

第一章 緒論

1.1 研究動機

在工業製造中一個普遍的趨勢是提升自動化的程度。在產品製造的過程中通常有著數百甚至更多的步驟，我們希望機器能自動的去執行這些步驟進而減少甚或達到無人為操作的製造過程[12]。以半導體製造為例，這是一個高度自動化且複雜的過程。在製程中每一個步驟都會針對不同的項目進行測量，我們會得到許多的數據資料(通常每個 IC 從晶圓到成品的製造過程會有數百個甚至上千個步驟，每個步驟中又包含數種項目的測量數據，所以一顆 IC 製造完成我們通常可以得到上千項數據資料)。

正如前面所述，因為在 IC 自動化的製造過程中有著數百甚至更多的步驟，而每個步驟都會有數種測量項目，這些測量的項目可能是晶圓片的原始資料，可能是某個步驟金屬層或多晶矽層的厚度，可能是某個步驟落在晶圓上塵埃微粒的大小或分佈個數，可能是某處的電流值電阻值或其他電氣特性，抑或是某個步驟所處的溫度壓力值等各項的測試資料。若將每個項目都視為一個可能影響製作成敗(例如良率)的變數，那麼每一塊晶圓經過完整製程的我們都可以得到一個包含數百甚至上千個變數的測試資料。

現在我們可能發現某顆 IC 製程從開始到第 100 或 150 片晶圓，良率都偏低，此時我們要面對的資料將是一個極龐大的二維矩陣(矩陣的行代表的是測試項目的總個數也就是變數個數，列代表的是從開始到第 N 片晶圓的樣本數，也就是經過幾次製程)，這麼大量的資料和這麼多的變數量不管是儲存、分析或運算都會是相當龐大的負擔，尤其是數據需要經過運算處理時甚至可能超出軟體所能負荷的量，因此在分析資料和建立模型我們首要考慮的是如何減少變數量。

其次，IC 製程中同一個步驟或數個步驟之間會存在著類似的測量項目。舉例來說，可能在某個步驟會測量特定幾處多晶矽層的厚度也可能連續幾個步驟度

量同一點金屬層的厚度，諸如此類。因為這些類似的測量項目造成了資料的許多變數間存在著高度的相關性(correlated)(見圖 1.1，Var11 和 Var12 以及 Var17 和 Var18)，因此會有不少多餘的資訊(redundancy)。這些資料的存在對於分析會產生相當大的影響，用這樣的資料去建立模型不但會造成誤差，甚至可能得到和真實情況完全相反的結果，所以除了減少變數的量之外，我們還要去考慮如何有效地濾除多餘的資料。

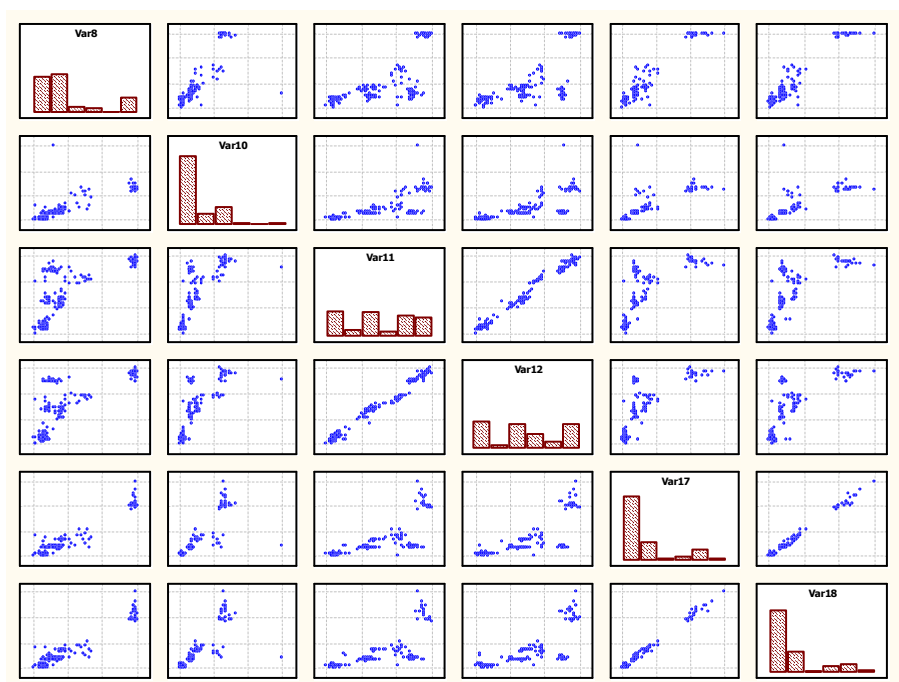


圖 1.1 資料相關的例子

從圖 1.1 中可以瞭解資料的分佈類型並不固定，同時還可以發現到某些資料會有歧異值(Outlier，例如圖 1.1 中 Var10 相對於其他變數的關係圖裡都會有一個點遠遠的離開大部分的資料點)不規則地分佈，歧異值的存在通常對於資料分析及模型建立會有著相當顯著的影響，這是我們在處理資料必須要注意和調整的。

最後是缺少資料項(Missing Data)的問題，有些時候因為一些特別的因素導致有些項目的數值無法測量或是沒有被記錄下來，這樣就有缺少資料項的產生，雖然在我們得到的數據中缺少的資料項不多，但是這種情形在資料的分析上很常發

生，因此也值得我們去討論。

總結來說，我們得到的數據會有以下的特性：

- (1) 資料變數量過於龐大，造成分析不易；
- (2) 許多變數之間會存在關聯性，因此會有多餘的資料存在；
- (3) 存在歧異點和缺少資料項的問題。

現在當製造過程中因為設計或其他因素導致良率的下降或不穩定，我們應該如何從大量的原始資料中去找出導致製造成敗的主要因素？如何從這些數據建立一套適合的模型提供我們去研究製造的過程，包括決定哪些步驟對良率有著相當顯著的關連性並讓我們能討論各步驟之間互相影響的關係？

1.2 研究方法

本論文主要是希望將 IC 製造所得的大量資料先經過包括特徵選取(Feature Selection)等的前置處理減少需要處理的資料量和一些異常資料分佈再藉由圖形化模型(Graphical Models, GMs)的方法去建立出模型，提供我們遇到生產上的問題(良率不穩定、偏低甚至完全不合乎我們期待)時能夠加以研究，找出影響良率的步驟或因素以及找出這些因素之間的關係，進而搭配其他方法達到調整或預測生產的效果。

所謂的圖形化模型是統計上一些方法的統稱，它已經被成功地應用在相當廣泛的領域上，從醫學診斷和決策的理論、社會行為科學的研究到時域的預測和訊號編碼等，用來闡述和研究種種問題或現象。GMs 結合了機率和圖學的理论，它概括了相當多種運用在統計分析與訊號處理的技術諸如 Kalman Filters[1], auto-regressive models[21],和許多資訊編碼理論[2]。GMs 以視覺圖形化的表述方式讓人們能清楚地研究和理解圖形背後所要表達隨機程序中的重要特性及實際物理的基本意義，同時它也提供了一些理論便於實現機率計算和決策。總括來說，GMs 包含了相當多完備的統計方法和技巧供我們應用[13,14]。

我們選擇使用的是 GMs 中的圖形化高斯模型(Graphical Gaussian Models, GGMs)，選擇的原因在於它處理的方向是針對共變異矩陣(covariance matrices)和關聯矩陣(correlation matrices)，對於多餘資料(redundancy)和資料間存在高度關聯性(highly correlated)的問題提供適當的解決方法。透過 GGMs 也能得到變數選擇的效果，同時圖形化的表示讓我們能夠很直觀地去瞭解變數之間互相的關聯性，對於資料的分析研究有很大的幫助。

現在就因應資料的特性選擇適合的實驗架構如圖 1.2 所示，其中前置處理包含了歧異值和缺少資料項的處理以及特徵選取的步驟去降低資料的變數量，藉著選取得到的排名，我們可以將和良率比較有關係的變數先拿來進行分析和建立模型；再來的圖形化模型則是針對變數之間的關聯性以及存在多餘資料的情形加以處理。

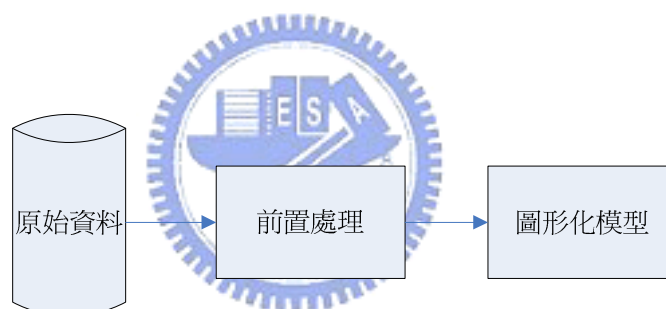


圖 1.2 針對資料特性提出的架構圖

雖然我們能夠藉由 GGMs 建立起模型，但是實際試驗的時候發現建立過程的運算量相當的大，變數越多，計算時間也變長。圖 1.3 是經過特徵選取過後的 30 個變數資料(共 154 片晶圓資料)，慢慢增加變數個數去測試個數的增加對於計算的影響，其中橫座標 2 代表的是依序將資料中變數兩兩送入 GGMs 中建立模型，3 代表的是依序將資料中變數每三個為一組送入 GGMs 中建立模型...依此類推，圖 1.3 顯示的是變數個數對於運算時間平均關係¹，可以看到運算時間幾乎呈

¹ 有的變數相關性較弱，可能檢驗中要保留的邊較少，計算量也較多；相對的相關性較強的話可能要測驗的邊較少，計算量也可能減少，因此在討論變數個數對時間的影響會將資料依序分組送入 GGMs 中測試並取平均值，以降低相關性造成的影響。

指數的成長。

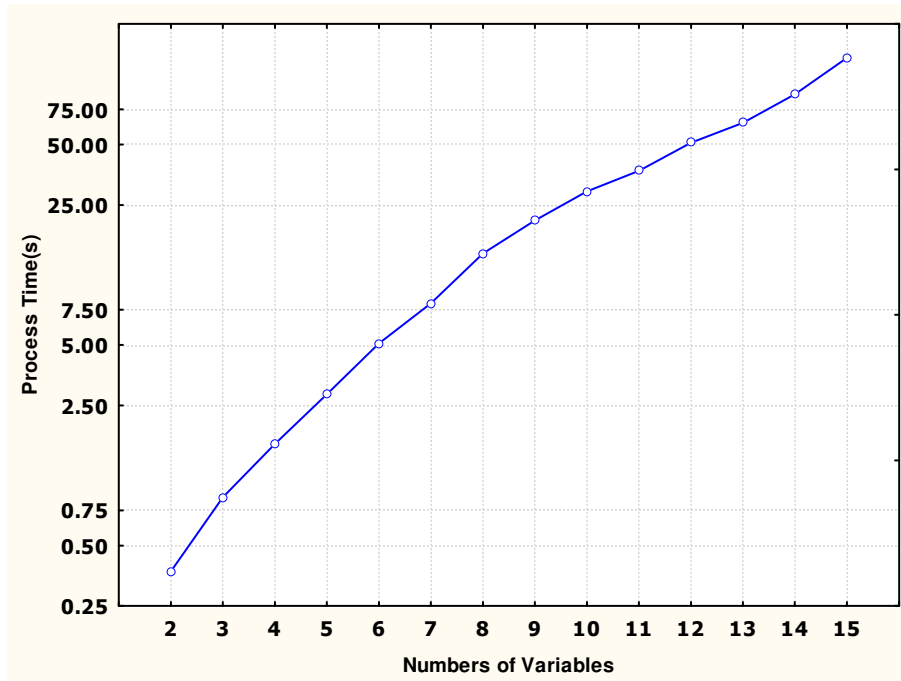


圖 1.3 變數增加對於時間的影響

值得注意的是圖中只顯示到送進 15 個變數去進行 GGMs 的建立，這是因為在實驗的時候，16 個變數的運算量已經超出程式編譯器所能負荷的量，造成記憶體不足而無法正常執行下去，因此若是以圖 1.3 中的趨勢來預測往後的運算量或運算時間可以發現用 20 個變數去建立 GGMs 運算時間可能約 1.3k 秒(相當於 22 分鐘)，而用 30 個變數運算時間則成長到 130k 秒(相當於 36 小時)，同時計算量更是遠遠超出電腦所能負荷的量，這是我們必須要改進的地方。

新的構想是原始資料經過特徵選取後，先進行變數的分群，將每一群的個數儘量限制在電腦所能負荷的變數量，每一群先各自建立 GGMs，然後各群之間再兩兩去建立連接的關係，以分群近似的方法去達到建立高維度資料模型的目標。

這邊分群所採用的方法是在多變量統計相當常用的因素分析(Factor Analysis, FA)和多維縮放比例(Multidimensional Scaling, MDS)兩種方法。因素分析最主要的用途在於將變數分類以及降低變數量。在此我們可以運用因素分析

找出新的潛在變數(例如因素, factor)或從該群中已知的資料裡選出最能代表該群的變數, 依照這些因素去將變數分群[5,23,24]。而多維縮放比例分析可以算是另一種形式的因素分析, 它簡單來說像是將變數重新安放在研究者給定的維度空間中, 因此我們需要在給定的維度空間中找到最近似的位址, 並且去估算變數間近似的相對位置, 依照變數間潛在的位置關係及彼此間的距離去解釋變數相似的程度。[15,16]

這兩種方法的差別在於FA的資料必須是多變量常態分佈且每一個變數為因素(factor)和誤差的線性組合, 而MDS則沒有這些限制, 因此MDS在實際上可應用的範圍較廣但是分類結果較簡單; FA雖然有所限制但是可能會比MDS抽取出更多的因素也更加準確, 有鑑於此我採用這兩種方法互相佐證。

現在整個研究方法的架構會如圖 1.4 所示, 原始資料先經過包含去除歧異值與特徵選取等的前置處理減少資料異常分佈的情形和降低變數量, 之後經過結構檢測(Structure Detection)也就是藉著因素分析和多維縮放比例分析將變數分群, 每一群先各自建立圖形化模型, 再去建立群和群之間的關係以完成整體的模型。

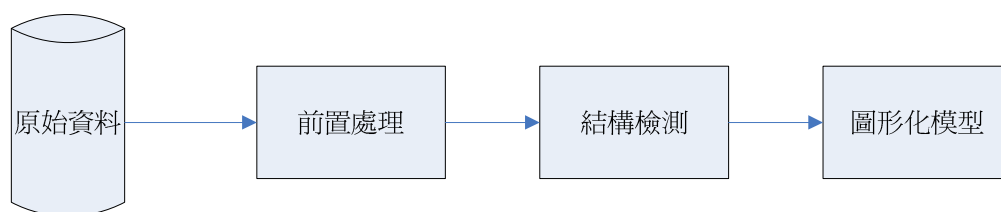


圖 1.4 簡化方法的整體架構

在實驗的過程中, 簡化方法相對於一般方法的確能縮短模型建立的時間, 運算量的降低有助於我們分析更多的變數。但是卻有一個新的問題產生, 即模型的偏差值並不符合我們的要求, 也就是運用簡化方法建立出來的模型會和原始資料存在的整體關係相差甚遠。探究其中的原因可以發現簡化方法中分群去建立圖形化模型以及各群兩兩建立連接關係的步驟, 事實上就是忽略了變數之中的部分

關聯性，可能會誤刪了一些在某一群中並不重要但對整體而言相當具有影響力的邊，因此要改進這項缺失我們進行圖形化模型的步驟之前，對所有的邊進行最大近似估測，把對於整體模型較重要的邊先保留起來再去建立模型。以圖 1.4 而言整體的架構並沒有改變，只是在結構檢測和圖形化模型的步驟之間加上一道估測和保留的保護機制，先從整體去討論對模型比較重要的邊，減少往後在分群建立模型過程發生誤刪的情形，如此使簡化方法在我們能處理更多變數的同時還能保持模型的偏差值在所要求的範圍之內。

建立完模型之後我們可以運用一些專家系統或是貝氏網路等方法去分析 [10]。舉例而言當製程的良率不符合我們的要求時，就可以運用結合貝氏網路的模型去找出對於良率影響最大的數種變數以供我們進行改良製程時的參考資料，也可以藉由這種結合的方法去預測調整某些參數或製造方法對於結果會有如何的影響，以減少嘗試錯誤的機會和成本的浪費。



1.3 論文結構

本論文共分為六章。第二章針對圖 1.4 前置處理和結構檢測的部分包括因素分析和多維縮放比例分析加以敘述。第三章則對於圖形化模型目前的發展和模型建立的原理有更進一步的描述，同時提出建立圖形化模型的演算流程。第四章將簡化方法遭遇到的問題提出說明和改進方法。第五章是實驗和討論的部分，會對一些資料特性的處理作說明，接著討論模型建立中計算的效能限制，最後再去比較簡化研究方法的結果。第六章是結論的部分。

第二章 前置處理和多變量分析

2.1 前置處理

2.1.1 前置處理簡介

原始數據在分析之前，通常會針對一些異常的資訊，諸如歧異值或資料項的缺失等，作過濾或處理以降低不正常數據對研究結果的影響，並且在變數過多的時候使用特徵選取的方式去選擇較重要的變數來分析。

2.1.2 歧異值的處理

所謂的歧異值就是在單變數或多變數中一些數值相當極端的樣本，通常這些對於整個資料的統計特性影響很大。舉例來說，在求取回歸函式的係數時，歧異值通常就會成為決定配適結果的主要因素，如圖 2.1 的二維分佈圖所示，圖中下面較聚集的點彼此差異不大，因此我們可以找到一條不錯的回歸線(圖中的虛線部分)，但是加上右上方的 A 點之後，為了考慮這個極端的值造成配適結果的改變，在這個圖中我們可以發現歧異值相對於同一群中的其他點的確有著相當顯著的影響。事實上歧異值的問題在資料分析是非常常見的，且歧異點對分析結果的影響程度通常並沒有一定的標準可循，而歧異點的產生大概有以下四種原因：

- (1) 資料輸入錯誤，這可能是人為輸入疏失造成資料項目不正確；
- (2) 缺少資料項引起的錯誤，在撰寫電腦程式分析資料時若忽略了對某些項目缺少資料的情形作適當的考慮就可能有錯誤的產生；
- (3) 取樣的錯誤，有時候歧異點可能根本不屬於我們所要取樣的樣本空間，因此在察覺到這種情形時這類的數值應該予以刪除。
- (4) 資料分佈的特性，可能有些資料就真的或多或少會存在一些極端的情形，這些是很難預測的。

對於連續分佈的變數，會先考慮單變量的歧異值，其判斷的依據是偏差值大於 3.29($p < 0.01$, two-tailed test)，不過資料極端的程度還是跟樣本的數量有關，當樣本的數量 N 很大時，偏差值大於 3.29 的樣本通常非常少。其他判斷歧異值的依據是藉著圖形的方法，我們可以將資料繪成統計圖、方塊圖、常態機率圖、趨勢圖或分佈圖。正常的資料通常是在平均值周圍大概呈對稱的分佈，而歧異點(或是一些比較少數的樣本)看起來似乎就和整個分佈群沒有連接在一起。

