Artificial Intelligence

Probability Distributions

Donghui Wang
AI Institute@ZJU
2015.03

Contents

- The Gaussian Distribution
- Other distributions
- The Exponential Family
- Nonparametric Methods

References:

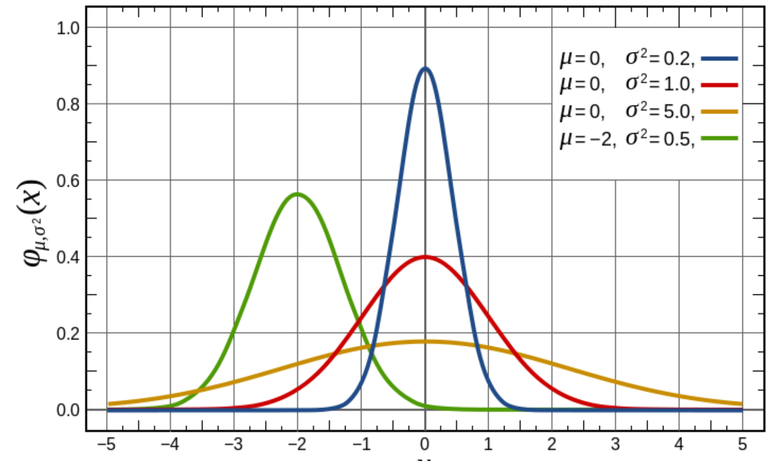
1. *Bishop. "Pattern Recognition and Machine Learning", Chapter 2. 2006.*
2. *Probability and Statistics Cookbook, http://matthias.vallentin.net/probability-and-statistics-cookbook/*
3. *http://cs229.stanford.edu/materials.html*

The Gaussian Distribution

The Gaussian Distribution

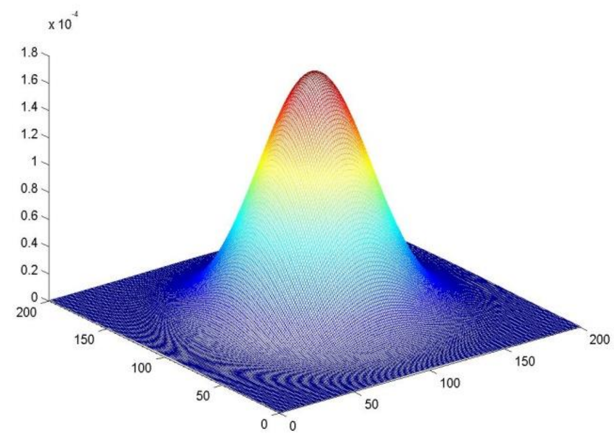
- Single variable Gaussian

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



- Multivariate Gaussian

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



The Gaussian Distribution

- Central limit theorem:

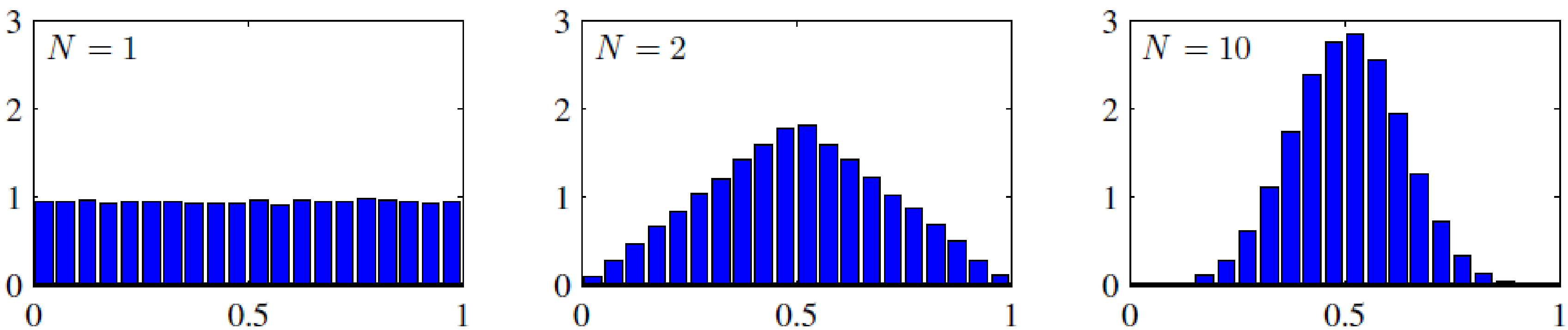


Figure 2.6 Histogram plots of the mean of N uniformly distributed numbers for various values of N. We observe that as N increases, the distribution tends towards a Gaussian.

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

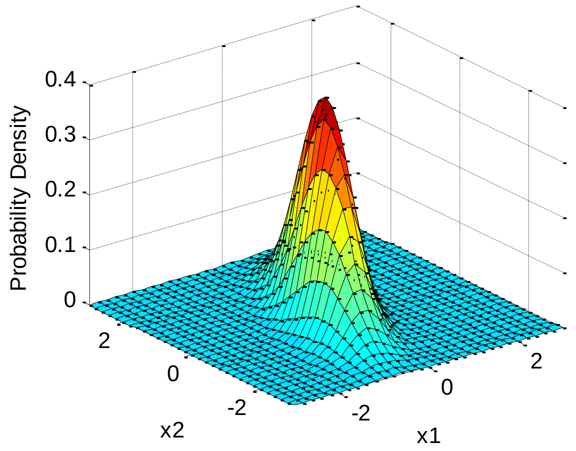
$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$

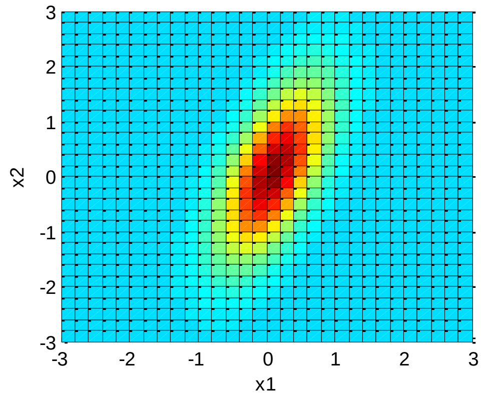
Multivariate Gaussian Distribution

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

- Mahalanobis distance Δ → Euclidean distance

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






```
mu = [0 0];
Sigma = [.25 .3; .3 1];
%Sigma = [.25 0; 0 1];
%Sigma = [0.5 0; 0 0.5];
x1 = -3:.1:3;
x2 = -3:.1:3;
[X1,X2] = meshgrid(x1,x2);
F = mvnpdf([X1(:) X2(:)],mu,Sigma);

F = reshape(F,length(x2),length(x1));
surf(x1,x2,F);
caxis([min(F(:))-.5*range(F(:)),max(F(:))]);
axis([-3 3 -3 3 0 .4])
xlabel('x1'); ylabel('x2');
zlabel('Probability Density');
```

Multivariate Gaussian Distribution

$$\Delta^2 = (\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) = (x-\mu)^T U^T \Lambda^{-1} U(x-\mu) = \left(U(x-\mu)\right)^T \Lambda^{-1}\left(U(x-\mu)\right) = y^T\Lambda^{-1}y$$

The matrix Σ can be taken to be symmetric, without loss of generality.

M is symmetric, so that $\mathbf{M}^{\mathrm{T}} = \mathbf{M}$. $\mathbf{M}\mathbf{M}^{-1} = \mathbf{I}$

$$\left(\mathbf{M}^{-1}\right)^{\mathrm{T}}\mathbf{M}^{\mathrm{T}} = \mathbf{I}^{\mathrm{T}} = \mathbf{I} \Rightarrow \left(\mathbf{M}^{-1}\right)^{\mathrm{T}}\mathbf{M} = \mathbf{I} \Rightarrow \left(\mathbf{M}^{-1}\right)^{\mathrm{T}} = \mathbf{M}^{-1}$$

so $\mathbf{M}^{-1}$ is also a symmetric matrix.

the eigenvector equation for the covariance matrix

$$\boldsymbol{\Sigma}\mathbf{u}_i = \lambda_i \mathbf{u}_i \quad \text{where } i = 1, \ldots, D \qquad \mathbf{u}_i^{\mathrm{T}}\mathbf{u}_j = I_{ij} \qquad I_{ij} = \begin{cases} 1, & \text{if } i = j \\ 0, & \text{otherwise.} \end{cases} \qquad \mathbf{U}\mathbf{U}^{\mathrm{T}} = \mathbf{I}$$

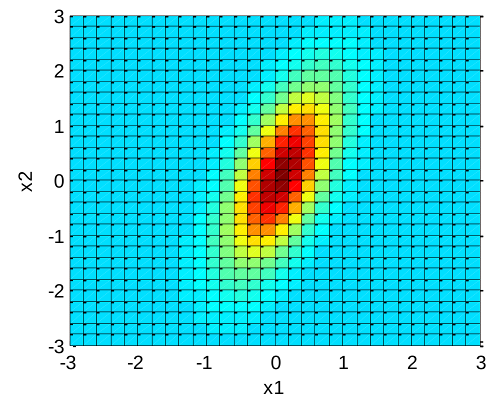
$$\boldsymbol{\Sigma} = \sum_{i=1}^{D} \lambda_i \mathbf{u}_i \mathbf{u}_i^{\mathrm{T}} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^{\mathrm{T}} \Rightarrow \mathbf{U}^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{U} = \mathbf{U}^{\mathrm{T}}\mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^{\mathrm{T}}\mathbf{U} = \boldsymbol{\Lambda} \qquad \text{U is orthonormal, } \mathbf{U}^{-1} = \mathbf{U}^{\mathrm{T}}$$

$$\boldsymbol{\Sigma}^{-1} = \left(\mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^{\mathrm{T}}\right)^{-1} = \left(\mathbf{U}^{\mathrm{T}}\right)^{-1}\boldsymbol{\Lambda}^{-1}\mathbf{U}^{-1} = \mathbf{U}\boldsymbol{\Lambda}^{-1}\mathbf{U}^{\mathrm{T}} = \sum_{i=1}^{D}\frac{1}{\lambda_i}\mathbf{u}_i\mathbf{u}_i^{\mathrm{T}}.$$

$$\Delta^2 = (\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) \xrightarrow{y_i = \mathbf{u}_i^{\mathrm{T}}(\mathbf{x}-\boldsymbol{\mu})} \Delta^2 = \sum_{i=1}^{D}\frac{y_i^2}{\lambda_i} \xrightarrow{\mathbf{y} = \mathbf{U}(\mathbf{x}-\boldsymbol{\mu})} \Delta^2 = \mathbf{y}^{\mathrm{T}}\boldsymbol{\Lambda}^{-1}\mathbf{y}$$

