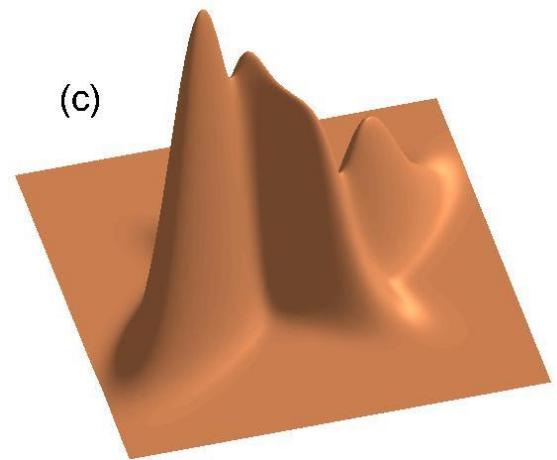
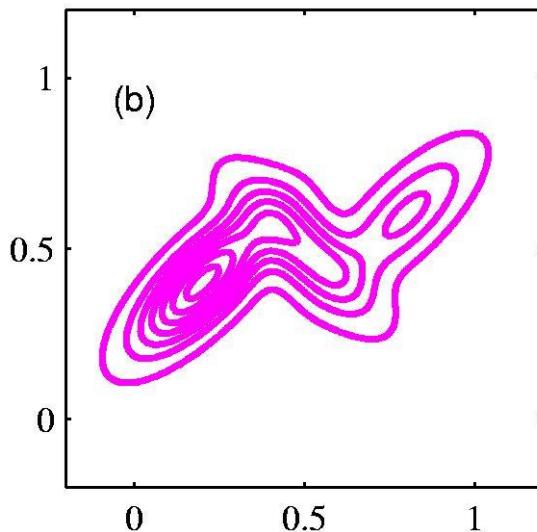
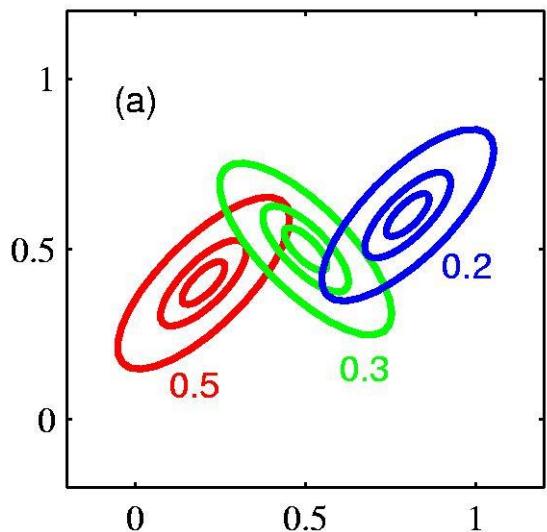
$$p(\mathbf{x}) = \sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{x} | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

↑
Component
Mixing coefficient



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

Mixtures of Gaussians (3)



Mixtures of Gaussians (4)

Determining parameters μ , Σ , and π using maximum log likelihood

$$\ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^N \ln \left\{ \sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right\}$$

Log of a sum; no closed form maximum.

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

Outlines

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

The Exponential Family (1)

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

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

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

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

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

The Exponential Family (2.1)

The Bernoulli Distribution

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

Comparing with the general form we see that

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

The Exponential Family (2.2)

The Bernoulli distribution can hence be written as

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

where

$$u(x) = x$$

$$h(x) = 1$$

$$g(\eta) = 1 - \sigma(\eta) = \sigma(-\eta).$$

The Exponential Family (3.1)

The Multinomial Distribution

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

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

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

NOTE: The 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

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

Here the η_k parameters are independent. Note that

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

The Exponential Family (3.3)

