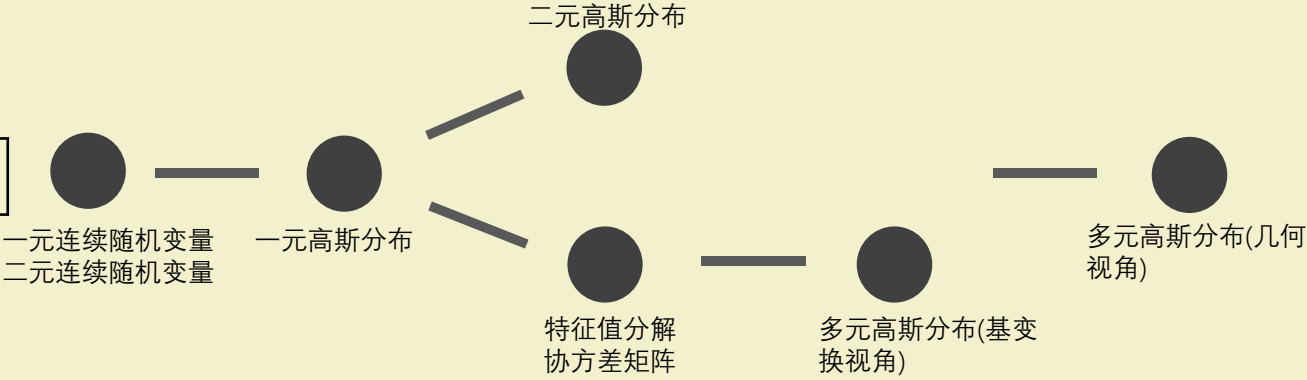


# PRML

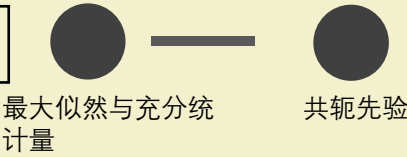
CHAPTER 02

# 研讨结构

## 2.3节



## 2.4节



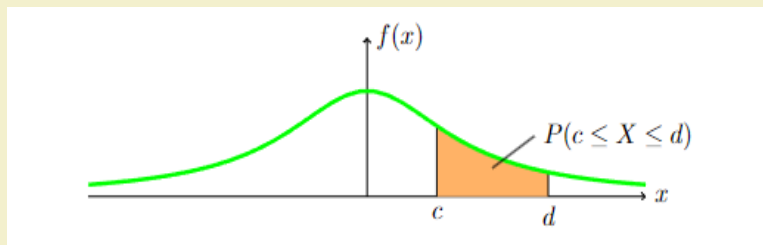
## 2.5节



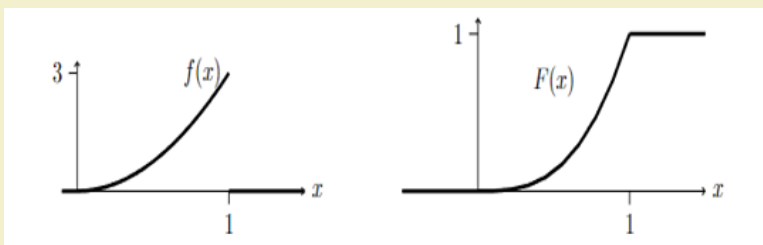
1. 由于章节内容较多，时间较为紧迫。重点放在捋清思路，以理解作者大致是在干什么的叙述过程为主，几乎不会有任何公式推导内容。希望这次的研讨能成为重读PRML的二三周目的良好开端。
2. 大量今日没有涉及的内容是非常重要的，由于水平有限，会尽可能更多提及效率更高的知识点而非最重要的知识点。
3. 最最关键：本人学术理论以及表述水平极其匮乏，如果有问题可以随意开麦问（大概率你问的问题我也不会），如果觉得我讲的有问题也随时欢迎批评



# 一元连续随机变量



一元概率密度函数(PDF)

