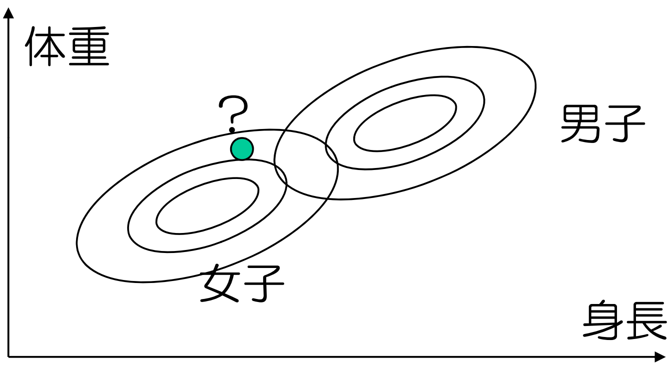
# 判別分析

## 判別分析とは？

２つ以上の群（クラスとも呼ぶ）が存在し，それぞれの群の観測値の統計的性質（平均や共分散）がわかっているとする.
この条件のもとで，あるテストサンプルの観測値から，そのサンプルがどちらの群に属するか，判別したい.

例）男子学生の身長，体重，女子学生の身長，体重の統計的分布がわかっているとする．このとき，あるテスト学生の身長，体重からその学生の性別を推定したい.






# 講義内容

マハラノビス距離
１．１変数の場合
２．相関のない２変数の場合
３．相関のある２変数の場合
４．一般的な表現

判別分析
１．１変数の場合
２．２変数の場合

# マハラノビス距離 －１変数の場合－



平均m，分散$\sigma^2$の正規母集団N(m，$\sigma^2$)からサンプルを1つ取り出し，その値がxであったとする．
このとき，このサンプルと，母集団平均mとの“基準化した距離”は，
基準化後の変数，

$$v = (x - m)/\sigma$$

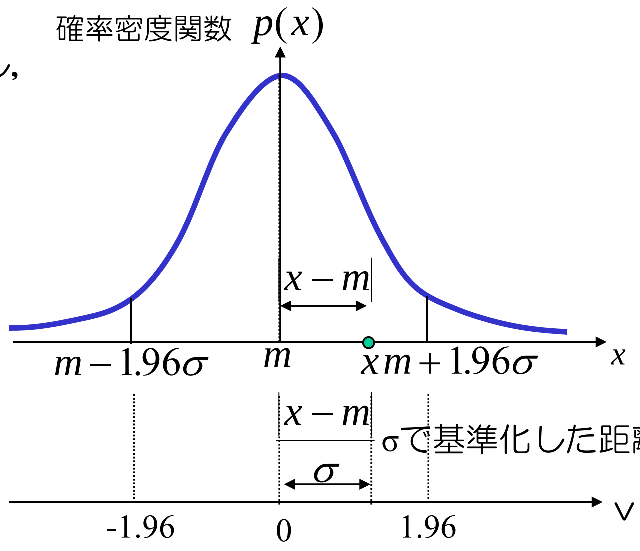
の絶対値，

$$d_M \equiv |u| = |x - m|/\sigma$$

で与えられる．また，平方距離は

$$d_M^2 = (x - m)^2/\sigma^2$$

で定義される．




このように，標準偏差σで補正した中心からの距離$d_M$をマハラノビス距離という

---

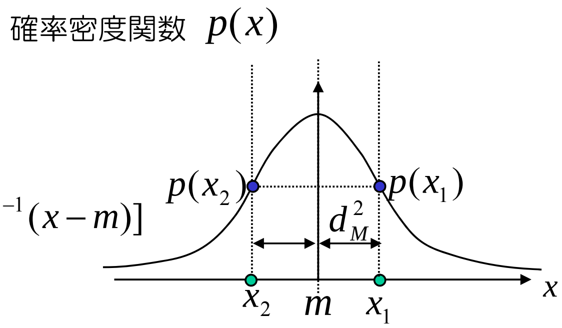
Mahalanobis：インドの数学者

# マハラノビス距離と正規分布



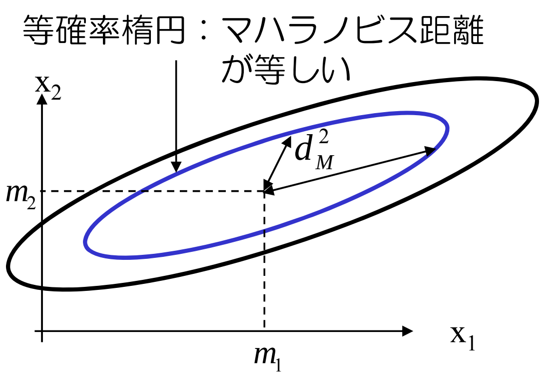
**１次元正規分布：**

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma}\exp[-\frac{1}{2}\overbrace{\frac{(x-m)^2}{\sigma^2}}^{d_M^2}]$$
$$= (2\pi)^{-1/2}(\sigma^2)^{-1/2}\exp[-\frac{1}{2}(x-m)(\sigma^2)^{-1}(x-m)]$$




