数字特征

特征化是人们压缩数据的一种方式，它能够反映一些群体的某方面的特点，它能够反映一些群体的某方面的特点。

数学期望

数学期望(mean)（或均值，亦简称期望）是试验中每次可能结果的概率乘以其结果的总和。它反映随机变量平均取值的大小。其公式如下：

$E(X)=\sum_{k=1}^{\infty} x_k * p_k$

数学期望反映的是平均水平。通过它，我们能够了解一个群体的平均水平.它所包含的信息也是十分有限的，首先是个体信息被压缩了，其次如果单纯看期望的话，是看不出样本的数量。

方差

方差是衡量随机变量或一组数据时离散程度的度量。概率论中方差用来度量随机变量和其数学期望（即均值）之间的偏离程度。计算公式如下：

$D(X)=Var(X)=E\{[X-E(X)]^2\}$

公式逐步解释：$[X-E(X)] \to [X-E(X)]^2 \to E\{[X-E(X)]^2\}$

$[X-E(X)]$是计算随机变量中各个值与期望的距离（反映的是以E(X)为基准计算的偏差）。但是只是将偏差进行求和，可能导致结果为0的情况（会产生离散程度较高，评价却为0的情况）。

$[X-E(X)]^2$可避免上述情况发生，但问题依据存在，不同的随机变量(比如，X,Y)之间在此级别是无法进行比较的，因为X,Y的数量空间是不同的（X可能有3个值，Y可能有1000个值），进而导致不具有可比性。

$E\{[X-E(X)]^2\}$则是将数量空间进行了统一，使得不同随机变量的方差具有了可比性。

协方差

期望与方差都是考察单个随机变量（1维）,但是事实上当我们考察一个群体的时候，往往事物的属性是多方面的（多维），这里只考察2维情况，形式如：(X,Y)。

(X,Y)的意思这类事物具有两个方面的属性，更进一步来说，一个样本有X,Y两方面的值，体现在数据库中，有两列（X列，Y列）。当Ｘ，Ｙ这两个属性出现在同一类事物中的时候，我们很自然想到Ｘ，Ｙ之间有某种关系，但是如何来刻画这种关系呢，这就是协方差所讨论的。

当样本含有大量维度（随机变量多）的时候，我们就需要使用矩阵来刻画各个维度之间的关联关系。

(X,Y) 是2维的，只考虑1维会无法从整体把握问题。而如果进行关联分析，有时候却需要对维度拆分来进行研究。协方差公式：

$$Cov(X,Y) = E\{[X - E(X)][Y - E(Y)]\}$$

[X−E(X)] 与[Y−E(Y)]都只考虑了各自随机变量这1维，通过相乘的方式使得上面两个离差建立起数值关系，[X−E(X)][Y−E(Y)]是两者共同作用的结果，即和X，Y都有关。又因为X,Y都是随机变量，所以自然[X−E(X)][Y−E(Y)]也是合成的新的随机变量。

根据相关性定义可知，如果X,Y独立，那么$[X - E(X)]$与$[Y - E(Y)]$也是独立的，那么

$$\because \text{随机变量} X, Y \text{相互独立（即，} P(X,Y) = P(X) * P(Y)\text{）}$$

$$\therefore Cov(X,Y) = E\{[X - E(X)][Y - E(Y)]\}$$

$$= E[X - E(X)] * E(Y - E(Y))$$

$$\because E[X - E(X)] = 0 \text{且} E(Y - E(Y)) = 0$$

$$\therefore X, Y \text{相互独立} => Cov(X,Y) = 0$$

如果X,Y有关系，那么关联性会使得某个变量的随机性不再那么随机。即，假如说X是随机的，X的值确定后会限定Y的随机性（将Y限定在某个范围）。这里举个简单的例子，假如学生具有（年龄，年级）两个属性，如果年龄是17岁，那么年级范围很可能是在高中范围内。年龄这个变量影响着年级这个变量。

