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?

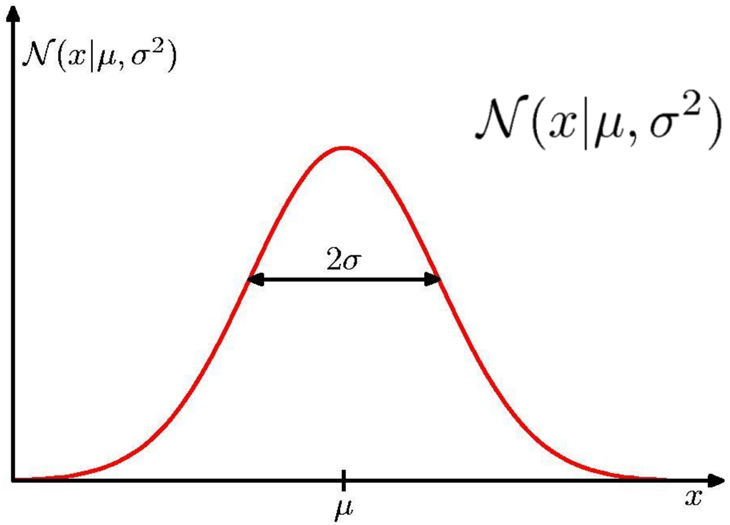
Outlines

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

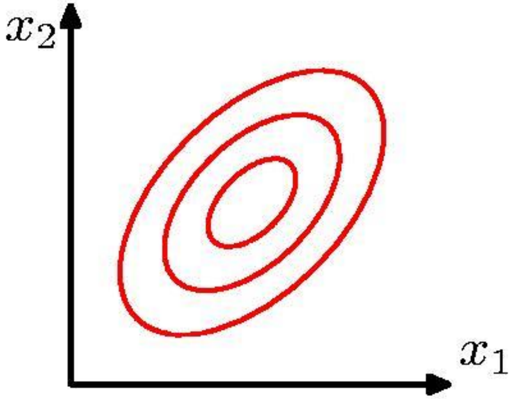
Outlines

- Gaussian Distributions and Central Limit Theorem
- Moments and Properties of Gaussians
- Posterior of the Linear Gaussian Variable
- Partitioned Gaussian Conditionals and Marginals
- ML Learning of Gaussian Distributions
- Sequential Learning of Gaussian Distributions
- Bayesian Learning of Gaussian Distributions
- Mixture of Gaussian Distributions

The Gaussian Distribution



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

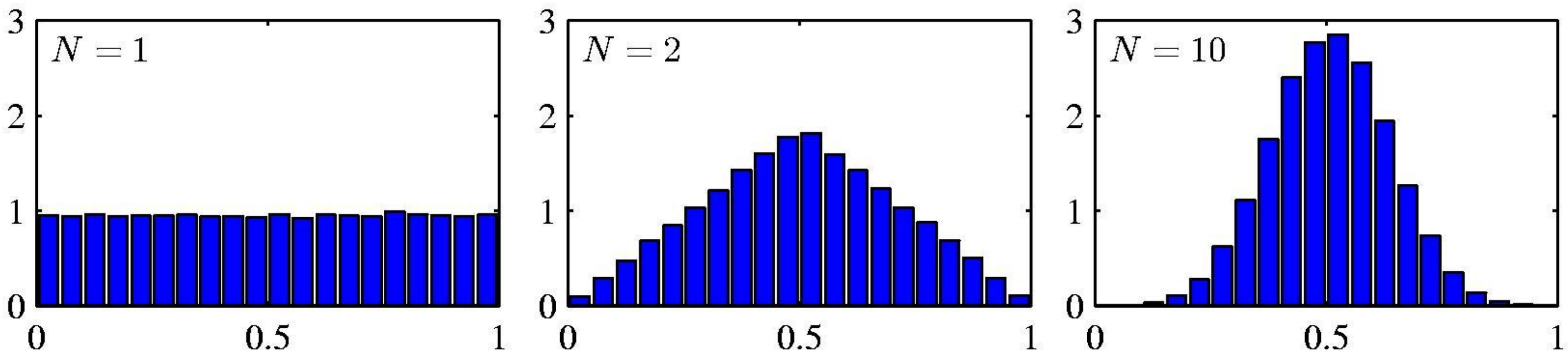


$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$

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.