The Multinomial distribution can then be written as

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

where

$$\boldsymbol{\eta} = (\eta_1, \dots, \eta_{M-1}, 0)^T$$

$$\mathbf{u}(\mathbf{x}) = \mathbf{x}$$

$$h(\mathbf{x}) = 1$$

$$g(\boldsymbol{\eta}) = \left(1 + \sum_{k=1}^{M-1} \exp(\eta_k) \right)^{-1}.$$

The Exponential Family (4)

The Gaussian Distribution

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

where

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

ML for the Exponential Family (1)

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

$$\nabla g(\boldsymbol{\eta}) \int h(\mathbf{x}) \exp \{ \boldsymbol{\eta}^T \mathbf{u}(\mathbf{x}) \} d\mathbf{x} + g(\boldsymbol{\eta}) \int h(\mathbf{x}) \exp \{ \boldsymbol{\eta}^T \mathbf{u}(\mathbf{x}) \} \mathbf{u}(\mathbf{x}) d\mathbf{x} = 0$$


$1/g(\boldsymbol{\eta})$ $\mathbb{E}[\mathbf{u}(\mathbf{x})]$

Thus

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

ML for the Exponential Family (2)

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

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

Thus we have

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

Sufficient statistic

Conjugate priors

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

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

Combining with the likelihood function, we get

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

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

Outlines

- Binary Distributions
 - Multinomial Distributions
 - Gaussian Distributions
 - Exponential Families
 - Non-information 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 $\lambda = \eta^2$:

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

Non-informative Priors (2)

Translation invariant priors. Consider

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

For a corresponding prior over μ , we have

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

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 $\hat{x} = cx$

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

For a corresponding prior over σ , we have

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

for any A and B . Thus $p(\sigma) \propto 1/\sigma$ 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 $p(\sigma) \propto 1/\sigma$, then $p(\lambda) \propto 1/\lambda$.

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

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

- 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)

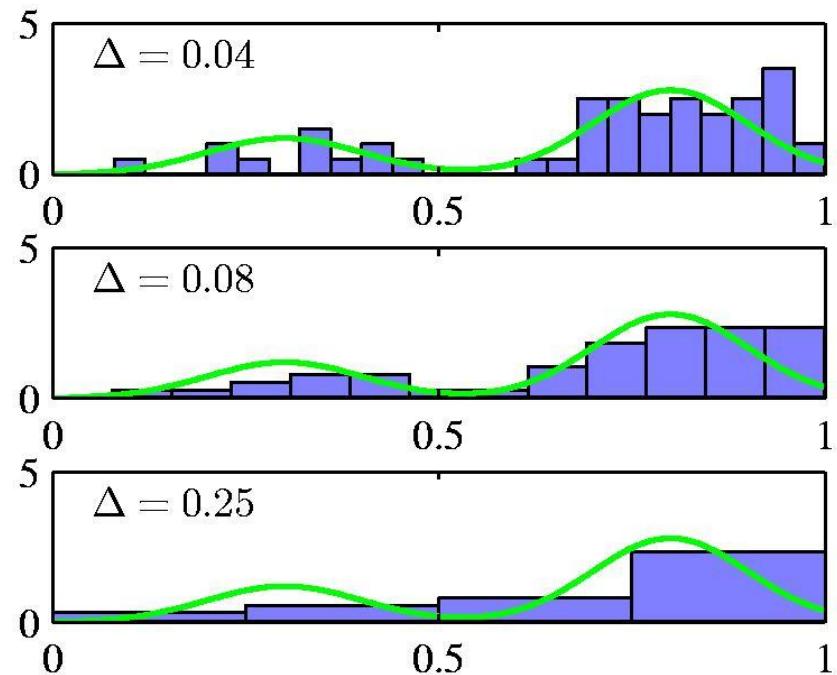
- Parametric distribution models are restricted to specific forms, which may not always be suitable; for example, consider modelling a multimodal distribution with a single, unimodal model.
- Non-parametric approaches make few assumptions about the overall shape of the distribution being modelled.