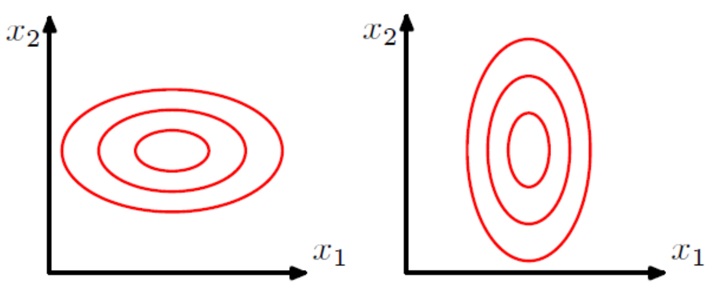
Multivariate Gaussian Distribution

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

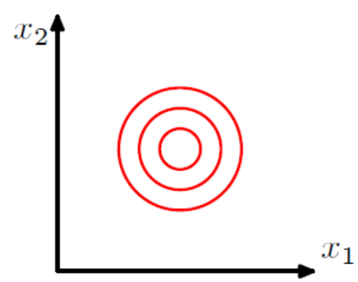


$$\Delta^2 = (\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) = \mathbf{y}^{\mathrm{T}}\boldsymbol{\Lambda}^{-1}\mathbf{y} \qquad \boldsymbol{\Sigma}^{-1} = \mathbf{U}\boldsymbol{\Lambda}^{-1}\mathbf{U}^{\mathrm{T}}$$

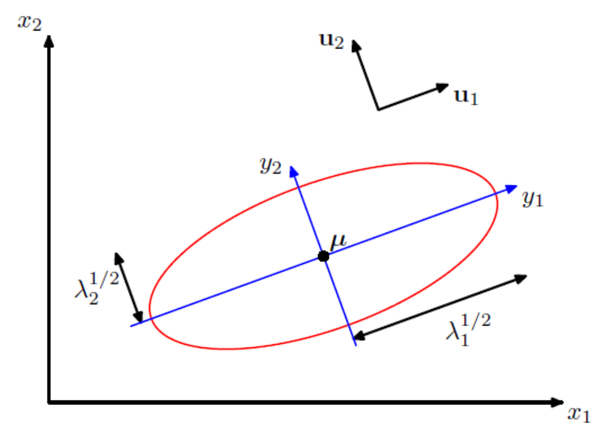
$$\Rightarrow \Delta^2 = \sum_{i=1}^{D}\frac{y_i^2}{\lambda_i} \qquad y_i = \mathbf{u}_i^{\mathrm{T}}(\mathbf{x}-\boldsymbol{\mu}) \qquad \mathbf{y} = \mathbf{U}(\mathbf{x}-\boldsymbol{\mu})$$



$$\boldsymbol{\Sigma} = \mathrm{diag}(\sigma_i^2)$$



$$\boldsymbol{\Sigma} = \sigma^2\mathbf{I}$$



Jacobian factor or matrix

Under a nonlinear change of variable, a probability density transforms differently from a simple function, due to the Jacobian factor. For instance, if we consider a change of variables $x = g(y)$, then a function $f(x)$ becomes $\widetilde{f}(y) = f(g(y))$. Now consider a probability density $p_x(x)$ that corresponds to a density $p_y(y)$ with respect to the new variable y, where the suffices denote the fact that $p_x(x)$ and $p_y(y)$ are different densities. Observations falling in the range $(x, x + \delta x)$ will, for small values of δx, be transformed into the range $(y, y + \delta y)$ where $p_x(x)\delta x \simeq p_y(y)\delta y$, and hence

$$p_y(y) = p_x(x)\left|\frac{\mathrm{d}x}{\mathrm{d}y}\right| = p_x(g(y))\,|g'(y)|\,.$$

$$\Delta^2 = \sum_{i=1}^{D} \frac{y_i^2}{\lambda_i} \qquad y_i = \mathbf{u}_i^{\mathrm{T}}(\mathbf{x}-\boldsymbol{\mu}) \qquad \mathbf{y} = \mathbf{U}(\mathbf{x}-\boldsymbol{\mu}) \Rightarrow \mathbf{x} = \mathbf{U}^{\mathrm{T}}\mathbf{y} + \boldsymbol{\mu} \Rightarrow J_{ij} = \frac{\partial x_i}{\partial y_j} = U_{ji}$$

$$\Rightarrow \mathbf{J} = \mathbf{U}^{\mathrm{T}} \Rightarrow |\mathbf{J}|^2 = \left|\mathbf{U}^{\mathrm{T}}\right|^2 = \left|\mathbf{U}^{\mathrm{T}}\right||\mathbf{U}| = \left|\mathbf{U}^{\mathrm{T}}\mathbf{U}\right| = |\mathbf{I}| = 1 \Rightarrow |\mathbf{J}| = 1$$

$$|\boldsymbol{\Sigma}|^{1/2} = \prod_{j=1}^{D} \lambda_j^{1/2} \longrightarrow p(\mathbf{y}) = p(\mathbf{x})|\mathbf{J}| = \prod_{j=1}^{D} \frac{1}{(2\pi\lambda_j)^{1/2}} \exp\left\{-\frac{y_j^2}{2\lambda_j}\right\}$$

Multivariate Gaussian Distribution

- It's normalized! $$\int_{-\infty}^{\infty} \mathcal{N}\left(x|\mu,\sigma^2\right)\,\mathrm{d}x = 1$$

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

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

$$|\mathbf{J}| = 1$$

$$\boldsymbol{\Sigma} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^{\mathrm{T}} \longrightarrow |\boldsymbol{\Sigma}| = |\mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^{\mathrm{T}}| = |\mathbf{U}||\boldsymbol{\Lambda}||\mathbf{U}^{\mathrm{T}}| = |\mathbf{U}||\mathbf{U}^{\mathrm{T}}||\boldsymbol{\Lambda}| = |\boldsymbol{\Lambda}| \longrightarrow |\boldsymbol{\Sigma}|^{1/2} = \prod_{j=1}^{D}\lambda_j^{1/2}$$

$$p(\mathbf{y}) = p(\mathbf{x})|\mathbf{J}| = \prod_{j=1}^{D}\frac{1}{(2\pi\lambda_j)^{1/2}}\exp\left\{-\frac{y_j^2}{2\lambda_j}\right\}$$

$$\longrightarrow \int p(\mathbf{y})\,\mathrm{d}\mathbf{y} = \prod_{j=1}^{D}\int_{-\infty}^{\infty}\frac{1}{(2\pi\lambda_j)^{1/2}}\exp\left\{-\frac{y_j^2}{2\lambda_j}\right\}\mathrm{d}y_j = 1 \longrightarrow \int p(\mathbf{y})\,\mathrm{d}\mathbf{y} = 1$$

Multivariate Gaussian Distribution

- Expectation of a random vector x:

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

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

Multivariate Gaussian Distribution

- The second order moments of the Gaussian

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

$$\underline{\underline{\mathbf{z} = \mathbf{x}-\boldsymbol{\mu}}} \quad \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \int \exp\left\{-\frac{1}{2}\mathbf{z}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{z}\right\} (\mathbf{z}+\boldsymbol{\mu})(\mathbf{z}+\boldsymbol{\mu})^{\mathrm{T}}\,\mathrm{d}\mathbf{z}$$

$$\boxed{\begin{gathered}\mathbf{z} = \sum_{j=1}^{D} y_j \mathbf{u}_j \\ \text{where } y_j = \mathbf{u}_j^{\mathrm{T}}\mathbf{z}. \\ \Delta^2 = \sum_{i=1}^{D} \frac{y_i^2}{\lambda_i}\end{gathered}}$$

$\boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}}$ is constant, $\boldsymbol{\mu}\mathbf{z}^{\mathrm{T}}$ and $\boldsymbol{\mu}^{\mathrm{T}}\mathbf{z}$ will again vanish by symmetry.

Consider the term involving $\mathbf{z}\mathbf{z}^{\mathrm{T}}$

$$\frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \int \exp\left\{-\frac{1}{2}\mathbf{z}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{z}\right\} \mathbf{z}\mathbf{z}^{\mathrm{T}}\,\mathrm{d}\mathbf{z}$$

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

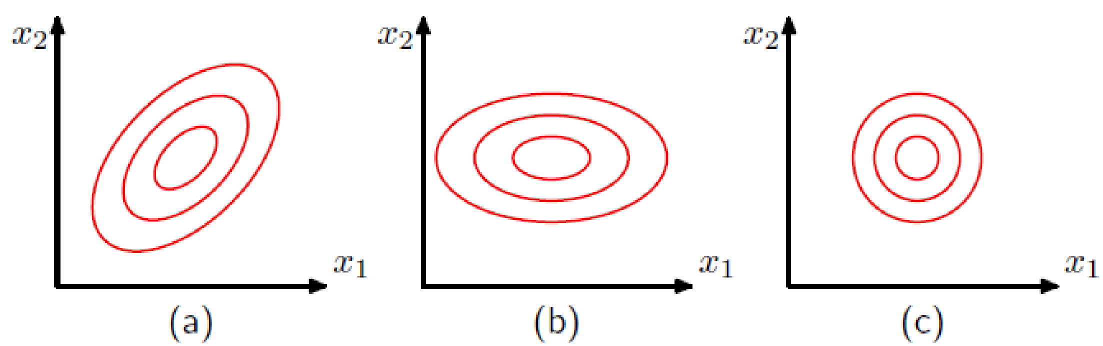
$$\boxed{\mathbb{E}[\mathbf{x}\mathbf{x}^{\mathrm{T}}] = \boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}} + \boldsymbol{\Sigma}}$$

Multivariate Gaussian Distribution

- The covariance of a random vector x:

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

A general symmetric covariance matrix Σ will have D(D + 1)/2 independent parameters, and there are another D independent parameters in μ, giving D(D + 3)/2 parameters in total.



