

圖形化高斯模型應用於 自動化生產資料之關聯性分析

學 生：郭宇豪

指導教授：周志成

國立交通大學電機與控制工程學系碩士班

摘 要

自動化生產過程有上百個步驟，每個步驟都包含相當多的測量項目，所以會得到相當龐大的原始數據。這些數據不但有極大量的變數，同時變數之間也存在高度的關聯性，因此會有多餘資訊的產生。依據這些特性我們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對於被解釋變數的影響以及解釋變數之間互相的關聯性。

以晶圓製造為例，論文中將資料作前置處理後，先就一般的方法作討論，並測試其計算的極限和一些影響的因素。接著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來簡化一般方法，並且針對偏差值提出改進。利用簡化方法所建立的模型在分析更多變數量的同時，更能使模型偏差值保持於使用者要求的範圍之內。

同樣的方法亦可應用於與製程類似自動化過程所測量的大量數值資料中，模型建立之後再結合專家系統或貝氏網路，將可以對結果進行預測與診斷之工作。

Application of Graphical Gaussian Models to Dependency Analysis with Automated Manufacturing Data

Student : Yu-Hao Kuo

Advisor: Dr. Chi-Cheng Jou

Department of Electrical and Control Engineering
National Chiao Tung University

Abstract

There are hundreds of steps in the process of automated manufacture operation. Every step contains lots of measurements. As a result a tremendous amount of data is available. These data have a great deal of variables, which are highly correlated. According redundancy exists. In order to provide analysts the influence of predictors upon dependents, and to explain the correlations of variables, we use Graphical Gaussian Models(GGMs) to establish models based on the characteristic of the gathered data.

Take manufacture of silicon wafers for example, data will be preprocessed first. Then we will discuss the measured limits and the factors of the general GGMs. Through the combination between factor analysis (FA)and multidimensional scaling (MDS), clusters of variables will be proceeded. According to the clusters, he procedure of modeling will be simplified and an improved method will be introduced to analyze more variables while maintaining the requested deviance.

This method also can be applied to the massive data gathered by the similar procedure like automated manufacturing operation. Combining Expert system or Bayesian Network, we can prognosis and diagnosis results after a model is built.

誌 謝

首先要感謝周志成老師給予我良好的研究環境，並且不論是學習、論文甚至處世都提供我許多不同的思考方向，讓我能更多方面去觀察研究問題。同時也感謝林進燈老師和張志永老師在口試時給我的指導。

其次感謝的是家人這幾年給我的支持和鼓勵，讓我度過不少的難關，尤其是父親常常往返台北和新竹，不時照顧我的健康，我非常的感謝他。

此外也要感謝同實驗室同學和研究所認識的好朋友，因為和他們的相處，使我的研究生活不再單調，許多不瞭解的課題也能經過互相討論學習而有所增長。

最後感謝的是從大學以及社團認識認識的一群好朋友和學長，他們陪伴我走過大學和研究所的學習過程，從彼此的互動中讓我學到許許多多課本以外的知識和想法，使我不單活在書本中，而能實際去看到這個世界，感謝他們，也感謝所有曾經相識、幫助或陪伴我的人。



目 錄

中文摘要.....	i
英文摘要.....	ii
誌 謝.....	iii
目 錄.....	iv
表 目 錄.....	vii
圖 目 錄.....	viii
第一章 緒論.....	1
1.1 研究動機.....	1
1.2 研究方法.....	3
1.3 論文結構.....	7
第二章 前置處理和多變量分析.....	8
2.1 前置處理.....	8
2.1.1 前置處理簡介.....	8
2.1.2 歧異值的處理.....	8
2.1.3 缺少資料項的處理.....	10
2.1.4 特徵選取.....	10
A. 使用特徵選取的原因.....	10
B. 特徵選取的方法.....	11
2.2 多變量分析.....	13
2.2.1 多變量分析簡介.....	13
2.2.2 因素分析.....	13
2.2.3 多維縮放比例.....	16
第三章 圖形化模型.....	19

