國立臺灣師範大學

資訊工程研究所碩士論文

指導教授：陳柏琳 博士

基於分類錯誤之線性鑑別式特徵轉換應用於大詞彙連續語音辨識

Classification Error-based Linear Discriminative Feature Transformation for Large Vocabulary Continuous Speech Recognition

研究生： 李鴻欣 撰

中華民國 九十八 年 六 月

摘要

線性鑑別分析(linear discriminant analysis, LDA)的目標在於尋找一個線性轉換，能將原始資料投射到較低維度的特徵空間，同時又能保留類別間的幾何分離度(geometric separability)。然而，LDA 並不能總是保證在分類過程中產生較高的分類正確率。其中一個可能的原因在於 LDA 的目標函式並非直接與分類錯誤率連接，因此它也就未必適合在某特定分類器控制下的分類規則，自動語音辨識(automatic speech recognition, ASR)就是一個很好的例子。在本篇論文中，我們藉著探索每一對容易混淆之音素類別間的經驗分類錯誤率(empirical classification error rate)與馬氏距離(Mahalanobis distance)的關係，擴展了傳統的 LDA，並且將原來的類別間散佈矩陣(between-class scatter)，從每一對類別間的歐式距離(Euclidean distance)估算，修改為它們的成對經驗分類正確率。這個新方法不僅保留了原本 LDA 就具有的輕省可解性，同時無須預設資料是為何種機率分佈。

另一方面，我們更進一步提出一種嶄新的線性鑑別式特徵擷取方法，稱之為普遍化相似度比率鑑別分析(generalized likelihood ratio discriminant analysis, GLRDA)，其旨在利用相似度比率檢驗(likelihood ratio test)的概念尋求一個較低維度的特徵空間。GLRDA 不僅考慮了全體資料的異方差性(heteroscedasticity)，即所有類別之共變異矩陣可被彈性地視為相異；並且在分類上，能藉由最小化類別間最混淆之情況（由虛無假設(null hypothesis)所描述）的發生機率，而求得有助於分類效果提升的較低維度特徵子空間。同時，我們也證明了 LDA 與異方差性線性鑑別分析(heteroscedastic linear discriminant analysis, HLDA)可被視為 GLRDA 的兩種特例。再者，為了增進語音特徵的強健性，GLRDA 更可進一步地與辨識器所提供的經驗混淆資訊結合。

實驗結果顯示，在中文大詞彙連續語音辨識系統中，我們提出的方法都比 LDA 或其它現有的改進方法，如 HLDA 等，有較佳的表現。

Abstract

The goal of linear discriminant analysis (LDA) is to seek a linear transformation that projects an original data set into a lower-dimensional feature subspace while simultaneously retaining geometrical class separability. However, LDA cannot always guarantee better classification accuracy. One of the possible reasons lies in that its criterion is not directly associated with the classification error rate, so that it does not necessarily accommodate itself to the allocation rule governed by a given classifier, such as that employed in automatic speech recognition (ASR). In this thesis, we extend the classical LDA by leveraging the relationship between the empirical phone classification error rate and the Mahalanobis distance for each respective phone class pair. To this end, we modify the original between-class scatter from a measure of the Euclidean distance to the pairwise empirical classification accuracy for each class pair, while preserving the lightweight solvability and taking no distributional assumption, just as what LDA does.

Furthermore, we also present a novel discriminative linear feature transformation, named generalized likelihood ratio discriminant analysis (GLRDA), on the basis of the likelihood ratio test (LRT). It attempts to seek a lower dimensional feature subspace by making the most confusing situation, described by the null hypothesis, as unlikely to happen as possible without the homoscedastic assumption on class distributions. We also show that the classical linear discriminant analysis (LDA) and its well-known extension – heteroscedastic linear discriminant analysis (HLDA) are just two special cases of our proposed method. The empirical class confusion information can be further incorporated into GLRDA for better recognition performance.

Experimental results demonstrate that our approaches yields moderate improvements over LDA and other existing methods, such as HLDA, on the Chinese large vocabulary continuous speech recognition (LVCSR) task.

致謝

感謝　神，仍賜給我一顆柔軟與受教的心，使我如同纔生的嬰孩，對於真理仍滿心切慕，對於未知仍驚訝不已。

感謝陳柏琳教授，我的良師，我的益友。從您的身上，我學習到以更鴻大的視野來看待研究題目，以更超脫、充滿樂趣地經歷研究生活、以更無私、平等的態度來面對周遭的人事物。因著我前面有這樣的標竿，使我期許自己未來能夠成為一個任何人都能託付重任的人。

感謝實驗室的同學們，無論是已畢業的或仍在學的，沒有你們，我無法再重新肯定自己，也無法以更包容、體貼的心來面對他人。

感謝我的母親，我們母子連心，我知道這一切有你暗中地為我禱告祈求。

鴻欣　謹誌

目錄

圖目錄

表目錄

常用專有名詞英文簡稱表

英文簡稱	中文
MS-DEF	基於特徵空間之鑑別式特徵擷取
FS-DFE	基於模型空間之鑑別式特徵擷取
LDA	線性鑑別分析
MFCC(s)	梅爾倒頻譜係數
PCA	主成分分析
HLDA	異方差線性鑑別分析
HDA	異方差鑑別分析
PLDA	基於乘冪平均的線性鑑別分析
WLDA	權重式線性鑑別分析
PWLDA	基於乘冪之權重式線性鑑別分析
aPTAC	近似成對理論正確標準
EER-WLDA	基於經驗錯誤率之權重式線性鑑別分析
DE-WLDA	距離－錯誤耦合之權重式線性鑑別分析
aPEAC	近似成對經驗正確率標準
RWW	關連權重式類別內共變異矩陣
EERW	基於經驗錯誤率之類別內共變異矩陣
GLRDA	普遍化相似度比率鑑別分析
CI-GLRDA	基於混淆資訊之普遍化相似度比率鑑別分析

本論文一致性的數學符號表

符號	中文意義
$\mathbf{x}$	某資料點或某特徵向量
N	資料總數或音框總數
C_i	第 i 類別
C	類別總數
n_i	屬於類別 C_i 之資料總數
p_i	類別 C_i 之事前機率
$\overline{\mathbf{m}}$	全部樣本資料之期望值向量
$\mathbf{m}_i$	屬於類別 C_i 之樣本資料的期望值向量
$\mathbf{S}_i$	屬於類別 C_i 之樣本資料的共變異矩陣
$\mathbf{S}_W$	類別內散佈矩陣
$\mathbf{S}_B$	類別間散佈矩陣
$\mathbf{S}_T$	整體散佈矩陣
$\boldsymbol{\mu}$	全部母體資料之期望值向量
$\boldsymbol{\mu}_i$	類別母體 C_i 之期望值向量
$\boldsymbol{\Sigma}$	全部母體資料之共變異矩陣
$\boldsymbol{\Sigma}_i$	類別母體 C_i 之共變異矩陣
$\boldsymbol{\Theta},\boldsymbol{\theta}$	轉換矩陣和轉換向量
n	原始特徵空間維度
d	投影子空間維度
$\widehat{\mathbf{Y}},\widehat{\mathbf{y}}$	表示矩陣 $\mathbf{Y}$ 或向量 $\mathbf{y}$ 經過白化轉換
$\widetilde{\mathbf{Y}},\widetilde{\mathbf{y}}$	表示矩陣 $\mathbf{Y}$ 或向量 $\mathbf{y}$ 經過降維轉換
Δ_{ij}	類別 C_i 與類別 C_j 的馬氏距離

第 1 章　研究目標與方法論

本論文的研究目標與方法論將會在本章詳細說明之。

1.1　基本目標與方法

語音，作為資訊傳遞媒介的優勢在於迅速與便利。因此，近幾十年來，語音處理技術之最前線－－自動語音辨識(automatic speech recognition, ASR)的研究日益重要，其發展亦突飛猛進。以大詞彙[1]連續語音辨識(large vocabulary continuous speech recognition, LVCSR)系統為例，針對外場記者所講述的中文電視新聞，其字錯誤率(character-error rate, CER)已能降至 20 %以下[1]；而對於小詞彙的英語數字辨識來說，字錯誤率(word-error rate, WER)甚至早已低於 1 % [2]。這樣的成果足能造就許多商業化的應用，例如，AT&T 所研發之電子詢答系統中的 VoiceTone，嘗試以人機對話機制取代傳統純人工服務，大量減少了企業的人力支出，自動語音辨識正是它主要的一環[3]；因著網路搜尋活動漸深入人類生活，Nuance、Google 和 Tellme 等公司也開始著手於包含音訊在內的多模式搜尋(multimodal search)，讓使用者能在不便於文字輸入的環境下，透過語音即可獲得想要的多媒體資訊，而自動語音辨識與文件搜尋技術即為音訊搜尋的兩大必要成分[4]。

儘管在實際應用上，自動語音辨識技術大致獲得了初步的成功，在學術研究上仍存有許多問題。眾所周知地，即使在乾淨無噪音的環境下，目前最先進之自動語音辨識系統的辨識效果依然比人類語音辨識(human speech recognition, HSR)還差。Morgan 等人認為，自動語音辨識的關鍵瓶頸在於現今廣為使用的前端聲

[1] 此處所指的大詞彙系統，意指能夠處理 5000 至 60000 字詞的系統，見[33]。

學特徵(acoustic feature)均十分相似而難以產生有助於後端分類的鑑別性(discriminability)，而特徵中所包含的前後文資訊(contextual information)或時間資訊(temporal information)又不如人耳所能掌握得多且精確[5]。Hermansky 也持有相似的觀點，他認為，欲解決自動語音辨識遠不如人類語音辨識的問題，我們可將人類的聽覺特性列為先備知識(prior knowledge)帶入自動語音辨識系統以彌補純統計式方法的不足[2]，並且可把握兩個原則：嘗試減少聲學特徵中人耳所聽不到的成分，且在能清晰聽見的成分中，找出更可靠、與音訊辨識有關的部分[6]。上面的敘述提供了本論文的研究動機與方法論基礎。在本論文中，如圖 1.1，我們將著重於藉著整合聲學特徵所具有的兩種資訊：前後文資訊與混淆資訊(confusion information)，透過線性轉換[3]並降維(dimensionality reduction)來進行特徵擷取(feature extraction)，將原本符合部分人類聽覺特性的聲學特徵轉換成真正具有分類鑑別性的聲學特徵。其中，前後文資訊、混淆資訊的使用具有以下的物理意涵：

一、就連人耳也無法對於時域過短（如小於 100 ms）的音素(phoneme)產生較好的辨識效果，那麼自動語音辨識系統在設計時的確需要考慮在現有幾近獨立的短期音框(short-term frame)之間，加入前後文相關資訊，使得系統所能處理的每一音框，其所含的資訊就不限於那短短的 10 至 20 ms [7]。常見的方法有：多向量輸入(multi-vector input)，也就是將每個音框與其前後各 4 至 5 個音框串接在一起，形成一個超級向量(super-vector)後再作降維處理[8-10]（見 5.2.6 節）；動態特徵(dynamic features)則是結合了時間導數(time derivatives)，試著捕捉短期音

[2] 筆者認爲，這種以人類聽覺知識爲本的自動語音辨識研究雖無法被證實爲語音訊號處理之必要進路，但卻有其在認知科學方面的解釋空間。事實上，人類聽覺系統的確受到某些物理限制，例如，人類聽不見某種頻率（如 20,000 Hz 以上）的聲音，而這些限制卻不會影響人腦對語言的辨識與理解，甚至是文字或文化的產生。文字或語言符號既源於人類的感知，那麼從人類的感覺經驗著手，極有可能是找出音訊與文字之間連結的最佳捷徑或基本原理，見[6]。

[3] 儘管在圖樣識別(pattern recognition)領域中，非線性技術於特徵處理中，特別是在影像處理的領域，也廣爲討論，如 Kernel PCA、Kernel LDA 等，見[21]，但在本論文中，特徵擷取主要是在線性轉換的框架下論述的，見[42]。而在語音處理的研究上，由於資料量十分龐大，現今非線性技術相較下較無突破性發展。

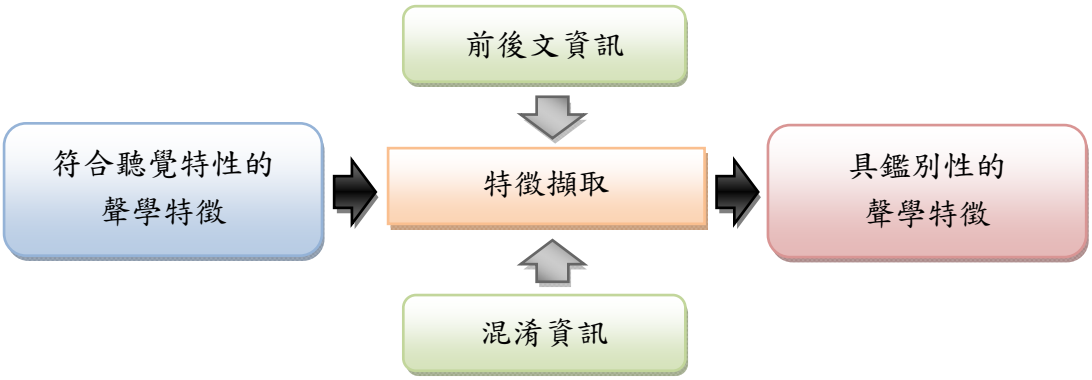


圖 1.1　本論文之研究目的與方法

框間的時間相關性[11]。

二、在認知科學上，部分學者認為，人類聽覺系統在處理音素時，具有一種反饋(feedback)的機制，使得人類能夠依據較高階的資訊，如詞彙知識(lexical knowledge)，來修正感知結果[4][12-17]。因此，許多自動語音辨識研究者已嘗試利用後端由辨識器提供之分類錯誤的混淆資訊，並根據一些準則，如最小分類錯誤(minimum classification error, MCE)[18-19]或最小音素錯誤(minimum phone error, MPE)[20]等，來產生較具鑑別性的聲學特徵。

而在語音辨識上，我們可把具分類資訊或混淆資訊的特徵轉換稱作鑑別式特徵擷取(discriminative feature extraction, DFE)。其它非鑑別式特徵擷取，如無分類資訊輔助的主成分分析(principal component analysis, PCA)[21-23]，雖不會遭遇訓練資料與測試資料不一致(mismatch)的情況，但在實務上卻無太大效果。綜合 Gales 與 Wang 的說法[24-25]，我們可依照操作空間的不同，將鑑別式特徵擷取分作兩大範疇：基於模型空間之鑑別式特徵擷取(model-space based DFE, MS-DEF)與基於特徵空間之鑑別式特徵擷取(feature-space based DFE, FS-DFE)[5]。在 MS-DEF 中，如 fMPE [20]、MCE [18]，其線性轉換的求取是與統計模型的參

[4] Norris 等人認爲，在人類聽者不會發生任何辨識錯誤的環境下，因著缺少分類錯誤資訊，反饋機制根本沒必要存在。但是 Tanenhaus 等人則反對這種奧坎剃刀(Occam's razors)式的論述，見[15, 16]。

[5] 本論文之後皆以 MS-DEF 與 FS-DFE 來分別簡稱『基於特徵空間之鑑別式特徵擷取』與『基於模型空間之鑑別式特徵擷取』。

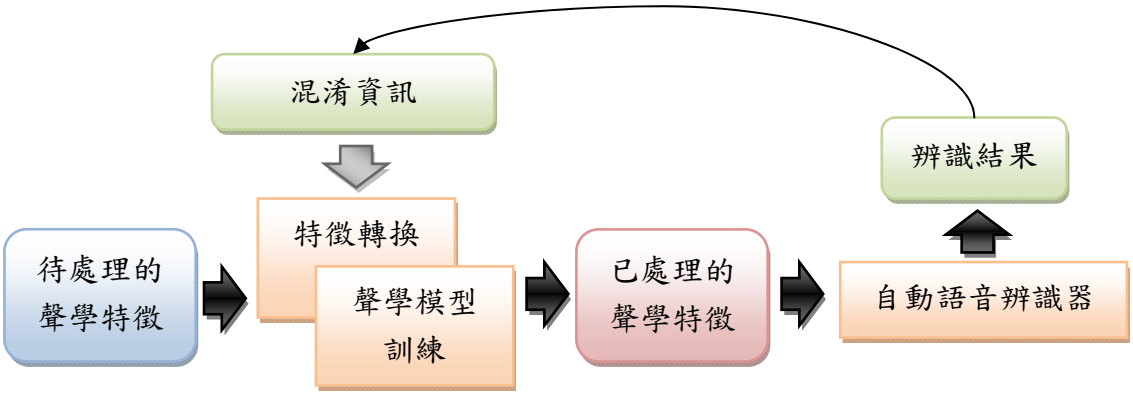


圖 1.2　以模型空間為基礎之鑑別性特徵擷取

數估計或分類器的分類規則緊密結合並一同進行的，如圖 1.2。相反地，FS-DFE 則是依據各種幾何或機率式的類別分離度(class separability)估量，獨立地求取線性轉換，以期轉換後的特徵在後端具有較好的分類效果，如圖 1.3。其常見的方法有線性鑑別分析(linear discriminant analysis, LDA)[6][9]、異方差線性鑑別分析(heteroscedastic linear discriminant analysis, HLDA)[26]、異方差鑑別分析(heteroscedastic discriminant analysis, HDA)[27]，或是以最小化分類錯誤或最大化交互資訊(maximum mutual information, MMI)為音素分離度量測的方法[28]等。

本論文的研究方向較傾向後者 FS-DFE，主要原因為：第一，它所需的計算複雜度(computational complexity)較低，因其線性轉換的求取不須同時處理所有的訓練資料；第二，在最佳化(optimization)過程中，某些方法，如 LDA，不僅具有輕省的可解性(lightweight solvability)而不需繁複的迭代(iterative)過程，也可保證所求出的轉換矩陣已是全域解(global solution)；第三，由於特徵擷取完全與聲學模型(acoustic models)分離，對於較複雜的自動語音辨識系統，聲學模型訓練模式的改變，就不會影響到前端的訊號處理，使得此系統易於被分析解構。當某些系統的聲學模型機制是固定的，或是以硬體方式呈現，那麼我們就能在無法更動硬體的情況下，對前端訊號處理進行研究或改善[29]。更重要的是，我們相信，根據 FS-DFE 所建立的方法較能更廣泛地應用在其它圖樣辨識(pattern recognition)

[6] 本論文之後皆以 LDA 來簡稱『線性鑑別分析』。

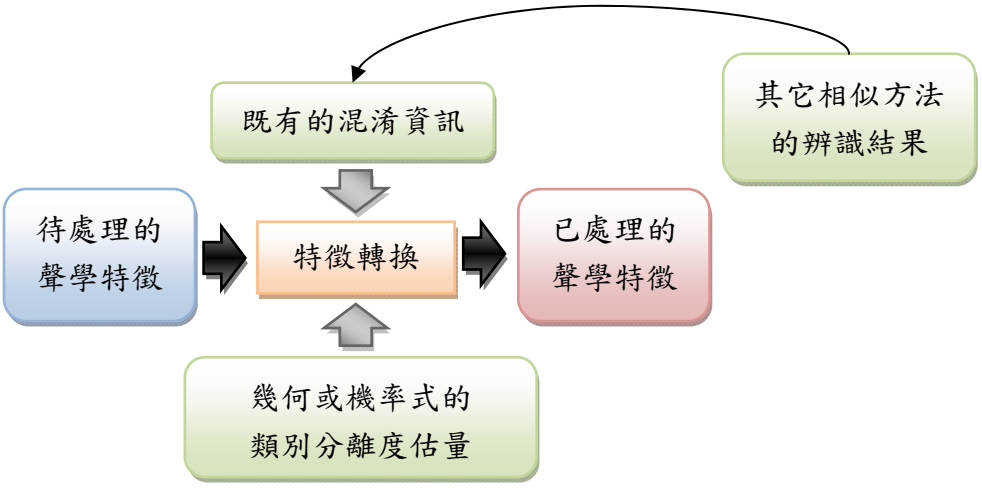


圖 1.3　以特徵空間為基礎之鑑別性特徵擷取

的領域中。

當然，我們無法避免在 FS-DFE 中既存的問題：類別分離度與辨識效果仍存有較大的差距。本論文的目標即在於如何使用辨識器產生的混淆資訊，在承繼傳統線性鑑別分析的可解性下，降低前端的特徵擷取和後端的辨識過程的不一致性。

1.2　研究基礎：線性鑑別分析

那麼，我們的研究為什麼要以線性鑑別分析(LDA)為基礎呢？LDA 是一種資料導引(data-driven)技術，它試著在類別內具有相同變異度(variability)的假設下，藉著最大化類別間的幾何分離度(geometric separability)來最佳化類別間的線性鑑別度。在前端聲學特徵擷取上，它最常被使用於由短期特徵向量(short-term feature vectors)所組成的多向量輸入，而就像一個二維的線性濾波器，應用於由所有聲學特徵向量所組成的特徵－時間平面，而後輸出具有鑑別性的特徵（見圖 5.3）[30]。

雖然 LDA 是一種資料導引的工程處理方法，也在許多大詞彙連續語音辨識

工作上獲得了不錯的成果[31]，但我們卻可發現事實上它具有某些符合人類聽覺系統的特性。由生理學可知，人類聽覺對於音頻的解析度會隨著頻率升高而遞減，且對於頻率在 4 Hz 之間的調變頻率(modulation frequency)十分敏感，而對每一特徵維度使用 LDA 所產生的調變頻率不僅具有相似的特性，其扮演的角色就像是中央頻率在 4 Hz 附近的帶通濾波器(band-pass filter)[32]。因此，本論文對於 LDA 的改善研究並不離開以人類聽覺特性為出發點的初衷。

1.3 論文貢獻

本論文的主要貢獻有三：

第一，我們完整地討論並分析了 LDA 的數學與物理意義，以及它在圖樣辨識中扮演的角色與功能，而這在一般教科書與學術論文中均屬少見。

第二，利用較複雜的分類器所產生之混淆資訊，改進 LDA 不利於分類的問題，並提出嶄新的類別分離度標準，將原本以最大化類別間幾何分離度為主的 LDA 目標函式，修改並擴展為以最大化類別間經驗分類正確率(empirical classification accuracy)為主的目標函式。

第三，在普遍化相似度比率檢定(generalized likelihood ratio test, GLRT)的框架下，設定與鑑別性有關、且打破同方差性(homoscedasticity)的兩個統計假設（虛無假設(null hypothesis)與對立假設(alternative hypothesis)），並據此求出使不具鑑別性之虛無假設愈趨於不成立的投影子空間。我們亦證明了傳統的 LDA 與有名的異方差線性鑑別分析(HLDA)都只是此框架下的一個特例。

以上貢獻之部份成果亦已發表於國際會議論文，其中，在利用混淆資訊改進 LDA 的部分也獲得了第六屆中文口語語言處理國際會議(ISCSLP 2008)的最佳學

生論文獎(best student paper award)。

1.4 論文架構

本論文接下來的架構如下：

第二章為背景知識的介紹。我們將概要說明統計式語音辨識的基本架構：特徵擷取、聲學模型、語言模型，以及聲學比對與語言解碼，之後並詳細描述特徵擷取的部分。我們也會探討近年來 FS-DFE 在語音辨識上的發展，特別是 LDA 的原理及其限制與改進。

第三章則描述我們對於傳統 LDA 的另一種改進，提出了加權式線性鑑別分析(weighting-based LDA, WLDA)的概念，並包括三種利用混淆資訊的方法：基於經驗錯誤率之權重式線性鑑別分析(empirical error rate based WLDA, EER-WLDA)、距離－錯誤耦合之權重式線性鑑別分析(distance-error coupled WLDA, DE-WLDA)、近似成對經驗正確標準(approximate pairwise empirical accuracy criterion, aPEAC)。此外，我們也利用混淆資訊提出了針對類別內散佈矩陣(within-class scatter)的重新估計：基於經驗錯誤率之類別內共變異矩陣(empirical error rate based within-class covariance matrix, EERW)。

在第四章中，我們將針對在第三章所提出的方法進行分析與異方差性(heteroscedasticity)的延伸，從相似度比例檢驗(likelihood ratio test, LRT)觀點來討論，並發展出嶄新的鑑別式分析方法，稱為普遍化相似度比率鑑別分析(generalized likelihood ratio discriminant analysis, GLRDA)。GLRDA 能普遍化現有的線性鑑別式分析，如 LDA、HLDA 等，並能與辨識器產生的混淆資訊結合，產生更具有普遍化能力(generalization)的特徵。

第五章介紹實驗環境與相關設定，包含了語料庫和大詞彙連續語音辨識、線性特徵轉換的應用，也整理了所有實驗結果，包含了基礎實驗、研究改進實驗等。

第六章為本論文的結論和未來展望。

第 2 章　背景介紹

在本章中，我們首先將概略地介紹統計式語音辨識的架構，特別是特徵擷取的部分。然後，我們也將會探討近年來線性特徵擷取技術在語音辨識領域的發展，特別是線性鑑別分析(linear discriminant analysis, LDA)的原理、限制、以及它常見的改進。

2.1　統計式語音辨識

自動語音辨識(automatic speech recognition, ASR)的目的在於給定一段人類所發出的語音訊號(speech signal)作為輸入後，期望系統能夠輸出其對應的正確文句(literal utterance)。作為圖形識別(pattern recognition)領域的一個重要分支，自動語音辨識基本上是由兩個主要階段所構成：特徵擷取(feature extraction)與分類(classification)，如圖 2.1。第一階段的特徵擷取牽涉到對於語音訊號的前端處理(front-end processing)，使得這些訊號能依據人類聽覺或其它有助於後端分類的特性，被壓縮成一連串的聲學特徵向量(acoustic feature vectors or observations)以作為後端分類用的輸入(input)。而在第二階段的分類過程中，當給定某一串特徵向量後，聲學模型(acoustic models)、語言模型(language models)與詞典(lexicon)則可用來推論出最可能代表它的文字語句。其中，聲學模型的作用在於能將每一串聲學特徵向量對應到個別的字(word)或次字元(sub-word unit)，如音節(syllable)或音素(phone)。詞典則作為聲學模型和語言模型間的橋樑，提供字與次字元之間之對應關係，並使得語言模型能利用含有句法(syntactic)或語意(semantic)的資訊，產生出最具有語言形式的文句。

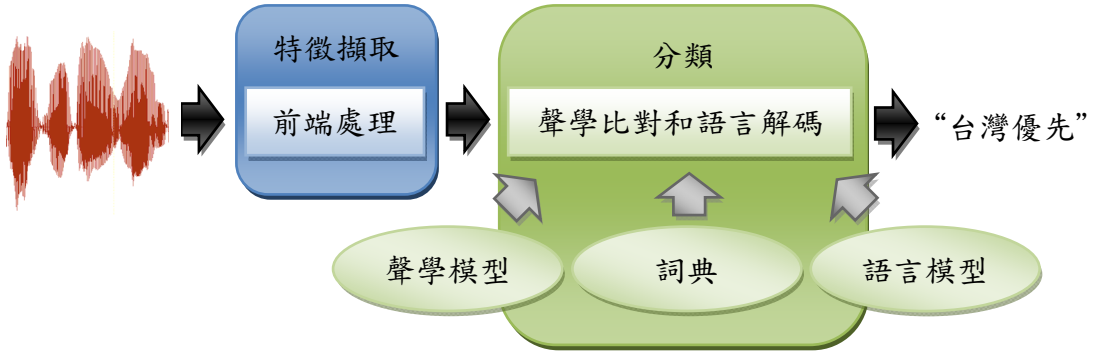


圖 2.1　語音辨識系統流程圖

要完成自動語音辨識系統，有許多實作的方式[7]。其中，基於隱藏式馬可夫模型(hidden Markov models, HMMs)的統計式模型可說是目前的主流技術，其主要精神在於藉由統計式架構的方法，在給定一串待測的聲學特徵向量 O 之前提下，找出最有可能發生的詞序列(word sequence)$\hat{W}$ 。我們可將機率式表達如下：

$$\hat{W} = \arg\max_{W \in \mathbf{W}_h} \underbrace{P(W \mid O)}_{\text{posterior}} \tag{2.1}$$

其中，$\mathbf{W}_h$ 代表所有可能詞序列之所構成的集合；$P(W \mid O)$ 為給定某串聲學特徵向量 $O = \{o_1, ..., o_t, ..., o_T\}$ 下，發生詞序列 W 的事後機率(posterior probability)，其中，o_t 為在某個時間點 t 的聲學特徵向量。因為搜尋具有最大事後機率的詞序列的過程並不會受到 O 本身發生的機率 $p(O)$ 所影響，我們便可應用貝氏定理(Bayes theorem)，將式(2.1)展開與簡化為：

$$\begin{aligned} \hat{W} &= \arg\max_{W \in \mathbf{W}_h} \frac{p(O \mid W)P(W)}{p(O)} \\ &= \arg\max_{W \in \mathbf{W}_h} \underbrace{p(O \mid W)}_{\text{likelihood}} \underbrace{P(W)}_{\text{prior}} \end{aligned} \tag{2.2}$$

由式(2.2)看出，最佳詞序列的計算是由兩種機率分佈所組成。其一是在給定某詞序列 W 下，產生聲學特徵向量 O 的機率或相似度(likelihood) $p(O \mid W)$ ，被稱之

[7] 其它異於統計式模型的 ASR 技術，如類神經網路系統(connectionist systems)或以知識爲基礎的系統(knowledge based systems)，見[35]。

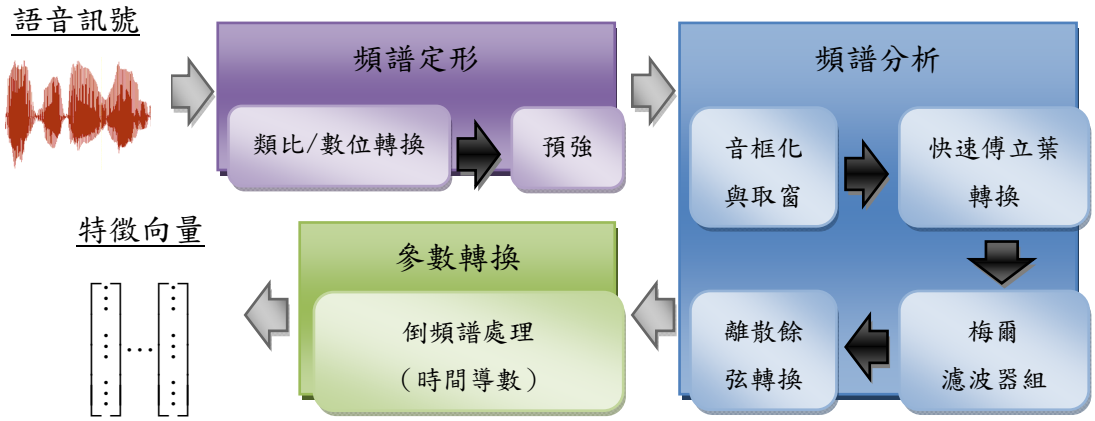


圖 2.2　梅爾倒頻譜係數產生流程圖

為聲學模型，一般被廣泛採用的是由左至右(left-to-right)的連續密度隱藏式馬可夫模型(continuous density HMMs, CDHMMs)。其二是在語言中發生此段詞序列 W 的事前機率(prior probability) $P(W)$，被稱之為語言模型，其最常見的實作方法為 n－連(n-gram)語言模型。而以式(2.2)為目標函數的搜尋演算法即是所謂的最大化事後機率(maximum a posteriori, MAP)解碼方法。

由於前端特徵擷取，也就是聲學特徵向量 O 的產生方式才是本論文所主要探討的部份，其它關於統計式聲學與語言學模型的詳細架構、模型訓練方法、聲學比對與語言解碼的過程可參考[33-36]。

2.2　聲學特徵擷取

如前所述，特徵擷取是自動語音辨識的關鍵成分，它負責精簡且扼要地捕捉潛藏在語音訊號中的聲學特徵，使這些特徵不僅具有人類感知上的物理意義，也能實用地與後端的統計模型或辨識器相互契合，以達到更高的辨識率。

若要使語音訊號能使用在隱藏式馬可夫模型架構下的語音辨識中，則此連續的語音訊號就必須被轉換成為一連串似穩定(quasi-stationary)的離散時間向量(discrete-time vector sequences)或音框(frame)序列[37-38]，這乃是隱藏式馬可夫模

型一個甚難打破的重要假設。基於此假設，語音訊號的頻譜包絡(spectral envelope)統計資訊要能夠從每一音框中被擷取出來，使得音框中含有最豐富的聲學資訊。

傳統上語音特徵擷取可以分為三大步驟[8][39]，頻譜定形(spectral shaping)、頻譜分析(spectral analysis)和參數轉換(parametric transform)，如圖 2.2，以下將詳細介紹之。

2.2.1 頻譜定形

在頻譜定形的過程中，類比音訊會先透過類比／數位轉換器(A/D converter)被轉為數位訊號，並經由一些數位濾波器，加強訊號中重要的頻率成分，稱之為預強(pre-emphasis)。預強可被視為一種高通濾波器(high-pass filter)，其 Z 轉換(Z transform)為

$$H[\mathbf{z}] = 1 - \alpha_{pre} \times z^{-1} \tag{2.3}$$

一般在時域上我們會以式(2.4)來處理：

$$\hat{s}(n) = s(n) - \alpha_{pre} \times s(n-1) \tag{2.4}$$

其中，$\hat{s}(n)$ 為第 n 個採樣點經預強後的輸出訊號，$s(n)$ 為第 n 個採樣點的輸入訊號，a_{pre} 為預強的參數，在本論文中設為 0.975。

預強的主要目的在於，語音在空氣中傳送時，高頻部分的能量會隨著時間快速遞減，而人耳的外聽道的共振作用恰可提高頻率區間為 2000～5000 Hz 的聲音強度，因此我們需要預強來模擬人耳外聽道的功能以彌補聲音高頻部分的能量損失。其次，也因著人耳對於音訊頻譜上超過 1000 Hz 的區域較為敏感，預強就能

[8] 其它非基於頻譜封包的方法，見[5, 36]。

加強這些高頻共振峰(formants)的重要性[39]。

2.2.2 頻譜分析

頻譜分析，顧名思義，即藉著分析語音訊號的頻譜(spectra)擷取出有用的聲學特徵。由於語音訊號的波形在時域上變化十分迅速且無一定的規則，而難以在後端作進一步處理。但若藉著快速傅立葉轉換(fast Fourier transform, FFT)，把語音訊號由時域轉成頻域，則可發現在短時間(20-40 ms)的情況下，頻譜呈現出週期性的變化。因此我們可假設語音訊號為短時間穩定(short-time stationary)或似穩定的(quasi-stationary)，便可每隔一小段時間對語音訊號取一個音框。為了讓相鄰音框與音框之間能保有相互關聯，相鄰音框間會重疊(overlap)一小段時間，這些動作稱為音框化(framing)。在本論文中一個音框長設定為 20 ms，音框間重疊為 10 ms。

又因著每個音框是在固定時間點被切割，其邊界便會造成不連續現象，這會使得音框經過後面的快速傅立葉轉換將產生高頻雜訊。為了減低此雜訊的產生，音框在快速傅立葉轉換前會乘上一個漢明窗(Hamming window)，稱之為取窗(windowing)，以增加音框附近的連續性。漢明窗的表達式如下，其中為 β_{Ham} 控制漢明窗的參數，在本論文設定為 0.46。

$$w(n)=\begin{cases}(1-\beta_{Ham})-\beta_{Ham}\cos\left(\dfrac{2\pi n}{N-1}\right), & n=0,1,...,N-1\\ 0, & \text{otherwise}\end{cases} \tag{2.5}$$

在音框化與取窗的過程後，我們就可藉由快速傅立葉轉換將語音訊號轉換為頻域上的功率頻譜(power spectrum)，其表達式如下：

$$X_i(e^{j2\pi k/N})=\sum_{n=0}^{N-1}x_i(n)e^{j2\pi k/N} \tag{2.6}$$

其中，x_i 是第 i 個音框向量，$x_i(n)$ 為第 i 個音框向量中的第 n 個值，N 是頻域上的取樣點數。

頻譜分析的方法有很多，梅爾倒頻譜係數(Mel-frequency cepstral coefficients, MFCCs)[9]是目前在語音辨識上最廣為使用的聲學特徵[40]。MFCCs 的演算法由三個主要部分組成：梅爾頻率尺度(Mel-frequency scale)、三角濾波器(triangular filters)與離散餘弦轉換(discrete cosine transform, DCT)。

