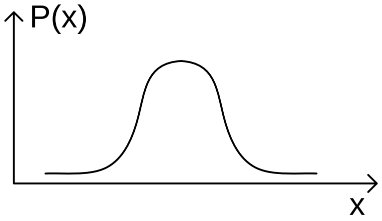
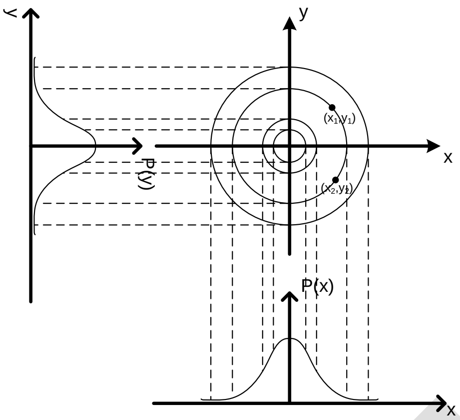
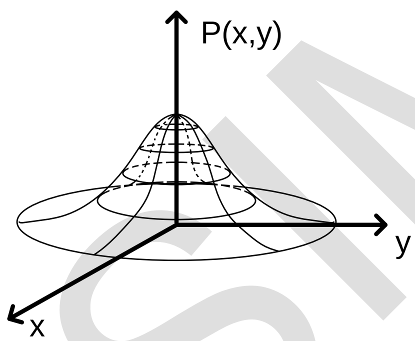
Multivariate Gaussian distribution

Univariate Gaussian(一維高斯)



Multivariate Gaussian(多維高斯)
以二維為例：





lesson6 筆記中有詳細說明

若變數彼此不獨立，就不會有 $P(x)P(y)=P(r)$ 的關係，那麼就不會是長短軸為 x,y 軸的橢圓的圖形

先定義一下等下用的變數
一樣假設為二維，X, Y 為一個二維分布 Z 映射到一維的分布 X, Y

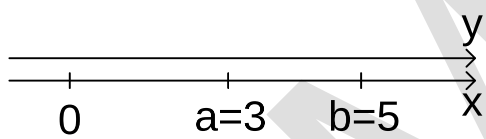
$$\begin{array}{ccc} Z & \mu & \Sigma \\ \begin{bmatrix} x \\ y \end{bmatrix} & \begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix} & \begin{bmatrix} \sigma_x{}^2 & \mathrm{cov}_{xy} \\ \mathrm{cov}_{xy} & \sigma_y{}^2 \end{bmatrix} \end{array}$$

則分布會變成 $P(\mu,\Sigma)=\dfrac{1}{\sqrt{2\pi}^{k}\,|\Sigma|^{0.5}}e^{\frac{-1}{2}(x-\mu)^{T}\Sigma^{-1}(x-\mu)}$

note: $(x-\mu)^{T}\Sigma(x-\mu)$ 在直角坐標(變數彼此獨立)中，可視為歐式距離(Euclidean distance)，而若變數彼此不獨立，$(x-\mu)^{T}\Sigma(x-\mu)$ 稱為 Mahalanobis distance(馬氏距離)

簡介馬氏距離：
當變數彼此間不獨立時，變數彼此間會存在某種關係，這時用歐式距離就會顯得比較不精確，因為歐式距離僅考量每個變數的差，馬氏距離會將其相關性考量進去，舉個比較極端的例子來說，如果我們想要判斷兩個成年人像不像，我們用三個標準來評量，一個是性別，一個是有沒有喉結，一個是身高，你應該會覺得這個標準很奇怪，因為其實這兩個標準是近乎完全相關的，如果我們看到一個男生和一個女生，用歐式空間來判斷的話，同樣的因素(男 or 女)會被我們考慮兩次，如此會顯得身高的差距比較沒有那麼重要，但是馬氏距離會將這兩個標準視為同一種標準，就不會有上面講的情況。以此類推，再舉個比較沒有那麼極端的例子，應該就比較能了解了，一樣是判斷兩個人像不像，三個標準分別是身高、體重、髮量，雖然身高和體重並不是完全相關，但是還是有一定的正相關，身高較高的人通常會較身高較矮的人重，這時將身高體重視為兩個標準(互相獨立，歐式距離)來看就有失公允，雖然這身高體重也沒辦法視為一個標準來看，但是可想成可視為 1~2 之間個標準來看。