如果X,Y有关系，从关系传递性角度来说，离差[X−E(X)]与[Y−E(Y)]也会有一定的关系。正常情况下随机变量[X−E(X)]与[Y−E(Y)]会在0水平附近波动，如果上述两个随机变量无关，那么两个随机变量的相乘的方式会在0附近波动（即Cov(X,Y)=0）;如果X,Y有关，那么[X−E(X)]∗[Y−E(Y)]波动范围将会受到影响，不再围绕0。

协方差计算公式

$$Cov(X,Y) = E\{[X-E(X)][Y-E(Y)]\}$$

$$Cov(X,Y) = E(XY) - E(X)E(Y)$$

$$D(X+Y) = D(X) + D(Y) + 2Cov(X,Y)$$

协方差性质

$$1^\circ \ Cov(aX,bY) = abCov(X,Y)$$

$$2^\circ \ Cov(X_1+X_2,Y) = Cov(X_1,Y) + Cov(X_2,Y)$$

(X,Y)是2元组，X,Y 共同出现，可能有关系。为度量这种相关性，制定了一个指标（协方差），来刻画X,Y之间关系。（将相关性映射到协方差）

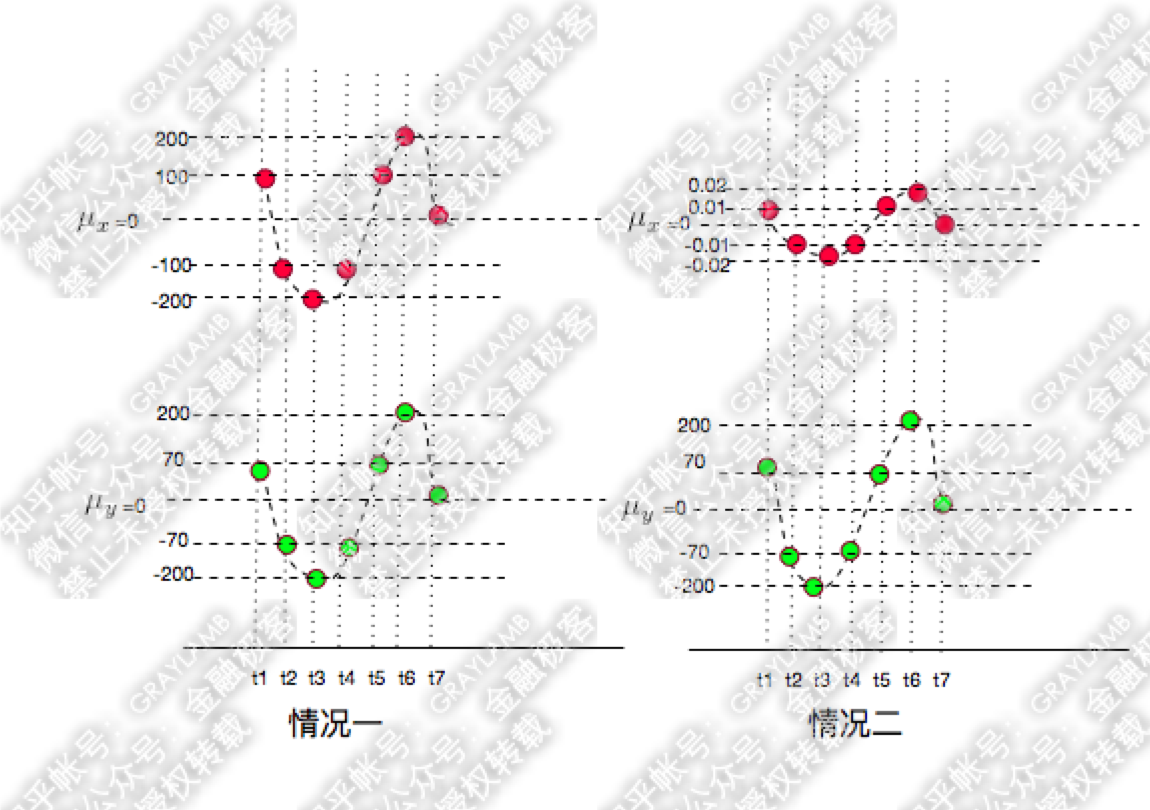
相关系数

相关系数是对协方差进行了归一化处理，使其区间处于[－１，１]范围内。

相关系数也可以看成协方差：一种剔除了两个变量量纲影响、标准化后的特殊协方差。它反映了两个变量变化是是同向的还是反向的，如果同向变化就为正，反向变化就为负。他消除了两个变量变化幅度的影响，而只是单纯反应两个变量每单位变化时的相似程度。