根據人類聽覺特性，梅爾頻率尺度藉著扭曲原本高頻（大於 1000 Hz）聲音的線性頻率尺度來模擬人耳內部基底膜(basiler membrane)傳遞到聽覺神經的現象，其表達式如下，在本論文中參數 γ 設為 1127：

$$\mathrm{Mel}(f_{Hz}) = \gamma \log_{100}\left(1 + \frac{f_{Hz}}{700}\right) \tag{2.7}$$

三角濾波器的功能，除了降低資料量外，則具有對頻譜進行平滑化並消除諧波(harmonics)的作用，以突顯原語音的共振峰。三角濾波器的表達式如下：

$$H_m[k] = \begin{cases} 0, & k < f[m-1] \\ \dfrac{k - f[m-1]}{f[m] - f[m-1]}, & f[m-1] \le k \le f[m] \\ \dfrac{f[m+1] - k}{f[m+1] - f[m]}, & f[m] \le k \le f[m+1] \\ 0, & k > f[m+1] \end{cases} \tag{2.8}$$

其中 $f[m]$ 為第 m 個三角濾波器的中心點，$H_m[k]$ 為 k 頻率在第 m 個三角濾波器的權重(weight)，N 為頻域上取樣點數。$f[m]$ 可進一步表示成：

$$f[m] = \left(\frac{N}{F_s}\right)\mathrm{Mel}^{-1}\left(\mathrm{Mel}(f_l) + m\frac{\mathrm{Mel}(f_h) - \mathrm{Mel}(f_l)}{M+1}\right) \tag{2.9}$$

其中 F_s 為取樣頻率，f_l 為三角濾波器組中最低的頻率，f_h 為三角濾波器組中最

[9] 本論文之後皆以 MFCC(s)來簡稱『梅爾倒頻譜係數』。

高的頻率，M 為三角濾波器組的個數。在本論文中共取 18（$M=18$）個三角濾波器。

為了模擬人耳對於頻率能量變化的遲鈍，我們會再將三角濾波器輸出的值作對數轉換，再經由離散餘弦轉換而成為 MFCCs：

$$C_t[n]=\sum_{k=1}^{N}\log|X_t(e^{j2\pi kn/N})|\cos\left(n(k-0.5)\frac{\pi}{k}\right),\ n=0,1,...,L \tag{2.10}$$

其中，其中 $X_t(\cdot)$ 是第 t 個音框向量在頻域的成分，N 是頻域上取樣點數，n 是第 n 個 MFCC。離散餘弦轉換是一種反傅立葉轉換(inverse Fourier transform, IFT)，因此我們將轉換後的特徵稱為倒頻譜(cepstrum)，其不僅能代表每一音框語音訊號的頻譜包絡變化資訊，更能降低特徵維度間的空間關聯性(spatial correlation)，使後端的隱藏式馬可夫模型在處理每一類別的共變異矩陣(covariance matrices)並作對角化假設時，資訊損失不會太多。再者，此降維的動作也能加快辨識的效率。

2.2.3 參數轉換

由於隱藏式馬可夫模型亦具有特徵向量獨立(observation independence)的假設，而忽略了語音訊號在時間上的關聯性。因此，我們不僅讓相鄰音框保有重疊聲學資訊外，更在每一音框間加入動態資訊(dynamic information)。常見的方法則是在目前音框的特徵向量之後，加入它與附近音框的一階和二階的時間差量[11]：

$$\Delta C_t[n]=\frac{\sum_{p=1}^{P}p(C_{t+p}[n]-C_{t-p}[n])}{2\sum_{p=1}^{P}p^2} \tag{2.11}$$

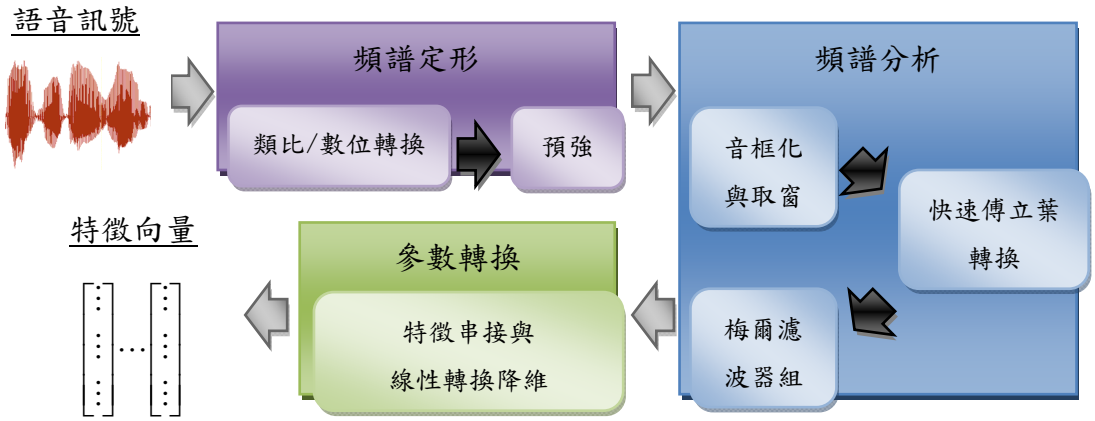


圖 2.3　多向量輸入處理流程圖

$$\Delta^2 C_t[n] = \frac{\sum_{p=1}^{P} p(\Delta C_{t+p}[n] - \Delta C_{t-p}[n])}{2\sum_{p=1}^{P} p^2} \tag{2.12}$$

其中，n 一般為 MFCCs 加上能量共 13 維，加入一階與二階的差量計算後，最後特徵擷取的維度為 39 維。

2.2.4　另一種框架：多向量輸入

除了以時間導數作為時間或動態資訊上的整合方式外，Makino 等人提出了另一種稱為多向量輸入(multi-vector input)的方法，能夠利用來自某一音框附近較長的語音段落，與原音框前後串接，形成一個新的時域－頻域(temporal-spectral)特徵向量，再對這些短期的語音段落進行分類[10][10]。無可避免地，這種處理短期特徵向量的方式有可能會嚴重破壞原有向量所屬的類別與機率分佈，因此，一般來說，多向量輸入還會再進一步結合鑑別分析(discriminant analysis)，依據類別資訊將原有的特徵經過線性或非線性轉換成更具鑑別性的特徵，例如，LDA 就在此方面取得不錯的效果[41]。

[10] Makino 等人的工作在於以多向量作爲後端多層式感知器(multi-layer perceptron, MLP)模型的輸入(input)，用來從事子音(consonant)辨識。Hermansky 認爲，這種時域－頻域的圖樣(pattern)較適用於特別設計的分類器，如多層式感知器，見[7]。

表 2.1　本論文中梅爾倒頻譜係數架構之特徵擷取使用到的係數

取樣頻率	16 kHz
音框長度	320 點, 20 ms
音框重疊	160 點, 10 ms
預強	0.975
漢明窗	0.46
三角濾波器	18 組
離散餘弦轉換	12 階
能量及差量	能量維 1 維，一階、二階差量倒頻譜各 13 維，總共 39 維

值得一提的是，以多向量輸入這種特徵串接(feature concatenation)的方式，搭配主成分分析(principal component analysis, PCA)[11]或其它鑑別分析（如 LDA），也可用來取代傳統離散餘弦轉換的工作[12][19]，如圖 2.3，因為此二者在一定程度上亦具特徵去相關(feature de-correlation)的作用（見 2.3.2 節）。

[11] 本論文之後皆以 PCA 來簡稱『主成分分析』。

[12] 亦有許多方法將多向量輸入應用在倒頻譜係數(cepstrum)之後，見[26, 27]，但若我們使用以線性鑑別分析爲基礎的降維方法提前應用在梅爾頻譜(Mel-spectrum)之後，一來可保留了與人類聽覺系統相似的特性，見[32]；二來則不必再重複與離散餘弦轉換相同性質的處理。

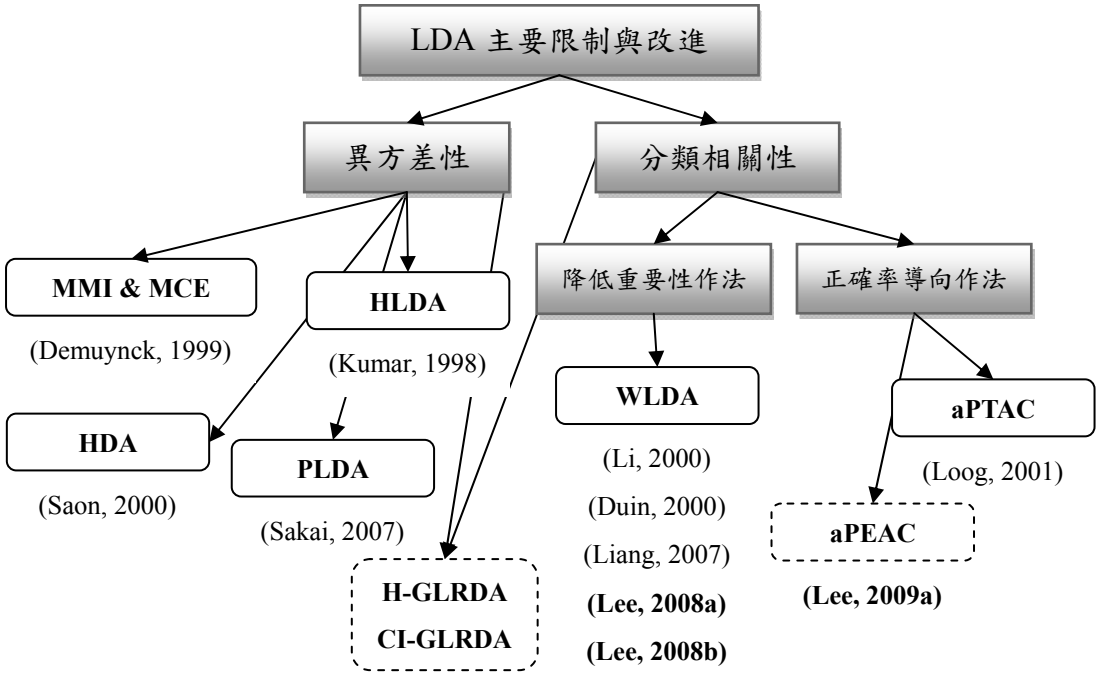


圖 2.4　近年來對 LDA 本質之改進方法
（虛線部分為本論文所提出）

2.3　線性鑑別分析(LDA)

如第 1 章所述，本論文著重在討論基於特徵空間之鑑別式特徵擷取(feature-space based DFE, FS-DFE)，其目標在於尋找一個線性轉換，將原來 n 維的特徵向量，投射至 d 維的子空間（$d < n$），使得新特徵在類別間具有的較好的鑑別力[42]。而在所有 FS-DFE 的方法中，要屬 LDA 最簡單且最廣為使用。據筆者所知，Hunt 為第一人將 LDA 用於分離由一連串聲學特徵向量所組成的字或音節(syllables)[43]。之後，Brown 在他的博士論文中也首次將 LDA 應用在多向量輸入的框架下，並證實了在使用離散隱藏式馬可夫模型(discrete HMMs, DHMMs)所建立的辨識器上，LDA 具有優於 PCA 的表現[30]。以後的幾年，LDA 應用在小詞彙連續語音辨識上均有不錯的成果[44-45]，但在大詞彙系統上的辨識結果卻有好有壞[46-47]。

近十年，語音處理研究者對於 LDA 的研究開始由實作上的細節部分，例如

最小分類單位的決定[31]、機率分佈與實際資料的一致性[48]等，漸漸轉為以 LDA 的本質缺陷為改良進路之研究。我們可將這些研究大致分為兩大類：異方差性(heteroscedasticity)－打破 LDA 中對於每一類別母體(population)具有相同共變異矩陣(covariance matrix)的假設；分類相關性(classification-related property)－將 LDA 更緊密聯於分類器(classifier)結構或分類規則(allocation rule)，如圖 2.4。其中，我們又可將分類相關性的部分再細分為兩種方案：調整類別間幾何距離之貢獻度來降低重要性的作法(de-emphasis scheme)，以及考慮分類結果的正確率導向作法(accuracy-driven scheme)。之後，我們將會介紹各個方法，特別是本論文的兩個主軸：近似成對經驗正確率標準(approximate pairwise empirical accuracy criterion, aPEAC)和能夠同時含有異方差性與分類相關性的普遍化相似度比率鑑別分析(generalized likelihood ratio discriminant analysis, GLRDA)。

2.3.1 目標函式

若我們將全部資料 $\{\mathbf{x} \in \Re^{n\times 1}\}$ 分為 C 個類別，n_i 為屬於類別 C_i 的資料總數，N 為全部資料總數，則類別間散佈矩陣(between-class scatter matrix) $\mathbf{S}_B \in \Re^{n\times n}$ 與類別內散佈矩陣(within-class scatter matrix) $\mathbf{S}_W \in \Re^{n\times n}$ 的定義分別如下[49]：

$$\mathbf{S}_B = \sum_{i=1}^{C} p_i(\mathbf{m}_i - \overline{\mathbf{m}})(\mathbf{m}_i - \overline{\mathbf{m}})^T \tag{2.13}$$

$$\mathbf{S}_W = \sum_{i=1}^{C} p_i \mathbf{S}_i \tag{2.14}$$

其中，$\overline{\mathbf{m}}$ 為全部資料的期望值（平均）向量，p_i、$\mathbf{m}_i$ 與 $\mathbf{S}_i$ 分別為類別 C_i 的事前機率、期望值向量與共變異矩陣，三者的數學定義如下[13]：

[13] 注意，這裡對於所有資料母體或每一類別母體之統計量的估測(estimation)均是基於最大相似度估測法(maximum likelihood estimation, MLE)而得。其中，共變異矩陣的估測是有偏差的(biased)，見[22]。但由於在語音處理的實務上，每一類別的資料數都不小，這種偏差(bias)便可忽略之。

$$p_i = \frac{n_i}{N} \tag{2.15}$$

$$\mathbf{m}_i = \frac{1}{n_i}\sum_{\mathbf{x}\in C_i}\mathbf{x} \tag{2.16}$$

$$\mathbf{S}_i = \frac{1}{n_i}\sum_{\mathbf{x}\in C_i}(\mathbf{x}-\mathbf{m}_i)(\mathbf{x}-\mathbf{m}_i)^T \tag{2.17}$$

此外，我們更可進一步將類別間散佈矩陣$\mathbf{S}_B$表示成

$$\begin{aligned}\mathbf{S}_B &= \frac{1}{2}\sum_{i=1}^{C}\sum_{j=1}^{C} p_i p_j (\mathbf{m}_i-\mathbf{m}_j)(\mathbf{m}_i-\mathbf{m}_j)^T \\ &= \frac{1}{2}\sum_{i=1}^{C}\sum_{j=1}^{C} p_i p_j \mathbf{S}_{ij}\end{aligned} \tag{2.18}$$

其中，$\mathbf{S}_{ij} = (\mathbf{m}_i-\mathbf{m}_j)(\mathbf{m}_i-\mathbf{m}_j)^T$，為類別配對(class-pair)$C_i$與$C_j$之期望值向量的並向量積(dyadic product)。由式(2.18)出發，我們可反推得式(2.13)：

$$\begin{aligned}&\frac{1}{2}\sum_{i=1}^{C}\sum_{j=1}^{C} p_i p_j (\mathbf{m}_i-\mathbf{m}_j)(\mathbf{m}_i-\mathbf{m}_j)^T \\ &= \frac{1}{2}\sum_{i=1}^{C}\sum_{j=1}^{C} p_i p_j \left((\mathbf{m}_i-\overline{\mathbf{m}})+(\overline{\mathbf{m}}-\mathbf{m}_j)\right)\left((\mathbf{m}_i-\overline{\mathbf{m}})+(\overline{\mathbf{m}}-\mathbf{m}_j)\right)^T \\ &= \frac{1}{2}\sum_{i=1}^{C}\sum_{j=1}^{C} p_i p_j \begin{pmatrix}(\mathbf{m}_i-\overline{\mathbf{m}})(\mathbf{m}_i-\overline{\mathbf{m}})^T+(\mathbf{m}_j-\overline{\mathbf{m}})(\mathbf{m}_j-\overline{\mathbf{m}})^T+ \\ (\overline{\mathbf{m}}-\mathbf{m}_j)(\mathbf{m}_i-\overline{\mathbf{m}})^T+(\mathbf{m}_i-\overline{\mathbf{m}})(\overline{\mathbf{m}}-\mathbf{m}_j)^T\end{pmatrix} \\ &= \frac{1}{2}\sum_{i=1}^{C} p_i (\mathbf{m}_i-\overline{\mathbf{m}})(\mathbf{m}_i-\overline{\mathbf{m}})^T + \frac{1}{2}\sum_{j=1}^{C} p_j (\mathbf{m}_j-\overline{\mathbf{m}})(\mathbf{m}_j-\overline{\mathbf{m}})^T \\ &= \sum_{i=1}^{C} p_i (\mathbf{m}_i-\overline{\mathbf{m}})(\mathbf{m}_i-\overline{\mathbf{m}})^T = \mathbf{S}_B\end{aligned} \tag{2.19}$$

此外，由於本論文處理的問題有關語音特徵擷取，通常資料量（見表 3.1）遠大於資料維度（最多 162 維），並不會遇到小資料量問題(small-sample-size problem, SSS problem)，因此不會發生共變異矩陣爲奇異(singular)的情形。

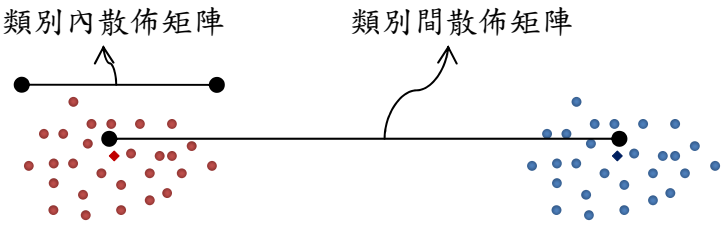


圖 2.5　線性鑑別分析的幾何示意圖

由式(2.14)與式(2.18)可看出，類別內散佈矩陣$\mathbf{S}_W$其實就是所有類別之共變異矩陣的算術平均(arithmetic average)，而類別間散佈矩陣$\mathbf{S}_B$被表示成類別配對的形式，有助於我們將來針對每一類別配對對於 LDA 求取的貢獻作調整(modification)。

LDA 也可被稱為典型變數分析(canonical variate analysis, CVA)，起初是被 Fisher 所引進，作為分離兩個類別母體的統計方法[50-51]，爾後由 Rao 將之延伸至多類別母體[52]。在資料降維上，它的基本精神在於尋求一個線性轉換矩陣$\mathbf{\Theta} \in \Re^{n \times d}$，藉著最大化其類別間散佈矩陣$\mathbf{S}_B$與類別內散佈矩陣$\mathbf{S}_W$的比率，能夠在 n 維的原始訓練資料$\{\mathbf{x} \in \Re^{n \times 1}\}$在經過轉換至 d 維的子空間$\Re^{d \times 1}$後（$d < n$），鑑別性資訊的損失降至最低[14][49, 53]，如圖 2.5。LDA 目標函式主要有兩種形式，分別以線性代數中跡數(trace)[54]與行列式(determinant)[55]來表示：

$$\begin{aligned} J_{\text{LDA\_TR}}(\mathbf{\Theta}) &= \text{trace}\left((\mathbf{\Theta}^T\mathbf{S}_W\mathbf{\Theta})^{-1}(\mathbf{\Theta}^T\mathbf{S}_B\mathbf{\Theta})\right) \\ &= \frac{1}{2}\sum_{i=1}^{C}\sum_{j=1}^{C}\text{trace}\left((\mathbf{\Theta}^T\mathbf{S}_W\mathbf{\Theta})^{-1}(\mathbf{\Theta}^T\mathbf{S}_{ij}\mathbf{\Theta})\right) \end{aligned} \tag{2.20}$$

$$J_{\text{LDA\_DET}}(\mathbf{\Theta}) = \frac{|\mathbf{\Theta}^T\mathbf{S}_B\mathbf{\Theta}|}{|\mathbf{\Theta}^T\mathbf{S}_W\mathbf{\Theta}|} \tag{2.21}$$

式(2.20)與式(2.21)可被證明具有相同的解集合[49]，$\mathbf{\Theta}$可經由處理輕省(lightweight)的普遍化本徵值問題(generalized eigenvalue problem)而求得：

[14] 資料經過降維處理後，所含資訊量不是維持不變，就是減少。

$$\mathbf{S}_B\boldsymbol{\theta}_i = \lambda_i \mathbf{S}_W\boldsymbol{\theta}_i \tag{2.22}$$

其中，λ_i 為 $\mathbf{S}_W^{-1/2}\mathbf{S}_B$ 之第 i 大的非零本徵值(nonzero eigenvalue)，而 $\boldsymbol{\theta}_i$ 則為其對應的本徵向量(eigenvector)。因此，$\boldsymbol{\Theta}$ 最後可被表示為 $[\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, ..., \boldsymbol{\theta}_d]$。

值得注意的是，d 值大小是受限制的，除了一定不大於資料原始維度外，它也與相異類別的總數有關。以下的命題 2.1 說明了資料中所有相關的距離資訊都包含在由類別期望值向量所展開(span)之最多 $C-1$ 維的子空間，這會影響到 LDA 在解決多類別分類問題的能力，我們將在 2.3.4 節討論之[56]。

命題 2.1：在 LDA 中，投影子空間之維度的限制為 $d \le \min(n, C-1)$，其中，d 為非零本徵值（或投影子空間的最大維度），n 為原始維度，C 為類別總數。

證明：C 個向量形成向量組 M 如下：

$$M = \{p_i(\mathbf{m}_i - \overline{\mathbf{m}}) \mid 1 \le i \le C\} \tag{2.23}$$

因為 $\sum_i p_i\mathbf{m}_i = \overline{\mathbf{m}}$，式(2.23)滿足了

$$p_1(\mathbf{m}_1 - \overline{\mathbf{m}}) + p_2(\mathbf{m}_2 - \overline{\mathbf{m}}) + ... + p_C(\mathbf{m}_C - \overline{\mathbf{m}}) = \mathbf{0} \tag{2.24}$$

因此，任一向量 $p_i(\mathbf{m}_i - \overline{\mathbf{m}})$ 均可表示為其它 $C-1$ 個向量的線性組合，也就是說，由向量組 M 生成(span)的空間維度 $d \le C-1$。假設有一向量 $\mathbf{g}$ 與向量組 M 中任一向量均正交，即 $p_i(\mathbf{m}_i - \overline{\mathbf{m}})^T\mathbf{g} = 0$，則根據式(2.13)中 $\mathbf{S}_B$ 的定義，

$$\mathbf{S}_B\mathbf{g} = \sum_{i=1}^{C} p_i(\mathbf{m}_i - \overline{\mathbf{m}})(\mathbf{m}_i - \overline{\mathbf{m}})^T\mathbf{g} = \sum_{i=1}^{C}(\mathbf{m}_i - \overline{\mathbf{m}})0 = \mathbf{0} = 0\mathbf{g} \tag{2.25}$$

將式(2.25)中等號兩邊同乘 $\mathbf{S}_W^{-1}$，可得 $\mathbf{S}_W^{-1}\mathbf{S}_B\mathbf{g} = 0\mathbf{g}$，說明了 $\mathbf{S}_W^{-1}\mathbf{S}_B$ 具有 $n-d$ 個本徵值為 0 的正交本徵向量(orthogonal eigenvectors)，這蘊含了 $\mathbf{S}_W^{-1}\mathbf{S}_B$ 只具有 d 個或比

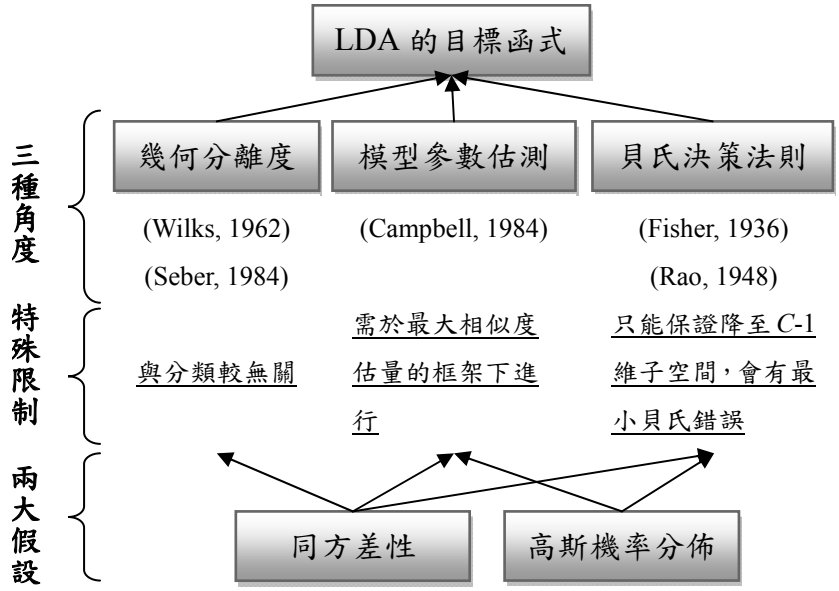


圖 2.6　LDA 目標函式產生的三種角度

d 更小的非零本徵值。又，$d \le C-1$，因此 $\mathbf{S}_W^{-1}\mathbf{S}_B$ 的非零本徵值必須滿足 $d \le \min(n, C-1)$。 ■

此外，值得一提的是，若我們對原始資料的做非奇異線性轉換(nonsingular linear transformation)[15]，也不會影響 $\boldsymbol{\Theta}$ 的求取與 LDA 目標函式的輸出值。因此，滿足於最大化 LDA 目標函式的 $\boldsymbol{\Theta}$ 並無唯一解[16]，且它的縮放比例(scaling)可以是任意的(arbitrary)，這使得每一類別資料在線性轉換前後的機率相似度可能會差距太大而影響辨識器的處理[57]。為了解決此縮放比例問題(scaling problem)，我們必須在 LDA 的求解過程中加上限制，使得 $\boldsymbol{\Theta}$ 的縮放比例適中，相關的技術將在 2.3.2 節討論之。

為了說明 LDA 目標函式的來源，我們可以從三個角度來看：使用貝氏決策法則(Bayes decision rule)來決定類別間最佳的決策邊界(decision boundary)、使用類別間幾何距離(geometrical distance)作為類別間分離度的量測(measurement)、使

[15] 換句話說，同構空間(isomorphic spaces)具有相同的 LDA 轉換。

[16] Prieto (2003) 也因此提出了 LDA 轉換矩陣的一般解形式。在 ASR 實驗中，的確可以證明，不同縮放比例的 LDA 轉換矩陣，會造成些微不同的辨識率。

用最大化相似度估計法(maximum likelihood estimation, MLE)對每一類別的統計模型和線性轉換作參數估計(parameter estimation)[17]。從這三個角度，加上其本身具有的假設，均可以推導出 LDA 目標函式。許多論文與教科書也常常因著這三個角度，而對 LDA 的限制和假設有所誤會，其中最有名的兩個問題為：

一、LDA 到底有沒有假設所有類別資料的機率分佈都遵循高斯分佈(Gaussian distribution)或其它機率分佈？

二、LDA 的同方差性(homoscedasticity)[18]假設，也就是假設所有類別母體的共變異矩陣都相同，是在甚麼意義下說的？

事實上，LDA 所處理的問題就是對轉換矩陣$\mathbf{\Theta}$作參數估計。經過文獻整理，我們可以歸納出圖 2.6，並發現，當我們使用貝氏決策法則或最大化相似度估計法，LDA 才有高斯機率分佈的假設。而至於同方差性假設，則是這三種角度都必要的假設，也因此是 LDA 的重要限制，但它的假設時機與意義隨三種角度而各異，將在之後章節討論之。此外，我們也可以發現 LDA 只有在貝氏決策法則的角度下才有分類(allocation)上的意義，且此分類是具有限制的。在同方差性和高斯機率分佈的假設下，LDA 只能決定出最佳的$C-1$維子空間，也就是$\mathbf{\Theta} \in \Re^{k\times d}$（$d = C-1 < k$），在此空間內，類別間會有最小的貝氏錯誤(Bayes error)[19]。

2.3.2 幾何分離度的意義與分析

儘管 LDA 目標函式之兩個形式（式(2.20)與式(2.21)）均具有相同的解集合，

[17] 有些教科書和論文最多只提及了前兩個角度，見[56]。

[18] 『同方差性』和『異方差性』分別是統計學詞彙『homoscedasticity』和『heteroscedasticity』的正式翻譯，而『同質性』和『異質性』則是在線性代數中分別對於『homogeneity』和『heterogeneity』的常見中文翻譯。

[19] 其實，在分類上較佳的標準爲最小貝氏風險(minimum Bayes risk, MBR)，不過我們在此缺乏關於分類花費(cost)的先備知識，因此只考慮 0/1 風險(0/1 risk)的情形。關於貝氏錯誤的定義與解說，見[56]。

它們在幾何分離度上的物理意義卻不一樣[20]。但在本節，我們只討論式(2.20)，它的物理意義如下之命題 2.2：

命題 2.2：假設所有類別母體均具有相同的共變異矩陣$\mathbf{S}_W$，也就是對任一類別C_i來說，$\mathbf{S}_i = \mathbf{S}_W$（即同方差性假設），則最大化式(2.20)等同於最大化在投影空間中之平均類別間的馬氏距離(Mahalanobis distance)平方。

證明：經過$\boldsymbol{\Theta}$轉換，任兩類別C_i和C_j之間之馬氏距離平方可被定義為

$$D_{ij}^2 = (\boldsymbol{\Theta}^T\mathbf{m}_i - \boldsymbol{\Theta}^T\mathbf{m}_j)^T(\boldsymbol{\Theta}^T\mathbf{S}_W\boldsymbol{\Theta})^{-1}(\boldsymbol{\Theta}^T\mathbf{m}_i - \boldsymbol{\Theta}^T\mathbf{m}_j) \tag{2.26}$$

則平均類別間之馬氏距離平方為$\frac{1}{2}\sum\sum p_i p_j D_{ij}^2$。根據式(2.18)並進一步推導：

$$\begin{aligned}
&\frac{1}{2}\sum_{i=1}^{C}\sum_{j=1}^{C} p_i p_j D_{ij}^2 \\
&= \frac{1}{2}\sum_{i=1}^{C}\sum_{j=1}^{C} p_i p_j (\boldsymbol{\Theta}^T\mathbf{m}_i - \boldsymbol{\Theta}^T\mathbf{m}_j)^T(\boldsymbol{\Theta}^T\mathbf{S}_W\boldsymbol{\Theta})^{-1}(\boldsymbol{\Theta}^T\mathbf{m}_i - \boldsymbol{\Theta}^T\mathbf{m}_j) \\
&= \frac{1}{2}\sum_{i=1}^{C}\sum_{j=1}^{C} p_i p_j \,\mathrm{trace}\Big((\boldsymbol{\Theta}^T\mathbf{m}_i - \boldsymbol{\Theta}^T\mathbf{m}_j)^T(\boldsymbol{\Theta}^T\mathbf{S}_W\boldsymbol{\Theta})^{-1}(\boldsymbol{\Theta}^T\mathbf{m}_i - \boldsymbol{\Theta}^T\mathbf{m}_j)\Big) \\
&= \frac{1}{2}\sum_{i=1}^{C}\sum_{j=1}^{C} p_i p_j \,\mathrm{trace}\Big((\boldsymbol{\Theta}^T\mathbf{S}_W\boldsymbol{\Theta})^{-1}(\boldsymbol{\Theta}^T\mathbf{m}_i - \boldsymbol{\Theta}^T\mathbf{m}_j)(\boldsymbol{\Theta}^T\mathbf{m}_i - \boldsymbol{\Theta}^T\mathbf{m}_j)^T\Big) \\
&= \frac{1}{2}\sum_{i=1}^{C}\sum_{j=1}^{C} p_i p_j \,\mathrm{trace}\Big((\boldsymbol{\Theta}^T\mathbf{S}_W\boldsymbol{\Theta})^{-1}(\boldsymbol{\Theta}^T\mathbf{S}_{ij}\boldsymbol{\Theta})\Big) = J_{LDA_TR}(\boldsymbol{\Theta})
\end{aligned} \tag{2.27}$$

由式(2.27)知，LDA 的目標函式恰等於平均類別間之馬氏距離平方。 ■

由命題 2.2，我們可以看出兩件事實：第一，LDA 視馬氏距離為類別間幾何分離度的量測標準，而這種距離標準並不會受到各個維度間縮放比例的差異所影響。第二，在同方差性的假設中，所有類別母體共有的共變異矩陣恰好就是它們

[20] 一個非奇異共變異矩陣的行列式被稱爲普遍化變異度(generalized variance)，可用來表示空間中所有資料點展開的體積(volume)，而式(2.21)即用此來描述資料的離散程度，相關的說明與證明，見[52, 55]。

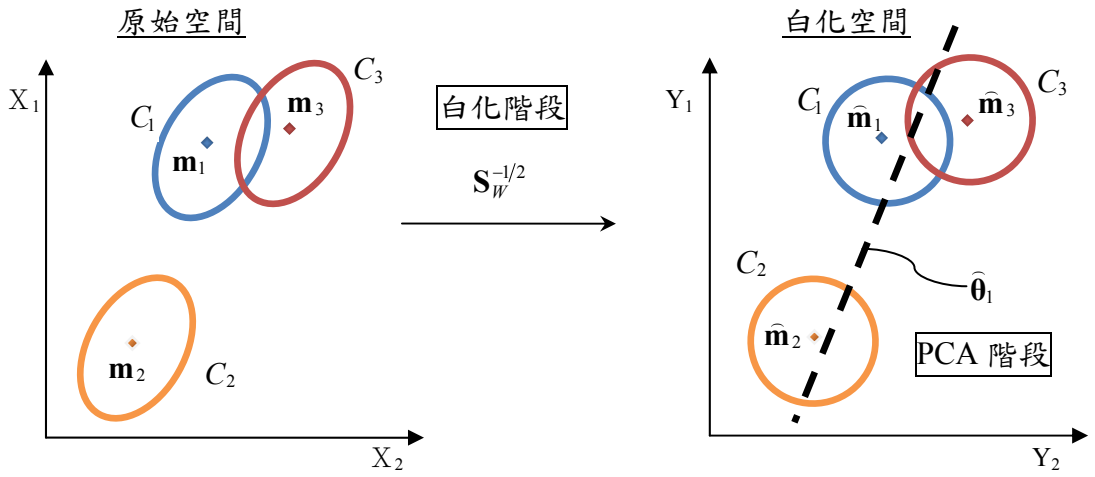


圖 2.7　線性鑑別分析的兩階段求解過程（幾何分析）

的算術平均$\mathbf{S}_W$[21]。

而在幾何分析方面，LDA 的求解可被視為一種兩階段的過程[58]，如圖 2.7。在第一階段，所有特徵向量會經過一次白化轉換(whitening transform)$\mathbf{S}_W^{-1/2}$，使得原來每一類別的分佈等值線圖(distribution contour)幾近於單位圓[22]，以便於我們合理的使用歐氏距離(Euclidean distance)作為類別間幾何分離度的量度標準。而在第二階段，藉著 PCA[23]作用在每一個被白化(whitened)的類別期望值向量$\mathbf{m}_i$，來決定一個使得所有類別期望值向量具有最大變異度(variation)的子空間（或投影方向），如圖 2.7 中的$\hat{\boldsymbol{\theta}}_1$。

因為對任何非奇異線性轉換矩陣，均不會影響 LDA 目標函式的求解，而$\mathbf{S}_W^{-1/2}$正好就是非奇異矩陣[24]，且由第二階段（PCA 階段）求得的投影方向$\hat{\boldsymbol{\Theta}}=[\hat{\boldsymbol{\theta}}_1,...,\hat{\boldsymbol{\theta}}_d]$均互為單範正交(orthonormal)，也就是$\hat{\boldsymbol{\Theta}}^T\hat{\boldsymbol{\Theta}}=\mathbf{I}_{(d\times d)}$，所以我們可將原始的 LDA