**２次元正規分布：**

$$p(\mathbf{x}) = (2\pi)^{-1}|\mathbf{C}|^{-\frac{1}{2}}\exp[-\frac{1}{2}\underbrace{(\mathbf{x}-\mathbf{m})^T\mathbf{C}^{-1}(\mathbf{x}-\mathbf{m})}_{d_M^2}]$$

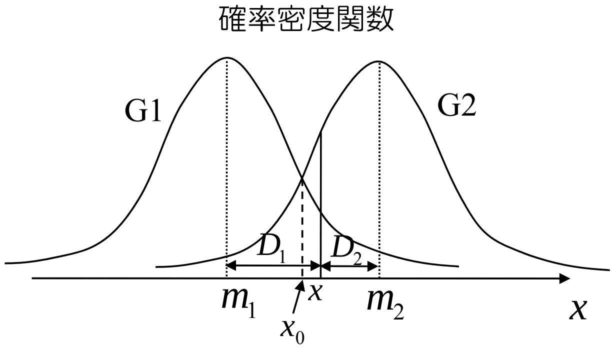



**多次元正規分布：**

$$p(\mathbf{x}) = (2\pi)^{-\frac{d}{2}}|\mathbf{C}|^{-\frac{1}{2}}\exp[-\frac{1}{2}\underbrace{(\mathbf{x}-\mathbf{m})^T\mathbf{C}^{-1}(\mathbf{x}-\mathbf{m})}_{d_M^2}]$$

# １変数による判別

２群の判別を考える．
仮定：分散σ²の値は同じ




$$D_1^2 = \frac{(x-m_1)^2}{\sigma^2} \qquad D_2^2 = \frac{(x-m_2)^2}{\sigma^2}$$

母平均からの距離によって判別を行う．
すなわち，

$$\begin{cases} x \in G_2 & if \;\; D_1^2 \geq D_2^2 \\ x \in G_1 & if \;\; D_1^2 < D_2^2 \end{cases}$$

$$D_1^2 - D_2^2 = \frac{1}{\sigma^2}\{(x-m_1)^2 - (x-m_2)^2\}$$
$$= \frac{1}{\sigma^2}\{-2(m_1-m_2) + (m_1^2 - m_2^2)\}$$
$$= \frac{-2}{\sigma^2}\boxed{(m_1-m_2)}(x - \boxed{\frac{m_1+m_2}{2}})$$