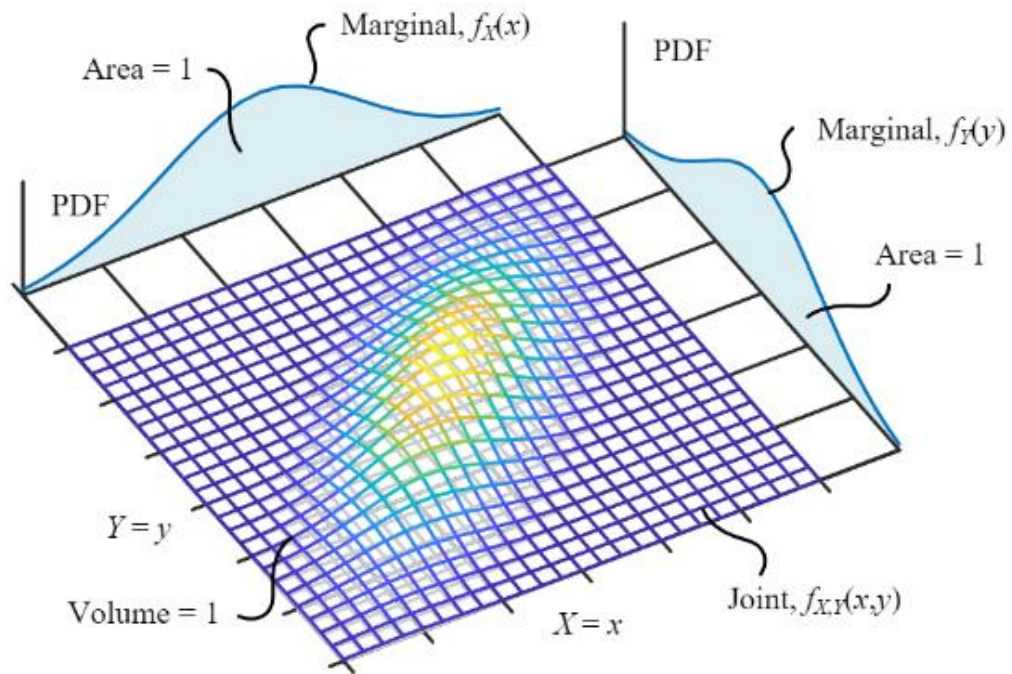


一元累积分布函数(CDF)

- 1.一元随机变量，概率用面积来表示（积分）
- 2.概率密度函数不表示概率，且必须非负，可大于1
- 3.单点与单点集合概率为0

跟分布没什么关系，是自然而然产生的。当讨论连续随机变量的时候，你无法讨论 $X$ =某个确切值的概率（恒为0无异议）。你需要划分区间，并计算概率。实际上CDF所做的事情**仅仅只是限定了随机变量的取值格式**（ $X$ 小于等于某个值的概率）

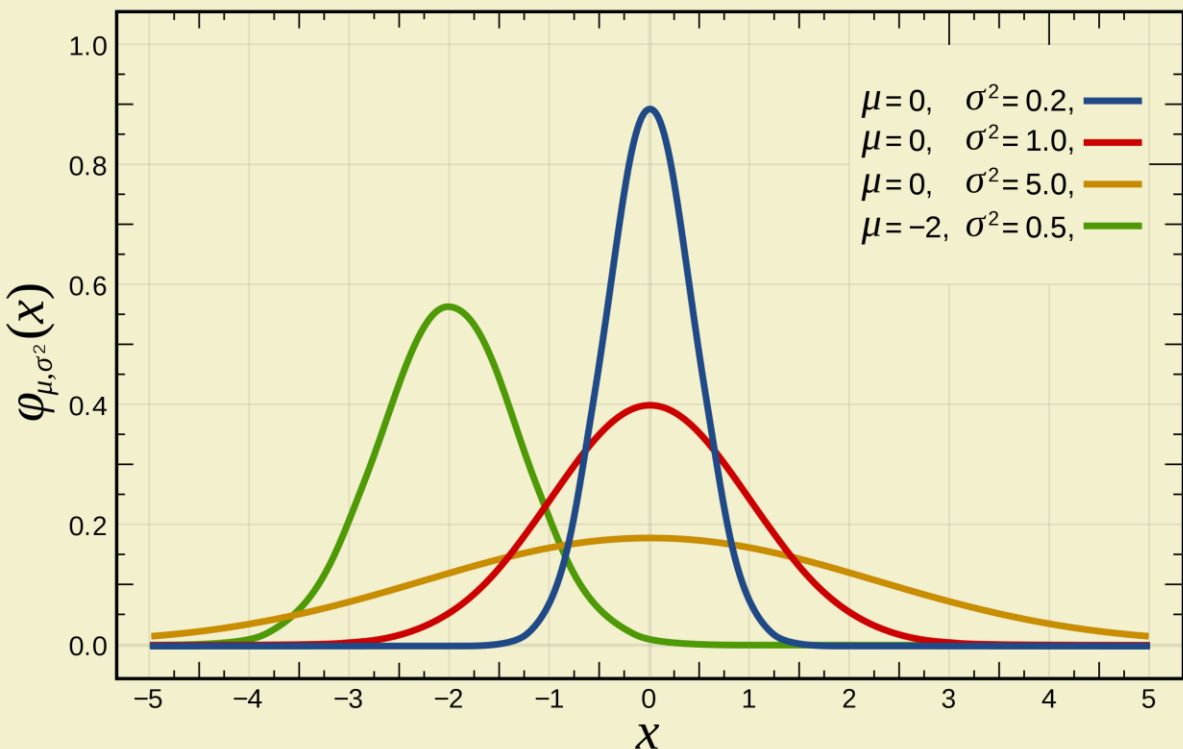
# 二元连续随机变量



二元概率密度函数(joint PDF)

1. 概率用体积来衡量
2. 截面为边缘概率密度函数
3. 通过二重积分求得概率

# 一元高斯分布



参数2: 方差 $\sigma^2$  (标准差 $\sigma$ )