[21] 此同方差性假設亦可從貝氏決策理論以及最大相似度估測的角度來看，見 2.3.3 和 2.3.4 節。
[22] 此處亦有 LDA 的同方差性假設。
[23] PCA 是一種正交(orthogonal)線性轉換，能將原始資料轉換到一個新的座標系統，使得投影到第一座標軸（或第一主成分(the first principal component)）的資料具有最大的變異度，投影到第二座標軸的資料具有次大的變異度，以此類推。
[24] 理論上，任何共變異矩陣均為正半定(positive semi-definite)，但在實務上，由於我們的資料數遠大於特徵維度，所有共變異矩陣均可視為正定(positive definite)，也就當然為非奇異矩陣。

目標函式(2.20)轉為第一階段（白化階段）之後的目標函式：

$$\begin{aligned}\hat{J}_{\text{LDA\_TR}}(\hat{\boldsymbol{\Theta}}) &= \text{trace}\left((\hat{\boldsymbol{\Theta}}^T\mathbf{S}_W^{-1/2}\mathbf{S}_W\mathbf{S}_W^{-1/2}\hat{\boldsymbol{\Theta}})^{-1}(\hat{\boldsymbol{\Theta}}^T\mathbf{S}_W^{-1/2}\mathbf{S}_B\mathbf{S}_W^{-1/2}\hat{\boldsymbol{\Theta}})\right) \\ &= \text{trace}\left(\hat{\boldsymbol{\Theta}}^T\mathbf{S}_W^{-1/2}\mathbf{S}_B\mathbf{S}_W^{-1/2}\hat{\boldsymbol{\Theta}}\right)\end{aligned} \tag{2.28}$$

而最大化式(2.28)等同於求解本徵值問題：

$$(\mathbf{S}_W^{-1/2}\mathbf{S}_B\mathbf{S}_W^{-1/2})\hat{\boldsymbol{\theta}}_i = \hat{\lambda}_i\hat{\boldsymbol{\theta}}_i \tag{2.29}$$

其中，$\hat{\lambda}_i$ 為矩陣 $\mathbf{S}_W^{-1/2}\mathbf{S}_B\mathbf{S}_W^{-1/2}$ 之第 i 大的本徵值，而 $\hat{\boldsymbol{\theta}}_i$ 則為其對應的本徵向量。因此，$\hat{\boldsymbol{\Theta}}$ 最後可被表示為 $[\hat{\boldsymbol{\theta}}_1,\hat{\boldsymbol{\theta}}_2,...,\hat{\boldsymbol{\theta}}_d]$。欲求得作用於原始空間的LDA轉換矩陣 $\boldsymbol{\Theta}$，我們可在 $\hat{\boldsymbol{\Theta}}$ 之前乘上 $\mathbf{S}_W^{-1/2}$ 形成 $\mathbf{S}_W^{-1/2}\hat{\boldsymbol{\Theta}}$。以下的命題 2.3 說明了以上兩階段的幾何分析與傳統 LDA 的普遍化本徵值問題求解是等價的。

命題 2.3：由 LDA 幾何分析兩階段所組成的轉換矩陣 $\mathbf{S}_W^{-1/2}\hat{\boldsymbol{\Theta}}$ 亦能夠最大化原始的 LDA 目標函式，亦即，$\mathbf{S}_W^{-1/2}\hat{\boldsymbol{\Theta}}$ 為式(2.20)中 $\boldsymbol{\Theta}$ 的其中一種形式。

證明：由式(2.29)，可推導出：

$$\begin{aligned}&(\mathbf{S}_W^{-1/2}\mathbf{S}_B\mathbf{S}_W^{-1/2})\hat{\boldsymbol{\theta}}_i = \hat{\lambda}_i\hat{\boldsymbol{\theta}}_i \\ &\Rightarrow \mathbf{S}_W^{-1/2}(\mathbf{S}_W^{-1/2}\mathbf{S}_B\mathbf{S}_W^{-1/2})\hat{\boldsymbol{\theta}}_i = \hat{\lambda}_i\mathbf{S}_W^{-1/2}\hat{\boldsymbol{\theta}}_i \\ &\Rightarrow \mathbf{S}_W^{-1}\mathbf{S}_B(\mathbf{S}_W^{-1/2}\hat{\boldsymbol{\theta}}_i) = \hat{\lambda}_i(\mathbf{S}_W^{-1/2}\hat{\boldsymbol{\theta}}_i)\end{aligned} \tag{2.30}$$

因此，式(2.30)證明了 $\mathbf{S}_W^{-1/2}\hat{\boldsymbol{\Theta}}$ 亦為式(2.22)中 $\mathbf{S}_W^{-1/2}\mathbf{S}_B$ 的本徵向量矩陣，也就是說，$\mathbf{S}_W^{-1/2}\hat{\boldsymbol{\Theta}}$ 亦能最大化 LDA 的目標函式。 ■

經過LDA的兩階段求解過程，我們會發現在投影空間中，類別間散佈矩陣 $\mathbf{S}_B$ 與類別內散佈矩陣 $\mathbf{S}_W$ 都被對角化了（見以下的式(2.31)和式(2.32)），說明了 LDA

在類別內(within classes)和類別間(between classes)都具有特徵去相關的功能[25][59]。

$$\mathbf{S}_W \xrightarrow{\mathbf{S}_W^{-1/2}\widehat{\boldsymbol{\Theta}}} (\mathbf{S}_W^{-1/2}\widehat{\boldsymbol{\Theta}})^T \mathbf{S}_W (\mathbf{S}_W^{-1/2}\widehat{\boldsymbol{\Theta}}) = \mathbf{I}_{(d\times d)} \tag{2.31}$$

$$\mathbf{S}_B \xrightarrow{\mathbf{S}_W^{-1/2}\widehat{\boldsymbol{\Theta}}}$$
$$(\mathbf{S}_W^{-1/2}\widehat{\boldsymbol{\Theta}})^T \mathbf{S}_B (\mathbf{S}_W^{-1/2}\widehat{\boldsymbol{\Theta}}) = (\mathbf{S}_W^{-1/2}\widehat{\boldsymbol{\Theta}})^T \mathbf{S}_W (\mathbf{S}_W^{-1/2}\widehat{\boldsymbol{\Theta}})\boldsymbol{\Lambda} = \boldsymbol{\Lambda}_{(d\times d)} = \begin{bmatrix} \lambda_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \lambda_d \end{bmatrix} \tag{2.32}$$

此外，由式(2.31)更可以看出，LDA 的兩階段求解過程也同時解決了$\boldsymbol{\Theta}$的縮放比例問題，意即，式(2.20)可改寫為具有限制條件(constraint)的目標函式[26]：

$$J'_{\text{LDA\_TR}}(\boldsymbol{\Theta}) = \text{trace}(\boldsymbol{\Theta}^T \mathbf{S}_B \boldsymbol{\Theta}),\ \boldsymbol{\Theta}^T \mathbf{S}_W \boldsymbol{\Theta} = \mathbf{I}_{(d\times d)} \tag{2.33}$$

我們稱此限制條件$\boldsymbol{\Theta}^T \mathbf{S}_W \boldsymbol{\Theta} = \mathbf{I}_{(d\times d)}$為對於$\mathbf{S}_W$共軛正交(conjugate orthogonal)條件[59]，它使得$\boldsymbol{\Theta}$的縮放比例受到控制，這樣未來在與其它基於 LDA 之改進方法比較時會較為公平[27]。

此外，LDA 的兩階段幾何分析也提供我們至少有兩個方向可用來進一步地普遍化或改進 LDA。一是在第一階段（白化階段）中，我們可以試著找出比$\mathbf{S}_W^{-1/2}$更佳的白化轉換，或是比$\mathbf{S}_W$更有幫助的類別內散佈矩陣；另一是在第二階段（PCA 階段）中，我們可以重新調整類別間散佈矩陣$\mathbf{S}_B$，使得經過 PCA 處理的類別更具有鑑別性。

---

[25] 但這不能保證對於個別(individual)的類別母體，LDA 也能達到完全特徵去相關的效果，也就是$\boldsymbol{\Theta}^T \mathbf{S}_i \boldsymbol{\Theta} = \text{diag}(\boldsymbol{\Theta}^T \mathbf{S}_i \boldsymbol{\Theta})$未必成立。

[26] 式(2.34)也可以直接一維一維地以拉氏乘數(Lagrange multipliers)作最佳化，得到的轉換矩陣也會同樣滿足傳統的 LDA 目標函式。

[27] 許多研究者如 Sammon (1970)、Foley (1975)、Duchene (1988)試著求出在正交空間下的 LDA，但就筆者所知，目前沒有研究可以證明，正交特徵空間對於語音辨識有明顯幫助。

2.3.3 限制與改進：異方差性(Heteroscedasticity)

若我們單純地從 LDA 的形式（式(2.20)與(2.21)）或它所蘊涵的最大化類別間幾何分離度來看，會發現 LDA 對資料並無任何機率分佈的假設，而只具有對資料統計資訊的要求（期望值與變異度）[56, 60]。但若我們假設每一類別的資料都遵循高斯分佈(Gaussian distribution)，則 Campbell 以一種普遍化的線性模型證明了 LDA 轉換矩陣的求取，等同於將所有具同方差性的 n 維資料置於最大化相似度的框架下作參數估計，並『假設』所有類別鑑別性資訊僅存於所欲投影的 d 維的子空間，而剩下的 $n-d$ 維則不具任何鑑別性[61]。

為了除去此同方差性的假設，Kumar 等人提出了異方差線性鑑別分析(heteroscedastic linear discriminant analysis, HLDA)[28][26, 62]，同樣是在線性模型下最大化所有資料的相似度，只是模型參數中的類別共變異矩陣不再視為相同。HLDA 中每一類別母體的期望值 $\boldsymbol{\mu}_i$ 與共變異矩陣 $\boldsymbol{\Sigma}_i$，以及經由最大化相似度估計法求出的估計子(estimators)可被表示如下：

$$\boldsymbol{\mu}_i = \begin{bmatrix} \mu_{i,1} \\ \vdots \\ \mu_{i,d} \\ \mu_{0,1} \\ \vdots \\ \mu_{0,(n-d)} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\mu}_i^d \\ \boldsymbol{\mu}_0^{(n-d)} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Theta}_d^T \mathbf{m}_i \\ \boldsymbol{\Theta}_{(n-d)}^T \overline{\mathbf{m}} \end{bmatrix} \tag{2.34}$$

$$\boldsymbol{\Sigma}_i = \begin{bmatrix} \boldsymbol{\Sigma}_i^d & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_0^{(n-d)} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Theta}_d^T \mathbf{S}_i \boldsymbol{\Theta}_d & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Theta}_{(n-d)}^T \mathbf{S}_T \boldsymbol{\Theta}_{(n-d)} \end{bmatrix} \tag{2.35}$$

[28] Saon 等人認爲 Kumar 等人的構想非原創，而是源於 Schukat-Talamazzini 等人於 ICASSP'95 提出的論文，但筆者不以爲然。事實上，Schukat-Talamazzini 等人所做的是在最大相似度的框架下，基於模型空間的線性轉換，也就是將線性轉換置於聲學模型參數估測中同時進行。因此，兩者的處理結構是不同的。Kumar 也在他博士論文的附錄中說明了他的方法如何聯於 LDA，見[62]。本論文之後皆以 HLDA 來簡稱『異方差線性鑑別分析』。

其中，$\boldsymbol{\mu}_i^d$ 為投影後第 i 個類別平均向量的前 d 維，$\boldsymbol{\mu}_0^{(n-d)}$ 為投影後第 i 個類別平均向量的後 $n-d$ 維，同樣地，$\boldsymbol{\Sigma}_i^d$ 和 $\boldsymbol{\Sigma}_0^{(n-d)}$ 分別為前 d 維與後 $n-d$ 維的共變異矩陣；換言之，所有類別的 $\boldsymbol{\mu}_0^{(n-d)}$ 和 $\boldsymbol{\Sigma}_0^{(n-d)}$ 都是相同的。而 $\mathbf{S}_T$ 為整體散佈矩陣(total scatter matrix)，其定義如下：

$$\mathbf{S}_T = \frac{1}{N}\sum_{\mathbf{x}}(\mathbf{x}-\overline{\mathbf{m}})(\mathbf{x}-\overline{\mathbf{m}})^T \tag{2.36}$$

我們很容易證明 $\mathbf{S}_T$ 、$\mathbf{S}_B$ 和 $\mathbf{S}_W$ 三者的關係為（見附錄 7.2.1）

$$\mathbf{S}_T = \mathbf{S}_B + \mathbf{S}_W \tag{2.37}$$

HLDA 的全秩(full-rank)轉換矩陣參數 $\boldsymbol{\Theta} = [\boldsymbol{\Theta}_d, \boldsymbol{\Theta}_{(n-d)}]$ 同樣可藉著最大化相似度估計法求出，其目標函式與一階偏導數為[26]

$$J_{\mathrm{HLDA}}(\boldsymbol{\Theta}) = -\frac{N}{2}\log|\boldsymbol{\Theta}_{(n-d)}^T\mathbf{S}_T\boldsymbol{\Theta}_{(n-d)}| - \sum_{i=1}^{C}\frac{n_i}{2}\log|\boldsymbol{\Theta}_d^T\mathbf{S}_i\boldsymbol{\Theta}_d| + N\log|\boldsymbol{\Theta}| \tag{2.38}$$

$$\frac{\partial J_{\mathrm{HLDA}}(\boldsymbol{\Theta})}{\partial\boldsymbol{\Theta}} = \left[\sum_{i=1}^{C}\frac{n_i}{N}\mathbf{S}_i\boldsymbol{\Theta}_d(\boldsymbol{\Theta}_d^T\mathbf{S}_i\boldsymbol{\Theta}_d)^{-1} \quad \mathbf{S}_T\boldsymbol{\Theta}_{(n-d)}(\boldsymbol{\Theta}_{(n-d)}^T\mathbf{S}_T\boldsymbol{\Theta}_{(n-d)})^{-1}\right] - \boldsymbol{\Theta}^{-T} \tag{2.39}$$

式(2.39)中的 $N\log|\boldsymbol{\Theta}|$ 扮演了控制 $\boldsymbol{\Theta}$ 之縮放比例的角色，同時，$\boldsymbol{\Theta}$ 並沒有固定形式(close-form)，必須藉著梯度下降(gradient descent)等遞迴式(iterative)的最佳化技術來求解。

如 2.3.2 節所述，從類別間幾何分離度的角度來看，LDA 本身的形式已隱含了同方差性的假設[29]。Saon 等人據此提出了異方差鑑別分析(heteroscedastic discriminant analysis, HDA)[27]，試圖藉著考慮每一類別母體的共變異矩陣，變

[29] LDA 的第二種目標函式（式(2.21)）亦含有類別間幾何分離度上的同方差性假設，見[55]。

更式(2.21)中的分母部分，而得出新的目標函式[30]：

$$J_{\mathrm{HDA}}(\boldsymbol{\Theta}) = \frac{|\boldsymbol{\Theta}^T \mathbf{S}_B \boldsymbol{\Theta}|^N}{\prod_{i=1}^{C} |\boldsymbol{\Theta}^T \mathbf{S}_i \boldsymbol{\Theta}|^{n_i}} \tag{2.40}$$

我們也可將式(2.40)寫成對數(logarithm)形式的目標函數及其一階偏導數如下[26]：

$$\log J_{\mathrm{HDA}}(\boldsymbol{\Theta}) = \sum_{i=1}^{C} n_i \log |\boldsymbol{\Theta}^T \mathbf{S}_i \boldsymbol{\Theta}| + N \log |\boldsymbol{\Theta}^T \mathbf{S}_B \boldsymbol{\Theta}| \tag{2.41}$$

$$\frac{\partial \log J_{\mathrm{HDA}}(\boldsymbol{\Theta})}{\partial \boldsymbol{\Theta}} = \sum_{i=1}^{C} -2n_i \mathbf{S}_i \boldsymbol{\Theta} (\boldsymbol{\Theta}^T \mathbf{S}_i \boldsymbol{\Theta})^{-1} + 2N \mathbf{S}_B \boldsymbol{\Theta} (\boldsymbol{\Theta}^T \mathbf{S}_B \boldsymbol{\Theta})^{-1} \tag{2.42}$$

與 HLDA 一樣，$\partial \log J_{\mathrm{HDA}}(\boldsymbol{\Theta}) / \partial \boldsymbol{\Theta} = 0$ 也沒有固定解，必須藉著梯度下降等遞迴式的最佳化技術來求解。此外，很明顯地，$\boldsymbol{\Theta}$ 於式(2.41)的最佳化中並不具有唯一解。因此，為了控制 $\boldsymbol{\Theta}$ 的縮放比例，我們可將式(2.41)進一步改寫為：

$$J'_{\mathrm{HDA}}(\boldsymbol{\Theta}) = \sum_{i=1}^{C} -n_i \, |\boldsymbol{\Theta}^T \mathbf{S}_i \boldsymbol{\Theta}|, \; |\boldsymbol{\Theta}^T \mathbf{S}_B \boldsymbol{\Theta}| = K \tag{2.43}$$

其中，K 為任意常數，其用意在於控制 $\boldsymbol{\Theta}$ 的縮放比例。由式(2.43)可看出，HDA 也具有在最大化相似度框架下的另一層意義：若每一群具有獨立期望值與獨立共變異矩陣的類別母體均呈高斯分佈，則最佳化 HDA 的目標函式，等同於在某種條件限制（也就是 $|\boldsymbol{\Theta}^T \mathbf{S}_B \boldsymbol{\Theta}|$ 為某定值）下，最大化所有資料在投影空間的相似度。

最近幾年，HDA 這種源於類別內散佈矩陣之變形的方法引起了 Sakai 等人

[30] 我們無法單單從式(2.40)來判斷 HDA 是否打破了 LDA 的同方差性假設，因爲之後 Sakai 等人證明了，HDA 所做的只是對類別內共變異矩陣 $\mathbf{S}_W$ 作重新估計(re-estimate)而已，見[63]。筆者認爲，嘗試從形式上或類別間幾何分離度來打破同方差性假設是很困難的，儘管 HDA 在發明之初似乎有此動機。不過，若賦予 HDA 在最大相似度估測法下的意義，則 HDA 的目標是成功的。本論文之後皆以 HDA 來簡稱『異方差鑑別分析』。

的注意。他們發現，LDA 中的類別內散佈矩陣$\mathbf{S}_W$（式(2.21)中的分母）可被視為所有類別共變異矩陣的算術平均(arithmetic average)，而 HDA 中的改良式類別內散佈矩陣（式(2.40)中的分母）則可被視為所有類別共變異矩陣的幾何平均(geometrical average)。基於這種『巧合』，他們提出了基於乘冪平均(power mean)的線性鑑別分析(power linear discriminant analysis, PLDA)[31][63-64]，其目標函式如下：

$$J_{\mathrm{PLDA}}(\mathbf{\Theta}, m) = \frac{|\mathbf{\Theta}^T \mathbf{S}_B \mathbf{\Theta}|}{\left| \left(\sum_{i=1}^{C} p_i (\mathbf{\Theta}^T \mathbf{S}_i \mathbf{\Theta})^m \right) \right|^{1/m}} \tag{2.44}$$

其中，m 為乘冪平均中可自由設定的參數，當$m=1$時，PLDA 將還原成 LDA；當$m=0$時，PLDA 將還原成 HDA。雖然 PLDA 在一定程度下普遍化了類別內散佈矩陣，但筆者認為它仍有一些缺點：第一，即使乘冪平均又可稱為普遍化平均(generalized mean)，意即它可以普遍化所有類型的平均，但其中最佳之 m 值卻甚難求取。第二，式(2.44)在非整數的 m 值設定下，並沒有其一階偏導數或二階偏導數的固定形式，這導致了最佳化技術的困難[32]，求出的值也只能達到相對最佳值(local optimum)。Sakai 等人只提供了式(2.45)在 m 為整數時的一階偏導數如下：

$$\frac{\partial J_{\mathrm{PLDA}}(\mathbf{\Theta}, m)}{\partial \mathbf{\Theta}} = 2\mathbf{S}_B \mathbf{\Theta} \widetilde{\mathbf{S}}_B^{-1} - 2D_m \tag{2.45}$$

[31] 本論文之後皆以 PLDA 來簡稱『基於乘冪平均的線性鑑別分析』。

[32] 即便使用單形法(simplex method)（如在 MATLAB 裡面的 fminsearch 函數），無須提供目標函式的一階偏導數，在維度過大（或變數過多）的情形還是難以處理。

$$D_m = \begin{cases} \dfrac{1}{m}\sum_{i=1}^{C} p_i \mathbf{S}_i \boldsymbol{\Theta} \sum_{j=1}^{m} X_{m,j,k}, & m > 0 \\ \sum_{i=1}^{C} p_i \mathbf{S}_i \boldsymbol{\Theta} \widetilde{\mathbf{S}}_i^{-1}, & m = 0 \\ -\dfrac{1}{m}\sum_{i=1}^{C} p_i \mathbf{S}_i \boldsymbol{\Theta} \sum_{j=1}^{|m|} Y_{m,j,i}, & \text{otherwise} \end{cases} \tag{2.46}$$

$$X_{m,j,k} = \widetilde{\mathbf{S}}_i^{m-j}\left(\sum_{l=1}^{C} p_l \widetilde{\mathbf{S}}_l^{m}\right)^{-1} \widetilde{\mathbf{S}}_i^{j-1},\ Y_{m,j,i} = \widetilde{\mathbf{S}}_i^{m+j-1}\left(\sum_{l=1}^{c} P_l \widetilde{\mathbf{S}}_l^{m}\right)^{-1} \widetilde{\mathbf{S}}_i^{-j} \tag{2.47}$$

其中，變數上的波浪號『～』表示已經過$\boldsymbol{\Theta}$的轉換。值得一提的是，Sakai 等人亦試著將 HLDA 還原成 PLDA，他們認為，LDA 目標函式中的$\mathbf{S}_B$與$\mathbf{S}_T$既然可以互通(也就是將式(2.21)中的$\mathbf{S}_B$更換為$\mathbf{S}_T$，並不會改變$\boldsymbol{\Theta}$)，那麼以$\mathbf{S}_B$取代 HLDA 目標函式中的$\mathbf{S}_T$勢必不會影響 HLDA 轉換矩陣[64]。但事實則不然。我們若將式(2.38)中$\mathbf{S}_T$的更換為$\mathbf{S}_B$，一來 HLDA 將喪失原來的物理意義，二來更換前後的兩種目標函式並不能保證具有相同的$\boldsymbol{\Theta}$解。真正對於 HLDA 的普遍化方法，將在本論文第 4 章提出。

2.3.4 限制與改進：分類相關性

由 2.3.3 節中 LDA 的幾何分析，我們可以看出在第二階段（PCA 階段）所得到的主軸，會由原本距離較大的類別配對所主導(dominate)。以類別配對(C_i, C_j)來說，若它們的馬氏距離平方$D_{ij}^2 = (\mathbf{m}_i - \mathbf{m}_j)^T \mathbf{S}_W^{-1}(\mathbf{m}_i - \mathbf{m}_j)$愈大，則在白化轉換後，$\mathbf{S}_W^{-1/2}(\mathbf{m}_i - \mathbf{m}_j)$的方向會對$\mathbf{S}_W^{-1/2}\mathbf{S}_B\mathbf{S}_W^{-1/2}$最大的本徵向量（也就是對應到最大的本徵值）產生愈大的貢獻。因此，LDA 會使原本距離較大的類別配對，在轉換後的子空間中，相對形成更大的距離優勢。如圖 2.7 中，類別C_2與類別C_3在白化空間中的距離最大，但經過 PCA 投影後，類別C_2與類別C_3還是維持相對大的距離，但在分類的實務上，原本距離較近的類別，如類別C_1與類別C_3，在

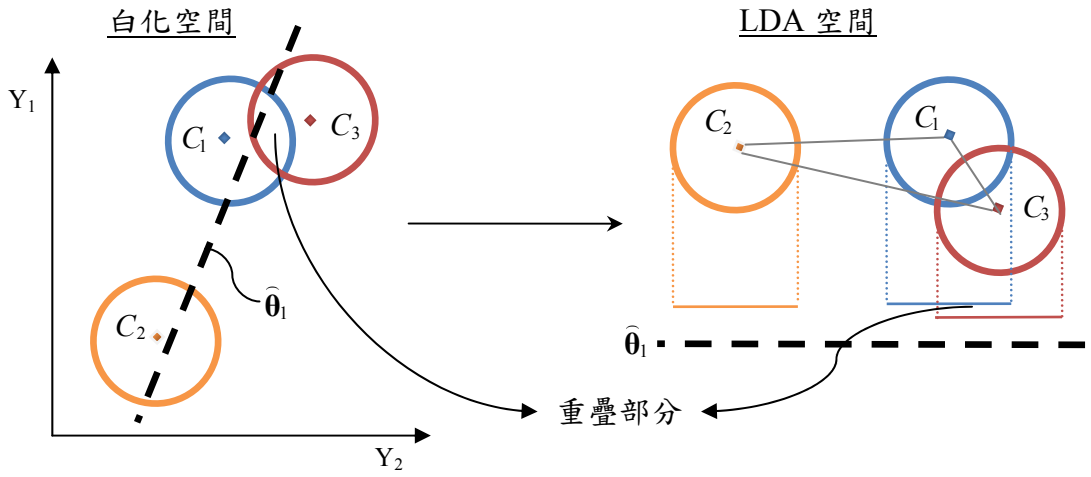


圖 2.8　LDA 之過度強調問題示意圖

分佈上具有較大的重疊，才是造成混淆或分類正確率降低的主要來源。這種相對上使遠者更遠，近者更近，造成分類困難的情形，我們稱之為 LDA 的過度強調問題(over-emphasis problem)[33]。除此之外，這也突顯了 LDA 在圖樣識別上的先天限制：它的目標函式似乎並不關聯於分類器的分類規則(allocation rules)或分類正確率，換句話說，就算線性鑑別分析能夠造成類別之間較高的幾何分離度，但也不能保證後端辨識器會產生較低的分類錯誤率。對於過度強調問題，大部分研究是以基於權重的線性鑑別分析(weighting-based LDA, WLDA)來解決，也就是對距離較大的類別配對加上較小的權重來去強調(de-emphasis)。權重大小的決定將在第 3 章討論之。

雖然 LDA 不利於分類，但這並不是說它與分類完全無關。事實上，LDA 本來就不是為了特定分類器所設計，但我們的確可以命題 2.4 來證明在 LDA 對於任何兩個高斯分佈、共變異矩陣均相同的類別母體，能夠產生在分類上最佳的投影子空間。

命題 2.4：假設有兩類資料，其母體為高斯分佈且具有相同的共變異矩陣 $\mathbf{S}_W$，則 LDA 能夠在貝氏錯誤(Bayes error)最小的意義下，決定出最佳的投影子空間。

[33] 筆者稱它作 LDA 的馬太效應(Matthew Effect)：『因爲凡有的，還要給他，他就充盈有餘；沒有的，連他所有的，也要從他奪去。』《新約聖經・馬太福音》第 25 章第 29 節。

證明：在兩類問題(two-class problem)中，貝氏錯誤ε可定義為[49]：

$$\varepsilon = \int \min\left(p_1 f_1(\mathbf{x}), p_2 f_2(\mathbf{x})\right) d\mathbf{x} \tag{2.48}$$

其中，p_1與p_2分別為類別C_1和類別C_2的事前機率，$f_1(\mathbf{x})$和$f_2(\mathbf{x})$則分別為類別C_1和類別C_2產生資料$\mathbf{x}$的機率密度函數(probability density function, PDF)。由於ε難以處理，我們僅能求其上界(upper bound)ε_{upper}：

$$\varepsilon_{upper} = p_1^s p_2^{1-s} \int f_1^s(\mathbf{x}) f_2^{1-s}(\mathbf{x}) d\mathbf{x} \tag{2.49}$$

現在，我們的目標在於求取一個線性轉換$\boldsymbol{\theta} \in \Re^{n\times 1}$，使得資料落在投影子空間後，具有最小的貝氏錯誤上界ε_{upper}。倘若這兩個類別的機率密度函數皆遵循高斯分佈，也就是$\mathbf{x} \sim N(\mathbf{m}_1, \mathbf{S}_1)$和$\mathbf{x} \sim N(\mathbf{m}_2, \mathbf{S}_2)$，則在投影子空間內，$\varepsilon_{upper}$[34]具有固定的表達式（$\tilde{\mathbf{x}} = \boldsymbol{\theta}^T \mathbf{x};\ \tilde{\mathbf{m}}_i = \boldsymbol{\theta}^T \mathbf{m}_i;\ \tilde{\mathbf{S}}_i = \boldsymbol{\theta}^T \mathbf{S}_i \boldsymbol{\theta};\ i = 1,2$）：

$$\varepsilon_{Chernoff} = p_1^s p_2^{1-s} \int f_1^s(\tilde{\mathbf{x}}) f_2^{1-s}(\tilde{\mathbf{x}}) d\mathbf{x} = p_1^s p_2^{1-s} e^{-\tilde{\mu}(s)},\ 0 \le s \le 1 \tag{2.50}$$

$$\begin{aligned} \tilde{\mu}(s) = {} & \frac{s(1-s)}{2} (\tilde{\mathbf{m}}_1 - \tilde{\mathbf{m}}_2)^T \left(s\tilde{\mathbf{S}}_1 + (1-s)\tilde{\mathbf{S}}_2\right)^{-1} (\tilde{\mathbf{m}}_1 - \tilde{\mathbf{m}}_2) \\ & + \frac{1}{2} \ln \frac{|\, s\tilde{\mathbf{S}}_1 + (1-s)\tilde{\mathbf{S}}_2 \,|}{|\, \tilde{\mathbf{S}}_1 \,|^s |\, \tilde{\mathbf{S}}_2 \,|^{1-s}} \end{aligned} \tag{2.51}$$

因為類別C_1和類別C_2均具有相同的共變異矩陣$\mathbf{S}_W$，也就是$\mathbf{S}_1 = \mathbf{S}_2 = \mathbf{S}_W$，所以式(2.51)可進一步化簡為：

$$\tilde{\mu}(s) = \frac{s(1-s)}{2} (\tilde{\mathbf{m}}_1 - \tilde{\mathbf{m}}_2)^T \tilde{\mathbf{S}}_W^{-1} (\tilde{\mathbf{m}}_1 - \tilde{\mathbf{m}}_2) \propto J_{LDA\_TR}(\boldsymbol{\Theta}) \tag{2.52}$$

其中，$\tilde{\mathbf{S}}_W = \boldsymbol{\Theta}^T \mathbf{S}_W \boldsymbol{\Theta}$。由式(2.52)知，最大化LDA目標函式，等同於最小化式(2.50)，

[34] 此時的ε_{upper}被稱做薛氏上界(Chernoff bound)。

也就達到了最小化貝氏分類錯誤。 ■

至於多類問題(multi-class problem)，若所有資料類別為高斯分佈且具同方差性，則 LDA 能產生最佳的$C-1$維子空間（C 為類別總數），使得在貝氏決策法則下，具有最小的貝氏錯誤[52]。不同於命題 2.4 以及 Rao 所提出的方式[65]，以下我們試著以較簡單的方式來證明：

命題 2.5：假設有多類資料，其母體為高斯分佈且具有相同的共變異矩陣$\mathbf{S}_W$，則 LDA 能夠在貝氏錯誤(Bayes error)最小的條件下，決定出最佳的投影子空間，但此子空間的維度必須要是$C-1$。

證明：在分類時，若我們要達到最小貝氏錯誤，則需滿足以下規則[66]：

將任一資料 $\mathbf{x}$ 分類至類別C_k，若

$$\ln\left(p_k f_k(\mathbf{x})\right) > \ln\left(p_i f_i(\mathbf{x})\right),\ \forall i \neq k \tag{2.53}$$

其中，我們將$\ln\left(p_i f_i(\mathbf{x})\right)$稱為類別$C_i$之鑑別分數(discriminant score)[65]，對於同方差性的高斯分佈，類別C_i之鑑別分數可寫成（見附錄 7.2.2）

$$\begin{aligned} d_i(\mathbf{x}) &= \ln\left(p_i f_i(\mathbf{x})\right) \\ &= -\frac{1}{2}\log|\mathbf{S}_W| - \frac{1}{2}(\mathbf{x}-\mathbf{m}_i)^T\mathbf{S}_W^{-1}(\mathbf{x}-\mathbf{m}_i) + \log p_i,\ i=1,...,C \end{aligned} \tag{2.54}$$

由於每一類別母體具有相同的共變異矩陣$\mathbf{S}_W$，因此$(-1/2)\log|\mathbf{S}_W| + (-1/2)\mathbf{x}^T\mathbf{S}_W^{-1}\mathbf{x}$項並不影響分類，我們可將式(2.54)展開並減去此項，得到新類別C_i之鑑別分數：

$$
\begin{aligned}
d_i(\mathbf{x}) &= (\mathbf{m}_i^T \mathbf{S}_W^{-1})\mathbf{x} \underbrace{-\frac{1}{2}\mathbf{m}_i^T \mathbf{S}_W^{-1}\mathbf{m}_i + \log p_i}_{\text{constant } \gamma_i},\ i = 1,...,C \\
&= \underbrace{(\mathbf{m}_i^T \mathbf{S}_W^{-1})}_{\text{decision boundary}}\ \mathbf{x} + \gamma_i,\ i = 1,...,C
\end{aligned}
\tag{2.55}
$$

式(2.55)說明了最佳的決策邊界是線性的，且分類規則為：

將任一資料 $\mathbf{x}$ 分至類別 C_k ，若

$$
\begin{aligned}
d_k(\mathbf{x}) - d_i(\mathbf{x}) &= (\mathbf{m}_k^T \mathbf{S}_W^{-1})\mathbf{x} + \gamma_k - (\mathbf{m}_i^T \mathbf{S}_W^{-1})\mathbf{x} - \gamma_i \\
&= \underbrace{\left((\mathbf{m}_k - \mathbf{m}_i)^T \mathbf{S}_W^{-1}\right)}_{\text{decision boundary}}\mathbf{x} + \underbrace{\gamma_k - \gamma_i}_{\text{constant}} > 0,\ \forall i \neq k
\end{aligned}
\tag{2.56}
$$

由式(2.56)可看出，所有決策邊界 $(\mathbf{m}_k - \mathbf{m}_i)^T \mathbf{S}_W^{-1},\ i \neq k$ 能夠展開一個 $C-1$ 維的子空間 $\Re^{(C-1)\times 1}$ ，所有鑑別分數均能夠無損地藉由這個子空間求得，而這個子空間剛好可以用 LDA 目標函式求出。

現在，讓我們思考兩種情況：一、若 LDA 所決定之子空間的維度小於 $C-1$ ，也就是 $d < C-1$ ，則顯然每一類別之鑑別分數（式(2.56)）勢必會有所損失。二、根據命題 2.1， $\mathbf{S}_W^{-1}\mathbf{S}_B$ 之本徵值為零的本徵向量 $\mathbf{g}_l,\ \forall l = 1,...,(n-C+1)$ 會與任一向量 $(\mathbf{m}_i - \overline{\mathbf{m}})$ 正交，也會正交於向量 $(\mathbf{m}_k - \overline{\mathbf{m}}) - (\mathbf{m}_i - \overline{\mathbf{m}}) = \mathbf{m}_k - \mathbf{m}_i,\ \forall i,k = 1,...,C$ 。因此，這會使得

$$
\mathbf{g}_l^T(\mathbf{m}_k - \mathbf{m}_i) = 0,\ \forall i,k = 1,...,C,\ \forall l = 1,...,(n-C+1) \tag{2.57}
$$

意即，在 $\mathbf{g}_l,\ \forall l = 1,...,(n-C+1)$ 所生成的空間中，不具有任何鑑別性資訊。 ■

而在 $d < C-1$ 的情形下（d 為子空間維度），Geisser 亦證明了這些類別若同時亦具有相同事前機率，其投影至 1 維空間的貝氏錯誤可表示為[67]：

$$2C^{-1}\sum_{i=1}^{C-1}\Phi\left(\frac{\eta_i-\eta_{i+1}}{2}\right) \tag{2.58}$$

$$\eta_i=\frac{\boldsymbol{\theta}^T\mathbf{m}_i}{\sqrt{\boldsymbol{\theta}^T\mathbf{S}_W\boldsymbol{\theta}}},\ \eta_1\le\eta_2\le\cdots\le\eta_C \tag{2.59}$$