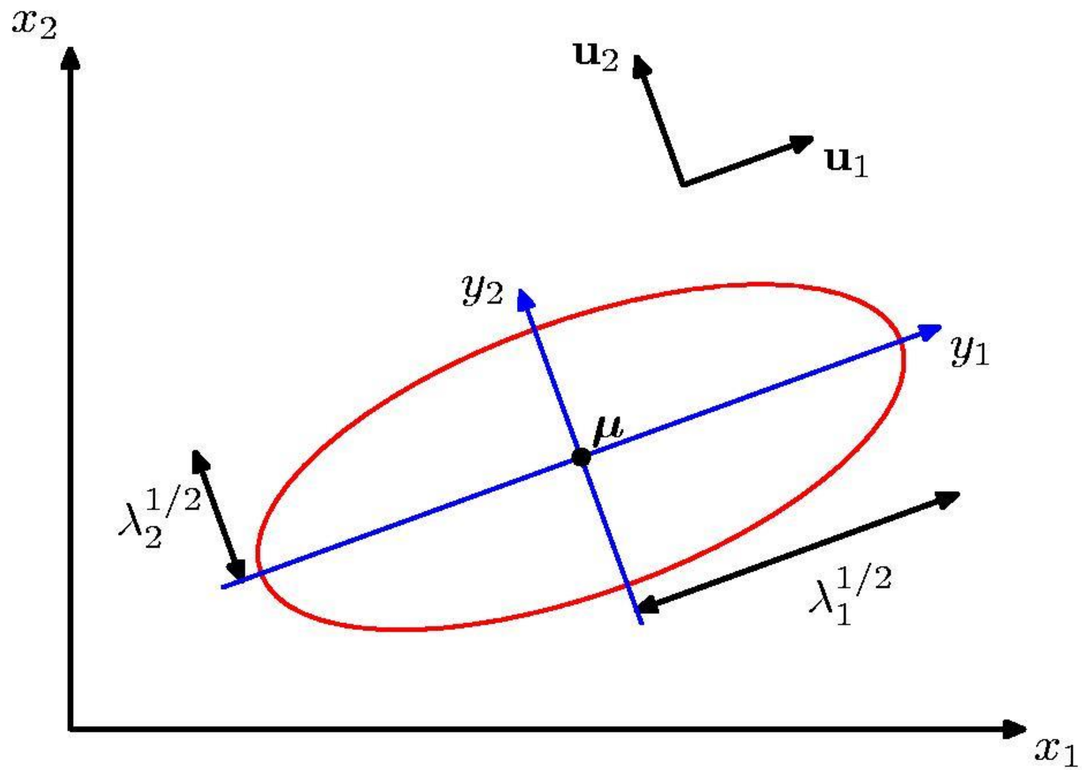
Geometry of the Multivariate Gaussian

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

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

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

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



Moments of the Multivariate Gaussian (1)

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

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

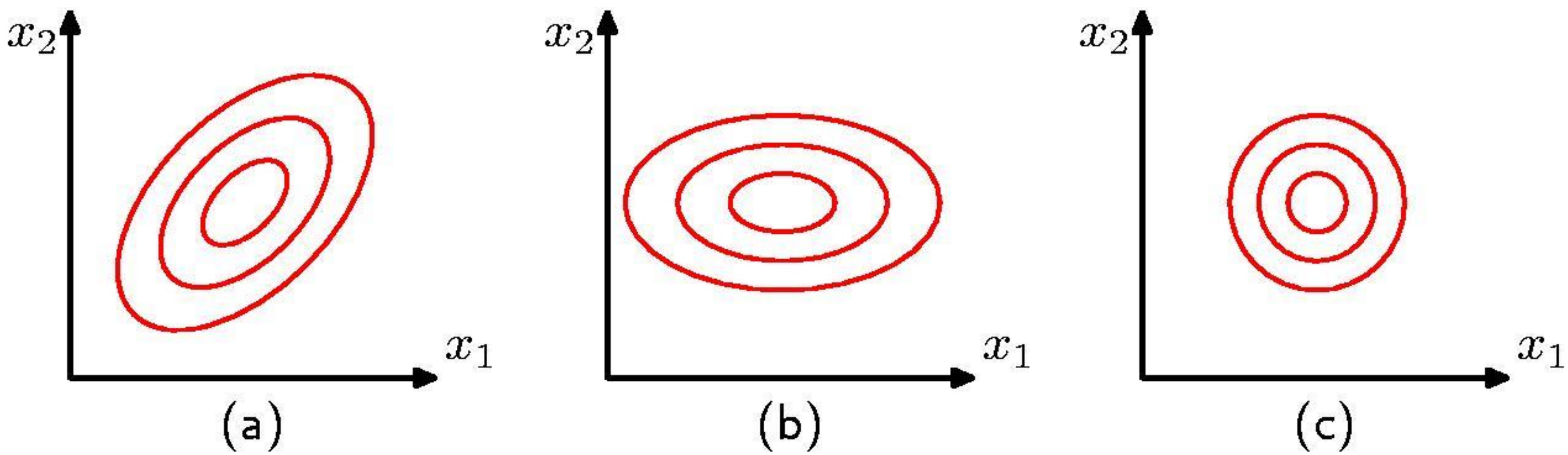
$$\mathbb{E}[\mathbf{x}] = \boldsymbol{\mu}$$

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

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

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

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

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

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

$$\Updownarrow$$

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

Quadratic terms of x (x^2) are equal

1st order terms of *x* are also equal

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

Properties of Multivariate Gaussians

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

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

(where division "–" denotes matrix inversion)

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

Properties of Multivariate Gaussians

$$p_X(x) \sim N(\mu,\ \Sigma) \quad \Longleftrightarrow \quad \begin{array}{l} p_{X_1}(x) \sim N(\mu_1,\ \Sigma_1) \\ p_{X_2}(x) \sim N(\mu_2,\ \Sigma_2) \end{array}$$

Precision

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

Mean

Bayes' Theorem for Gaussian Variables (1)

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 (2)

Given $$x = m + u$$

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

we have

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

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

Bayes' Theorem for Gaussian Variables (3)

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

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

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

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

Bayes' Theorem for Gaussian Variables (4)

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

where

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

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 (5)

Then

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

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

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

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

Bayes' Theorem for Gaussian Variables (6)*

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

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

Partitioned Gaussian Distributions (1)

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

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

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

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

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

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

Partitioned Gaussian Distributions (2)