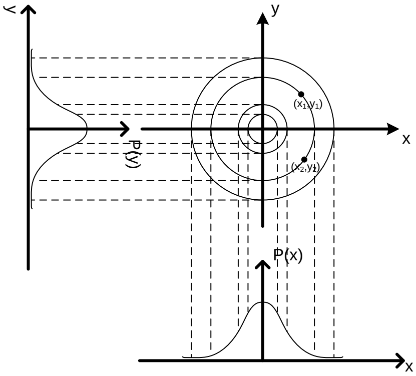
如果用直角坐標想這件事情，兩個完全相關的維度就好像是我們將 y 軸定義為 x 軸，我們重複算了兩次 x 軸



假設 x 軸和 y 軸完全相關，我們想知道 ab 兩點的距離，如果我們將兩個維度都視為獨立，計算出來的距離為 $\sqrt{(5-3)^{2}+(5-3)^{2}}=2\sqrt{2}$，但是其實這是兩個一樣的維度，距離應該是 2，只看其中一個軸就好

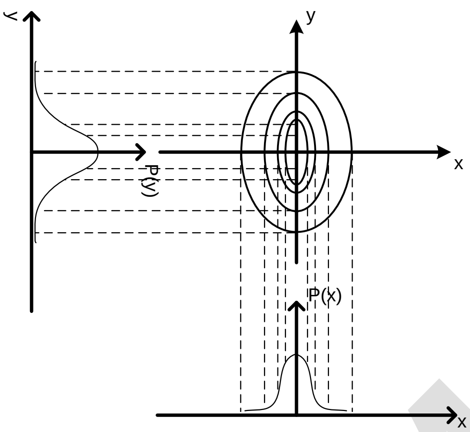
case1: 若兩變數分布一樣，就好像是兩個高斯函數一樣，$\Sigma=\mathrm{I}=\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}$

兩變數合成的分布為



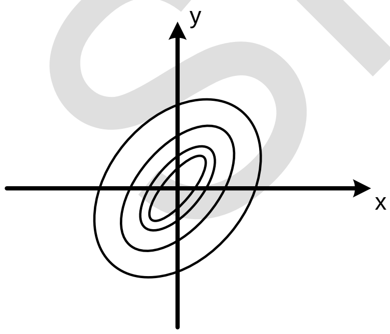
case2: 若兩變數還是獨立，但是 variance 不同，$\Sigma = D = \begin{bmatrix} {\sigma_x}^2 & 0 \\ 0 & {\sigma_y}^2 \end{bmatrix}$

兩變數合成之分布為



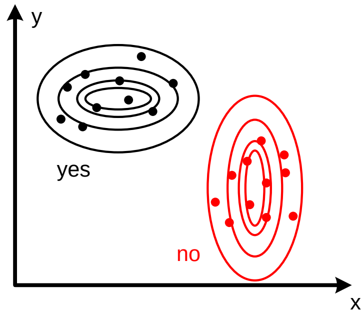
case3: 若兩變數不獨立，$\Sigma = symmetric = \begin{bmatrix} {\sigma_x}^2 & \mathrm{cov}_{xy} \\ \mathrm{cov}_{xy} & {\sigma_y}^2 \end{bmatrix}$

兩變數合成之分布就不會是方正的，可能會有旋轉，平移…



note: 還記得 naïve Bayes classifier 吧！如果是使用 naïve Bayes classifier 計算各事件機率，因為我們會針對不同的 outcome 去分析，會將各變數視為獨立，計算其 posterior，所以各個 outcome 我們都會得到一群 contour(可想成等高線或是山丘，也就是上面那張圖)，這也是 naïve Bayes classifier 缺陷所

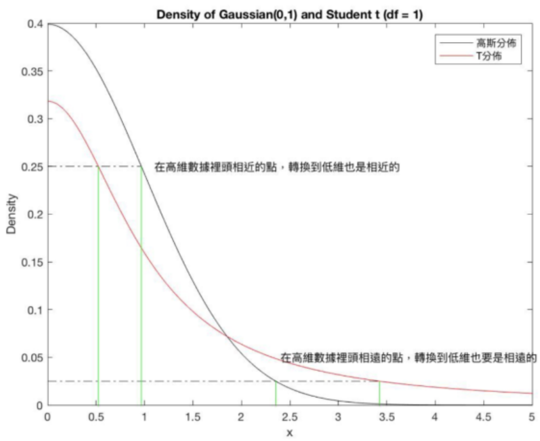
在
e.g. 以打網球那個例子為例，得出來的分布可能會是這樣，稍微注意的是因為我們將所有變數視為獨立，故 contour 只會是方正的，不會有上面那張圖一樣是斜的(各變數有相關)



student-t distribution

講之前就要先提一下 z test，z test 是在已經知道總體的 mean 及 variance 時，我們假設總體的分布為常態分布，再看先在發生的 outcome 在此分布下發生更極端事件的機率，但是通常我們在取樣時無法知道總體的分布為何，也不知道總體的 mean 及 variance，所以沒辦法假設其分布，student-t distribution 就是適用在這樣的情形，尤其適用在小樣本取樣時，student-t distribution 簡單來說就是較不確定、能容許更多例外版的 Gaussian distribution，會用小樣本中的 mean, variance 得到其分布，其分布外觀較 Gaussian distribution”寬”，詳細的分布公式這裡就不給，有概念就好。