其中，$\Phi(\cdot)$ 為變異度(variance)為 1 之標準高斯分佈(standard Gaussian distribution)的累積分佈函數(cumulative distribution function, CDF)，η_i 為已被白化之類別 C_i 期望值向量，$\boldsymbol{\theta}\in\Re^{n\times 1}$ 為我們所要求的投影方向。此時，LDA 目標函式為

$$\frac{1}{2}\sum_{i=1}^{C}\sum_{j=1}^{C}(\eta_i-\eta_j)^2 \tag{2.60}$$

很明顯地，能夠最大化式(2.60)的 1 維子空間並不會使式(2.58)最小化。也就是說，在多類問題中，當轉換後維度小於 $C-1$ 時，LDA 無法決定使貝氏錯誤最小的最佳子空間[35]，我們可將此問題稱為正確率無關問題(accuracy-irrelation problem)。之後，在第 3 章中，我們將會提出許多近似方法，能夠有效且迅速的將 LDA 關聯至分類正確率。

[35] Geisser 並沒有對式(2.44)提出解決方式。之後，Schervish (1984) 提出，若原始空間爲 $\Re^2$，則三類問題(three-class problem)的貝氏錯誤可被定義爲一個凸函數(convex function)，也就具有全域最佳解(global optimum)。不過，若原始空間維度大小超過 $\Re^2$ 或類別數大於 3，目前並無一般解。最近，Hamsici 等人 (2008) 提出另一個能夠找出任一維度之最佳子空間的方法，並將之延伸至異方差分佈(heteroscedastic distributions)的情形。不過對於類別數過多的問題，例如 $C>100$，他們的方法就顯得非常不實用了。

第 3 章　基於經驗資訊之線性鑑別分析

在本章中，我們將先提出基於權重之線性鑑別分析以及相關的技術，它們的主要好處在於承繼了 LDA 輕省的可解性，又同時解決了第 2 章所提到的過度強調問題。並且，為了充分將 LDA 與後端辨識器緊密結合，我們更提出了基於經驗混淆資訊之線性鑑別分析，以期能夠解決傳統 LDA 推導時不能納入之分類正確率資訊的問題。

3.1　權重式線性鑑別分析

為了要解決 LDA 的過度強調問題(over-emphasis problem)，一個可能的簡單做法是在原始的 LDA 目標函式，針對每一類別配對的並向量積 $(\mathbf{m}_i-\mathbf{m}_j)(\mathbf{m}_i-\mathbf{m}_j)^T$ 加上適當的權重值(weight)，使得原本距離較大的類別配對不會被過度強調，而距離較小的類別配對也不會被忽略。於是，在白化空間中的類別間共變異矩陣可表示為：

$$\begin{aligned}\widehat{\mathbf{S}}_B &= \frac{1}{2}\sum_{i=1}^{C}\sum_{j=1}^{C} p_i p_j \,\mathrm{w}(i,j)(\widehat{\mathbf{m}}_i-\widehat{\mathbf{m}}_j)(\widehat{\mathbf{m}}_i-\widehat{\mathbf{m}}_j)^T \\ &= \frac{1}{2}\sum_{i=1}^{C}\sum_{j=1}^{C} p_i p_j \,\mathrm{w}(i,j)\widehat{\mathbf{S}}_{ij}\end{aligned} \tag{3.1}$$

其中，$\widehat{\mathbf{m}}_i=\mathbf{S}_W^{-1/2}\mathbf{m}_i$，$\widehat{\mathbf{S}}_{ij}=\mathbf{S}_W^{-1/2}\mathbf{S}_{ij}\mathbf{S}_W^{-1/2}$，$\mathrm{w}(i,j)$ 為類別 C_i 與類別 C_j 之間的權重因子(weighting factor)，乃用來控制它們對於投影方向的貢獻。之後，這種權重式線性鑑別分析(weighting-based LDA, WLDA)[36]的目標函式可被定義為：

[36] 本論文之後皆以 WLDA 來簡稱『權重式線性鑑別分析』。

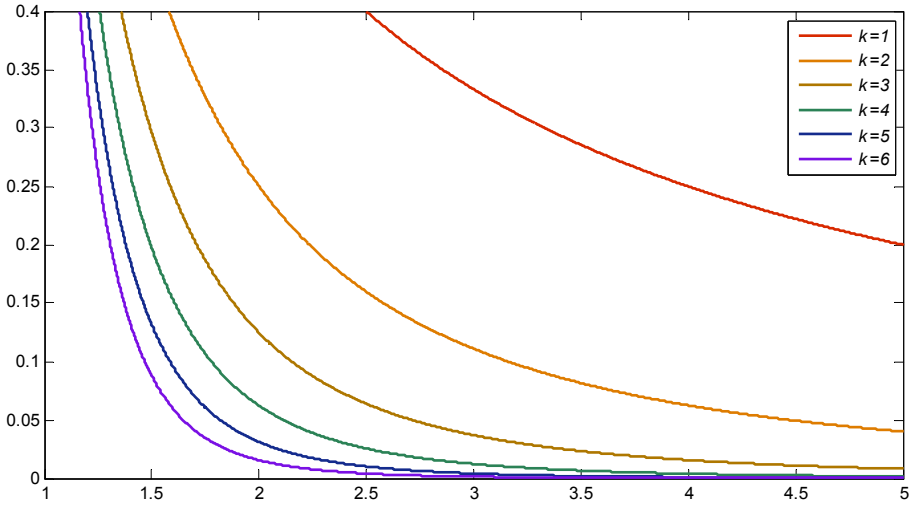


圖 3.1　PWLDA 之距離與權重關係圖

（橫軸為 Δ_{ij}，縱軸為 $\mathrm{w}_{\mathrm{PWLDA}}(i,j)$）

$$J_{\mathrm{WLDA}}(\widehat{\boldsymbol{\Theta}}) = \frac{1}{2}\sum_{i=1}^{C}\sum_{j=1}^{C} p_i p_j \,\mathrm{w}(i,j)\,\mathrm{trace}(\widehat{\boldsymbol{\Theta}}^T \widehat{\mathbf{S}}_{ij} \widehat{\boldsymbol{\Theta}}) \tag{3.2}$$

很明顯地，若 $\mathrm{w}(i,j)$ 獨立於任何非奇異矩陣的線性轉換，例如 $\mathbf{S}_W^{-1/2}$，則類似於 LDA，WLDA 的轉換矩陣 $\boldsymbol{\Theta}$ 也可簡單地表示為 $\mathbf{S}_W^{-1/2}\widehat{\boldsymbol{\Theta}}$，其中 $\widehat{\boldsymbol{\Theta}}$ 為矩陣 $\widehat{\mathbf{S}}_B$ 之本徵值前 d 大之本徵向量 $[\widehat{\boldsymbol{\theta}}_1,...,\widehat{\boldsymbol{\theta}}_d]$ 所組成的矩陣。

在許多關於決定權重因子的研究中，許多研究者直接用類別間之馬氏距離的乘冪作為形成權重的依據，在此我們稱之為基於乘冪之權重式線性鑑別分析 (power WLDA, PWLDA)[37][68-70]，其 $\mathrm{w}(i,j)$ 可普遍化如下：

$$\mathrm{w}_{\mathrm{PWLDA}}(i,j) = \Delta_{ij}^{-k},\ k>0 \tag{3.3}$$

其中，k 為可自由設定的調節常數，Δ_{ij} 為類別 C_i 與類別 C_j 之間的馬氏距離，定義為：

$$\Delta_{ij} = \sqrt{(\mathbf{m}_i - \mathbf{m}_j)^T S_W^{-1/2}(\mathbf{m}_i - \mathbf{m}_j)} \tag{3.4}$$

[37] 本論文之後皆以 PWLDA 來簡稱『基於乘冪之權重式線性鑑別分析』。

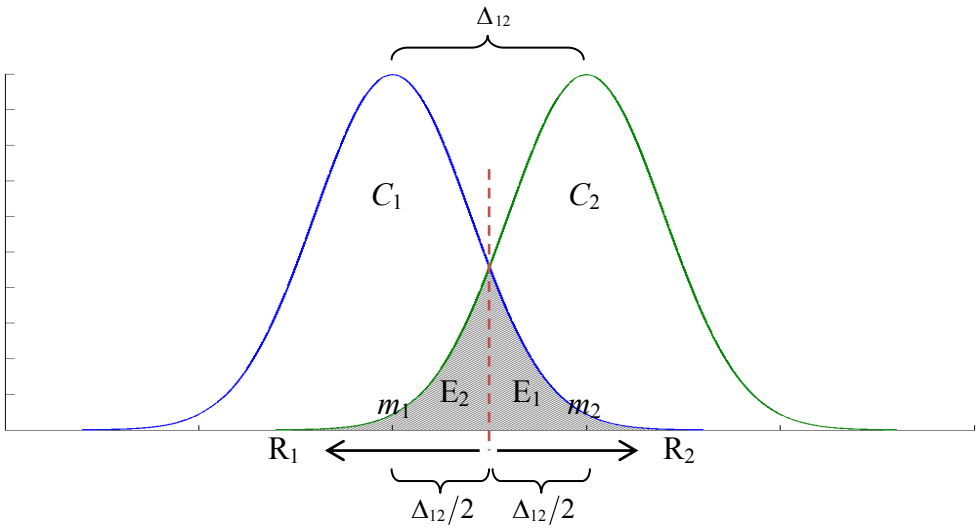


圖 3.2　兩個單變量高斯分佈及其貝氏錯誤示意圖

由圖(3.1)可看出，在各種 k 值的設定下，馬氏距離愈大的類別配對，所配置的權重愈低，反之，馬氏距離愈小的類別配對，所配置的權重愈高。這種治標的方式似乎可解決 LDA 的過度強調問題，但卻只能憑經驗地(empirically)設定 k 以降低距離大的類別配對所造成的影響力，並不能連於分類實務本身。

因此，Loog 等人考慮任兩類別母體 C_i 與 C_j，二者皆為同方差性的高斯分佈，經過白化過程後，當投影到方向 $\hat{\boldsymbol{\theta}} = (\hat{\mathbf{m}}_i - \hat{\mathbf{m}}_j)/\|\hat{\mathbf{m}}_i - \hat{\mathbf{m}}_j\|$ 時，其貝氏準確率(Bayes accuracy)為：

$$A_{ij}(\hat{\boldsymbol{\theta}}) = \frac{1}{2} + \frac{1}{2}\mathrm{erf}\left(\frac{\Delta_{ij}}{2\sqrt{2}}\right) \tag{3.5}$$

其證明如以下之命題 3.1：

命題 3.1：任兩白化後的類別母體 C_i 與 C_j，二者皆為同方差性的高斯分佈，當投影到方向 $\hat{\boldsymbol{\theta}} = (\hat{\mathbf{m}}_i - \hat{\mathbf{m}}_j)/\|\hat{\mathbf{m}}_i - \hat{\mathbf{m}}_j\|$（$\hat{\boldsymbol{\theta}}$ 的長度已被正規化為 1）時，其貝氏準確率為 $A_{ij}(\hat{\boldsymbol{\theta}}) = 1/2 + (1/2)\,\mathrm{erf}(\Delta_{ij}/2\sqrt{2})$。

證明：不失一般性，令兩類別母體為C_1與C_2，白化後投影至 1 維空間後，顯然均成為單變量高斯分佈，且變異度均為 1，即資料$X \sim N_1(m_1,1)$且$X \sim N_2(m_2,1)$。且當投影方向$\hat{\mathbf{\theta}}$為$(\hat{\mathbf{m}}_i - \hat{\mathbf{m}}_j)/\left\|\hat{\mathbf{m}}_i - \hat{\mathbf{m}}_j\right\|$，也就是兩類別母體期望值向量之連線方向時，會有最小的貝氏錯誤，即圖 3.2 中的灰色重疊部分。在圖 3.2 中，紅色虛線為最佳的決策邊界[49]，恰為兩類別母體期望值向量之中線，也就是說，若資料x落於區間R_1，則分作類別C_1，反之落於區間R_2，則分作類別C_2。在灰色重疊部分中，E_1表示資料原本屬於類別C_1，卻分到類別C_2的錯誤率，而E_2表示資料原本屬於類別C_2，卻分到類別C_1的錯誤率。

因此，為了方便，若我們將$N_1(m_1,1)$正規化(normalize)為$N(0,1)$，決策邊界的座標位置為a，則對於原本屬於類別C_1的資料x來說，其分類正確率可表示如下（對於類別C_2亦同）：

$$\begin{aligned} A_{12}(\hat{\mathbf{\theta}}) &= \int_{-\infty}^{a} N(x,0,1)dx = \int_{-\infty}^{0} N(x,0,1)dx + \int_{0}^{a} N(x,0,1)dx \\ &= \frac{1}{2} + \frac{1}{\sqrt{2\pi}}\int_{0}^{a} e^{-\frac{x^2}{2}}dx \\ &= \frac{1}{2} + \frac{1}{\sqrt{2\pi}}\int_{0}^{\frac{t}{\sqrt{2}}} \sqrt{2}e^{-t^2}dt = \frac{1}{2} + \frac{1}{\sqrt{\pi}}\int_{0}^{\frac{t}{\sqrt{2}}} e^{-t^2}dt \\ &= \frac{1}{2} + \frac{1}{\sqrt{\pi}} \times \frac{\sqrt{\pi}}{2}\mathrm{erf}\left(\frac{t}{\sqrt{2}}\right) = \frac{1}{2} + \frac{1}{2}\mathrm{erf}\left(\frac{t}{\sqrt{2}}\right) \end{aligned} \tag{3.6}$$

其中，erf(·)為錯誤函數(error function)，被定義為

$$\mathrm{erf}(z) = \frac{2}{\sqrt{\pi}}\int_{0}^{z} e^{-t^2}dt \tag{3.7}$$

將$t = \Delta_{12}/2$，也就是正規化後之決策邊界的座標，代入式(3.7)，可得：

$$A_{12}(\hat{\mathbf{\theta}}) = \frac{1}{2} + \frac{1}{2}\mathrm{erf}\left(\frac{\Delta_{12}}{2\sqrt{2}}\right) \tag{3.8}$$

式(3.8)即為所求之貝氏準確率。 ■

根據式(3.5)，Loog 等人找出了馬氏距離與貝氏正確率的關係。而回到 WLDA，對於任兩類別母體 C_i 與 C_j，他們提出了如下的 $\mathrm{w}(i,j)$，並稱這種方法為近似成對理論正確標準(approximate pairwise theoretical accuracy criterion, aPTAC)[38]：

$$\mathrm{w}_{\mathrm{aPTAC}}(i,j)=\frac{1}{2\Delta_{ij}^2}\mathrm{erf}\left(\frac{\Delta_{ij}}{2\sqrt{2}}\right) \tag{3.9}$$

若所有資料只被分作兩類，在已知最佳投影方向為 $\hat{\boldsymbol{\theta}}=(\hat{\mathbf{m}}_i-\hat{\mathbf{m}}_j)/\left\|\hat{\mathbf{m}}_i-\hat{\mathbf{m}}_j\right\|$ 時，aPTAC 目標函式可寫成[39]：

$$\begin{aligned}
&J_{\mathrm{aPTAC}}(\hat{\boldsymbol{\theta}})=p_i p_j\,\mathrm{w}_{\mathrm{aPTAC}}(i,j)\,\mathrm{trace}(\hat{\boldsymbol{\theta}}^T\hat{\mathbf{S}}_{ij}\hat{\boldsymbol{\theta}})\\
&=\frac{1}{2\Delta_{ij}^2}\mathrm{erf}\left(\frac{\Delta_{ij}}{2\sqrt{2}}\right)\mathrm{trace}\left(\frac{(\hat{\mathbf{m}}_i-\hat{\mathbf{m}}_j)^T}{\left\|\hat{\mathbf{m}}_i-\hat{\mathbf{m}}_j\right\|}(\hat{\mathbf{m}}_i-\hat{\mathbf{m}}_j)(\hat{\mathbf{m}}_i-\hat{\mathbf{m}}_j)^T\frac{(\hat{\mathbf{m}}_i-\hat{\mathbf{m}}_j)^T}{\left\|\hat{\mathbf{m}}_i-\hat{\mathbf{m}}_j\right\|}\right)\\
&=\frac{1}{2}\mathrm{erf}\left(\frac{\Delta_{ij}}{2\sqrt{2}}\right)=p_i p_j\left(A_{ij}(\hat{\boldsymbol{\theta}})-\frac{1}{2}\right)
\end{aligned} \tag{3.10}$$

因此，對於兩類問題，最大化 $J_{\mathrm{aPTAC}}(\hat{\boldsymbol{\theta}})$，相當於最大化貝氏正確率 $A_{ij}(\hat{\boldsymbol{\theta}})$。現在，我們進一步考慮多類別的情形。由於 $\hat{\boldsymbol{\Theta}}=[\hat{\boldsymbol{\theta}}_1,...,\hat{\boldsymbol{\theta}}_d]$，且

$$\mathrm{trace}(\hat{\boldsymbol{\Theta}}^T\mathbf{S}_{ij}\hat{\boldsymbol{\Theta}})=\hat{\boldsymbol{\theta}}_1^T\mathbf{S}_{ij}\hat{\boldsymbol{\theta}}_1+\hat{\boldsymbol{\theta}}_2^T\mathbf{S}_{ij}\hat{\boldsymbol{\theta}}_2+...+\hat{\boldsymbol{\theta}}_d^T\mathbf{S}_{ij}\hat{\boldsymbol{\theta}}_d=\sum_{m=1}^{d}\hat{\boldsymbol{\theta}}_m^T\mathbf{S}_{ij}\hat{\boldsymbol{\theta}}_m \tag{3.11}$$

[38] 原本 Loog 等人稱之爲近似成對正確標準(approximate pairwise accuracy criterion, aPAC)，往後許多的他人著作亦如此稱之。但爲了與本論文提出『經驗』的方法有所區隔，我們在此加上『理論』二字。本論文之後皆以 aPTAC 來簡稱『近似成對理論正確標準』。在 aPTAC 中，『近似』兩字有兩種含意：一、因爲只考慮每一類別配對，使得產生的正確率並不含多類別之間的關係，將造成正確率的估計過高。二、在多類別時，aPTAC 所決定的子空間也不能保證在某投影方向會產生『最大』的正確率，見 2.3.4 節最後的說明。

[39] 注意，兩類問題只能考慮一個投影方向，即 $\Re^{n\times 1}$ 的子空間，因爲 $d\le\min(n,C-1)=1$。

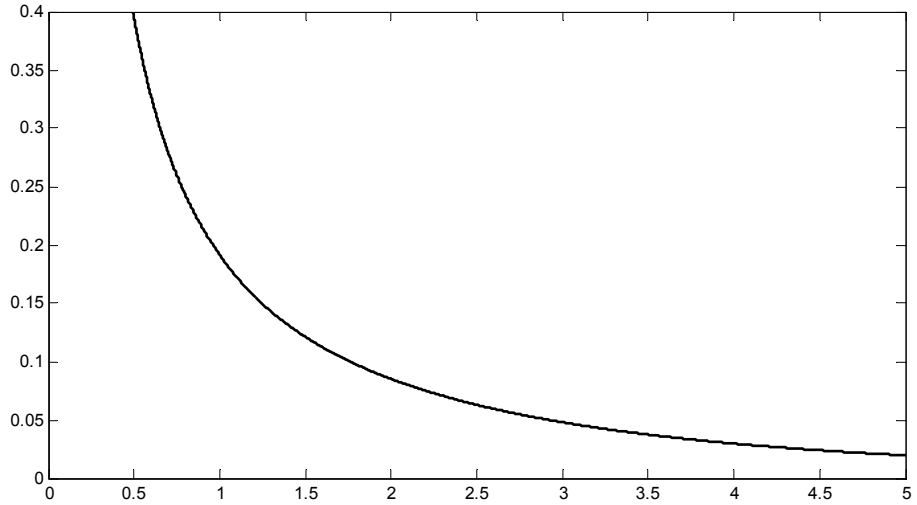


圖 3.3　aPTAC 之距離與權重關係圖

（橫軸為 Δ_{ij} ，縱軸為 $\mathrm{w}_{\mathrm{aPTAC}}(i,j)$ ）

aPTAC 目標函式便可寫成：

$$\begin{aligned} J_{\mathrm{aPTAC}}(\widehat{\boldsymbol{\Theta}}) &= \frac{1}{2}\sum_{i=1}^{C}\sum_{j=1}^{C} p_i p_j \, \mathrm{w}_{\mathrm{aPTAC}}(i,j)\,\mathrm{trace}(\widehat{\boldsymbol{\Theta}}^T \mathbf{S}_{ij}\widehat{\boldsymbol{\Theta}}) \\ &= \frac{1}{2}\sum_{i=1}^{C}\sum_{j=1}^{C} p_i p_j \, \mathrm{w}_{\mathrm{aPTAC}}(i,j)\,\mathrm{trace}\left(\sum_{m=1}^{p}\widehat{\boldsymbol{\theta}}_m^T \mathbf{S}_{ij}\widehat{\boldsymbol{\theta}}_m\right) \\ &= \frac{1}{2}\sum_{m=1}^{d}\sum_{i=1}^{C}\sum_{j=1}^{C} \underbrace{p_i p_j \, \mathrm{w}_{\mathrm{aPTAC}}(i,j)\,\mathrm{trace}(\widehat{\boldsymbol{\theta}}_m^T \mathbf{S}_{ij}\widehat{\boldsymbol{\theta}}_m)}_{\text{approximate accuracy}} \end{aligned} \tag{3.12}$$

其中，式(3.12)中的 $p_i p_j \, \mathrm{w}_{\mathrm{aPTAC}}(i,j)\,\mathrm{trace}(\widehat{\boldsymbol{\theta}}_m^T \mathbf{S}_{ij}\widehat{\boldsymbol{\theta}}_m)$ 可被視為類別母體 C_i 與 C_j 在投影方向 $\widehat{\boldsymbol{\theta}}_m$ 的『近似』貝氏正確率。因此，aPTAC 的目標在於找出 $\widehat{\boldsymbol{\Theta}} = [\widehat{\boldsymbol{\theta}}_1,...,\widehat{\boldsymbol{\theta}}_d]$ ，使得所有類別配對在這些投影方向中之『近似』貝氏正確率的總和最大。而由於 $\mathrm{w}_{\mathrm{aPTAC}}(i,j)$ 顯然並不受到任何非奇異線性轉換影響[40]，因此 aPTAC 的求解完全可依循 WLDA，等同於求解輕省的本徵值問題。

由圖 3.3 可看出，aPTAC 對於馬氏距離較大的類別配對，所配置的權重也會較低，因此解決了 LDA 的過度強調問題。此外，由於權重因子 $\mathrm{w}_{\mathrm{aPTAC}}(i,j)$ 的設

[40] 值得一提的是，馬氏距離大小並不會因任何非奇異線性轉換而改變。

計考量了分類正確率，儘管是個近似的方法，也在一定程度上解決了 LDA 的分類正確率無關問題。但是，aPTAC 仍具有潛在的限制：它假設了所有類別母體均遵循高斯分佈，且分類器的分類規則需嚴格遵照貝氏決策法則才會有較佳的分類效果。若是分類器較複雜，如自動語音辨識系統，則 aPTAC 便無法保證有較佳的辨識率。至於目標函式以類別兩兩配對(pairwise)的方式組成，亦有對正確率高估的情形，這點將在 3.2 節進一步解釋之。

3.2 基於混淆資訊之權重式線性鑑別分析

為了同時解決過度強調的問題，又能將 LDA 適切的聯於後端分類器，以期更高的辨識率，對於 WLDA 中之權重因子 $w(i,j)$ 設計就必須遵循兩個重要法則，在此稱之為權重法則(weighting rules)：

第一，$w(i,j)$ 與馬氏距離需具有某種程度的負相關性(negative correlation)。

第二，由分類器本身提供的資訊必須適當地嵌入 $w(i,j)$。

根據以上法則，我們將循序提出三種方法。

3.2.1 基於經驗錯誤率之權重式線性鑑別分析

第一種方法稱為基於經驗錯誤率之權重式線性鑑別分析(empirical error rate based WLDA, EER-WLDA)[41][71]。首先，我們必須先定義由辨識器所產生之混淆矩陣(confusion matrix) $\mathbf{M}_{\text{EER–WLDA}} = [m_{ij}^{ERR}]_{C\times C}$ 如下：

[41] 本論文之後皆以 EER-WLDA 來簡稱『基於經驗錯誤率之權重式線性鑑別分析』。

$$m_{ij}^{\mathrm{ERR}} = \begin{cases} \dfrac{e_{ij}}{n_i}, & \text{if } i \neq j \\ 0, & \text{if } i = j \end{cases} \tag{3.13}$$

其中，e_{ij} 為原本屬於類別 C_i，卻被錯誤分類至類別 C_j 的特徵向量個數，m_{ij}^{ERR} 則為矩陣 $\mathbf{M}_{\mathrm{EER-WLDA}}$ 的第 i 列、第 j 行之元素，表示原本屬於類別 C_i，卻被錯誤分類至類別 C_j 的經驗分類錯誤率(empirical classification error rate)。m_{ij}^{ERR} 在某種程度上可用來量度類別 C_i 與類別 C_j 之間的混淆度大小。也就是說，對於類別 C_i，若 m_{ij} 愈大，則它與類別 C_j 愈容易發生分類錯誤[42]。

於是，EER-WLDA 相關的權重因子 $\mathrm{w}(i,j)$ 可定義如下：

$$\mathrm{w}_{\mathrm{EER-WLDA}}(i,j) = \alpha + (1-\alpha) \times m_{ij}^{\mathrm{ERR}},\ 0 \leq \alpha \leq 1 \tag{3.14}$$

其中，α 為可自由設定的調整因子，可用來調節距離與經驗分類錯誤率之間的比重。若將式(3.14)代入式(3.2)則可形成 EER-WLDA 目標函式：

$$J_{\mathrm{EER-WLDA}}(\widehat{\boldsymbol{\Theta}}) = \frac{1}{2}\sum_{i=1}^{C}\sum_{j=1}^{C} p_i p_j \left(\alpha + (1-\alpha) \times m_{ij}^{\mathrm{ERR}}\right) \mathrm{trace}(\widehat{\boldsymbol{\Theta}}^T \widehat{\mathbf{S}}_{ij} \widehat{\boldsymbol{\Theta}}) \tag{3.15}$$

由式(3.15)可看出，α 和 m_{ij}^{ERR} 在控制 $\widehat{\mathbf{S}}_{ij}$ 上扮演了很重要的角色。我們可先考慮兩種極端的情況：若 $\alpha = 1$，則 EER-WLDA 顯然還原成 LDA；若 $\alpha = 0$，則 $m_{ij}^{\mathrm{ERR}} \neq 0$ 的類別配對將主導整個權重，而 $m_{ij}^{\mathrm{ERR}} = 0$ 的類別配對的權重貢獻將完全被忽略。此外，由於 m_{ij} 在某種程度上代表了類別配對 C_i 與 C_j 之間的分離度，或在幾何空間上的重疊部分，若類別 C_i 擁有較小的 m_{ij}^{ERR}，則我們可合理推論它與類別 C_j 具

[42] m_{ij}^{ERR} 又可被稱爲生產者錯誤率(producer's error rate)，著重於遺漏錯誤(error of omission)的計算。從生產者的觀點，原本各個類別的總數均視爲已知。與生產者錯誤率相對的是使用者錯誤率(user's error rate)，可被定義爲 e_{ij}/q_i，q_i 爲實際上分至類別 C_i 的總數；它強調判斷錯誤(error of commission)的計算。目前在語音特徵擷取上並無任何理論能證明這兩種錯誤率的計算方式孰優孰劣，不過從實驗結果來看，我們在此選擇生產者錯誤率。

有較好的分離度，那麼權重因子$\left(\alpha+(1-\alpha)\times m_{ij}^{\mathrm{ERR}}\right)$也會變得相對的小；反之，擁有較大$m_{ij}^{\mathrm{ERR}}$的類別配對，則應在距離上得到較大的權重。

不過，儘管 EER-WLDA 提供了一個新嘗試，使得再複雜的辨識器都能夠簡單地介入前端特徵擷取的過程，但它還是一個過於啟發性(heuristic)的方法。除了調整因子α只能隨實驗結果自由設定之外，它並沒有提出任何證據可證明資料的幾何分離度確實與經驗分類錯誤率具有正相關性(positive correlation)。

3.2.2 距離－錯誤耦合之權重式線性鑑別分析

EER-WLDA 的啟發性預設將在第二種方法中獲得解決。首先，我們發現在 ERR-WLDA 中，雖然m_{ij}^{ERR}與m_{ji}^{ERR}並不一定相同，但在其目標函式中，因為幾何距離是對稱的，所以類別配對C_i與C_j之間就沒有順序性(order)。若我們只考慮此類別配對的部分，式(3.15)可寫成：

$$\begin{aligned}&\frac{1}{2}p_i p_j\left(\alpha+(1-\alpha)\times m_{ij}^{\mathrm{ERR}}\right)\mathrm{trace}(\hat{\boldsymbol{\Theta}}^T\hat{\mathbf{S}}_{ij}\hat{\boldsymbol{\Theta}})\\&+\frac{1}{2}p_j p_i\left(\alpha+(1-\alpha)\times m_{ji}^{\mathrm{ERR}}\right)\mathrm{trace}(\hat{\boldsymbol{\Theta}}^T\hat{\mathbf{S}}_{ji}\hat{\boldsymbol{\Theta}})\\&=p_i p_j\left(\alpha+\frac{(1-\alpha)\times(m_{ij}^{\mathrm{ERR}}+m_{ji}^{\mathrm{ERR}})}{2}\right)\mathrm{trace}(\hat{\boldsymbol{\Theta}}^T\hat{\mathbf{S}}_{ij}\hat{\boldsymbol{\Theta}})\end{aligned}\tag{3.16}$$

也就是說，根據式(3.13)，真正代表類別配對C_i與C_j之間混淆程度的是：

$$m_{ij}^{\mathrm{ERR}}+m_{ji}^{\mathrm{ERR}}=\frac{e_{ij}}{n_i}+\frac{e_{ji}}{n_j}\tag{3.17}$$

但是，我們很難從式(3.17)看出此分類錯誤率的計算來自什麼統計基礎或有什麼數學意義。為了更適切地表達類別配對C_i與C_j之間的混淆程度，我們可將混淆

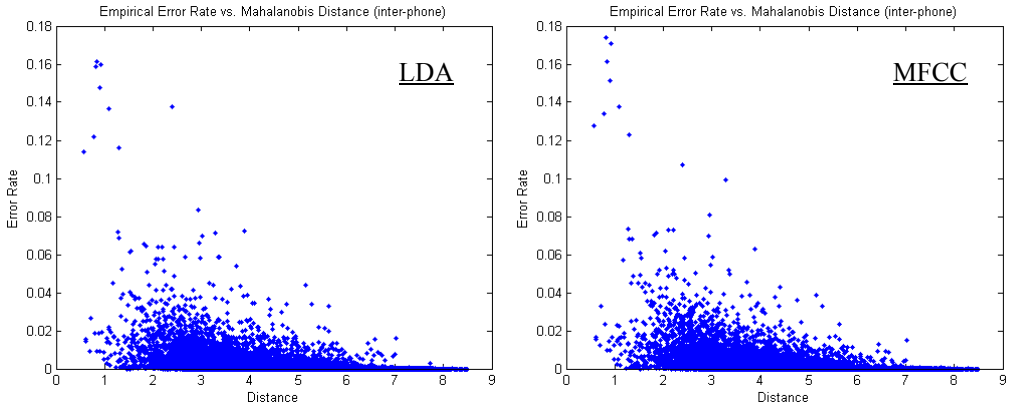


圖 3.4　經驗分類錯誤率與馬氏距離的關係圖（一）
（類別配對屬於不同的音素模型）

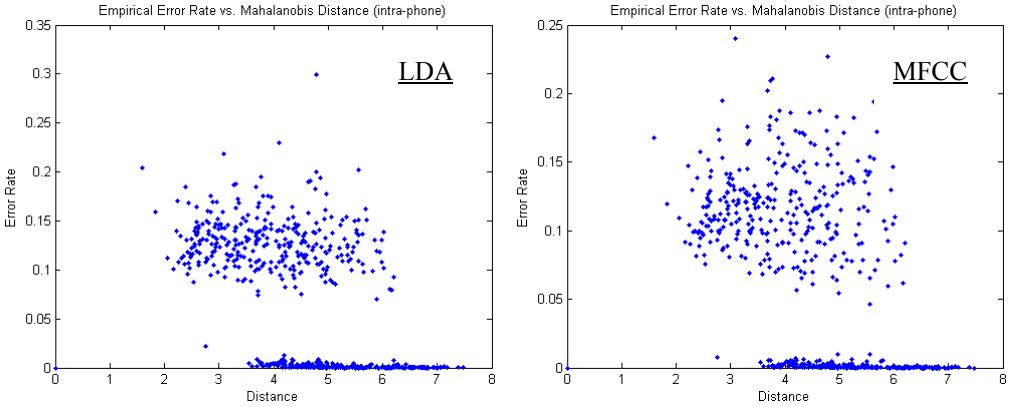


圖 3.5　經驗分類錯誤率與馬氏距離的關係圖（二）
（類別配對屬於相同的音素模型）

矩陣 $\mathbf{M}_{\text{DE–WLDA}} = [m_{ij}^{\text{DE}}]_{C\times C}$ 改寫成：

$$m_{ij}^{\text{DE}} = \begin{cases} \dfrac{e_{ij}}{n_i + n_i}, & \text{if } i \neq j \\ 0, & \text{if } i = j \end{cases} \tag{3.18}$$

式(3.18)的數學意義在於將類別配對 C_i 與 C_j 視為一個整體，計算此整體內發生的錯誤率。當然，這裡的 m_{ij}^{DE} 在某種程度上也可用來量度類別 C_i 與類別 C_j 之間的混淆度大小。

在以隱藏式馬可夫模型中的狀態(state)作為分類基本單位的自動語音辨識實

表 3.1　LDA 與 MFCC 對於 MATBN 訓練語料之音素辨識統計

特徵擷取方法	LDA	MFCC
總音框數	9,183,440	
音素模型數	151	
狀態模型數（基本類別）	455	
所有分類錯誤數／錯誤率	3,896,185 / 42.43%	4,012,326 / 43.69%
不同音素分類錯誤數／錯誤率	2,400,806 / 26.14%	2,637,002 / 28.71%

驗中[31]，我們可以針對 LDA 與 MFCC[43]兩種常見的特徵擷取方法分別繪出所有類別配對之經驗分類錯誤率與馬氏距離的關係圖，如圖 3.4 與圖 3.5，二者的橫軸為類別配對 C_i 與 C_j 的馬氏距離 Δ_{ij}，縱軸則為經驗分類錯誤率 m_{ij}^{DE}。圖 3.4 中的每一點代表隸屬於不同音素模型(phone models)[44]的類別配對，而圖 3.5 中的每一點則代表隸屬於相同音素模型的類別配對。

由圖 3.4 我們可以觀察到一個明顯的現象：馬氏距離較短的類別配對（例如 $\Delta_{ij} < 4$），易於產生較高的經驗分類錯誤率（例如 $m_{ij}^{\mathrm{DE}} > 0.07$），而馬氏距離較長的類別配對（例如 $\Delta_{ij} > 4$），則可能擁有較低的經驗分類錯誤率（例如 $m_{ij}^{\mathrm{DE}} < 0.07$）。值得一提的是，這種現象似乎並不會因著特徵擷取的方式不同而改變，因此我們可以合理地認為這等關係具有一致性。但在圖 3.5 中，經驗分類錯誤率與馬氏距離的關係並不那麼明確。不過幸運的是，我們真正關心的是後端分類器所產生的音素錯誤率(phone error rate, PER)，因為它會影響到真正決定語音辨識效果的字錯誤率(word error rate, WER)，而在同一個音素模型中的狀態分類錯誤，並不會影響最終的音素錯誤率，所以在圖 3.5 中難以捉摸的關係可以放心地被忽略，因此在表 3.1，關於 MATBN 訓練語料之音素辨識統計中，我們只需考慮約 26%的分類錯誤音框，而在同一音素模型中約 16%的分類錯誤音框是可以不考慮的，因