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

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

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

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

Inverse Covariance Matrix*

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

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

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

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

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

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*

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

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

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

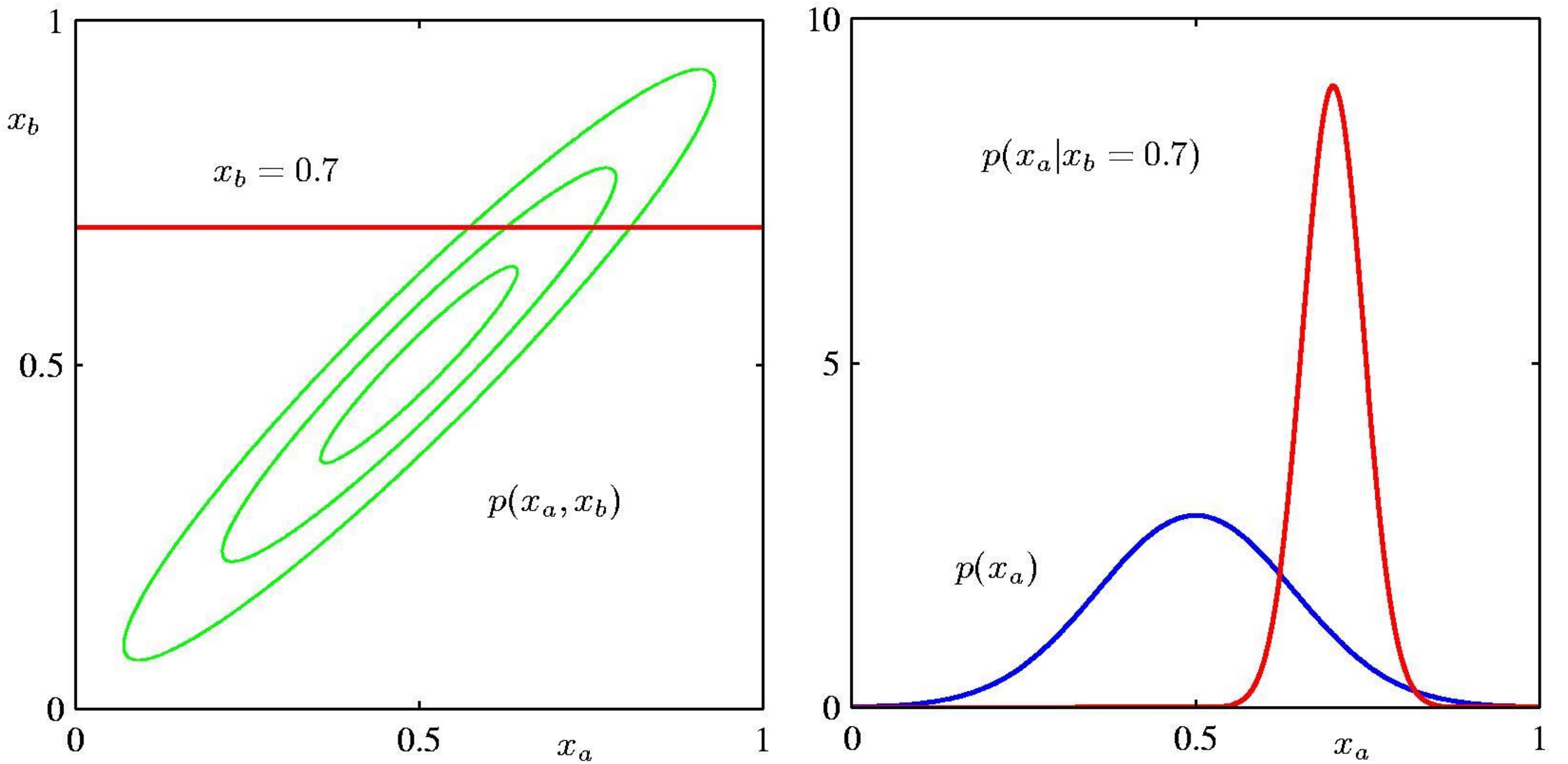
Partitioned Conditionals and Marginals*

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

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

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

Partitioned Conditionals and Marginals



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

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

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)

The trace is invariant under cyclic permutation of matrix product

$$\mathrm{tr}\{ABC\} = tr\{CAB\} = tr\{BCA\}$$

as follows

$$x^T\Sigma^{-1}x = tr\{x^T\Sigma^{-1}x\} = tr\{xx^T\Sigma^{-1}\}$$

Besides

$$\frac{\partial tr\{AB\}}{\partial A} = B^T$$

then

$$\frac{\partial tr\{xx^T\Sigma^{-1}\}}{\partial \Sigma^{-1}} = xx^T$$

Maximum Likelihood for the Gaussian (4)

Given $$\frac{\partial \ln |A|}{\partial A} = (A^{-1})^T = (A^T)^{-1}$$

then $$\frac{\partial \ln |\Sigma^{-1}|}{\partial \Sigma^{-1}} = \Sigma$$

As follows

$$\frac{\partial \ln p(X|\mu, \Sigma)}{\partial \Sigma^{-1}} = \frac{N}{2}\Sigma - \frac{1}{2}\sum_{n=1}^{\mathrm{N}} (x_n - \mu)(x_n - \mu)^T = 0$$

Maximum Likelihood for the Gaussian (5)