如果直接用 z test 做小樣本分析，誤差會很大，可以想成我們想要知道全台灣人的身高分布，但我們只取 10 個人來做為我們的總體分布，這樣做出來的分布離實際上的分布一定有很大的差距，但若是 student-t distribution，若只取 10 個人，我不會太輕易以這 10 個人的數據做為絕對依據，會給予一定的”不確定”空間，所以這也是為何 student-t distribution 為何適用於小樣本的原因。



T分佈與高斯分佈的機率密度函數

圖片來源：https://medium.com/d-d-mag/淺談兩種降維方法-pca-與-t-sne-d4254916925b

黑線是 Gaussian distribution，越寬的分布是自由度越低(這裡先不用管自由度是啥，只要知道 student-t distribution 比 Gaussian distribution 還寬就好)的 student-t distribution

T-SNE(補充)

T-SNE 是一種降維的方法，將一筆資料由較高的維度降到較低的維度，怎麼降維的就省略不談，T-SNE 用較精準的 Gaussian distribution 表示，但是在降維之後必定會遺失某些資訊，故使用 student-t distribution 來表示。詳細的內容可參考
https://medium.com/d-d-mag/淺談兩種降維方法-pca-與-t-sne-d4254916925b

Affine transformation(仿射轉換)

若將一空間線性轉換為另一空間，稱為線性轉換(linear transformation)，旋轉、放大縮小是線性轉換，而仿射轉換增加”平移”這一個功能，原本線性轉換可表示為 $T(\vec{x}) = A\vec{x}$，而仿射轉換可表示為

$$T(\vec{x}) = A\vec{x} + \vec{b}$$

而在這裡我們想看看 Multivariate Gaussian 是不是也擁有 Affine 的性質，也就是

$$x \sim N(\mu, \sigma^2) \quad \Rightarrow f(x) = Ax + b \sim N(A\mu + b, ACA^T)$$

在這裡我們先看 mean 和 variance 會有怎樣的轉變，詳細證明見下方提供的網址

mean

$$E(x) = \mu = \int xP(x)dx$$
$$E(Ax+b) = \int (Ax+b)P(x)dx = \int AxP(x)dx + \int bP(x)dx = A\int xP(x)dx + b\int P(x)dx$$
$$= AE(x) + b \cdot 1 = A\mu + b$$

variance

$$\mathrm{cov}(x) = \Sigma = E\{(x-\mu)(x-\mu)^T\}$$
$$\mathrm{cov}(Ax+b) = E\{(Ax-b-\mu)(Ax-b-\mu)^T\}$$
$$= E\{(Ax-b-(A\mu-b))(Ax-b-(A\mu-b))^T\}$$
$$= E\{(A(x-\mu)(A(x-\mu))^T)$$
$$= E\{(A(x-\mu)(x-\mu)^T A^T\}$$
$$= AE\{(x-\mu)(x-\mu)^T\}A^T = A\Sigma A^T$$

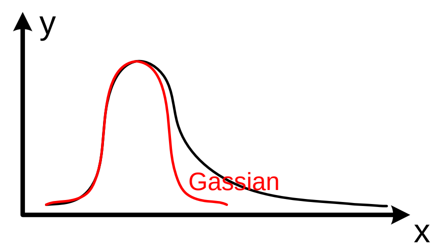
若想看詳細證明，見 http://www.mast.queensu.ca/~stat353/slides/5-multivariate_normal17_4.pdf

Moment generating function(MGF)　　(補充)

其實在 lesson3 中已經有提到此概念，是不知道為何上課上到這還要再拿出來講一遍

我們使用 mean 及 variance 其實還沒辦法完整的描述一個分布，我們在此之前大都是用最普遍的 Gaussian distribution 做為假設但是常常分布會有偏斜或是 peak 較尖…，這時候我們通常都要使用第三、第四階動差函數才能描述。