3.1	圖形化模型簡介	19
3.2	圖形化高斯模型原理	20
3.2.1	圖學簡介.....	20
3.2.2	圖形化模型的理論根據.....	23
3.2.3	最大近似估測.....	25
3.2.4	最大基數搜尋測試.....	28
3.2.5	χ^2 測試和偏差值.....	30
3.3	圖形化高斯模型建立流程	31
第四章	簡化方法的問題和改進.....	34
4.1	最大近似估計的問題	34
4.2	降低簡化方法的偏差值	34
4.3	簡化方法的完整步驟	35
第五章	實驗與討論.....	38
5.1	實驗資料簡介	38
5.2	前置處理	39
5.2.1	歧異值的問題.....	39
	A. 找尋歧異值.....	39
	B. 歧異值對特徵選取的影響.....	39
	C. 歧異值對模型建立的影響.....	40
5.2.2	缺少資料項的問題.....	42
5.3	GGMs效能檢測	42
5.3.1	變數個數對計算時間的影響.....	42
5.3.2	不同檢測值對模型的影響.....	44
5.3.3	GGMs過程中模型的變化.....	48
5.4	簡化方法	49
5.4.1	資料分群.....	49

5.4.2 簡化GGMs的建立.....	53
A. 分群建立模型.....	53
B. 簡化方法的改進.....	54
第六章 結論.....	58
附錄A F-測試.....	60
參考文獻.....	62



表 目 錄

表 2.1 圖2.1中F值與相關係數r的變化.....	11
表 4.1 7個變數的ECM矩陣.....	50
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.....	39
表 5.2 去除歧異點前後對特徵選取排名的影響.....	40
表 5.3 比較去除歧異值前後偏差值的差異.....	41
表 5.4 特徵選取的排名.....	43
表 5.5 以 $\alpha = 0.05$ 去建立模型.....	47
表 5.6 以 $\alpha = 0.01$ 去建立模型.....	47
表 5.7 因素分析的結果.....	49
表 5.8 多維縮放比例分析的結果.....	50
表 5.9 分群建立GGMs的結果，與未分群的數據作比較.....	54
表 5.10先整體計算最大近似估測再進行分群，與未分群的結果作比較.....	56



圖 目 錄

圖 1.1	資料相關的例子.....	2
圖 1.2	針對資料特性提出的架構圖.....	4
圖 1.3	變數增加對於時間的影響.....	5
圖 1.4	簡化方法的整體架構.....	6
圖 2.1	歧異點對二維分佈圖配適的影響.....	9
圖 2.2	分段去計算相關F值的例子.....	12
圖 2.3	模型條件獨立的關係.....	14
圖 2.4	旋轉因素結構的例子.....	16
圖 3.1	圖形化模型的種類.....	20
圖 3.2	(a)非完整圖形.....	21
	(b)完整圖形.....	21
圖 3.3	兩個包含無弦4-循環的圖形.....	22
圖 3.4	//ABC,BCD,BDE的圖形.....	23
圖 3.5	$q=4, \omega^3 = \omega^{24} = 0$ 的圖形.....	25
圖 3.6	分解的例子.....	28
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.....	30
圖 3.8	GGMs的建立流程圖.....	33
圖 4.1	簡化方法的完整架構.....	36
圖 4.2	簡化模型的建立步驟.....	36
圖 5.1	歧異值對於資料分佈的影響.....	40
圖 5.2	去除歧異值對於模型偏差值的影響.....	41
圖 5.3	變數個數對計算時間的關係圖.....	43
圖 5.4	(a)不同檢測值與去除邊數的比較.....	45
	(b) $\alpha = 0.05$ 之下邊數的變化.....	45
	(c) $\alpha = 0.01$ 之下邊數的變化.....	46

圖 5.5	不同變數個數的情形下，每一回合邊的去除相對整體模型偏差值的 變化.....	48
圖 5.6	(a)將表4.8繪製於二維座標上的位置關係.....	50
	(b)圖(a)中A區塊放大後的位置圖.....	51
	(c)圖(a)中B區塊放大後的位置圖.....	51
	(d)圖(a)中C區塊放大後的位置圖.....	52
	(e)圖(a)中D區塊放大後的位置圖.....	52
圖 5.7	改進後再分群和未分群計算時間的比較.....	56
圖 5.8	7個變數的圖形化高斯模型.....	57