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, $\mathbf{x}_N$

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

correction given $\mathbf{x}_N$

correction weight

old estimate

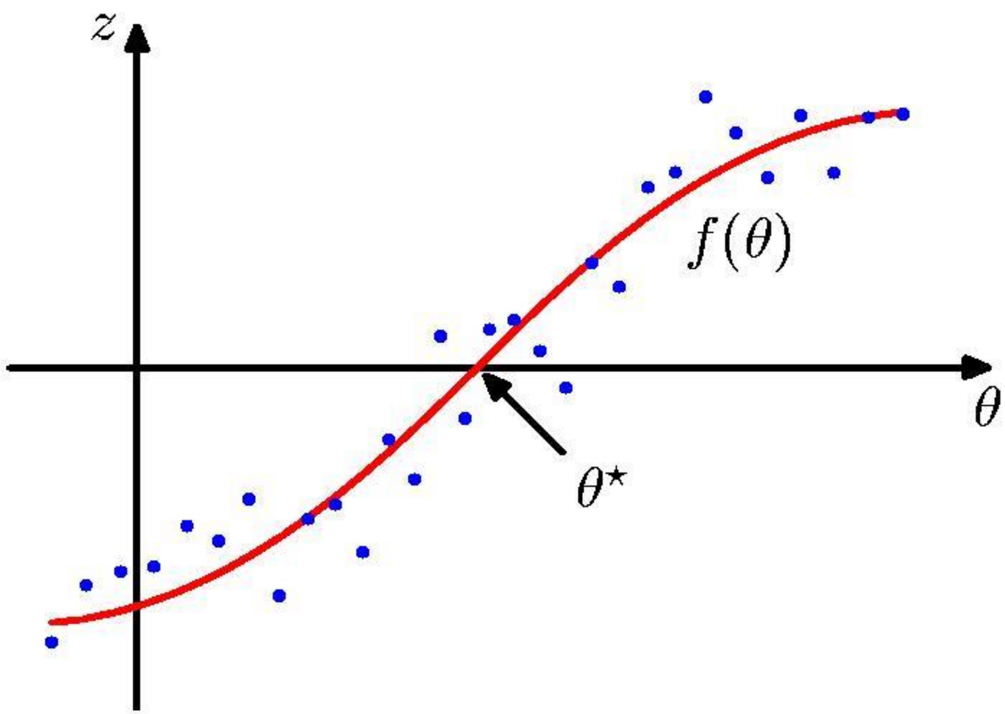
The Robbins-Monro Algorithm (1)*

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

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

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

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

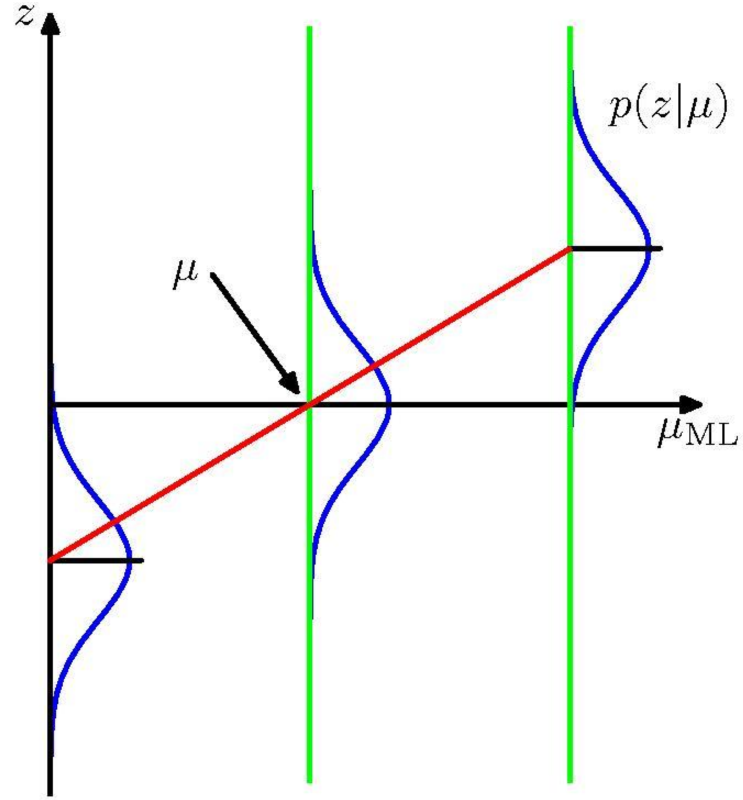
Robbins-Monro for Maximum Likelihood (2)*

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

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

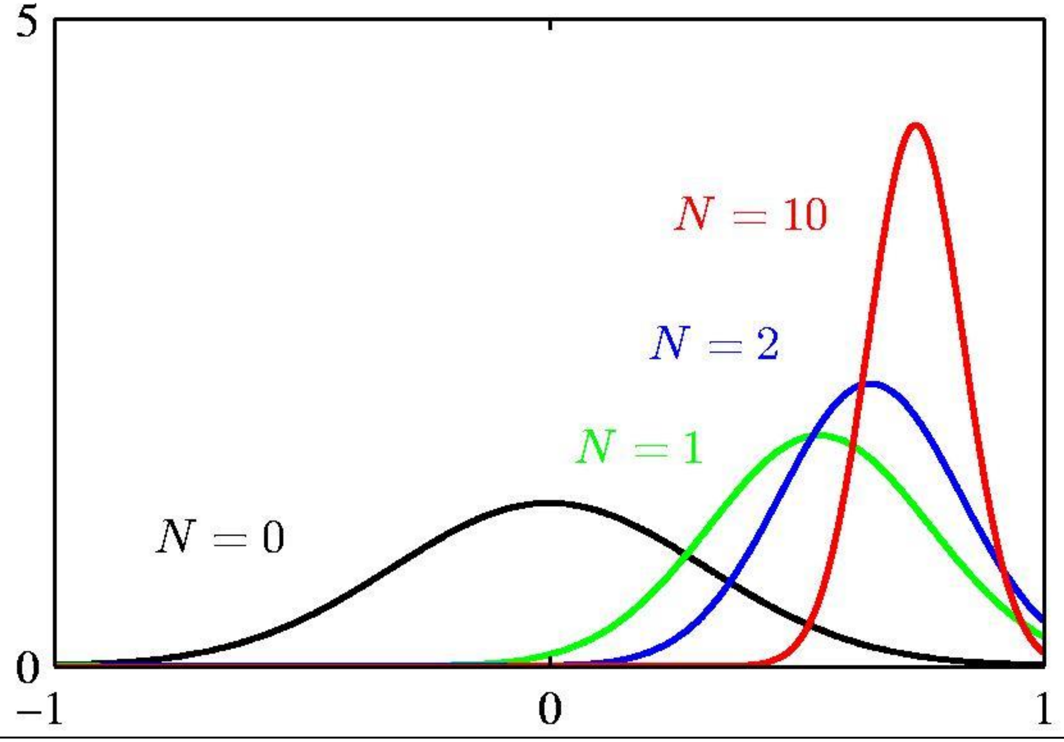
$$\frac{1}{\sigma_N^2} = \frac{1}{\sigma_0^2} + \frac{N}{\sigma^2}.$$