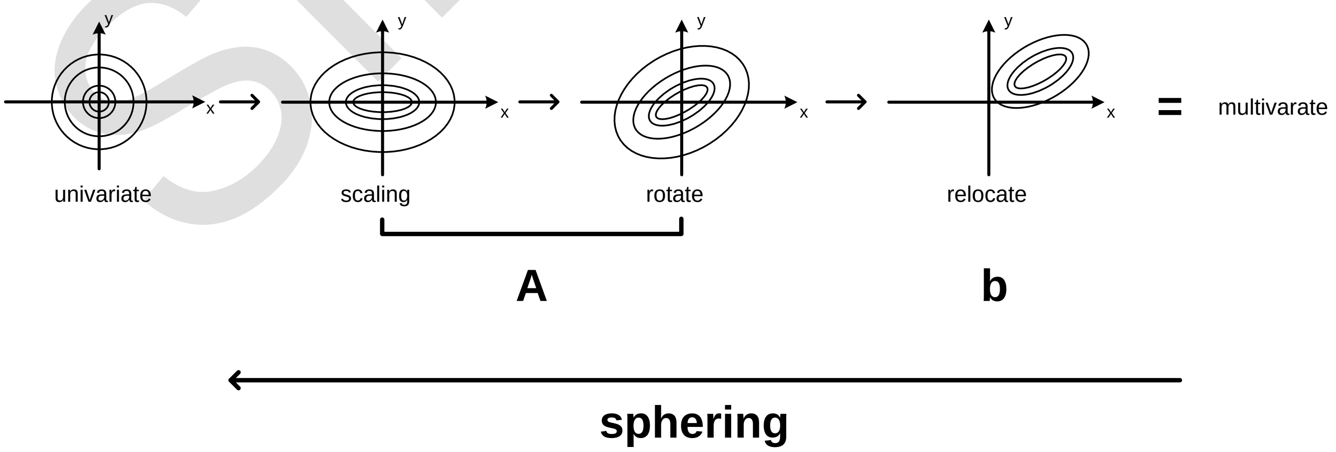
e.g.



以 univariate Gaussian 經由 Affine transformation 來轉為 Multivariate Gaussian

假設原始資料有 n 個維度

$$\text{此n維皆為獨立}\begin{cases} X_0 \sim N(0,1) \\ X_1 \sim N(0,1) \\ \quad\cdots \\ X_n \sim N(0,1) \end{cases}$$



故所有的 multivariate Gaussian distribution 都可藉由 univariate Gaussian distribution 經由仿射轉換得

到，以下由代數證明

Linear transformation of Gaussian distribution

要證明有線性轉換須證明兩件事

1. $$\forall x \in X, \forall y \in Y$$
$$T(x+y) = T(x) + T(y)$$

2. $$\forall x \in X, \forall y \in Y, a \in R$$
$$T(ax) = aT(x) \qquad \Rightarrow T(ax_1 + bx_2 + \ldots) = aT(x_1) + bT(x_2) + \ldots$$

第一個條件

另一 Y 分布為兩分布 X_1, X_2 的合，即

$$Y = X_1 + X_2 \qquad X_1 \sim N(\mu_1, \sigma_1^{\ 2}), X_2 \sim N(\mu_2, \sigma_2^{\ 2})$$

$$E(X) = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \Sigma = \begin{bmatrix} \sigma_1^{\ 2} & 0 \\ 0 & \sigma_2^{\ 2} \end{bmatrix}$$

則取 $A = \begin{bmatrix} 1 & 1 \end{bmatrix}, b = [0]$

則根據 affine property

$$AX + b = \begin{bmatrix} 1 & 1 \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} + 0 = X_1 + X_2 \sim N(\mu_1 + \mu_2, A\Sigma A^T)$$

$$A\mu + b = \begin{bmatrix} 1 & 1 \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} + 0 = \mu_1 + \mu_2$$

$$A\Sigma A^T = \begin{bmatrix} 1 & 1 \end{bmatrix} \begin{bmatrix} \sigma_1^{\ 2} & 0 \\ 0 & \sigma_2^{\ 2} \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \sigma_1^{\ 2} + \sigma_2^{\ 2}$$

$$\Rightarrow Y \sim N(\mu_1 + \mu_2, \sigma_1^{\ 2} + \sigma_2^{\ 2})$$

衍伸：

$$Y = X_1 + X_2 + \ldots + X_k = \sum_{i=1}^{k} X_i$$

$$\Rightarrow Y \sim N(\sum_{i=1}^{k} \mu_i, \sum_{i=1}^{k} \sigma_i^{\ 2})$$

第二個條件

$Y = aX_1 + bX_2 + cX_3 + ...$

則$Y \sim N(a\mu_1 + b\mu_2 + \mathrm{c}\mu_3 + ..., a\,\sigma_1{}^2 + b\sigma_2{}^2 + c\sigma_3{}^2 + ...)$