$x$によらない定数　　$x_0$

であるから，たとえば

$m_1 < m_2$　の場合

$$if \;\; x \geq x_0 = \frac{m_1+m_2}{2}, \; then \;\; x \in G_2$$

$$if \;\; x < x_0 = \frac{m_1+m_2}{2}, \; then \;\; x \in G_1$$

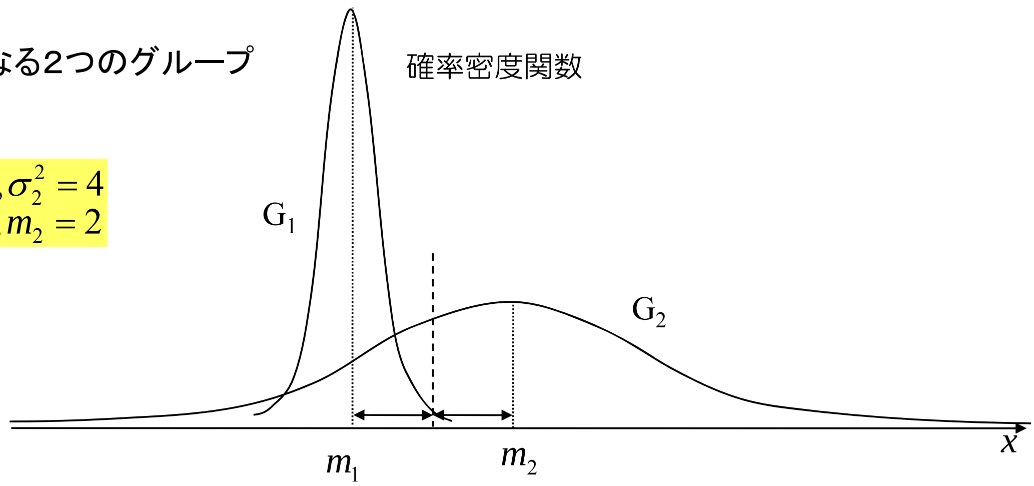
すなわち，２つの母平均の中点のどちら側に
ｘがあるかで，属する群が決まる．



# １変数による判別（分散が異なる場合）

平均と分散の異なる2つのグループ
の判別を考える．

分散 : $\sigma_1^2 = 1, \sigma_2^2 = 4$
平均 : $m_1 = 0, m_2 = 2$

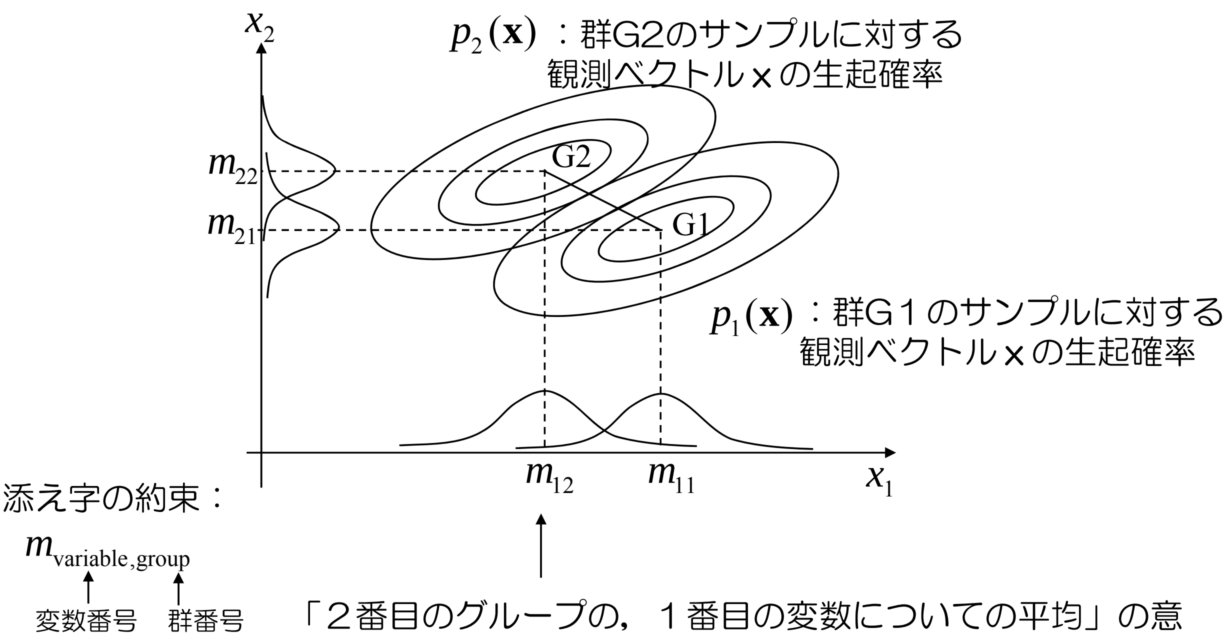



課題1：マハラノビス距離を用いて，グループ分けの境界となる点（x座標）を求めなさい．

課題2：データの母集団は与えられた平均，分散をもつ正規確率分布に従うものとする．確率密度関数の大小によって，判別の境界を求めた場合，マハラノビス距離を用いた場合と比べて，どのようになるか考えなさい．

# ２変数による判別

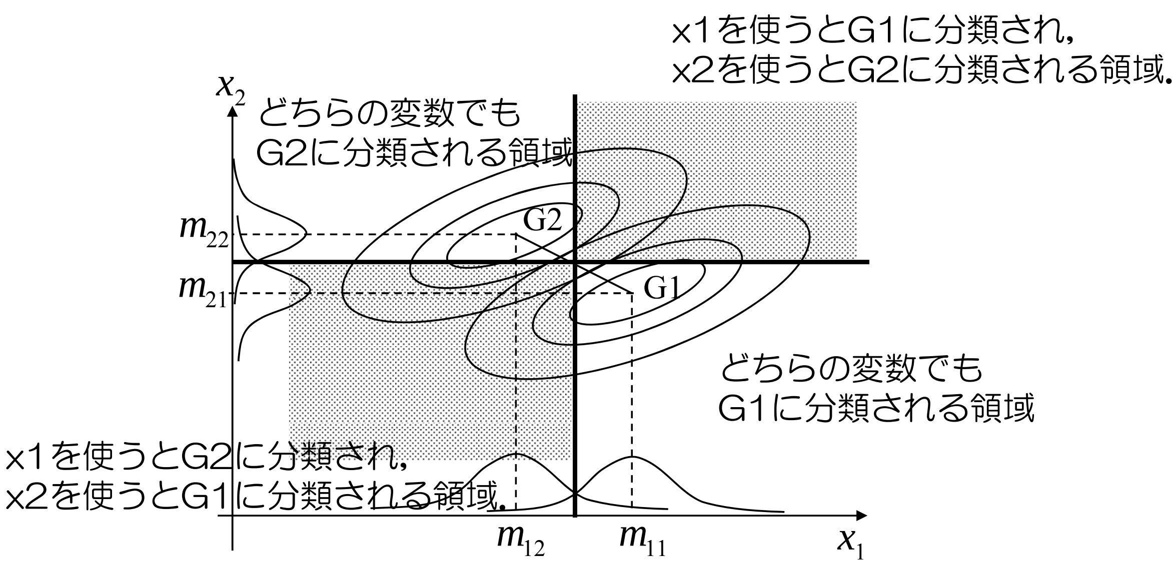
共分散行列が等しく，平均ベクトルが異なる
２つの群,G1,G2を判別する式を求めたい.