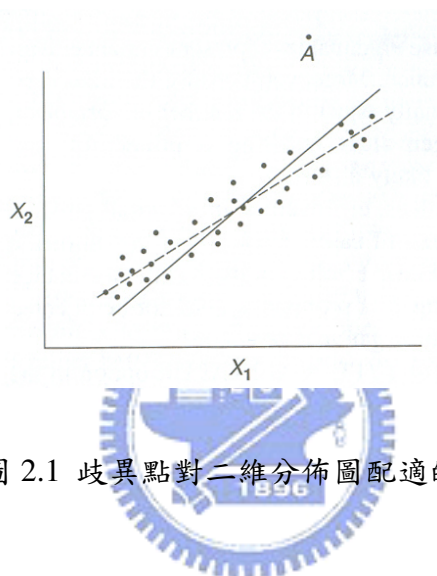


圖 2.1 歧異點對二維分佈圖配適的影響

找到單變量的歧異值之後，通常會循著這個方向去考慮多變量的情形。判斷多變量的歧異值也可以依照單變量的作法藉著偏差值或圖形方法來進行，但是在多變量的情形下直接用肉眼去區分其中差異幾乎是不可能的，我們可以藉著統計的方法對每一個樣本計算馬哈拉諾畢斯距離(Mahalanobis distance)。這個距離指的是某個樣本到其他樣本所組成重心(centroid)的位置，所謂的重心是由所有變數的平均值構成的，詳細的討論見[7]。

最後值得注意的是，有時候一些歧異值會被某些歧異值所隱蔽，舉例而言我們發現了一些歧異值並且去除，雖然會使得此項資料分佈正常但是其他資料卻出現相當極端的分佈，因此比較謹慎的作法是多方去比較歧異點去除之後的效果，多去嘗試直到無法再找到新的數值或是去除歧異點之後的結果足夠讓我們滿意為止。

2.1.3 缺少資料項的處理

在資料分析上另一個很常見的問題是缺少資料項，產生的原因很多，可能是機械發生問題或電氣效應無法測量，也可能是人為的疏失。通常缺少的資料項所分佈的形式比數量重要，如果缺少的某些數據在整個資料中呈隨機分佈那麼造成的問題並不大，但是若有一定的分佈形式，我們就必須要特別注意了。如果在大量的資料中只有少數的資料缺失且呈隨機分佈，這樣的情形比較不嚴重且任何一種處理缺少資料項的方法計算結果都相當類似；但是在資料集合的大小屬於中等或更小的情形下若有不少的資料缺失，將會是極嚴重的問題，不幸的是現在並沒有明確的準則讓我們瞭解資料集合的大小對於缺少資料項能容忍的程度在哪裡。

遇到缺少資料項的情形時，可能採用的方法有三種：[7]

- (1) Case-wise Deletion：僅考慮沒有包含缺少資料項的樣本(case)，也就是只要某個樣本的其中一個以上變數資料缺乏，則忽略此樣本。
- (2) Pair-wise Deletion：統計值的計算例如相關係數的計算會是兩兩變數為一組來計算，若某一個樣本在兩變數間有一個以上變數有缺少資料項的情形，那麼計算這兩個變數時會忽略該樣本。因為這樣的計算是針對個別情形考慮的，因此若計算相關係數的時候使用 pairwise 的方式，那麼計算每個係數所用的樣本數將會隨著缺少資料項的與否而有不同。
- (3) Mean Substitution：將樣本缺乏的資料項以該變數的平均值補上。

2.1.4 特徵選取

A. 使用特徵選取的原因

在預測性資料掘取(predictive data mining)¹中常碰到的工作就是從原始資料

¹ 預測性資料掘取是資料掘取(Data mining)的一種，主要是指從資料經過統計或類神經等的方式建立模型，以便於往後遇到同類資料可以預測可能發生的結果來進行決策或其他動作。

大量的候選變數集合中找出適當個數且適合用來預測整個資料趨勢的部分變數集合。舉例而言，經由自動化過程或電腦處理收集所得的原始資料，通常就包含了數百或上千甚至更多的變數量，但是一些基本的資料分析方法，像是類神經網路分析、分類樹狀圖或一般的線性模型沒辦法應付如此龐大的變數量，如此高的維度，因此都會需要先做特徵選取的動作[17]。

大部分用來預測或分類的方法若是有 k 個變數和 N 個個案(a)要儲存的矩陣大小最少是 $k \times (k+1)/2$ ，也就是下三角矩陣的資料量；(b)要重複處理資料檔案以得到最佳化的預測模型，通常至少會需要將大小是 $k \times N$ 的完整資料集合儲存在記憶體中以加快計算時間。舉例而言，進行一些簡單的線性回歸分析，其基本的準備工作，像是算出資料的相關聯矩陣，在變數相當多時很快就會達到程式或硬體所能負荷的極限，再加上有時進行線性模型的配適(fitting)或最佳化會要求相關聯矩陣(或共變異矩陣)的反矩陣，這樣的計算在變數量大的時候會相當困難而且運算中因為有效數字造成的誤差(round-off error)會不斷的累積最後嚴重影響計算造成結果相當不穩定，所以在數值計算之前我們也會傾向作特徵選取的動作。

模型配適，參數估測或多項式的最佳化等在高維度的時候通常會有維度的困擾(“Curse” of Dimensionality)¹，當輸入資料的維度(變數個數)增加的時候，要找尋最佳解的工作困難度會成指數增長，複雜度會大大的提高，因此實際研究就會將原始資料過濾，不會拿太多的變數量作分析。

B. 特徵選取的方法

特徵選取其實隨著問題的不同也會有不一樣的作法，假如候選的變數群和我們有興趣的變數之間呈線性關係的話，可以很簡單的求相關係數(correlation coefficients)，和我們有興趣的變數相關係數越高的話排名越前面；而若不是呈線

¹ “Curse” of Dimensionality(Bellman, 1961, Bishop, 1995)。

性關係的話則已有許多不同的文獻加以論述，在此不多加介紹。

另外在連續相關變數的前提之下，可將變數的分佈範圍分為 k 段，之後去求出每一個變數的 F 和 p 分佈值¹，依據這兩種數值去做排名的依據。以圖 2.1(a)

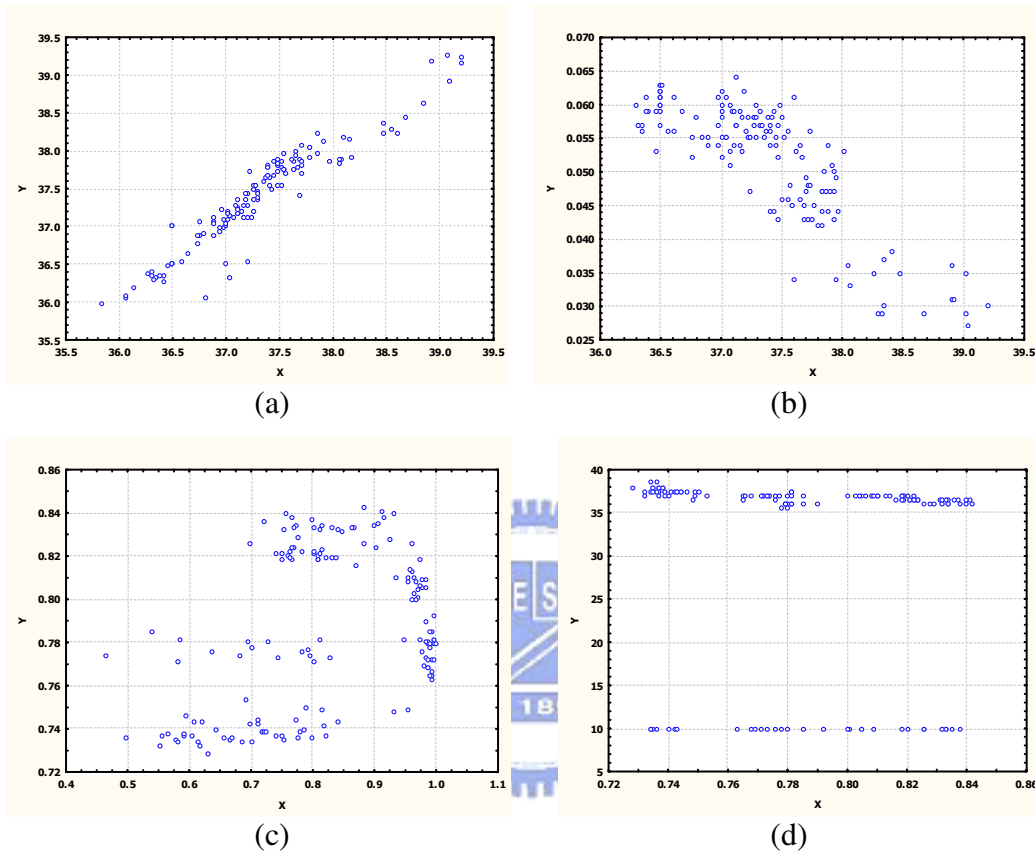


圖 2.2 分段去計算相關 F 值的例子

	(a)	(b)	(c)	(d)
r	0.9492	-0.8429	0.3917	-0.0681
F-value	181.8921	109.2766	17.4037	0.5258

表 2.1 圖 2.1 中 F 值與相關係數 r 的變化

而言，我們就是以 Y(目標變數)為主用特徵選取的方法去決定到底 X(待測變數)

¹ F 值和 p 值計算見附錄 A。

對於 Y 影響大不大。假設此時的特徵選取是會將待決定的變數分佈範圍分為五段，在圖 2.1(a)中就是將 X 的分佈範圍平均分為五個部分(Var1 的分佈是 35.5~39.5，所以分五等分的話是以 0.8 為間隔，依序是 35.5，36.3，...，39.5)，每一個部分的 X 都可以各自求出對 Y 的平均值，F 值的計算就是以這五個平均值求得。其中 F 值越大的話代表的是每個部分所得 Y 平均值之間的差異越大，也就是當 X 改變的時候 Y 就會跟著改變，而且 F 值越大 X 的改變會造成 Y 更大的變化。

以分段去計算 F 值的特徵選取方法資料間的分佈就不一定要像圖 2.1(a)(b) 呈近似線性的分佈，對於(c)(d)呈現不規則的非線性分佈，我們仍然能夠將資料分段，計算 F 值去探討每一段變化的大小以決定待測變數是不是影響目標變數甚大。如表 2.1 所示，可以發現 F 值越大其兩個變數的相關性越強。



2.2 多變量分析

2.2.1 多變量分析簡介

藉由多變量分析我們可以將變數分群，目的是希望以分群進行 GGMs 的方式簡化模型建立的過程，會使用到的有因素分析和多維縮放比例兩種方法。

2.2.2 因素分析

因素分析(Factor Analysis, FA)最主要的用途在於將變數分類以及降低變數量。我們可以運用因素分析將相近的變數分成同一群，並且找出新的潛在變數(例如因素, factor)或從該群中已知的資料裡選出最能代表該群的變數，這樣的動作也更有利於我們對每一群變數特性的瞭解；或者我們也可以運用因素分析從一大群變數當中找出適量(小群的)足夠代表原始資料的變數以進行其他的研究[5,23]。

因素分析是依據模型

$$y = Ax + e, \quad (2.1)$$

其中 $y = [y_1 \dots y_m]^T$ 我們已知的 m 個變數是由數個潛在的因素 $x = [x_1 \dots x_n]^T$ 的線性

組合加上干擾或誤差的部分 $e = [e_1 \dots e_m]^T$ 。在這個模型中假設：

- x 的平均值為零，且其變異量為有限值。
- e (平均值為零) 為彼此不相關的隨機變數也就是

$$\Sigma_e := E_{ee^T} = \text{diag}\{\sigma_1^2, \dots, \sigma_m^2\} \quad (2.2a)$$

且 x 和 e 之間存在如下的關係

$$\Sigma_{xe} := E_{xe^T} = 0 \quad (2.2b)$$

A 又稱為因素矩陣，

這個模型主要的目的是要藉著較少數的共同因素 x 在不考慮干擾或變化誤差的情況下去解釋已知變數之間彼此依賴的關係，由這個觀點我們可以將原來的式子改成

$$\hat{y}_i := a_i^T x_i \quad (2.3)$$

其中 a_i^T 代表的是矩陣 A 中第 i 個列，且對於 $i \neq j$ 我們有

$$E_{y_i y_j} = E_{\hat{y}_i \hat{y}_j} \quad (2.4)$$

的關係，這是給定 X 隨機變數 $\{y_1, \dots, y_m\}$ 垂直(conditional orthogonality)的關係或在高斯分佈的例子稱條件獨立的關係，簡單來說，就是 y_i 彼此間僅透過共同的因素 X 而相關。在此定義一個表示向量 $\{v_1, \dots, v_m\}$ 線性組合的 m 維隨機變數，

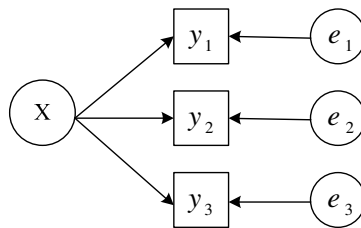


圖 2.3 模型條件獨立的關係

其內積空間為 $H(v)$ 。因此垂直的特性相當於是由 $\hat{y} := Ax$ 線性組合的隨機變數，其子空間(記做 $X := H(\hat{y})$)的特性，且任何類型的 X 集合都成立。一個能讓 y 內部元素呈條件獨立的子空間 X 稱為 $\{y_1, \dots, y_m\}$ 的分離子空間(splitting subspace)，而變數 \hat{y}_i 只是垂直的投影 $\hat{y}_i = E\{y_i | X\}$ 。

任何 X 的變數集合都可以當作是因素向量。在不失一般情形下選擇正規化的基底，也就是

$$E_{xx^T} = I \quad (2.5)$$

是不錯的選擇。因此我們可以知道 $n = \dim X$ 為模型線性獨立的列數(rank)。明顯的 $\text{rank}A = n$ 。而對於兩個表示同一群變數 y 的因素分析模型若因素形成相同的分離子空間 X ，則這兩個模型相等。

通常因素不是給定的，而且選擇的方法也隨著模型的特性和複雜度有所不同，我們通常希望使用 $n \ll m$ 的模型，找到最少量的必要因素去表示 y 。最少量因素的模型相對的其分離子空間 X 的維度也最小。

事實上給定已知的變數 $\{y_1, \dots, y_m\}$ 將會有無數個最小分離子空間，因此很多不同的因素分析模型可以同時表達固定的 m 維給定變數 y 。舉例來說選擇對於所有 $k \in \{1, \dots, m\}$ 、 $x := [y_1 \dots y_{k-1} y_{k+1} \dots y_m]^T$ ，我們可以得到一個獨特的因素分析模型稱為基礎回歸模型(elementary regressions)，形式如下：

$$\begin{cases} y_1 = [1 \dots 0]x + 0 \\ \vdots \\ y_k = \hat{a}_k^T x + e_k \\ \vdots \\ y_m = [0 \dots 1]x + 0 \end{cases} \quad (2.6)$$

其中 $\hat{a}_k^T = E_{y_k} x^T (E_{xx^T})^{-1}$ 。值得注意的是在這個模型中誤差變異矩陣只有一項不為零。

現在另一個問題是如何決定我們的因素，在決定因素個數的時候除了要 n 最

少之外也希望選擇的因素量足夠表達給定的變數 y ，那究竟多少因素才足夠呢？在這裡我們可以根據主要成分(principal components)的方法去決定。

所謂的主要成分是其實也是因素分析的一種，它假設因素的是量和給定的變數數量相同，而每一個因素是所有給定變數的線性組合， $x_i = \omega_1 y_1 + \omega_2 y_2 + \dots$ 。我們藉由求取主要成分分析中的特徵值(eigenvalue)來決定因素的數量。

此外因為在主要成分的方法中，第一個因素通常對於變數的模型影響最大，所以在決定完因素的個數之後我們通常會希望將因素的影響力平均分散到每個因素上，我們會將因素的結構加以旋轉。

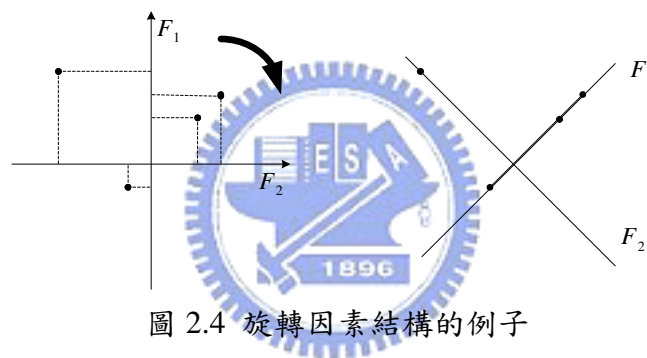


圖 2.4 旋轉因素結構的例子

最後，事實上因素不一定要彼此互相垂直，有時候為了找尋較適當且簡單的結構可以選擇傾斜(oblique)的因素，選擇互相垂直的因素主要是因為計算處理較簡單，而選擇傾斜的因素整個結構可能會更簡單，但是因為因素之間會存在相依關係，計算上會相對的更複雜。

2.2.3 多維縮放比例

多維縮放比例(Multidimensional Scaling, MDS)分析可以算是另一種形式的因素分析，通常這個分析方法的主要目的是要找尋變數間潛在的位置關係，用距離來解釋變數彼此相似的程度[15]。多維縮放比例使用的主要前提是變數資料可

以多維度的數值表示，因此變數可以在多維度的空間中定址，變數間相似的關係就和多維空間中的距離成反比關係。明可夫司基距離計算式(Minkowski distance metric)提供了一個普遍在多維度空間計算距離的方法：

$$d_{ij} = [\sum_{k=1}^n |x_{ik} - x_{jk}|^r]^{1/r} \quad (2.7)$$

其中 n 是維度， x_{ik} 是變數 i 在維度 k 時的數據資料。當 $r=2$ 時，計算式相當於二維空間的歐基理德距離計算式(Euclidian distance metric)。

多維縮放比例可以用在很多不同的用途上：用於資料的分析上，如果將要分析的變數放在較低維度的空間，資料觀察的複雜度有時候可以從原來的資料矩陣降低同時仍能保留必要的資訊；用於心理醫療資料上，多維縮放比例可以顯示人類決策是如何產生的，提供心理資料一些合理的解釋；多維縮放比例也常用在分類或辨識的數學模型中當作資料處理的基礎工作。

多維縮放比例不那麼像是完整的計算處理程序，而比較像是將變數重新安放在研究者給定的維度空間中，因此我們就需要在給定的維度空間中找到最近似的位址，並且去估算變數間近似的相對位置。

Kruskal(1964)提出了一個用來評斷距離 d_{ij} 和所要觀察相似程度 o_{ij} 的 stress 函式如下，

$$s = \sqrt{\frac{\sum_{ij} (d_{ij} - d_{ij}^*)^2}{\sum_{ij} d_{ij}^2}} \quad (2.8)$$