推廣到矩陣仍適用

X_i: multivariate Gaussian distribution(前面的推論的 X 都是 univariate)

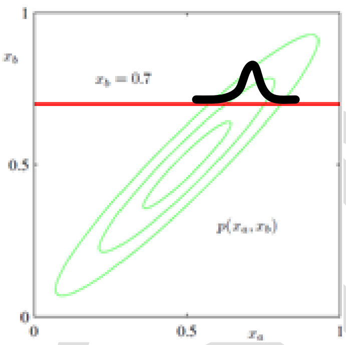
$$E(Y) = \sum_{i=1}^{k} B_i\mu_i, Var(\mathrm{Y}) = \sum_{i=1}^{k} B_i\Sigma_i B_i^T$$

$$Y \sim N(\sum_{i=1}^{k} B_i\mu_i, \sum_{i=1}^{k} B_i\Sigma_i B_i^T)$$

Conditional Gaussian distribution

當一分布有許多變量時(multivariate Gaussian distribution)，當我們已確定某變量時，其餘變量所形成的分布仍為高斯分布

用圖來表示就像這樣



來自課本 p90

此圖有 x_a, x_b 兩個變量，左圖是兩個變量形成的分布。當我們確定 $x_b = 0.7$時， $P(x_a \mid x_b = 0.7)$ 的分布如右圖，仍為一個高斯分布。

這裡在課本的 p85-87 之間，推導部分我看不懂為何經過條件機率後仍為高斯分布，但是我可以解釋 mean 和 variance 為何是那些值，會有些複雜。

Derivation

假設 X 中有 X_a, X_b 兩個變量集合

$$\left.\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \ldots \\ x_k \end{bmatrix}\right] \mathbf{X_a}$$

$$\left.\begin{bmatrix} x_{k+1} \\ x_{k+2} \\ \ldots \\ x_n \end{bmatrix}\right] \mathbf{X_b}$$

假設 X_b 為定值的向量

原始 multivariate Gaussian distribution 的式子為 $P(\mu,\Sigma) = \frac{1}{\sqrt{2\pi}^k |\Sigma|^{0.5}} e^{\frac{-1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)}$，我們只專注於 exponential 的指數，因為前面的係數只是為了讓機率加總為 1 而已，假設 X 為我們假設的那樣，指數變為

$$\frac{-1}{2}(X-\mu)^T\Sigma^{-1}(X-\mu) = \frac{-1}{2}\left(\begin{bmatrix} X_a \\ X_b \end{bmatrix} - \begin{bmatrix} \mu_a \\ \mu_b \end{bmatrix}\right)^T \Sigma^{-1}\left(\begin{bmatrix} X_a \\ X_b \end{bmatrix} - \begin{bmatrix} \mu_a \\ \mu_b \end{bmatrix}\right) = \frac{-1}{2}\begin{bmatrix} X_a-\mu_a \\ X_b-\mu_b \end{bmatrix}^T \begin{bmatrix} \Lambda_{aa} & \Lambda_{ab} \\ \Lambda_{ba} & \Lambda_{bb} \end{bmatrix}\begin{bmatrix} X_a-\mu_a \\ X_b-\mu_b \end{bmatrix}$$

$$= \frac{-1}{2}\begin{bmatrix} X_a-\mu_a & X_b-\mu_b \end{bmatrix}\begin{bmatrix} \Lambda_{aa} & \Lambda_{ab} \\ \Lambda_{ba} & \Lambda_{bb} \end{bmatrix}\begin{bmatrix} X_a-\mu_a \\ X_b-\mu_b \end{bmatrix} = \frac{-1}{2}\begin{bmatrix} X_a-\mu_a & X_b-\mu_b \end{bmatrix}\begin{bmatrix} \Lambda_{aa}(X_a-\mu_a)+\Lambda_{ab}(X_b-\mu_b) \\ \Lambda_{ba}(X_a-\mu_a)+\Lambda_{bb}(X_b-\mu_b) \end{bmatrix}$$

$$= \frac{-1}{2}\Lambda_{aa}(X_a-\mu_a)^2 + \frac{-1}{2}\Lambda_{ab}(X_a-\mu_a)(X_b-\mu_b) + \frac{-1}{2}\Lambda_{ba}(X_b-\mu_b)(X_b-\mu_b) + \frac{-1}{2}\Lambda_{bb}(X_b-\mu_b)(X_b-\mu_b)$$