# ２変数による判別－１変数のみによる判別での問題点－

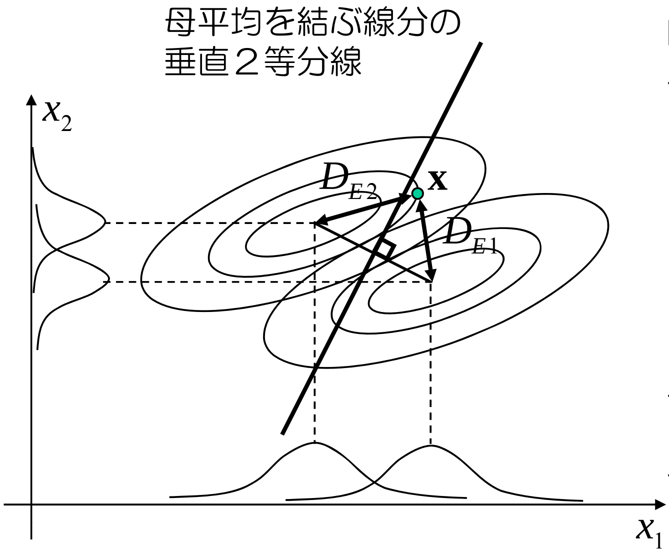
仮にいずれか一方の変数のみで判別すると．．．




の領域では分類に矛盾が生じる

# ２変数による判別－単純なユークリッド距離では？－

単純なユークリッド距離で判別することにすると．．．




図において，観測値ベクトルxの点は
ユークリッド距離では

$$D_{E1} < D_{E2}$$

となり，より近いG1に分類される．
しかし，実際の生起確率は

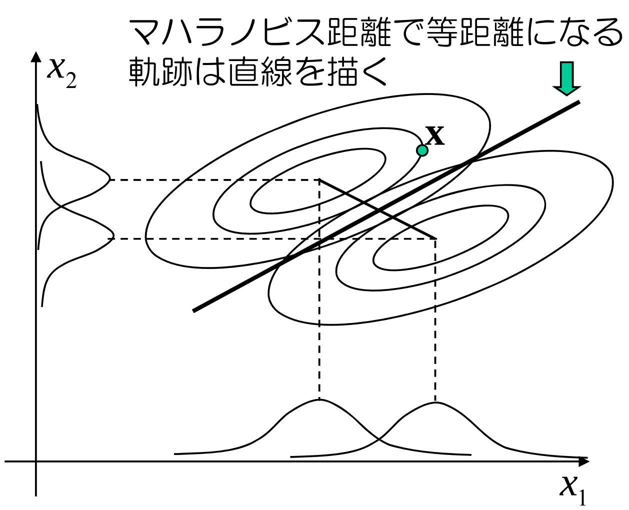
$$p_2(\mathbf{x}) > p_1(\mathbf{x})$$

であり，G2に分類すべきである．
このように，単純なユークリッド距離では望ましい判別ができない．



# ２変数による判別－マハラノビス距離による判別－

そこで，等生起確率を与えるマハラノビス距離で判別する




$$p_1(\mathbf{x}) = (2\pi)^{-1}|\mathbf{C}|^{-\frac{1}{2}} \exp[-\frac{1}{2}(\mathbf{x}-\mathbf{m}_1)^T \mathbf{C}^{-1}(\mathbf{x}-\mathbf{m}_1)]$$

$$p_2(\mathbf{x}) = (2\pi)^{-1}|\mathbf{C}|^{-\frac{1}{2}} \exp[-\frac{1}{2}(\mathbf{x}-\mathbf{m}_2)^T \mathbf{C}^{-1}(\mathbf{x}-\mathbf{m}_2)]$$