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

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

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

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

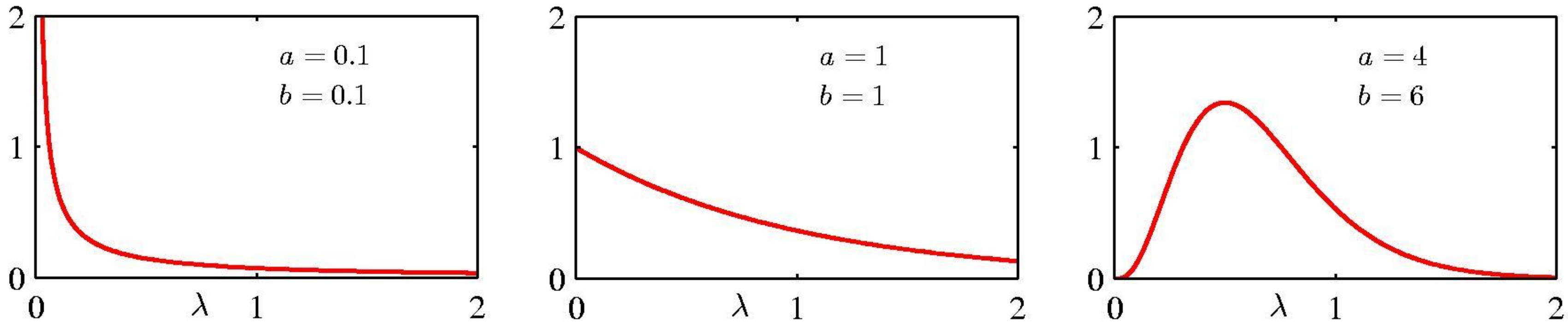
This has a Gamma shape as a function of λ.

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

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

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

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

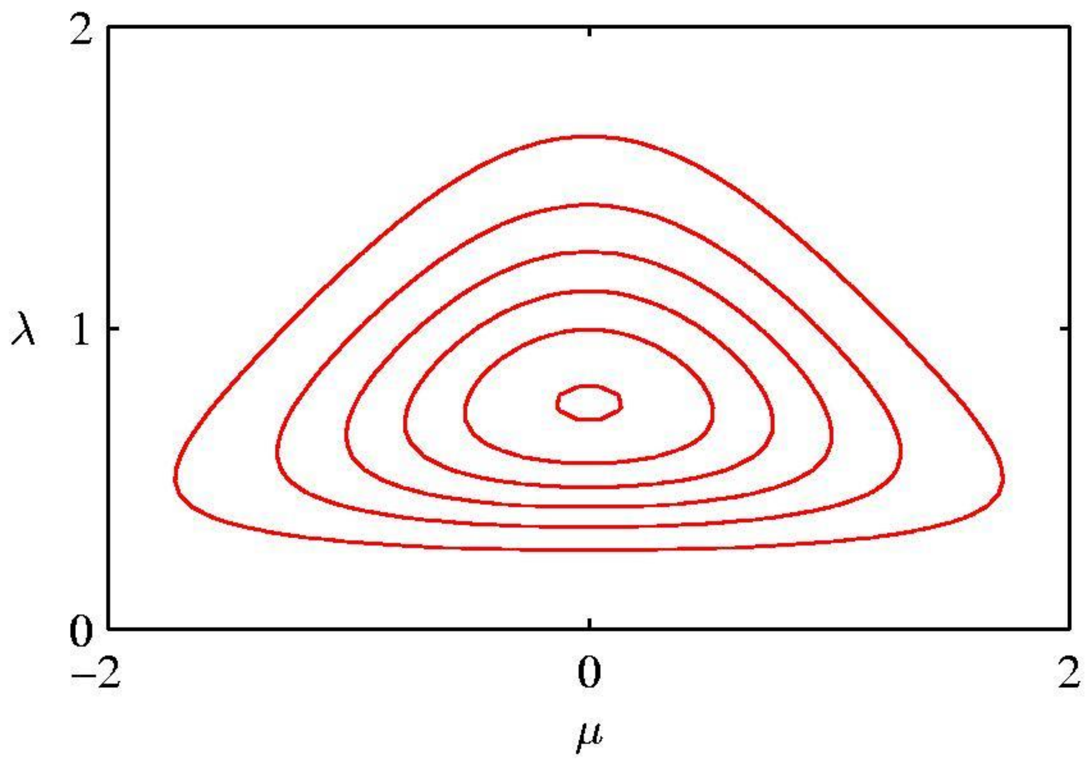
- Quadratic in μ.
- Linear in λ.

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

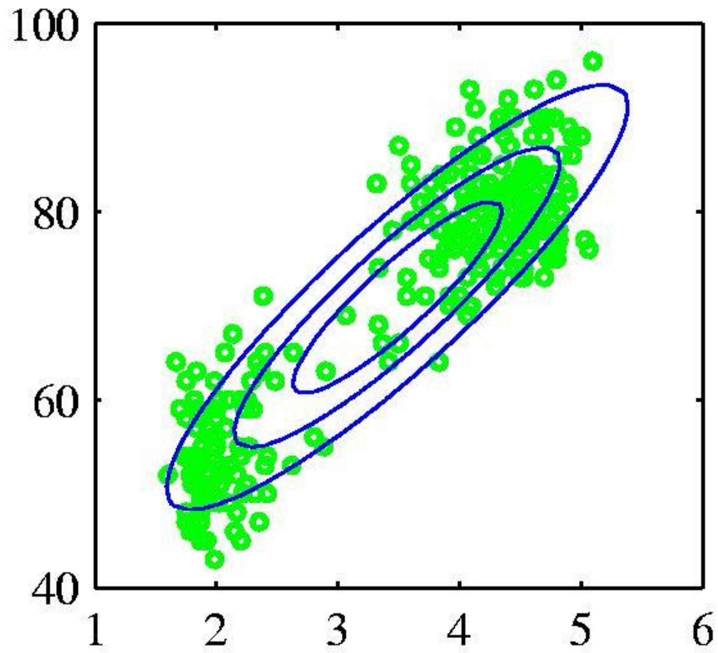
$$\mathcal{W}(\boldsymbol{\Lambda}|\mathbf{W},\nu) = B|\boldsymbol{\Lambda}|^{(\nu-D-1)/2}\exp\left(-\frac{1}{2}\mathrm{Tr}(\mathbf{W}^{-1}\boldsymbol{\Lambda})\right).$$