而由於 X_b 是一個定值的向量，現在我們的變量只有 X_a 一個，假設我們現在已經知道經由條件機率後的分布仍為 multivariate Gaussian distribution，我們就要試著將

$$\frac{-1}{2}(X_a-\mu_a)^T\Lambda_{aa}(X_a-\mu_a) + \frac{-1}{2}(X_a-\mu_a)^T\Lambda_{ab}(X_b-\mu_b) + \frac{-1}{2}(X_b-\mu_b)^T\Lambda_{ba}(X_a-\mu_a) + \frac{-1}{2}(X_b-\mu_b)^T\Lambda_{bb}(X_b-\mu_b)$$

化為 $\frac{-1}{2}(X-\mu)^T\Sigma_{X_a|X_b}{}^{-1}(X-\mu)$ 形式，而在這個式子展開後會有 quadratic 項、linear 項，展開後的形式為

$$\frac{-1}{2}X^T\Sigma_{X_a|X_b}{}^{-1}X + \frac{1}{2}X^T\Sigma_{X_a|X_b}{}^{-1}\mu + \frac{1}{2}\mu^T\Sigma_{X_a|X_b}{}^{-1}X + \frac{-1}{2}\mu^T\Sigma_{X_a|X_b}{}^{-1}\mu = \frac{-1}{2}X^T\Sigma_{X_a|X_b}{}^{-1}X + X^T\Sigma_{X_a|X_b}{}^{-1}\mu + const.$$

中間的兩項($\frac{1}{2}X^T\Sigma_{12}{}^{-1}\mu$ 、$\frac{1}{2}\mu^T\Sigma_{21}{}^{-1}X$)為何可以合併，請見 lesson1，證明方式是一樣的。

$$\frac{-1}{2}(X_a-\mu_a)^T\Lambda_{aa}(X_a-\mu_a) + \frac{-1}{2}(X_a-\mu_a)^T\Lambda_{ab}(X_b-\mu_b) + \frac{-1}{2}(X_b-\mu_b)^T\Lambda_{ba}(X_a-\mu_a) + \frac{-1}{2}(X_b-\mu_b)^T\Lambda_{bb}(X_b-\mu_b)$$

中，第一項中會有 quadratic 項，第二項和第三項可以合併，合併後和第一項的某幾項中會有 linear 項，第四項為常數，故

quadratic 項：

$$\Sigma_{X_a|X_b} = \Lambda_{aa}{}^{-1}$$

linear 項：

$$\frac{-1}{2}(X_a-\mu_a)^T\Lambda_{aa}(X_a-\mu_a)+\frac{-1}{2}(X_a-\mu_a)^T\Lambda_{ab}(X_b-\mu_b)+\frac{-1}{2}(X_b-\mu_b)^T\Lambda_{ba}(X_a-\mu_a)$$

其中的 linear 項為

$$X_a{}^T\Lambda_{aa}\mu_a + X_a{}^T\Lambda_{ab}(X_b-\mu_b)=X_a{}^T(\Lambda_{aa}\mu_a+\Lambda_{ab}(X_b-\mu_b))$$

故

$$\Sigma_{X_a|X_b}{}^{-1}\mu = \Lambda_{aa}\mu_a+\Lambda_{ab}(X_b-\mu_b)$$
$$\Rightarrow \mu = \Sigma_{X_a|X_b}(\Lambda_{aa}\mu_a+\Lambda_{ab}(X_b-\mu_b)) = \mu_a+\Lambda_{aa}{}^{-1}\Lambda_{ab}(X_b-\mu_b)$$

解出 mean 和 covariance matrix 了！剩下的，就只有要解出 Λ 矩陣了，我們要用 Σ 來表示，所需要的技巧是方塊矩陣的反矩陣，可以只要記結論就好

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

其中

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

$$\begin{bmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{bmatrix}^{-1} = \begin{bmatrix} \Lambda_{aa} & \Lambda_{ab} \\ \Lambda_{ba} & \Lambda_{bb} \end{bmatrix}$$
$$\Rightarrow \Sigma_{a|b} = \Lambda_{aa}{}^{-1} = \Sigma_{aa}-\Sigma_{ab}\Sigma_{bb}{}^{-1}\Sigma_{ba}$$
$$\mu = \mu_a+\Lambda_{aa}{}^{-1}\Lambda_{ab}(X_b-\mu_b) = \mu_a-(\Sigma_{aa}-\Sigma_{ab}\Sigma_{bb}{}^{-1}\Sigma_{ba})^{-1}(\Sigma_{aa}-\Sigma_{ab}\Sigma_{bb}{}^{-1}\Sigma_{ba})^{-1}\Sigma_{ab}\Sigma_{bb}{}^{-1}(X_b-\mu_b)$$
$$= \mu_a+\Sigma_{ab}\Sigma_{bb}{}^{-1}(X_b-\mu_b)$$