[43] 這兩種方法所產生的投影子空間在此均屬於 $\Re^{39\times 1}$。

[44] 在本論文的中文大詞彙連續語音辨識實驗中，每個音素模型具有 3 至 5 個隱藏式馬可夫模型狀態。

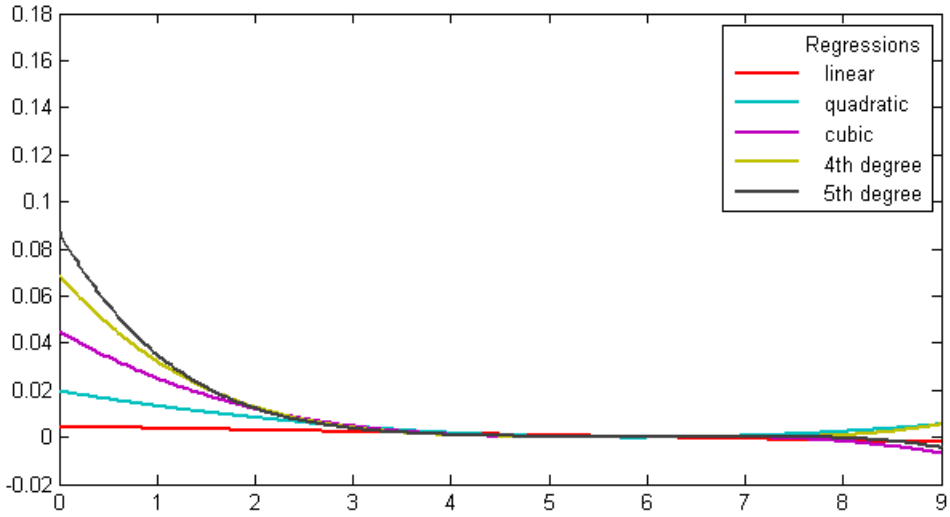


圖 3.6　根據圖 3.5 所繪出不同階數的多項式回歸曲線
（橫軸為 Δ_{ij} ，縱軸為 m_{ij}^{DE} ）

為拉大這些音框所代表之狀態類別的距離並無助於最後的辨識率[45]。

為了能夠更方便地描述圖 3.4 的現象，我們首先想到的是應用資料擬合(data fitting)的方法找出馬氏距離的函數 $\mathrm{E}(\Delta_{ij})$ ，使我們藉著這個函數，不僅歸納出馬氏距離與經驗分類錯誤率的關係，也能夠預測出特定馬氏距離所對應的經驗分類錯誤率 m_{ij}^{DE} 。資料擬合是一種數學最佳化方法，試著在給定一系列已有類別標記之資料點 $\{(u_i, v_i) \mid i = 1,...,n\}$ 的情況下，找出一個函數 $\mathrm{G}(u_i)$ ，使得其輸出 $\tilde{v}_i$ 近似於 $\tilde{v}$ 。換句話說，資料擬合能夠最小化所有資料點 $(u_i, \tilde{v}_i)$ 與 (u_i, v_i) 之間的平方錯誤和(sum of squared error)。

例如，若 $\mathrm{E}(\Delta_{ij})$ 的形式是 2 階多項式(quadratic polynomial)，例如 $\mathrm{E}(\Delta_{ij}) = a\Delta_{ij}^2 + b\Delta_{ij} + c$ ，則給定所有圖 3.4 中的資料點 $\{(\Delta_{ij}, m_{ij}^{\mathrm{DE}}) \mid i = 1,...n\}$ ，我們可以藉著最小化所有 $\left(\mathrm{E}(\Delta_{ij}) - m_{ij}^{\mathrm{DE}}\right)$ 的平方和，來估計 $\mathrm{E}(\Delta_{ij})$ 的參數 a、b、c，整個估計方法可寫成下式：

$$\{\hat{a}, \hat{b}, \hat{c}\} = \arg\min_{a,b,c} \sum_{i=1}^{C-1} \sum_{j=i}^{C} \left((a\Delta_{ij}^2 + b\Delta_{ij} + c) - m_{ij}^{\mathrm{DE}} \right)^2 \tag{3.19}$$

[45] 每一語句內的音素邊界(phone boundary)或音素內的狀態邊界(state boundary)是靠強迫對齊(forced alignment)技術所決定，見[34]。

最後，$\hat{\mathrm{E}}(\Delta_{ij}) = \hat{a}\Delta_{ij}^2 + \hat{b}\Delta_{ij} + \hat{c}$ 即為我們所求的二階經驗分類錯誤率函數。我們將第 1 階到第 5 階的多項式回歸(regression)函數繪於圖 2.6，也證明了馬氏距離和經驗分類錯誤率在較高階的多項式回歸曲線上，具有某種程度的負相關性。

因此，我們便可以設計一種權重因子如下（以 2 階多項式為例）：

$$\mathrm{w}_{\mathrm{DE-WLDA}}(i,j) = \hat{\mathrm{E}}(\Delta_{ij}) = \hat{a}\Delta_{ij}^2 + \hat{b}\Delta_{ij} + \hat{c} \tag{3.20}$$

式(3.20)不僅滿足了第一權重法則：馬氏距離愈大的類別配對，獲得較小的權重，也藉著實際上距離與錯誤率的關係，將混淆資訊適當的嵌入了權重因子。若將式(3.20)代入 WLDA 的式(3.2)則可形成所謂的距離－錯誤耦合之權重式線性鑑別分析(distance-error coupled WLDA, DE-WLDA)[46][72]目標函式：

$$J_{\mathrm{DE-WLDA}}(\widehat{\boldsymbol{\Theta}}) = \frac{1}{2}\sum_{i=1}^{C}\sum_{j=1}^{C} p_i p_j (\hat{a}\Delta_{ij}^2 + \hat{b}\Delta_{ij} + \hat{c}) \operatorname{trace}(\widehat{\boldsymbol{\Theta}}^T \widehat{\mathbf{S}}_{ij} \widehat{\boldsymbol{\Theta}}) \tag{3.21}$$

3.2.3 近似成對經驗正確率標準

前面兩種方法的物理意義仍離不開 WLDA，也就是最大化權重式的平均馬氏距離平方。儘管 DE-WLDA 已嘗試將錯誤率與距離較緊密的連結在一起，整個目標函式仍舊與分類正確率有差距。為了解決此分類正確率無關問題，我們可利用式(3.19)先得出給定某馬氏距離 Δ_{ij} 的經驗分類正確率 $\hat{\mathrm{A}}(\Delta_{ij})$ 如下：

$$\hat{\mathrm{A}}(\Delta_{ij}) = 1 - \hat{\mathrm{E}}(\Delta_{ij}) \tag{3.22}$$

根據式(3.22)，我們可設計另一種新的權重因子：

[46] 本論文之後皆以 DE-WLDA 來簡稱『距離—錯誤耦合之權重式線性鑑別分析』。

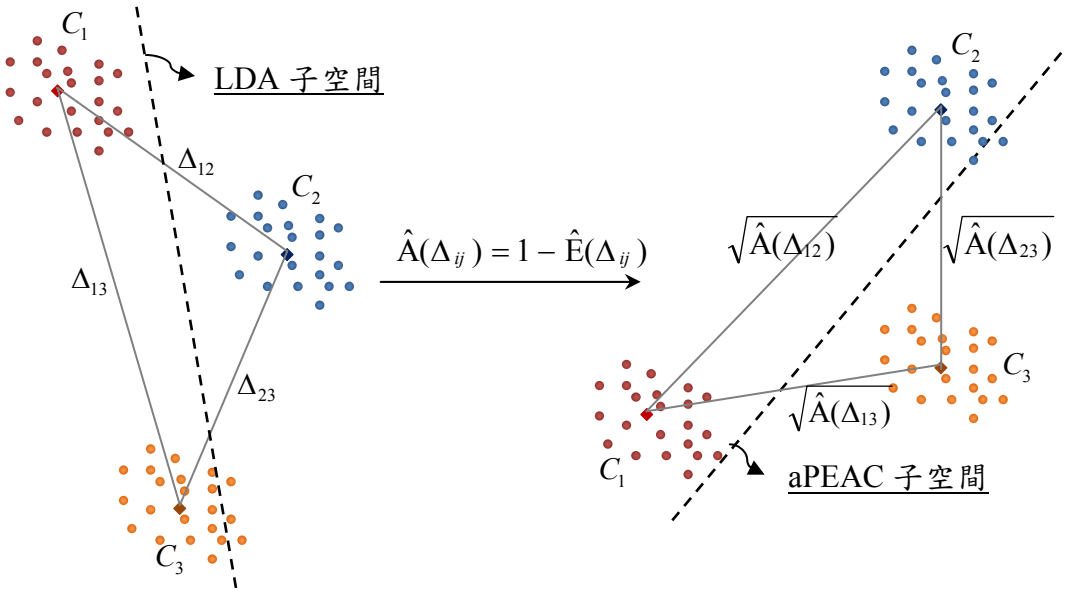


圖 3.7　由 LDA 子空間轉換至 aPEAC 子空間示意圖

$$w_{aPEAC}(i,j) = \frac{\hat{A}(\Delta_{ij})}{\Delta_{ij}^2} = \frac{1-\hat{E}(\Delta_{ij})}{\Delta_{ij}^2} \tag{3.23}$$

將式(3.23)代入 WLDA 的式(3.2)則可形成所謂的近似成對經驗正確率標準(approximate pairwise empirical accuracy criterion, aPEAC)[47][73]目標函式：

$$J_{aPEAC}(\widehat{\boldsymbol{\Theta}}) = \frac{1}{2}\sum_{i=1}^{C}\sum_{j=1}^{C} p_i p_j \left(\frac{1-\hat{E}(\Delta_{ij})}{\Delta_{ij}^2}\right) \mathrm{trace}(\widehat{\boldsymbol{\Theta}}^T \widehat{\mathbf{S}}_{ij} \widehat{\boldsymbol{\Theta}}) \tag{3.24}$$

從數學意義來看，我們可以證明 aPEAC 能夠成功的將原本 LDA 的平均成對馬氏距離平方最大化轉為平均成對經驗分類正確率最大化。若 $\widehat{\boldsymbol{\Theta}} = [\widehat{\boldsymbol{\theta}}_1,...,\widehat{\boldsymbol{\theta}}_d]$，則我們可推導出以下的式(3.25)：

[47] 本論文之後皆以 aPEAC 來簡稱『近似成對經驗正確率標準』。

表 3.2　LDA、aPTAC 和 aPEAC 性質比較表

	LDA	**aPTAC**	**aPEAC**
類別分離度標準	馬氏距離	貝氏正確率	經驗正確率
機率分佈假設	無	高斯分佈	無
同方差性假設	有	有	有
可解性	本徵值問題	本徵值問題	本徵值問題
考慮分類器性質	無	無	有

$$
\begin{aligned}
J_{\text{aPEAC}}(\widehat{\boldsymbol{\Theta}}) &= \frac{1}{2}\sum_{i=1}^{C}\sum_{j=1}^{C} p_i p_j \left(\frac{1-\hat{\mathrm{E}}(\Delta_{ij})}{\Delta_{ij}^2}\right)\mathrm{trace}(\widehat{\boldsymbol{\Theta}}^T\widehat{\mathbf{S}}_{ij}\widehat{\boldsymbol{\Theta}}) \\
&= \frac{1}{2}\sum_{i=1}^{C}\sum_{j=1}^{C} p_i p_j \left(1-\hat{\mathrm{E}}(\Delta_{ij})\right)\mathrm{trace}\left(\widehat{\boldsymbol{\Theta}}^T\left(\frac{\widehat{\mathbf{m}}_i-\widehat{\mathbf{m}}_j}{\Delta_{ij}}\right)\left(\frac{\widehat{\mathbf{m}}_i-\widehat{\mathbf{m}}_j}{\Delta_{ij}}\right)^T\widehat{\boldsymbol{\Theta}}\right) \\
&= \frac{1}{2}\sum_{m=1}^{d}\sum_{i=1}^{C}\sum_{j=1}^{C} p_i p_j \left(1-\hat{\mathrm{E}}(\Delta_{ij})\right)\left\|\boldsymbol{\theta}_m^T\left(\frac{\widehat{\mathbf{m}}_i-\widehat{\mathbf{m}}_j}{\Delta_{ij}}\right)\right\|^2 \\
&= \frac{1}{2}\sum_{m=1}^{d}\sum_{i=1}^{C}\sum_{j=1}^{C} p_i p_j \left\|\boldsymbol{\theta}_m^T\sqrt{1-\hat{\mathrm{E}}(\Delta_{ij})}\,\hat{\mathbf{m}}_{ij}\right\|^2
\end{aligned}
\tag{3.25}
$$

由式(3.25)我們可看出 aPEAC 完全將『距離大小』這個因子從目標函式中剔除，取而代之的是經驗分類正確率。也就是說，aPEAC 的目標在於找出一個投影子空間，使得所有資料在此空間中具有最大的成對經驗正確率。如圖 3.7，我們可以看出，原本在 LDA 的求解中，其子空間（或投影方向）主要是由在原始空間中馬氏距離較大的類別配對（如類別 C_1 和類別 C_3）所決定，如前所述，這會造成過度強調問題。但是經由 aPEAC 的轉換，子空間的決定因素便不是馬氏距離了，取而代之的是在分類上更有助益的經驗分類正確率。因此，在轉換後空間中具有較大成對經驗分類正確率的類別配對（如類別 C_1 和類別 C_2）反而對主要的投影方向提供較大的貢獻。

3.2.4　aPTAC 與 aPEAC 之比較

如表 3.2，我們可以看出 aPTAC 和 aPEAC 都是基於分類正確率所發展出來

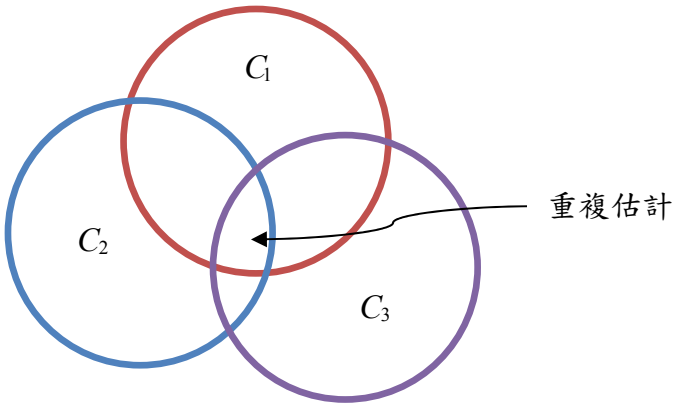


圖 3.8　aPTAC 與 aPEAC 之重複估測問題

的方法，二者並承繼了傳統 LDA 輕省的可解性。相對於 aPTAC，aPEAC 所受的限制顯然比較少，並且它實際地藉由觀察考慮了分類器本身的性質。因此，當資料分佈並不遵循高斯分佈，分類器內的分類規則並非簡單的貝氏決策法則，則 aPTAC 產生的特徵勢必不利於分類結果。

除此之外，不只是 aPTAC，aPEAC 也會遇到重複估計的問題(over-estimation problem)。如圖 3.8，aPTAC 和 aPEAC 會將整體的分類錯誤率視為類別配對 C_1 和 C_2、C_1 和 C_3、C_2 和 C_3 的分類錯誤率總和，如此，三個類別重疊的部分便會被重複估計。因此，更正確的說，aPTAC 和 aPEAC 所做的只是最小化全部分類錯誤率的上界(upper bound)或最大化全部分類正確率的下界(lower bound)。

3.3　基於經驗錯誤率之類別內共變異矩陣

我們曾在 2.3.2 節中，基於 LDA 的幾何分析，提出兩種針對 LDA 的改進方向。其中之一是試著找出比 $\mathbf{S}_W^{-1/2}$ 更佳的白化轉換，或是比 $\mathbf{S}_W$ 更有幫助的類別內散佈矩陣。Tang 等人認為，雖然 $\mathbf{S}_W$（見式(2.14)）在統計上是無偏差的估計量(unbiased estimator)，但對於分類正確率就未必合適了。我們可以設想一個最糟的情況，若某一類別母體之共變異矩陣內的元素，例如第 1 行、第 1 列的元素，

異常地較其它類別母體之共變異矩陣內所對應元素還大，則此類別將會主導整個 $\mathbf{S}_W$ 的計算，而導致所產生的 $\mathbf{S}_W$ 只對此類別具有代表性。一旦此類別同時又是離群類別(outlier)或噪音(noise)，則會使 LDA 的工作只在於使此類別在投影後的變異度最小，而無助於分類。因此，為了減少離群類別的影響力，他們提出了關連權重式類別內共變異矩陣(relevance weighted within-class covariance matrix, RWW)[48]的概念[74]：

$$\mathbf{S}_W^{\mathrm{RWW}} = \sum_{i=1}^{C} p_i r_i \mathbf{S}_i \tag{3.26}$$

其中，r_i 是基於關連性的權重因子，其定義如下：

$$r_i = \sum_{j \neq i} \frac{1}{L_{ij}} \tag{3.27}$$

L_{ij} 是類別 C_i 與類別 C_j 的相異度(dissimilarity)，用來估計在原始空間中，類別 C_i 與類別 C_j 的分離程度。常見的 L_{ij} 設定有歐氏距離、馬氏距離或貝氏正確率等。由式(3.27)可知，當類別 C_i 與其它類別的相異度都較高時，可被視為離群類別，因此所得的權重因子 r_i 較低。

筆者認為，RWW 具有理論上的缺陷。當我們假設資料中有離群類別時，RWW 所做的是降低它的影響力，但若實際上離群類別不止 1 個，且這些離群類別之間的相異度均很低，則彼此反而都能獲得較高的權重。在 RWW 的框架下，我們甚至可以將之比喻成『兩粒老鼠屎足以壞了一鍋粥』。因此，面對離群類別的存在可能性，我們選擇不去降低它的影響力，而是著重它在分類上的角色。對於每一類別，可藉由分類器產生的混淆資訊來判斷它是否具有分類上的重要性。利用式(3.13)的混淆矩陣，基於經驗錯誤率之類別內共變異矩陣(empirical error

[48] 本論文之後皆以 RWW 來簡稱『關連權重式類別內共變異矩陣』。

rate based within-class covariance matrix, EERW)[49]可被定義如下：

$$\mathbf{S}_W^{\mathrm{EERW}} = \sum_{i=1}^{C} p_i \left(\sum_{j \neq i} m_{ij}^{\mathrm{DE}} \right) \mathbf{S}_i \tag{3.28}$$

在式(3.28)中，$\sum_{j \neq i} m_{ij}^{\mathrm{DE}}$ 代表類別 C_i 的經驗分類錯誤率，它與事前機率結合，反映了此類別中容易造成錯誤的資料數（或音框數）。若這樣的數目很大，則此類別不該被視為離群類別，反之應加以強調，以有利於後端分類器對此類別的處理。

[49] 本論文之後皆以 EERW 來簡稱『基於經驗錯誤率之類別內共變異矩陣』。

第 4 章　普遍化相似度比率鑑別分析

在 3.2 節中，我們所提出的三種方法全都具有同方差性假設。在本章，我們將提出另一種方法，在普遍化相似度比率的框架下進行鑑別式特徵擷取，並試著打破同方差性假設，以及進一步結合經驗分類資訊。

4.1　相似度比率檢定

根據統計式假設檢定(statistical hypothesis testing)的定義[59]，相似度比率檢定(likelihood ratio test, LRT)是一種廣為使用的方法，藉著它我們可獲得虛無假設(null hypothesis) H_0與完全普遍化之對立假設(alternative hypothesis) H_1間的檢定統計量。在本論文中，虛無假設H_0通常表示不利於我們的目標設定，或我們不願見到的情況，在鑑別式特徵擷取上，即為使類別母體不具鑑別性的情況。值得一提的是，虛無假設和對立假設之聯集(union)恰為完整的參數空間。

若Ω表示完整的參數空間(parameter space)，而ω表示被虛無假設H_0所限制的參數空間，則相似度比率檢定針對虛無假設H_0和對立假設H_1之間的標準為

$$LR = \frac{\sup L_\omega}{\sup L_\Omega} \tag{4.1}$$

其中，L表示資料的相似度，$\sup L_S$則表示以S為參數空間時的最大相似度。由式(4.1)可看出，相似度比率檢定是由兩個部分組成：最大相似度與比率。使用最大化相似度估計法的用意在於找出最適合兩個統計假設或最具代表性的參數估計量。而相似度比率背後的邏輯則在於，若我們不考慮任何信心度量測(confidence measure)且虛無假設H_0絕對為真(true)，則在完整參數空間Ω中的最大相似度必定發生在參數空間為ω的情況；因此，$\sup L_\omega$與$\sup L_\Omega$必定會非常接

近，使 LR 趨近於 1。反之，若 H_0 絕對為假(false)，則最大相似度發生的參數空間必定不是 ω；因此，$\sup L_\omega$ 將會小於 $\sup L_\Omega$。

4.2 普遍化相似度比率鑑別分析

相似度比率檢定在語音處理上的應用並不多見，近年來它常被用於評估音素間的混淆程度(phonetic confusions)[75]或是語音活動偵測(voice activity detection, VAD)[76]。在鑑別式語音特徵擷取技術中，我們並不打算緊密遵照相似度比率檢定的過程，目標也不在依據統計量來檢定虛無假設是真是假。我們的目的在於尋找一個投影子空間，使得虛無假設在此子空間中盡可能不會為真。為了使所有類別母體在此子空間中具有鑑別性，我們設計了以下的鑑別式統計假設：

$$\begin{cases} H_0：\text{所有類別母體都相同。} \\ H_1：\text{所有類別母體都相異。} \end{cases}$$

因此，我們所找到的子空間（為 $\boldsymbol{\Theta} \in \Re^{n\times d}$ 的行向量(column vectors)所生成），必須盡可能的推翻不具鑑別性的虛無假設 H_0，換句話說，即使得所有聲學特徵在此虛無假設下的相似度要盡可能的小。根據這個原則，我們提出了普遍化相似度比率鑑別分析(generalized likelihood ratio discriminant analysis, GLRDA)[50]，其目標函式可寫成

$$J_{\text{GLRDA}}(\boldsymbol{\Theta}) = \arg\min_{\boldsymbol{\Theta}} \frac{\sup L_{\text{假設所有類別母體均相同的參數空間}}(\boldsymbol{\Theta})}{\sup L_{\text{完整的參數空間}}(\boldsymbol{\Theta})} \tag{4.2}$$

最後，GLRDA 的轉換矩陣 $\boldsymbol{\Theta} \in \Re^{n\times d}$ 可藉著最小化 $J_{\text{GLRDA}}(\boldsymbol{\Theta})$ 求得。

[50] 本論文之後皆以 GLRDA 來簡稱『普遍化相似度比率鑑別分析』。

4.2.1 同方差性(Homoscedasticity)

一般來說，我們會以類別母體之期望值向量的估計量異同作為所有類別母體是否相同的參數空間。若所有類別母體遵循高斯分佈且具同方差性，$\boldsymbol{\mu}_i$ 為每一類別母體 C_i 的期望值向量，$\boldsymbol{\Sigma}_i$ 為每一類別母體 C_i 的共變異矩陣，則 H_0^{homo} 和 H_1^{homo} 可設定為：

$$\begin{cases} H_0^{\text{homo}}：\text{每一類別母體 } C_i \text{ 均呈高斯分佈，且 } \boldsymbol{\Sigma}_i = \boldsymbol{\Sigma}\text{，}\boldsymbol{\mu}_i = \boldsymbol{\mu}\text{。} \\ H_1^{\text{homo}}：\text{每一類別母體 } C_i \text{ 均呈高斯分佈，且 } \boldsymbol{\Sigma}_i = \boldsymbol{\Sigma}\text{，}\boldsymbol{\mu}_i \text{ 不受任何限制。} \end{cases}$$

H_0^{homo} 代表了一種極端的情況，若它為真，則所有類別母體幾近完全重疊，也就沒有任何鑑別性。因此，GLRDA 的任務即在 H_0^{homo} 最不可能為真的情況下，找出合適的投影子空間。

以下的命題 4.1 說明了 LDA 轉換矩陣，等同於將 H_0^{homo} 與 H_1^{homo} 置於 GLRDA 的框架下求解。

命題 4.1：若每一類別母體 C_i 都具有高斯機率分佈 $N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$，則最小化同方差性之 GLRDA 的目標函式

$$J_{\text{GLRDA}}^{\text{homo}}(\boldsymbol{\Theta}) = \frac{\sup L_{H_0^{\text{homo}}}(\boldsymbol{\Theta})}{\sup L_{H_1^{\text{homo}}}(\boldsymbol{\Theta})} \tag{4.3}$$

等同於最大化 LDA 目標函式，即式(2.21)。

證明：為了方便起見，我們先將式(4.3)取對數，這並不影響 $\boldsymbol{\Theta}$ 的求解：

$$\log J_{\mathrm{GLRDA}}^{\mathrm{homo}}(\boldsymbol{\Theta}) = \sup \log L_{H_0^{\mathrm{homo}}}(\boldsymbol{\Theta}) - \sup \log L_{H_1^{\mathrm{homo}}}(\boldsymbol{\Theta}) \tag{4.4}$$

而 $\sup \log L_{H_0^{\mathrm{homo}}}(\boldsymbol{\Theta})$ 和 $\sup \log L_{H_1^{\mathrm{homo}}}(\boldsymbol{\Theta})$ 可被進一步分別表示為所有資料 $\mathbf{x}_1^N$ 在每一高斯分佈的類別母體下的相似度（見附錄 7.2）：

$$\begin{aligned} &\sup \log L_{H_0^{\mathrm{homo}}}(\boldsymbol{\Theta}) = \max \log p(\mathbf{x}_1^N, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\Theta}) \\ &= \max\left(g(N,d) - \sum_{i=1}^{C} \frac{n_i}{2} \left((\tilde{\mathbf{m}}_i - \tilde{\boldsymbol{\mu}})^T \tilde{\boldsymbol{\Sigma}}^{-1} (\tilde{\mathbf{m}}_i - \tilde{\boldsymbol{\mu}}) + \mathrm{trace}(\tilde{\boldsymbol{\Sigma}}^{-1} \tilde{\mathbf{S}}_i) + \log |\tilde{\boldsymbol{\Sigma}}| \right) \right) \end{aligned} \tag{4.5}$$

$$\begin{aligned} &\sup \log L_{H_1^{\mathrm{homo}}}(\boldsymbol{\Theta}) = \max \log p(\mathbf{x}_1^N, \{\boldsymbol{\mu}_i\}, \boldsymbol{\Sigma}, \boldsymbol{\Theta}) \\ &= \max\left(g(N,d) - \sum_{i=1}^{C} \frac{n_i}{2} \left((\tilde{\mathbf{m}}_i - \tilde{\boldsymbol{\mu}}_i)^T \tilde{\boldsymbol{\Sigma}}^{-1} (\tilde{\mathbf{m}}_i - \tilde{\boldsymbol{\mu}}_i) + \mathrm{trace}(\tilde{\boldsymbol{\Sigma}}^{-1} \tilde{\mathbf{S}}_i) + \log |\tilde{\boldsymbol{\Sigma}}| \right) \right) \end{aligned} \tag{4.6}$$

其中，$g(N,d) = (-Nd/2)\log(2\pi)$，$d$ 為投影特徵子空間的維度（或特徵數）；$\tilde{\mathbf{m}}_i$ 和 $\tilde{\mathbf{S}}_i$ 分別為經過 $\boldsymbol{\Theta}$ 轉換後之樣本(sample)的期望值向量與共變異矩陣，而 $\tilde{\boldsymbol{\mu}}$、$\{\tilde{\boldsymbol{\mu}}_i\}$ 和 $\tilde{\boldsymbol{\Sigma}}$ 則是根據統計假設 H_0^{homo} 與 H_1^{homo} 而設定，經過 $\boldsymbol{\Theta}$ 轉換後之母體(population)期望值向量與共變異矩陣，即我們所要估計的參數。

欲求在假設 H_0^{homo} 下，$\tilde{\boldsymbol{\mu}}_0^{\mathrm{homo}}$ 和 $\tilde{\boldsymbol{\Sigma}}_0^{\mathrm{homo}}$ 的最大相似度估計子(ML estimator)，可將 $\log L_{H_0^{\mathrm{homo}}}(\boldsymbol{\Theta})$ 分別對 $\tilde{\boldsymbol{\mu}}$ 和 $\tilde{\boldsymbol{\Sigma}}$ 偏微分，並令其為 0，可得：

$$
\begin{aligned}
&\frac{\partial \log L_{H_0^{\text{homo}}}(\boldsymbol{\Theta})}{\partial \widetilde{\boldsymbol{\mu}}} \\
&= \frac{\partial\left(-\sum_{i=1}^{C}\frac{n_i}{2}\left((\widetilde{\mathbf{m}}_i-\widetilde{\boldsymbol{\mu}})^T\widetilde{\boldsymbol{\Sigma}}^{-1}(\widetilde{\mathbf{m}}_i-\widetilde{\boldsymbol{\mu}})+\text{trace}(\widetilde{\boldsymbol{\Sigma}}^{-1}\widetilde{\mathbf{S}}_i)+\log|\widetilde{\boldsymbol{\Sigma}}|\right)\right)}{\partial\widetilde{\boldsymbol{\mu}}} \\
&= \frac{\partial\left(-\sum_{i=1}^{C}\frac{n_i}{2}\left((\widetilde{\mathbf{m}}_i-\widetilde{\boldsymbol{\mu}})^T\widetilde{\boldsymbol{\Sigma}}^{-1}(\widetilde{\mathbf{m}}_i-\widetilde{\boldsymbol{\mu}})\right)\right)}{\partial\widetilde{\boldsymbol{\mu}}} \\
&= -\sum_{i=1}^{C} n_i\left(\widetilde{\boldsymbol{\Sigma}}^{-1}(\widetilde{\mathbf{m}}_i-\widetilde{\boldsymbol{\mu}})\right)=0 \\
&\Rightarrow \sum_{i=1}^{C} n_i\widetilde{\boldsymbol{\Sigma}}^{-1}\widetilde{\boldsymbol{\mu}} = \sum_{i=i}^{C} n_i\widetilde{\boldsymbol{\Sigma}}^{-1}\widetilde{\mathbf{m}}_i \\
&\Rightarrow \widetilde{\boldsymbol{\mu}}_0^{\text{homo}} = \sum_{i=i}^{C}\frac{n_i}{N}\widetilde{\mathbf{m}}_i = \widetilde{\mathbf{m}}
\end{aligned}
\tag{4.7}
$$

$$
\begin{aligned}
&\frac{\partial \log L_{H_0^{\text{homo}}}(\boldsymbol{\Theta})}{\partial \widetilde{\boldsymbol{\Sigma}}} \\
&= \frac{\partial\left(-\sum_{i=1}^{C}\frac{n_i}{2}\left((\widetilde{\mathbf{m}}_i-\widetilde{\mathbf{m}})^T\widetilde{\boldsymbol{\Sigma}}^{-1}(\widetilde{\mathbf{m}}_i-\widetilde{\mathbf{m}})+\text{trace}(\widetilde{\boldsymbol{\Sigma}}^{-1}\widetilde{\mathbf{S}}_i)+\log|\widetilde{\boldsymbol{\Sigma}}|\right)\right)}{\partial\widetilde{\boldsymbol{\Sigma}}} \\
&= -\sum_{i=1}^{C} n_i\widetilde{\boldsymbol{\Sigma}}^{-1} + \sum_{i=1}^{C} n_i\widetilde{\boldsymbol{\Sigma}}^{-1}\widetilde{\mathbf{S}}_i\widetilde{\boldsymbol{\Sigma}}^{-1} + \sum_{i=1}^{C} n_i\widetilde{\boldsymbol{\Sigma}}^{-1}(\widetilde{\mathbf{m}}_i-\widetilde{\mathbf{m}})\widetilde{\boldsymbol{\Sigma}}^{-1}(\widetilde{\mathbf{m}}_i-\widetilde{\mathbf{m}})^T\widetilde{\boldsymbol{\Sigma}}^{-1} \\
&= -N\widetilde{\boldsymbol{\Sigma}}^{-1} + \widetilde{\boldsymbol{\Sigma}}^{-1}\left(\sum_{i=1}^{C} n_i\widetilde{\mathbf{S}}_i\right)\widetilde{\boldsymbol{\Sigma}}^{-1} + \widetilde{\boldsymbol{\Sigma}}^{-1}\left(\sum_{i=1}^{C} n_i(\widetilde{\mathbf{m}}_i-\widetilde{\mathbf{m}})(\widetilde{\mathbf{m}}_i-\widetilde{\mathbf{m}})^T\right)\widetilde{\boldsymbol{\Sigma}}^{-1}=0 \\
&\Rightarrow \widetilde{\boldsymbol{\Sigma}}_0^{\text{homo}} = \left(\sum_{i=1}^{C}\frac{n_i}{N}\widetilde{\mathbf{S}}_i\right) + \left(\sum_{i=1}^{C}\frac{n_i}{N}(\widetilde{\mathbf{m}}_i-\widetilde{\mathbf{m}})(\widetilde{\mathbf{m}}_i-\widetilde{\mathbf{m}})^T\right) = \widetilde{\mathbf{S}}_W + \widetilde{\mathbf{S}}_B = \widetilde{\mathbf{S}}_T
\end{aligned}
\tag{4.8}
$$

將 $\widetilde{\boldsymbol{\mu}}_0^{\text{homo}} = \widetilde{\mathbf{m}}$ 和 $\widetilde{\boldsymbol{\Sigma}}_0^{\text{homo}} = \widetilde{\mathbf{S}}_T$ 代入式(4.5)，可得在假設 H_0^{homo} 下的最大對數相似度：

$$
\sup \log L_{H_0^{\text{homo}}}(\boldsymbol{\Theta}) = \max \log p(\mathbf{x}_1^N, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\Theta}) = g(N,d) - \frac{N}{2}\log|\widetilde{\mathbf{S}}_T| - \frac{Nd}{2} \tag{4.9}
$$

同理，欲求在假設 H_1^{homo} 下，$\{\widetilde{\boldsymbol{\mu}}_{1,i}^{\text{homo}}\}$ 和 $\widetilde{\boldsymbol{\Sigma}}_1^{\text{homo}}$ 的最大相似度估計子(ML estimator)，可將 $\log L_{H_1^{\text{homo}}}(\boldsymbol{\Theta})$ 分別對 $\widetilde{\boldsymbol{\mu}}_i$ 和 $\widetilde{\boldsymbol{\Sigma}}$ 偏微分，並令其為 0，可得：

$$
\begin{aligned}
&\frac{\partial \log L_{H_1^{\text{homo}}}(\boldsymbol{\Theta})}{\partial \widetilde{\boldsymbol{\mu}}_i} \\
&= \frac{\partial\left(-\sum_{i=1}^{C}\frac{n_i}{2}\left((\widetilde{\mathbf{m}}_i-\widetilde{\boldsymbol{\mu}}_i)^T\widetilde{\boldsymbol{\Sigma}}^{-1}(\widetilde{\mathbf{m}}_i-\widetilde{\boldsymbol{\mu}}_i)+\text{trace}(\widetilde{\boldsymbol{\Sigma}}^{-1}\widetilde{\mathbf{S}}_i)+\log|\widetilde{\boldsymbol{\Sigma}}|\right)\right)}{\partial \widetilde{\boldsymbol{\mu}}_i} \\
&= \frac{\partial\left(-\sum_{i=1}^{C}\frac{n_i}{2}\left((\widetilde{\mathbf{m}}_i-\widetilde{\boldsymbol{\mu}}_i)^T\widetilde{\boldsymbol{\Sigma}}^{-1}(\widetilde{\mathbf{m}}_i-\widetilde{\boldsymbol{\mu}}_i)\right)\right)}{\partial \widetilde{\boldsymbol{\mu}}_i} \\
&= n_i\left(\widetilde{\boldsymbol{\Sigma}}^{-1}(\widetilde{\mathbf{m}}_i-\widetilde{\boldsymbol{\mu}}_i)\right)=0 \\
&\Rightarrow \widetilde{\boldsymbol{\mu}}_{1,i}^{\text{homo}}=\widetilde{\mathbf{m}}_i
\end{aligned}
\tag{4.10}
$$

$$
\begin{aligned}
&\frac{\partial \log L_{H_1^{\text{homo}}}(\boldsymbol{\Theta})}{\partial \widetilde{\boldsymbol{\Sigma}}} \\
&= \frac{\partial\left(-\sum_{i=1}^{C}\frac{n_i}{2}\left((\widetilde{\mathbf{m}}_i-\widetilde{\mathbf{m}}_i)^T\widetilde{\boldsymbol{\Sigma}}^{-1}(\widetilde{\mathbf{m}}_i-\widetilde{\mathbf{m}}_i)+\text{trace}(\widetilde{\boldsymbol{\Sigma}}^{-1}\widetilde{\mathbf{S}}_i)+\log|\widetilde{\boldsymbol{\Sigma}}|\right)\right)}{\partial \widetilde{\boldsymbol{\Sigma}}} \\
&= \frac{\partial\left(-\sum_{i=1}^{C}\frac{n_i}{2}\left(\text{trace}(\widetilde{\boldsymbol{\Sigma}}^{-1}\widetilde{\mathbf{S}}_i)+\log|\widetilde{\boldsymbol{\Sigma}}|\right)\right)}{\partial \widetilde{\boldsymbol{\Sigma}}} \\
&= -\sum_{i=1}^{C} n_i\widetilde{\boldsymbol{\Sigma}}^{-1}+\sum_{i=1}^{C} n_i\widetilde{\boldsymbol{\Sigma}}^{-1}\widetilde{\mathbf{S}}_i\widetilde{\boldsymbol{\Sigma}}^{-1} \\
&= -N\widetilde{\boldsymbol{\Sigma}}^{-1}+\widetilde{\boldsymbol{\Sigma}}^{-1}\left(\sum_{i=1}^{C} n_i\widetilde{\mathbf{S}}_i\right)\widetilde{\boldsymbol{\Sigma}}^{-1}=0 \\
&\Rightarrow \widetilde{\boldsymbol{\Sigma}}_1^{\text{homo}}=\left(\sum_{i=1}^{C}\frac{n_i}{N}\widetilde{\mathbf{S}}_i\right)=\widetilde{\mathbf{S}}_W
\end{aligned}
\tag{4.11}
$$