直線の方程式は

$$\delta_1 = m_{11} - m_{21}$$

$$\delta_2 = m_{12} - m_{22}$$

$$\overline{m}_1 = \frac{m_{11} + m_{12}}{2}$$

$$\overline{m}_2 = \frac{m_{21} + m_{22}}{2}$$

として，

$$f(x_1, x_2) \equiv (\sigma_{11}\delta_1 + \sigma_{12}\delta_2)(x_1 - \overline{m}_1) + (\sigma_{21}\delta_1 + \sigma_{22}\delta_2)(x_2 - \overline{m}_2) = 0$$

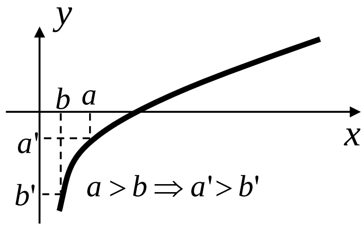
と書ける．

# 生起確率の大小とマハラノビス距離の関係

２つの群を仮定したときの確率の大小を比較する.

$$p_1(\mathbf{x}) = (2\pi)^{-1}|\mathbf{C}|^{-\frac{1}{2}} \exp[-\frac{1}{2}(\mathbf{x}-\mathbf{m}_1)^T \mathbf{C}^{-1}(\mathbf{x}-\mathbf{m}_1)]$$

$$p_2(\mathbf{x}) = (2\pi)^{-1}|\mathbf{C}|^{-\frac{1}{2}} \exp[-\frac{1}{2}(\mathbf{x}-\mathbf{m}_2)^T \mathbf{C}^{-1}(\mathbf{x}-\mathbf{m}_2)]$$




比をとって比較

$$\frac{p_1(\mathbf{x})}{p_2(\mathbf{x})} = \frac{(2\pi)^{-1}|\mathbf{C}|^{-\frac{1}{2}} \exp[-\frac{1}{2}(\mathbf{x}-\mathbf{m}_1)^T \mathbf{C}^{-1}(\mathbf{x}-\mathbf{m}_1)]}{(2\pi)^{-1}|\mathbf{C}|^{-\frac{1}{2}} \exp[-\frac{1}{2}(\mathbf{x}-\mathbf{m}_2)^T \mathbf{C}^{-1}(\mathbf{x}-\mathbf{m}_2)]}$$

$$= \exp[-\frac{1}{2}\{(\mathbf{x}-\mathbf{m}_1)^T \mathbf{C}^{-1}(\mathbf{x}-\mathbf{m}_1) - (\mathbf{x}-\mathbf{m}_2)^T \mathbf{C}^{-1}(\mathbf{x}-\mathbf{m}_2)\}]$$

$$\frac{p_1(\mathbf{x})}{p_2(\mathbf{x})} > 1 \Leftrightarrow \mathbf{x} \in G_1$$

$$\frac{p_1(\mathbf{x})}{p_2(\mathbf{x})} < 1 \Leftrightarrow \mathbf{x} \in G_2$$

対数をとって比較

$$\log \frac{p_1(\mathbf{x})}{p_2(\mathbf{x})} = -\frac{1}{2}\{(\mathbf{x}-\mathbf{m}_1)^T \mathbf{C}^{-1}(\mathbf{x}-\mathbf{m}_1) - (\mathbf{x}-\mathbf{m}_2)^T \mathbf{C}^{-1}(\mathbf{x}-\mathbf{m}_2)\}$$

$$= -\frac{1}{2}(D_{M,1}^2 - D_{M,2}^2)$$

$$\log \frac{p_1(\mathbf{x})}{p_2(\mathbf{x})} = -\frac{1}{2}(D_{M,1}^2 - D_{M,2}^2) > 0 \Leftrightarrow \mathbf{x} \in G_1$$

$$\log \frac{p_1(\mathbf{x})}{p_2(\mathbf{x})} = -\frac{1}{2}(D_{M,1}^2 - D_{M,2}^2) < 0 \Leftrightarrow \mathbf{x} \in G_2$$

⇨ 生起確率の大小がマハラノビス距離の大小に対応づけられた.



# 判別関数の表現

判別関数を２つの群の平均までのマハラノビス距離の差で定義する.

$$f(\mathbf{x}) \equiv D_2^2(\mathbf{x}) - D_1^2(\mathbf{x})$$

ただし

$$\begin{cases} D_1^2(\mathbf{x}) = (\mathbf{x}-\mathbf{m}_1)^T \mathbf{C}^{-1}(\mathbf{x}-\mathbf{m}_1) \\ D_2^2(\mathbf{x}) = (\mathbf{x}-\mathbf{m}_2)^T \mathbf{C}^{-1}(\mathbf{x}-\mathbf{m}_2) \end{cases}$$

f(**x**)を**C**,**m**$_1$,**m**$_2$を使って表すと

$$f(\mathbf{x}) = (\mathbf{x}-\mathbf{m}_2)^T \mathbf{C}^{-1}(\mathbf{x}-\mathbf{m}_2) - (\mathbf{x}-\mathbf{m}_1)^T \mathbf{C}^{-1}(\mathbf{x}-\mathbf{m}_1)$$

