國 立 中 央 大 學

數 學 研 究 所

碩 士 論 文

神經細胞訊號的一種分類方法

研 究 生：李 易 霖

指導教授：單 維 彰 博 士

中 華 民 國 九 十 四 年 四 月 三 十 日

國立中央大學圖書館
碩博士論文電子檔授權書

(93 年 5 月最新修正版)

本授權書所授權之論文全文電子檔，為本人於國立中央大學，撰寫之碩/博士學位論文。(以下請擇一勾選)

(✓)同意 (立即開放)

( )同意 (一年後開放)，原因是：_____

( )同意 (二年後開放)，原因是：_____

( )不同意，原因是：_____

以非專屬、無償授權國立中央大學圖書館與國家圖書館，基於推動讀者間「資源共享、互惠合作」之理念，於回饋社會與學術研究之目的，得不限地域、時間與次數，以紙本、微縮、光碟及其它各種方法將上列論文收錄、重製、公開陳列、與發行，或再授權他人以各種方法重製與利用，並得將數位化之上列論文與論文電子檔以上載網路方式，提供讀者基於個人非營利性質之線上檢索、閱覽、下載或列印。

研究生簽名: 李 易 霖

論文名稱: 神 經 細 胞 訊 號 的 一 種 分 類 方 法

指導教授姓名：單 維 彰

系所 ：數學研究 所 ☐博士 ☑碩士班

學號：92221015

日期：民國 94 年 7 月 20 日

備註：

1. 本授權書請填寫並**親筆**簽名後，裝訂於各紙本論文封面後之次頁（全文電子檔內之授權書簽名，可用電腦打字代替）。
2. 請加印一份單張之授權書，填寫並親筆簽名後，於辦理離校時交圖書館（以統一代轉寄給國家圖書館）。
3. 讀者基於個人非營利性質之線上檢索、閱覽、下載或列印上列論文，應依著作權法相關規定辦理。

摘要

要研究大腦的活動，我們就要研究神經細胞的活性。神經元活動時會發出能被微電極探測到的動作電位，記錄這些訊號並加以分析，才能研究大腦的活動。但通常一根微電極探測到的動作電位會來自數個神經元，而且合成了不可預測的雜訊。因此，分析神經活性訊號的第一個步驟就是要探測動作電位的發生，其次就是動作電位的分類。

這篇論文不談探測，而提出一種新的分類方法。我們先介紹神經細胞活性與動作電位的背景知識，然後介紹一種常被使用的分類方法：主成分分析 (簡稱 PCA)。PCA 屬於多變量分析領域，很早就被應用在動作電位的分類。這個分類法雖然能在許多情況下分類動作電位，但依然有些問題無法解決。當數個神經元發出的動作電位差異不大時，PCA 的分類結果就會受到操作者的影響而變得不客觀。

本論文提出一個新的分類方法，這個方法的目的是當碰到動作電位差異不大的情況時，可以用一個較客觀的方式分類動作電位。這個新方法的另一個優點是可以估計分類正確率，因此我們能知道分類結果是否可以信賴。除了推導這個新方法以外，我們也設計了三組數據，用以示範新方法的功效，並與 PCA 做比較。