$\boldsymbol{\Sigma} = \mathrm{diag}(\sigma_i^2)$ 2D independent parameters

$\boldsymbol{\Sigma} = \sigma^2\mathbf{I}$ isotropic covariance, D + 1 independent parameters

Conditional Gaussian Distributions

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

- If two sets of variables are jointly Gaussian, then the conditional distribution of one set conditioned on the other is again Gaussian.

$$\mathbf{x} = \begin{pmatrix}\mathbf{x}_a\\ \mathbf{x}_b\end{pmatrix} \quad \boldsymbol{\mu} = \begin{pmatrix}\boldsymbol{\mu}_a\\ \boldsymbol{\mu}_b\end{pmatrix} \quad \boldsymbol{\Sigma} = \begin{pmatrix}\boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab}\\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb}\end{pmatrix} \xleftrightarrow{\boldsymbol{\Lambda} \equiv \boldsymbol{\Sigma}^{-1}} \boldsymbol{\Lambda} = \begin{pmatrix}\boldsymbol{\Lambda}_{aa} & \boldsymbol{\Lambda}_{ab}\\ \boldsymbol{\Lambda}_{ba} & \boldsymbol{\Lambda}_{bb}\end{pmatrix}$$

covariance matrix

precision matrix

Both of Σ and Λ can be taken to be symmetric, without loss of generality.

$p(\mathbf{x}_a|\mathbf{x}_b) \Rightarrow$

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

completing the square

Completing the square:

$$\mu_{a|b} \quad \Sigma_{a|b} \quad p(\mathbf{x}_a|\mathbf{x}_b) \Rightarrow -\frac{1}{2}\mathbf{x}^{\mathrm{T}}\Sigma^{-1}\mathbf{x} + \mathbf{x}^{\mathrm{T}}\Sigma^{-1}\mu + \text{const}$$

$$-\frac{1}{2}(\mathbf{x}_a-\mu_a)^{\mathrm{T}}\Lambda_{aa}(\mathbf{x}_a-\mu_a) - \frac{1}{2}(\mathbf{x}_a-\mu_a)^{\mathrm{T}}\Lambda_{ab}(\mathbf{x}_b-\mu_b) - \frac{1}{2}(\mathbf{x}_b-\mu_b)^{\mathrm{T}}\Lambda_{ba}(\mathbf{x}_a-\mu_a) - \frac{1}{2}(\mathbf{x}_b-\mu_b)^{\mathrm{T}}\Lambda_{bb}(\mathbf{x}_b-\mu_b)$$

$$-\frac{1}{2}\mathbf{x}_a^{\mathrm{T}}\Lambda_{aa}\mathbf{x}_a \Rightarrow \boxed{\Sigma_{a|b} = \Lambda_{aa}^{-1}}$$

$$\mathbf{x}_a^{\mathrm{T}}\{\Lambda_{aa}\mu_a - \Lambda_{ab}(\mathbf{x}_b-\mu_b)\} = \mathbf{x}_a^{\mathrm{T}}\Sigma_{a|b}^{-1}\mu_{a|b} \Rightarrow \Sigma_{a|b}^{-1}\mu_{a|b} = \Lambda_{aa}\mu_a - \Lambda_{ab}(\mathbf{x}_b-\mu_b)$$

$$\Rightarrow \mu_{a|b} = \Sigma_{a|b}\{\Lambda_{aa}\mu_a - \Lambda_{ab}(\mathbf{x}_b-\mu_b)\} = \boxed{\mu_a - \Lambda_{aa}^{-1}\Lambda_{ab}(\mathbf{x}_b-\mu_b)}$$

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

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

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

$$\Rightarrow \begin{aligned} \Lambda_{aa} &= (\Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba})^{-1} \\ \Lambda_{ab} &= -(\Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba})^{-1}\Sigma_{ab}\Sigma_{bb}^{-1} \end{aligned}$$

$$\Rightarrow \boxed{\begin{aligned} \mu_{a|b} &= \mu_a + \Sigma_{ab}\Sigma_{bb}^{-1}(\mathbf{x}_b-\mu_b) \\ \Sigma_{a|b} &= \Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba}. \end{aligned}}$$

Marginal Gaussian Distributions

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

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

1. considering the terms involving x_b and then completing the square:

$$-\frac{1}{2}\mathbf{x}_b^{\mathrm{T}}\boldsymbol{\Lambda}_{bb}\mathbf{x}_b+\mathbf{x}_b^{T}\mathbf{m} = -\frac{1}{2}(\mathbf{x}_b-\boldsymbol{\Lambda}_{bb}^{-1}\mathbf{m})^{\mathrm{T}}\boldsymbol{\Lambda}_{bb}(\mathbf{x}_b-\boldsymbol{\Lambda}_{bb}^{-1}\mathbf{m})+\frac{1}{2}\mathbf{m}^{\mathrm{T}}\boldsymbol{\Lambda}_{bb}^{-1}\mathbf{m}$$

$$\int \exp\left\{-\frac{1}{2}(\mathbf{x}_b-\boldsymbol{\Lambda}_{bb}^{-1}\mathbf{m})^{\mathrm{T}}\boldsymbol{\Lambda}_{bb}(\mathbf{x}_b-\boldsymbol{\Lambda}_{bb}^{-1}\mathbf{m})\right\}\mathrm{d}\mathbf{x}_b \qquad \mathbf{m} = \boldsymbol{\Lambda}_{bb}\boldsymbol{\mu}_b-\boldsymbol{\Lambda}_{ba}(\mathbf{x}_a-\boldsymbol{\mu}_a)$$

2. considering the remaining terms that depend on x_a:

$$\frac{1}{2}\left[\boldsymbol{\Lambda}_{bb}\boldsymbol{\mu}_b-\boldsymbol{\Lambda}_{ba}(\mathbf{x}_a-\boldsymbol{\mu}_a)\right]^{\mathrm{T}}\boldsymbol{\Lambda}_{bb}^{-1}\left[\boldsymbol{\Lambda}_{bb}\boldsymbol{\mu}_b-\boldsymbol{\Lambda}_{ba}(\mathbf{x}_a-\boldsymbol{\mu}_a)\right]-\frac{1}{2}\mathbf{x}_a^{\mathrm{T}}\boldsymbol{\Lambda}_{aa}\mathbf{x}_a+\mathbf{x}_a^{\mathrm{T}}(\boldsymbol{\Lambda}_{aa}\boldsymbol{\mu}_a+\boldsymbol{\Lambda}_{ab}\boldsymbol{\mu}_b)+\text{const}$$

$$= -\frac{1}{2}\mathbf{x}_a^{\mathrm{T}}(\boldsymbol{\Lambda}_{aa}-\boldsymbol{\Lambda}_{ab}\boldsymbol{\Lambda}_{bb}^{-1}\boldsymbol{\Lambda}_{ba})\mathbf{x}_a+\mathbf{x}_a^{\mathrm{T}}(\boldsymbol{\Lambda}_{aa}-\boldsymbol{\Lambda}_{ab}\boldsymbol{\Lambda}_{bb}^{-1}\boldsymbol{\Lambda}_{ba})^{-1}\boldsymbol{\mu}_a+\text{const}$$

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

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

$$\boldsymbol{\Sigma}_a(\boldsymbol{\Lambda}_{aa}-\boldsymbol{\Lambda}_{ab}\boldsymbol{\Lambda}_{bb}^{-1}\boldsymbol{\Lambda}_{ba})\boldsymbol{\mu}_a = \boldsymbol{\mu}_a$$

$$\boxed{\begin{aligned}\mathbb{E}[\mathbf{x}_a] &= \boldsymbol{\mu}_a\\ \operatorname{cov}[\mathbf{x}_a] &= \boldsymbol{\Sigma}_{aa}\end{aligned}}$$

Partitioned Gaussians

Partitioned Gaussians

Given a joint Gaussian distribution $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Lambda} \equiv \boldsymbol{\Sigma}^{-1}$ and

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

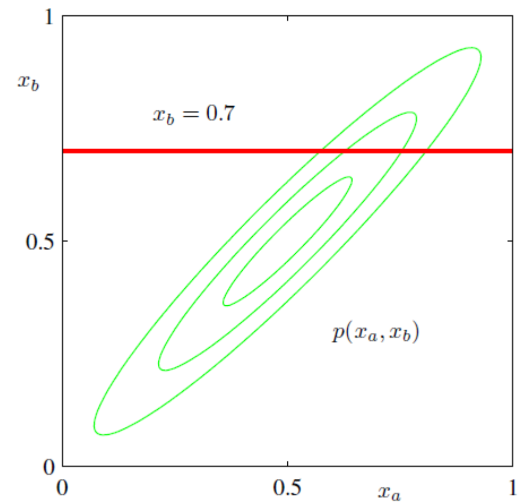
$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb} \end{pmatrix}, \quad \boldsymbol{\Lambda} = \begin{pmatrix} \boldsymbol{\Lambda}_{aa} & \boldsymbol{\Lambda}_{ab} \\ \boldsymbol{\Lambda}_{ba} & \boldsymbol{\Lambda}_{bb}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