控制曲线的形状

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

参数1: 均值 $\mu$

控制曲线的位置

高斯分布的产生

貌似比掷硬币要难理解一点

熵取最大值的时候为高斯分布

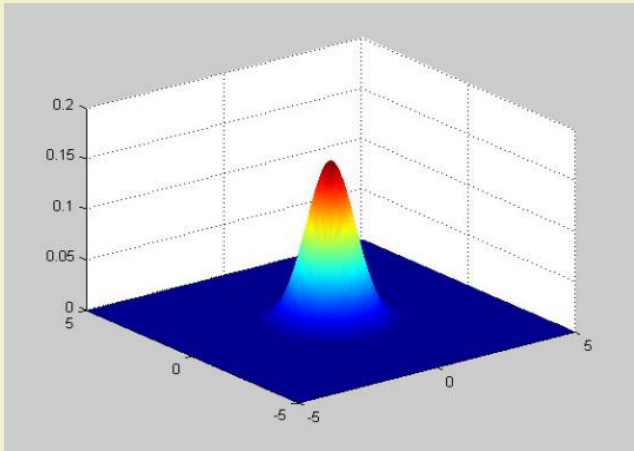
中心极限定理: 随机变量之和会产生高斯分布, 其他概率分布可以用高斯分布作为近似

高斯对测量误差研究中发现了高斯分布

- 应用:
1. 参数估计
  2. QQ图与经验分布函数
  3. 生成正态分布随机变量



# 二元高斯分布



完成归一化，使围成的体积总为一

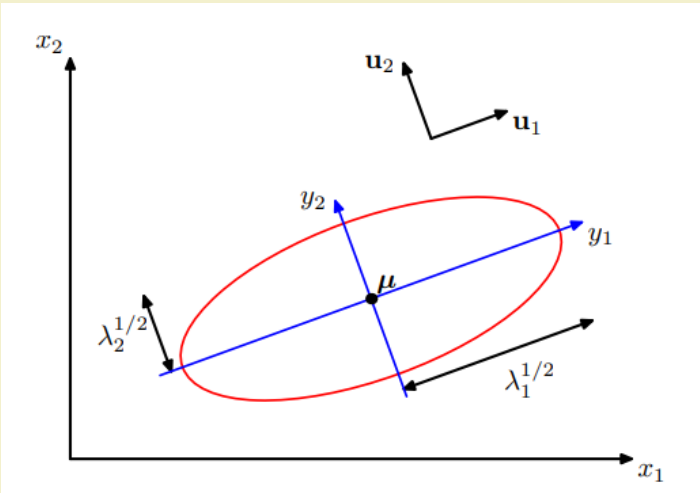
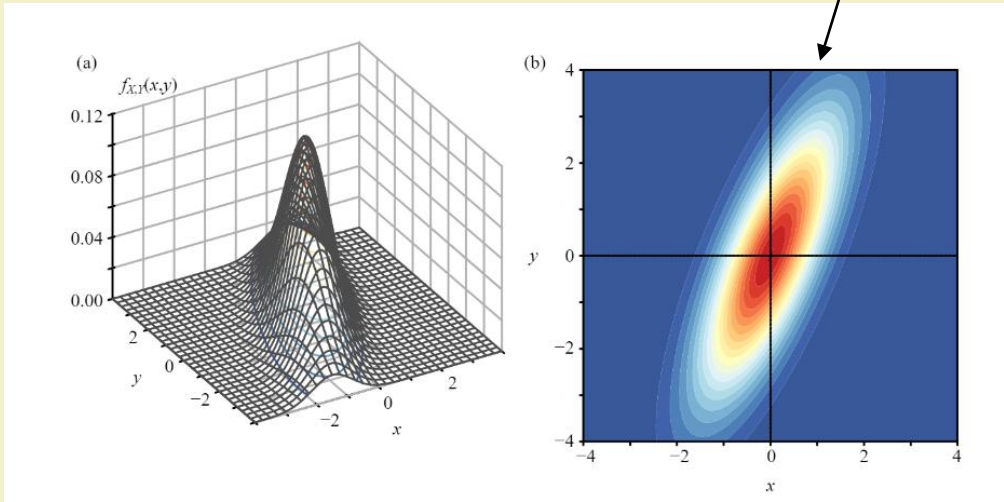
参数 $\rho$ 是XY两个随机变量的相关系数，与方差一起提供了曲面的扭曲程度。

PDF: 
$$f(x,y) = \left(2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}\right)^{-1} \exp\left[-\frac{1}{2(1-\rho^2)}\left(\frac{(x-\mu_1)^2}{\sigma_1^2} - \frac{2\rho(x-\mu_1)(y-\mu_2)}{\sigma_1\sigma_2} + \frac{(y-\mu_2)^2}{\sigma_2^2}\right)\right]$$

$$\frac{x_1^2}{m^2} + \frac{x_2^2}{n^2} - 2\rho\frac{x_1x_2}{mn} = 1$$

**椭圆解析式：**由双曲线与椭圆叠加而成， $x_1 \times x_2$ 实际上为一个旋转双曲线，系数 $\rho$ 可以调节双曲线对椭圆曲线的影响程度与 $\rho$ 成正比。当相关系数 $\rho$ 为0时，椭圆等高线为正椭圆（但 $\rho$ 不能推断出两变量独立，仅能推断出不存在线性关系）