其中計算了預測的距離 d_{ij} 和我們所要達到的目標距離 d_{ij}^* 之間差異程度。藉由反覆計算，我們可以找到變數間最佳近似的位置。Borg 和 Groenen 提出了用來在多維空間中移動變數位置找尋最佳解的遞迴方法[11]，不過最主要還是要讓 stress 越小越好。

接下來一個重要的問題是如何選擇適當的維度？當我們配置變數的時候使用高維度雖然可以得到很低的 stress 值，不過結果卻很難由人眼觀察，而且還容易受到雜訊的影響而產生不好的結果；反之使用維度太少可能會喪失太多的資

訊。

最常用來選擇使用維度的方法是透過 scree 測試，這是藉由 stress (或相關評量最佳近似的方法)對維度的作圖來決定。這個測試方法最先由 Cattell 針對因素分析中因素個數的討論而提出的[20]；Kruskal 和 Wish 這個測試方法用在維度的決定上[19]。



第三章 圖形化模型

3.1 圖形化模型簡介

圖形化模型可以稱的上是結合了機率和圖學的理论，大致上來說，它提供了兩項吸引人的因素：首先圖形化模型可以視為一個具抽象特性但條理分明的視覺化語言，藉著這項語言，我們可以從多變量隨機程序的觀點去描述真實系統或訊號的一些重要特性(像是條件獨立的特性)。圖形化模型之中的每種圖形都有自己一系列相對應的數學方法提供我們去解釋圖形的意義。其次，伴隨著圖形化模型的是一些有利於我們做機率推斷和決策處理的理论，相較於傳統的機率推測方法，藉著圖形化模型的推論過程和其圖形上的抽象特性我們能大大的減少計算量 [14]。

簡單來說，圖形化模型描述的是隨機變數中條件獨立(conditional independence)的特性，給定圖形化模型也就是提供了我們一連串變數條件獨立的關係。所謂的條件獨立就是指在某種給定的機率分佈 $p(\cdot)$ 下三個不同的隨機變數 X 、 Y 、 Z 對於所有的 x 、 y 、 z 有著以下的關係：

$$p(X = x, Y = y | Z = z) = p(X = x | Z = z)p(Y = y | Z = z) \quad (3.1)$$

則稱為條件獨立，記作 $X \perp\!\!\!\perp Y | Z$ 。換句話說當我們知道 Z 的情形下， Y 對 X 並不會產生影響，反之亦然。條件獨立是一個強而有力的概念，透過這個概念統計的模型可以經過一連串的程序加以簡化或轉變成我們所需的型態。

一個圖形化模型是由圖形(graph) $\mathcal{G} = (V, \mathcal{E})$ 所組成的，其中 V 是頂點(vertices，也稱節點或隨機變數)的集合，而邊的集合 \mathcal{E} 則是 $V \times V$ 的子集合。一個圖形描述的就是變數集合 V 之間所有的機率分佈關係，而邊 \mathcal{E} 的型態則隨著圖形的種類不同有所差異(有無方向性或其他類型)。圖形化模型的種類很多，每一種使用的狀況都不同，種類包括了有方向性的模型(DGMs Directed Graphical Models 或稱貝氏網路, Bayesian Networks)、沒有方向性的模型(UGMs, Undirected

Graphical Models)、鍊結網路(chain graphs)、可分解模型(decomposable models)等許多種類型。一般而言，不同的圖形選用的條件都有各自相對應的資料分佈，不過有時候也會有重複的地方(也就是一種資料分佈可以用兩種圖形以不同的觀點去解釋)[14]。本文所選用的圖形化高斯模型則是屬於 UGMs 的一種。

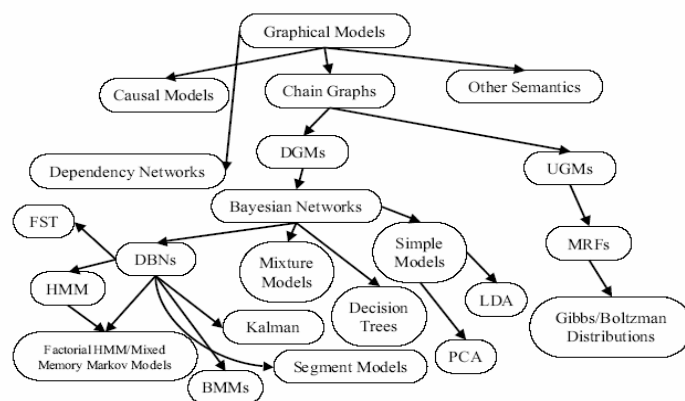


圖 3.1 圖形化模型的種類

3.2 圖形化高斯模型原理

3.2.1 圖學簡介

在討論模型的原理之前，必須對圖形理論的某些定義有基本的認識，因此這個小節把往後會用到的概念簡單做介紹。

一個圖形， $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ ，是由有限項的頂點(或稱節點)集合 \mathcal{V} 和有限項的邊(或稱弧，arcs)集合 \mathcal{E} 所組成的結構，用邊 \mathcal{E} 來表示頂點 \mathcal{V} 連接的關係。在本篇論文裡頂點是以 x_1, x_2, x_3, \dots 的方式表示的，而每個頂點在要處理資料中都有各自對應的變數。我們將邊表示成 $[x_1 x_2]$ 或相同於 $[x_2 x_1]$ ，而且邊是沒有方向性的(見圖 3.2(a))。

假如我們說兩個頂點 $x_1, x_2 \in \mathcal{V}$ 接鄰(adjacent)，記做 $x_1 \sim x_2$ ，表示的是兩者之間有邊連接，也就是 $[x_1 x_2] \in \mathcal{E}$ ，舉例而言，圖 3.2(a)中 x_2 和 x_3 接鄰而 x_3 和 x_4 則沒有接鄰。

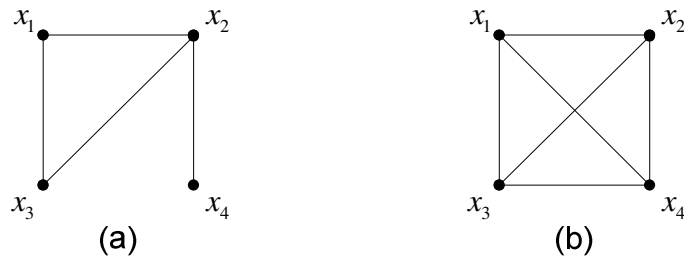


圖 3.2 (a)非完整圖形(b)完整圖形

一個圖形完整(complete)代表的是圖形中的每個頂點彼此之間都有邊相連接，舉例而言 3.2(b)則為完整。

任何 V 的子集合 $u \subseteq V$ 形成 G 的次圖形(subgraph)，即圖形 $G_u = (u, \mathcal{F})$ ，其中邊集合 \mathcal{F} 是由集合 \mathcal{E} 裡和 u 的頂點連接相關的邊所組成的。若稱子集合 $u \subseteq V$ 完整，則表示該集合能形成完整的次圖形，換句話說假如所有在 u 中的頂點彼此接鄰則形成完整圖形，例如圖 3.2(a)裡子集合 $\{x_1, x_2, x_3\}$ 完整。

假如一個子集合 $u \subseteq V$ 在某個圖形中為最大的完整圖形(maximally complete)則稱 u 為圖群(clique)，也就是假設 u 是完整且 $u \subset w$ 則 w 不是完整的，圖群的概念對於圖形化模型相當的重要，常常我們會需要辨識出圖形中有哪些圖群存在，舉例而言圖 3.2(a)中的圖群是 $\{x_1, x_2, x_3\}$ 和 $\{x_2, x_4\}$ 。

頂點 x_0, \dots, x_n 若對於 $i=1 \sim n$ ， $x_{i-1} \sim x_i$ ，則稱此為介於 x_0 和 x_n 之間長度為 n 的路徑(path)。舉例來說，圖 3.2(a)中 x_4, x_2, x_3, x_1 為介於 x_4 和 x_1 長度為 3 的路徑，而 x_4, x_2, x_1 為介於 x_4 和 x_1 長度為 2 的路徑。

一個路徑 $x_1, x_2, \dots, x_n, x_1$ 稱做 n -循環(n -cycle)或長度為 n 的循環，例如圖 3.2(b)中 x_1, x_2, x_4, x_3, x_1 為一個 4-循環。假若 n 個頂點 x_1, x_2, \dots, x_n 組成 n -循環 $x_1, x_2, \dots, x_n, x_1$ 且其中 $x_j \sim x_k$ ， $|j-k|=1$ 或 $n-1$ 則稱此為無弦循環(chordless cycle)，如圖 3.3(a)(b)都包含了無弦 4-循環，其中第一個圖中若 x_1 和 x_3 有邊連接或 x_2 和 x_4 有邊相連接則不為無弦循環。所謂的弦(chord)是在無方向性的圖形中，若每個長度大於三的

循環就會有弦的產生，也就是循環中兩個不連貫頂點間的連線，如圖 3.3(b)的 $[x_9, x_7]$ 或 $[x_8, x_6]$ 。

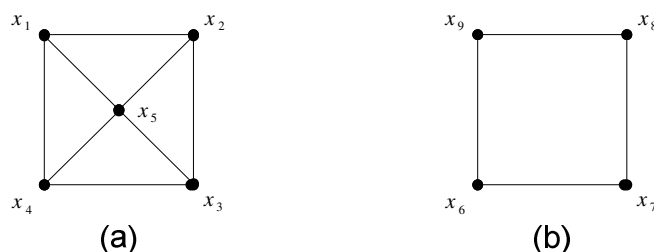


圖 3.3 兩個包含無弦 4-循環的圖形

如果一個圖形沒有包含長度大於或等於 4 的無弦循環則稱此圖形三角化 (triangulated)，以圖 3.3(a) 為例，若加上 $[x_1, x_3]$ 則該圖形有三角化，三角化的特性對最大近似估計的動作有著相當重要的關係，所以是我們必須特別注意的。

對於三個屬於 V 的子集合 a, b 和 s 假如所有從 a 到 b 的路徑所組成的集合 $m \subseteq s$ 則稱 s 將 a 和 b 分隔 (separate)。

到目前為止，模型究竟是如何和圖形扯上關係的呢？在圖形化模型中我們注重的是模型條件獨立的特性，例如 $X \perp\!\!\!\perp Y \mid \omega$ (其他變數)，這裡其他變數指的是模型說除了 X, Y 之外的所有變數。這樣的模型我們就可以建立一個沒有方向性的圖形 (V, \mathcal{E}) ，其中 V 代表模型中變數所形成的集合， \mathcal{E} 表示的是和條件獨立無關的其他邊組成的，也就是說對於所有 $\{X, Y\}$ 若 $X \perp\!\!\!\perp Y \mid \omega$ (其他變數)，則介於 X 和 Y 之間的邊將被去除或忽略，而其他的頂點間存在邊相連。因此從圖形可以發現，假使兩個變數沒有接鄰那麼給定其他變數的情況下兩變數條件獨立。這是用於非方向性圖形針對兩兩討論的馬可夫特性 (pairwise Markov property for undirected graphs)。而主要用來解釋這類圖形的是用於非方向性的全域性馬可夫特性 (global Markov property)，敘述如下：

若兩個變數集合 u 和 v 被第三個變數集合 ω 分隔，則 $u \perp\!\!\!\perp v \mid \omega$ 。

舉例而言，我們可能有四個變數 x_1, x_2, x_3, x_4 組成的模型，且 $x_1 \perp\!\!\!\perp x_4 \mid (x_2, x_3)$ ，

$x_3 \perp\!\!\!\perp x_4 | (x_1, x_2)$ ，則邊 $[x_1x_4]$ 和 $[x_3x_4]$ 必須從 $\{x_1, x_2, x_3, x_4\}$ 的圖形中去除，見圖 3.2(a)，從圖形我們可以推斷 $x_1 \perp\!\!\!\perp x_4 | x_2$ 且 $x_3 \perp\!\!\!\perp x_4 | x_2$ 。因此全域性馬可夫特性提供我們圖學特性中分隔和統計特性中條件獨立很簡單的概念轉換[3]。

全域型馬可夫特性反過來說也是成立的：假如對於任何包含於 V 的集合 u, v, ω ， $u \perp\!\!\!\perp v | \omega$ 在這個模型裡所有情形下都成立則 ω 在圖形中將 u, v 分隔。所以我們可以直接從圖形得到模型中獨立(marginal independence)和條件獨立的特性。這樣的性質稱為馬可夫的完整性(Markov perfection)¹。

最後給定一個圖形，我們藉由分別找出圖形中的圖群來建立模型的表示式，舉例來說，圖 3.4 中圖群為 $\{A, B, C\}$ ， $\{B, C, D\}$ ， $\{B, D, E\}$ 則相對應的表示式為 $//ABC, BCD, BDE$ ， $//$ 僅是本文代表模型表示式的符號。

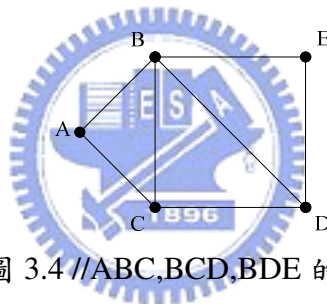


圖 3.4 $//ABC, BCD, BDE$ 的圖形

3.2.2 圖形化模型的理論根據

假設 $Y = (Y_1, \dots, Y_q)'$ 是一個維度為 q 呈多變量常態分佈的隨機變數，平均值為

$$\mu = \begin{pmatrix} \mu^1 \\ \vdots \\ \mu^q \end{pmatrix} \quad (3.2)$$

而共變異矩陣為

¹ 見 Frydenberg, 1990 或 Geiger 和 Pearl, 1993。

$$\Sigma = \begin{pmatrix} \sigma^{11} & \cdots & \sigma^{1q} \\ \vdots & \ddots & \vdots \\ \sigma^{q1} & \cdots & \sigma^{qq} \end{pmatrix} \quad (3.3)$$

我們有興趣的是共變異矩陣的反矩陣 $\Omega = \Sigma^{-1}$ ，記做

$$\Omega = \begin{pmatrix} \omega^{11} & \cdots & \omega^{1q} \\ \vdots & \ddots & \vdots \\ \omega^{q1} & \cdots & \omega^{qq} \end{pmatrix} \quad (3.4)$$

這個矩陣通常稱為精確矩陣(precision matrix)或濃縮矩陣(concentration matrix)。

現在，給定 (Y_3, \dots, Y_q) 後 (Y_1, Y_2) 的條件分佈將會呈二維的常態分佈且共變異矩陣為

$$\begin{pmatrix} \omega^{11} & \omega^{12} \\ \omega^{21} & \omega^{22} \end{pmatrix}^{-1} = \frac{1}{\omega^{11}\omega^{22} - (\omega^{12})^2} \begin{pmatrix} \omega^{22} & -\omega^{21} \\ -\omega^{12} & \omega^{11} \end{pmatrix} \quad (3.5)$$

在這二維常態分佈下的相關係數為

$$\rho^{12 \cdot 3 \dots q} = \frac{-\omega^{12}}{(\omega^{11}\omega^{22})^{1/2}} \quad (3.6)$$

又稱為部分相關係數(partial correlation coefficient)。我們可以發現

$$\rho^{12 \cdot 3 \dots q} = 0 \Leftrightarrow \omega^{12} = 0 \quad (3.7)$$

換句話說，兩個變數其在精確矩陣相對應的項數值為零，則我們可以說”給定其他剩餘的變數之下此兩個變數獨立”。

圖形化高斯模型就是設定某些精確矩陣特定的項為零(也就是其部分相關係數為零)。舉例而言，假如 $q=4$ 並且假定模型中 $\omega^{13} = \omega^{24} = 0$ ，則精確矩陣為

$$\Omega = \begin{pmatrix} \omega^{11} & \omega^{12} & 0 & \omega^{14} \\ \omega^{21} & \omega^{22} & \omega^{23} & 0 \\ 0 & \omega^{32} & \omega^{33} & \omega^{34} \\ \omega^{41} & 0 & \omega^{43} & \omega^{44} \end{pmatrix} \quad (3.8)$$

則這個模型的圖形將是由每兩個頂點相對應的部分相關係數不為零者所組成的

[3]，如圖 3.4 所示。

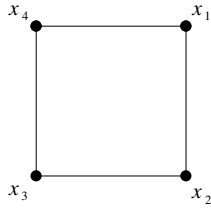
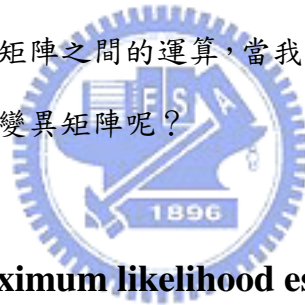


圖 3.5 $q=4$ ， $\omega^{13} = \omega^{24} = 0$ 的圖形

現在我們知道共變異矩陣和條件獨立的關係，接著就是要慢慢建立出適合於系統且最簡化的模型，在此採用的是最基本的貪婪策略(Greedy Method)，也就是剛開始假設所有變數彼此相關聯，接著每一回合去掉一個最不影響整個模型的邊(由偏差值和 p 值為比較基準)，在精確矩陣則是設定該項為零。因此接下來就是要討論由共變異矩陣和精確矩陣之間的運算，當我們設定精確矩陣的某一個項為零，如何去求得相對應的共變異矩陣呢？



3.2.3 最大近似估測(Maximum likelihood estimate, MLE)

假設有 N 個樣本 $y^{(1)}, y^{(2)}, \dots, y^{(N)}$ ，且 $\bar{y} = \sum_{k=1}^N y^{(k)} / N$ 為樣本平均值組成的向量， $S = \sum_{k=1}^N (y^{(k)} - \bar{y})(y^{(k)} - \bar{y})' / N$ 是樣本的共變異矩陣。那麼對數密度(log density)可以寫成

$$\ln f(y) = -q \ln(2\pi) / 2 - \ln |\Sigma| / 2 - (y - \mu)' \Sigma^{-1} (y - \mu) / 2,$$

其中 $f(y)$ 為 Y 的密度函數，可寫成

$$f(y) = |2\pi\Sigma|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(y - \mu)' \Sigma^{-1} (y - \mu)\right\},$$

所以樣本的對數近似值將會是

$$l(\mu, \Omega) = -Nq \ln(2\pi) / 2 - N \ln |\Sigma| / 2 - \sum_{k=1}^N (y^{(k)} - \mu)' \Omega (y^{(k)} - \mu) / 2,$$

其中最後一項使用跡(trace)函式(也就是方陣中對角線元素的總和)加以化簡

$$\sum_{k=1}^N (y^{(k)} - \mu)' \Omega (y^{(k)} - \mu) = \sum_{k=1}^N (y^{(k)} - \bar{y})' \Omega (y^{(k)} - \bar{y}) + N(\bar{y} - \mu)' \Omega (\bar{y} - \mu)$$

$$\sum_{k=1}^N (y^{(k)} - \bar{y})' \Omega (y^{(k)} - \bar{y}) = N \text{tr}(\Omega S),$$

則我們可以得到對數近似函數(log likelihood)

$$l(\mu, \Omega) = -Nq \ln(2\pi) / 2 - N \ln |\Sigma| / 2 - N \text{tr}(\Omega S) / 2 - N(\bar{y} - \mu)' \Omega (\bar{y} - \mu) / 2 \quad (3.9)$$