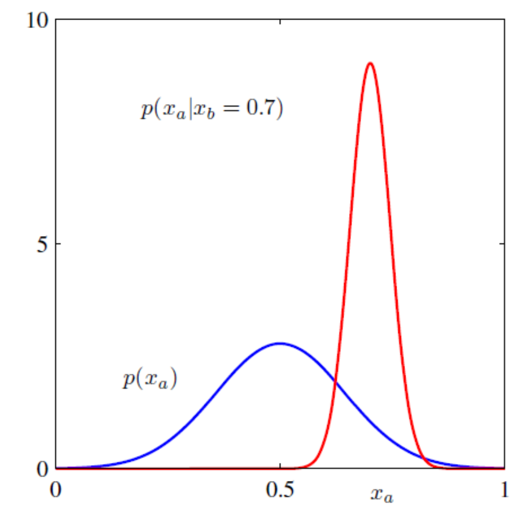
Conditional distribution:

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

Marginal distribution:

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





Bayes' Theorem for Gaussian Variables

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

$$p(\mathbf{z}) = p(\mathbf{y}|\mathbf{x})p(\mathbf{x}) \qquad \mathbf{z} = \begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix}$$

$$p(\mathbf{x}|\mathbf{y}) \qquad\qquad p(\mathbf{y})$$

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

Bayes' Theorem for Gaussian Variables

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

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

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

$$\Rightarrow$$

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

$$\operatorname{cov}[\mathbf{z}] = \mathbf{R}^{-1} = \begin{pmatrix}\boldsymbol{\Lambda}^{-1} & \boldsymbol{\Lambda}^{-1}\mathbf{A}^{\mathrm{T}}\\ \mathbf{A}\boldsymbol{\Lambda}^{-1} & \mathbf{L}^{-1}+\mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^{\mathrm{T}}\end{pmatrix}$$

$$\mathbf{x}^{\mathrm{T}}\boldsymbol{\Lambda}\boldsymbol{\mu} - \mathbf{x}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{L}\mathbf{b} + \mathbf{y}^{\mathrm{T}}\mathbf{L}\mathbf{b} = \begin{pmatrix}\mathbf{x}\\ \mathbf{y}\end{pmatrix}^{\mathrm{T}}\begin{pmatrix}\boldsymbol{\Lambda}\boldsymbol{\mu}-\mathbf{A}^{\mathrm{T}}\mathbf{L}\mathbf{b}\\ \mathbf{L}\mathbf{b}\end{pmatrix}$$

$$\Rightarrow$$

$$\mathbb{E}[\mathbf{z}] = \mathbf{R}^{-1}\begin{pmatrix}\boldsymbol{\Lambda}\boldsymbol{\mu}-\mathbf{A}^{\mathrm{T}}\mathbf{L}\mathbf{b}\\ \mathbf{L}\mathbf{b}\end{pmatrix} = \begin{pmatrix}\boldsymbol{\mu}\\ \mathbf{A}\boldsymbol{\mu}+\mathbf{b}\end{pmatrix}$$

Conditional distribution:

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

Marginal distribution:

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

$$\Rightarrow$$

$$\begin{aligned}\mathbb{E}[\mathbf{y}] &= \mathbf{A}\boldsymbol{\mu}+\mathbf{b} \\ \operatorname{cov}[\mathbf{y}] &= \mathbf{L}^{-1}+\mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^{\mathrm{T}}\end{aligned}$$

$$\begin{aligned}\mathbb{E}[\mathbf{x}|\mathbf{y}] &= (\boldsymbol{\Lambda}+\mathbf{A}^{\mathrm{T}}\mathbf{L}\mathbf{A})^{-1}\left\{\mathbf{A}^{\mathrm{T}}\mathbf{L}(\mathbf{y}-\mathbf{b})+\boldsymbol{\Lambda}\boldsymbol{\mu}\right\} \\ \operatorname{cov}[\mathbf{x}|\mathbf{y}] &= (\boldsymbol{\Lambda}+\mathbf{A}^{\mathrm{T}}\mathbf{L}\mathbf{A})^{-1}.\end{aligned}$$

Maximum Likelihood for the Gaussian

- Given a data set $\mathbf{X} = (x_1, \ldots, x_N)^T$ in which the observations $\{x_n\}$ are assumed to be drawn independently from a multivariate Gaussian distribution, how to estimate the parameters of the distribution by maximum likelihood?

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

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

Maximum Likelihood for the Gaussian

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

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

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

$$\boxed{\frac{\partial \mathbf{a}^T\mathbf{X}^{-1}\mathbf{b}}{\partial \mathbf{X}} = -\mathbf{X}^{-T}\mathbf{a}\mathbf{b}^T\mathbf{X}^{-T}}$$

$$\boxed{\frac{\partial}{\partial\mathbf{A}}\ln|\mathbf{A}| = \left(\mathbf{A}^{-1}\right)^{\mathrm{T}}} \quad \Rightarrow \quad -\frac{N}{2}\frac{\partial}{\partial\boldsymbol{\Sigma}}\ln|\boldsymbol{\Sigma}| = -\frac{N}{2}\left(\boldsymbol{\Sigma}^{-1}\right)^{\mathrm{T}} = -\frac{N}{2}\boldsymbol{\Sigma}^{-1}$$

$$\Rightarrow \quad -\frac{1}{2}\frac{\partial}{\partial\boldsymbol{\Sigma}}\sum_{n=1}^{N}(\mathbf{x}_n-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}_n-\boldsymbol{\mu}) = \frac{N}{2}\boldsymbol{\Sigma}^{-1}\mathbf{S}\boldsymbol{\Sigma}^{-1} \qquad \mathbf{S} = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n-\boldsymbol{\mu})(\mathbf{x}_n-\boldsymbol{\mu})^{\mathrm{T}}$$

$$\frac{N}{2}\boldsymbol{\Sigma}^{-1} = \frac{N}{2}\boldsymbol{\Sigma}^{-1}\mathbf{S}\boldsymbol{\Sigma}^{-1} \quad \Rightarrow \quad \boldsymbol{\Sigma} = \mathbf{S} \qquad \boldsymbol{\Sigma}_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n-\boldsymbol{\mu}_{\mathrm{ML}})(\mathbf{x}_n-\boldsymbol{\mu}_{\mathrm{ML}})^{\mathrm{T}}$$

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

$$\boxed{\frac{\partial}{\partial \mathbf{A}} \ln|\mathbf{A}| = \left(\mathbf{A}^{-1}\right)^{\mathrm{T}}} \Rightarrow -\frac{N}{2}\frac{\partial}{\partial \boldsymbol{\Sigma}} \ln|\boldsymbol{\Sigma}| = -\frac{N}{2}\left(\boldsymbol{\Sigma}^{-1}\right)^{\mathrm{T}} = -\frac{N}{2}\boldsymbol{\Sigma}^{-1}$$

$$\sum_{n=1}^{N} (\mathbf{x}_n - \boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{x}_n - \boldsymbol{\mu}) = N \operatorname{Tr}\left[\boldsymbol{\Sigma}^{-1}\mathbf{S}\right]$$

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

$$\Rightarrow$$

$$\begin{aligned}
\frac{\partial}{\partial \Sigma_{ij}} \sum_{n=1}^{N} (\mathbf{x}_n - \boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{x}_n - \boldsymbol{\mu}) &= N \frac{\partial}{\partial \Sigma_{ij}} \operatorname{Tr}\left[\boldsymbol{\Sigma}^{-1}\mathbf{S}\right] \\
&= N \operatorname{Tr}\left[\frac{\partial}{\partial \Sigma_{ij}} \boldsymbol{\Sigma}^{-1}\mathbf{S}\right] = -N \operatorname{Tr}\left[\boldsymbol{\Sigma}^{-1} \frac{\partial \boldsymbol{\Sigma}}{\partial \Sigma_{ij}} \boldsymbol{\Sigma}^{-1}\mathbf{S}\right] \\
&= -N \operatorname{Tr}\left[\frac{\partial \boldsymbol{\Sigma}}{\partial \Sigma_{ij}} \boldsymbol{\Sigma}^{-1}\mathbf{S}\boldsymbol{\Sigma}^{-1}\right] = -N\left(\boldsymbol{\Sigma}^{-1}\mathbf{S}\boldsymbol{\Sigma}^{-1}\right)_{ij}
\end{aligned}$$

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

$$\frac{N}{2}\boldsymbol{\Sigma}^{-1} = \frac{N}{2}\boldsymbol{\Sigma}^{-1}\mathbf{S}\boldsymbol{\Sigma}^{-1} \Rightarrow \boldsymbol{\Sigma} = \mathbf{S} \qquad \boldsymbol{\Sigma}_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N} (\mathbf{x}_n - \boldsymbol{\mu}_{\mathrm{ML}})(\mathbf{x}_n - \boldsymbol{\mu}_{\mathrm{ML}})^{\mathrm{T}}$$

Maximum Likelihood for the Gaussian

- Estimate the parameters of the distribution by maximum likelihood:

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

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

$$\begin{aligned}
\mathbb{E}[\boldsymbol{\Sigma}_{\mathrm{ML}}] &= \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\left[\left(\mathbf{x}_n - \frac{1}{N}\sum_{m=1}^{N}\mathbf{x}_m\right)\left(\mathbf{x}_n^{\mathrm{T}} - \frac{1}{N}\sum_{l=1}^{N}\mathbf{x}_l^{\mathrm{T}}\right)\right] \\
&= \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\left[\mathbf{x}_n\mathbf{x}_n^{\mathrm{T}} - \frac{2}{N}\mathbf{x}_n\sum_{m=1}^{N}\mathbf{x}_m^{\mathrm{T}} + \frac{1}{N^2}\sum_{m=1}^{N}\sum_{l=1}^{N}\mathbf{x}_m\mathbf{x}_l^{\mathrm{T}}\right] \\
&= \left\{\boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}} + \boldsymbol{\Sigma} - 2\left(\boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}} + \frac{1}{N}\boldsymbol{\Sigma}\right) + \boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}} + \frac{1}{N}\boldsymbol{\Sigma}\right\} \\
&= \left(\frac{N-1}{N}\right)\boldsymbol{\Sigma}
\end{aligned}$$