すべて展開

$$= \underline{\mathbf{x}^T \mathbf{C}^{-1}\mathbf{x}} - \mathbf{x}^T \mathbf{C}^{-1}\mathbf{m}_2 - \mathbf{m}_2^T \mathbf{C}^{-1}\mathbf{x} + \mathbf{m}_2^T \mathbf{C}^{-1}\mathbf{m}_2$$

$$- \{\underline{\mathbf{x}^T \mathbf{C}^{-1}\mathbf{x}} - \mathbf{x}^T \mathbf{C}^{-1}\mathbf{m}_1 - \mathbf{m}_1^T \mathbf{C}^{-1}\mathbf{x} + \mathbf{m}_1^T \mathbf{C}^{-1}\mathbf{m}_1\}$$

・下線部消える
・xでくくる

$$= \mathbf{x}^T \mathbf{C}^{-1}(\mathbf{m}_1 - \mathbf{m}_2) + (\mathbf{m}_1 - \mathbf{m}_2)^T \mathbf{C}^{-1}\mathbf{x}$$

$$- (\mathbf{m}_1^T \mathbf{C}^{-1}\mathbf{m}_1 - \mathbf{m}_2^T \mathbf{C}^{-1}\mathbf{m}_2)$$

$$= 2\mathbf{x}^T \mathbf{C}^{-1}(\mathbf{m}_1 - \mathbf{m}_2)$$

$$- \{(\mathbf{m}_1 + \mathbf{m}_2)^T \mathbf{C}^{-1}\mathbf{m}_1 - (\mathbf{m}_1 + \mathbf{m}_2)^T \mathbf{C}^{-1}\mathbf{m}_2\}$$

第１項：Cが対称行列のとき **a$^T$Cb=b$^T$Ca** を利用
第２項：テクニック

$$= 2\mathbf{x}^T \mathbf{C}^{-1}(\mathbf{m}_1 - \mathbf{m}_2)$$

$$- (\mathbf{m}_1 + \mathbf{m}_2)^T \mathbf{C}^{-1}(\mathbf{m}_1 - \mathbf{m}_2)$$

$$= 2(\mathbf{x} - \overline{\mathbf{m}})^T \mathbf{C}^{-1}(\mathbf{m}_1 - \mathbf{m}_2)$$

ただし, $\overline{\mathbf{m}} = \dfrac{\mathbf{m}_1 + \mathbf{m}_2}{2}$

# 判別関数の表現

$$f(\mathbf{x}) = 2(\mathbf{x}-\overline{\mathbf{m}})^T \mathbf{C}^{-1}(\mathbf{m}_1 - \mathbf{m}_2)$$

$$\mathbf{C} = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{bmatrix} \quad \text{とおくと，} \quad \mathbf{C}^{-1} = |\mathbf{C}|^{-1} \begin{bmatrix} \sigma_{22} & -\sigma_{12} \\ -\sigma_{12} & \sigma_{11} \end{bmatrix}$$

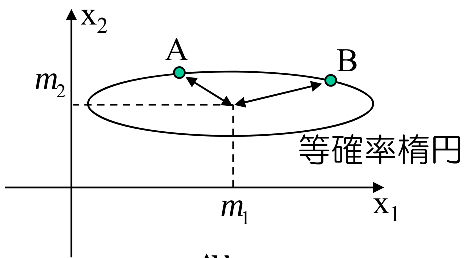
$$\begin{aligned} f(\mathbf{x}) &= 2(\mathbf{x}-\overline{\mathbf{m}})^T \mathbf{C}^{-1}(\mathbf{m}_1 - \mathbf{m}_2) \\ &= (\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - \begin{bmatrix} \overline{m}_1 \\ \overline{m}_2 \end{bmatrix})^T |\mathbf{C}|^{-1} \begin{bmatrix} \sigma_{22} & -\sigma_{12} \\ -\sigma_{12} & \sigma_{11} \end{bmatrix} (\begin{bmatrix} m_{1,1} \\ m_{1,2} \end{bmatrix} - \begin{bmatrix} m_{2,1} \\ m_{2,2} \end{bmatrix}) \\ &= (\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - \begin{bmatrix} \overline{m}_1 \\ \overline{m}_2 \end{bmatrix})^T |\mathbf{C}|^{-1} \begin{bmatrix} \sigma_{22} & -\sigma_{12} \\ -\sigma_{12} & \sigma_{11} \end{bmatrix} \begin{bmatrix} \delta_1 \\ \delta_2 \end{bmatrix} \\ &= |\mathbf{C}|^{-1}[(x_1 - \overline{m}_1)(\sigma_{22}\delta_1 - \sigma_{12}\delta_2) + (x_2 - \overline{m}_2)(-\sigma_{12}\delta_1 - \sigma_{11}\delta_2)] \end{aligned}$$