等高线剖面为椭圆形状



椭圆蕴含的参数信息

# 多元高斯分布(序)

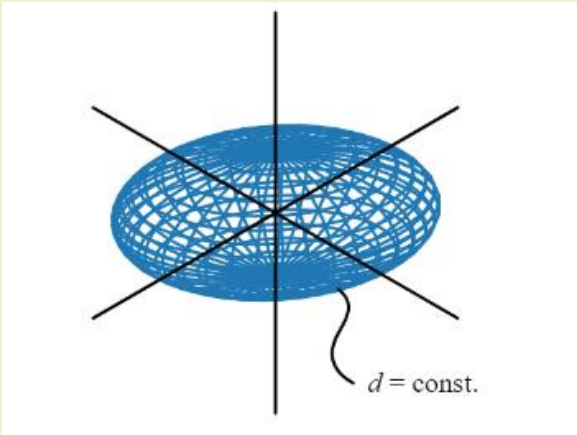
## 三元高斯分布

假设我们有 $\sigma_1=\sigma_2=\sigma_3 = 1$ ,  $\mu_1=\mu_2=\mu_3 = 0$ , 我们有三元高斯函数的解析式如下。  
与一二元相同, 前面为归一因子, 指数函数部分为椭球的解析式

$$f_{X1,X2,X3}(x_1,x_2,x_3) = \frac{\exp\left(\frac{-1}{2}d^2\right)}{(2\pi)^{\frac{3}{2}}\sqrt{1+2\rho_{1,2}\rho_{1,3}\rho_{2,3}-(\rho_{1,2}^2+\rho_{1,3}^2+\rho_{2,3}^2)}}$$

$$d^2 = \frac{x_1^2(\rho_{2,3}^2-1)+x_2^2(\rho_{1,3}^2-1)+x_3^2(\rho_{1,2}^2-1)+2[x_1x_2(\rho_{1,2}-\rho_{1,3}\rho_{2,3})+x_1x_3(\rho_{1,3}-\rho_{1,2}\rho_{2,3})+x_2x_3(\rho_{2,3}-\rho_{1,3}\rho_{2,3})]}{(\rho_{1,2}^2+\rho_{1,3}^2+\rho_{2,3}^2-2\rho_{1,2}\rho_{1,3}\rho_{2,3}-1)}$$

当d为确定值时, 我们知道三元高斯分布的PDF为嵌套的椭球。从一元到三元我们经历了由线到椭圆到椭球的过程。然而对于更高维度的情况下, 想要写出代数展开式过于复杂。我们需要使用矩阵算式。





# 多元高斯分布(破)

二次型:

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

马氏距离

书上给出了详细的多元高斯分布的解析式

D维向量

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_D \end{bmatrix}$$
$$\mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{\frac{D}{2}}} \frac{1}{|\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp \left\{ -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right\}$$

D维随机向量

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_D \end{bmatrix},$$

D\*D矩阵

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{1,1} & \sigma_{1,2} & \cdots & \sigma_{1,D} \\ \sigma_{2,1} & \sigma_{2,2} & \cdots & \sigma_{2,D} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{D,1} & \sigma_{D,2} & \cdots & \sigma_{D,D} \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & \rho_{1,2}\sigma_1\sigma_2 & \cdots & \rho_{1,D}\sigma_1\sigma_D \\ \rho_{1,2}\sigma_1\sigma_2 & \sigma_2^2 & \cdots & \rho_{2,D}\sigma_2\sigma_D \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{1,D}\sigma_1\sigma_D & \rho_{2,D}\sigma_2\sigma_D & \cdots & \sigma_D^2 \end{bmatrix}$$

矩阵的要求:

- 1.  $\Sigma$  的行列式能开根号, 说明一定大于等于0
- 2.  $\Sigma$  的行列式位于分母, 说明不能等于0

矩阵为正定矩阵

# 特征值分解

如果矩阵对某一个向量或某些向量只发生伸缩变换，不对这些向量产生旋转的效果，那么这些向量就称为这个矩阵的特征向量，伸缩的比例就是特征值。

谱分解：在特殊情况下，如果A为对称矩阵，则特征值分解可以写成

Handwritten derivation of spectral decomposition for a symmetric matrix  $A$ :

$$A = V \Lambda V^T = [v_1 \ v_2 \ \dots \ v_D] \begin{bmatrix} \lambda_1 & & \\ & \lambda_2 & \\ & & \ddots \\ & & & \lambda_D \end{bmatrix} \begin{bmatrix} v_1^T \\ v_2^T \\ \vdots \\ v_D^T \end{bmatrix}$$

外积(叉积)

$$= \lambda_1 (v_1 v_1^T) + \lambda_2 (v_2 v_2^T) + \dots + \lambda_D (v_D v_D^T)$$
$$= \lambda_1 v_1 \otimes v_1 + \lambda_2 v_2 \otimes v_2 + \dots + \lambda_D v_D \otimes v_D$$
$$= \sum_{j=1}^D \lambda_j v_j \otimes v_j$$