比较抽象，下面还是举个例子来说明：

首先，还是承接上文中的变量X、Y变化的示意图（X为红点，Y为绿点），来看两种情况：



很容易就可以看出以上两种情况X，Y都是同向变化的，而这个“同向变化”，有个非常显著特征：**X、Y同向变化的过程，具有极高的相似度！**无论第一还是第二种情况下，都是：t1时刻X、Y都大于均值，t2时刻X、Y都变小且小于均值，t3时刻X、Y继续变小且小于均值，t4时刻X、Y变大但仍小于均值，t5时刻X、Y变大且大于均值……

可是，计算一下他们的协方差，

第一种情况下：

$$[(100-0)\times(70-0)+(-100-0)\times(-70-0)+(-200-0)\times(-200-0)\ldots]\div 7\approx 15428.57$$

第二种情况下：

$$[(0.01-0)\times(70-0)+(-0.01-0)\times(-70-0)+(-0.02-0)\times(-200-0)\ldots]\div 7\approx 1.542857$$

协方差差出了一万倍，只能从两个协方差都是正数判断出两种情况下X、Y都是同向变化，但是，一点也看不出两种情况下X、Y的变化都具有相似性这一特点。

这是为什么呢？

因为以上两种情况下，在X、Y两个变量同向变化时，X变化的幅度不同，这样，两种情况的协方差更多的被变量的变化幅度所影响了。

所以，为了能够准确的研究两个变量在变化过程中的相似程度，我们就要把变化幅度对协方差的影响从协方差中剔除掉。

下面看看相关系数ρ_{XY}的计算公式：

$$\rho_{XY} = \frac{Cov(X,\ Y)}{\sqrt{D(X)} * \sqrt{D(Y)}}$$

$$其中, Cov(X,Y) = E\{[X - E(X)][Y - E(Y)]\}$$

定理

$$1^\circ\ |\rho_{XY}| \le 1$$

$$2^\circ\ |\rho_{XY}| = 1 的充要条件是，存在常数 a, b，使得$$

$$P\{Y = a + bX\} = 1$$

$$(2^\circ\ 的含义：Y可完全用随机变量X线性表示。X确定，Y唯一确定)$$

需要注意的一些事情

- **【线性】**ρ_{XY}表示的是X,Y之间**线性相关程度。**（不适用于多次方，指数等）
- $\rho_{XY} = 0$，我们称X,Y不相关。
- **【独立，相关】**X, Y相互独立 $=> \rho_{XY} = 0$
- **【独立，相关】**X,Y相互独立，则$\rho_{XY} = 0$;$\rho_{XY} = 0$,不能推出X,Y相互独立。（$\rho_{XY} = 0$*只能说明非"线性相关"，但X,Y可能是"非线性"相关*）

> 协方差除以标准差，也就是把协方差中变量变化幅度对协防擦汗的影响剔除掉，从而标准化协方差，反应的也就是两个变量每单位变化时的情况。

我们暂且先做出一个简单的假设：X,Y完全线性相关,

$$设: Y = a * X + b, a、b都不为0$$

$$将上式带入Cov(X,Y)得：$$

$$分子：Cov(X,Y) = E(XY) - E(X)E(Y)$$

$$= E[X * (a * X + b)] - E(X)E(a * X + b)$$

$$= aD(X)$$

$$另一方面，分母：\sqrt{D(X)} * \sqrt{D(Y)}$$

$$= \sqrt{D(X)} * \sqrt{D(a * X + b)}$$

$$= |a|D(X)$$

$$所以在线性相关的前提下，导致了相关系数只与a的符号相关。$$

再接下来，让我们放开那个非常强的假设（完全线性相关在现实生活中几乎不太可能存在，总会有些干扰的），去掉“完全”这个假设，留下“线性”这个假设。这里只是定性的分析下，定量的证明请参考数学书。 分母这里认为是正的，那么这里先只考虑分子的正负。