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

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

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

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

$$K \simeq NP.$$

Thus

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

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)

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

It follows that

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

Non-parametric Methods (5)

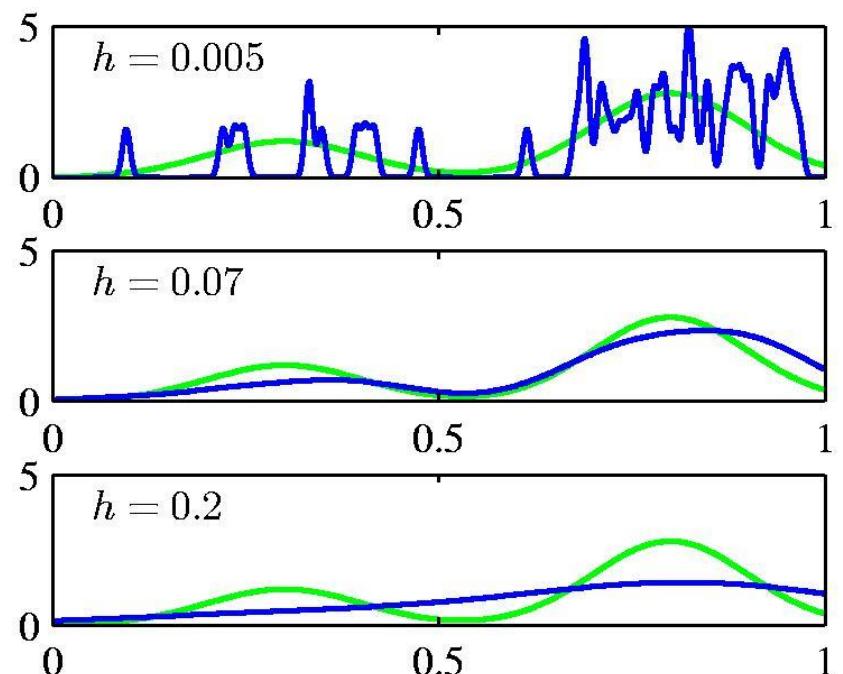
To avoid discontinuities in $p(x)$,
use a smooth kernel, e.g. a
Gaussian

$$p(\mathbf{x}) = \frac{1}{N} \sum_{n=1}^N \frac{1}{(2\pi h^2)^{D/2}} \exp \left\{ -\frac{\|\mathbf{x} - \mathbf{x}_n\|^2}{2h^2} \right\}$$

Any kernel such that

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

will work.



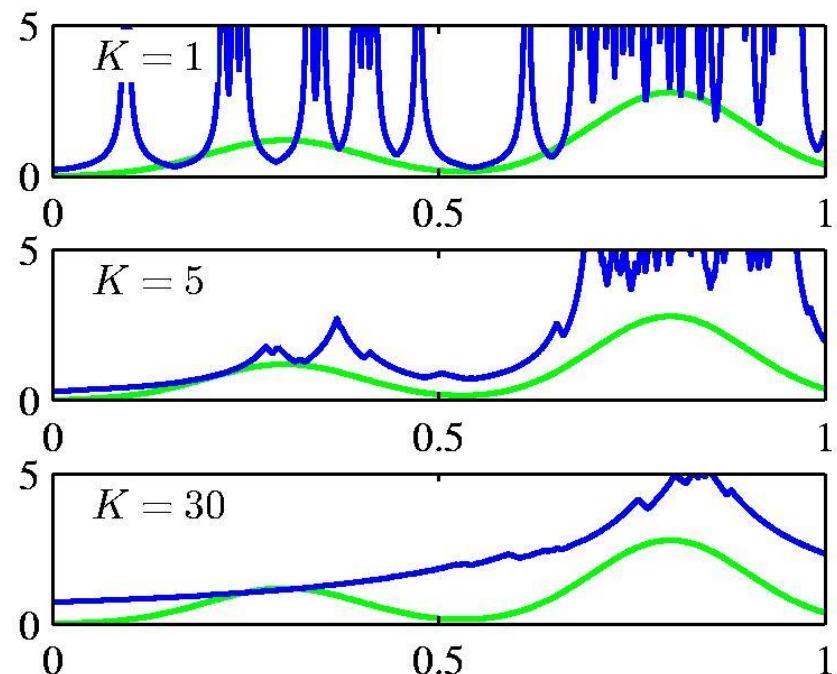
h acts as a smoother.