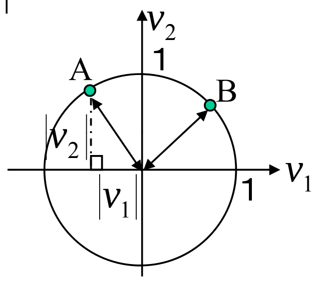
# マハラノビス距離 －2変数の場合－

(I) ２変数が互いに無相関の場合

右図のような無相関な２次元正規母集団からサンプルが発生する場合を考える．




平均： $\mathbf{m} = \begin{bmatrix} m_1 \\ m_2 \end{bmatrix}$ 共分散行列： $\mathbf{C} = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}$




このサンプルの観測値 x =[$x_1$,$x_2$]$^T$を以下のように基準化する．

$$v_1 = (x_1 - m_1)/\sigma_1$$

$$v_2 = (x_2 - m_2)/\sigma_2$$

このとき，平方距離は

$$D^2 \equiv v_1^2 + v_2^2 = \frac{(x_1 - m_1)^2}{\sigma_1^2} + \frac{(x_2 - m_2)^2}{\sigma_2^2}$$

Dをマハラノビス距離とよぶ．

特徴：マハラノビス距離が等しい２点は同じ生起確率をもつ

（注）もとの観測値空間での単純なユークリッド(Euclid)距離では

$$D_E^2 \equiv (x_1 - m_1)^2 + (x_2 - m_2)^2$$

このとき，図から明らかなように

$$D_{E,A}^2 \neq D_{E,B}^2$$

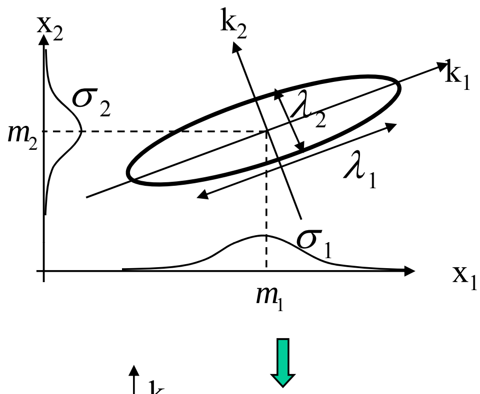
# マハラノビス距離 －2変数の場合（つづき）－

(II) ２変数に相関がある場合

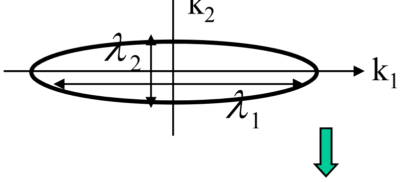
右図のような相関のある２次元正規母集団からサンプルが発生する場合を考える.

統計量：

$$\mathbf{m} = \begin{bmatrix} m_1 \\ m_2 \end{bmatrix}, \quad \mathbf{C} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}$$




分布の主成分分析を行い，第１主成分および第２主成分を求める.
次に，各主成分方向を新しい座標系にとる．このとき平方距離は

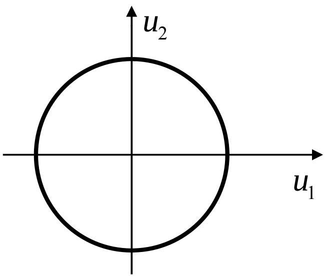



$$\mathbf{C}_k = \begin{bmatrix} \lambda_1^2 & 0 \\ 0 & \lambda_2^2 \end{bmatrix}$$

$$D^2 \equiv u_1^2 + u_2^2 = \frac{k_1^2}{\lambda_1^2} + \frac{k_2^2}{\lambda_2^2}$$

Dをマハラノビス距離とよぶ.

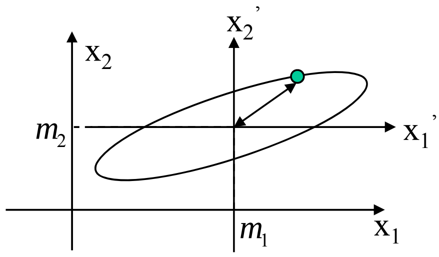
$$v_1 = k_1 / \lambda_1$$
$$v_2 = k_2 / \lambda_2$$






# マハラノビス距離 －一般的表現－

一般に，２変数のマハラノビス距離は以下のステップで変換した座標系におけるユークリッド距離で与えられる.