將 $\widetilde{\boldsymbol{\mu}}_{1,i}^{\text{homo}}=\widetilde{\mathbf{m}}_i$ 和 $\widetilde{\boldsymbol{\Sigma}}_1^{\text{homo}}=\widetilde{\mathbf{S}}_W$ 代入式(4.6)，可得在假設 H_1^{homo} 下的最大對數相似度：

$$
\begin{aligned}
&\sup \log L_{H_1^{\text{homo}}}(\boldsymbol{\Theta})=\log\max\, p(\mathbf{x}_1^N,\{\boldsymbol{\mu}_i\},\boldsymbol{\Sigma},\boldsymbol{\Theta}) \\
&= g(N,d)-\frac{N}{2}\log|\widetilde{\mathbf{S}}_W|-\frac{Nd}{2}
\end{aligned}
\tag{4.12}
$$

最後，將式(4.9)與式(4.12)代入式(4.4)，可得：

$$
\begin{aligned}
&\log J_{\text{GLRDA}}^{\text{homo}}(\boldsymbol{\Theta}) \\
&= \left(g(N,d) - \frac{N}{2}\log|\widetilde{\mathbf{S}}_T| - \frac{Nd}{2} \right) - \left(g(N,d) - \frac{N}{2}\log|\widetilde{\mathbf{S}}_W| - \frac{Nd}{2} \right) \\
&= \frac{N}{2}(\log|\widetilde{\mathbf{S}}_W| - \log|\widetilde{\mathbf{S}}_T|) \\
&= \frac{N}{2}(\log|\widetilde{\mathbf{S}}_W| - \log(|\widetilde{\mathbf{S}}_B| + |\widetilde{\mathbf{S}}_W|) \qquad (4.13) \\
&= \frac{N}{2}\log\frac{|\widetilde{\mathbf{S}}_W|}{|\widetilde{\mathbf{S}}_B| + |\widetilde{\mathbf{S}}_W|} \\
&= \frac{N}{2}\log\frac{1}{|\widetilde{\mathbf{S}}_B| \big/ |\widetilde{\mathbf{S}}_W| + 1}
\end{aligned}
$$

$\boldsymbol{\Theta}$ 可經由最小化式(4.13)而求出。又因為對數函數為單調遞增(monotonically increasing)函數，所以

$$
\begin{aligned}
\boldsymbol{\Theta} &= \arg\min_{\boldsymbol{\Theta}} \log J_{\text{GLRDA}}^{\text{homo}}(\boldsymbol{\Theta}) = \arg\min_{\boldsymbol{\Theta}} \frac{N}{2}\log\frac{1}{|\widetilde{\mathbf{S}}_B| \big/ |\widetilde{\mathbf{S}}_W| + 1} \\
&= \arg\min_{\boldsymbol{\Theta}} \frac{1}{|\widetilde{\mathbf{S}}_B| \big/ |\widetilde{\mathbf{S}}_W| + 1} = \arg\max_{\boldsymbol{\Theta}} \frac{|\widetilde{\mathbf{S}}_B|}{|\widetilde{\mathbf{S}}_W|} \qquad (4.14) \\
&= \arg\max_{\boldsymbol{\Theta}} \frac{|\boldsymbol{\Theta}^T \mathbf{S}_B \boldsymbol{\Theta}|}{|\boldsymbol{\Theta}^T \mathbf{S}_W \boldsymbol{\Theta}|} = \arg\max_{\boldsymbol{\Theta}} J_{\text{LDA\_DET}}(\boldsymbol{\Theta})
\end{aligned}
$$

由式(4.14)知，最小化同方差性之 GLRDA 的目標函式，等同於最大化 LDA 的目標函式。 ■

4.2.2 異方差性(Heteroscedasticity)

現在，我們考慮異方差性的統計假設[77]：

$$\begin{cases} H_0^{\text{heter}}：\text{每一類別母體 } C_i \text{ 均呈高斯分佈，且 } \boldsymbol{\mu}_i = \boldsymbol{\mu} \text{ ，} \boldsymbol{\Sigma}_i \text{ 不受任何限制。} \\ H_1^{\text{heter}}：\text{每一類別母體 } C_i \text{ 均呈高斯分佈，且 } \boldsymbol{\Sigma}_i \text{ 與 } \boldsymbol{\mu}_i \text{ 均不受任何限制。} \end{cases}$$

仿照命題 4.1 的方式，異方差性之 GLRDA 的目標函式可寫成：

$$\log J_{\text{GLRDA}}^{\text{heter}}(\boldsymbol{\Theta}) = \sup \log L_{H_0^{\text{heter}}}(\boldsymbol{\Theta}) - \sup \log L_{H_1^{\text{heter}}}(\boldsymbol{\Theta}) \tag{4.15}$$

而 $\sup \log L_{H_0^{\text{heter}}}(\boldsymbol{\Theta})$ 和 $\sup \log L_{H_1^{\text{heter}}}(\boldsymbol{\Theta})$ 可被分別進一步表示為

$$\begin{aligned} &\sup \log L_{H_0^{\text{heter}}}(\boldsymbol{\Theta}) = \max \log p(\mathbf{x}_1^N, \boldsymbol{\mu}, \{\boldsymbol{\Sigma}_i\}, \boldsymbol{\Theta}) \\ &= \max\left(g(N,d) - \sum_{i=1}^{C} \frac{n_i}{2} \left((\widetilde{\mathbf{m}}_i - \widetilde{\boldsymbol{\mu}})^T \widetilde{\boldsymbol{\Sigma}}_i^{-1} (\widetilde{\mathbf{m}}_i - \widetilde{\boldsymbol{\mu}}) + \text{trace}(\widetilde{\boldsymbol{\Sigma}}_i^{-1} \widetilde{\mathbf{S}}_i) + \log |\widetilde{\boldsymbol{\Sigma}}_i| \right) \right) \end{aligned} \tag{4.16}$$

$$\begin{aligned} &\sup \log L_{H_1^{\text{heter}}}(\boldsymbol{\Theta}) = \max \log p(\mathbf{x}_1^N, \{\boldsymbol{\mu}_i\}, \{\boldsymbol{\Sigma}_i\}, \boldsymbol{\Theta}) \\ &= \max\left(g(N,d) - \sum_{i=1}^{C} \frac{n_i}{2} \left((\widetilde{\mathbf{m}}_i - \widetilde{\boldsymbol{\mu}}_i)^T \widetilde{\boldsymbol{\Sigma}}_i^{-1} (\widetilde{\mathbf{m}}_i - \widetilde{\boldsymbol{\mu}}_i) + \text{trace}(\widetilde{\boldsymbol{\Sigma}}_i^{-1} \widetilde{\mathbf{S}}_i) + \log |\widetilde{\boldsymbol{\Sigma}}_i| \right) \right) \end{aligned} \tag{4.17}$$

欲求在假設 H_0^{heter} 下，$\widetilde{\boldsymbol{\mu}}_0^{\text{heter}}$ 和 $\widetilde{\boldsymbol{\Sigma}}_{0,i}^{\text{heter}}$ 的最大相似度估計子(ML estimator)，可將 $\log L_{H_0^{\text{heter}}}(\boldsymbol{\Theta})$ 分別對 $\widetilde{\boldsymbol{\mu}}$ 和 $\widetilde{\boldsymbol{\Sigma}}_i$ 偏微分，並令其為 0，可得：

$$
\begin{aligned}
&\frac{\partial \log_{H_0^{\text{heter}}} L(\boldsymbol{\Theta})}{\partial \tilde{\boldsymbol{\mu}}} \\
&= \frac{\partial\left(-\sum_{i=1}^{C}\frac{n_i}{2}\left((\tilde{\mathbf{m}}_i-\tilde{\boldsymbol{\mu}})^T\tilde{\boldsymbol{\Sigma}}_i^{-1}(\tilde{\mathbf{m}}_i-\tilde{\boldsymbol{\mu}})+\text{trace}(\tilde{\boldsymbol{\Sigma}}_i^{-1}\tilde{\mathbf{S}}_i)+\log|\tilde{\boldsymbol{\Sigma}}_i|\right)\right)}{\partial\tilde{\boldsymbol{\mu}}} \\
&= \frac{\partial\left(-\sum_{i=1}^{C}\frac{n_i}{2}\left((\tilde{\mathbf{m}}_i-\tilde{\boldsymbol{\mu}})^T\tilde{\boldsymbol{\Sigma}}_i^{-1}(\tilde{\mathbf{m}}_i-\tilde{\boldsymbol{\mu}})\right)\right)}{\partial\tilde{\boldsymbol{\mu}}} \\
&= -\sum_{i=1}^{C} n_i\left(\tilde{\boldsymbol{\Sigma}}_i^{-1}(\tilde{\mathbf{m}}_i-\tilde{\boldsymbol{\mu}})\right)=0 \\
&\Rightarrow \sum_{i=1}^{C} n_i\tilde{\boldsymbol{\Sigma}}_i^{-1}\tilde{\boldsymbol{\mu}} = \sum_{i=1}^{C} n_i\tilde{\boldsymbol{\Sigma}}_i^{-1}\tilde{\mathbf{m}}_i \\
&\Rightarrow \tilde{\boldsymbol{\mu}}_0^{\text{heter}} = \left(\sum_{i=1}^{C} n_i\tilde{\boldsymbol{\Sigma}}_i^{-1}\right)^{-1}\sum_{i=1}^{C} n_i\tilde{\boldsymbol{\Sigma}}_i^{-1}\tilde{\mathbf{m}}_i
\end{aligned}
\tag{4.18}
$$

$$
\begin{aligned}
&\frac{\partial \log L_{H_0^{\text{heter}}}(\boldsymbol{\Theta})}{\partial \tilde{\boldsymbol{\Sigma}}_i} \\
&= \frac{\partial\left(-\sum_{i=1}^{C}\frac{n_i}{2}\left((\tilde{\mathbf{m}}_i-\tilde{\boldsymbol{\mu}}_0^{\text{heter}})^T\tilde{\boldsymbol{\Sigma}}_i^{-1}(\tilde{\mathbf{m}}_i-\tilde{\boldsymbol{\mu}}_0^{\text{heter}})+\text{trace}(\tilde{\boldsymbol{\Sigma}}_i^{-1}\tilde{\mathbf{S}}_i)+\log|\tilde{\boldsymbol{\Sigma}}_i|\right)\right)}{\partial\tilde{\boldsymbol{\Sigma}}_i} \\
&= -\frac{n_i}{2}\left(-\tilde{\boldsymbol{\Sigma}}_i^{-1}(\tilde{\mathbf{m}}_i-\tilde{\boldsymbol{\mu}}_0^{\text{heter}})(\tilde{\mathbf{m}}_i-\tilde{\boldsymbol{\mu}}_0^{\text{heter}})^T\tilde{\boldsymbol{\Sigma}}_i^{-1}-\tilde{\boldsymbol{\Sigma}}_i^{-1}\tilde{\mathbf{S}}_i\tilde{\boldsymbol{\Sigma}}_i^{-1}+\tilde{\boldsymbol{\Sigma}}_i^{-1}\right)=0 \\
&\Rightarrow \tilde{\boldsymbol{\Sigma}}_{0,i}^{\text{heter}} = (\tilde{\mathbf{m}}_i-\tilde{\boldsymbol{\mu}}_0^{\text{heter}})(\tilde{\mathbf{m}}_i-\tilde{\boldsymbol{\mu}}_0^{\text{heter}})^T+\tilde{\mathbf{S}}_i = \tilde{\mathbf{B}}_i+\tilde{\mathbf{S}}_i
\end{aligned}
\tag{4.19}
$$

其中，$\tilde{\mathbf{B}}_i = (\tilde{\mathbf{m}}_i-\tilde{\boldsymbol{\mu}}_0^{\text{heter}})(\tilde{\mathbf{m}}_i-\tilde{\boldsymbol{\mu}}_0^{\text{heter}})^T$。將式(4.18)和式(4.19)中估計出的$\tilde{\boldsymbol{\mu}}_0^{\text{heter}}$和$\tilde{\boldsymbol{\Sigma}}_{0,i}^{\text{heter}}$代入式(4.16)，可得在假設$H_0^{\text{heter}}$下的最大對數相似度：

$$
\begin{aligned}
&\sup\log L_{H_0^{\text{heter}}}(\boldsymbol{\Theta}) = \max\log p(\mathbf{x}_1^N,\boldsymbol{\mu},\{\boldsymbol{\Sigma}_i\},\boldsymbol{\Theta}) \\
&= g(N,d)-\sum_{i=1}^{C}\frac{n_i}{2}\begin{pmatrix}(\tilde{\mathbf{m}}_i-\tilde{\boldsymbol{\mu}}_0^{\text{heter}})^T\tilde{\boldsymbol{\Sigma}}_{0,i}^{\text{heter}\,-1}(\tilde{\mathbf{m}}_i-\tilde{\boldsymbol{\mu}}_0^{\text{heter}})+\\ \text{trace}(\tilde{\boldsymbol{\Sigma}}_{0,i}^{\text{heter}\,-1}\tilde{\mathbf{S}}_i)+\log|\tilde{\boldsymbol{\Sigma}}_{0,i}^{\text{heter}}|\end{pmatrix} \\
&= g(N,d)-\sum_{i=1}^{C}\frac{n_i}{2}\begin{pmatrix}(\tilde{\mathbf{m}}_i-\tilde{\boldsymbol{\mu}}_0^{\text{heter}})^T(\tilde{\mathbf{B}}_i+\tilde{\mathbf{S}}_i)^{-1}(\tilde{\mathbf{m}}_i-\tilde{\boldsymbol{\mu}}_0^{\text{heter}})+\\ \text{trace}((\tilde{\mathbf{B}}_i+\tilde{\mathbf{S}}_i)^{-1}\tilde{\mathbf{S}}_i)+\log|\tilde{\mathbf{B}}_i+\tilde{\mathbf{S}}_i|\end{pmatrix} \\
&= g(N,d)-\frac{Nd}{2}-\sum_{i=1}^{C}\frac{n_i}{2}\log|\tilde{\mathbf{B}}_i+\tilde{\mathbf{S}}_i|
\end{aligned}
\tag{4.20}
$$

欲求在假設 H_1^{heter} 下，$\widetilde{\boldsymbol{\mu}}_1^{\text{heter}}$ 和 $\widetilde{\boldsymbol{\Sigma}}_{1,i}^{\text{heter}}$ 的最大相似度估計子(ML estimator)，可將 $\log L_{H_1^{\text{heter}}}(\boldsymbol{\Theta})$ 分別對 $\widetilde{\boldsymbol{\mu}}_i$ 和 $\widetilde{\boldsymbol{\Sigma}}_i$ 偏微分，並令其為 0，可得：

$$
\begin{aligned}
&\frac{\partial \log L_{H_1^{\text{heter}}}(\boldsymbol{\Theta})}{\partial \widetilde{\boldsymbol{\mu}}_i} \\
&= \frac{\partial\left(-\sum_{i=1}^{C}\frac{n_i}{2}\left((\widetilde{\mathbf{m}}_i-\widetilde{\boldsymbol{\mu}}_i)^T\widetilde{\boldsymbol{\Sigma}}_i^{-1}(\widetilde{\mathbf{m}}_i-\widetilde{\boldsymbol{\mu}}_i)+\text{trace}(\widetilde{\boldsymbol{\Sigma}}_i^{-1}\widetilde{\mathbf{S}}_i)+\log|\widetilde{\boldsymbol{\Sigma}}_i|\right)\right)}{\partial \widetilde{\boldsymbol{\mu}}_i} \\
&= \frac{\partial\left(-\frac{n_i}{2}\left((\widetilde{\mathbf{m}}_i-\widetilde{\boldsymbol{\mu}}_i)^T\widetilde{\boldsymbol{\Sigma}}_i^{-1}(\widetilde{\mathbf{m}}_i-\widetilde{\boldsymbol{\mu}}_i)\right)\right)}{\partial \widetilde{\boldsymbol{\mu}}_i} \\
&= n_i\widetilde{\boldsymbol{\Sigma}}_i^{-1}(\widetilde{\mathbf{m}}_i-\widetilde{\boldsymbol{\mu}}_i)=0 \\
&\Rightarrow \widetilde{\boldsymbol{\mu}}_{1,i}^{\text{heter}}=\widetilde{\mathbf{m}}_i
\end{aligned}
\tag{4.21}
$$

$$
\begin{aligned}
&\frac{\partial \log L_{H_1^{\text{heter}}}(\boldsymbol{\Theta})}{\partial \widetilde{\boldsymbol{\Sigma}}_i} \\
&= \frac{\partial\left(-\sum_{i=1}^{C}\frac{n_i}{2}\left((\widetilde{\mathbf{m}}_i-\widetilde{\mathbf{m}}_i)^T\widetilde{\boldsymbol{\Sigma}}_i^{-1}(\widetilde{\mathbf{m}}_i-\widetilde{\mathbf{m}}_i)+\text{trace}(\widetilde{\boldsymbol{\Sigma}}_i^{-1}\widetilde{\mathbf{S}}_i)+\log|\widetilde{\boldsymbol{\Sigma}}_i|\right)\right)}{\partial \widetilde{\boldsymbol{\Sigma}}_i} \\
&= -\frac{n_i}{2}\left(-\widetilde{\boldsymbol{\Sigma}}_i^{-1}\widetilde{\mathbf{S}}_i\widetilde{\boldsymbol{\Sigma}}_i^{-1}+\widetilde{\boldsymbol{\Sigma}}_i^{-1}\right)=0 \\
&\Rightarrow \widetilde{\boldsymbol{\Sigma}}_{1,i}^{\text{heter}}=\widetilde{\mathbf{S}}_i
\end{aligned}
\tag{4.22}
$$

將式(4.21)和式(4.22)中估計出的 $\widetilde{\boldsymbol{\mu}}_1^{\text{heter}}$ 和 $\widetilde{\boldsymbol{\Sigma}}_{1,i}^{\text{heter}}$ 代入式(4.17)，可得在假設 H_1^{heter} 下的最大對數相似度：

$$
\begin{aligned}
&\sup\log L_{H_1^{\text{heter}}}(\boldsymbol{\Theta})=\max\log p(\mathbf{x}_1^N,\{\boldsymbol{\mu}_i\},\{\boldsymbol{\Sigma}_i\},\boldsymbol{\Theta}) \\
&= g(N,d)-\sum_{i=1}^{C}\frac{n_i}{2}\begin{pmatrix}(\widetilde{\mathbf{m}}_i-\widetilde{\mathbf{m}}_i)^T\widetilde{\mathbf{S}}_i^{-1}(\widetilde{\mathbf{m}}_i-\widetilde{\mathbf{m}}_i)+\\ \text{trace}(\widetilde{\mathbf{S}}_i^{-1}\widetilde{\mathbf{S}}_i)+\log|\widetilde{\mathbf{S}}_i|\end{pmatrix} \\
&= g(N,d)-\sum_{i=1}^{C}\frac{n_i}{2}\left(d+\log|\widetilde{\mathbf{S}}_i|\right) \\
&= g(N,d)-\frac{Nd}{2}-\sum_{i=1}^{C}\frac{n_i}{2}\log|\widetilde{\mathbf{S}}_i|
\end{aligned}
\tag{4.23}
$$

最後，將式(4.20)與式(4.23)代入式(4.15)，可得：

$$
\begin{aligned}
&\log J_{\mathrm{GLRDA}}^{\mathrm{heter}}(\boldsymbol{\Theta}) \\
&= \left(g(N,d) - \frac{Nd}{2} - \sum_{i=1}^{C} \frac{n_i}{2} \log|\widetilde{\mathbf{B}}_i + \widetilde{\mathbf{S}}_i| \right) - \left(g(N,d) - \frac{Nd}{2} - \sum_{i=1}^{C} \frac{n_i}{2} \log|\widetilde{\mathbf{S}}_i| \right) \\
&= -\sum_{i=1}^{C} \frac{n_i}{2} (\log|\widetilde{\mathbf{B}}_i + \widetilde{\mathbf{S}}_i| - \log|\widetilde{\mathbf{S}}_i|) \\
&= -\sum_{i=1}^{C} \frac{n_i}{2} \log|\mathbf{I}_{(p\times p)} + \widetilde{\mathbf{S}}_i^{-1}\widetilde{\mathbf{B}}_i| \\
&= -\sum_{i=1}^{C} \frac{n_i}{2} \log|\mathbf{I}_{(p\times p)} + \widetilde{\mathbf{S}}_i^{-1}(\widetilde{\mathbf{m}}_i - \widetilde{\boldsymbol{\mu}}_0^{\mathrm{heter}})(\widetilde{\mathbf{m}}_i - \widetilde{\boldsymbol{\mu}}_0^{\mathrm{heter}})^T| \\
&= -\sum_{i=1}^{C} \frac{n_i}{2} \log\left(1 + (\widetilde{\mathbf{m}}_i - \widetilde{\boldsymbol{\mu}}_0^{\mathrm{heter}})^T \widetilde{\mathbf{S}}_i^{-1}(\widetilde{\mathbf{m}}_i - \widetilde{\boldsymbol{\mu}}_0^{\mathrm{heter}})\right)
\end{aligned}
\tag{4.24}
$$

值得一提的是，在式(4.24)中，$\widetilde{\boldsymbol{\mu}}_0^{\mathrm{heter}}$ 中含有未知形式的 $\widetilde{\boldsymbol{\Sigma}}_i$，因此在實作上，我們只能先令 $\widetilde{\boldsymbol{\Sigma}}_i = \widetilde{\mathbf{S}}_i$，得到 $\widetilde{\boldsymbol{\mu}}_0^{\mathrm{heter}}$ 的近似估計量(estimate)。而在原始空間中，參照式(4.18)，此近似估計量可被還原成

$$
\boldsymbol{\mu}_0^{\mathrm{heter}} = \left(\sum_{i=1}^{C} n_i \mathbf{S}_i^{-1} \right)^{-1} \sum_{i=1}^{C} n_i \mathbf{S}_i^{-1} \mathbf{m}_i \tag{4.25}
$$

因此，我們可得到異方差性的 GLRDA 目標函式：

$$
G_{\mathrm{H}}(\boldsymbol{\Theta}) = -\sum_{i=1}^{C} \frac{n_i}{2} \log\left(1 + (\boldsymbol{\Theta}^T\mathbf{m}_i - \boldsymbol{\Theta}^T\boldsymbol{\mu}_0^{\mathrm{heter}})^T (\boldsymbol{\Theta}^T\mathbf{S}_i\boldsymbol{\Theta})^{-1} (\boldsymbol{\Theta}^T\mathbf{m}_i - \boldsymbol{\Theta}^T\boldsymbol{\mu}_0^{\mathrm{heter}})\right) \tag{4.26}
$$

為了使用梯度下降等遞迴式的最佳化技術求解 $\boldsymbol{\Theta}$，式(4.25)對 $\boldsymbol{\Theta}$ 的一階偏導數可寫成：

$$
\frac{\partial G_{\mathrm{H-GLRDA}}(\boldsymbol{\Theta})}{\partial \boldsymbol{\Theta}} = -\sum_{i=1}^{C} n_i \frac{(-\mathbf{S}_i\boldsymbol{\Theta}\widetilde{\mathbf{S}}_i^{-1}\widetilde{\mathbf{B}}_i + \mathbf{B}_i\boldsymbol{\Theta})\widetilde{\mathbf{S}}_i^{-1}}{1 + \mathrm{trace}(\widetilde{\mathbf{S}}_i^{-1}\widetilde{\mathbf{B}}_i)} \tag{4.27}
$$

其中，$\mathbf{B}_i = (\mathbf{m}_i - \boldsymbol{\mu}_0^{\mathrm{heter}})(\mathbf{m}_i - \boldsymbol{\mu}_0^{\mathrm{heter}})^T$，$\widetilde{\mathbf{B}}_i = \boldsymbol{\Theta}^T\mathbf{B}_i\boldsymbol{\Theta}$，$\widetilde{\mathbf{S}}_i = \boldsymbol{\Theta}^T\mathbf{S}_i\boldsymbol{\Theta}$。

表 4.1　GLRDA 在不同假設下的統計量歸納表

統計假設	期望值向量 估計子	共變異矩陣 估計子	含重要項之 最大對數相似度
$H_0^{\text{homo}}\begin{cases}\boldsymbol{\Sigma}_i=\boldsymbol{\Sigma}\\ \boldsymbol{\mu}_i=\boldsymbol{\mu}\end{cases}$	$\tilde{\mathbf{m}}$	$\tilde{\mathbf{S}}_T$	$-\frac{N}{2}\log\lvert\tilde{\mathbf{S}}_T\rvert$
$H_1^{\text{homo}}\begin{cases}\boldsymbol{\Sigma}_i=\boldsymbol{\Sigma}\\ \boldsymbol{\mu}_i:\text{無限制}\end{cases}$	$\tilde{\mathbf{m}}_i$	$\tilde{\mathbf{S}}_W$	$-\frac{N}{2}\log\lvert\tilde{\mathbf{S}}_W\rvert$
$H_0^{\text{heter}}\begin{cases}\boldsymbol{\Sigma}_i:\text{無限制}\\ \boldsymbol{\mu}_i=\boldsymbol{\mu}\end{cases}$	$\tilde{\boldsymbol{\mu}}_0^{\text{heter}}$	$\tilde{\mathbf{B}}_i+\tilde{\mathbf{S}}_i$	$-\frac{1}{2}\sum_{i=1}^{C}n_i\log\lvert\tilde{\mathbf{B}}_i+\tilde{\mathbf{S}}_i\rvert$
$H_1^{\text{heter}}\begin{cases}\boldsymbol{\Sigma}_i:\text{無限制}\\ \boldsymbol{\mu}_i:\text{無限制}\end{cases}$	$\tilde{\mathbf{m}}_i$	$\tilde{\mathbf{S}}_i$	$-\frac{1}{2}\sum_{i=1}^{C}n_i\log\lvert\tilde{\mathbf{S}}_i\rvert$

4.2.3　討論與比較

表 4.1 歸納了 GLRDA 在不同假設下的統計量。GLRDA 是一個較大之相似度比率的框架，它不僅是 LDA 的普遍化形式（LDA 轉換矩陣可由 H_0^{homo} 與 H_1^{homo} 的相似度比率得到），以下命題 4.2 亦證明了它也是 HLDA 的普遍化形式。

命題 4.2：異方差線性鑑別分析(HLDA)轉換矩陣可由 GLRDA 中的 H_0^{homo} 與 H_1^{heter} 的相似度比率得到。也就是說，GLRDA 是 HLDA 的普遍化形式。

證明：HLDA 目標函式為：

$$J_{\text{HLDA}}(\boldsymbol{\Theta})=-\frac{N}{2}\log\lvert\boldsymbol{\Theta}_{(n-d)}^T\mathbf{S}_T\boldsymbol{\Theta}_{(n-d)}\rvert-\sum_{i=1}^{C}\frac{n_i}{2}\log\lvert\boldsymbol{\Theta}_d^T\mathbf{S}_i\boldsymbol{\Theta}_d\rvert+N\log\lvert\boldsymbol{\Theta}\rvert \qquad (4.28)$$

因為在此 $\boldsymbol{\Theta}=\boldsymbol{\Theta}_{(n\times n)}=[\boldsymbol{\Theta}_d,\boldsymbol{\Theta}_{(n-d)}]$，我們可證明[64]：

$$
\begin{aligned}
&|\boldsymbol{\Theta}^T\mathbf{S}_T\boldsymbol{\Theta}| = |\boldsymbol{\Theta}_d^T\mathbf{S}_T\boldsymbol{\Theta}_d| \times |\boldsymbol{\Theta}_{(n-d)}^T\mathbf{S}_T\boldsymbol{\Theta}_{(n-d)}| \\
&\Rightarrow \log|\boldsymbol{\Theta}^T\mathbf{S}_T\boldsymbol{\Theta}| = \log|\boldsymbol{\Theta}_d^T\mathbf{S}_T\boldsymbol{\Theta}_d| + \log|\boldsymbol{\Theta}_{(n-d)}^T\mathbf{S}_T\boldsymbol{\Theta}_{(n-d)}| \\
&\Rightarrow \log|\boldsymbol{\Theta}_{(n-d)}^T\mathbf{S}_T\boldsymbol{\Theta}_{(n-d)}| = \log|\boldsymbol{\Theta}^T\mathbf{S}_T\boldsymbol{\Theta}| - \log|\boldsymbol{\Theta}_d^T\mathbf{S}_T\boldsymbol{\Theta}_d|
\end{aligned}
\tag{4.29}
$$

$$
\begin{aligned}
&N\log|\boldsymbol{\Theta}| - \frac{N}{2}\log|\boldsymbol{\Theta}^T\mathbf{S}_T\boldsymbol{\Theta}| \\
&= N\log|\boldsymbol{\Theta}| - \frac{N}{2}\log|\boldsymbol{\Theta}| - \frac{N}{2}\log|\mathbf{S}_T| - \frac{N}{2}\log|\boldsymbol{\Theta}| \\
&= -\frac{N}{2}\log|\mathbf{S}_T|
\end{aligned}
\tag{4.30}
$$

將式(4.29)代入式(4.28)，且考慮式(4.30)所推導出的結果，我們可得出新的 HLDA 目標函式：

$$
\begin{aligned}
&J_{HLDA}(\boldsymbol{\Theta}) \\
&= -\frac{N}{2}(\log|\boldsymbol{\Theta}^T\mathbf{S}_T\boldsymbol{\Theta}| - \log|\boldsymbol{\Theta}_d^T\mathbf{S}_T\boldsymbol{\Theta}_d|) - \sum_{i=1}^{C}\frac{n_i}{2}\log|\boldsymbol{\Theta}_d^T\mathbf{S}_i\boldsymbol{\Theta}_d| + N\log|\boldsymbol{\Theta}| \\
&= \frac{N}{2}\log|\boldsymbol{\Theta}_d^T\mathbf{S}_T\boldsymbol{\Theta}_d| - \sum_{i=1}^{C}\frac{n_i}{2}\log|\boldsymbol{\Theta}_d^T\mathbf{S}_i\boldsymbol{\Theta}_d| + \left(N\log|\boldsymbol{\Theta}| - \frac{N}{2}\log|\boldsymbol{\Theta}^T\mathbf{S}_T\boldsymbol{\Theta}|\right) \\
&= \underbrace{-\sum_{i=1}^{C}\frac{n_i}{2}\log|\boldsymbol{\Theta}_d^T\mathbf{S}_i\boldsymbol{\Theta}_d|}_{\sup\log L_{H_1^{\text{heter}}}(\boldsymbol{\Theta})} \underbrace{-\left(-\frac{N}{2}\log|\boldsymbol{\Theta}_d^T\mathbf{S}_T\boldsymbol{\Theta}_d|\right)}_{\sup\log L_{H_0^{\text{homo}}}(\boldsymbol{\Theta})} - \frac{N}{2}\log|\mathbf{S}_T|
\end{aligned}
\tag{4.31}
$$

很明顯地，若不考慮常數 $(-N/2)\log|\mathbf{S}_T|$，則式(4.31)等同於 GLRDA 中，H_0^{homo} 與 H_1^{heter} 含重要項之最大對數相似度比率（見表 4.1）。 ■

由命題 4.1 可看出，HLDA 與異方差性之 GLRDA 的主要差別在於虛無假設 H_0 的設定，HLDA 的虛無假設 H_0^{homo} 較為嚴格，相對地要使它盡可能不發生的難度也較低。反之，異方差性之 GLRDA 的虛無假設 H_0^{heter} 所產生的參數空間就比較大，使得找出的投影子空間在類別鑑別性上就較為強健(robust)，這一點我們可以由實驗結果看出。

表 4.2　MATBN 訓練語料之音素辨識中前 10 組最易混淆之音素模型配對

K	類別（音素）配對　（RCD 模型）		錯誤音框數
1	in　(ㄧㄣ)	ing　(ㄧㄥ)	66,353
2	an　(ㄢ)	eng　(ㄥ)	42,550
3	i　(O ㄧ)	sil	31,796
4	u　(O ㄨ)	sil	29,082
5	sic_e　(ㄜ)	sil	26,134
6	sic_i　(ㄧ)	sil	25,709
7	ing　(ㄧㄥ)	sil	21,629
8	g_u　(ㄍㄨ)	sil	19,197
9	ian　(ㄧㄢ)	ie　(ㄧㄝ)	17,212
10	sic_i　(ㄧ)	i　(O ㄧ)	17,022

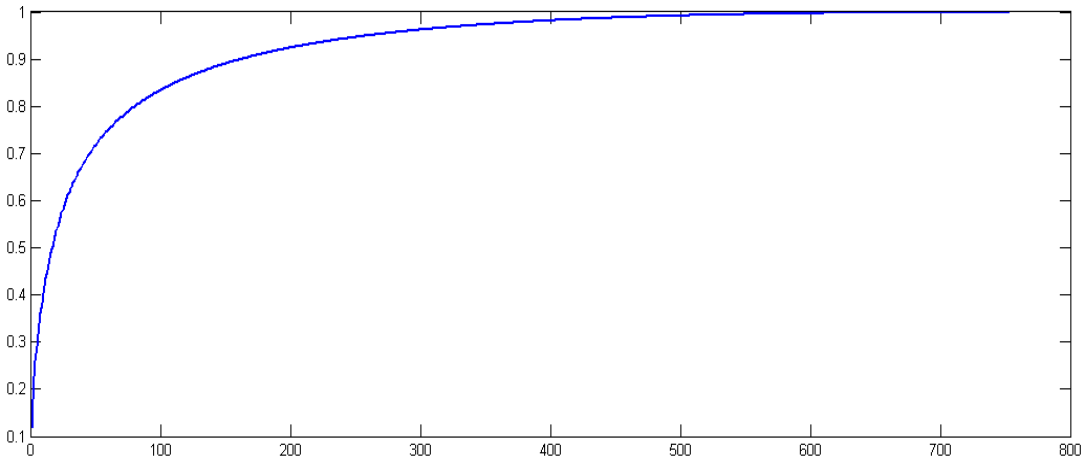


圖 4.1　前 K 組易於混淆之類別配對與累積錯誤音框比率圖
（橫軸為 $K/10$，縱軸為累積錯誤音框比率）

4.3　混淆資訊的延伸

由異方差性之 GLRDA 的虛無假設 H_0^{heter} 來看，它設想每一類別母體的期望值向量在投影子空間中幾乎重疊在一起，但事實上這個假設的設定未必精確。在使用 LDA 作為聲學特徵擷取方法之 MATBN 訓練語料的音素辨識結果中，我們將所有類別配對，依其混淆程度由大至小排序，可得到前 K 個類別配對，如表 4.2。舉例來說，在表 4.2 中，最易於混淆的類別類對是(in, ing），其錯誤音框數