章節目錄

圖例目錄

第一章、背景知識

1.1 動作電位分類

神經元是神經系統的構造和機能單位，能將各種感覺由身體各處傳到大腦，或是從大腦將運動的訊息傳到身體各處。在沒有傳遞訊息時，神經元會保持靜止狀態。當神經元處於靜止狀態時，細胞膜內外通常會保持一定的電位差，稱作靜止膜電位。哺乳動物的膜外電位通常會比膜內電位高 70 mV 左右，此時的神經元處於極化狀態。造成膜內外的電位差的原因有兩個，一個是因為細胞膜通常會將一些生物大分子包圍在膜的內部，這些生物性的分子通常會帶有負電性。而另一個原因是神經細胞膜上具有鈉離子跟鉀離子通道以及鈉鉀離子泵，細胞會藉由離子通道及離子泵的調節作用使得膜內外的鈉離子及鉀離子分佈不均。當細胞處於靜止狀態時，鉀離子容易由細胞內流到細胞外，鈉離子不容易由細胞外流入細胞內。而鈉鉀細胞泵消耗一個 ATP 時會從膜外將兩個鉀離子帶入膜內，並且將三個鈉離子從膜內帶到膜外。這些因素造成膜內外的鈉鉀離子分佈不均，使得膜內外有電位差。神經元在受到適當的刺激後，會由相對靜止狀態轉變成顯著活躍狀態。當細胞受到刺激由靜止狀態進入活躍狀態時，細胞膜的鈉離子通透性會突然增加。由於濃度梯度以及內外電位差的作用，鈉離子將會迅速流入膜內，使得細胞膜去極化。鈉離子的流入會降低膜內外的電位差，這會造成更多的鈉離子通道打開，使得更多的鈉離子流入。去極化達到高峰時，膜內外的電位差將會從膜外比膜內高 70 mV 變成膜內比膜外高 35 mV。在這個時候，鈉離子通道會關閉，因此細胞膜對鈉離子的通透性就會下降。此時鉀離子通道也會打開，鉀離子會從細胞內流出到細胞外，細胞膜會很快的再極化，膜電位會恢復到 -80mV 左右。接著鈉鉀細胞泵會迅速發揮作用，使得膜內外的電位差再回到原本靜止狀態時的電位差。這個由靜止到活躍再回到靜止的電位變化，即是一個動作電位。

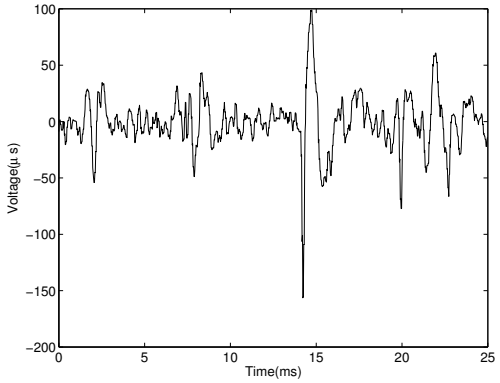
神經元會從樹突接受刺激，受到刺激後，如果刺激夠大，刺激點就會產生動作電位。由於納離子的流入，刺激點附近的細胞膜的納離子通道會跟著打開，然後如同上一段的描述，開始發生動作電位。這個流程會如同骨牌效應，讓動作電位沿著細胞膜傳遞。當鈉離子通道開啟過後，在短暫時間內，那個鈉離子通道不會因刺激而再度打開，因此動作電位不會回流。當動作電位傳到軸突時，會刺激軸突會發出神經傳導物質。靠近軸突的其他神經元的樹突會受到刺激，開始發生動作電位。經由這樣的流程，神經訊號可以傳達到身體的各部位。

動作電位的產生與否遵循全有全無律，當刺激夠大，動作電位就發生，否則就不發生，不會有介於兩者之間的情況。在刺激大到足以發生動作電位時，不論刺激有多大，動作電位都不會發生變化。但是較強烈的刺激會產生較多次的動作電位，也就是說當刺激較強烈時，動作電位的發生頻率也會比較大。但是動作電位的發生頻率有上限，所以當刺激的強度大到某種程度以後，動作電位的發生頻率不會再隨著刺激強度的增加而增加。

要研究大腦的活動，我們必須記錄大腦中的神經活性。要紀錄神經活性，其中一個方法就是記錄神經細胞發出的動作電位。神經細胞所發出的動作電位可以被微電極記錄到，因此我們可以將神經活性以電訊號的形式記錄下來。神經元在傳遞訊息時，細胞膜周圍的鈉和鉀離子濃度會發生變化。由於這兩種都是帶電離子，因此如果我們將微電極放到神經元附近，當神經元傳遞訊息時，細胞膜附近的電位變化就會被微電極紀錄下來。這個電位變化在被微電極紀錄下來時，會以電訊號的形式儲存。雖然在神經元發生動作電位時，微電極可以探測到細胞膜附近的電位變化，但隨著微電極與神經元相對位置的不同，探測到的動作電波形也不一樣。為了以下描述的方便，當我們說神經元發出的動作電位，我們是指神經元發出動作電位時微電極所收到的電位變化。