假如X,Y线性相关，接下来看看会对 $Cov(X,Y) = E(XY) - E(X)E(Y)$造成什么影响。
这里我们设X是自由的，那么X确定之后，则限定了Y的自由活动的空间（见前面年龄、年级的例子），即Y不再自由了。造成的后果是

$$在E(XY)中Y被限制住了，（因为这两个同时出现，构成了新的随机变量）$$

$$而在E(Y)中Y没有被限制住。$$

$$于是，Cov(X,Y) = E(XY) - E(X)E(Y)$$

$$Cov(X,Y) = E(X * a * X + 干扰因子) - E(X)E(a * X + 干扰因子),$$

$$假设干扰因子是随机的，此处我们暂且忽略。$$

$$于是, Cov(X,Y) = aE(X^2) + aE(X)^2 = a\sqrt{D(X)}$$

$$所以，相关系数的正负和正负线性相关性有很大的关联性。$$

协方差矩阵

当样本含有大量维度（随机变量多）的时候，我们就需要使用矩阵来刻画各个维度之间的关联关系。

协方差矩阵的特点：

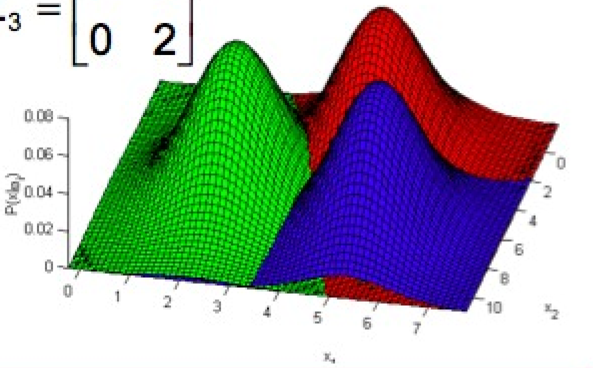
- 对角线元素(i,i)为数据第 i 维的方差。
- 非对角线元素(i,j)为第 i 维和第 j 维的协方差。
- 协方差矩阵是对称阵

协方差矩阵取值对于图形形状的影响：

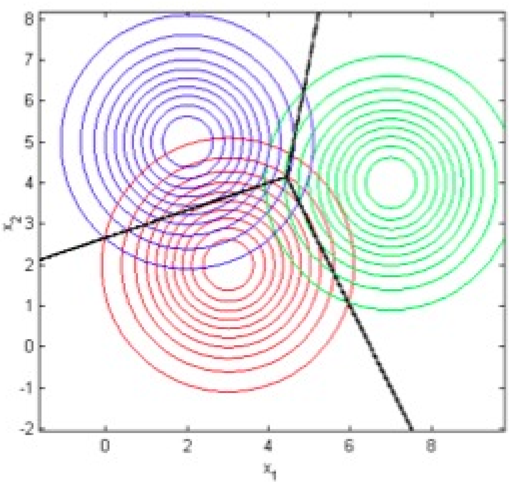
- 均值为分布的中心点位置。
- 对角线元素决定了分布图形是圆还是扁。
- 非对角线元素决定了分布图形的轴向（扁的方向）。

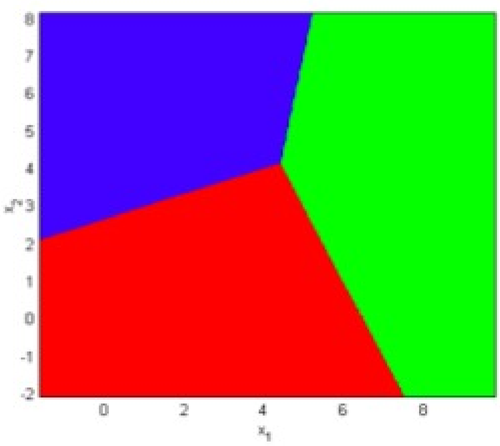
$$\mu_1 = [3 \quad 2]^T \quad \mu_2 = [7 \quad 4]^T \quad \mu_3 = [2 \quad 5]^T$$

$$\Sigma_1 = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix} \quad \Sigma_2 = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix} \quad \Sigma_3 = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$$