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

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



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-information Priors
 - Non-parametric Methods
 - KNN
-

K-Nearest-Neighbours for Classification (1)

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

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

number of total data $p(\mathbf{x})$ $\frac{K}{NV}$ number of data in a region
 volume of the region

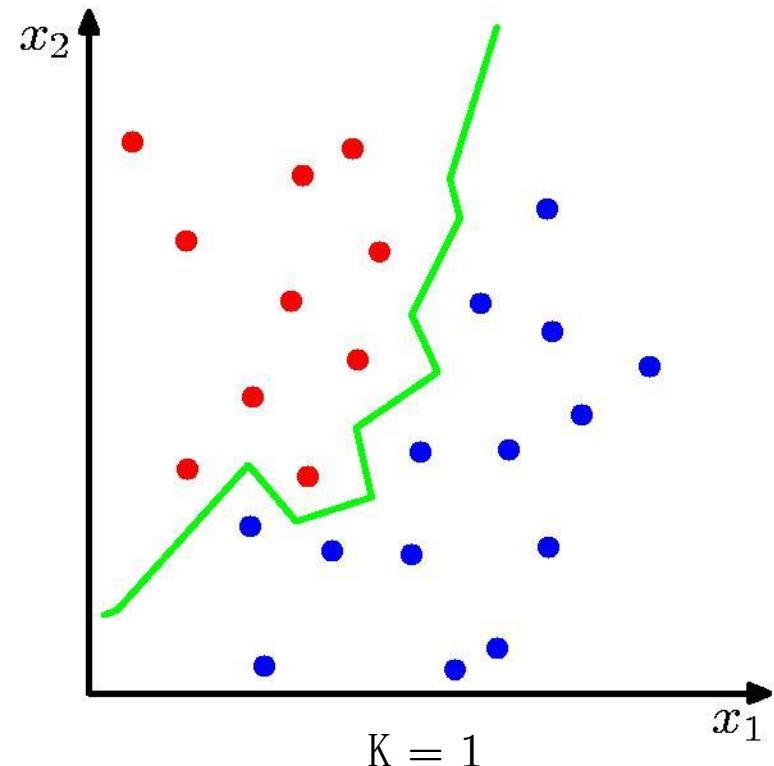
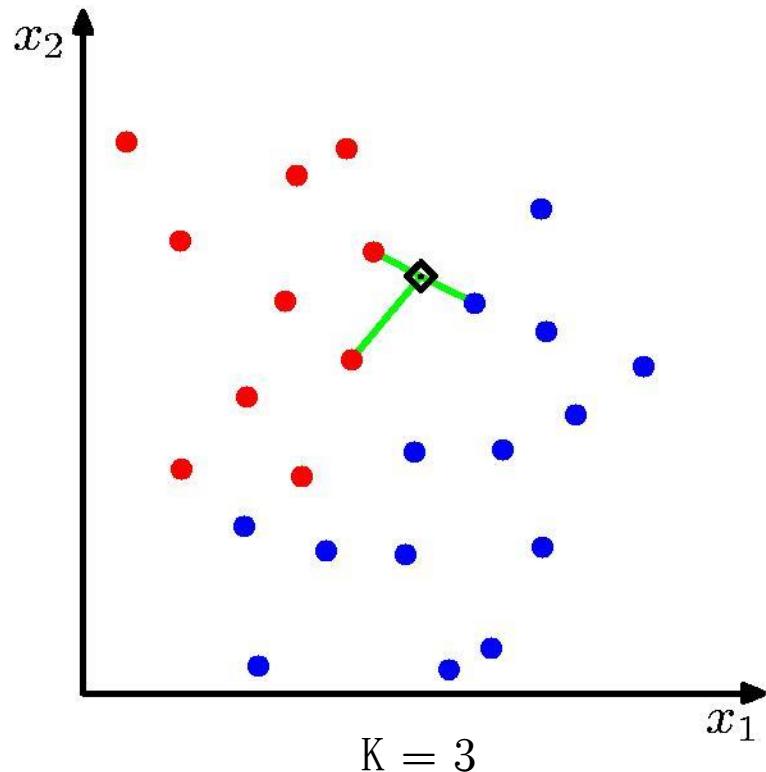
and correspondingly