一根微電極通常會收到多個神經元發出的動作電位，隨著研究目的的不同，我們會需要將不同的神經元所發出的動作電位分開。通常我們會靠著比對動作電位的波形將不同神經元發出的動作電位分開。這個作法來自兩個假設。第一個假設是我們紀錄到的不同的神經元的動作電位波形會有差異。第二個假設是同一個神經元發出的動作電位波形不會隨著時間改變。要是這兩個假設有一個不成立，分開動作電位就會變得非常困難。由於微電極探測到的動作電位的波形會隨著電極與神經元的距離和相對位置而有所不同，也就是說即使是同一個神經元，只要微電極的位置發生變化，我們還是會探測到不同波形的動作電位。根據這個事實，我們可以知道不同神經元發出的動作電位波形不會完全相同，也就是說第一個假設總是會成立。但第二個假設卻不一定成立，有些神經元發出的動作電位波形會隨著時間而改變。在這種情況下，我們很難把這個神經元的動作電位分開，甚至於可能會將這些動作電位誤判成兩個以上的神經元的動作電位。

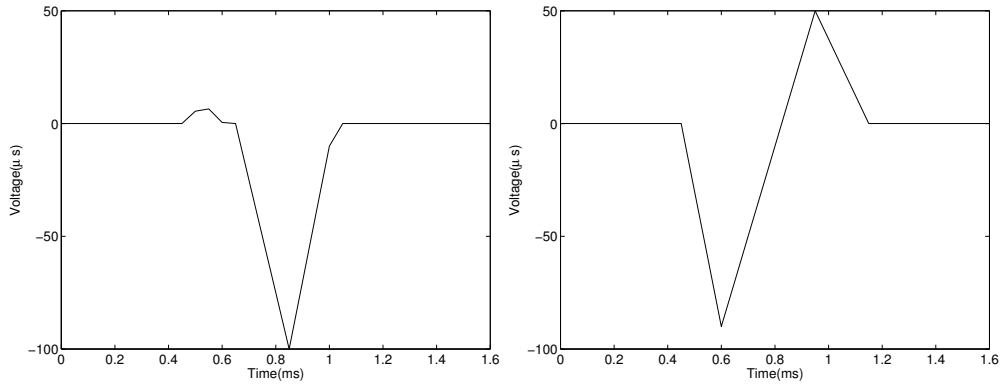
在分類動作電位時，我們還要考慮雜訊的干擾。雜訊的干擾有兩種，一種是環境雜訊。由於神經細胞產生的電位相當微弱，因此對環境雜訊的干擾會較敏感。環境雜訊的強度並不固定，當雜訊較大時，分類就會比較困難。另一種雜訊的干擾則是生物體



圖一

內各種活動的干擾。生物體的干擾有幾個情況，最受注意的是神經元彼此的干擾。通常微電極會探測到多個神經元的動作電位，如果有複數神經元同時產生動作電位，則這些動作電位會疊在一起，妨礙彼此之間的分類。不同神經元的動作電位在被微電極探測到時，其最大振幅會有差異。造成這個差異的原因之一就是微電極與神經元的距離。較靠近微電極的神經元，被探測到的動作電位也會比較大。因為微電極跟不同的神經元之間的距離會有很大的差異，所以探測到的不同神經元的動作電位的振幅也會有很大的變化。隨著動作電位振幅大小的不同，情況可能會分成兩類。一種是疊在一起的動作電位的振幅都比較大，此時我們要將這些動作電位挑出來，採用不同的方法分開這些動作電位，但本論文不討論這種情況。另一種情況則是只有一個動作電位較大，其他動作電位都比較小時，我們會把小動作電位當成雜訊，只分類大動作電位。下面提到的 thresholding 探測就是一種忽視小動作電位的探測。

Thresholding 探測的功用是找尋動作電位，是一種相當簡單的方法。這個方法的原理是這樣的：先決定一個數，把這個數當成一個波形該不該被視為動作電位的門檻。我們就將那個波形當成動作電位的波形，否則就當成雜訊。以圖一為例，如果我們認為振福要超過 100μV 才是動作電位，那麼圖一中只有 15ms 附近有一個動作電位，剩下的都被視為雜訊，即使我們知道其中可能包含了某些神經元的動作電位。並沒有一個硬性的方法告訴我們這個門檻要設多大才是最好的。門檻的大小決定了我們要處理的動作電位的數目。如果門檻設得比較高，則要被處理的動作電位就會比較少。門檻要是設得太高，我們就會遺失許多動作電位，後續的分析就會受到影響。反過來說，門檻設得比較低，則會有較多的動作電位要被處理，分類也會比較困難。要是門檻設得太低，則會有一些太小的動作電位被取進來，這些極難分類的動作電位會造成分類的負擔。所以我們必須要觀察探測到的電訊號，才能決定門檻的大小。通常一個動作