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

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

Bayesian Inference for the Gaussian

- Maximum likelihood framework → Bayesian treatment
  - Input:

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

	Known	**To infer**
$\mathcal{N}(x\|\mu, \sigma^2)$	variance σ^2	mean μ
	mean μ	variance σ^2
$\mathcal{N}(\mathbf{x}\|\boldsymbol{\mu}, \boldsymbol{\Sigma})$		mean μ variance σ^2

Bayesian Inference for the Gaussian

1. Known the variance, to infer the mean:

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

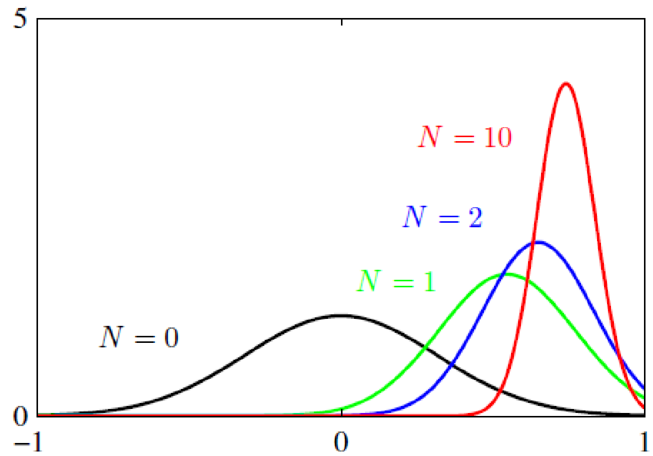
Prior: $p(\mu) = \mathcal{N}\left(\mu|\mu_0, \sigma_0^2\right)$

Posterior: $p(\mu|\mathbf{X}) \propto p(\mathbf{X}|\mu)p(\mu)$

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

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

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



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

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

Posterior: $$p(\mu|\mathbf{X}) \propto p(\mathbf{X}|\mu)p(\mu) \qquad p(\mu|\mathbf{X}) = \mathcal{N}\left(\mu|\mu_N, \sigma_N^2\right)$$

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

$$-\frac{1}{2\sigma_0^2}(\mu-\mu_0)^2 - \frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2$$

$$= -\frac{\mu^2}{2}\left(\frac{1}{\sigma_0^2} + \frac{N}{\sigma^2}\right) + \mu\left(\frac{\mu_0}{\sigma_0^2} + \frac{1}{\sigma^2}\sum_{n=1}^{N} x_n\right) + \text{const}$$

$$\frac{1}{\sigma_N^2} = \frac{N}{\sigma^2} + \frac{1}{\sigma_0^2} \qquad \mu_N = \left(\frac{N}{\sigma^2} + \frac{1}{\sigma_0^2}\right)^{-1}\left(\frac{\mu_0}{\sigma_0^2} + \frac{1}{\sigma^2}\sum_{n=1}^{N} x_n\right)$$

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

Bayesian Inference for the Gaussian

2. Known the mean, to infer the variance: $\lambda \equiv 1/\sigma^2$

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

Prior: $\mathrm{Gam}(\lambda|a_0, b_0)$ *gamma distribution*

Posterior: $p(\lambda|\mathbf{X}) \propto p(\mathbf{X}|\lambda)\,\mathrm{Gam}(\lambda|a_0, b_0)$

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

$$p(\lambda|\mathbf{X}) \propto \lambda^{a_0-1}\lambda^{N/2} \exp\left\{-b_0\lambda - \frac{\lambda}{2}\sum_{n=1}^{N}(x_n-\mu)^2\right\} \Rightarrow \mathrm{Gam}(\lambda|a_N, b_N)$$

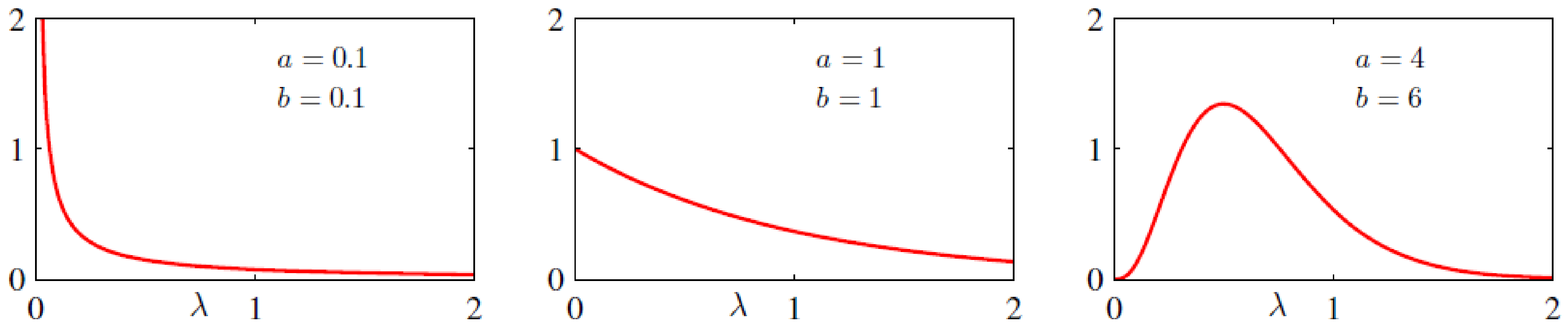
$$a_N = a_0 + \frac{N}{2}$$

$$b_N = b_0 + \frac{1}{2}\sum_{n=1}^{N}(x_n-\mu)^2 = b_0 + \frac{N}{2}\sigma^2_{\mathrm{ML}}$$

Gamma distribution:

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

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



$$\begin{aligned}
\int_0^\infty \mathrm{Gam}(\tau|a,b)\,\mathrm{d}\tau &= \frac{1}{\Gamma(a)} \int_0^\infty b^a \tau^{a-1} \exp(-b\tau)\,\mathrm{d}\tau \\
&= \frac{1}{\Gamma(a)} \int_0^\infty b^a u^{a-1} \exp(-u) b^{1-a} b^{-1}\,\mathrm{d}u \\
&= 1
\end{aligned}$$

$$\boxed{b\tau = u}$$

$$\begin{aligned}
\mathbb{E}[\tau] &= \frac{1}{\Gamma(a)} \int_0^\infty b^a \tau^{a-1} \tau \exp(-b\tau)\,\mathrm{d}\tau \\
&= \frac{1}{\Gamma(a)} \int_0^\infty b^a u^a \exp(-u) b^{-a} b^{-1}\,\mathrm{d}u \\
&= \frac{\Gamma(a+1)}{b\Gamma(a)} = \frac{a}{b}
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}[\tau^2] &= \frac{1}{\Gamma(a)} \int_0^\infty b^a \tau^{a-1} \tau^2 \exp(-b\tau)\,\mathrm{d}\tau \\
&= \frac{1}{\Gamma(a)} \int_0^\infty b^a u^{a+1} \exp(-u) b^{-a-1} b^{-1}\,\mathrm{d}u \\
&= \frac{\Gamma(a+2)}{b^2\Gamma(a)} = \frac{(a+1)\Gamma(a+1)}{b^2\Gamma(a)} = \frac{a(a+1)}{b^2}
\end{aligned}$$

$$\mathrm{var}[\tau] = \mathbb{E}[\tau^2] - \mathbb{E}[\tau]^2 = \frac{a(a+1)}{b^2} - \frac{a^2}{b^2} = \frac{a}{b^2}$$

Bayesian Inference for the Gaussian

3. Both unknown, to infer the mean and the variance: $\lambda \equiv 1/\sigma^2$

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

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

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

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

$$= \exp\left\{-\frac{\beta\lambda}{2}(\mu - c/\beta)^2\right\}\lambda^{\beta/2}\exp\left\{-\left(d-\frac{c^2}{2\beta}\right)\lambda\right\}$$

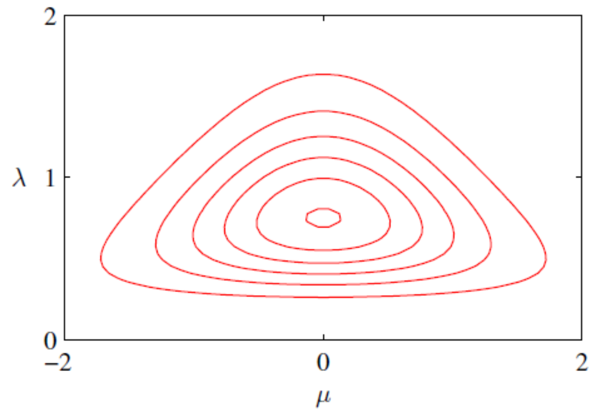
$$= \mathcal{N}(\mu|\mu_0,(\beta\lambda)^{-1})\mathrm{Gam}(\lambda|a,b)$$ *normal-gamma or Gaussian-gamma*

Posterior: $$p(\mu,\lambda|\mathbf{X}) \propto p(\mathbf{X}|\mu,\lambda)p(\mu,\lambda)$$

- *Normal-gamma or Gaussian-gamma distribution:*

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

$$= \exp\left\{-\frac{\beta\lambda}{2}(\mu - c/\beta)^2\right\} \lambda^{\beta/2} \exp\left\{-\left(d - \frac{c^2}{2\beta}\right)\lambda\right\}$$



$$\boxed{\mu_0 = c/\beta \quad a = 1 + \beta/2 \quad b = d - c^2/2\beta}$$

Conjugacy: *If we choose a prior, then the posterior distribution will have the same functional form as the prior.*

- *Normal-Wishart or Gaussian-Wishart distribution :*

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

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

$$B(\mathbf{W}, \nu) = |\mathbf{W}|^{-\nu/2}\left(2^{\nu D/2}\,\pi^{D(D-1)/4}\prod_{i=1}^{D}\Gamma\left(\frac{\nu+1-i}{2}\right)\right)^{-1}$$

