

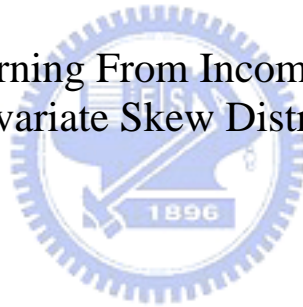
國立交通大學

統計學研究所

博士論文

多變量偏斜分佈對於不完整資料之研究

Topics on Learning From Incomplete Data Using
Multivariate Skew Distributions



研究生：林資荃

指導教授：陳鄰安 博士

林宗儀 博士

中華民國九十八年一月

多變量偏斜分佈對於不完整資料之研究
Topics on Learning From Incomplete Data Using
Multivariate Skew Distributions

研究生：林資荃

Student : Tzy-Chy Lin

指導教授：陳鄰安 博士

Advisor : Dr. Lin-An Chen

林宗儀 博士

Dr. Tsung-I Lin

國立交通大學
統計學研究所
博士論文



A Dissertation
Submitted to Institute of Statistics
College of Science

National Chiao Tung University

in partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy

in

Statistics

January 2009

Hsinchu, Taiwan, Republic of China

中華民國九十八年一月

學生：林資荃

指導教授：陳鄰安 博士
林宗儀 博士

國立交通大學統計學研究所

摘 要

對於經常面對實際資料的研究者來言，處理具複雜遺失形態的多變量資料之統計學習方法是一個非常重要課題。在許多應用的資料分析上，為了數學上的便利，研究者習慣地假設資料為常態分佈。然而，當資料具有非典型或遠離中心的觀察值，常態性假設將會不成立，因而產生無效的推論。本篇論文包含三部份論說：分別在多變量的偏斜常態和偏斜 t 分佈的架構下，提供處理具異質性母體與遺失性資料的方法。

第一部份論說中，我們發展便利的計算的工具來分析具遺失訊息的混合多變量偏斜常態模型，在隨機遺失機制下，我們提供可解析的 EM 演算法，用以處理模型的參數與遺失資料的監督學習。本文所提出的混合分析器，包含了高斯混合模式這個特例，對於處理不完整高維度資料的從事者，提供較寬廣的考慮層面。

第二部份論說提出一些分析高維偏斜 t 模型的方法來處理資料同時具有厚尾，不對稱與觀察值遺失現象。我們提出一個蒙地卡羅 ECM 演算法，用來估計參數與填補遺失資料值。此外，我們也發展一個有效率的資料擴增(data augmentation)演算法，藉由多重填補法說明參數與遺失觀察值的不確定性。

最後的論說中，我們提供以混合多變量偏斜 t 模型為基準的方法，用於不完整實驗資料的穩健性分類。我們也使用數個實際例子與模擬資料來闡述所提出的方法。

關鍵字：EM 演算法，多變量偏斜常態模型，多變量截斷常態分佈，資料擴增，MCECM 演算法，隨機遺失，多變量偏斜 t 模型，多重填補，多變量截斷 t 分佈。

Topics on Learning From Incomplete Data Using Multivariate Skew Distributions

student : Tzy-Chy Lin

Advisors : Dr. Lin-An Chen
Dr. Tsung-I Lin

Institute of Statistics
National Chiao Tung University

ABSTRACT

Statistical learning of multivariate data with complex missing patterns is a very important issue for researchers who act on the real-life problems encountered in their own practice. In many applied data analyses, outcomes are routinely assumed to be normally distributed by practitioners for mathematical convenience. However, such a normality assumption is vulnerable to atypical or outlying observations and subsequently yields invalid inferences. My dissertation consists of three essays on the use of multivariate skew normal and skew t distributions to deal with data in the presence of population heterogeneity and possible missing values.

In the first essay, we develop computationally flexible tools for the analysis of multivariate skew normal mixtures when missing values occur in data. Under missing at random mechanisms, we present an analytically feasible EM algorithm for the supervised learning of parameters as well as missing observations. The proposed mixture analyzer, including the most commonly used gaussian mixtures as a special case, allowing practitioners to handle incomplete multivariate data sets in a wide variety of considerations.

The second essay presents some analytical devices for multivariate skew t models when fat-tailed, asymmetric and missing observations may simultaneously occur in the input data. We present a Monte Carlo version of the ECM algorithm, which is performed to estimate the parameters and retrieves the missing observation with a single imputation. Additionally, an efficient data augmentation scheme is developed to account for the uncertainty based on multiply imputed parameters and missing outcomes.


In the last essay, we offer a multivariate skew t mixture-based approach to robust clustering for experimental data with incomplete observations. Several real data sets as well as simulations are illustrated in my dissertation.

Keywords: EM algorithm, MSN model, Multivariate truncated normal, data augmentation, MCECM algorithm, missing at random, MST model, multiple imputation, truncated t distribution.

誌 謝

時間過的真快，5 年半博士班的研究時光一轉眼就過去，在交大統計所的師資優良、設備齊全的環境下，我的課業與研究皆有相當大的收穫。

首先，我要感謝我的指導老師陳鄰安與林宗儀教授。陳鄰安教授在生活上悉心照顧讓我可以專心撰寫我的論文，林宗儀教授盡心盡力指導我的論文，在研究上給予我許多的幫忙與協助，有時還利用假日與下班時間與我一起討論，幫我克服許多研究上障礙；其間並鼓勵我參加研討會讓我視野更為開廣。



再者，我要感謝所上的每一位老師課業上的指導還有行政助理郭碧芬小姐事務上的熱心幫忙以及中興應數所學妹(姜伶、淑君)論文上幫忙以及女友乙力在英文上面協助與精神上鼓勵。此外我也感謝口試委員們(本所的彭南夫老師、台中科大林淑惠老師以及淡江大學統計系蔡宗儒主任)於百忙撥空參加我的口試，校正並且提供許多寶貴意見。

最後，我要感謝家人，在我求學期間不斷給我鼓勵及支持，讓我在這 5 年半的研究生涯中能夠順利完成。在此，對於所有幫助過我的朋友及老師，獻上最高的敬意。