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

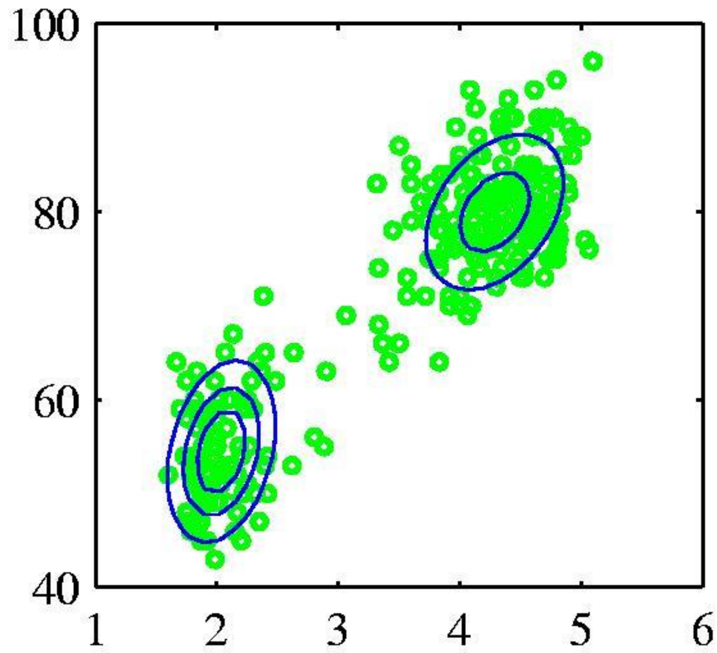
$$\mathcal{N}(\boldsymbol{\mu}|\boldsymbol{\mu}_0,(\beta\boldsymbol{\Lambda})^{-1})\,\mathcal{W}(\boldsymbol{\Lambda}|\mathbf{W},\nu)$$

Mixtures of Gaussians (1)

Old Faithful data set



Single Gaussian



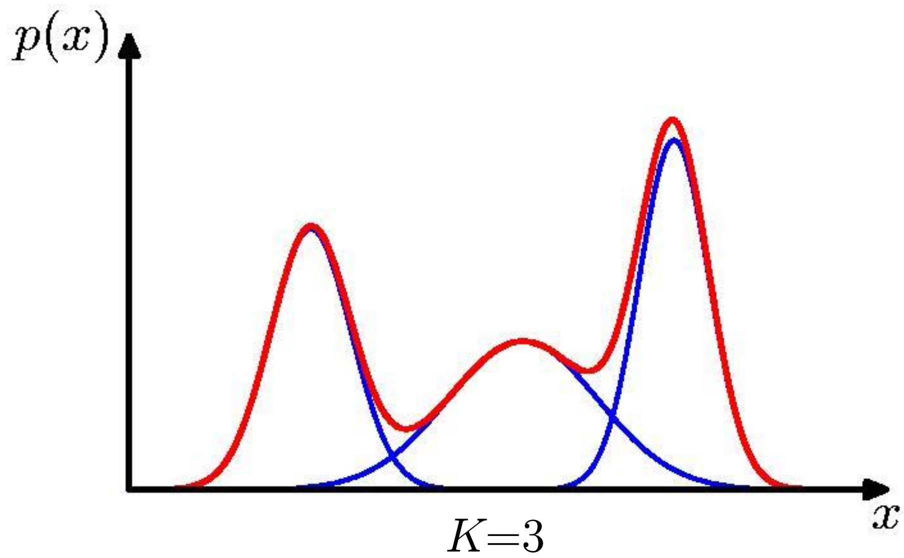
Mixture of two Gaussians

Mixtures of Gaussians (2)

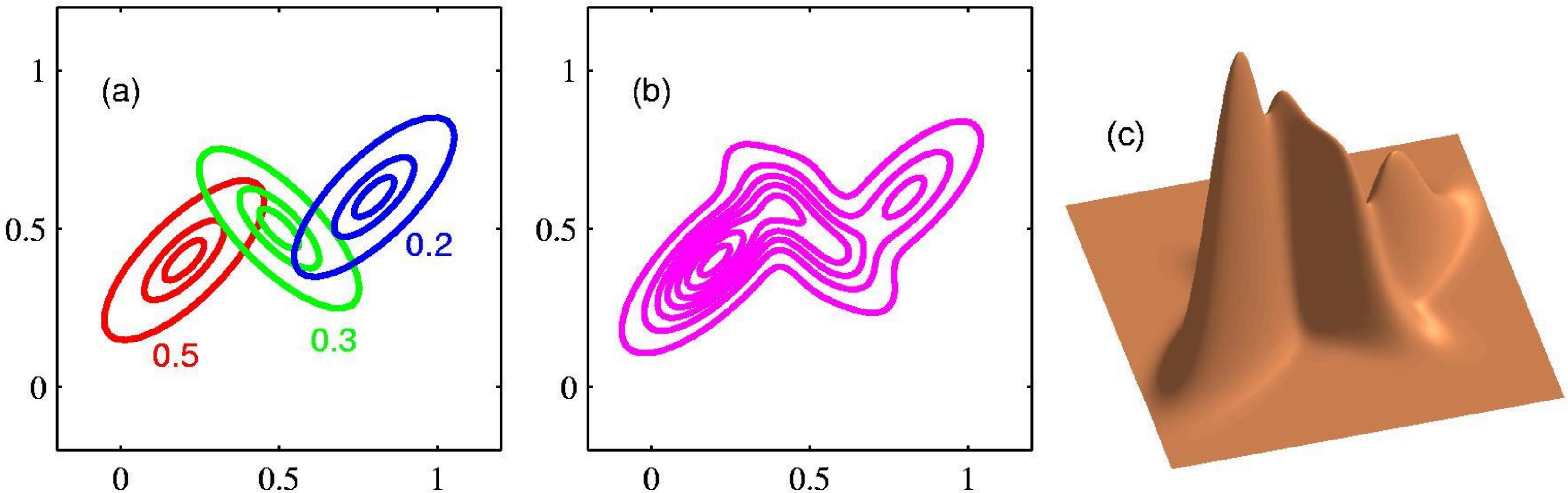
Combine simple models
into a complex model:

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

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



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

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

Student's t-Distribution*

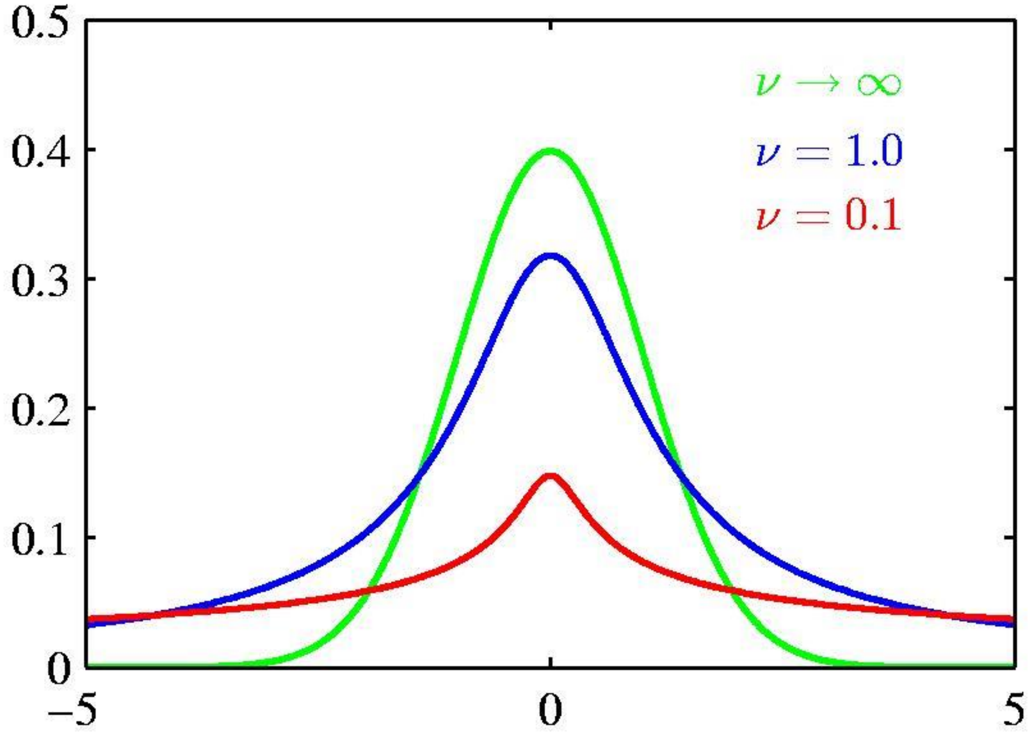
$$\begin{aligned}
p(x|\mu, a, b) &= \int_0^\infty \mathcal{N}(x|\mu, \tau^{-1})\mathrm{Gam}(\tau|a, b)\,\mathrm{d}\tau \\
&= \int_0^\infty \mathcal{N}\left(x|\mu, (\eta\lambda)^{-1}\right)\mathrm{Gam}(\eta|\nu/2, \nu/2)\,\mathrm{d}\eta \\
&= \frac{\Gamma(\nu/2 + 1/2)}{\Gamma(\nu/2)}\left(\frac{\lambda}{\pi\nu}\right)^{1/2}\left[1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

where

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

Infinite mixture of Gaussians.