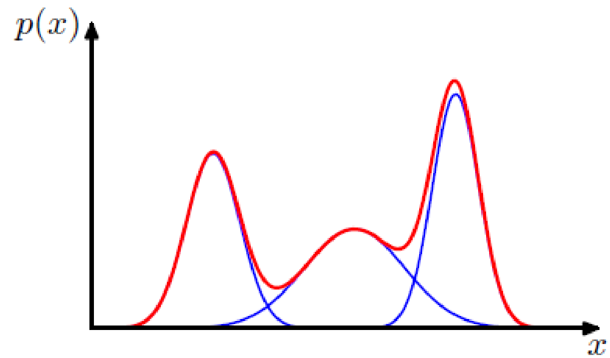
Mixture of Gaussians

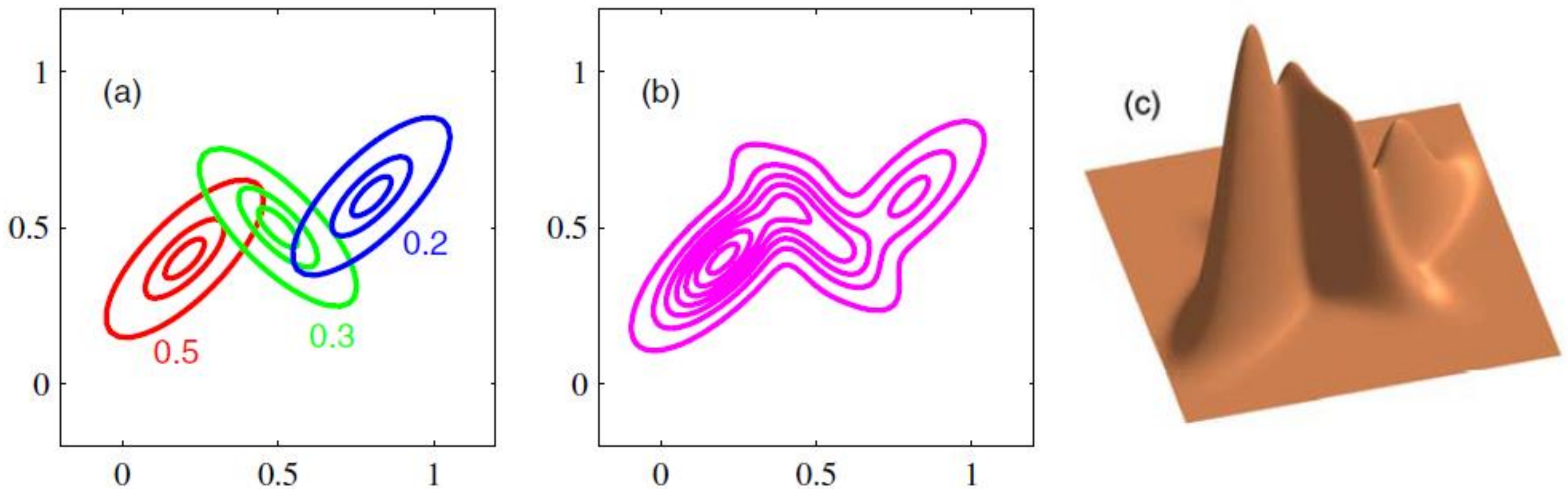
- Component and mixing coefficients

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

$$\sum_{k=1}^{K} \pi_k = 1$$

$$0 \leqslant \pi_k \leqslant 1$$





Other distributions

Binary Variables

- Bernoulli distribution:

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

$$\boxed{\begin{gathered} x \in \{0,1\} \\ p(x=1|\mu) = \mu \\ p(x=0|\mu) = 1-\mu \\ 0 \leqslant \mu \leqslant 1 \end{gathered}}$$

$$p(\mathcal{D}|\mu) = \prod_{n=1}^{N} p(x_n|\mu) = \prod_{n=1}^{N} \mu^{x_n}(1-\mu)^{1-x_n} \qquad \mathcal{D} = \{x_1, \ldots, x_N\}$$

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

- Binomial distribution:

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

$$\binom{N}{m} \equiv \frac{N!}{(N-m)!m!}$$

$$\Rightarrow$$

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

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

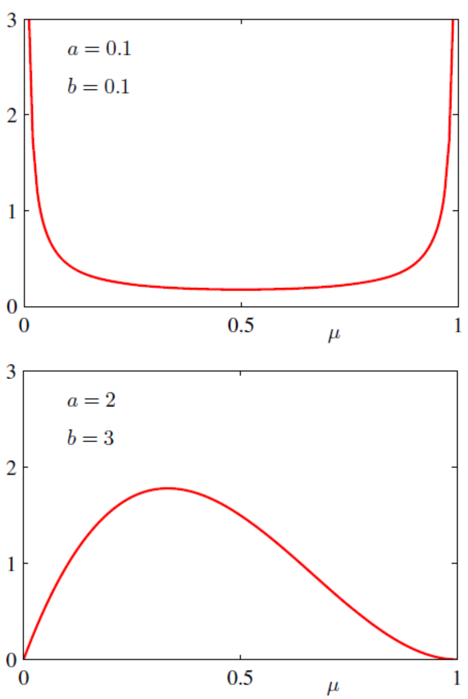
Binary Variables

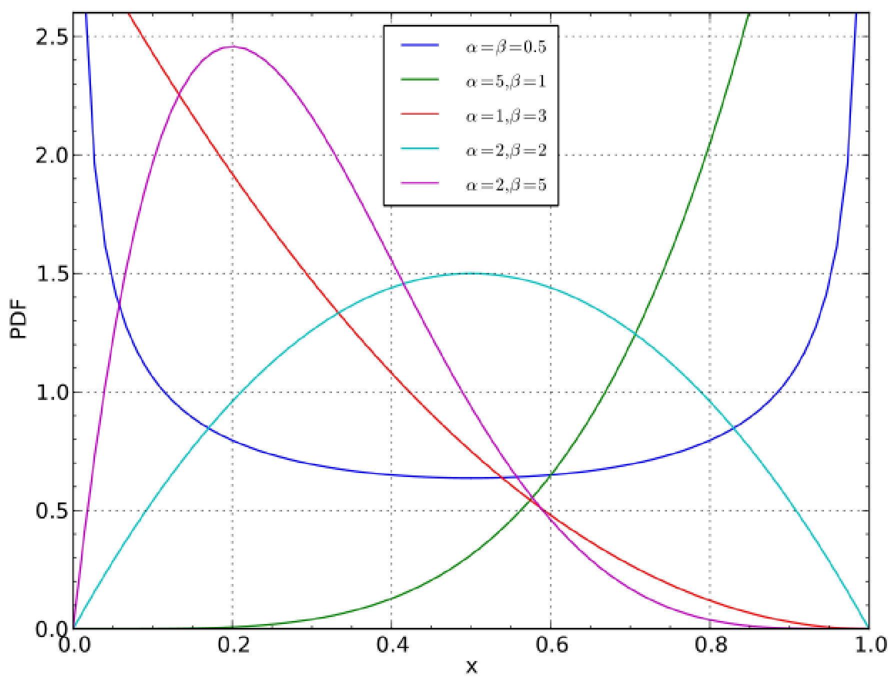
- Beta distribution:

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

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

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





Multinomial Variables

- Multinomial distribution:

$$\mathrm{Mult}(m_1, m_2, \ldots, m_K | \boldsymbol{\mu}, N) = \binom{N}{m_1 m_2 \ldots m_K} \prod_{k=1}^{K} \mu_k^{m_k}$$

$$\binom{N}{m_1 m_2 \ldots m_K} = \frac{N!}{m_1! m_2! \ldots m_K!} \qquad \sum_{k=1}^{K} m_k = N.$$

$$\mathbf{x} = (0,0,1,0,0,0)^{\mathrm{T}} \qquad \textstyle\sum_{k=1}^{K} x_k = 1$$

$$p(\mathbf{x}|\boldsymbol{\mu}) = \prod_{k=1}^{K} \mu_k^{x_k} \qquad \boldsymbol{\mu} = (\mu_1, \ldots, \mu_K)^{\mathrm{T}} \qquad \textstyle\sum_k \mu_k = 1$$

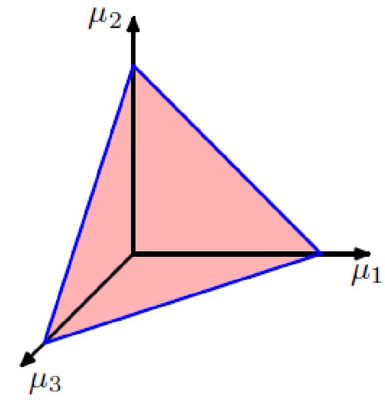
$$\sum_{\mathbf{x}} p(\mathbf{x}|\boldsymbol{\mu}) = \sum_{k=1}^{K} \mu_k = 1 \qquad \mathbb{E}[\mathbf{x}|\boldsymbol{\mu}] = \sum_{\mathbf{x}} p(\mathbf{x}|\boldsymbol{\mu})\mathbf{x} = (\mu_1, \ldots, \mu_M)^{\mathrm{T}} = \boldsymbol{\mu}$$

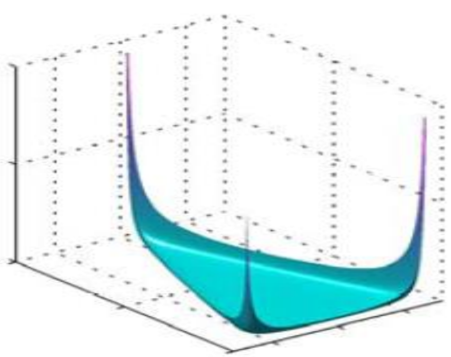
Multinomial Variables

- The Dirichlet distribution:

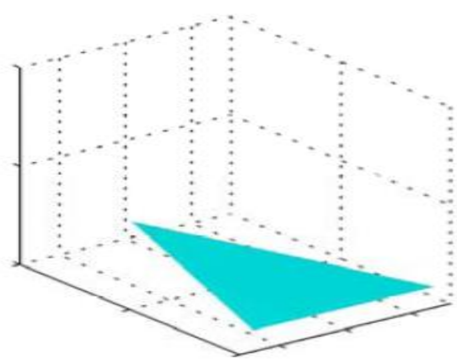
$$p(\boldsymbol{\mu}|\boldsymbol{\alpha}) \propto \prod_{k=1}^{K} \mu_k^{\alpha_k - 1} \qquad 0 \leqslant \mu_k \leqslant 1 \text{ and } \textstyle\sum_k \mu_k = 1$$

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

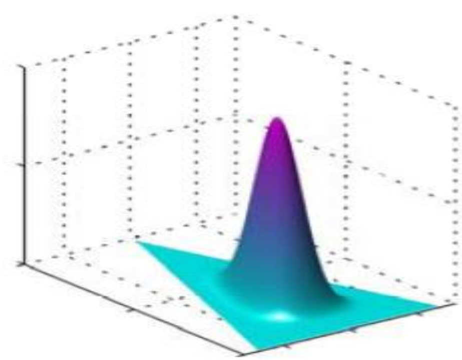




$\{\alpha_k\} = 0.1$



$\{\alpha_k\} = 1$



$\{\alpha_k\} = 10$

Student's t-distribution

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

$$= \int_0^{\infty} \mathcal{N}(x|\mu,\tau^{-1})\mathrm{Gam}(\tau|a,b)\,\mathrm{d}\tau$$

$$\nu = 2a \qquad \lambda = a/b$$

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

$$\mathbb{E}[\mathbf{x}] = \mu, \qquad \text{if} \quad \nu > 1$$

$$\mathrm{cov}[\mathbf{x}] = \frac{\nu}{(\nu-2)}\mathbf{\Lambda}^{-1}, \qquad \text{if} \quad \nu > 2$$

- *Precision λ and degrees of freedom ν*

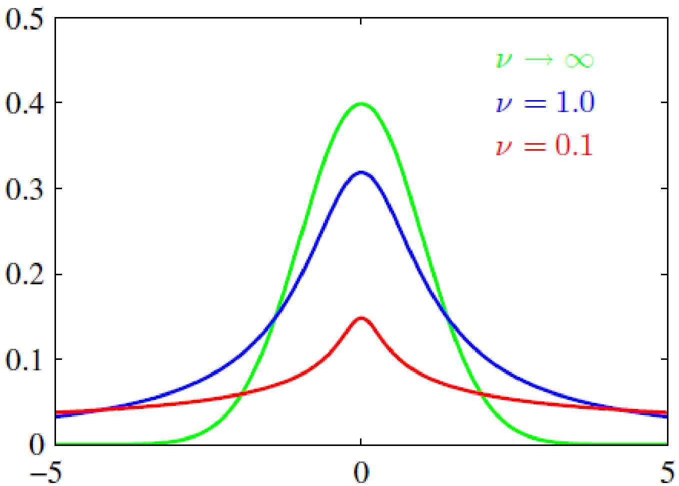
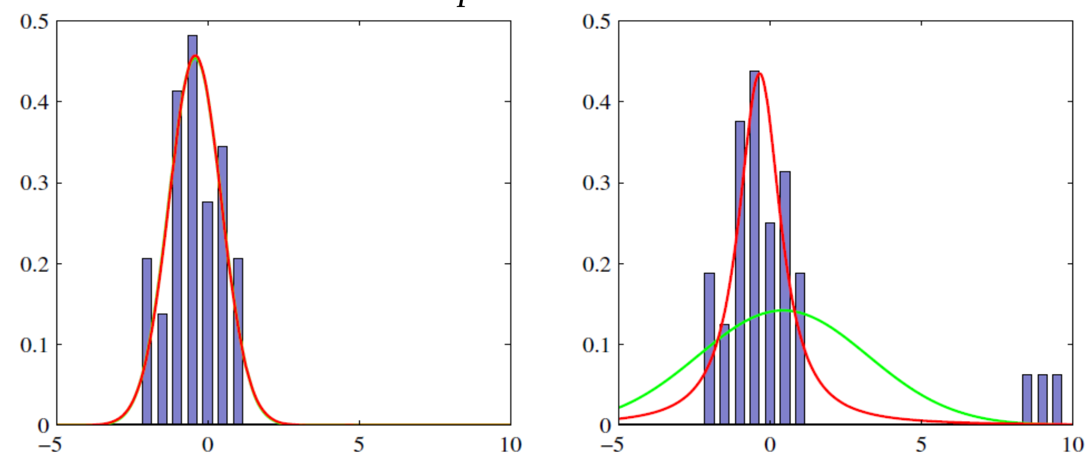


Illustration of the robustness of Student's t-distribution compared to a Gaussian



The Exponential Family

The Exponential Family

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

- A pdf or pmf $p(\mathbf{x}|\boldsymbol{\theta})$, for $\mathbf{x} = (x_1, \ldots, x_m) \in \mathcal{X}^m$ and $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^d$, is said to be in the **exponential family** if it is of the form

$$p(\mathbf{x}|\boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})}h(\mathbf{x})\exp[\boldsymbol{\theta}^T\phi(\mathbf{x})] \tag{9.1}$$

$$= h(\mathbf{x})\exp[\boldsymbol{\theta}^T\phi(\mathbf{x}) - A(\boldsymbol{\theta})] \tag{9.2}$$

where

$$Z(\boldsymbol{\theta}) = \int_{\mathcal{X}^m} h(\mathbf{x})\exp[\boldsymbol{\theta}^T\phi(\mathbf{x})]d\mathbf{x} \tag{9.3}$$

$$A(\boldsymbol{\theta}) = \log Z(\boldsymbol{\theta}) \tag{9.4}$$

Here $\boldsymbol{\theta}$ are called the **natural parameters** or **canonical parameters**, $\phi(\mathbf{x}) \in \mathbb{R}^d$ is called a vector of **sufficient statistics**, $Z(\boldsymbol{\theta})$ is called the **partition function**, $A(\boldsymbol{\theta})$ is called the **log partition function** or **cumulant function**, and $h(\mathbf{x})$ is the a scaling constant, often 1. If $\phi(\mathbf{x}) = \mathbf{x}$, we say it is a **natural exponential family**.

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

- Bernoulli distribution:

$$p(x|\mu) = \mathrm{Bern}(x|\mu) = \mu^x(1-\mu)^{1-x} = \exp\{x\ln\mu + (1-x)\ln(1-\mu)\}$$

$$= (1-\mu)\exp\left\{\ln\left(\frac{\mu}{1-\mu}\right)x\right\} = p(x|\eta) = \sigma(-\eta)\exp(\eta x)$$

Logistic sigmoid function

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

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

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

- Multinomial distribution:

$$p(\mathbf{x}|\boldsymbol{\mu}) = \prod_{k=1}^{M}\mu_k^{x_k} = \exp\left\{\sum_{k=1}^{M}x_k\ln\mu_k\right\} \quad \frac{\boldsymbol{\eta} = (\eta_1,\ldots,\eta_M)^{\mathrm{T}}}{\mathbf{x} = (x_1,\ldots,x_N)^{\mathrm{T}}} \quad p(\mathbf{x}|\boldsymbol{\eta}) = \exp(\boldsymbol{\eta}^{\mathrm{T}}\mathbf{x})$$

$$\frac{\ln\left(\frac{\mu_k}{1-\sum_j\mu_j}\right) = \eta_k}{0 \leqslant \mu_k \leqslant 1,\ \sum_{k=1}^{M-1}\mu_k \leqslant 1} \quad \left(1+\sum_{k=1}^{M-1}\exp(\eta_k)\right)^{-1}\exp(\boldsymbol{\eta}^{\mathrm{T}}\mathbf{x})$$

Softmax function

$$\mu_k = \frac{\exp(\eta_k)}{1+\sum_j\exp(\eta_j)}$$

- Gaussian distribution:

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

$$\Rightarrow$$

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

Maximum likelihood and sufficient statistics

- To estimate $\boldsymbol{\eta}$ by ML:

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

$$\Rightarrow \quad -\frac{1}{g(\boldsymbol{\eta})}\nabla g(\boldsymbol{\eta}) = g(\boldsymbol{\eta}) \int h(\mathbf{x}) \exp\left\{\boldsymbol{\eta}^{\mathrm{T}}\mathbf{u}(\mathbf{x})\right\} \mathbf{u}(\mathbf{x})\,\mathrm{d}\mathbf{x} = \mathbb{E}[\mathbf{u}(\mathbf{x})]$$

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

$$\begin{aligned} -\nabla\nabla \ln g(\boldsymbol{\eta}) &= g(\boldsymbol{\eta}) \int h(\mathbf{x}) \exp\left\{\boldsymbol{\eta}^{\mathrm{T}}\mathbf{u}(\mathbf{x})\right\} \mathbf{u}(\mathbf{x})\mathbf{u}(\mathbf{x})^{\mathrm{T}}\,\mathrm{d}\mathbf{x} + \nabla g(\boldsymbol{\eta}) \int h(\mathbf{x}) \exp\left\{\boldsymbol{\eta}^{\mathrm{T}}\mathbf{u}(\mathbf{x})\right\} \mathbf{u}(\mathbf{x})\,\mathrm{d}\mathbf{x} \\ &= \mathbb{E}[\mathbf{u}(\mathbf{x})\mathbf{u}(\mathbf{x})^{\mathrm{T}}] - \mathbb{E}[\mathbf{u}(\mathbf{x})]\mathbb{E}[\mathbf{u}(\mathbf{x})^{\mathrm{T}}] \quad = \quad \mathrm{cov}[\mathbf{u}(\mathbf{x})] \end{aligned}$$

$$p(\mathbf{X}|\boldsymbol{\eta}) = \left(\prod_{n=1}^{N} h(\mathbf{x}_n)\right) g(\boldsymbol{\eta})^N \exp\left\{\boldsymbol{\eta}^{\mathrm{T}} \sum_{n=1}^{N} \mathbf{u}(\mathbf{x}_n)\right\} \quad \Rightarrow \quad -\nabla \ln g(\boldsymbol{\eta}_{\mathrm{ML}}) = \frac{1}{N}\sum_{n=1}^{N} \mathbf{u}(\mathbf{x}_n)$$

$$N \to \infty \qquad \mathbb{E}[\mathbf{u}(\mathbf{x})]$$

Conjugate priors

- Conjugacy: If we choose a prior, then the posterior distribution will have the same functional form as the prior.
- For any member of the exponential family: $p(\mathbf{x}|\boldsymbol{\eta}) = h(\mathbf{x})g(\boldsymbol{\eta})\exp\left\{\boldsymbol{\eta}^{\mathrm{T}}\mathbf{u}(\mathbf{x})\right\}$
- there exists a conjugate prior: $p(\boldsymbol{\eta}|\boldsymbol{\chi},\nu) = f(\boldsymbol{\chi},\nu)g(\boldsymbol{\eta})^{\nu}\exp\left\{\nu\boldsymbol{\eta}^{\mathrm{T}}\boldsymbol{\chi}\right\}$