$D \times D$  矩阵

$$A = V \Lambda V^{-1}$$



特征向量矩阵的逆等于特征向量矩阵的转置

$$A = V \Lambda V^T$$

# 协方差矩阵

方差是用来度量单个随机变量的离散程度，而协方差则一般用来刻画两个随机变量的共性程度

$$\Sigma = \begin{bmatrix} \sigma(x_1, x_1) & \cdots & \sigma(x_1, x_d) \\ \vdots & \ddots & \vdots \\ \sigma(x_d, x_1) & \cdots & \sigma(x_d, x_d) \end{bmatrix} \in \mathbb{R}^{d \times d}$$

协方差

方差

协方差:  $\text{cov}(X, Y) = E[(X - \mu_X)(Y - \mu_Y)]$

去量纲化

相关系数:  $\rho_{XY} = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y}$

# 协方差矩阵

## 二元：协方差

现有数据集：3条数据，每条数据2个属性。意味着2元分布。

图表表示：

$$D = \{X_1, X_2, X_3\}, \quad X_i = \begin{bmatrix} x_i \\ y_i \end{bmatrix} \quad \text{随机向量}$$

	x	y
$X_1$	6	9
$X_2$	12	3
$X_3$	15	3

$$\Rightarrow \begin{aligned} \mu_x &= \frac{6+12+15}{3} = 11 \\ \mu_y &= \frac{9+3+3}{3} = 5 \end{aligned}$$

$$\text{差值矩阵} \Rightarrow \begin{bmatrix} -5 & 4 \\ 1 & -2 \\ 4 & -2 \end{bmatrix}$$

$$\Rightarrow \text{cov}(x, y) = \frac{-5 \cdot 4 + 1 \cdot (-2) + 4 \cdot (-2)}{3} = -10 \quad \text{负相关}$$

## 二元以上：协方差矩阵

现有数据集：3条数据，每条数据3个属性。意味着3元分布。

图表表示：

$$D = \{X_1, X_2, X_3\}, \quad X_i = \begin{bmatrix} x_i \\ y_i \\ z_i \end{bmatrix} \quad \begin{array}{l} \text{三元随机向量} \\ \text{作为分布函数的输入} \end{array}$$

	x	y	z
$X_1$	12	9	3
$X_2$	3	6	3
$X_3$	9	3	3

$$\mu_x = \frac{12+3+9}{3} = 8$$

$$\Rightarrow \mu_y = \frac{9+6+3}{3} = 6$$

$$\mu_z = \frac{3+3+3}{3} = 2$$

$$\mu = \begin{bmatrix} 8 \\ 6 \\ 2 \end{bmatrix} \quad \text{均值向量}$$

$$\begin{array}{ccc|c} (x-\mu_x) & (y-\mu_y) & (z-\mu_z) & \\ \hline 4 & 3 & 1 & X_1 \\ -5 & 0 & 1 & X_2 \\ 1 & -3 & 1 & X_3 \end{array}$$

$$\text{方差}(\sigma_i^2) \Rightarrow \begin{array}{l} \sigma_{1,1} = \frac{4^2+(-5)^2+1^2}{3} \\ \sigma_{2,1} = \frac{4 \cdot 0 + (-5) \cdot 1 + 1 \cdot 1}{3} \\ \sigma_{3,1} = \frac{4 \cdot 1 + 0 \cdot 1 + 1 \cdot 1}{3} \\ \sigma_{2,2} = \frac{0^2+0^2+(-3)^2}{3} \\ \sigma_{3,2} = \frac{0 \cdot 1 + (-3) \cdot 1 + 1 \cdot 1}{3} \\ \sigma_{3,3} = \frac{1^2+1^2+1^2}{3} \end{array}$$

协方差矩阵

# 多元高斯分布(Q)

$\Delta^2 = (x - \mu)^T (\Sigma^{-1}) (x - \mu)$  二次型  
前提: I.  $\Sigma$  可取也必须是实对称矩阵

II. 由于实对称矩阵性质可知

①  $\mu_i^T \mu_i = 1$  ②  $\mu_i^T \mu_j = 0$

分析: 考虑将  $\Sigma$  特征值分解.

Step 1:  $\Sigma \mu_i = \lambda_i \mu_i$   
特征值                  特征向量

谱分解

Step 2:  $\Sigma = \sum_{i=1}^D \lambda_i \mu_i \otimes \mu_i$   
 $= \sum_{i=1}^D \lambda_i \mu_i \mu_i^T$

Step 3: 求逆

$$\begin{aligned}\Sigma^{-1} &= (V \Lambda V^T)^{-1} \quad \text{已知 } V^T = V^T \\ &= (V^T)^T \Lambda^{-1} V^T \\ &= V \Lambda^{-1} V^T \\ &= \sum_{i=1}^D \frac{1}{\lambda_i} \mu_i \mu_i^T\end{aligned}$$