假如一個變數集合 $a \subseteq \Gamma$ ，且 Σ^{aa} 和 S^{aa} 為 Σ 和 S 相對於 a 的子矩陣，其中 S 代表的是原始樣本的共變異矩陣，而 Σ 是 S 經過反矩陣運算，因為圖形某些邊被去除而精確矩陣相對應的項被設為零，再求反矩陣所得到的矩陣，這樣的運算在變數量大的時候很難執行，所以我們考慮使用最大近似估測的方式。現在對於我們所要討論的變數 $a = q_1, \dots, q_t$ ，我們希望從相對應的 S^{aa} 去求 Σ^{aa} ，但是為了計算

考量我們去求 $\hat{\Sigma}^{aa}$ 近似真實的 Σ^{aa} 值，在此近似值就可以藉由模型已知的值來表示， $\hat{\mu} = \bar{y}$ ， $\hat{\Sigma}^{aa} = S^{aa}$ 。舉例而言，假設變數集合 $\Gamma = \{X, Y, Z\}$ ，其原始的共變異矩陣是 S

$$S = \begin{pmatrix} 3.023 & 1.258 & 1.004 \\ 1.258 & 1.709 & 0.842 \\ 1.004 & 0.842 & 1.116 \end{pmatrix}$$

在 $//XY, YZ$ 的模型下 Σ 的最大近似估測為

$$\hat{\Sigma} = \begin{pmatrix} 3.023 & 1.258 & 0.620 \\ 1.258 & 1.709 & 0.842 \\ 0.620 & 0.842 & 1.116 \end{pmatrix}$$

而其精確矩陣為

$$\hat{\Omega} = \begin{pmatrix} 0.477 & -0.351 & 0.000 \\ -0.351 & 1.190 & -0.703 \\ 0.000 & -0.703 & 1.426 \end{pmatrix}$$

近似估計的計算限定了 $\hat{\Sigma}$ 中除了 $\hat{\sigma}^{xz}$ 項之外，其他數值都和 S 相對應的項相同，

因為模型中設定 $\hat{\omega}^{xz} = 0$ ，所以造成 S 相對應的項數值不同。

Speed 和 Kiiveri 提出了一套用於圖形化高斯模型計算最大近似估計的演算法[22]。這個方法是藉著慢慢增加精確矩陣 Ω 的某部分子矩陣去滿足近似的要求。精確矩陣剛開始被設為單位矩陣，每一回合計算是針對我們要討論的 a 去計算增加量 E ，讓現在的 Ω 儘量達到 $\hat{\Sigma}^{aa} = S^{aa}$ 的要求。我們要求的增加量 E 由

$$\begin{pmatrix} S^{aa} & * \\ * & * \end{pmatrix} = \begin{pmatrix} \Omega^{aa} + E & \Omega^{ab} \\ \Omega^{ba} & \Omega^{bb} \end{pmatrix}^{-1} \quad (3.10)$$

其中*代表沒討論的部分，運用反矩陣的計算結果可得

$$S^{aa} = (\Omega^{aa} + E - \Omega^{ab} (\Omega^{bb})^{-1} \Omega^{ba})^{-1} \quad (3.11)$$

$$E = (S^{aa})^{-1} - (\Omega^{aa} - \Omega^{ab} (\Omega^{bb})^{-1} \Omega^{ba}) \quad (3.12)$$

$$= (S^{aa})^{-1} - (\Sigma^{aa})^{-1} \quad (3.13)$$

通常計算很快就會收斂，Buhl(1993)對於最大近似估計的存在條件有詳細的討論。另外因為我們計算所針對的 a ，其實就是由模型中的圖群所組成的，每一回合的每一個步驟都是針對某一個圖群去計算增加量，每一回合分別計算所有圖群造成的增加量，如此去逼近 $\hat{\Sigma}^{aa} = S^{aa}$ 的要求，所以我們在建立模型的步驟中會特別去注意保持模型是以圖群組成的。維持圖群的另一個原因是機率分佈可依全域型馬可夫特性分解(factorize)成許多完整項的乘積。也就是機率分佈存在

$$P(X) = \prod_{a \text{ is complete}} \psi_a(X) \quad (3.14)$$

的分解形式，舉例來說，圖 3.6(a)可分解成

$$P(X) = \psi_{v_1, v_2}(X) \psi_{v_2, v_3}(X) \psi_{v_3, v_4}(X) \psi_{v_4, v_5}(X)$$

的形式，而圖 3.6(b)可分解為

$$P(v_1, v_2, v_3, v_4) = \psi_{v_1} \psi_{v_2} \psi_{v_3} \psi_{v_4} \psi_{v_1, v_2} \psi_{v_1, v_2} \psi_{v_2, v_3} \psi_{v_2, v_3} \psi_{v_2, v_4} \psi_{v_3, v_4} \psi_{v_1, v_2, v_3} \psi_{v_2, v_3, v_4}$$

這個分解式包括了所有完整項的乘積，不過式子還能更加簡化，因為有些 ψ 會等

於 1，而且當 ψ_A, ψ_B ， $B \subset A$ 時我們可以用 $\psi'_A = \psi_B \psi_A$ 取代；在所有 ψ_B 都被取代之後，每個 ψ 並不包含其他的項，則此式子成為唯一的簡化且封閉型態(closed form)的表示式，也就是由圖群所組成的表示式，因此我們會維持模型以圖群的方式表示。

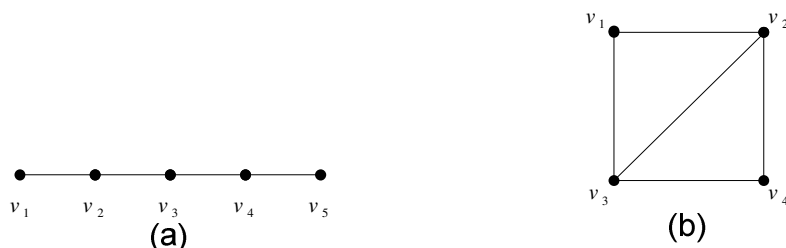


圖 3.6 分解的例子

3.2.4 最大基數搜尋測試(Maximum Cardinality Search, MCS)

此外要計算最大近似估計，模型還要保持可分割(decomposable)，所謂可分割的條件是圖形為三角化的。使用可分割模型主要的原因是在於可分割模型封閉型態的表示式有利於最大近似估計的計算，使得數值的估測較有效率；另外比起一般的模型，可分割模型在理論推理上和計算上都來的較易處理，像是之後用於判斷是否要去除邊的條件測試(例如卡方檢定測試)，雖然仍可以使用一般的模型去測試，不過計算量將會大增。

在模型的建立中我是以最大基數搜尋的方法去判斷一個模型是否可分割。最大基數搜尋的方法在圖學中主要是用於無方向性圖形中循環是否有弦存在的偵測[8,18]，由 3.2.2.1 弦的定義可知有弦圖形(chordal graph)指的是無方向性圖形中的子圖形皆是長度小於或等於三的循環，也就是三角化的圖形。最大基數搜尋的說明如下：

輸入：無方向性圖形 $\mathcal{G} = (V, \mathcal{E})$ ，且起始的 x_j 為 $V = \{x_1, \dots, x_n\}$ 中一個節點。

輸出：圖形是否為三角化的圖形？布林變數 complete 若為 true 則是三角化圖

形，否則為非三角化圖形。

演算法：Maximum Cardinality Search

1. 設定執行的步驟 $i=1$ ，且將 x_j 列為已編號節點中的的第一個點，也就是 $\alpha(1) = x_j$ 。初始化 $complete=true$ 。
2. 在未編號的節點裡找出跟已編號節點相連邊數最多者 x_k ，將 x_k 編號為 $i+1$ ，也就是 $\alpha(i+1) = x_k$ 。
3. 如果 $Nbr(x_k) \cap \{\alpha(1), \dots, \alpha(i)\}$ 中的節點沒有完整，則圖形不為三角化， $complete=false$ ，結束整個演算法。其中 $Nbr(x_k)$ 是與 x_k 相鄰節點的集合，完整即是任兩個節點都要連接在一起。
4. 假如 $i=n$ 演算法就停止，此時 $complete=true$ ；否則令 $i=i+1$ ，回到第二步驟繼續搜尋。其中 n 為節點的總數。

以圖 3.7 為例，輸入的頂點是 $V = \{x_A, x_B, x_C, x_D, x_E\}$ (在此以 A, ..., E 代表輸入時點的順序)，假設初始點為 x_C ，則當 $i=1$ ， $\alpha(i) = x_C$ (如圖 3.7(b))。第二步驟，從未編號的點中找尋與已編號節點相連邊數最多的節點，在這邊我們選到 x_B, x_C, x_E ，因為三者皆和一個已編號的節點 x_C 相連，此時由輸入的順序決定 x_B 為第二個編號點， $\alpha(2) = x_B$ (圖 3.7(c))。第三步驟要檢查的集合是 $\{x_C\} \cap \{x_C, x_B\} = \{x_C\}$ 完整，接下來 i 增加並回到步驟二去尋找。

值得注意的是在我們找到第四個編號點，也就是 $\alpha(4) = x_D$ ，要檢查完整性的集合， $\{x_A, x_C\} \cap \{x_C, x_B, x_A, x_D\} = \{x_A, x_C\}$ ，發現缺少了 $[AC]$ ，則 MCS 停止，此圖形為非三角化的圖形。若再以另一個圖形(圖 3.7(e)，圖(a)和(e)僅差 $[BD]$) 為例，則 MCS 的結果將如圖 3.7(f)，為三角化的圖形。

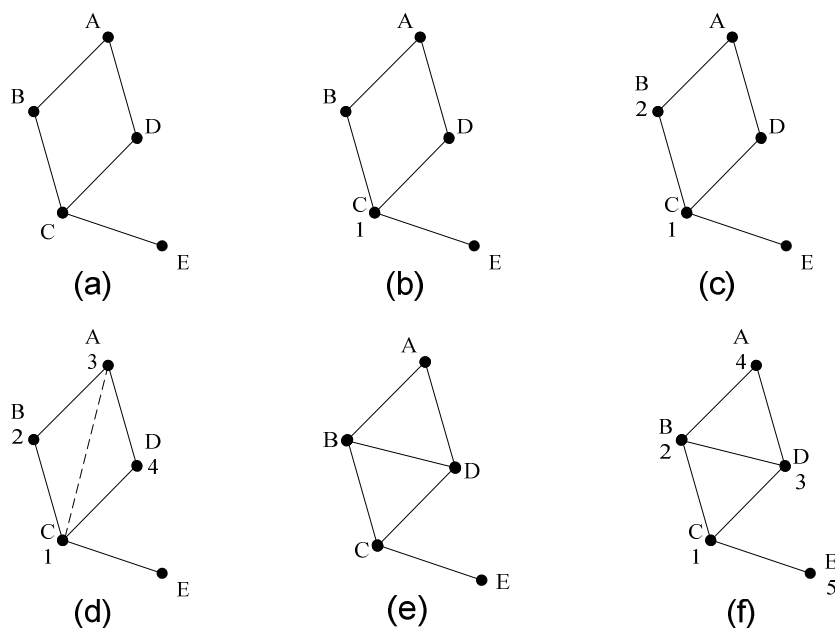


圖 3.7 運用 MCS 判斷圖形是否三角化，分別以(a)(e)兩個圖形為例

3.2.5 χ^2 測試和偏差值

某些邊通過最大基數搜尋的測試(也就是單獨去除某個邊後整個模型仍然維持可分割)，之後又經過最大近似估計的運算得到估計的共變異矩陣，接下來就是從這些候選的邊裡找出對於整個模型影響最小的邊予以去除。在此我們是計算原始和估測後的共變異矩陣兩者之間的偏差值(deviance)，之後再由偏差值做 χ^2 測試得到 p 值，根據 p 值的大小推斷邊的重要性。

使用我們求取最大近似估計前的假設，也就是 $\hat{\mu} = \bar{y}$ ，則(3.9)式中的最後一項消失，又因為 $\hat{\Sigma}$ 和 S 不同處僅在於 $\omega^{jj} = 0$ 的項，因此 $tr(\hat{\Omega}S) = tr(\hat{\Omega}\hat{\Sigma}) = q$ 。藉由以上的關係我們能將最大對數近似值簡化為

$$\hat{l}_m = -Nq \ln(2\pi)/2 - N \ln |\hat{\Sigma}|/2 - Nq/2$$

在完整的模型也就是 $\hat{\Sigma} = S$ 的情形下最大對數近似值可以寫成

$$\hat{l}_f = -Nq \ln(2\pi)/2 - N \ln |S|/2 - Nq/2,$$

因此模型的偏差為

$$G^2 = 2(\hat{l}_f - \hat{l}_m) \\ = N \ln(|\hat{\Sigma}| / |S|) \quad (3.15)$$

歸納之後若模型 $\mathcal{M}_0 \subset \mathcal{M}_1$ 則偏差值為

$$d = N \ln(|\hat{\Sigma}_0| / |\hat{\Sigma}_1|) \quad (3.16)$$

其中 $\hat{\Sigma}_0$ 和 $\hat{\Sigma}_1$ 分別為在 \mathcal{M}_0 和 \mathcal{M}_1 之下對 Σ 的估測矩陣。在 \mathcal{M}_0 的模型下， d 假設是呈 \mathcal{N}^2 的分佈 (chi-square distribution)，且其自由度是由 \mathcal{M}_0 和 \mathcal{M}_1 差異的邊數來決定的 (若 \mathcal{M}_0 是 \mathcal{M}_1 去除三個邊後得的新矩陣，則自由度為 3，依此類推)。

因為我們的偏差值假設的前提是 d 呈 \mathcal{N}^2 分佈，因此我們在求取 p 值就依據 \mathcal{N}^2 測試，也就是以

$$f(x) = \left[\frac{1}{2^{v/2} \times \Gamma(v/2)} \right] \times [x^{(v/2)-1} \times e^{-x/2}] \quad (3.17)$$

其中 v 為自由度而 Γ 為 gamma 函式， Γ 的定義 [6] 是

$$\Gamma(\alpha) \triangleq \int_0^{\infty} x^{\alpha-1} e^{-x} dx, \quad \alpha > 0 \quad (3.18)$$

經過部分積分運算之後可得 $\Gamma(\alpha) = (\alpha-1)! \Gamma(1)$ ，又 $\Gamma(1) = \int_0^{\infty} e^{-x} dx = 1$ 所以對於 $\alpha = 1, 2, 3, \dots$ ， $\Gamma(\alpha) = (\alpha-1)!$ ，而另外經由原始定義的計算能夠得到 $\Gamma(0.5) = \sqrt{\pi}$ ， $\Gamma(1.5) = \sqrt{\pi}/2$ 。

3.3 圖形化高斯模型建立流程

GGMs 方法針對的是共變異矩陣，因此我們要先計算資料的共變異矩陣。GGMs 採用逐步減去圖形中最適合邊的方式來建立模型。最剛開始會將每個變數視為彼此互相關聯的，在圖形中即是每個節點之間互相連接，因此 n 個變數將會

有 $C_2^n = \frac{n!}{(n-2)!2!} = \frac{n \times (n-1)}{2}$ 條邊。在每一回合中針對每一條連接的邊(未被去除的邊)去做最大近似估計的測試，所謂的測試是假設某一邊去除的話對於整個結構是否差異甚大，也就是求出假設已去除某一邊的共變異矩陣與未去除的相比較，求出兩者的偏差值，並且用卡方檢定求出一個 p 值，用於將偏差值較廣的分佈範圍濃縮到 0 和 1 之間便於找出最大的 p 值作為最適合去除邊的根據。在此我們規定 p 值若小於某個檢定量(critical value) α 的話(內定是 0.05，我們也可以視自己需要設定 α 為 0.01 或更低，較低的檢定量會增加計算的負擔和運算時間)必須保留¹。另外能列為候選的邊還必須符合可分解性，也就是去除掉候選的邊之後還必須保持整個圖形是三角化的，所採用的是最大基數搜尋的測試方法。

一個連接的邊在某一回合若是之前沒有被保留(前一步驟中 p 值 $> \alpha$)，若去除這個邊符合 MCS 的測試，而且經過 MLE 的運算發現去除這個邊計算出來的 p 值最大，則這個邊在這一回合將被除去，在去除的同時我們還要讓整個模型維持圖群的型態。如此一直測試和計算，每一回合去除一個邊並且更新整個模型和共變異矩陣。整個程序不斷重複直到某一回合再也找不到可以去除的邊為止。

GGMs 完整的建立流程見圖 2.1，其中 ECM(Edge Connection Matrix)儲存的是每個邊連接的情形(1 代表邊有連結，0 代表邊被打斷)，初始因為假設所有節點互相連接，所以每個項都是 1；SEM(Significant Edge Matrix)儲存的是每回合中 p 值 $< \alpha$ 的邊(1 代表被保留起來，之後都不會列入考慮；0 則沒有被保留)，以作為每回合尋找候選邊的根據。

¹ P 值高於 0.05 代表的是 deviance 值在高斯分佈中超過 95% 的偏差量，表示去除此邊會造成整個模型劇烈的改變，所以 p 值相當顯著，這樣的邊絕對不能忽略，必須保留下來，而且一旦一個步驟中某個邊被保留下來，則往後這個邊將不再列入可能被去除的候選測試當中。