Case 1: $\Sigma_i = \sigma^2 I$, *example*

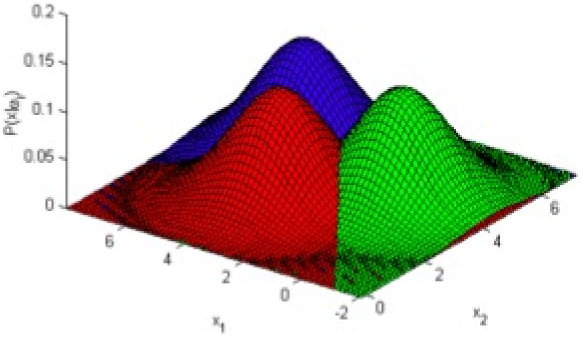




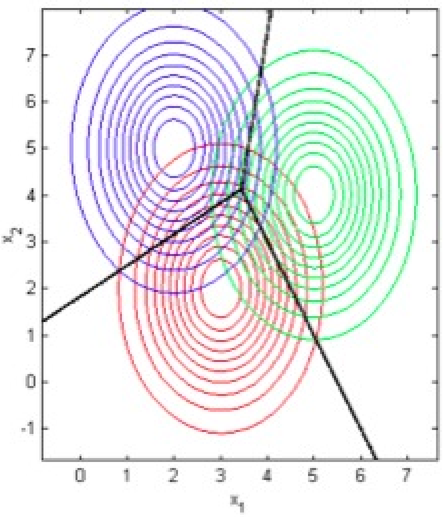
三个协方差矩阵相同，都为对角阵，对角线元素相同

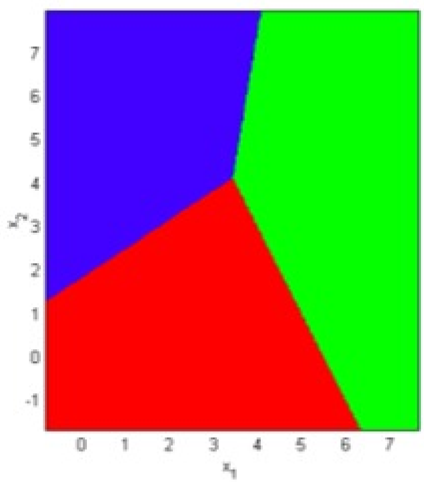
$$\mu_1 = [3 \quad 2]^T \quad \mu_2 = [5 \quad 4]^T \quad \mu_3 = [2 \quad 5]^T$$

$$\Sigma_1 = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix} \quad \Sigma_2 = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix} \quad \Sigma_3 = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}$$