方塊矩陣反矩陣證明(補充)

詳細見 https://ccjou.wordpress.com/2010/08/02/分塊矩陣的解題案例/

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix} = \begin{bmatrix} I & O \\ CD^{-1} & I \end{bmatrix} \begin{bmatrix} A-BD^{-1}C & O \\ O & D \end{bmatrix} \begin{bmatrix} I & BD^{-1} \\ O & I \end{bmatrix}$$ (將$\begin{bmatrix} A & B \\ C & D \end{bmatrix}$藉由行運算、列運算變為對角矩陣)

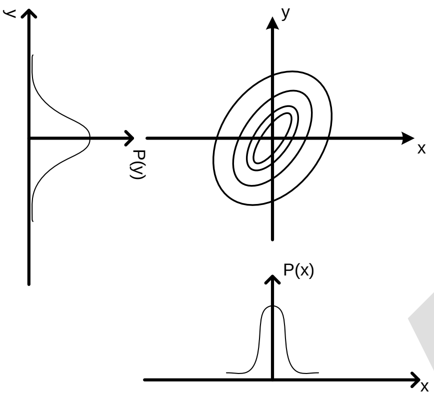
$$\Rightarrow \begin{bmatrix} A & B \\ C & D \end{bmatrix}^{-1} = \begin{bmatrix} I & BD^{-1} \\ O & I \end{bmatrix}^{-1} \begin{bmatrix} A-BD^{-1}C & O \\ O & D \end{bmatrix}^{-1} \begin{bmatrix} I & O \\ CD^{-1} & I \end{bmatrix}^{-1}$$

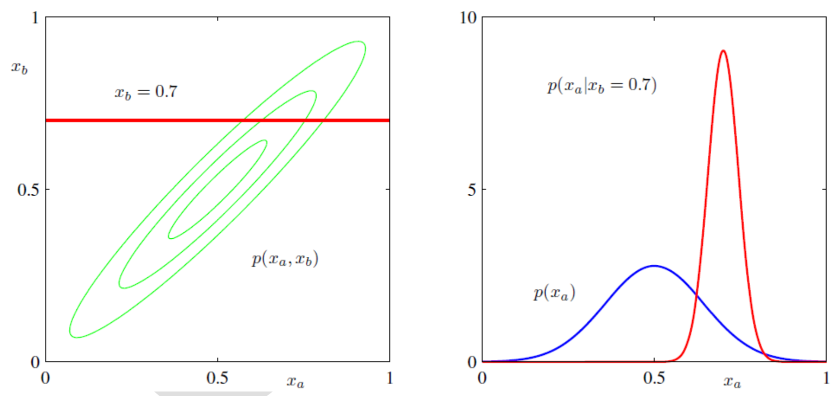
$$= \begin{bmatrix} I & -BD^{-1} \\ O & I \end{bmatrix} \begin{bmatrix} (A-BD^{-1}C)^{-1} & O \\ O & D^{-1} \end{bmatrix} \begin{bmatrix} I & O \\ -CD^{-1} & I \end{bmatrix} = \begin{bmatrix} M & -MBD^{-1} \\ -D^{-1}CM & D^{-1}+D^{-1}CMBD^{-1} \end{bmatrix}$$

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

Marginal Gaussian distribution

還記的 marginal 的意義嗎？是只看我們有興趣的維度，每個維度的值的機率為該維度值的前提下，將其餘維度的機率做加總，看不懂的話回去翻一下 lesson3，而實際上做完 marginal 後，分布仍然為 Gaussian distribution，詳細證明在課本 p88-89，但我看不太懂，用圖來表現的話就像下圖





來自課本 p90

上圖可看到綠色的線是原始的二維 multivariate Gaussian distribution，藍色線為 marginal Gaussian distribution(紅色是 conditional Gaussian，檢驗一下自己是不是看的懂這張圖)，藍色線會在 x_a=0.5 的地方有最大值是因為我們將 x_a=0.5 畫一條鉛直線，這條線上的機率加總是最大的

那麼這些被拆開來的維度的 mean 和 variance 如何表示？
和前面的假設一樣

$$X=\begin{bmatrix} X_a \\ X_b \end{bmatrix}, \mu=\begin{bmatrix} \mu_a \\ \mu_b \end{bmatrix}, \Sigma=\begin{bmatrix} \sigma_a{}^2 & B \\ B & \sigma_b{}^2 \end{bmatrix}$$