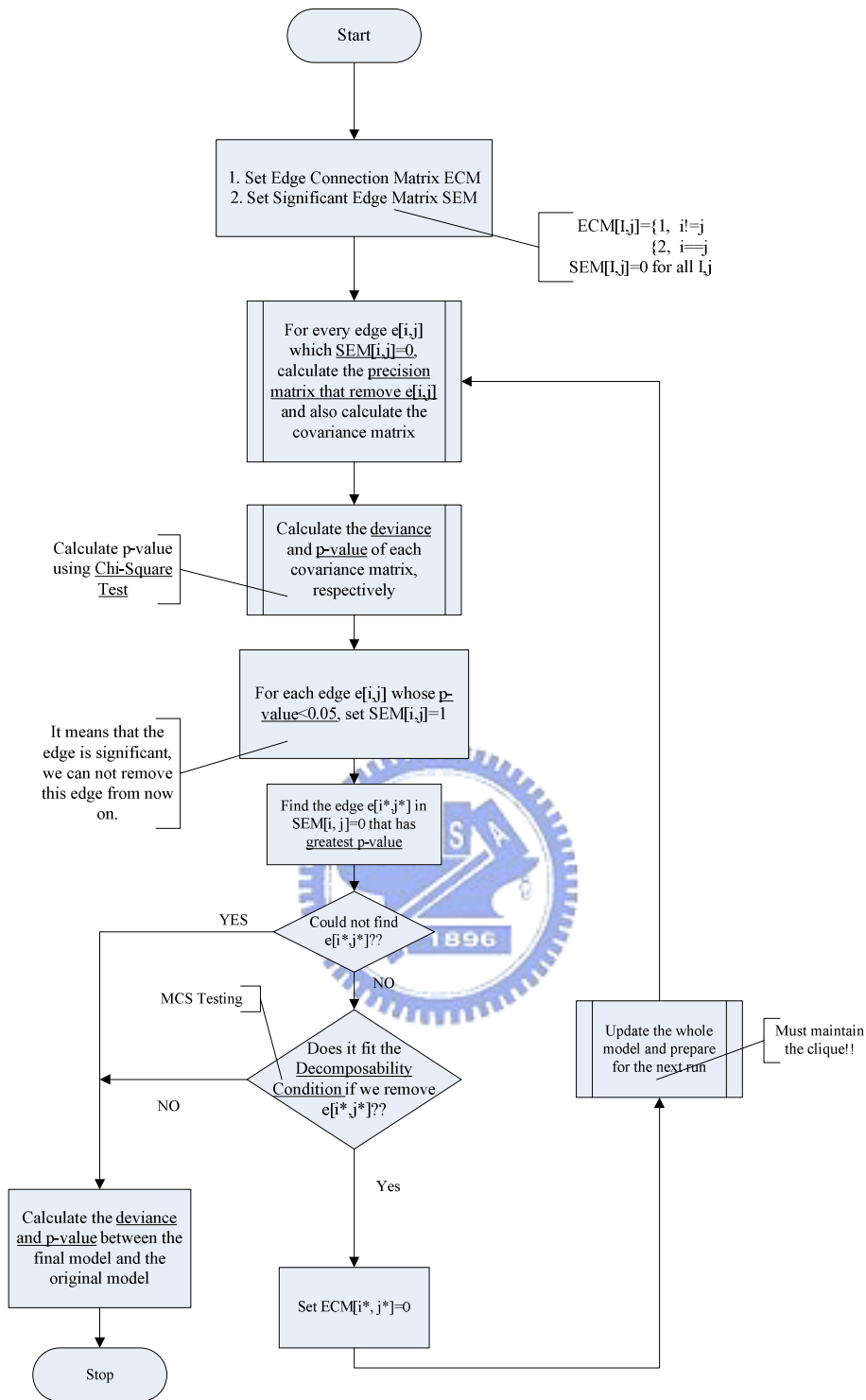


圖 3.8 GGMs 的建立流程圖

第四章 簡化方法的問題和改進

本章會針對簡化方法會遭遇到的問題以及一般方法中最大近似估測的問題作討論，並提出改進的作法，最後完整的提出簡化方法的步驟。

4.1 最大近似估計的問題

第 3.2.3 節所述的最大近似估測可說是建立 GGMs 中運算量最大的步驟，因為其中牽涉了大量逆矩陣運算，而且當一個模型中的圖群越多的時候每一回合所要估計的運算也會增加，這也就是建立 GGMs 會遭遇到的瓶頸。此外，因為最大近似估測是一個反覆運算的過程，每一回合都依照模型的圖群去作矩陣的估測，當本回合所得的結果和上回合的矩陣相比較，誤差小於一個特定值(δ)估測才算完成。現在又出現另一個問題：究竟 δ 應該取多少？原本使用的是固定數值 (10^{-5})，但是常常仍然會有計算過久，也就是估測值不易收斂的情形發生，所以我們另外採用 $n^2 \times 10^{-5}$ ，其中 n 是矩陣的行(或列)個數，也就是變數個數。這樣的好處是當矩陣越大， δ 也會跟著放大，放寬計算的標準減少不易收斂的情形，可是相對的是整個計算會累積較大的誤差，當 $n^2 \times 10^{-5}$ 的誤差並不是平均分散在矩陣的每一個項而是集中在某些特定的項，這樣就會造成估測嚴重的錯誤，是另外一個計算估測的難題。

4.2 降低簡化方法的偏差值

仔細地去檢查簡化模型的最大近似估測的過程，檢驗去除每個邊造成的偏差值可以發現較大偏差值的產生會集中在某些去除的邊上，以 7 個變數為例，經過分群建立模型的結果，ECM 矩陣如表 4.1 所示。其中 1 表示相對應的兩個變數之間有邊連接，0 表示邊被去除。

	Yield	Var4	Var7	Var10	Var13	Var21	Var22
Yield	/	0	1	0	1	1	0
Var4	0	/	1	1	1	0	1
Var7	1	1	/	1	1	1	1
Var10	0	1	1	/	1	0	1
Var13	1	1	1	1	/	0	1
Var21	1	0	1	0	0	/	1
Var22	0	1	1	1	1	1	/

表 4.1 7 個變數的 ECM 矩陣

7 個變數將良率獨自分為一群，Var4、Var7、Var10 為第二群，Var13 為第三群，Var21、Var22 為第四群。整個模型偏差值為 15.8848，但是實際上僅去除了良率和 Var4 之間的連線偏差值就已經到達 10.6147，去除這樣的邊對於整個模型就造成很大的影響，檢查計算過程中的數據也就是第一群和第二群變數之間去計算邊的關係之過程可以發現去除該線的確符合我們定的條件。這樣的訊息表示每群各自去建立連接關係的作法雖然降低了運算量卻缺乏了整體的考量，也就是在每一群之中或某兩群之間去除某些邊(此時不考慮其他的變數)並不會造成多大的偏差，但是這些邊卻對整個模型有相當大的影響。因此改進的方法是在分群建立模型之初使用最大近似估測將所有邊去除後所造成的偏差值預估出來，將會產生過大偏差值(去除後的 p 值會大於 0.95)的邊先保留下來，再去進行分群建立模型的動作，從巨觀的角度去衡量哪些邊對整體模型較為重要，以減少分群建立簡化模型造成的偏差。

4.3 簡化方法的完整步驟

本節將論文中提到的簡化方法以及 4.2 節所提的改進程序再作更清楚的說明，現在整體的架構如圖 4.1 所示，原始資料會先經過包括找尋歧異值和特徵選取的前置處理以去除異常的資料分佈，並找出跟良率關係較密切的變數排名。以

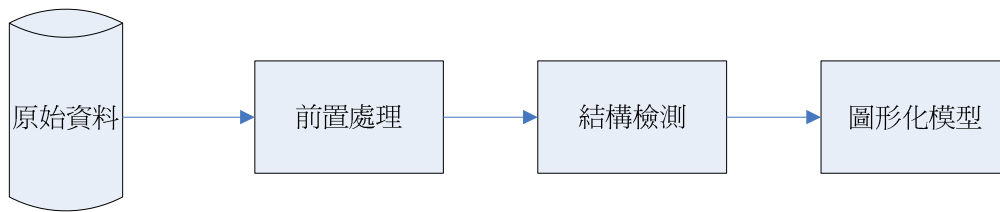


圖 4.1 簡化方法的完整架構

圖 4.2(a)為例，經過前置處理後我們將特徵選取中前五名變數與良率來建立模型。圖 4.2(b)顯示的是結構檢測部分，我們將圖 4.2(a)選擇出來的變數經過 FA 和 MDS 的分群步驟，資料分佈較類似的變數分為同一群，在此我們僅把解釋變數加以分群，而如良率的被解釋變數則自成一類。再來為了避免分群會造成偏差值

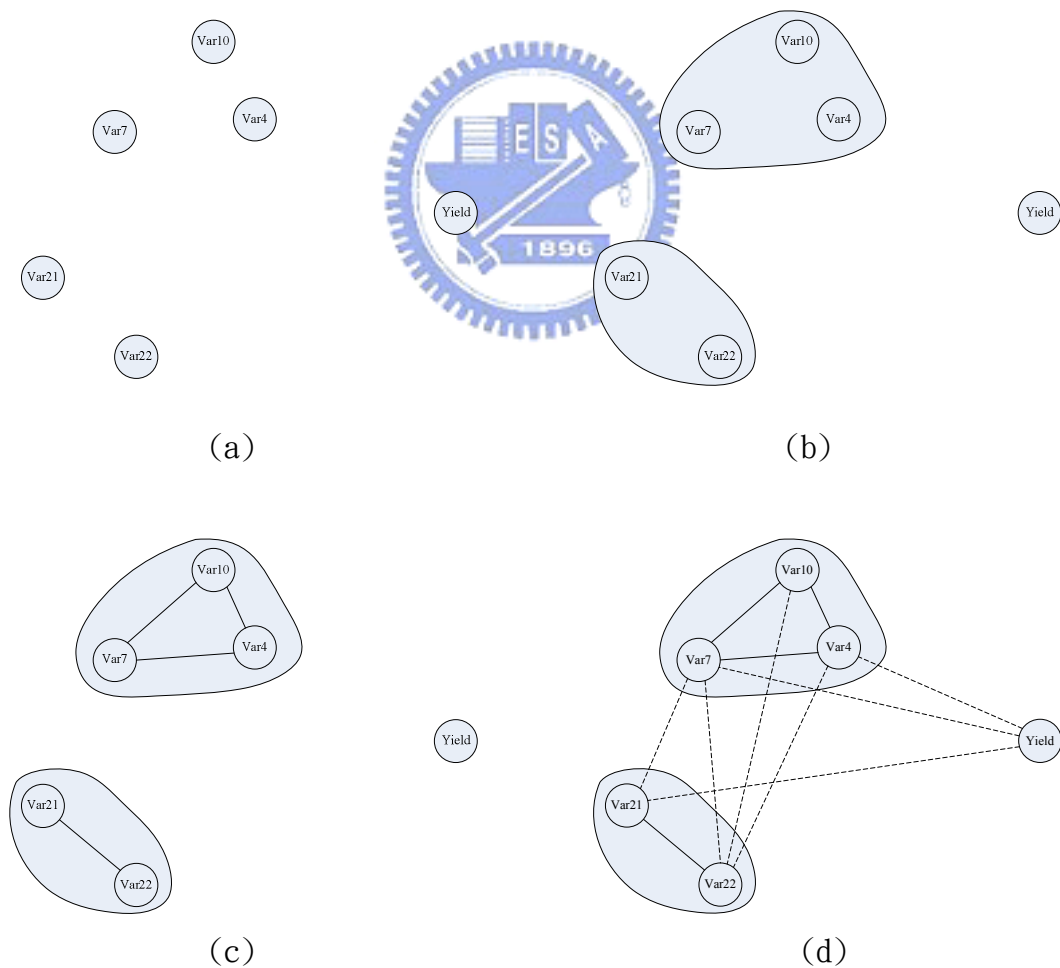


圖 4.2 簡化模型的建立步驟

不正常分佈，我們根據 4.2 節所述的方法將圖形中的每一條邊進行 3.2.3 節所提最大近似估測的手續，以圖 4.2(b) 為例，我們會檢查 $C_2^6 = \frac{6!}{(6-2)!2!} = 15$ 條邊，把去除後會造成偏差值大於 3.84 的邊(因為本例設定 $\alpha = 0.05$)保留起來，往後的過程中這些保留的邊將不再列入刪除的考量。

接著是依照第三章的方法以群為單位去建立 GGMs，結果如圖 4.2(c) 所示。最後是各群間建立連接的關係，如圖 4.2(d) 所示的虛線部分，在此步驟僅去討論各群之間的關係，而不再去討論每一群中邊的連接與否(也就是不會去更改由圖 4.2(c) 所建立出來實線的部分)。藉著以上的過程，我們將簡化的模型建立出來，除了降低模型的建立時間和運算量，同時維持整體的偏差值在我們允許的範圍之內。



第五章 實驗與討論

5.1 實驗資料簡介

實驗所使用的數據是由一個被解釋變數也就是良率以及 30 個解釋變數構成，他們的名稱分別是 Var1(Yield)、Var2、Var3、...、Var31，採用的是 Statistica 資料集中的數據資訊，這些資料是實際晶圓自動化生產製程所得的原始數據(如 1.1 節所述的原始資料可能有上千個變數)，經過一些類似特徵選取的前置處理之後所得的較小部分資料集合，總共有 154 個樣本數(晶圓數)且皆為數據資料而非分類後的結果。資料的特性和一些我們在處理上可能遭遇到的問題在 1.1 節已經有概略的討論，在此不再描述。

本論文主要的目的是以上述資料建立符合我們要求的 GGMs 模型，藉著模型來探討解釋變數對於良率的影響和各解釋變數之間互相的關聯性，之後能再結合其他方法，諸如專家系統或貝氏網路進行錯誤的診斷或良率的預測。

因此實驗的部分會從原始資料出發，先進行前置處理，包括歧異值的問題、缺少資料項的問題和特徵選取的過程將解釋變數根據變數與良率相關程度的高低做出排名，接著對一般 GGMs 作測試，也就是找出其所能負擔的最大運算量或變數個數，再討論不同檢定量對於模型建立的時間、去除邊的個數和偏差值的影響，並且觀察每去除一邊對於整體模型造成的變化。之後是簡化方法的探討，包括分群的問題，也就是以因素分析和多維縮放比例互相印證的分群結果去建立簡化的 GGMs，並對簡化方法會遇到的問題進行改進，最後比較一般和改進後的簡化方法兩者偏差值和時間的差異。實驗中的部分方法會使用 Statistica¹ 軟體進行運算，主要 GGMs 部分則是使用 Borland C ++ Builder² 撰寫程式進行模型的建立。

¹ <http://www.statsoft.com/>

² <http://www.borland.com/bcppbuilder/index.html>

5.2 前置處理

5.2.1 歧異值的問題

A. 找尋歧異值

根據 2.1.1 節，多變量的分析我們可以藉著對每一個樣本計算馬哈拉諾畢斯距離找出歧異值，表 5.1 依照距離大小進行排名，列出前 10 名的樣本，依照排名第 25 和 83 個樣本相當偏離正常資料的分佈，很可能是歧異值的所在，但是再根據良率和所有解釋變數的二維分佈圖可以發現雖然刪除第 83 個樣本對於整體資料分佈影響不大反而是馬哈拉諾畢斯距離只有 18.2 的第 90 個樣本對於某些解釋變數的分佈造成相當大的影響，因此選定第 25 及 90 個樣本為歧異值，在往後的過程中這兩個樣本會予以刪除。

樣本編號	25	83	55	64	143	23	58	68	142	51
距離	138.7654	128.6784	55.60106	52.25884	42.33746	37.24694	37.08688	36.1336	33.65033	33.43491

表 5.1 154 個樣本各自去計算馬哈拉諾畢斯距離找出的前 10 名

圖 5.1 分別顯示了去除這兩個歧異值之後對部分資料分佈的影響，去除歧異值之後我們更能瞭解資料正常的分佈情形，其中(b1)(b2)明顯的顯示出去除第 90 個樣本對於資料分佈造成的影響。正因為歧異值對於變數分佈的影響不小因此去除歧異值會連帶的影響到特徵選取的排名。

B. 歧異值對特徵選取的影響

表 5.2 顯示的是所有變數對於良率的特徵選取排名(見 2.1.3，使用 pairwise deletion，並將資料分為五個區段)，其中部分變數的排名互有先後，而最明顯的差別在於原先只有 17 名的第 10 個變數以及在 20 名外的第 14 個變數(排名 26)

在去除歧異值之後名次大大的提升了。(可以對照圖 5.1 的分佈變化)

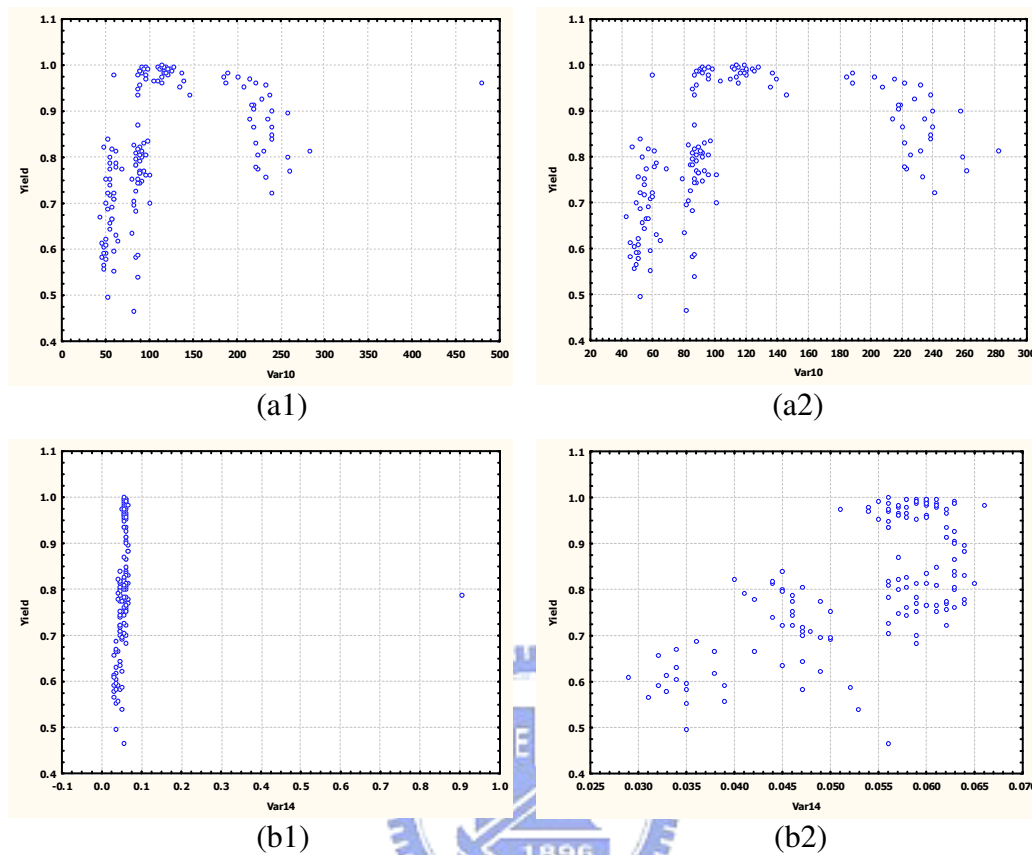


圖 5.1 歧異值對於資料分佈的影響

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
去除前	7	22	21	4	13	9	12	5	2	20	11	3	31	26	6	30	10	18	23	17
去除後	7	22	21	4	10	13	9	5	12	14	2	11	3	20	31	26	6	16	30	23

表 5.2 去除歧異點前後對特徵選取排名的影響

C. 歧異值對模型建立的影響

圖 5.2 和表 5.3 是依照表 5.2 的排名討論去除歧異值對於建立模型的影響，其中圖 5.2 橫軸是變數個數，3 表示的是包含良率以及第 7 和 22 個變數、4 表示的是包含良率以及 7、22、21 個變數...依此類推。縱軸是偏差值也就是由 GGMs 建立的模型其共變異矩陣以及原始資料的共變異矩陣之間的差異(建立模型使用