圖二

電位從發生到結束不超過 1.6ms，因此我們取出的波形不該比這個時間長。如果發現波形持續時間超過 1.6ms，那麼這個波形可能不是動作電位，也可能是數個動作電位重疊的結果。

1.2 實驗訊號的製作

本論文的訊號並非真實的訊號，而是由人工的動作電位加上真實的雜訊。雖然利用人工訊號得到的結果比不上真實訊號來得有說服力，但是人工訊號比較容易控制，使得我們可以避開一些問題。例如我們可以控制動作電位的波形，使得上述的兩個假設一定會成立。人工訊號的另一個好處是我們可以預先得知正確的分類結果，使得我們能夠驗證本論文的方法是否可靠。如果以真實訊號做實驗，我們就無法得知正確的分類結果，這是因為目前存在的分類方法並無法 100% 正確地將動作電位分類。在 [4] 中利用了硬體的技術取得已知分類結果的真實訊號，這也是一種適合用來檢驗分類方法是否有效的訊號。但由於我們手上並沒有這類的訊號，因此只能採用人工訊號。本論文的人工動作電位依照下述的原則製造。

圖二的兩張圖是很常見的動作電位波形，動作電位發生時波形都是先下後上，不同的是，左邊的圖下去再回到水平線就結束了（比較常碰到的情況其實是上來時會稍微超過水平線一點點，但我們忽略這點差異），右邊的圖則是回到水平線後會繼續往上一段時間，然後才回到水平線，不過往上的振幅通常不超過往下的振幅。本論文採用的動作電位波形是圖二的左圖，原因是這種波形比較容易製作跟控制。雖然也有製作類似圖二右圖的動作電位，但沒有實際用在這篇論文的實驗中。除了形狀外，波形的持續時間也有限制。§1.1 有提過，一個動作電位的發生通常不超過 1.6ms，本論文的要求再稍微多一點。如果是左圖的動作電位，我們要求從發生到結束該在 0.8ms 內

完成；右圖的動作電位我們則分別要求下去和上來的波形都不超過 0.8ms。以圖二為例，此訊號是假設取樣頻率為每秒 20000 個點而製做的動作電位，每張圖的橫軸都是 32 個點。左圖不為 0 的部份有 12 個點，右圖有 16 個點。人工動作電位製造出來後，加上真實的雜訊，就是本論文採用的實驗訊號。雜訊是從真實訊號 中取得的，取得的方法應用到 thresholding 探測。首先用這個方法找到大動作電位，將這些大動作電位拿掉，剩下的就是雜訊。雜訊無法人工製造，必須由實驗訊號中取得，因為真正的神經雜訊很難用簡單的模型描述。本論文的實驗訊號由國立宜蘭大學生物機電學系的蔡孟利教授所提供，平均值是 0μV，標準差是 11μV。而實驗中採用的人工動作電位的最大振幅約從 80μV 到 120μV。

在第二章會介紹一個常見的分類法，此方法應用了 PCA 方法。所以第二章會介紹關於 PCA 方法的細節及優點，然後說明此方法如何用來分類動作電位。然候我們會說明 PCA 方法碰到的問題，這是本論文試著要去解決的問題。我們也會說明兩個 PCA 方法無法處理的問題。最後說明得到實驗訊號後，該先用什麼流程處理，然後再進入本論文提供的方法。

第二章、主成分分析

2.1 PCA 簡介

假設我們手上有 n 筆資料，命名為 $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_n$，而且對任意一個 i，$\boldsymbol{v}_i$ 都是 m 維向量。在我們分析 $\{\boldsymbol{v}_i\}_{i=1}^n$時，$m$ 越小，分析就越容易。但由於 m 的大小是不能控制的，所以為了分析的方便，我們通常會選擇一個程序 $\mathcal{L}$，使得 $\mathcal{L}$ 會從每一個 $\boldsymbol{v}_i$ 得到一個 s 維的向量 $\boldsymbol{w}_i$。例如我們可以選擇 $\mathcal{L}$ 是求平均值，則 $s = 1$，而對於每一個 i，w_i 是 $\boldsymbol{v}_i$ 的平均值。一般來說，$\{\boldsymbol{w}_i\}_{i=1}^n$ 包含的資訊量會比 $\{\boldsymbol{v}_i\}_{i=1}^n$ 少，但是前者通常可以表現出—甚至於突顯出—後者的某些特徵。我們希望 $\mathcal{L}$ 能夠表現出足夠的特徵，使得在只考慮這些特徵的情況下，對 $\{\boldsymbol{w}_i\}_{i=1}^n$ 的分析等同於對 $\{\boldsymbol{v}_i\}_{i=1}^n$ 的分析。求平均值雖然是一個簡單且普遍的程序，但是它表現的特徵並不足以應付大多數的需求。在多變量分析領域裡，有幾個程序供我們選擇，主成分分析 (Principal Component Analysis，簡稱 PCA) 即是其中一種。