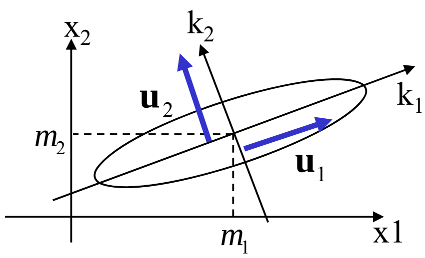



1．平均を0にする.

$$\mathbf{x'} = \mathbf{x} - \mathbf{m}$$

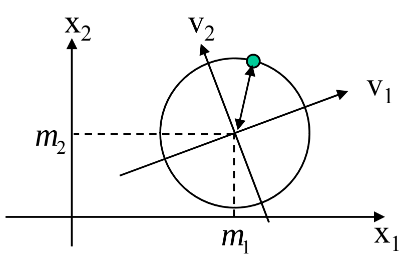
2．座標系を回転して（Hotelling変換）
主成分の方向を新しい軸とする.

$$\mathbf{k} = \mathbf{U}\mathbf{x'} \qquad \mathbf{U} = [\mathbf{u}_1 \quad \mathbf{u}_2]^T$$




3．各主成分を，それぞれの標準偏差で割って正規化する（白色化：whitening).

$$\mathbf{v} = \mathbf{C}_k^{-1/2}\mathbf{k}$$




4．新しい座標系でのユークリッド距離を算出する.

$$D^2 \equiv \|\mathbf{v}\|^2 = \mathbf{v}^T\mathbf{v}$$

# マハラノビス距離 ーー般的表現（つづき）ー

目標：D$^2$を母集団の平均ベクトルと共分散行列で表す.

1．平均を0にする.

$$\mathbf{x}' = \mathbf{x} - \mathbf{m}$$

2．回転する.　　ただし $\mathbf{C}_k = \mathbf{U}\mathbf{C}\mathbf{U}^T$

$$\mathbf{k} = \mathbf{U}\mathbf{x}'$$

$$\mathbf{C}_k = \begin{bmatrix} \lambda_1^2 & 0 \\ 0 & \lambda_2^2 \end{bmatrix}$$

3．各主成分を正規化する.

$$\mathbf{v} = \mathbf{C}_k^{-1/2}\mathbf{k} \quad \text{ただし} \quad \mathbf{C}_k^{-1/2} \equiv \begin{bmatrix} 1/\lambda_1 & 0 \\ 0 & 1/\lambda_2 \end{bmatrix}$$

要素で書けば

$$\begin{bmatrix} 1/\lambda_1 & 0 \\ 0 & 1/\lambda_2 \end{bmatrix}\begin{bmatrix} k_1 \\ k_2 \end{bmatrix} = \begin{bmatrix} k_1/\lambda_1 \\ k_2/\lambda_2 \end{bmatrix}$$

4．新しい座標系でのユークリッド距離を算出する.

$$D^2 \equiv \|\mathbf{v}\|^2 = \mathbf{v}^T\mathbf{v}$$

D$^2$を書き下すと

$$\begin{aligned} D^2 &= \mathbf{v}^T\mathbf{v} \\ &= (\mathbf{C}_k^{-1/2}\mathbf{k})^T\mathbf{C}_k^{-1/2}\mathbf{k} \\ &= (\mathbf{C}_k^{-1/2}\mathbf{U}(\mathbf{x}-\mathbf{m}))^T\mathbf{C}_k^{-1/2}\mathbf{U}(\mathbf{x}-\mathbf{m}) \\ &= (\mathbf{x}-\mathbf{m})^T\mathbf{U}^T(\mathbf{C}_k^{-1/2})^T\mathbf{C}_k^{-1/2}\mathbf{U}(\mathbf{x}-\mathbf{m}) \\ &= (\mathbf{x}-\mathbf{m})^T\mathbf{U}^T\mathbf{C}_k^{-1}\mathbf{U}(\mathbf{x}-\mathbf{m}) \\ &= (\mathbf{x}-\mathbf{m})^T\mathbf{U}^T(\mathbf{U}\mathbf{C}^{-1}\mathbf{U}^T)\mathbf{U}(\mathbf{x}-\mathbf{m}) \\ &= (\mathbf{x}-\mathbf{m})^T\mathbf{C}^{-1}(\mathbf{x}-\mathbf{m}) \end{aligned}$$

ただし，以下を用いた.

$$\begin{aligned} \mathbf{U}^T &= \mathbf{U}^{-1} \\ \mathbf{C}_k^{-1} &= (\mathbf{U}\mathbf{C}\mathbf{U}^T)^{-1} \\ &= (\mathbf{U}^T)^{-1}\mathbf{C}^{-1}\mathbf{U}^{-1} \\ &= \mathbf{U}\mathbf{C}^{-1}\mathbf{U}^T \end{aligned}$$