Step 4: 代入  $\Delta^2$

$$\begin{aligned}\Delta^2 &= (x - \mu)^T \Sigma^{-1} (x - \mu) \\ &= \sum_{i=1}^D (x - \mu)^T \left[ \frac{1}{\lambda_i} \mu_i \mu_i^T \right] (x - \mu) \\ &= \sum_{i=1}^D \frac{1}{\lambda_i} \left\{ \left[ \mu_i (x - \mu)^T \right]^T \left[ \mu_i^T (x - \mu) \right] \right\} \\ &= \sum_{i=1}^D \frac{y_i^2}{\lambda_i} \quad \boxed{y_i = \mu_i^T (x - \mu)}\end{aligned}$$



# 基变换

已知:  $y_i = M_i^T (x - u)$  这不同! ② 基变换

$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_D \end{bmatrix} = \begin{bmatrix} M_1^T (x - u) \\ M_2^T (x - u) \\ \vdots \\ M_D^T (x - u) \end{bmatrix} = V^T (x - u)$

③ 如果一开始基向量就不同?

单位向量 线性变换

前后位置改变

共通的问题: 无论我们以何种眼光(基)看待向量, 向量始终在那

eg. 对于如下向量函数

$Y = f(X)$

$X = \begin{bmatrix} u \\ v \end{bmatrix}, M = f(x, y, z) = x^2 + 2xy + z^2$

$V = f(x, y, z) = x - y^2 + z^2$

雅可比矩阵为

$\begin{bmatrix} \frac{\partial M}{\partial x} & \frac{\partial M}{\partial y} & \frac{\partial M}{\partial z} \\ \frac{\partial V}{\partial x} & \frac{\partial V}{\partial y} & \frac{\partial V}{\partial z} \end{bmatrix} = \begin{bmatrix} 2xy & x^2 & 2z \\ 1 & -2y & 2z \end{bmatrix}$

def: 若有向量  $x \in \mathbb{R}^n, y \in \mathbb{R}^n$  以及  $A \in \mathbb{R}^{m \times n}$ , 则对线性映射  $y = Ax$

有雅可比矩阵  $\frac{\partial y}{\partial x} = A$

斜视 正常 上帝视角

并无不同

$\vec{v} = 3 \cdot \begin{bmatrix} 1 \\ 0 \end{bmatrix} + 2 \cdot \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 3 \\ 2 \end{bmatrix}$

$\vec{v} = 3 \cdot \begin{bmatrix} 1 \\ \frac{1}{2} \end{bmatrix} + 2 \cdot \begin{bmatrix} -\frac{1}{2} \\ 1 \end{bmatrix} = \begin{bmatrix} 2 \\ \frac{7}{2} \end{bmatrix}$

我们眼中斜视 我们眼中的基

$T(x) = \begin{bmatrix} 1 & -\frac{1}{2} \\ \frac{1}{2} & 1 \end{bmatrix} \begin{bmatrix} 3 \\ 2 \end{bmatrix} = \begin{bmatrix} 2 \\ \frac{7}{2} \end{bmatrix}$

我们眼中的斜视 我们眼中的基

eg.  $2m = 2 \cdot 10cm = 20cm$

# 多元高斯分布 (终)

已有  $V^T V = I$ ,  $\Sigma = V \Lambda V^T$ ,  $\Sigma^{-1} = V \Lambda^{-1} V^T$

则现有  $I$  下的坐标与  $V$  下的坐标  $y$  满足转换

$$y = \Lambda^{-1/2} V^T (x - \mu) = V^T (x - \mu)$$

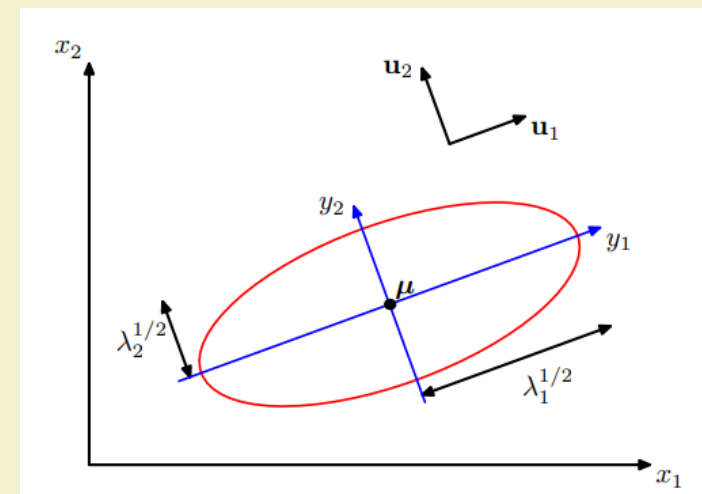
所以  $\Delta^2$  是将  $x$  平移  $\mu$  且各单位按  $\sqrt{\lambda_i}$  缩放后在以  $V$  为基的坐标系下到原点的距离。

马氏距离

将  $I$  与  $V$  视为两组基，得到转换  $V = I V$

转换矩阵为  $A = V$

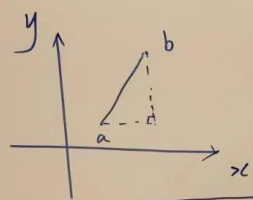
可推  $\Sigma = V \Lambda V^T$

$$\Sigma (\Lambda V^T)^T = V (\Lambda V^T) (\Lambda V^T)^T$$
$$(V^T)^T = (V^T)^T \quad \Sigma V \Lambda^{-1} = V$$
$$\Sigma [V \Lambda^{-1} V^T] = V V^T$$
$$\Sigma \Sigma^{-1} = V V^T$$
$$I = V V^T$$