PCA 是一種分析資料的方法，應用於許多地方—包含一些多變量分析的主題，如 factor analysis、discriminant analysis、cluster analysis。PCA 將 $\{\boldsymbol{v}_i\}_{i=1}^n$ 轉換成 $\{\boldsymbol{y}_k\}_{k=1}^m$，$\boldsymbol{y}_i$ 的維度都是 n。在 [1] 中，$\boldsymbol{y}_k$ 被稱作第 k 個主成分 (kth principal component)。有了 $\{\boldsymbol{y}_k\}_{k=1}^m$ 之後，我們可以定義

$$\boldsymbol{w}_j = (\boldsymbol{y}_{k_1}(j), \boldsymbol{y}_{k_2}(j), \ldots, \boldsymbol{y}_{k_s}(j)), \quad 1 \le k_1 < k_2 < \ldots k_s \le m, \quad 1 \le j \le n$$

其中 $\boldsymbol{y}_{k_l}(j)$ 是 $\boldsymbol{y}_{k_l}$ 的第 j 個元素。亦即 $\{\boldsymbol{w}_j\}_{i=j}^n$ 是由部分的主成分所決定的。根據 [1]，只要原始資料 $\{\boldsymbol{v}_i\}_{i=1}^n$ 不要太糟糕，我們只需要前面數個主成分就可以代表 $\{\boldsymbol{v}_i\}_{i=1}^n$。

以下是產生主成分的流程。我們先以一個矩陣表示原始資料 $\{\boldsymbol{v}_i\}_{i=1}^n$：

$$V = \begin{bmatrix} \boldsymbol{v}_1(1) & \boldsymbol{v}_1(2) & \ldots & \boldsymbol{v}_1(m) \\ \boldsymbol{v}_2(1) & \boldsymbol{v}_2(2) & \ldots & \boldsymbol{v}_2(m) \\ \cdot & \cdot & \ldots & \cdot \\ \cdot & \cdot & \ldots & \cdot \\ \boldsymbol{v}_n(1) & \boldsymbol{v}_n(2) & \ldots & \boldsymbol{v}_n(m) \end{bmatrix}_{n \times m}$$

其中 $\boldsymbol{v}_i(k)$ 是 $\boldsymbol{v}_i$ 的第 k 個元素。然後定義 $\boldsymbol{u}_j$ 是 V 的第 j 個行向量，因此我們得到 m 個 n 維向量 $\{\boldsymbol{u}_j\}_{j=1}^m$。為了簡化以下的描述，我們先假設

$$\sum_{k=1}^{n} \boldsymbol{u}_j(k) = 0, \quad 1 \le j \le m$$

亦即 $\boldsymbol{u}_j$ 的平均值都是 0。我們會在後面說明 $\boldsymbol{u}_j$ 的平均值不為 0 的情況。現在定義矩陣 $A=(a_{ij})$：

$$a_{ij}=\frac{\boldsymbol{u}_i^T\boldsymbol{u}_j}{n},\quad 1\le i,j\le m$$

對任意的 i、j，a_{ij} 就是 $\boldsymbol{u}_i$ 跟 $\boldsymbol{u}_j$ 的共變數 (covariance)，所以 A 被稱作一個共變矩陣 (covariance matrix)。A 會是一個 $m\times m$ 的對稱半正定矩陣，所以 A 是正交可對角化，這表示存在一個正交矩陣 (orthogonal matrix) P 使得

$$A=PDP^T$$

其中 D 是對角線矩陣，對角線元素 $\lambda_1,\lambda_2,\ldots,\lambda_m$ 為 A 的特徵值。在不失去一般性的情況下，我們令 $\lambda_1\ge\lambda_2\ge\ldots\ge\lambda_m$；如果 A 的特徵值不重複，那麼 P和D 就會唯一。由於 P 的行向量 $\{\boldsymbol{p}_k\}_{k=1}^m$ 組成了 $\mathbb{R}^m$ 的一組正則基底 (orthnormal basis)，所以任何一個 $\boldsymbol{x}\in\mathbb{R}^m$ 都可以寫成

$$\boldsymbol{x}=\sum_{k=1}^m\langle\boldsymbol{x},\boldsymbol{p}_k\rangle\boldsymbol{p}_k$$