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

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

$$\Rightarrow \quad p(\boldsymbol{\eta}|\mathbf{X},\boldsymbol{\chi},\nu) \propto p(\mathbf{X}|\boldsymbol{\eta})\,p(\boldsymbol{\eta}|\boldsymbol{\chi},\nu)$$

$$= \left(\prod_{n=1}^{N} h(\mathbf{x}_n)\right) g(\boldsymbol{\eta})^N \exp\left\{\boldsymbol{\eta}^{\mathrm{T}} \sum_{n=1}^{N} \mathbf{u}(\mathbf{x}_n)\right\} f(\boldsymbol{\chi},\nu)g(\boldsymbol{\eta})^{\nu}\exp\left\{\nu\boldsymbol{\eta}^{\mathrm{T}}\boldsymbol{\chi}\right\}$$

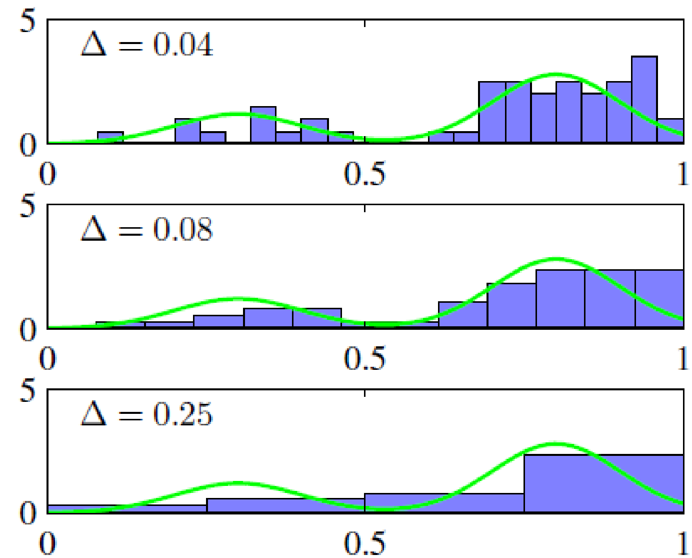
$$\propto g(\boldsymbol{\eta})^{\nu+N} \exp\left\{\boldsymbol{\eta}^{\mathrm{T}}\left(\sum_{n=1}^{N} \mathbf{u}(\mathbf{x}_n) + \nu\boldsymbol{\chi}\right)\right\}$$

Nonparametric Methods

Nonparametric Methods

- How to estimate unknown probability density p(x):

$$P = \int_{\mathcal{R}} p(\mathbf{x})\,\mathrm{d}\mathbf{x} \quad \Rightarrow \quad p(\mathbf{x}) = \frac{K}{NV}$$



Histogram approach to density estimation

- Kernel density estimator
  - Fix V, determine K from the data
- KNN density estimator
  - K-nearest-neighbour
  - Fix K, determine the value of V from the data

Kernel density estimators

- Parzen window (an example of a Kernel function)

$$k(\mathbf{u}) = \begin{cases} 1, & |u_i| \leqslant 1/2, \qquad i = 1, \ldots, D \\ 0, & \text{otherwise} \end{cases}$$

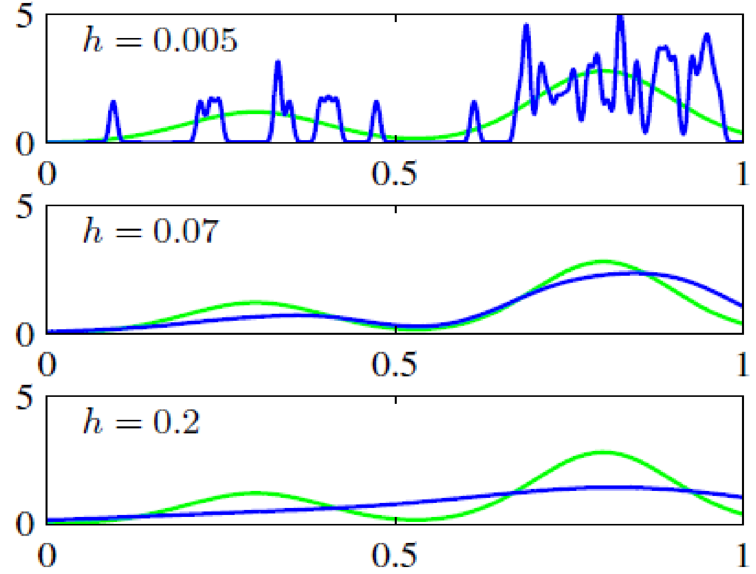
  - The total number of data points lying inside this cube:

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

  - The estimated density at x:

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

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

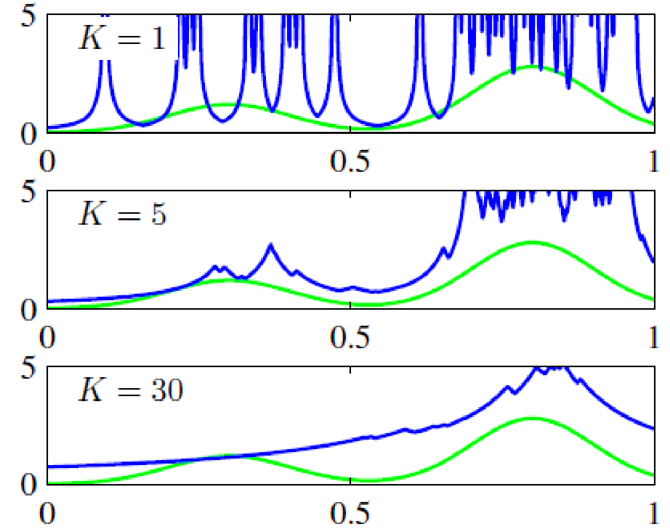


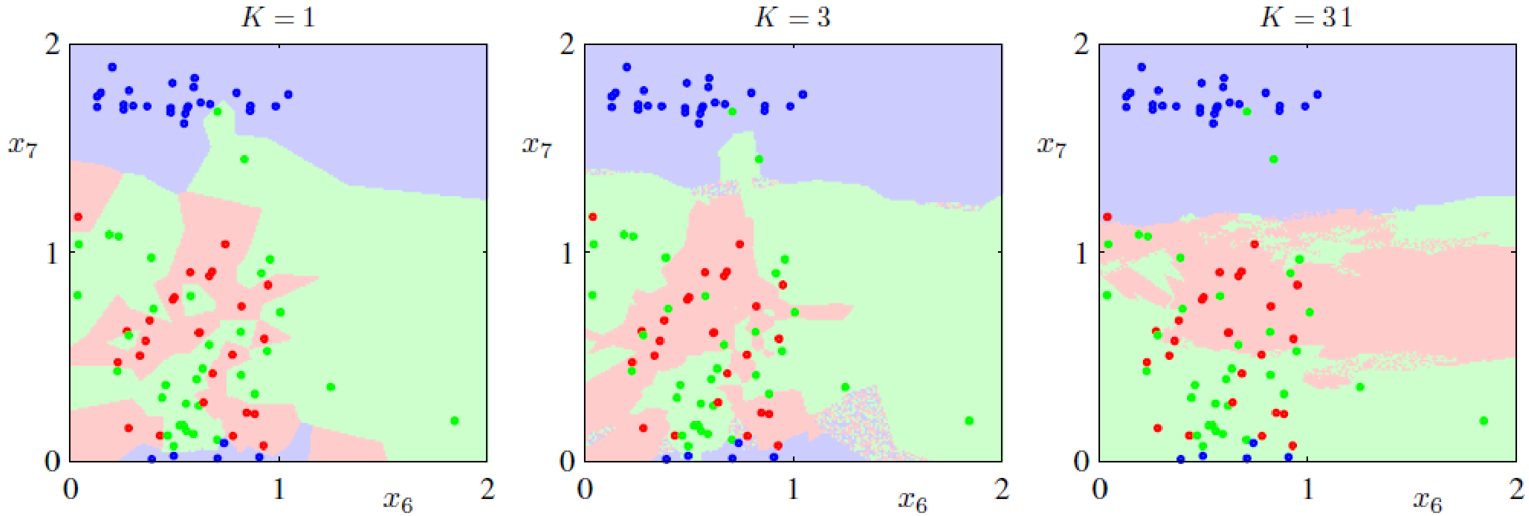
Nearest-neighbour methods

- KNN density estimation
  - K govern the radius of the sphere
- KNN classifier

$$p(\mathbf{x}|\mathcal{C}_k) = \frac{K_k}{N_k V} \qquad p(\mathbf{x}) = \frac{K}{NV} \qquad p(\mathcal{C}_k) = \frac{N_k}{N}$$

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





Next: Linear Models for Regression

- HW2:
  - 2.17, 2.19, 2.24, 2.26, 2.29, 2.30, 2.41, 2.47
  - Use KNN classifier to determine the class of handwritten digits.(find the details from course website)