# 欧氏距离

欧氏距离  $\Leftarrow L_2$  范数



二维情况下

$$\text{dist}(a, b) = \sqrt{(x_b - x_a)^2 + (y_b - y_a)^2}$$

满足条件:

- I. x轴与y轴垂直
- II. x轴与y轴单位一致

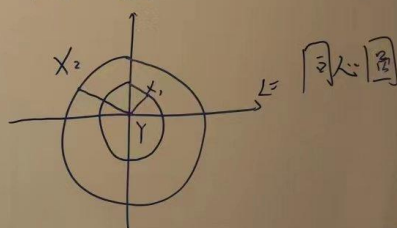
$$\text{dist}(x, y) = \sqrt{(x-y)^T (x-y)}$$

D维列向量

$$= \sqrt{\begin{bmatrix} x_1 - y_1 & x_2 - y_2 & \dots & x_D - y_D \end{bmatrix} \begin{bmatrix} x_1 - y_1 \\ x_2 - y_2 \\ \vdots \\ x_D - y_D \end{bmatrix}}$$

$$= \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \dots + (x_D - y_D)^2}$$

几何 取Y为定值. 对数据集X:  $\{x_1, x_2, \dots\}$



未标准化欧氏距离

$$\text{dist}(x, y) = \sqrt{(x-y)^T V^{-1} (x-y)}$$

$$V = \text{diag}(\text{diag}(\Sigma))$$

$$= \begin{bmatrix} \sigma_1^2 & & \\ & \sigma_2^2 & \\ & & \ddots \\ & & & \sigma_D^2 \end{bmatrix}$$

$$\text{dist}(x, y) = \sqrt{\begin{bmatrix} x_1 - y_1 & x_2 - y_2 & \dots & x_D - y_D \end{bmatrix} \begin{bmatrix} \frac{1}{\sigma_1^2} & & \\ & \frac{1}{\sigma_2^2} & \\ & & \ddots \\ & & & \frac{1}{\sigma_D^2} \end{bmatrix} \begin{bmatrix} x_1 - y_1 \\ x_2 - y_2 \\ \vdots \\ x_D - y_D \end{bmatrix}}$$

$$= \sqrt{\frac{(x_1 - y_1)^2}{\sigma_1^2} + \frac{(x_2 - y_2)^2}{\sigma_2^2} + \dots + \frac{(x_D - y_D)^2}{\sigma_D^2}}$$

$$= \sqrt{\sum_{j=1}^D \frac{(x_j - y_j)^2}{\sigma_j^2}} \quad \leftarrow Y = \frac{X - \mu}{\sigma}$$

假设 D=2

$$\text{dist}(x, y)^2 = \frac{(x_1 - y_1)^2}{\sigma_1^2} + \frac{(x_2 - y_2)^2}{\sigma_2^2}$$

# 马氏距离

马氏距离 ← 统计距离  $\Sigma = V^T \Lambda V$


$$\text{dist}(X, Y) = \sqrt{(X - Y)^T \Sigma^{-1} (X - Y)}$$
$$\Delta^2 = (X - Y)^T \Sigma^{-1} (X - Y)$$
$$= [(X - Y)^T \sqrt{\Sigma^{-1}}] [\sqrt{\Sigma^{-1}} (X - Y)]$$
$$= [\underbrace{\Lambda^{-\frac{1}{2}} V}_{\text{缩放}} \underbrace{V(X - Y)}_{\substack{\text{旋转} \\ \text{中心化}}}] [\underbrace{\Lambda^{-\frac{1}{2}} V}_{\text{缩放}} (X - Y)]$$

如果协方差矩阵为**单位矩阵**，马哈拉诺比斯距离就简化为欧氏距离；如果协方差矩阵为**对角阵**，其也可称为正规化的欧氏距离

马氏距离也可以定义为两个服从同一分布并且其协方差矩阵为 $\Sigma$ 的随机变量 $X$ 和 $Y$ 的差异程度

物理意义就是在规范化的主成分空间中的欧氏距离

# 指数族分布



指数族分布:

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

归一化:

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

标题隐含的信息:

1. 概率分布可以通过某些特定的解析式进行分类, 用族来称呼
2. 指数族分布最重要的项是指数项
3. 能够被分到一个类别当中说明他们一定具备某些共性



# MLE&充分统计量

## 指数族分布的性质

目标: 估计  $\eta$  参数 (MLE)

求偏导:  $\frac{\partial}{\partial \eta} \int h(x) \exp\{\eta^T \mu(x)\} dx = 0$

对  $\mu$  偏导  $\frac{\partial \mu}{\partial \eta}$

梯度

$\frac{\partial}{\partial \eta} \int h(x) \exp\{\eta^T \mu(x)\} dx + \int h(x) \exp\{\eta^T \mu(x)\} \cdot \mu(x) dx = 0 \leftarrow$  乘法求导