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

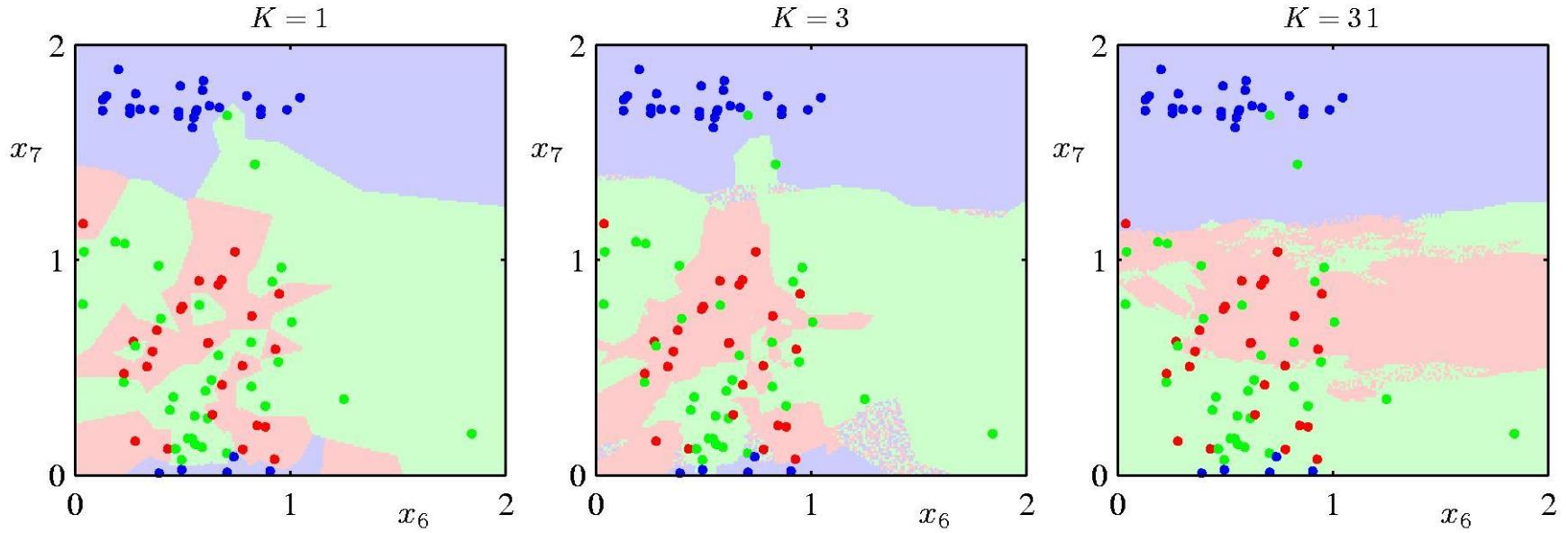
- Since $p(C_k) = N_k/N$, Bayes' theorem gives

$$p(C_k|\mathbf{x}) = \frac{p(\mathbf{x}|C_k)p(C_k)}{p(\mathbf{x})} = \frac{K_k}{K}.$$

K-Nearest-Neighbours for Classification (2)



K-Nearest-Neighbours for Classification (3)



- K acts as a smother
 - For $N \rightarrow \infty$, 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
-