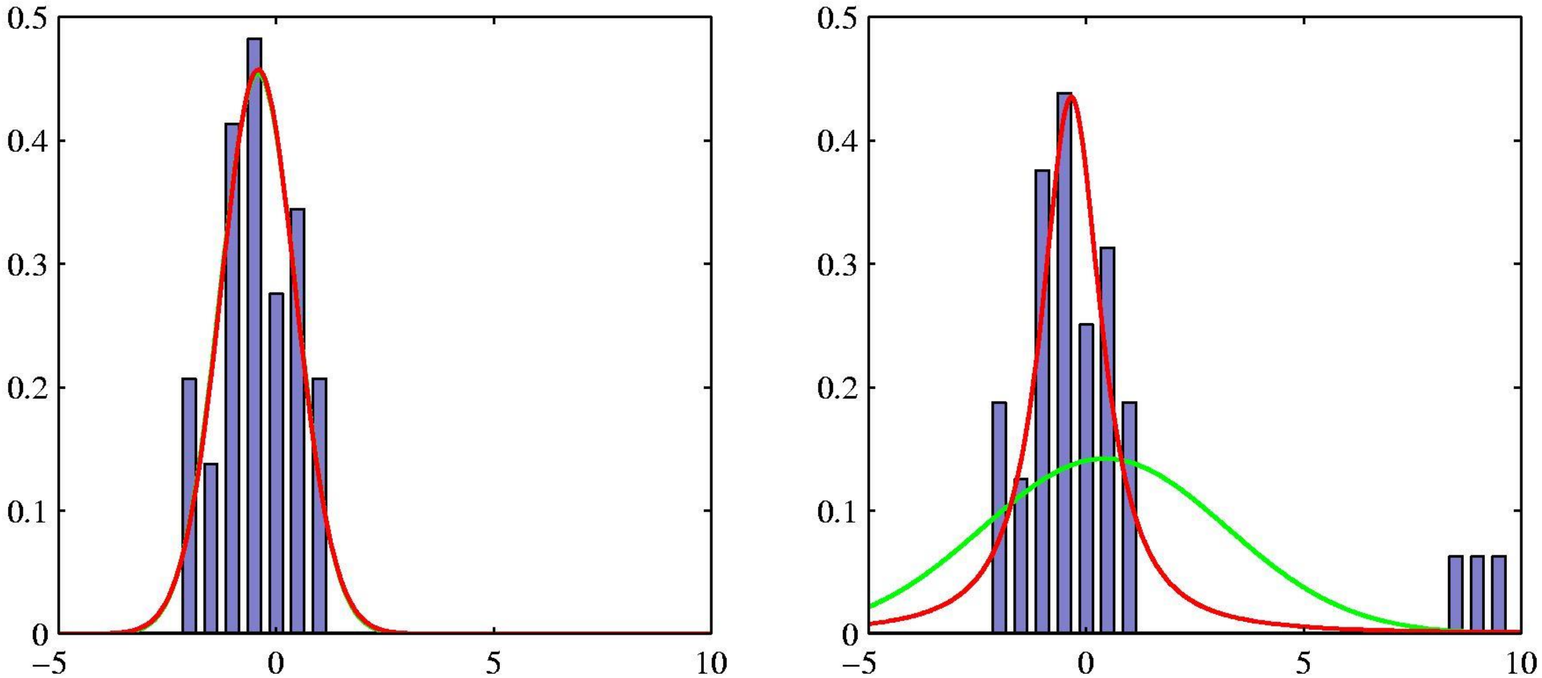
Student's t-Distribution*



	$\nu = 1$	$\nu \to \infty$
$\mathrm{St}(x\|\mu, \lambda, \nu)$	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

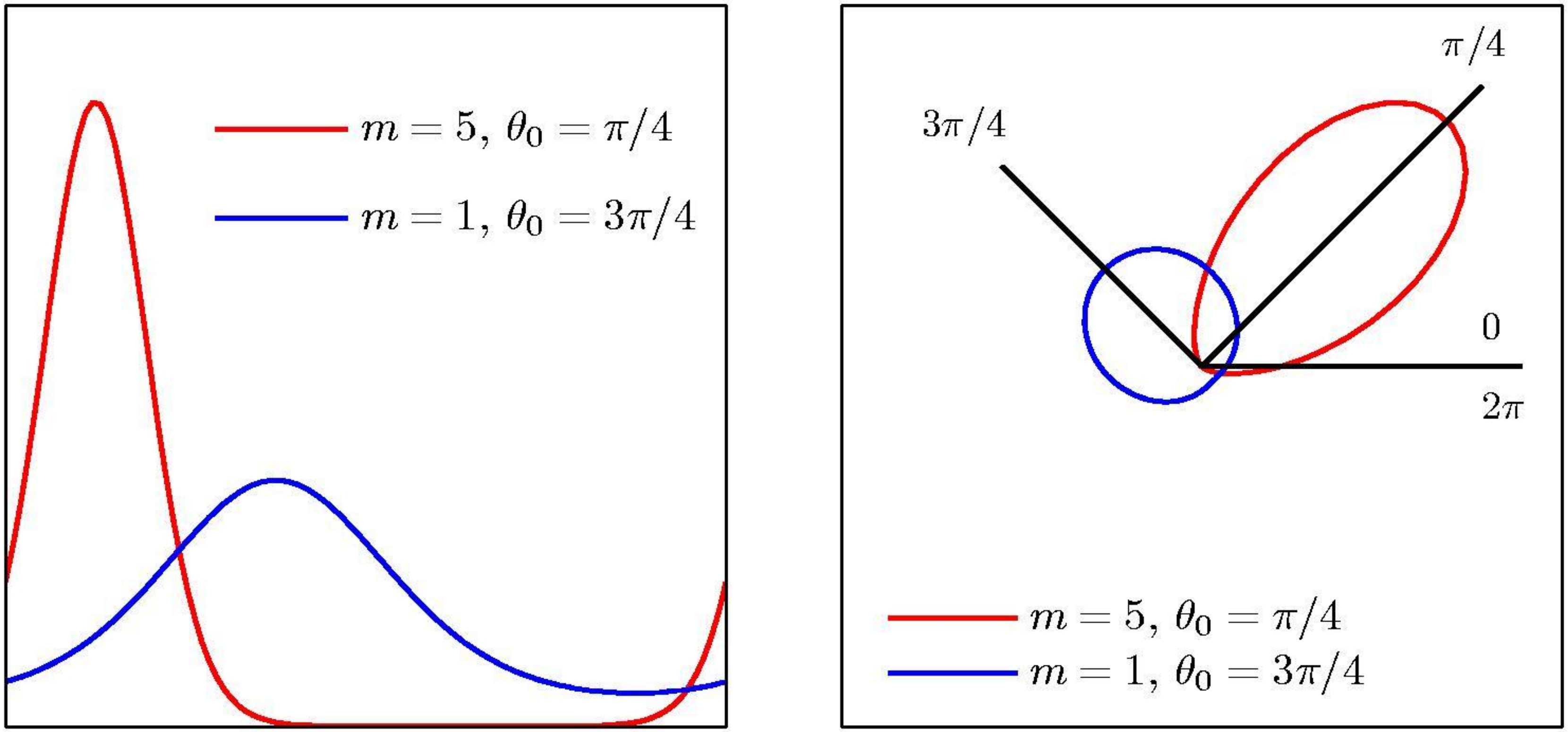
$$p(\theta|\theta_0, m) = \frac{1}{2\pi I_0(m)} \exp\left\{m\cos(\theta-\theta_0)\right\}$$

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}.