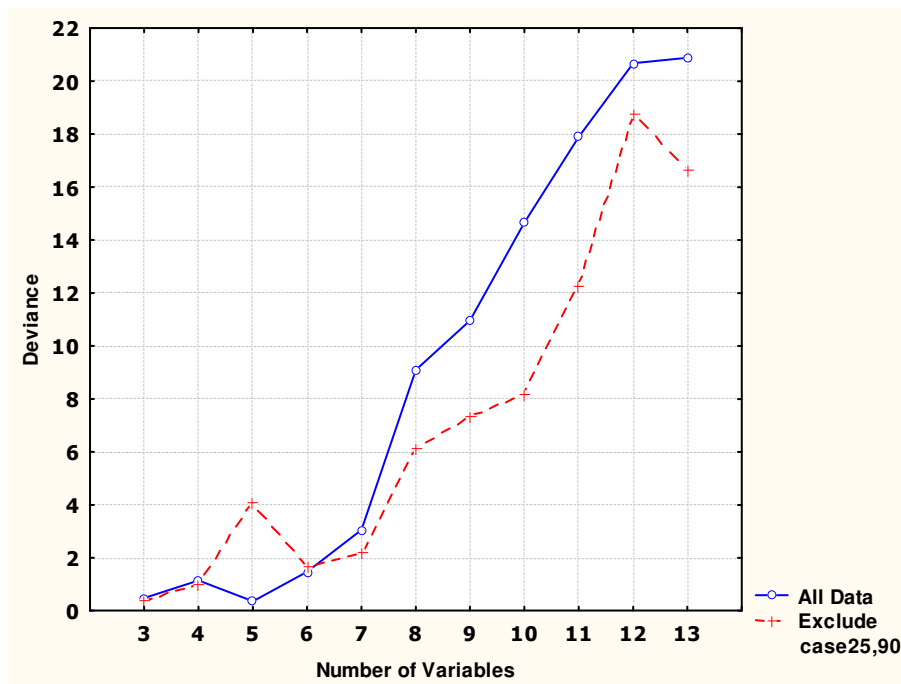


圖 5.2 去除歧異值對於模型偏差值的影響



	去除前偏差值	去除的邊數	去除後偏差值	去除的邊數
3	0.4707655	1	0.3958895	1
4	1.1396931	2	1.0166468	2
5	0.3564144	1	4.1072828	2
6	1.4714728	2	1.6490840	3
7	3.0501623	5	2.1607883	5
8	9.0991226	7	6.0971328	9
9	10.9555117	10	7.3053182	12
10	14.6618081	14	8.1663816	14
11	17.895209	21	12.2297745	19
12	20.676783	23	18.7506294	23
13	20.873949	28	16.47515	28

表 5.3 比較去除歧異值前後偏差值的差異

的是 pairwise deletion，檢定量為 0.05)。由圖 5.2 可以看出基本上去除歧異值之後的偏差值都會比去除前小，表示適當地處理歧異值的確有助於模型的建立。值得注意的是在建立 5 個變數以及 12 個變數的模型時去除歧異值後的偏差值會不正常的上升剛好發生在第 10 以及第 14 個變數的附近，也就是因為特徵排名的改變所導致的。

5.2.2 缺少資料項的問題

缺少資料項時一般的處理方式在 2.1.2 節已經介紹，其中因為 pair-wise 和 case-wise 兩種方法對於特徵選取產生的結果大致相同(如表 5.4 所列的 25 個變數的排名，計算時將資料分為 5 個區段)，因此分別以兩種方法建立的模型相差也不會很大，在此也不詳加比較。在往後的實驗中會偏向使用 pair-wise 的方式處理缺少資料項的問題，因為依這個方法所考慮的資料量通常會較 case-wise 的多。



5.3 GGMs 效能檢測

5.3.1 變數個數對計算時間的影響

剛開始要測試的是 GGMs 以程式實現出來後，程式究竟能承受多大的運算量，數值計算的極限在哪？

先將所得的數據經 2.1.3 節所述的特徵選取(使用 pairwise deletion，並去除歧異值--第 25、90 項的樣本數據，計算時將資料分佈分為 5 個區段)得到如表 5.4 左邊的變數排名，再分別使用包括良率在內的 3 個變數、4 個變數、5 個變數... 依序增加變數個數依照 3.3 節的流程去建立 GGMs(如圖 5.3 橫軸所示，每次以多少個變數建立模型)，以觀察變數個數對時間的影響，同時針對不同的檢定量 0.05 和 0.01 各去作實驗以供比較。

	使用 pair-wise 的排名	F-value	使用 case-wise 的排名	F-value
1	Var7	51.57072	Var7	51.28700
2	Var22	45.67306	Var22	45.20295
3	Var21	44.14635	Var4	43.82512
4	Var4	43.82512	Var21	43.60171
5	Var10	42.42450	Var10	42.42450
6	Var13	40.12993	Var13	40.06991
7	Var9	39.47857	Var5	38.81629
8	Var5	38.81629	Var9	38.20432
9	Var12	38.69620	Var12	37.99092
10	Var14	37.52217	Var2	37.28449
11	Var2	37.28449	Var14	36.90579
12	Var11	37.27260	Var11	36.58411
13	Var3	36.28633	Var3	36.28633
14	Var20	36.11199	Var20	36.09016
15	Var31	35.91279	Var31	35.24033
16	Var26	34.91119	Var6	34.37583
17	Var6	34.67190	Var26	34.25128
18	Var16	29.99843	Var16	29.99843
19	Var30	29.41678	Var30	28.86566
20	Var23	28.91856	Var23	28.56968
21	Var18	28.54709	Var18	26.24499
22	Var17	24.76398	Var15	25.06418
23	Var15	22.76162	Var17	22.01629
24	Var27	19.15109	Var27	18.73854
25	Var28	16.79223	Var28	17.40010

表 5.4 特徵選取的排名(分別透過 pair-wise 和 case-wise deletion)

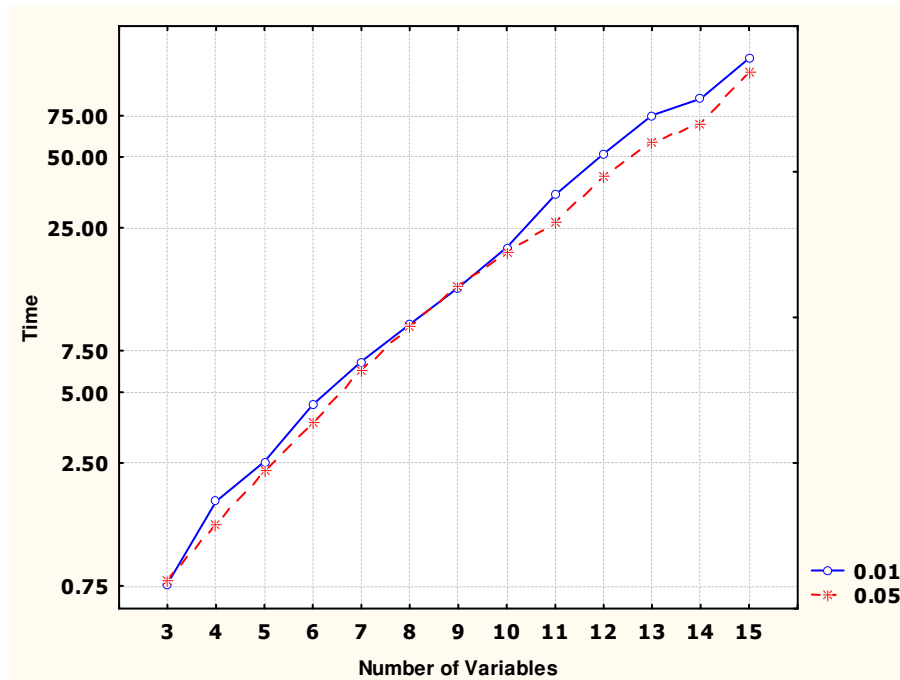


圖 5.3 變數個數對計算時間的關係圖

圖 5.3 顯示出計算的時間正如 1.2 節所述，隨著變數個數的增加呈指數的成長。檢測值為 0.01 的曲線只用了 15 個變數是因為 16 個變數其運算超過了程式所能負荷的量，造成記憶體不足程式異常終止。檢測值 0.05 的曲線雖然也只顯示到 15 個變數，但是實際上在 17 個變數的時候才会有異常終止的現象，不過因為使用 16 個變數去建立模型時，在計算 p 值會產生因數值計算造成的錯誤，因此並不列入參考。

5.3.2 不同檢測值對模型的影響

接下來以不同的檢定量去建立模型，並且像前一節一樣逐一增加變數個數去建立模型去觀察不同檢定量對於去除的邊以及偏差值的影響，圖 5.4(a)是不同檢測值和去除邊數的比較，(b)(c)分別是表示不同檢測值之下變數個數和去除邊數的關係圖。表 5.5、5.6 分別是以不同檢定量去建立模型而得的數據。

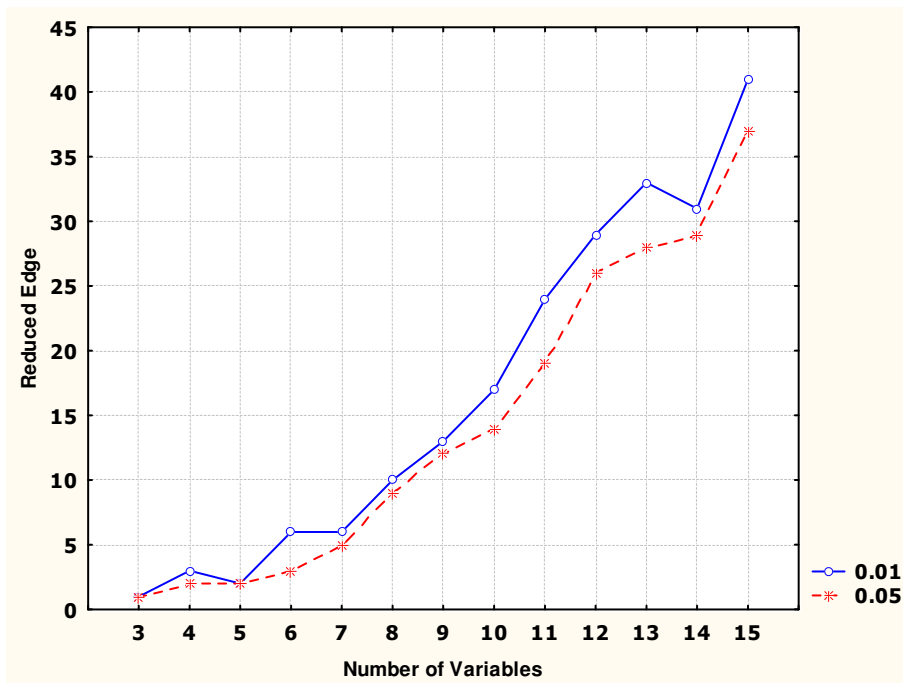


圖 5.4 (a)不同檢測值與去除邊數的比較

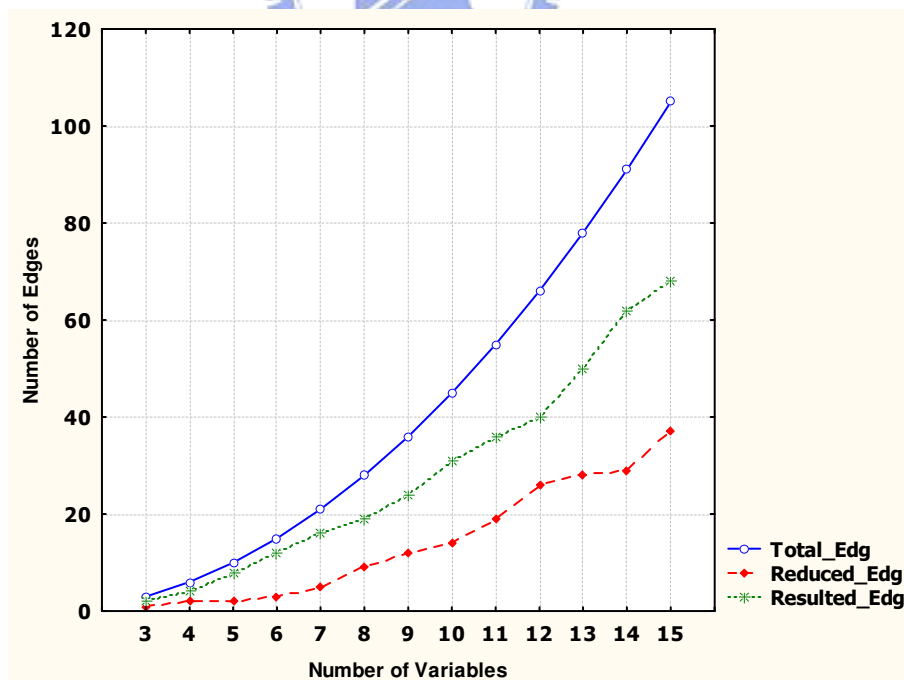


圖 5.4 (b) $\alpha = 0.05$ 之下邊數的變化

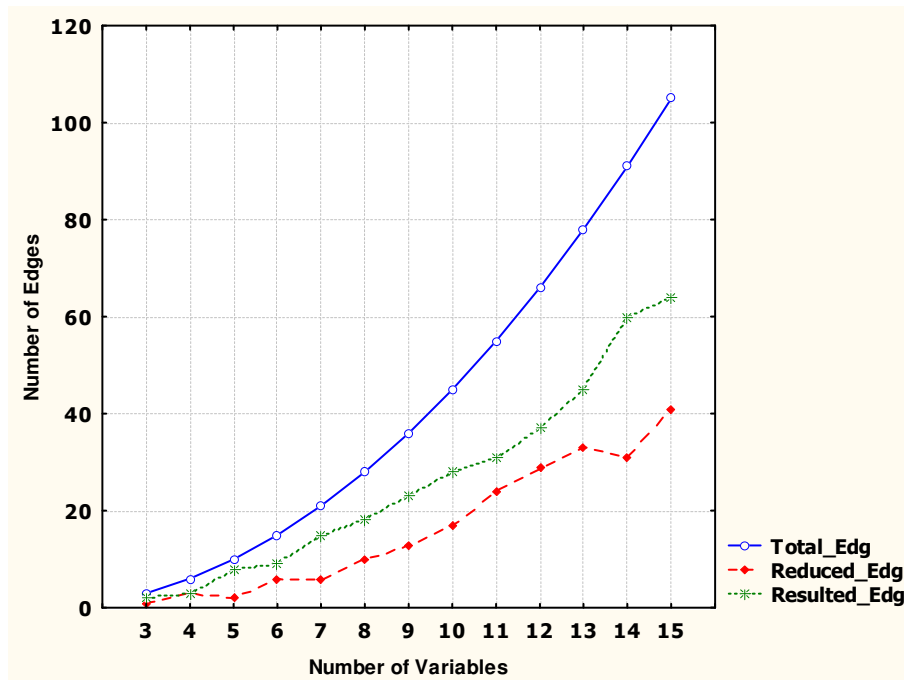


圖 5.4 (c) $\alpha = 0.01$ 之下邊數的變化

值得注意的是圖 5.4(a)中以 0.01 為檢測值的情形下使用 5 個變數和 14 個變數去建立模型，去除邊數都會有下降的情形，伴隨的是表 5.5 裡偏差值的減少。影響這種情形發生的原因可能是如同 5.2.1 中 B 和 C 小節所述，也就是去除歧異值處產生的影響，另外一個可能的原因則是因為變數之間的關聯性增強。以 14 個變數為例，增加了第 14 個變數造成了之前的 13 個變數關係更加的緊密，因此保留的邊數變多了，能夠去除的邊也相對的減少，所以會有這樣如圖 5.4(a)的情形產生。

綜合圖 5.3、圖 5.4(a)和表 5.5、表 5.6 我們可以知道檢定量越小也就是每次可能保留的邊越少，則計算時間會越長、去除的邊會越多，但是相對的建立出來的模型會較不準確也就是偏差值較大，因此這是我們在往後建立模型的時候需要取捨的地方，我們需要的是較簡化的模型(損失的資訊較多)或是較準確的模型(維度較高，可能產生的圖群也越大)？

	Time	Total_Edg	Reduced_Edg	Resulted_Edg	deviance	p-value
3	0.797	3	1	2	0.395889577	0.470775
4	1.375	6	2	4	1.01664688	0.398495
5	2.343	10	2	8	4.10728282	0.871733
6	3.75	15	3	12	1.64908402	0.351687
7	6.203	21	5	16	2.16078839	0.17352
8	9.485	28	9	19	6.09713282	0.26984
9	14.015	36	12	24	7.30531826	0.163204
10	19.484	45	14	31	8.16638167	0.119521
11	26.156	55	19	36	12.2297746	0.124435
12	41.36	66	26	40	18.7506294	0.153215
13	57.594	78	28	50	16.6475153	0.044769
14	69.453	91	29	62	15.2883955	0.017365
15	114.562	105	37	68	27.3826759	0.124364

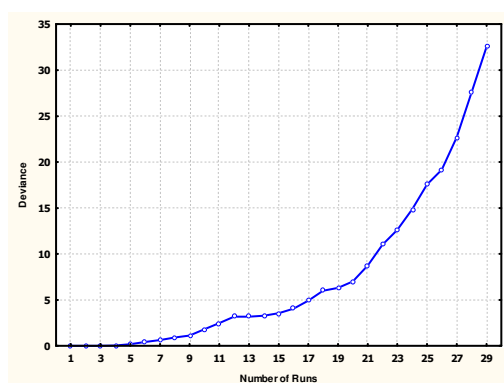
表 5.5 以 $\alpha=0.05$ 去建立模型

	Time	Total_Edg	Reduced_Edg	Resulted_Edg	deviance	p-value
3	0.766	3	1	2	0.395889577	0.470775
4	1.719	6	3	3	6.35386666	0.904392
5	2.531	10	2	8	4.10728282	0.871733
6	4.422	15	6	9	15.9814569	0.986146
7	6.75	21	6	15	7.2324832	0.70018
8	9.703	28	10	18	11.1688276	0.655482
9	13.937	36	13	23	12.3770131	0.50301
10	20.609	45	17	28	19.9830225	0.724901
11	34.61	55	24	31	26.7965633	0.685975
12	51.703	66	29	37	32.5504492	0.70376
13	75.031	78	33	45	33.9510482	0.578455
14	88.453	91	31	60	22.7241962	0.141034
15	131.313	105	41	64	45.8333084	0.721438

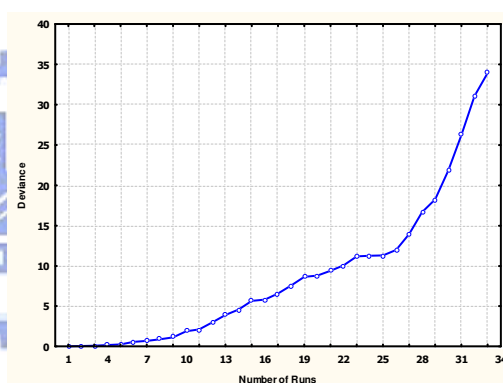
表 5.6 以 $\alpha=0.01$ 去建立模型

5.3.3 GGMs 過程中模型的變化

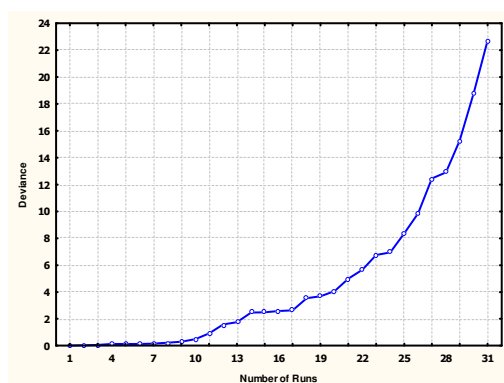
接著要討論的是每去除一邊對於整體模型的影響，我固定變數個數分別為 12、13、14、15 去觀察偏差值的變化情形。圖 5.5 分別表示的是不同變數個數的情形下每回合去除一個邊之後的模型相對於最初原始模型(也就是沒有去除任何一個邊，每個點彼此互相都有連接)的偏差值變化，從這四個圖我們可以大致看出邊的去對於整體模型的影響：剛開始的偏差值變化並不大，表示這些邊相對於整體模型來說屬於多餘的資訊，其所連接的頂點之間事實上關聯性並不強；慢慢的可以刪除的邊越來越少，每個邊的重要性也越來越大，因此到後來每刪除一個邊會讓模型的偏差值大幅的增加。