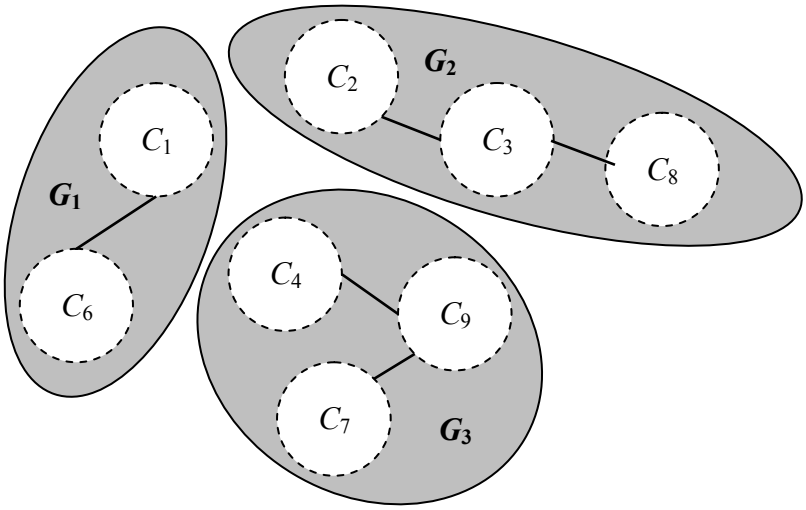


圖 4.2　類別配對與群聚形成示意圖

為 66,353，表示原本屬於音素 in 和 ing，卻分別被辨識器錯分至音素 ing 和 in 的音框總數。也就是說，以經驗資訊產生的事實來看，我們最不願意見到的假設，就是使這些易於混淆之類別配對的期望值向量幾乎重疊在一起，而非更廣泛的假設全體類別母體之期望值向量重疊，因為這些易於混淆之類別配對才是全體錯誤的主要來源。如圖 4.1 中，我們可以發現約前 10%易於混淆之類別配對主導了約 80%的錯誤音框總數。

但是，若我們只考慮這些類別配對之期望值向量重疊的假設，則會發生以下情況：若類別配對 C_1 與 C_2 為最混淆之配對，則虛無假設可設定為 $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$；而類別配對 C_2 與 C_3 為次混淆之配對，則虛無假設可增加 $\boldsymbol{\mu}_2 = \boldsymbol{\mu}_3$。因此，整個虛無假設可合併為 $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \boldsymbol{\mu}_3$。也就是說，我們必須在類別配對集合中找到所有相關的類別配對以組成混淆群聚(confusable cluster)，如圖 4.2。若我們把所有類別視為圖形(graph)中的點(vertex)，而由易於混淆之類別配對所建立的關係視為兩點之間的邊(edge)，則混淆群聚的產生可被視為尋找圖形(graph)中所有的連通子圖(connected subgraph)。所以，我們可以使用一些圖論中的演算法，如滿水填充演算法(flood fill algorithm)[78]，來解決這個問題。

因此，我們可以將異方差性之 GLRDA 改良成基於混淆資訊之普遍化相似度

比率鑑別分析(confusion information based GLRDA, CI-GLRDA)[51]：令 $G:\{G_k\}$ 為所有根據前 K 組易於混淆之類別配對，並利用滿水填充演算法求出之群聚的集合，則 CI-GLRDA 的虛無假設與對立假設可設定如下：

$$\begin{cases} H_0^{\mathrm{CI}}：每一類別母體 C_i 均呈高斯分佈，且 \boldsymbol{\Sigma}_i 不受任何限制，而若 C_i \in G_{l_i}，則 \boldsymbol{\mu}_i = \boldsymbol{\mu}_{l_i}。 \\ H_1^{\mathrm{CI}}：每一類別母體 C_i 均呈高斯分佈，且 \boldsymbol{\Sigma}_i 與 \boldsymbol{\mu}_i 均不受任何限制。 \end{cases}$$

其中，l_i 為群聚編號，用來標示類別 C_i 所屬的群聚。

相似於 4.2.2 節的最大化相似度估計法，我們可得到 CI-GLRDA 目標函式：

$$G_{\mathrm{CI}}(\boldsymbol{\Theta}) = \sum_{i=1, G_{l_i} \in G}^{C} -\frac{n_i}{2} \log\left(1 + (\boldsymbol{\Theta}^T \mathbf{m}_i - \boldsymbol{\Theta}^T \boldsymbol{\mu}_{l_i}^{\mathrm{CI}})^T (\boldsymbol{\Theta}^T \mathbf{S}_i \boldsymbol{\Theta})^{-1} (\boldsymbol{\Theta}^T \mathbf{m}_i - \boldsymbol{\Theta}^T \boldsymbol{\mu}_{l_i}^{\mathrm{CI}})\right) \quad (4.32)$$

其中，在實作上，對於某混淆群聚母體 G_{l_i} 之 $\boldsymbol{\mu}_{l_i}^{\mathrm{CI}}$ 的估計子可表示為

$$\boldsymbol{\mu}_{l_i}^{\mathrm{CI}} = \left(\sum_{C_i \in G_{l_i}} n_i \mathbf{S}_i^{-1} \right)^{-1} \sum_{C_i \in G_{l_i}} n_i \mathbf{S}_i^{-1} \mathbf{m}_i \quad (4.33)$$

為了最佳化的需要，式(4.32)對 $\boldsymbol{\Theta}$ 的一階偏導數亦可表示成：

$$\frac{\partial G_{\mathrm{CI}}(\boldsymbol{\Theta})}{\partial \boldsymbol{\Theta}} = -\sum_{i=1, G_{l_i} \in G}^{C} n_i \frac{(-\mathbf{S}_i \boldsymbol{\Theta} \widetilde{\mathbf{S}}_i^{-1} \widetilde{\mathbf{B}}_{l_i} + \mathbf{B}_{l_i} \boldsymbol{\Theta}) \widetilde{\mathbf{S}}_i^{-1}}{1 + \mathrm{trace}(\widetilde{\mathbf{S}}_i^{-1} \widetilde{\mathbf{B}}_{l_i})} \quad (4.34)$$

其中，$\mathbf{B}_{l_i} = (\mathbf{m}_i - \boldsymbol{\mu}_{l_i}^{\mathrm{CI}})(\mathbf{m}_i - \boldsymbol{\mu}_{l_i}^{\mathrm{CI}})^T$，$\widetilde{\mathbf{B}}_{l_i} = \boldsymbol{\Theta}^T \mathbf{B}_{l_i} \boldsymbol{\Theta}$，$\widetilde{\mathbf{S}}_i = \boldsymbol{\Theta}^T \mathbf{S}_i \boldsymbol{\Theta}$。

[51] 本論文之後皆以 CI-GLRDA 來簡稱『基於混淆資訊之普遍化相似度比率鑑別分析』。

第 5 章　實驗架構與實驗結果

首先，我們會介紹本論文所使用的語料庫以及台灣師大廣播新聞轉寫系統。之後我們會詳細說明實驗評估方式以及如何將線性特徵轉換應用在多向量輸入（頻域－時域的特徵空間）。最後則是基礎實驗結果，以及本論文所提出新方法的結果。

5.1　實驗語料庫

本論文主要使用的語料庫為 MATBN 中文電視新聞語料[79]，為中央研究院資訊所口語小組(SLG)耗時三年與公共電視台(PTS)合作錄製完成。錄製的對象為公視晚間新聞，其每天的長度皆為一個小時，收錄了 200 天（約 200 小時）的新聞語料，其中包含 2001 年的新聞 30 小時、2002 年 146 小時及 2003 年 24 小時。所有的新聞語料都有正確的人工轉寫以及其它的標註資訊（如音樂、背景雜訊、停頓、語助詞、呼吸、強調語氣、反覆、不適當的發音等），所有的人工轉寫與標注均使用 DGA&LDC 的轉寫器(transcriber)[80]來完成。

每天的新聞約含有二十多則報導，每則報導為一完整主題。除了語音資料，文字語料在其它應用上也有很大的價值（如資訊檢索、文件摘要等）。此新聞語料大致上可分內場及外場兩個部份，內場部分主要為主播(studio anchors)的語料，外場部分則可分為採訪記者(field reporters)與受訪者(interviewees)的語料。在篩選實驗語料時，考量新聞的特性，主播多為同一人所擔任。如表 5.1 所示，葉明蘭主播的語料在本語料庫中約佔了所有主播語料的 85%，這將使得實驗偏向語者相依(speaker-dependent)的環境，加上女性主播約佔了所有主播語料的 94%，也造成了性別相依(gender-dependent)的問題。如果使用主播語料的話，缺少足夠的變

表 5.1　MATBN 主播語料分佈表

語者姓名	性別	句數	語音總長度(秒)	所含語音百分比(%)
余佳璋-主播	男	36	452.20	0.50
林建成-主播	男	427	5,298.10	5.70
某主播	女	1	7.90	0.008
洪蕙竹-主播	女	89	1,407.40	1.50
洪蕙竹-氣象主播	女	155	1,443.60	1.50
徐惠玲-主播	女	225	3,208.20	3.40
馬紹-主播	男	35	465.60	0.50
黃明明-主播	女	175	2,932.60	3.10
葉明蘭-主播	女	5,101	78,584.70	83.60
蘇怡如-氣象主播	女	17	213.80	0.20

異來提供良好的訓練與客觀的評估，故本實驗不採用主播語料。此系統可檢索語句的統計資訊，如語者資訊、語音長度、所含背景雜訊、說話速度及正確轉譯文句等資訊，適合用來分析且定義出實驗的訓練集(training set)與評估集(evaluation set)。目前本研究初步地只選擇外場採訪記者部份作為實驗語料，在未來將會納入受訪者的部份。

本研究所使用外場記者的語料總共約 27 小時，其中 24.5 小時（5774 句，再切成 34,672 個短句供聲學模型訓練之用）做為聲學模型訓練的語料，1 小時（230 句）則為辨識評估的資料，而 1.5 小時（292 句）則為發展集(developing set)，其用途乃在於決定各種方法中，需要調整出的最佳參數。訓練語料由 2001 及 2002 年的新聞中篩選出不含語助詞(particle)的語料，為了建立性別平衡(gender-balanced)的訓練環境，男、女語料分別篩取 12.2 小時。為了準確客觀地評估研究方法，我們採用由中研院從 2003 年語料庫中所選定的外場記者語料作為評估語料，部份語音片段含有語助詞，關於訓練語料及測試語料詳細的資訊可見表 5.2。

而在外場受訪者部份，由於有較多的語助詞出現，如表 5.3 所示，因此如果

表 5.2　外場記者訓練與測試語料分佈表

性別	訓練語料總長(分)	評估語料總長(分)
男生	766.69	21.68
女生	766.79	65.23

表 5.3　語助詞出現次數統計表

語者型別	所含語音百分比(%)	語助詞出現次數(句)	每句平均語助詞出現次數(次)
外場採訪記者	48.69	877	0.07
外場受訪者	29.33	18,991	2.03
內場主播	21.98	771	0.12

直接只採用不含語助詞的外場語料的話，測試資料（只從 2003 年的語料擷取）將會非常的稀少。而訓練資料（從 2001 及 2002 年的語料擷取）為了要顧及性別平衡的因素，也會有訓練資料量不足的問題，所以本論文就沒有使用外場受訪者的語料。另外所有語料的詳細統計資訊可經由台灣師大資工語音實驗室所開發的公視新聞語料檢索系統獲得。此系統可檢索語句的統計資訊，如語者資訊、語音長度、所含背景雜訊、說話速度及正確轉寫文等內容，極為適合用來篩選聲學模型訓練所需的語料。

5.2　臺灣師大之中文大詞彙連續語音辨識系統

以下將分別介紹臺灣師大的中文大詞彙連續語音辨識系統採用的前端處理、聲學模型、詞典建立、語言模型以及詞彙樹複製搜尋等部份。

5.2.1　前端處理

在本論文中使用梅爾倒頻譜係數（MFCCs）作為最基本的語音特徵參數。

在求取梅爾倒頻譜係數時，將語音資料切割成一連串部分重疊的音框，每一個音框由 13 維的梅爾倒頻譜係數加上其一階與二階的時間導數(time derivatives)所形成的 39 維語音特徵向量所組成。其中 13 維的梅爾倒頻譜係數是由 18 個梅爾頻譜上濾波器組(filter banks)的輸出經離散餘弦轉換求得。

5.2.2 聲學模型

聲學模型是採用傳統的連續密度隱藏式馬可夫模型（CDHMM），模型內狀態的轉移情形只有兩種，一種是停留在原狀態，一種是由左至右跳到下一個相鄰的狀態。模型的總數量有 151 個，其中包含了 1 個靜音模型(silence)，112 個聲母模型(INITIAL)，以及 38 個韻母模型(FINAL)。每個模型的狀態數分別為 3 至 6 個不等，每個狀態皆為高斯混合分佈，其中每個高斯混合分佈的分佈個數分別為 1 至 128 個不等。此外，聲母和韻母共有 403 種不同的音節組合。

5.2.3 詞典建立與語言模型訓練

在中文裡約有 7,000 個單字詞，新詞可由此 7,000 個單字詞合併產生，則可根據字詞在語料中的統計特性，以自動化的方式產生新的複合詞(compound words)。新增複合詞的自動產生方式如下面所述：對於語料中任意相鄰的兩個詞 (w_i, w_j)，可以分別計算它們的前二連(forward bigram)機率 $P_f(w_j \mid w_i)$，與後二連(backward bigram)機率 $P_f(w_i \mid w_j)$，並以前後二連(forward and backward bigrams)的機率幾何平均(geometric average) $FB(w_i, w_j) = \sqrt{P_f(w_j \mid w_i) P_b(w_i \mid w_j)}$，作為 (w_i, w_j) 是否合併的依據。文字語料先經由一個含有一至四字詞約六萬八千個詞的詞典來斷詞，然後利用上述的公式，經數次的迭代以及不同的基準閥值(threshold)設定，產生約五千個二至十字詞的複合詞，使得最後的語音辨識詞典約含有七萬二千個

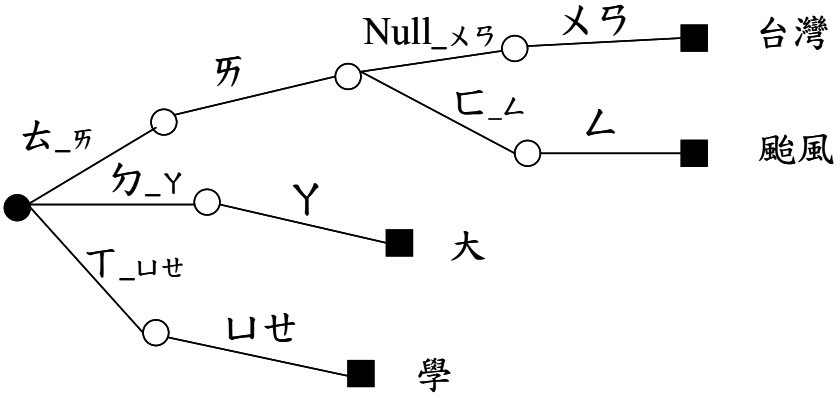


圖 5.1　詞彙樹範例

一至十字詞。本系統使用詞二連以及詞三連語言模型(word bigram and trigram language models)，並以從中央通訊社(Central News Agency，CNA)2001 與 2002 年所收集到的約一億七千萬個中文字語料作為背景語言模型訓練時的訓練資料(LDC)。本論文中的語言模型使用 Katz 語言模型平滑技術，語言模型訓練工具採用 SRI Language Modeling Toolkit (SRILM)[81]。

5.2.4　詞彙樹複製搜尋

本系統是採用由左至右(left-to-right)且音框同步(frame synchronous)的詞彙樹複製搜尋方式[82]。詞彙樹的架構如圖 5.1（取自[83]）所示，樹中的每個分枝(arc)代表一個聲母(INITIAL)、韻母(FINAL)或靜音(silence)模型。由樹的根節點（圖 5.1 的圓形實心點）走到樹的葉節點（圖 5.1 的方形實心點）的某一條完整路徑代表走完一個或一組發音相同的詞。而路徑上的每一個分枝正好對應到這些詞的一組聲學模型。詞彙樹複製搜尋在執行時，每個音框會同時存在數棵詞彙樹複製(tree copies)，而每棵詞彙樹代表來自不同的語言歷史或限制(language model history or constraint)。在同一棵詞彙樹裡，會進行隱藏式馬可夫模型狀態層次(state level)維特比(Viterbi)動態規劃搜尋。在詞彙樹搜尋中，只有在走到葉節點時，才能確定所搜尋的一個完整詞為何。另外，當具有相同語言模型歷史之不同詞彙樹

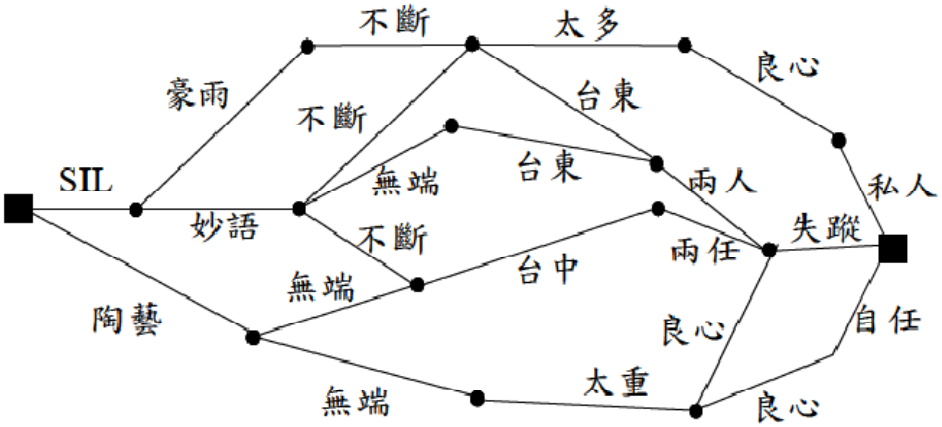


圖 5.2　詞圖範例

分別都已經走到自己所屬那棵樹的葉節點時，則會進行結合(recombination)，只保留其中分數最大者，並針對留下來的詞彙樹繼續執行詞彙樹複製搜尋。然而，真正在實作時，並不需要產生如此多的詞彙樹，僅需建立一棵詞彙樹作為參考之用，並分別記錄搜尋時存活下來之隱藏式馬可夫模型狀態節點的相關資訊（如到目前為此所累積的分數及前一狀態為何）。

另外一方面，由於存活的狀態節點通常會隨著音框數呈指數倍成長，因而必須以光束剪裁(beam pruning)技術將分數較低的狀態節點做剪裁的動作。在對每個狀態節點執行光束剪裁時，會依此節點所有可拜訪的葉節點之最大單連語言模型往前觀測分數(unigram language model look-ahead score)[82]及聲學往前觀測分數(Acoustic Look-ahead Score)[84]做為剪裁與否的依據。

此外，在每個音框，利用存活的詞彙樹複製樹其葉節點（代表可能的候選詞）所儲存的語言模型歷史、開始音框、結束音框及其聲學解碼的分數等資訊建立詞圖(word graph)，如圖 5.2（取自[85]），而後使用更高階的語言模型，如詞三連(trigram)或詞四連(fourgram)語言模型，抑或採用更複雜的聲學模型，如三連音素(triphone)，進行詞圖重評分(word graph rescoring)搜尋[86]，找出最佳的詞序列。在本論文中，詞彙樹複製搜尋階段是採用詞二連語言模型，詞圖搜尋階段則是使

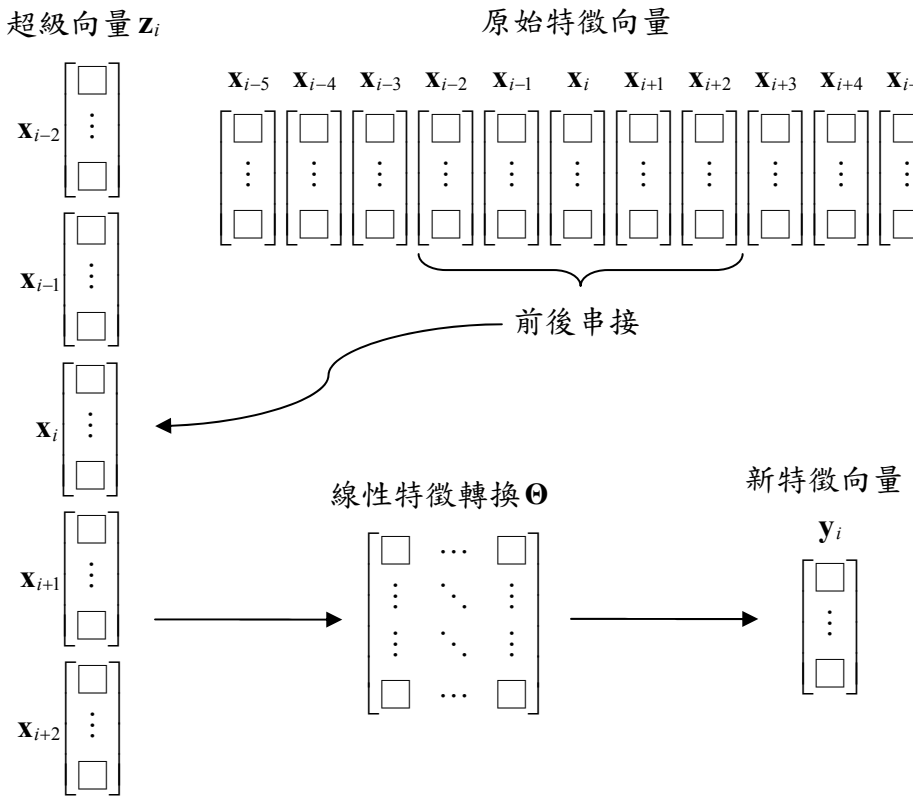


圖 5.3　多向量輸入（頻域－時域特徵擷取）示意圖

用詞三連語言模型。

5.2.5　實驗評估方式

此評估法則是採用美國國家標準與技術中心(National Institute of Standards and Technology, NIST)所訂立的評估標準來進行正確答案的詞序列與辨識詞序列的比較。此評估標準需要使用動態規畫(dynamic programming)來做詞序列比對。由於在中文會有斷詞不一致的問題，因此在本論文的實驗中主要是以字為比對單位。令 H 為正確答案詞序列與辨識詞序列比對後相同(match)的字的個數、I 為辨識詞序列多餘插入(insertion)的字的個數、N 為正確答案詞序列的字的個數，則語音辨識系統的正確率(accuracy)的計算方式為 $\frac{H-I}{N}\times 100\%$，而錯誤率(error rate)則為 1-正確率。在進行動態規畫比對時，替代(substituion)錯誤的懲罰權重(penalty

weight)為 10 分，插入及刪除的權重則皆為 7 分。因為中文是以字(character)為單位，所以在以下的實驗數據中，都是以字正確率(character accuracy)來呈現實驗結果。

5.2.6 多向量輸入（頻域－時域特徵擷取）

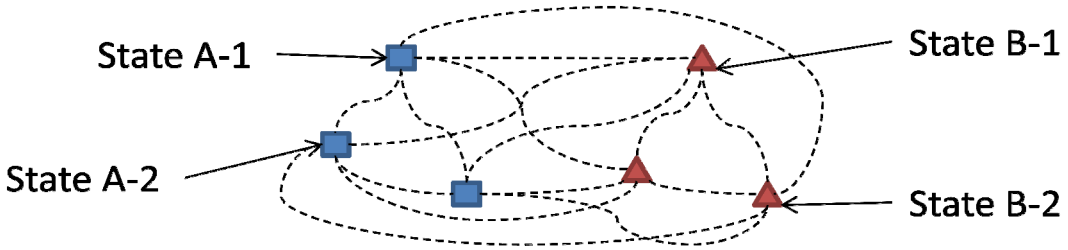
多向量輸入（頻域－時域特徵擷取）是希望藉由線性特徵轉換來同時擷取頻域上與時域上最要或具鑑別力的特徵向量[10]。如圖 5.3 所示，首先由特徵向量 $\mathbf{x}_i$ 本身加上前後各取 k 個特徵向量形成超級向量 $\mathbf{z}_i$ (feature super-vector)，此處的 k 為 2。超級特徵向量 $\mathbf{z}_i$ 經由基底矩陣 $\mathbf{\Theta}$ 線性轉換後可得新特徵向量 $\mathbf{y}_i$，其中的 $\mathbf{\Theta}$ 就是由前面章節所敘述之線性特徵轉換方法求得。在本論文的實驗中，資料相關線性特徵轉換皆是應用在多向量輸入（頻域－時域特徵擷取），並且我們將前後串接的數目 k 定為 4，也就是原始特徵向量維度會從梅爾濾波器組產生的 18 維，變成超級向量 $18\times(4\times2+1)=162$ 維。而目標維度則設定為 39 維($d=39$)，其目的則在於使我們能在子空間維度固定的情況下，定性地比較各種方法的優劣。

5.3 實驗結果

由於本論文中所提到之特徵轉換的方法，其求取出之轉換矩陣並不會使得各個類別的共變異矩陣為對角化，而會造成後端隱藏式馬可夫模型(HMM)參數的估計失真，所以我們嘗試在這些方法後各自加上最大化相似度線性轉換(maximum likelihood linear transformation, MLLT) $\mathbf{A}_{\text{MLLT}}\in\Re^{d\times d}$ [57]，其滿足了

$$\mathbf{A}_{\text{MLLT}}=\arg\max_{\mathbf{A}}\sum_{i=1}^{C}-\frac{n_i}{2}|\operatorname{diag}(\mathbf{A}^T\widetilde{\mathbf{S}}_i\mathbf{A})|+N\log|\mathbf{A}| \tag{5.1}$$

(a)



(b)

圖 5.4　以狀態和音素為類別定義的示意圖

5.3.1　關於類別定義的進一步討論

在本論文中，類別的最小單位則是以隱藏式馬可夫模型(HMMs)中的狀態(state)為主[31]，並經由一個辨識效果較高的系統針對每一訓練語句進行強制校準(forced alignment)，從而產生語句中的類別（音素(phone)和狀態）分界。而關於此類別單位的定義亦已有其它研究涉入[52]，它對線性鑑別分析(LDA)的重要性在於，不同的類別單位對於所有資料會形成不同的類別間散佈矩陣$\mathbf{S}_B$，因而決定出不同的幾何分離度。由歷年來許多語音辨識的實驗結果得知，以狀態作為類別單位的效果較音素為佳[31]。以下我們僅就最廣泛使用的兩種類別單位：狀態(state)與音素(phone)分別討論之。

[52] Haeb-Umbach 等人認爲，最適當的類別定義並不直覺，就算系統是以音素爲基礎的辨識器，我們也不能妄加論斷音素是最佳的分類單位，見[31]。而由 Haeb-Umbach 等人實驗結果可看出，以狀態和隱藏馬可夫模型(HMM)中的混和(mixture)爲類別單位的實驗結果十分接近，也都較以音素爲單位的結果好。在本論文中，由於實驗語料有限，所以暫不考慮以混合爲類別單位的情況。

表 5.4　LDA 在不同類別定義與子空間限制下之自由音節辨識正確率(%)

LDA		無 MLLT	有 MLLT
單範 (Normal)	狀態－狀態 （圖 5.4 (a)）	50.79	53.8
	狀態－音素 （圖 5.4 (b)）	50.73	53.4
	音素－音素	46.55	49.37
共軛正交 (Conjugate orthogonal)	狀態－狀態 （圖 5.4 (a)）	51.07	54.17
	狀態－音素 （圖 5.4 (b)）	50.99	53.21
	音素－音素	47.05	49.28

如圖 5.4 (a)，若我們以狀態作為類別的唯一單位，則類別間的幾何分離度將由所有狀態間的關係（以虛線表示）決定。但是，就語音辨識來說，我們可以合理地認為，最終的辨識錯誤率與音素之間的分離度較有關係，而在同一音素之內的狀態，並無必要考慮其分離度。因此，我們設計了另一種類別的關係，如圖 5.4 (b)所示，狀態間的關係僅存在於相異音素之間。而如 2.3.1 節所述，LDA 所決定的投影子空間並非唯一，我們在此列舉了兩種對此子空間加以限制的方法，一是單範(normal)空間，意指轉換矩陣 $\boldsymbol{\Theta} \in \Re^{n\times d}$ 之行向量的長度(norm)均為 1，即 $\|\boldsymbol{\theta}_i\| = 1,\ \forall i$；二是共軛正交(Conjugate orthogonal)空間，意指轉換矩陣 $\boldsymbol{\Theta} \in \Re^{n\times d}$ 之任兩個行向量均滿足

$$\boldsymbol{\theta}_i^T \mathbf{S}_W \boldsymbol{\theta}_j = \begin{cases} 0, & \text{if } i \neq j \\ 1, & \text{if } i = j \end{cases} \tag{5.2}$$

表 5.4 顯示了 LDA 在不同類別定義與子空間限制下之自由音節辨識(free syllable decoding)正確率，其中，『狀態－狀態』表示以『狀態』為最小的分類單位，而類別間散佈矩陣 $\mathbf{S}_B$ 則考慮所有相異的『狀態』，以此類推。首先，我們可以發現，在上述提到的兩種子空間限制下，其正確率之差距均不大，表示屬於

表 5.5　基本特徵擷取方法在大詞彙連續語音辨識之正確率(%)

基本方法	無 MLLT 之正確率			有 MLLT 之正確率		
	音節	字	詞	音節	字	詞
MFCC	76.09	**67.84**	58.62	75.06	**67.29**	58.35
LDA	76.81	**68.97**	60.10	79.06	**71.56**	62.74
HLDA	75.07	**66.96**	58.07	79.51	**71.87**	63.13
HDA	76.60	**68.51**	59.52	79.62	**72.07**	63.23
PLDA	77.23	**69.35**	60.5	79.90	**72.38**	63.38
PWLDA	77.41	**69.61**	60.68	79.36	**72.07**	63.39
aPTAC	77.06	**69.10**	60.08	79.05	**71.46**	62.79

同一同構空間的投影子空間對於語音辨識結果的影響有限。再者，我們也可以看出圖 5.4 (a)與(b)所表示的兩種結果差距也不大，二者均優於以『音素』為最小分類單位的實驗結果，這些都說明了無論是考慮狀態間還是音素間的分離度，以『狀態』為最小分類單位都是使辨識率提高的關鍵因素。

5.3.2　基礎實驗結果

本小節主要是比較幾種常見的語音特徵在相同的聲學模型訓練方法下，對外場記者語料辨識字正確率的影響。我們所比較的語音特徵有七種，包含梅爾倒頻譜係數(MFCC)、線性鑑別分析(LDA)、異方差線性鑑別分析(HLDA)、異方差鑑別分析(HDA)、基於乘冪平均的線性鑑別分析(PLDA)、基於乘冪之權重式線性鑑別分析(PWLDA)、近似成對理論正確標準(aPTAC)。

本論文實驗中的聲學模型訓練是使用傳統的最大化相似度估計法(maximum likelihood estimation, MLE)[87]，也就是利用最大化相似度使其聲學模型與其對應的訓練語音特徵量序列愈像愈好，即語音特徵向量序列落在其對應的聲學模型之相似度會最大，來估計聲學模型參數，並利用波氏重估(Baum-Welch re-estimation,

表 5.6　PLDA 在不同 m 值設定下之正確率(%)

PLDA	無 MLLT 之正確率			有 MLLT 之正確率		
	音節	字	詞	音節	字	詞
$m=3$	74.05	**66.34**	57.40	78.56	**71.15**	62.44
$m=2$	77.23	**69.35**	60.5	79.47	**71.89**	63.42
$m=1$ (LDA)	76.81	**68.97**	60.10	79.06	**71.56**	62.74
$m=0$ (HDA)	76.60	**68.51**	59.52	79.62	**72.07**	63.23
$m=-1$	73.42	**64.98**	55.62	79.90	**72.38**	63.38
$m=-2$	76.49	**68.56**	59.54	77.2	**69.45**	60.56
$m=-3$	75.43	**67.49**	58.26	77.42	**69.7**	61

BW)演算法[88]，經過5次迭代來訓練聲學模型中的參數。最後經過測試語料得到最後的音節、字、詞正確率，如表5.5。

其中，在 HLDA 和 HDA 的部分，其最佳化技術均為使用 10,000 次迭代過程的梯度下降法(gradient descent)。而在 PLDA 的部分，由於不同的整數 m 值均對映到不同的乘冪平均而產生不同的目標函式，因此我們將式(2.44)中的整數 m 設定為-3 至 3，所得的結果如表 5.6。此外，PWLDA 的目標函式（式(3.3)）亦會隨著 k 值不同而改變，因此我們將 k 設定為 1 至 6，所得的結果如表 5.7。在表 5.5 中，PLDA 和 PWLDA 的部分實驗結果均分別以它們在表 5.6 和表 5.7 中最佳的辨識結果填入。

由表 5.5 我們可以看出幾個現象。

第一，從 PLDA 的結果可知，類別內散佈矩陣的重新估計，是以調和平均(harmonic mean)（即 $m=-1$）較好，此結果亦與 Sakai 等人的實驗相似[64]。

第二，在無 MLLT 作對角化轉換下，HLDA 的字正確率較其它方法，甚至是 MFCC 為低，原因可能在於當我們針對 HLDA 的目標函式進行最佳化的過程

表 5.7　PWLDA 在不同 k 值設定下之正確率(%)

PWLDA	無 MLLT 之正確率			有 MLLT 之正確率		
	音節	字	詞	音節	字	詞
$k=1$	76.87	**68.99**	60.13	79.25	**71.68**	62.71
$k=2$	76.64	**68.86**	59.90	79.36	**72.07**	63.39
$k=3$	77.33	**69.57**	60.60	79.16	**71.70**	62.94
$k=4$	77.00	**69.15**	59.93	78.93	**71.47**	62.69
$k=5$	77.41	**69.61**	60.68	79.20	**71.72**	62.94
$k=6$	76.95	**69.35**	60.66	79.24	**71.86**	63.16

時，並不會保證所求取的轉換矩陣能夠使每一類別之共變異矩陣趨於對角化。因此，在聲學模型訓練時，HLDA 轉換反而會造成較多的資訊損失。

第三，有 MLLT 的實驗結果在任何特徵擷取的方法中都明顯較無 MLLT 為佳，顯示前端處理過程的確需要有特徵去相關(feature decorrealation)的處理，才能與後端聲學模型之共變異矩陣的對角化形式一致。這也另外說明了我們在此比較這些方法孰優孰劣時，當以結合 MLLT 的結果會比較公允。

第四，我們發現，在 MLLT 作對角化轉換下，除了 MFCC 之外，其它以 LDA 為基礎的方法，其字正確率的差距都不大，也都遠較傳統的 MFCC 好，顯示出線性鑑別式特徵轉換，特別是 LDA，在語音特徵擷取上已經是不錯的方法。

5.3.3　基於混淆資訊之權重式線性鑑別分析實驗結果

在三種基於混淆資訊之權重式線性鑑別分析的方法中，混淆矩陣是依下列步驟取得：我們先以 LDA 產生訓練語料的語音特徵，再以訓練出來的聲學模型對訓練語料本身進行自由音節辨識(free syllable decoding)並以強迫對齊(forced alignment)技術產生每一語句內的音素邊界(phone boundary)和狀態邊界(state boundary)。最後，混淆矩陣則是依據所有音框之類別標記與原正確答案作對比

表 5.8　EER-WLDA 在不同α值設定下之正確率(%)

EER-WLDA	無 MLLT 之正確率			有 MLLT 之正確率		
	音節	字	詞	音節	字	詞
$\alpha = 0.0$	76.57	**68.56**	59.46	78.96	**71.36**	62.45
$\alpha = 0.1$	76.83	**68.96**	60.12	79.03	**71.39**	62.70
$\alpha = 0.2$	77.01	**68.91**	59.90	79.07	**71.57**	62.65
$\alpha = 0.3$	77.06	**69.18**	60.21	79.16	**71.51**	62.71
$\alpha = 0.4$	77.45	**69.60**	60.75	78.81	**71.15**	62.24
$\alpha = 0.5$	77.30	**69.72**	60.59	78.96	**71.34**	62.50
$\alpha = 0.6$	77.03	**69.12**	60.00	79.31	**71.69**	62.93
$\alpha = 0.7$	76.73	**68.96**	60.26	79.14	**71.61**	62.83
$\alpha = 0.8$	76.88	**69.09**	60.26	79.17	**71.53**	62.82
$\alpha = 0.9$	77.09	**69.32**	60.28	78.92	**71.33**	62.39

表 5.9　DE-WLDA 在不同階數之多項式回歸曲線下的正確率(%)

DE-WLDA	無 MLLT 之正確率			有 MLLT 之正確率		
	音節	字	詞	音節	字	詞
table lookup	76.81	**68.93**	59.85	78.96	**71.36**	62.41
linear	63.27	**55.31**	46.88	66.38	**58.52**	50.24
quadratic	76.98	**69.32**	60.26	78.99	**71.54**	62.76
cubic	76.61	**68.95**	60.07	78.59	**71.23**	62.46
4^{th} degree	76.86	**69**	59.74	78.77	**71.12**	62.07
5^{th} degree	76.93	**69.04**	59.97	78.83	**71.38**	62.64
6^{th} degree	76.84	**68.99**	59.79	79.14	**71.64**	62.78