因为归一化  $\int g(\eta) h(x) \exp\{\eta^T \mu(x)\} dx = 1 \Rightarrow \int h(x) \exp\{\eta^T \mu(x)\} dx = \frac{1}{g(\eta)}$  代入上式

$-\frac{\nabla g(\eta)}{g(\eta)} = \int \boxed{g(\eta) h(x) \exp\{\eta^T \mu(x)\}} \mu(x) dx \Rightarrow \frac{\partial \ln(g(\eta))}{\partial \eta} = \frac{\nabla g(\eta)}{g(\eta)}$

$-\frac{\nabla g(\eta)}{g(\eta)} = E[\mu(x)] \uparrow \text{概率密度 } p(\mu(x)) \therefore -\nabla \ln g(\eta) = E[\mu(x)]$

最大似然估计

$p(x|\eta) = \prod_{n=1}^N h(x_n) g(\eta)^N \exp\left\{\eta^T \sum_{n=1}^N \mu(x_n)\right\}$

$\ln p(x|\eta) = \sum_{n=1}^N \ln h(x_n) + N \ln g(\eta) + \eta^T \sum_{n=1}^N \mu(x_n)$

$\frac{\partial \ln p(x|\eta)}{\partial \eta} = N \frac{\nabla g(\eta)}{g(\eta)} + \sum_{n=1}^N \mu(x_n)$

令导数为0

$-\frac{\nabla g(\eta)}{g(\eta)} = \frac{1}{N} \sum_{n=1}^N \mu(x_n)$

$-\ln g(\eta) = \frac{1}{N} \sum_{n=1}^N \ln h(x_n)$

充分统计量

两个充分统计量

① 样本和

② 样本平方和

N阶矩的性质 (补充)



We have already encountered the concept of a conjugate prior several times, for example in the context of the Bernoulli distribution (for which the conjugate prior is the beta distribution) or the Gaussian (where the conjugate prior for the mean is a Gaussian, and the conjugate prior for the precision is the Wishart distribution). In general, for a given probability distribution  $p(\mathbf{x}|\boldsymbol{\eta})$ , we can seek a prior  $p(\boldsymbol{\eta})$  that is conjugate to the likelihood function, so that the posterior distribution has the same functional form as the prior. For any member of the exponential family (2.194), there exists a conjugate prior that can be written in the form

$$p(\boldsymbol{\eta}|\boldsymbol{\chi}, \nu) = f(\boldsymbol{\chi}, \nu) g(\boldsymbol{\eta})^\nu \exp \{ \nu \boldsymbol{\eta}^T \boldsymbol{\chi} \} \quad (2.229)$$

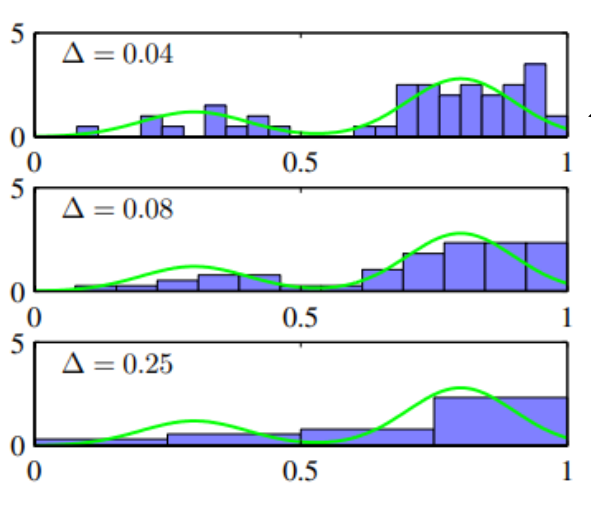
where  $f(\boldsymbol{\chi}, \nu)$  is a normalization coefficient, and  $g(\boldsymbol{\eta})$  is the same function as appears in (2.194). To see that this is indeed conjugate, let us multiply the prior (2.229) by the likelihood function (2.227) to obtain the posterior distribution, up to a normalization coefficient, in the form

$$p(\boldsymbol{\eta}|\mathbf{X}, \boldsymbol{\chi}, \nu) \propto g(\boldsymbol{\eta})^{\nu+N} \exp \left\{ \boldsymbol{\eta}^T \left( \sum_{n=1}^N \mathbf{u}(\mathbf{x}_n) + \nu \boldsymbol{\chi} \right) \right\}. \quad (2.230)$$

This again takes the same functional form as the prior (2.229), confirming conjugacy. Furthermore, we see that the parameter  $\nu$  can be interpreted as a effective number of pseudo-observations in the prior, each of which has a value for the sufficient statistic  $\mathbf{u}(\mathbf{x})$  given by  $\boldsymbol{\chi}$ .

# 核密度估计

核密度估计实际上就是寻找合适的随机变量概率密度函数，使其尽可能贴近样本的数据分布情况



统计频数

缺点:不够平滑

核密度估计需要指定一个核函数来描述每一个样本点