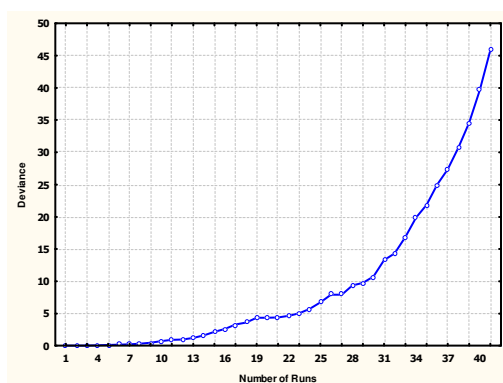
(a)12 個變數—去除 29 個邊



(b)13 個變數—去除 33 個邊



(c)14 個變數—去除 31 個邊



(d)15 個變數—去除 41 個邊

圖 5.5 不同變數個數的情形下，每一回合邊的去
相對整體模型偏差值的變化

5.4 簡化方法

5.4.1 資料分群

首先進行因素分析的程序，原理如 2.2.1 節所述，其中因素的選取依照主要成分的方式去決定，再經過旋轉之後得到表 5.7 的結果。接著再藉著 2.2.2 節的多維縮放比例進行分群，結果如表 5.8。將所有的變數描繪在如圖 5.6(a)-(e)的二維座標上我們更能清楚地看出它們之間的關係。在這邊只選擇兩個維度主要的原因是當維度過大的時候根本很難以肉眼去作分群的工作，二維的空間已經足夠我們將變數作區分了。

Factor Loadings (Varimax raw) (30PredictorsOfYield.sta)									
Extraction: Principal components (Marked loadings are >.700000)									
	Factor	Factor	Factor	Factor	Factor	Factor	Factor	Factor	Factor
Var2	0.85008	0.282309	0.277366	0.140888	Var17	0.87729	0.402282	0.161767	0.066260
Var3	0.86506	0.245299	0.248495	0.086286	Var18	0.88223	0.391163	0.167904	0.066127
Var4	0.89759	0.224115	0.289025	0.151257	Var19	0.08782	0.030614	-0.775028	0.308675
Var5	0.87149	0.291260	0.276854	0.087092	Var20	-0.60875	0.111426	-0.697444	-0.251274
Var6	0.86937	0.295964	0.247640	0.067255	Var21	-0.57485	0.119034	-0.710362	-0.251340
Var7	0.87877	0.235438	0.275866	0.112019	Var22	-0.65582	0.074104	-0.655997	-0.257583
Var8	0.88856	0.381478	0.139554	0.023352	Var23	-0.61343	0.145775	-0.677503	-0.198122
Var9	0.88560	0.316962	0.193123	0.134578	Var24	-0.04622	-0.078569	0.056289	-0.200686
Var10	0.87726	0.269787	0.195458	0.135539	Var25	-0.37334	-0.897367	0.009078	-0.124825
Var11	0.81991	-0.104568	0.034712	0.498829	Var26	-0.22892	-0.947648	0.117268	0.059676
Var12	0.80533	-0.073032	0.002406	0.514374	Var27	-0.25094	-0.951319	0.078193	0.016183
Var13	0.60909	-0.140560	0.519307	0.451448	Var28	-0.25228	-0.947625	0.058765	0.040611
Var14	0.61556	-0.137410	0.487766	0.476053	Var29	0.01213	0.265615	0.246906	0.811155
Var15	0.89159	0.365007	0.163898	0.106675	Var30	-0.35461	0.137383	-0.072806	-0.769478
Var16	0.88197	0.316754	0.174575	0.111553	Var31	0.80876	-0.055299	-0.133404	0.467815

表 5.7 因素分析的結果

	DIM. 1	DIM. 2		DIM. 1	DIM. 2		DIM. 1	DIM. 2
Var2	-0.773309	-0.122859	Var12	-0.545592	0.261813	Var22	1.558754	-0.369783
Var3	-0.761926	-0.141191	Var13	-0.558723	0.369536	Var23	1.511114	-0.382610
Var4	-0.744681	-0.096816	Var14	-0.565870	0.366494	Var24	0.380377	-0.778981
Var5	-0.760075	-0.120230	Var15	-0.739697	-0.163010	Var25	1.218191	0.727777
Var6	-0.752041	-0.151234	Var16	-0.724971	-0.148420	Var26	0.975378	0.802313
Var7	-0.750014	-0.102327	Var17	-0.758816	-0.179809	Var27	1.009190	0.805083
Var8	-0.702805	-0.189946	Var18	-0.755416	-0.176575	Var28	0.998846	0.809261
Var9	-0.737951	-0.144187	Var19	0.618627	-0.266721	Var29	-0.240361	0.658635
Var10	-0.721157	-0.138767	Var20	1.527836	-0.411773	Var30	1.260303	-0.705195
Var11	-0.516987	0.265820	Var21	1.509649	-0.419199	Var31	-0.457876	0.142903

表 5.8 多維縮放比例分析的結果

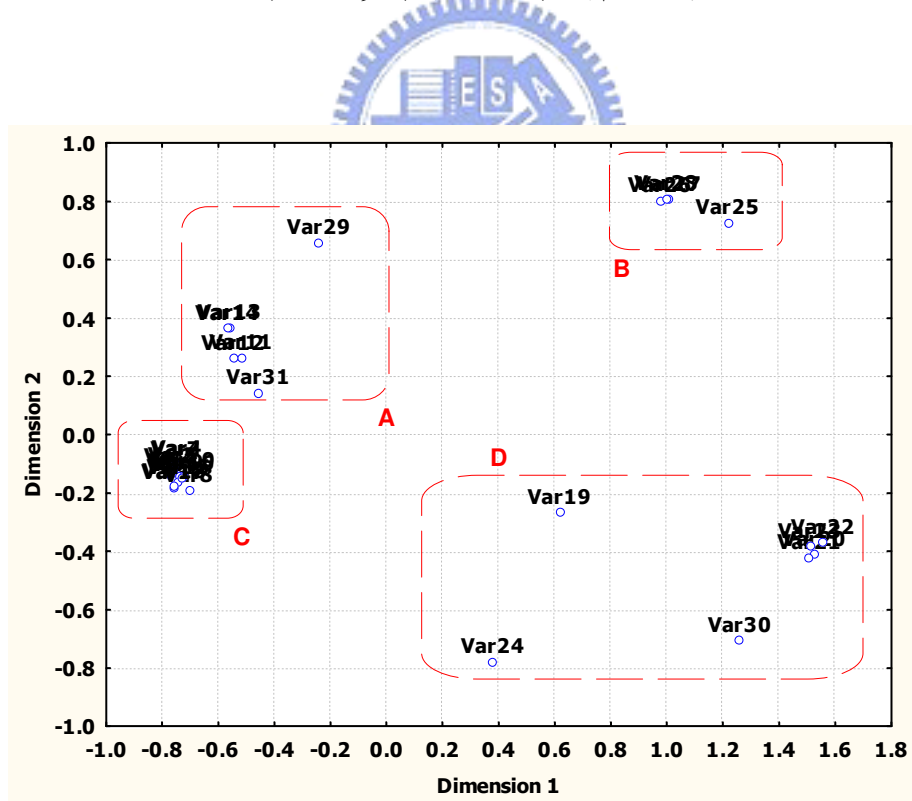


圖 5.6 (a)將表 4.8 繪製於二維座標上的位置關係

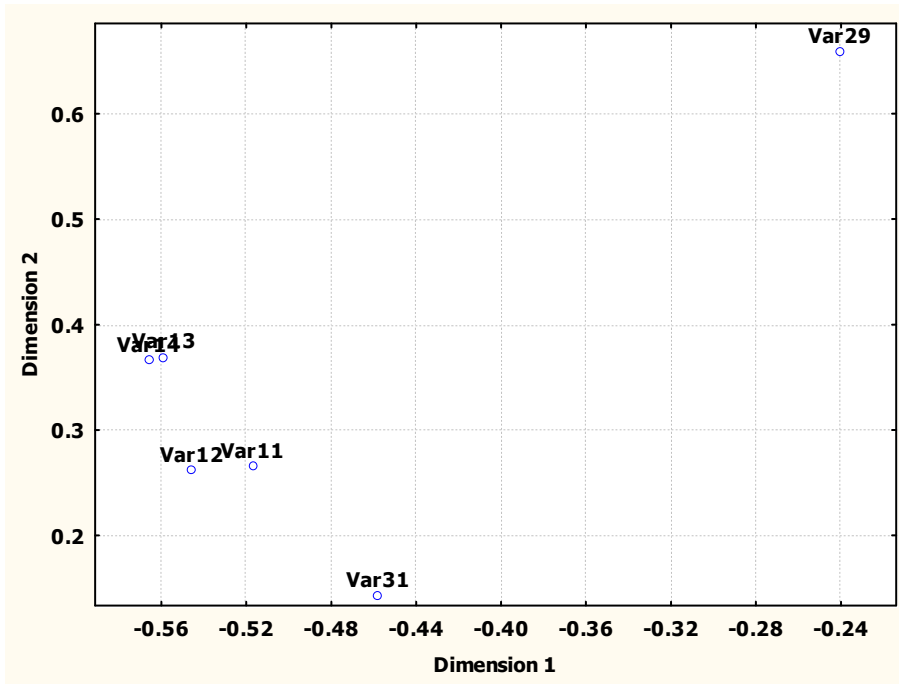


圖 5.6 (b)圖(a)中 A 區塊放大後的位置圖

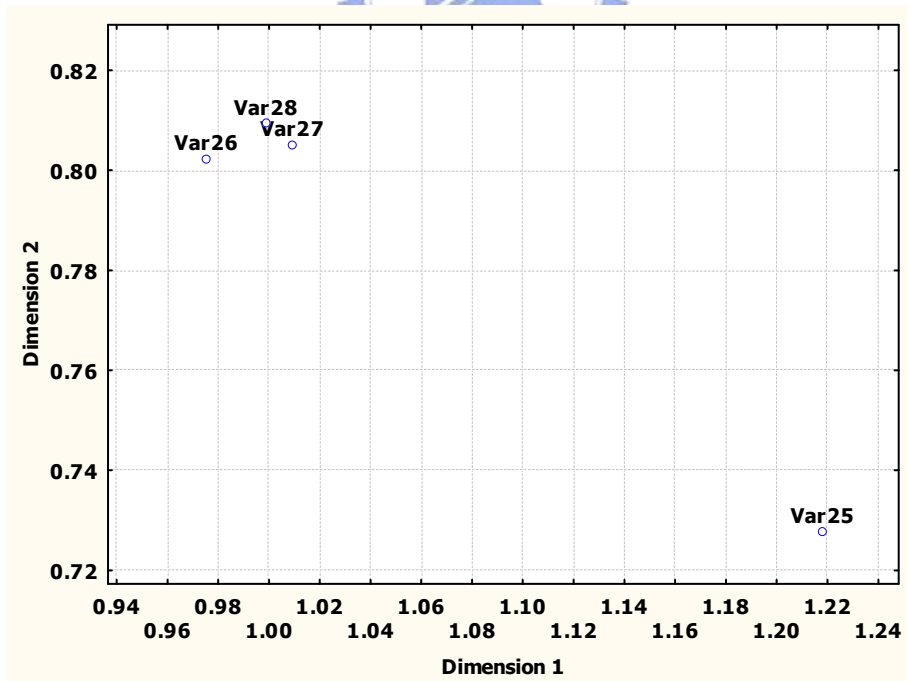


圖 5.6 (c)圖(a)中 B 區塊放大後的位置圖

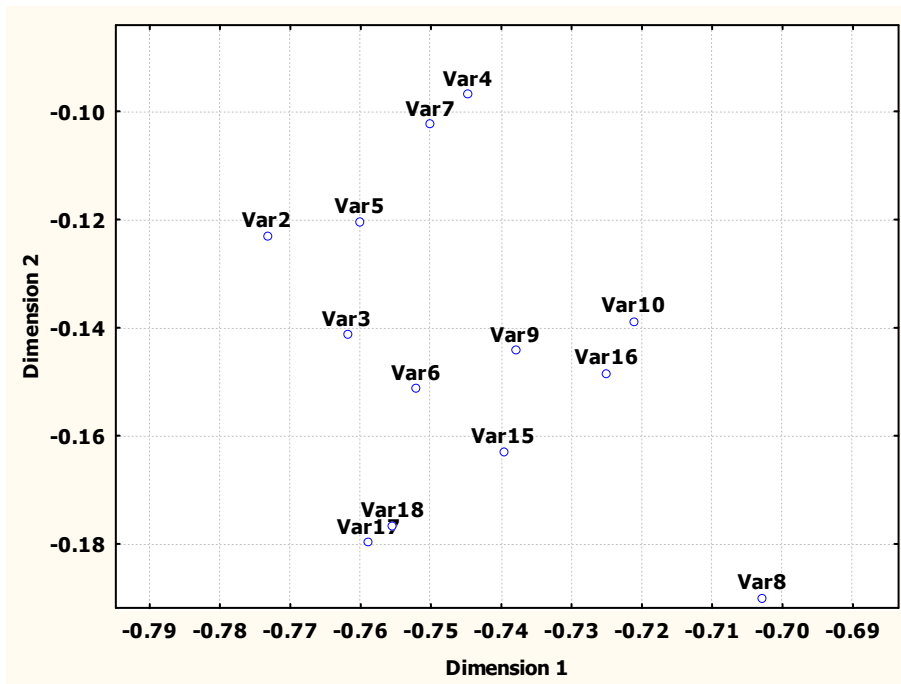


圖 5.6 (d)圖(a)中 C 區塊放大後的位置圖

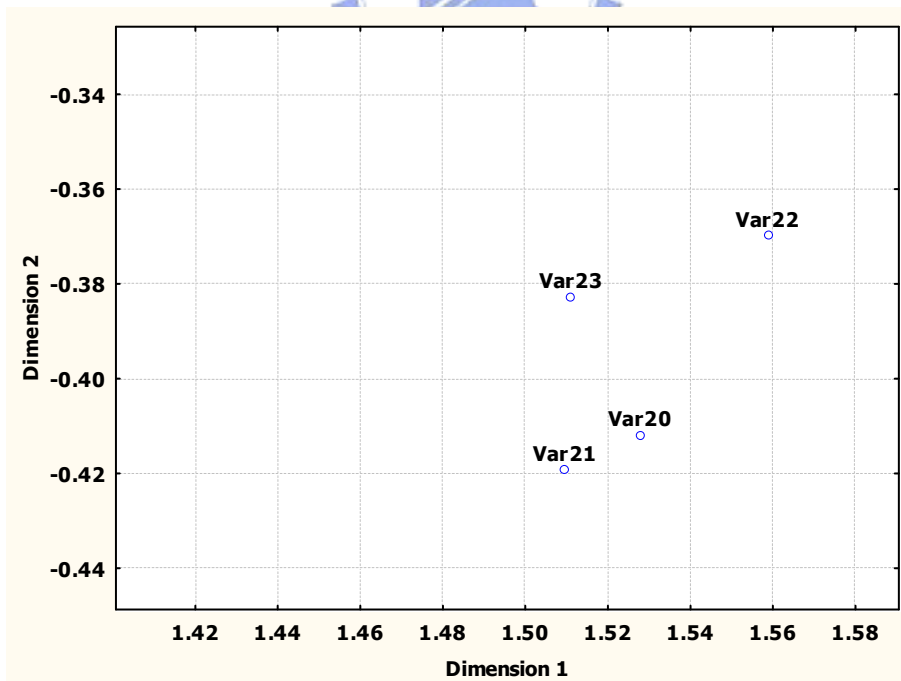
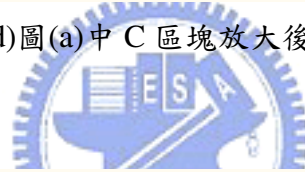


圖 5.6 (e)圖(a)中 D 區塊放大後的位置圖

(因為 D 區塊變數較分散所以只將其中分佈較密的地方加以放大)


綜合表 5.7、5.8 以及圖 5.6(a)-(e)的結果我們可以把變數分成如下幾類：

- (1) Var2、Var3、Var4、Var5、Var6、Var7、Var8、Var9、Var10、Var15、Var16、Var17、Var18。
- (2) Var11、Var12、Var13、Var14、Var29、Var31。
- (3) Var19、Var20、Var21、Var22、Var23、Var24、Var30。
- (4) Var25、Var26、Var27、Var28。

其中第三群大致分佈在圖 5.6(a)的 D 區塊中，不過因為 Var19、Var24、Var30 相較於同區塊中的其他變數來說相當分散，因此若要再細分也可以將這些變數再分別出來。

5.4.2 簡化 GGMs 的建立

A. 分群建立模型



再來就是依照特徵選取的排名以及上面的分群結果進行 GGMs 的建立，結果列於表 5.9，因為未分群之前的 GGMs 因為運算量的關係僅能計算和建立 15 個變數的 GGMs，因此表 5.9 中只列出分群後計算到 15 個變數的結果作為比較。表 5.9 最左邊的變數個數是依照特徵選取的排名慢慢去增加變數，舉例來說，排名前三名的變數是 Var7、Var22、Var21，計算時 3 個變數就是將良率資料和 Var7、Var22 拿去建立模型，4 個變數則是良率和 Var7、Var22、Var21 四個變數進行 GGMs 的建立，依此類推。

分群去建立 GGMs 的過程是先將每一群各自建立 GGMs，之後各群之間兩兩建立連接的關係(兩步驟都使用 3.3 節的流程)，所以最後完整模型的共變異矩陣就不能光是從建立 GGMs 過程中運算使用的共變異矩陣獲得，我們必須另外去計算完整模型的共變異矩陣。計算的根據還是依照 3.2.3 節所述的最大近似估測，建立 GGMs 之後我們可以得到整個模型所有變數之間的連接情形，可以知道哪些邊被去除，因此完整模型共變異矩陣的計算就是從最原始的共變異矩陣一