Case 2: $\Sigma_i = \Sigma$ (Σ diagonal), *example*

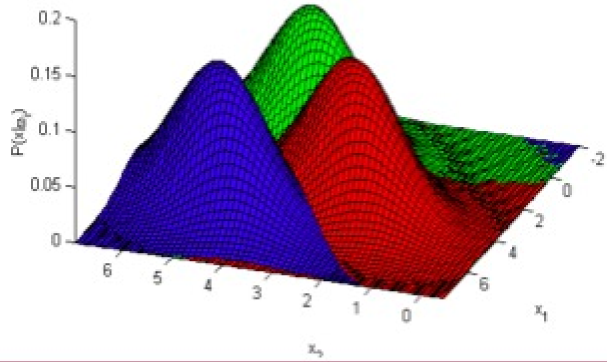




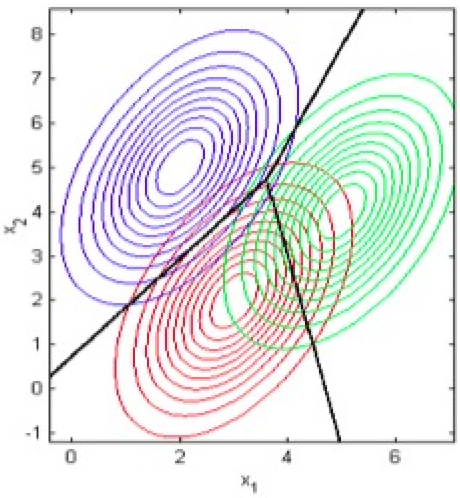
三个协方差矩阵相同，都为对角阵，对角线元素不同

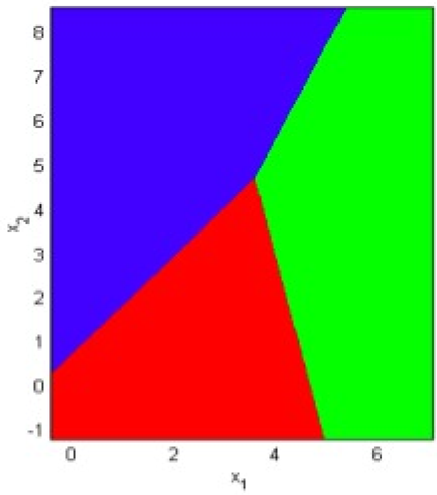
$\mu_1 = [3 \quad 2]^T \qquad \mu_2 = [5 \quad 4]^T \qquad \mu_3 = [2 \quad 5]^T$

$$\Sigma_1 = \begin{bmatrix} 1 & 0.7 \\ 0.7 & 2 \end{bmatrix} \quad \Sigma_2 = \begin{bmatrix} 1 & 0.7 \\ 0.7 & 2 \end{bmatrix} \quad \Sigma_3 = \begin{bmatrix} 1 & 0.7 \\ 0.7 & 2 \end{bmatrix}$$