而得。

在第一種方法：基於經驗錯誤率之權重式線性鑑別分析(EER-WLDA)中，我們將α值在 0 到 1 之間設定了 10 種，它們之字正確率的分佈並無明顯規則，如表 5.7。在第二種方法：距離－錯誤耦合之權重式線性鑑別分析(DE-WLDA)與第三種方法：近似成對經驗正確率標準(approximate pairwise empirical accuracy criterion, aPEAC)中，則比較了 6 種不同階數形成的多項式回歸曲線，用來描述馬氏距離與經驗錯誤率之關係，其字正確率分別如表 5.8 與表 5.9，似乎可看出

表 5.10　aPEAC 在不同階數之多項式回歸曲線下的正確率(%)

aPEAC	無 MLLT 之正確率			有 MLLT 之正確率		
	音節	字	詞	音節	字	詞
table lookup	77.34	**69.54**	60.73	78.94	**71.31**	62.55
linear	74.31	**66.68**	57.89	75.97	**68.32**	59.56
quadratic	77.55	**70.02**	61.28	78.98	**71.22**	62.24
cubic	77.09	**69.18**	60	78.65	**71.02**	61.97
4^{th} degree	76.87	**69.14**	60.24	78.6	**71.14**	62.18
5^{th} degree	76.84	**69.05**	60.01	78.5	**70.89**	61.97
6^{th} degree	77.05	**69.51**	60.65	78.97	**71.41**	62.5

隨著多項式階數增加，字正確率也有增加的趨勢。

在 DE-WLDA 與 aPEAC 的實驗中，我們也嘗試用查表法(table lookup)，也就是不對資料作回歸分析，而直接使用圖 3.4 中的資料點所對應到的經驗分類錯誤率。這種方法也可視為一種變形的 EER-WLDA，只是它不僅在錯誤率的估計方式與 EER-WLDA 相異，亦不考慮與原始距離之間的權重比較。實驗結果顯示，查表法的效果並不是最好的，原因可能在於它並不能代表真正資料分佈的趨勢，且易於受到離群類別(outlier)或噪音(noise)干擾。由圖 3.4 可知，大量資料點仍集中在經驗分類錯誤率小於 0.01 的區域，若我們使用查表法，則這些資料點會嚴重限制目標函數解決 LDA 過度強調問題的能力。

與 LDA 做比較，我們也發現這三種基於混淆資訊之權重式線性鑑別分析的辨識結果並沒有很突出。以 LDA 作為基線(baseline)，在最好的情況下，EER-WLDA（$(\alpha = 0.6)$+MLLT）的相對進步率(relative improvement)只有 0.6%，DE-WLDA（(6^{th} degree)+MLLT）的相對進步率更只有 0.4%，aPEAC（+MLLT）則完全沒進步。其中可能的原因有二：

第一，在圖 3.4 中，仍有大量的資料點，其馬氏距離與經驗分類錯誤率是無關的。例如，有許多馬氏距離 $\Delta_{ij} > 3$ 的資料點，其經驗分類錯誤率 $m_{ij}^{\mathrm{DE}} > 0.02$ ，

表 5.11　GLRDA 在異方差性與基於混淆資訊下之應用的正確率(%)

GLRDA	無 MLLT 之正確率			有 MLLT 之正確率		
	音節	字	詞	音節	字	詞
H-GLRDA 式(4.25)	67.74	**59.19**	48.93	79.69	**72.19**	63.32
CI-GLRDA 式(4.31)	68.85	**60.01**	49.86	79.64	**72.25**	63.39

可是圖 3.6 所有的多項式回歸曲線在 $\Delta_{ij} > 3$ 時的變動卻極低。這使得多項式回歸曲線受到這些不尋常的資料點所主導，無法真正呈現出負相關性或處理 $\Delta_{ij} > 3$ 之類別配對產生的過度強調問題。

第二，在預期資料點具有一定程度負相關性的情況下，aPEAC 完全去除類別配對的距離因子，完全倚靠圖 3.4 產生的負相關性，使得實際情況不如預期時，實驗結果反而會有得不償失的情況。

5.3.4　普遍化相似度比率鑑別分析實驗

我們主要是將普遍化相似度比率鑑別分析(GLRDA)的概念應用在異方差性的統計假設上（H-GLRDA，式(4.25)），其最佳化技術為使用 10,000 次迭代過程的梯度下降法(gradient descent)，辨識結果如表 5.10。其中，在基於混淆資訊之普遍化相似度比率鑑別分析(CI-GLRDA)的部分，我們考慮了 10 種前 K 組易於混淆的類別配對，$K = 10,20,...,100$，如表 5.12，並將其中最佳的結果填入表 5.10 中。而表 5.11 說明了，我們並不需要將 K 延伸至全部類別配對，因為易於混淆的類別配對中，其所含的音框分類錯誤數，在 $K = 100$ 時即佔了全部音框分類錯誤數 40%，並且它們本身的音框總數也佔了全體音框總數的 80%以上，也就是說，我們所處理的類別已佔了所有訓練音框中的絕大多數，而它們之中所含的配對音框錯誤數，也佔了頗大的比例（見表 3.1）。

表 5.12　CI-GLRDA 中前 K 組易於混淆之類別配對的相關統計

CI-GLRDA	事前機率	所含錯誤音框比率	混淆群聚數
$K=10$	28.84 %	11.65 %	3
$K=20$	43.48 %	17.92 %	3
$K=30$	53.14 %	22.56 %	1
$K=40$	62.02 %	26.00 %	1
$K=50$	67.85 %	28.97 %	1
$K=60$	72.58 %	31.66 %	1
$K=70$	73.82 %	34.02 %	1
$K=80$	75.97 %	36.25 %	1
$K=90$	78.41 %	38.37 %	1
$K=100$	81.21 %	40.25 %	1

表 5.13　CI-GLRDA 在只考慮前 K 組易於混淆之類別配對下之正確率(%)

CI-GLRDA	無 MLLT 之正確率			有 MLLT 之正確率		
	音節	字	詞	音節	字	詞
$K=10$	66.39	**57.39**	47.54	78.08	**70.61**	61.93
$K=20$	67.15	**57.98**	48.03	79.1	**71.6**	62.76
$K=30$	68.56	**59.66**	49.81	79.5	**71.85**	63.06
$K=40$	68.73	**59.95**	50.15	79.02	**71.51**	62.62
$K=50$	67	**58.1**	48.27	79.32	**71.88**	62.96
$K=60$	67.48	**58.59**	48.65	79.52	**72.14**	63.36
$K=70$	68.77	**59.98**	49.84	79.64	**72.25**	63.39
$K=80$	68.85	**60.01**	49.86	79.35	**72.09**	63.21
$K=90$	67.75	**58.87**	48.73	79.63	**72.18**	63.16
$K=100$	68.63	**59.98**	49.79	79.4	**71.95**	63.08

我們也可以發現，異方差性的 GLRDA（+MLLT）和 CI-GLRDA（+MLLT）對於 LDA 能夠分別有 2.21%和 2.43%的相對進步率。其中，前者些微超越了廣為使用的 HLDA（+MLLT）0.32%絕對字正確率，印證了我們在 4.2.3 節針對二者的比較與討論。而 CI-GLRDA（+MLLT）的成功，也說明了混淆資訊的另一種方式的運用，的確有鑑別性的效果。這也給了我們另一種啟示：要使辨識率進步，也可針對最混淆的音素來處理，特別是靜音(silence)模型（見表 3.1）。另一

表 5.14　本論文中各種特徵擷取方法於 MPE 聲學模型訓練下之正確率(%)

MPE training	有 MLLT 之正確率		
	音節	字	詞
MFCC	79.05	**71.51**	62.83
LDA	83.16	**75.64**	67.15
aPTAC	83.01	**75.43**	66.96
HLDA	83.10	**75.57**	67.24
HDA	83.37	**75.94**	67.45
PLDA ($m=-1$)	83.57	**76.01**	67.37
PWLDA ($k=2$)	83.47	**76.23**	67.69
DE-WLDA (6^{th} degree)	82.18	**74.65**	66.04
EER-WLDA ($\alpha=0.6$)	83.10	**75.48**	66.65
GLRDA (Heteroscedasticity)	83.42	**75.94**	67.55
CI-GLRDA ($K=70$)	83.58	**76.11**	67.77

方面，這也可以突顯出語音強健性(robustness)技術與語音活動偵測(voice activity detection, VAD)的重要性，因為上述二者均可以降低靜音模型與其他音素模型的混淆。

5.3.5　最小化音素錯誤(MPE)實驗

在聲學模型訓練中，除了傳統的最大化相似度訓練(ML training)之外，我們也試著將各種特徵擷取的方法應用在近來當紅的鑑別式訓練：最小化音素錯誤(minimum phone error, MPE)[89-90]。傳統的最大化相似度訓練並沒有考慮語音辨識時聲學模型彼此間的關係，在調整聲學模型參數之後，可以使得相關的語音特徵落在此聲學模型的相似度變大，卻也可能同時讓非相關的語音特徵落在此聲學

模型的相似度更大，造成辨識上的混淆。因此，像最小化音素錯誤這種鑑別式訓練補足了最大化相似度訓練的缺點，它是以全面風險(overall risk)為出發，目標函數變成是最大化語音辨識器對所有訓練語句（語音特徵向量序列）O_z的可能辨識出候選詞序列W_i的期望音素正確率（也就是最小化語音辨識器對所有訓練語句可能辨識出候選詞序列W_i的期望錯誤率），實驗結果如表 5.13。

第 6 章　結論與未來展望

鑑別式聲學特徵擷取在大詞彙連續語音辨識的研究上一直扮演著重要的角色。本論文旨在改善傳統線性鑑別分析(LDA)，相關研究內容與成果可從下面三個面向來作討論:

一、首先，本論文詳細探討了LDA的基本原理與其統計和分類上的意義。在一般的教科書上，對於LDA的描述僅只於特徵擷取與線性模型的基本介紹，而鮮見LDA與分類實務的分析。而在本論文中，我們利用了LDA的幾何分析與證明，點出了LDA在分類上的兩大問題：過度強調問題與分類正確率無關問題，並提供了後續在圖樣辨識領域中，關於LDA之研究的基本方向。

二、本論文充分利用了以辨識器產生的混淆資訊，並以此修正及解決了上述之LDA兩大問題，並且保有LDA輕省的可解性。而相較於在圖樣辨識領域有名的aPTAC方法，我們所提出的方法並不需有機率分布的假設，並且在實務上，能使得擷取出的特徵更貼近分類器的特性，使得圖樣辨識中的前端處理與後端分類更能緊密結合。

三、為了打破LDA對於各類別之共變異矩陣的限制，本論文亦參考了統計學上相似度比率檢驗的概念，提出了一種新式線性鑑別分析（GLRDA），不僅能普遍化現有的技術，如LDA與HLDA，更可進一步結合混淆資訊而產生更好的效果。不限於語音處理，我們相信GLRDA技術可以更廣泛地應用在所有需要鑑別性特徵的領域，特別是影像處理。

而未來與本論文相關的研究可分作兩方面：

一、在混淆資訊的使用上，我們所定義之類別配對的經驗分類錯誤率(pairwise empirical classification error rate)未必是最佳、最合乎實驗評估的定義。未來在語音辨識上，我們會嘗試參考最小化音素錯誤(MPE)中對於音素錯誤的定義，發展出更合適的經驗分類錯誤率。

二、GLRDA是一種很廣泛的概念，其產生之特徵的鑑別力強度，取決於虛無假設的設定與混淆資訊的使用。未來在混淆群聚(confusable cluster)的決定上，我們會將每一類別配對的混淆強度（或錯誤音框數）納入考慮。

儘管實驗結果並不十分突出，本論文提出之基於錯誤分析的方式的確提供了我們一個新的視野來看目前的線性特徵轉換。而以類別配對的角度嘗試最大化分類正確率雖然只能逼近全部分類正確率，但未來我們希望能夠以成對性的標準當作發展跳板，拉近與全部分類正確率的距離。

第 7 章　附錄

7.1　重要的向量微分公式

令 $\mathbf{x}$ 為向量(vector)（$\mathbf{x} \in \Re^{n\times 1}$），$\mathbf{B}$ 為任意矩陣(matrix)，$\mathbf{A}$ 、$\mathbf{S}_1$ 、$\mathbf{S}_2$ 為對稱矩陣 (symmetric matrices)（$\{\mathbf{A},\mathbf{S}_1,\mathbf{S}_2\} \in \Re^{n\times n}$, $\{\mathbf{A},\mathbf{S}_1,\mathbf{S}_2\} = \{\mathbf{A},\mathbf{S}_1,\mathbf{S}_2\}^T$），則[49]：

1. $$\frac{\partial \mathbf{x}\mathbf{A}\mathbf{x}}{\partial \mathbf{x}} = 2\mathbf{A}\mathbf{x} \tag{7.1}$$

2. $$\frac{\partial \log|\mathbf{A}|}{\partial \mathbf{A}} = \mathbf{A}^{-1} \tag{7.2}$$

3. $$\frac{\partial \mathbf{x}^T\mathbf{A}^{-1}\mathbf{x}}{\partial \mathbf{A}} = -\mathbf{A}^{-1}\mathbf{x}\mathbf{x}^T\mathbf{A}^{-1} \tag{7.3}$$

4. $$\frac{\partial \operatorname{trace}(\mathbf{A}^{-1}\mathbf{X})}{\partial \mathbf{A}} = -\mathbf{A}^{-1}\mathbf{X}^T\mathbf{A}^{-1} \tag{7.4}$$

5. $$\frac{\partial \operatorname{trace}\left((\mathbf{B}^T\mathbf{S}_2\mathbf{B})^{-1}(\mathbf{B}^T\mathbf{S}_1\mathbf{B})\right)}{\partial \mathbf{B}} = \\ -2\mathbf{S}_2\mathbf{B}(\mathbf{B}^T\mathbf{S}_2\mathbf{B})^{-1}(\mathbf{B}^T\mathbf{S}_1\mathbf{B})(\mathbf{B}^T\mathbf{S}_2\mathbf{B})^{-1} + 2\mathbf{S}_1\mathbf{B}(\mathbf{B}^T\mathbf{S}_2\mathbf{B})^{-1} \tag{7.5}$$

7.2 一些證明推導

7.2.1 證明式(2.37)

$$
\begin{aligned}
\mathbf{S}_T &= \frac{1}{N}\sum_{\mathbf{x}\in X}(\mathbf{x}-\overline{\mathbf{m}})(\mathbf{x}-\overline{\mathbf{m}})^T = \frac{1}{N}\sum_{i=1}^{C}\sum_{\mathbf{x}\in C_i}(\mathbf{x}-\overline{\mathbf{m}})(\mathbf{x}-\overline{\mathbf{m}})^T \\
&= \frac{1}{N}\sum_{i=1}^{C}\sum_{\mathbf{x}\in C_i}\big((\mathbf{x}-\mathbf{m}_i)+(\mathbf{m}_i-\overline{\mathbf{m}})\big)\big((\mathbf{x}-\mathbf{m}_i)+(\mathbf{m}_i-\overline{\mathbf{m}})\big)^T \\
&= \frac{1}{N}\sum_{i=1}^{C}\sum_{\mathbf{x}\in C_i}\begin{pmatrix}(\mathbf{x}-\mathbf{m}_i)(\mathbf{x}-\mathbf{m}_i)^T+(\mathbf{m}_i-\overline{\mathbf{m}})(\mathbf{m}_i-\overline{\mathbf{m}})^T+\\(\mathbf{x}-\mathbf{m}_i)(\mathbf{m}_i-\overline{\mathbf{m}})^T+(\mathbf{m}_i-\overline{\mathbf{m}})(\mathbf{x}-\mathbf{m}_i)^T\end{pmatrix} \\
&= \frac{1}{N}\sum_{i=1}^{C}\sum_{\mathbf{x}\in C_i}\big((\mathbf{x}-\mathbf{m}_i)(\mathbf{x}-\mathbf{m}_i)^T+(\mathbf{m}_i-\overline{\mathbf{m}})(\mathbf{m}_i-\overline{\mathbf{m}})^T\big) \\
&= \frac{1}{N}\sum_{i=1}^{C}n_i p_i\mathbf{S}_i + \frac{1}{N}\sum_{i=1}^{C}n_i(\mathbf{m}_i-\overline{\mathbf{m}})(\mathbf{m}_i-\overline{\mathbf{m}})^T = \mathbf{S}_W+\mathbf{S}_B
\end{aligned}
\tag{7.6}
$$

7.2.2 有關高斯分佈與相似度的證明

對於屬於類別 C_{l_i} 的資料 $\{\mathbf{x}_i\}$，其多變量高斯分佈(multivariate Gaussian distribution)為：（l_i 為類別編號，用來標示資料 $\mathbf{x}$ 所屬的類別。）

$$
p_{C_{l_i}}(\mathbf{x}_i,\boldsymbol{\mu}_{l_i},\boldsymbol{\Sigma}_{l_i}) = \frac{1}{\sqrt{(2\pi)^d\,|\boldsymbol{\Sigma}_{l_i}|}}\exp\left(-\frac{1}{2}(\mathbf{x}_i-\boldsymbol{\mu}_{l_i})^T\boldsymbol{\Sigma}_{l_i}^{-1}(\mathbf{x}_i-\boldsymbol{\mu}_{l_i})\right) \tag{7.7}
$$

將式(7.7)之等號兩邊取對數，可得

$$
\log p_{C_{l_i}}(\mathbf{x}_i,\boldsymbol{\mu}_{l_i},\boldsymbol{\Sigma}_{l_i}) = -\frac{d}{2}\log(2\pi)-\frac{1}{2}\log|\boldsymbol{\Sigma}_{l_i}|-\frac{1}{2}(\mathbf{x}_i-\boldsymbol{\mu}_{l_i})^T\boldsymbol{\Sigma}_{l_i}^{-1}(\mathbf{x}_i-\boldsymbol{\mu}_{l_i}) \tag{7.8}
$$

將每一筆資料 $\mathbf{x}_i$ 代入式(7.8)，並取其和，即為所有資料 $\{\mathbf{x}_i\}$ 在此類別 C_{l_i} 分佈下之相似度(likelihood)：

$$\log p_{C_{l_i}}(\mathbf{x}_1^{n_{l_i}}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \\ -\frac{n_{l_i} d}{2}\log(2\pi) - \frac{n_{l_i}}{2}\log|\boldsymbol{\Sigma}_{l_i}| - \frac{1}{2}\sum_{i=1}^{n_{l_i}}(\mathbf{x}_i - \boldsymbol{\mu}_{l_i})^T \boldsymbol{\Sigma}_{l_i}^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_{l_i}) \tag{7.9}$$

首先，我們可證明：

$$\begin{aligned} &\sum_{i=1}^{n_{l_i}}(\mathbf{x}_i - \boldsymbol{\mu}_{l_i})^T \boldsymbol{\Sigma}_{l_i}^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_{l_i}) \\ &= \text{trace}(\sum_{i=1}^{n_{l_i}}(\mathbf{x}_i - \boldsymbol{\mu}_{l_i})^T \boldsymbol{\Sigma}_{l_i}^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_{l_i})) \\ &= \text{trace}(\boldsymbol{\Sigma}_{l_i}^{-1}\sum_{i=1}^{N}(\mathbf{x}_i - \boldsymbol{\mu}_{l_i})^T(\mathbf{x}_i - \boldsymbol{\mu}_{l_i})) \\ &= n_{l_i}\,\text{trace}(\boldsymbol{\Sigma}_{l_i}^{-1}\mathbf{S}_{l_i}) \end{aligned} \tag{7.10}$$

將式(7.10)代入式(7.9)，可得

$$\log p_{C_{l_i}}(\mathbf{x}_1^{n_{l_i}}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = -\frac{n_{l_i} d}{2}\log(2\pi) - \frac{n_{l_i}}{2}\log|\boldsymbol{\Sigma}_{l_i}| - \frac{n_{l_i}}{2}\text{trace}(\boldsymbol{\Sigma}_{l_i}^{-1}\mathbf{S}_{l_i}) \tag{7.11}$$

若要求出所有類別的所有資料 $\mathbf{x}_1^N$ 的總相似度，我們可將所有類別之統計量代入式(7.11)，並取其和，可得：

$$\log p(\mathbf{x}_1^N, \{\boldsymbol{\mu}_k\}, \{\boldsymbol{\Sigma}_k\}) = -\frac{Nd}{2}\log(2\pi) - \\ \frac{1}{2}\sum_{k=1}^{C} n_k\left((\overline{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1}(\overline{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_k) + \text{trace}(\boldsymbol{\Sigma}_k^{-1}\overline{\boldsymbol{\Sigma}}_k) + \log|\boldsymbol{\Sigma}_k|\right) \tag{7.12}$$

參考文獻

[1] H.-S. Chiu*, et al.*, "Position information for language modeling in speech recognition," in *Proc. ISCSLP*, 2008, pp. 101-104.

[2] J. Li*, et al.*, "Soft margin estimation of hidden markov model parameters," in *Proc. Interspeech*, 2006, pp. 2422-2425.

[3] M. Gilbert*, et al.*, "Intelligent virtual agents for contact center automation," *IEEE Signal Processing Magazine,* vol. 22, pp. 32-41, 2005.

[4] M. Gilbert and J. Feng, "Speech and language processing over the web," *IEEE Signal Processing Magazine,* vol. 25, pp. 18-28, 2008.

[5] N. Morgan*, et al.*, "Pushing the envelope - aside," *IEEE Signal Processing Magazine,* vol. 22, pp. 81-88, 2005.

[6] H. Hermansky, "Should recognizers have ears?," *Speech Communication,* vol. 25, pp. 3-27, 1998.

[7] H. Hermansky, "Exploring temporal domain for robustness in speech recognition," in *Proc. ICA*, 1995, pp. 61-64.

[8] H. Bourlard and N. Morgan, *Connectionist Speech Recognition: A Hybrid Approach*: Springer, 1994.

[9] M. J. Hunt and C. Lefdbvre, "A comparison of several acoustic representations for speech recognition with degraded and undegraded speech," in *Proc. ICASSP*, 1989, pp. 262-265.

[10] S. Makino*, et al.*, "Recognition of consonant based on the Perceotron model," in *Proc. ICASSP*, 1983, pp. 738-741.

[11] S. Furui, "Speaker-independent isolated word recognizer using dynamic features of speech spectrum," *IEEE Trans. Acoustics, Speech, and Signal*

Processing, vol. 34, pp. 52-59, 1986.

[12] J. S. Bowers and C. J. Davis, "Is speech perception modular or interactive?," *Trends in Cognitive Sciences,* vol. 8, pp. 3-5, 2004.

[13] A. G. Samuel, "Knowing a word affects the fundamental perception of the sounds within it," *Psychological Science,* vol. 12, pp. 348-351, 2001.

[14] J. Obleser and F. Eisner, "Pre-lexical abstraction of speech in the auditory cortex," *Trends in Cognitive Sciences,* vol. 13, pp. 14-19, 2009.

[15] D. Norris*, et al.*, "Merging information in speech recognition: Feedback is never necessary," *Behavioral and Brain Sciences,* vol. 23, pp. 299-370, 2000.

[16] M. K. Tanenhaus*, et al.*, "No compelling evidence against feedback in spoken word recognition," *Behavioral and Brain Sciences,* vol. 23, pp. 348-349, 2000.

[17] D. B. Pisoni and R. E. Remez, *The Handbook of Speech Perception*. Oxford: Blackwell, 2005.

[18] R. Chengalvarayan and L. Deng, "HMM-based speech recognition using state-dependent, discriminatively derived transforms on mel-warped DFT features," *IEEE Trans. Speech and Audio Processing,* vol. 12, pp. 19-26, 1997.

[19] X.-B. Li*, et al.*, "Dimensionality reduction using MCE-optimized LDA transformation," in *Proc. ICASSP*, 2004, pp. 137-140.

[20] D. Povey*, et al.*, "fMPE: discriminatively trained features for speech recogntion," in *Proc. ICASSP*, 2005, pp. 961-964.

[21] B. Schölkopf and A. J. Smola, *Learning with Kernels - Support Vector Machines, Regularization, Optimization, and Beyond*. Cambridge, Massachusetts: The MIT Press, 2002.

[22] E. Alpaydin, *Introduction to Machine Learning*. Cambridge, MA: The MIT Press, 2004.

[23] A. R. Webb, *Statistical Pattern Recognition*, 2nd ed.: John Wiley and Sons,

2002.

[24] M. Gales, "Semi-tied covariance matrices for hidden Markov models," *IEEE Trans. Speech and Audio Processing,* vol. 7, pp. 272-281, 1999.

[25] X. Wang and K. K. Paliwal, "Feature extraction and dimensionality reduction algorithms and their applications in vowel recognition," *Pattern Recognition,* vol. 36, pp. 2429-2439, 2003.

[26] N. Kumar and A. G. Andreou, "Heteroscedastic discriminant analysis and reduced rank HMMs for improved speech recognition," *Speech Communication,* vol. 26, pp. 283-297, 1998.

[27] G. Saon*, et al.*, "Maximum likelihood discriminant feature spaces," in *Proc. ICASSP*, 2000, pp. 1129-1132.

[28] K. Demuynck*, et al.*, "Optimal feature sub-space selection based on discriminant analysis " in *Proc. Eurospeech*, 1999, pp. 1311-1314.

[29] X. Cui*, et al.*, "Stereo-based stochastic mapping with discriminative training for noise robust speech recognition," in *Proc. ICASSP*, 2009, pp. 2933-2936.

[30] P. F. Brown, "The acoustic-modelling problem in automatic speech recognition," Ph.D. dissertation, Carnegie Mellon University, 1987.

[31] R. Haeb-Umbach and H. Ney, "Linear discriminant analysis for improved large vocabulary continuous speech recognition," in *Proc. ICASSP*, 1992, pp. 13-16.

[32] H. Hermansky, "Stochastic techniques in deriving perceptual knowledge," in *Proc. SAPA*, 2004.

[33] D. Jurafsky and J. H. Martin, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition*, 2nd ed.: Prentice Hall, 2008.

[34] S. Young*, et al.*, *The HTK Book (for HTK Version 3.4)*: Cambridge University

Engineering Department, 2006.

[35] L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*: Prentice Hall, 1993.

[36] W. Chou and B.-H. Juang, *Pattern Recognition in Speech and Language Processing*: CRC Press, 2003.

[37] X. Liu, "Discriminative complexity control and linear projections for large vocabulary speech recognition," Ph.D. dissertation, University of Cambridge, 2005.

[38] M. N. Stuttle, "A Gaussian mixture model spectral representation for speech recognition," Ph.D. dissertation, University of Cambridge, 2003.

[39] J. W. Picone, "Signal modeling techniques in speech recognition," in *Proc. the IEEE*, 1993, pp. 1214-1247.

[40] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoustics, Speech, and Signal Processing,* vol. 28, pp. 357-366, 1980.

[41] R. Haeb-Umbach*, et al.*, "Improvements in connected digit recognition using linear discriminant analysis and mixture densities," in *Proc. ICASSP*, 1994, pp. 239-242.

[42] B. D. Ripley, *Pattern Recognition and Neural Networks*. New York: Cambridge University Press, 1996.

[43] M. J. Hunt, "A statistical approach to metrics for word and syllable recognition," presented at the the 98th Meeting of the Acoustical Society of America, 1979.

[44] G. R. Doddington, "Phonetically sensitive discriminants for improved speech recognition," in *Proc. ICASSP*, 1989, pp. 556-559.

[45] M. J. Hunt*, et al.*, "An investigation of PLP and IMELDA acoustic

representations and of their potential for combination," in *Proc. ICASSP*, 1991, pp. 881-884.

[46] L. Wood*, et al.*, "Improved vocabulary-independent sub-word HMM modelling," in *Proc. ICASSP*, 1991, pp. 181-184.

[47] G. Yu*, et al.*, "Discriminant analysis and supervised vector quantization for continuous speech recognition," in *Proc. ICASSP*, 1990, pp. 685-688.

[48] C. M. Ayer*, et al.*, "A discriminately derived linear transform for improved speech recognition," in *Proc. Eurospeech*, 1993, pp. 583-586.

[49] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, 2nd ed. New York: Academic Press, 1990.

[50] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of Eugenics,* vol. 7, pp. 179-188, 1936.

[51] R. A. Fisher, "The statistical utilization of multiple measurements," *Annals of Eugenics,* vol. 8, pp. 376-386, 1938.

[52] C. R. Rao, "The utilization of multiple measurements in problems of biological classification," *Journal of the Royal Statistical Society, Series B,* vol. 10, pp. 159-203, 1948.

[53] R. O. Duda*, et al.*, *Pattern Classification*. New York: John & Wiley, 2000.

[54] G. A. F. Seber, *Multivariate Observations*. New York: John Wiley & Sons, 1984.

[55] S. S. Wilks, *Mathematical Statistics*. New York: John Wiley & Sons, 1962.

[56] R. A. Johnson and D. W. Wichern, *Applied Multivariate Statistical Analysis*, 5th ed. New Jersey: Prentice Hall, 2002.

[57] R. A. Gopinth, "Maximum likelihood modeling with Gaussian distributions for classification," in *Proc. ICASSP*, 1998, pp. 661-664.

[58] N. A. Campbell and W. R. Atchley, "The geometry of canonical variate

analysis," *Systematic Zoology,* vol. 30, pp. 268-280, 1981.

[59] W. J. Krzanowski, *Principles of Multivariate Analysis: A User's Perspective*. New York: Oxford University Press, 1988.

[60] D. J. Hand, *Construction and Assessment of Classification Rules*. New York: John Wiley & Sons, 1997.

[61] N. A. Campbell, "Canonical Variate Analysis - A General Model Formulation," *Australian Journal ofStatistics,* vol. 26, pp. 86-96, 1984.

[62] N. Kumar, "Investigation of silicon auditory models and generalization of linear discriminant analysis for improved speech recognition," Ph.D. dissertation, Johns Hopkins University, 1997.

[63] M. Sakai*, et al.*, "Generalization of linear discriminant analysis used in segmental unit input hmm for speech recognition," in *Proc. ICASSP*, 2007, pp. 333-336.

[64] M. Sakai*, et al.*, "Linear discriminant analysis using a generalized mean of class covariances and its application to speech recognition," *IEICE Trans. Information and Systems,* vol. E91-D, pp. 478-487, 2008.

[65] C. R. Rao, *Linear Statistical Inference and Its Applications*, 2nd ed. New York: John Wiley & Sons, 2002.

[66] T. W. Anderson, *An Introduction to Multivariate Statistical Methods*, 2nd ed. New York: John Wiley & Sons, 1984.

[67] S. Geisser, "Discrimination, Allocatory, and Separatory Linear Aspects," in *Classification and Clustering*, J. V. Ryzin, Ed., ed, 1977, pp. 301-330.

[68] Y. Li*, et al.*, "Weighted pairwise scatter to improve linear discriminant analysis," in *Proc. ICSLP*, 2000, pp. 608-611.

[69] Y. Liang*, et al.*, "Uncorrelated linear discriminant analysis based on weighted pairwise Fisher criterion," *Pattern Recognition,* vol. 40, pp. 3606-3615, 2007.

[70] M. Loog and R. Haeb-Umbach, "Multi-class linear dimension reduction by generalized Fisher criteria," in *Proc. ICSLP*, 2000, pp. 1069-1072.

[71] H.-S. Lee and B. Chen, "Linear discriminant feature extraction using weighted classification confusion information," in *Proc.Interspeech*, 2008, pp. 2254-2257.

[72] H.-S. Lee and B. Chen, "Improved linear discriminant analysis considering empirical pairwise classification error rates," in *Proc. ISCSLP*, 2008, pp. 149-152.

[73] H.-S. Lee and B. Chen, "Empirical error rate minimization based linear discriminant analysis," in *Proc. ICASSP*, 2009.

[74] E. K. Tang*, et al.*, "Linear dimensionality reduction using relevance weighted LDA," *Pattern Recognition,* vol. 38, pp. 485-493, 2005.

[75] Y. Liu and P. Fung, "Acoustic and phonetic confusions in accented speech recognition," in *Proc. Interspeech*, 2005, pp. 3033-3036.

[76] J. M. Górriz*, et al.*, "Generalized LRT-based voice activity detector," *IEEE Signal Processing Letters,* vol. 13, pp. 636-639, 2006.

[77] N. A. Campbell, "Canonical variate analysis with unequal covariance matrices - generalizations of the usual solution," *Mathematical Geology,* vol. 16, pp. 109-124, 1984.

[78] J. D. Foley*, et al.*, *Computer Graphics: Principles and Practice in C*, 2nd ed.: Addison-Wesley, 1995.

[79] H.-M. Wang*, et al.*, "MATBN: A mandarin Chinese broadcast news corpus," *International Journal of Computational Linguistics and Chinese Language Processing,* vol. 10, pp. 219-235, 2005.

[80] C. Barras*, et al.*, "Transcriber : Development and use of a tool for assisting speech corpora production," *Speech Communication,* vol. 33, pp. 5-22, 2001.

[81] A. Stolcke, *SRI Language Modeling Toolkit (Version 1.5.2)*: http://www.speech.sri.com/projects/srilm/.

[82] X. Aubert, "An overview of decoding techniques for large vocabulary continuous speech recognition," *Computer Speech and Language,* vol. 16, pp. 89-114, 2002.

[83] 劉士弘, "改善鑑別式聲學模型訓練於中文連續語音辨識之研究," 碩士論文: 國立台灣師範大學, 2007.

[84] B. Chen*, et al.*, "Lightly supervised and data-driven approaches to mandarin broadcast news transcription," in *Proc. ICASSP*, 2004, pp. 777-780.

[85] 張志豪, "強健性和鑑別力語音特徵擷取技術於大詞彙連續語音辨識之研究," 碩士論文: 國立台灣師範大學, 2005.

[86] S. Ortmanns*, et al.*, "A word graph algorithm for large vocabulary continuous speech recognition," *Computer Speech and Language,* vol. 11, pp. 43-72, 1997.

[87] L. R. Bahl*, et al.*, "A maximum likelihood approach to continuous speech recognition," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. PAMI-5, pp. 179-190, 1983.

[88] L. E. Baum, "An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov Processes," *Inequalities,* vol. 3, pp. 1-8, 1972.

[89] D. Povey, "Discriminative training for large vocabulary speech recognition," Ph.D. Dissertation, University of Cambridge, 2004.

[90] D. Povey and P. C. Woodland, "Minimum phone error and I-smoothing for improved discriminative training," in *Proc. ICASSP*, 2002, pp. 105-108.

作者相關學術著作

1. **Hung-Shin Lee** and Berlin Chen, "Generalized likelihood ratio discriminant Analysis," *the 11th IEEE workshop on Automatic Speech Recognition and Understanding (ASRU 2010)*, submitted.
2. **Hung-Shin Lee** and Berlin Chen, "Likelihood ratio based discriminant analysis for large vocabulary continuous speech recognition," *ROCLING XXI: Conference on Computational Linguistics and Speech Processing (ROCLING 2009)*, September 1-2, 2009. (in Chinese)
3. **Hung-Shin Lee** and Berlin Chen, "Empirical error rate minimization based linear discriminant analysis," *the 34th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2009)*, Taipei, Taiwan, April 19-24, 2009.
4. **Hung-Shin Lee** and Berlin Chen, "Improved linear discriminant analysis considering empirical pairwise classification error rates," *the 6th International Symposium on Chinese Spoken Language Processing (ISCSLP 2008)*, pp. 149-152, Kunming, China, December 16-19, 2008. **(Best Student Paper Award 最佳學生論文獎)**
5. **Hung-Shin Lee** and Berlin Chen, "Linear discriminant feature extraction using weighted classification confusion information," *the 9th Annual Conference of the International Speech Communication Association (Interspeech 2008 - ICSLP)*, pp. 2254-2257, Brisbane, Australia, September 22-26, 2008.
6. Shih-Hung Liu, Fang-Hui Chu, Shih-Hsiang Lin, **Hung-Shin Lee**, and Berlin Chen, "Training data selection for improving discriminative training of acoustic models," *IEEE workshop on Automatic Speech Recognition and Understanding (ASRU 2007)*, pp. 284-289, Kyoto, Japan, December 9-13, 2007.