步一步將邊去除，每去除一個邊就作最大近似的估測，如此求得所需的矩陣。

變數 個數	未分群				分群後			
	去除邊數	偏差值	計算時間	p	去除邊數	偏差值	計算時間	p
3	1	0.395889577	0.797	0.470775	0	0	1	
4	2	1.01664688	1.375	0.398495	1	0.865668937	1.563	0.647843
5	2	4.10728282	2.343	0.871733	3	16.040137	2.203	0.998887
6	3	1.64908402	3.75	0.351687	5	16.1024	3.5	0.993442
7	5	2.16078839	6.203	0.17352	6	15.8848	4.859	0.985614
8	9	6.09713282	9.485	0.26984	11	22.0362	6.797	0.975906
9	12	7.30531826	14.015	0.163204	15	20.9047	8.75	0.860098
10	14	8.16638167	19.484	0.119521	18	30.2693	11.313	0.965091
11	19	12.2297746	26.156	0.124435	21	35.1878	14.36	0.973076
12	26	18.7506294	41.36	0.153215	28	40.5336	22.219	0.940777
13	28	16.6475153	57.594	0.044769	31	34.0867	25.969	0.678613
14	29	15.2883955	69.453	0.017365	38	43.17	33.391	0.740159
15	37	27.3826759	114.562	0.124364	46		38.688	

表 5.9 分群建立 GGMs 的結果，與未分群的數據作比較

在這裡我們藉由最大近似估測來計算分群建立 GGMs 後的共變異矩陣會遇到 5.1 節所述的計算問題，因此表 5.9 中分群後的偏差值只計算到 14 個變數，在 15 個變數的時候因為誤差嚴重累積造成計算的不準確或是因為計算量過大而無法得到所需的估測矩陣，因此並不列入考慮。不過我們藉由表 5.9 可以看到分群的確達到我們的一個要求：希望縮短整個運算的時間，也就是降低建立 GGMs 的運算量，但是分群的結果卻會造成相當大的偏差(在表 5.9 我們要求 p 值要小於 0.95)。

B. 簡化方法的改進

依照 4.2 節所討論和提出的方法，改進之後的結果如表 5.10 所示，偏差值的確有明顯的改進，所有的 p 值都在 0.95 以下(因為這個實驗的檢測值訂為 0.05，也就是希望所有去除的邊偏差值計算所得的 p 值都小於 95%)合於我們所要求的範圍。

最後對於改進後的分群建立模型方法進行時間的比較，如圖 5.7 所示。這樣的方法對於我們所使用的製程資料能夠處理到 20 個變數，以分群的方法而言隨著變數量的增加，計算量和計算時間的確會增加，不過另一個影響的因素是分群的結果，若每一群的各能夠平均分配，不會有某一群特別龐大，這樣相對來說應該能處理更多的變數。不過對於我們使用的資料來說有些變數的分佈的確相當接近(見 p.51 的第(1)群變數)而且特徵選取的排名又比較前面，因此在處理到 20 個變數的時候已經有一群達到 9 個變數，這樣的分群結果使得整個計算受到影響。

此外圖 5.8 顯示的是我們建立出來的圖形化模型，使用的軟體是 aiSee。較多變數所繪製出來的圖形其實圖形以及連接的邊過於複雜，肉眼很不容易從其中觀察資訊，因此在這邊以 7 個變數為例繪製圖形化模型，當變數量過多的時候則傾向去研究記錄邊連接關係的 ECM 矩陣。

變數 個數	未分群				改進的分群方法			
	R_Edge	Deviance	Time	p	R_Edge	Deviance	Time	p
3	1	0.395889577	0.797	0.470775	0	0	0.453	
4	2	1.01664688	1.375	0.398495	1	0.865668	0.594	1.563
5	2	4.10728282	2.343	0.871733	2	5.2983	0.609	0.929289
6	3	1.64908402	3.75	0.351687	4	5.3605	1.281	0.747734
7	5	2.16078839	6.203	0.17352	7	6.7002	2.297	0.539252
8	9	6.09713282	9.485	0.26984	9	7.18657	4.172	0.382298
9	12	7.30531826	14.015	0.163204	13	8.9954	5.609	0.226709
10	14	8.16638167	19.484	0.119521	16	20.7576	7.313	0.811983
11	19	12.2297746	26.156	0.124435	21	21.714	10.438	0.583852
12	26	18.7506294	41.36	0.153215	25	25.537	16.454	0.567375
13	28	16.6475153	57.594	0.044769	30	26.1437	21.093	0.332223
14	29	15.2883955	69.453	0.017365	37	34.097	29.437	0.394116
15	37	27.3826759	114.562	0.124364	43		34.875	

表 5.10 先整體計算最大近似估測再進行分群，同樣和未分群的結果作比較。

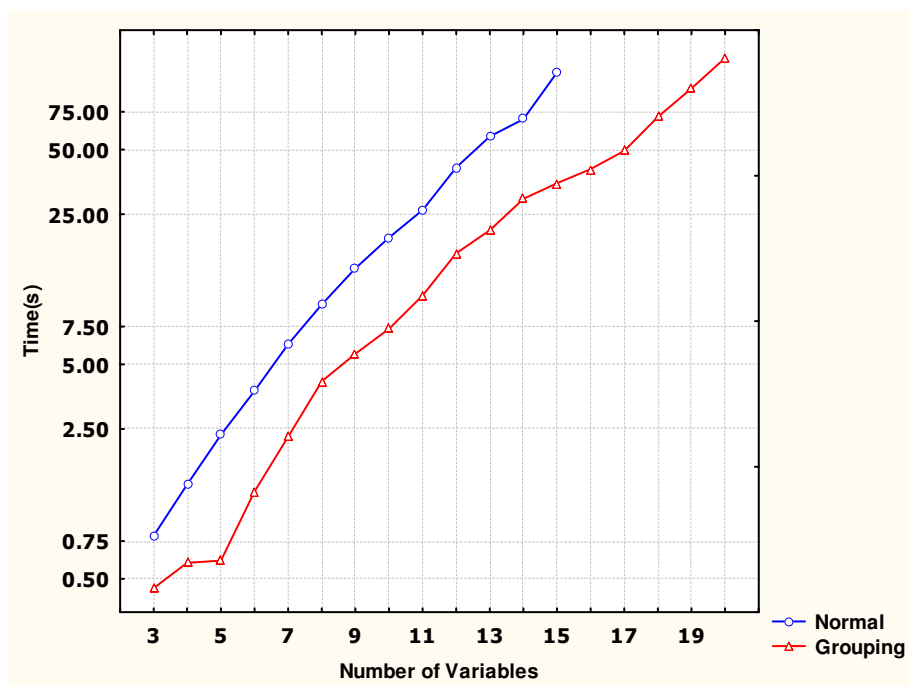
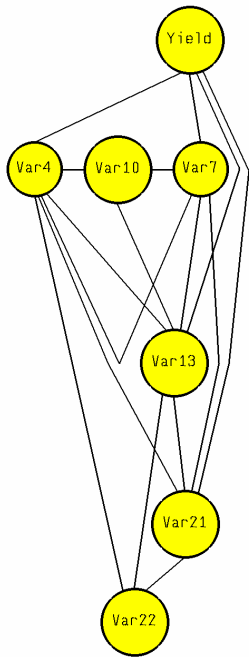
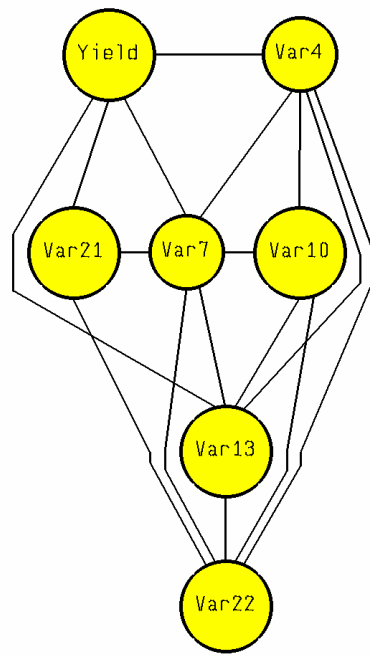


圖 5.7 改進後再分群和未分群計算時間的比較



(a)未分類建立模型



(b)分類建立模型

圖 5.8 7 個變數的圖形化高斯模型



第六章 結論

本論文主要是以晶圓製程資料為對象，依據如 1.1 節所述的資料特性選擇使用圖形化高斯模型的方法建立模型，提供研究者分析資料中解釋變數(Var2-Var31)對於被解釋變數(Var1)的影響以及解釋變數之間彼此互相的關聯性。雖然本論文僅使用晶圓製程的資訊為研究方向，但是同樣的方法依然可以應用在和製程類似自動化過程所測量的大量數值資料中。

在實驗中從前置處理開始討論歧異點對資料分佈以及特徵選取的影響並提出找尋的方法，接著對一般圖形化高斯模型的方法針對變數個數的檢測，找出程式所能負荷的極限，並且探討不同檢定量對於模型(包括計算時間、去除邊數、和偏差值)的影響以及運算過程中模型偏差值的變化情形。其次結合因素分析和多維縮放比例的方法進行變數的分群，藉由分群建立模型的方式簡化一般的方法，並對最大近似估測所遭遇到的難題進行討論，最後針對偏差值提出改進的方法，讓簡化方法所建立的模型於分析更多變數量的同時還能使模型偏差值保持在使用者要求的範圍之內。

從實驗的結果我們可以得到以下幾點：

- (1) 歧異值的存在對於變數分佈會產生相當大的影響，連帶會影響特徵選擇的排名順序，進而對模型造成相當大的改變，因此歧異值的檢驗和去除是前置處理中相當重要的步驟。
- (2) 檢定量的大小對於模型建立會有不少的影響，我們需要較簡化的模型(損失的資訊較多)或是較準確的模型(維度較高，可能產生的圖群也越大)都可以藉由調整檢定量的方式來達成。
- (3) 運用分群建立模型的簡化方法的確能相當程度的增加分析變數的個數，對於大型的模型的研究和探討會有一定程度的幫助，對於分群的建議是儘量將變數平均分佈在各群，讓每一群大小相類似而不至於有某一

群過於龐大徒然增加運算量。

- (4) 分群建立模型的方法即是忽略了變數部分的關聯性，如此可能會導致結果偏差值過大，因此模型建立之前可從大方向找尋所有對整體模型重要的邊加以保留來控制往後偏差值的變化以達到增加變數個數同時保持偏差值在我們的要求範圍之內。

最後談到本論文的改進能從以下幾點著手：

- (a) 圖形化高斯模型使用的前提會假設資料呈常態分佈，但是我們使用的資料卻仍有一些不符合假設的分佈(見圖 1.1)，這樣的問題是因為有時候樣本資料會是某幾群分佈的組合，因此可以考慮先將樣本分類，再針對相同類資料進行模型的建立以降低資料分佈不均造成的影響。
- (b) 本文所討論的變數關係是針對數據而言，因此分群的動作是依照資料分佈的相似程度作區分，將資料型態類似的變數分作一群，這樣的作法可能會忽略了變數原先所代表的物理量，因此在分群的時候也可以根據物理意義增加一些限制和考量，使分群的結果更合理。
- (c) 在計算偏差值和 p 值我們使用的是 χ^2 測試，但是實際上可以依照不同的情形(例如樣本數的多寡、特殊的分佈狀態等)嘗試使用不同的測試方法，以提升估測的準確率。
- (d) 模型的建立使用到大量的矩陣運算且主要影響計算量和運算時間的步驟在於最大近似估測，因此找尋更有效率的矩陣函式庫和發展更好的估測演算法將會是改進效能相當大的關鍵。

附錄 A F-測試

Eriksen(1996)對於樣本數量較少的情況提出完整的討論，他建議使用 F 測試的方法。要討論 F 測試，我們假設一個連續的變數 X 且一個模型中去除 X 之後的變數集合為 $a = V \setminus \{X\}$ ，其中 V 為所有的變數集合。首先考慮一個模型 \mathcal{M} 對於給定 a 後 X 條件分佈變異量是呈現固定分佈的，因此其條件密度函式會是

$$(2\pi\sigma_{x|a})^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(x - \mu_{x|a})^2 / \sigma_{x|a}\right\}, \quad (\text{A.1})$$

其中 $\mu_{x|a}$ 和 $\sigma_{x|a}$ 分別是條件平均值和變異量。因此條件對數近似值會是

$$-\left(\frac{N}{2}\right)\ln(2\pi) - \left(\frac{N}{2}\right)\ln(\sigma_{x|a}) - \frac{RSS}{2\sigma_{x|a}},$$

其中 RSS 是(A.1)是中計算的餘數平方和，我們也可以從推導得 $\hat{\sigma}_{x|a} = RSS / N$ ，

因此最大的對數近似值可以寫為

$$-\left(\frac{N}{2}\right)\ln(2\pi) - \left(\frac{N}{2}\right)\ln(RSS / N) - \frac{N}{2} \quad (\text{A.2})$$

現在假設兩個模型的測試 $\mathcal{M}_0 \subseteq \mathcal{M}_1$ ，且兩個模型都符合上述的的假設，則 \mathcal{M}_0 和 \mathcal{M}_1 之間的偏差值可以寫為

$$\begin{aligned} d &= 2(\hat{l}_V^0 - \hat{l}_V^1) \\ &= 2(\hat{l}_{x|a}^0 - \hat{l}_{x|a}^1) \\ &= N \ln(RSS_0 / RSS_1) \end{aligned} \quad (\text{A.3})$$

假如 r_0 代表的是介於 \mathcal{M}_0 和 \mathcal{M}_1 之間自由變數(同時也會介於 $\mathcal{M}_{x|a}^0$ 和 $\mathcal{M}_{x|a}^1$ 之

間)，且 r_1 代表 $\mathcal{M}_{x|a}^1$ 中自由的變數，則 $\mathcal{M}_{x|a}^0$ 相對於 $\mathcal{M}_{x|a}^1$ 的 F 測試可寫成

$$F = \frac{(RSS_0 - RSS_1) / r_0}{RSS_1 / (N - r_1)}$$

使用(A.3)式我們可以得到

$$F = (e^{d/N} - 1)((N - r_1) / r_0) \quad (\text{A.4})$$

簡化之後一般性的結果如下：假如 \mathcal{M}_0 和 \mathcal{M}_1 都是可分解模型，且兩個模型只相差一個邊，那麼在 $\mathcal{M}_0 \subseteq \mathcal{M}_1$ 情形下的 F 測試可以使用

$$F = (N - k)(e^{d/N} - 1) \quad (\text{A.1})$$

其中 k 是 \mathcal{M}_1 中包含唯一相差邊的圖群之頂點個數， N 是樣本的總數。而 \mathcal{M}_0 就會有自由度為 1 和 $N - k$ 的 F 分佈。

Whittaker(1990)曾經提出偏差值 d 和要移除的邊(假如從完整的模型中要移除邊 $[VW]$ ，而其部分相關係數是 $\hat{\rho}^{VW.b}$ ，其中 $b = \Gamma / \{V, W\}$)的關係將是

$$d = -N \ln \{1 - (\hat{\rho}^{VW.b})^2\} \quad (\text{A.2})$$

而由 F 值去求取 p 值可以用以下的式子來求得的，

$$f(x) = \{\Gamma[(v+w)/2]\} / \{\Gamma(v/2) \times \Gamma(w/2)\} \times (v/w)^{v/2} \times x^{(v/2)-1} \times \{1 + [(v/w) \times x]\}^{-(v+w)/2} \quad (\text{A.3})$$

其中 v 和 w 是自由度， Γ 是 Gamma 函式。



參考文獻

- [1] M. S. Grewal and A. P. Andrews., Kalman Filtering: Theory and Practice, Prentice-Hall, 1993.
- [2] B. Frey., Graphical Models for Machine Learning and Digital Communication, MIT Press, 1998.
- [3] David Edwards, Introduction to Graphical Modelling, Second Edition, Springer-Verlag, 2000.
- [4] J. Whittaker, Graphical Models in Applied Multivariate Statistics, John Wiley, New York, 1990.
- [5] T. W. Anderson, An Introduction to Multivariate Statistical Analysis, Second Edition, John Wiley, New York, 1984.
- [6] Henry Stark, John W. Woods, Probability and Random Processes with Applications to Signal Processing, Third Edition, Prentice-Hall, 2002.
- [7] Barbara G. Tabachnick, Linda S. Fidell, Using Multivariate Statistics, Third Edition, Harper Collins, New York, 1996.
- [8] Castillo E., Gutiérrez J. M., Ali S. Hadi, Expert Systems and Probabilistic Network Models, Springer-Verlag, 1997.
- [9] Schoichiro Nakamura, Applied Numerical Methods in C, Prentice-Hall, 1995.
- [10] Robert G. Cowell, et al., Probabilistic Networks and Expert Systems, Springer-Verlag, 1999.
- [11] Borg I., Goren P., Modern Multidimensional Scaling : Theory and Applications, Springer-Verlag, 1997.
- [12] 曹永誠,「自動化經驗談」, 工業電機&自動控制裝置與設計雜誌, 第十五期, 全國工業電機自動控制雜誌社, 1998年6月

- [13] Jeff A. Bilmes, “Graphical Models and Automatic Speech Recognition” , Mathematical Foundations of Speech and Language Processing, Institute of Mathematical Analysis Volumes in Mathematics Series, Springer-Verlag, 2003.
- [14] Jeff A. Bilmes, et al., “Discriminatively Structured Graphical Models for Speech”, UWEE Technical Report Series, Summer 2001.
- [15] Mark S. “Multidimensional Scaling”, Encyclopedia of Cognitive Science, pp. 1-7, Macmillan Reference Ltd., March 2001.
- [16] Man-Suk Oh, Adrian E. Raftery, “Bayesian Multidimensional Scaling and Choice of Dimension”, Journal of the American Statistical Association, 455, 96, pp. 1031-1044, American Statistical Association, September 2001.
- [17] Isabelle Guyon, André Elisseeff, “An Introduction to Variable and Feature Selection”, Journal of Machine Learning Research, 3, pp. 1157-1182, MIT Press, March 2003.
- [18] Anne Berry, Jean R. S. Blair, Pinar Heggernes, “Maximum Cardinality Search for Computing Minimal Triangulations for Graph”, Algorithmica, 4, 39, pp. 287-298, Springer-Verlag, New York, May 2004.
- [19] Kruskal J. B., Wish M., “Multidimensional Scaling”, Sage University Paper series on Quantitative Applications in the Social Sciences, pp. 7-11. Sage Publications, Newbury Park, CA, 1978.
- [20] Cattell R. B., “The scree test for the number of factors”, Multivariate Behavioral Research, 1, pp. 245-276, 1966.
- [21] J. Makhoul, “Linear prediction: A tutorial review”, Proceedings of the IEEE, 63, 4 , pp. 561-580, April 1975.
- [22] T. P. Speed, H. T. Kiiveri, “Gaussian markov distributions over finite graphs” Annals of Statistics, 14, 1, pp. 138-150, 1986

- [23] Picci G., Pinzoni S., “On Joint Identification and Latent Variable Estimation in Factor Analysis Models”, Proceedings of the IEEE Conference on Decision and Control, 4, pp. 3877-3882 , Lake Buena Vista, FL USA, December 1994.
- [24] Baek K., Draper, B. A., “Factor Analysis for Background Supression”, Proceedings of the IEEE Conference on Pattern Recognition, 2, pp. 643-646, August 2001.