Case 3: Σ=Σ (Σ non-diagonal), example

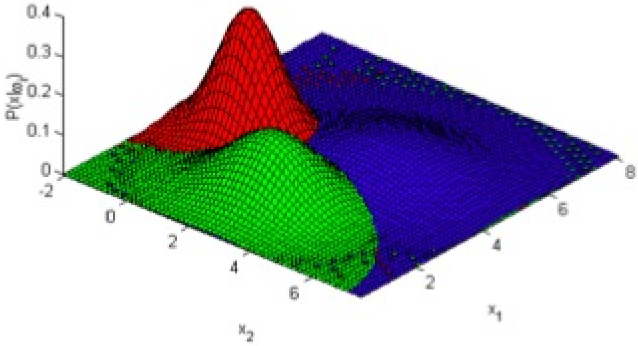




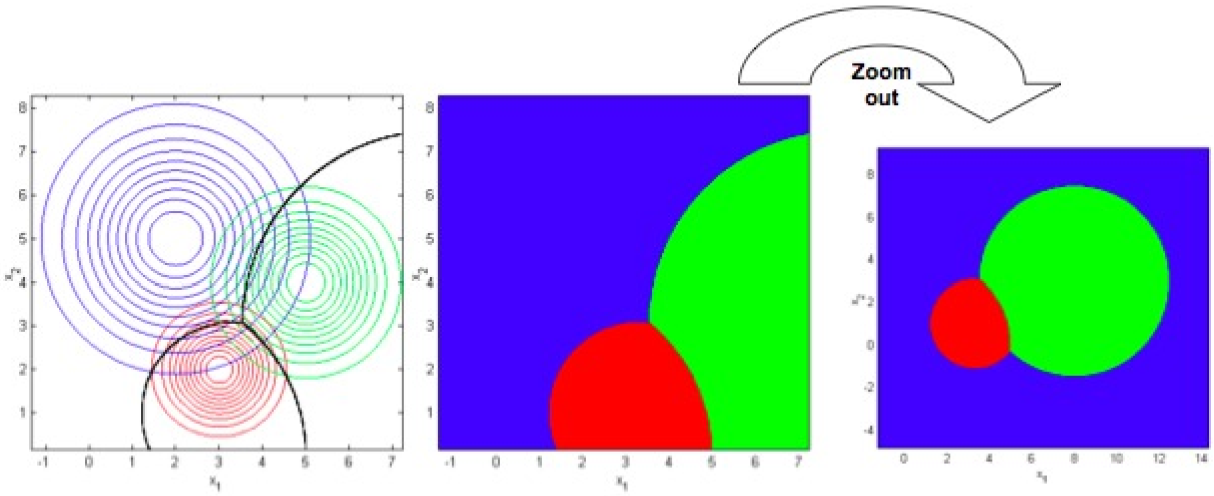
三个协方差矩阵相同，不是对角阵，对角线元素不同

$$\mu_1 = [3 \quad 2]^T \qquad \mu_2 = [5 \quad 4]^T \qquad \mu_3 = [2 \quad 5]^T$$

$$\Sigma_1 = \begin{bmatrix} 0.5 & 0 \\ 0 & 0.5 \end{bmatrix} \quad \Sigma_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \quad \Sigma_3 = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$$



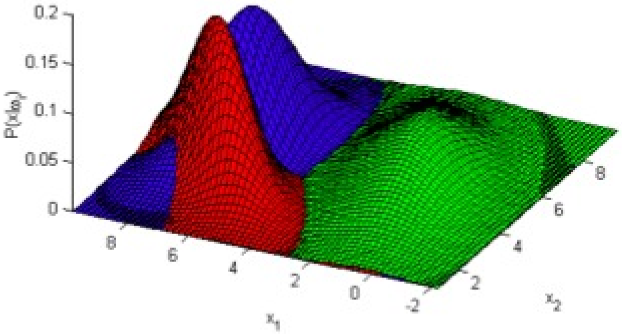
Case 4: $\Sigma_i=\sigma_i^2 I$, example



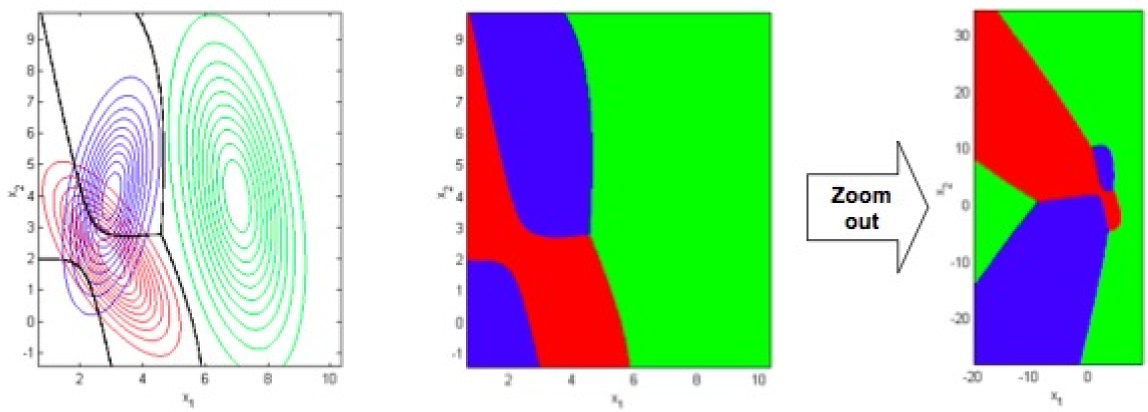
三个协方差矩阵不同，都是对角阵，对角线元素相同

$$\mu_1 = [3 \quad 2]^T \qquad \mu_2 = [5 \quad 4]^T \qquad \mu_3 = [2 \quad 5]^T$$

$$\Sigma_1 = \begin{bmatrix} 1 & -1 \\ -1 & 2 \end{bmatrix} \qquad \Sigma_2 = \begin{bmatrix} 1 & -1 \\ -1 & 7 \end{bmatrix} \qquad \Sigma_3 = \begin{bmatrix} 0.5 & 0.5 \\ 0.5 & 3 \end{bmatrix}$$



Case 5: $\Sigma_i \neq \Sigma_j$ general case, example



三个协方差矩阵不同，不是对角阵，对角线元素不同