取 $A=\begin{bmatrix} I & O \end{bmatrix}, b=0$

$AX+b=\begin{bmatrix} I & O \end{bmatrix}\begin{bmatrix} X_a \\ X_b \end{bmatrix}=X_a$，故 marginal Gaussian distribution 仍為一種 affine property

$$\mu=A\mu+b=\begin{bmatrix} I & O \end{bmatrix}\begin{bmatrix} \mu_a \\ \mu_b \end{bmatrix}+0=\mu_a$$

$$\Sigma_a=A\Sigma A^T=\begin{bmatrix} I & O \end{bmatrix}\begin{bmatrix} \sigma_a{}^2 & B \\ B & \sigma_b{}^2 \end{bmatrix}\begin{bmatrix} I \\ O \end{bmatrix}=\begin{bmatrix} I & O \end{bmatrix}\begin{bmatrix} \sigma_a{}^2 \\ B \end{bmatrix}=\sigma_a{}^2$$

故

$$X_a \sim N(\mu_a, \sigma_a{}^2)$$
$$X_b \sim N(\mu_b, \sigma_b{}^2)$$

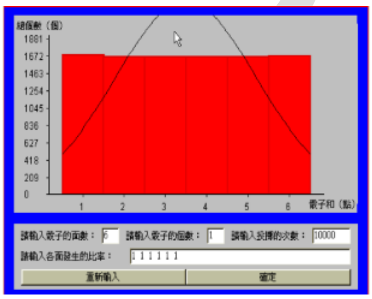
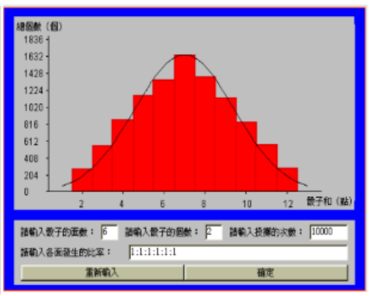
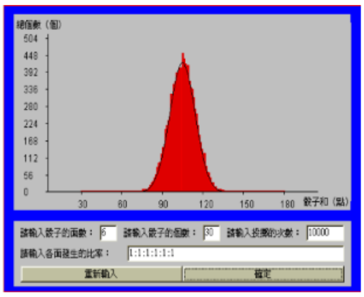
中央極限定理(central limit theorem)

無論原始的分布為何，只要我們一次取樣點越多，將取樣點取平均或加總，最終得到的分布會越趨近於高斯分布。

參考網站：http://www.math.nsysu.edu.tw/StatDemo/CentralLimitTheorem/CentralLimit.html

假設我們用骰子做實驗，如果我們一次骰一次骰子，做 10000 次取樣，得到的會趨近於 uniform distribution，而若我們一次骰的骰子越多，將一次採到的點數做加總，做 10000 次取樣後，所得到的分布會越趨近於高斯分布

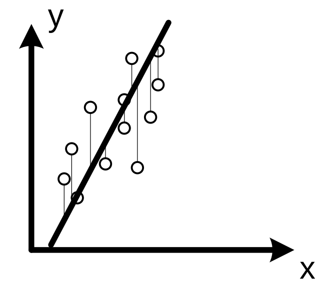
骰一顆骰子　　　　　　骰兩顆骰子　　　　　　骰 30 顆骰子

圖片來源：http://www.math.nsysu.edu.tw/StatDemo/CentralLimitTheorem/CentralLimit.html

Bayesian linear regression

回到 lesson1 的最後面，我們在計算 LSE 的地方，我們使用 LSE 是希望找出一個方程式，這個方程式的直線離我們現在的 data 的所有點有最短的距離。



就是希望得到有最小黑色細線長度平方的直線

但是，若我們是用 Bayesian 來想這件事情，我們的確知道這條線是最佳的，但是只有這種可能嗎？是不是也有可能發生其他條線的時候，但只是機率沒有最佳的那條直線高呢？LSE 可以知道實際的最佳直線為何，我們會希望離直線越近，機率應該是越大，離直線越遠機率越小，假設我們可以知道每個 X 值發生時 Y 值的 mean 及 covariance matrix，那麼 entropy 最大時的分布為 Gaussian distribution，所以如果我們直線中每點都取一個 Gaussian distribution，data 每點都可以得到一個機率，將每個點的機率做相乘，得到的就是這 data 能得到這條直線的機率。

如此我們就可以將問題轉換成另外一個問題，有哪條直線在我們手中的 data 中可以得到最大的機率，這條直線就是最佳的直線，這個是 MLE。更甚者，我們可以決定一個區間，是 outcome 可能會發生的地方(藉由已知的 variance)，也就是所謂的 predictive distribution