而 $\boldsymbol{x}^TP$ 就是 $\boldsymbol{x}$ 在基底 $\{\boldsymbol{p}_k\}_{k=1}^m$ 下的表達向量。因此若定義

$$Y=VP$$

則 Y 的第 i 個列向量就是 $\boldsymbol{v}_i$ 在基底 $\{\boldsymbol{p}_k\}_{k=1}^m$ 下的表達向量；而 Y 的第 k 個行向量則是我們要的 $\boldsymbol{y}_k$，也就是第 k 個主成分*。

$\{\boldsymbol{y}_k\}_{k=1}^m$ 有一些好的性質：第一、$\boldsymbol{y}_k$ 的變異數為 λ_k，第二、對於任意 $k\ne l$，$\boldsymbol{y}_k$ 跟 $\boldsymbol{y}_l$ 的共變數都是 0。第一個性質暗示了 $\boldsymbol{y}_1$ 最能表現 $\{\boldsymbol{v}_i\}_{i=1}^n$ 的差距：當 $\boldsymbol{v}_i$ 與 $\boldsymbol{v}_j$ 差異比較大時，$\boldsymbol{y}_1(i)$ 與 $\boldsymbol{y}_1(j)$ 的差一定也會比較大。而第二個性質則指出對於任意 $i\ne j$，$\boldsymbol{y}_i$、$\boldsymbol{y}_j$ 互不相關。所以當我們分析 $\boldsymbol{y}_i$ 時，不需要考慮到其他的 $\boldsymbol{y}_j$。

之前我們假設對於任意的 j，$\boldsymbol{u}_j$ 的平均值都是 0。如果原始資料不滿足這個假設，那我們有兩種處理的方法，第一個方法就是將矩陣 V 的每一個行向量都扣掉該行的平均值，第二個方法則是直接用 V 做前面敘述的流程。用第二個方法要注意的是共變矩陣 $A=(a_{ij})$ 的計算：

$$a_{ij}=\mathrm{cov}(\boldsymbol{u}_i,\boldsymbol{u}_j)=\frac{(\boldsymbol{u}_i-\overline{\boldsymbol{u}}_i)^T(\boldsymbol{u}_j-\overline{\boldsymbol{u}}_j)}{n},\quad 1\le i,j\le m$$

---

* 在 matlab 的 statistics toolbox 內的函式 princomp() 和某些論文中，認為主成分是 P 的行向量而非 Y 的行向量。

還是要減去平均值，因此不論是用哪個方法，我們都會得到相同的共變矩陣 A。而且經過簡單的推論，我們可以發現不論是否減去平均值，上一段描述的 $\{\boldsymbol{y}\}_{k=1}^{m}$ 的兩個性質都會保留下來。我們使用的 PCA 方法都是第一種處理方法，也就是事先將 $\boldsymbol{V}$ 的行向量的平均值都扣掉。以下對幾組資料做 PCA，以此驗證一些 PCA 的性質。

第一個例子定義

$$V = \begin{bmatrix} 1 & 100 \\ 2 & 99 \\ . & . \\ . & . \\ 100 & 1 \end{bmatrix}$$

我們可以由 V 得到兩個主成分：$\boldsymbol{y}_1$ 和 $\boldsymbol{y}_2$，但其實 $\boldsymbol{y}_2 = \boldsymbol{0}$，也就是說 V 其實只有一個主成分。這是可以預測的：既然 V 的每個列向量的兩個元素加起來都等於 101，所以只要知道其中一個元素，另一個元素的值也可以輕易地推得。這個例子顯示如果原始資料有某種程度的相關度，則後面數個主成分就不是那麼重要。

第二個例子定義

$$V = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1.1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1.1 & 1 \end{bmatrix}$$

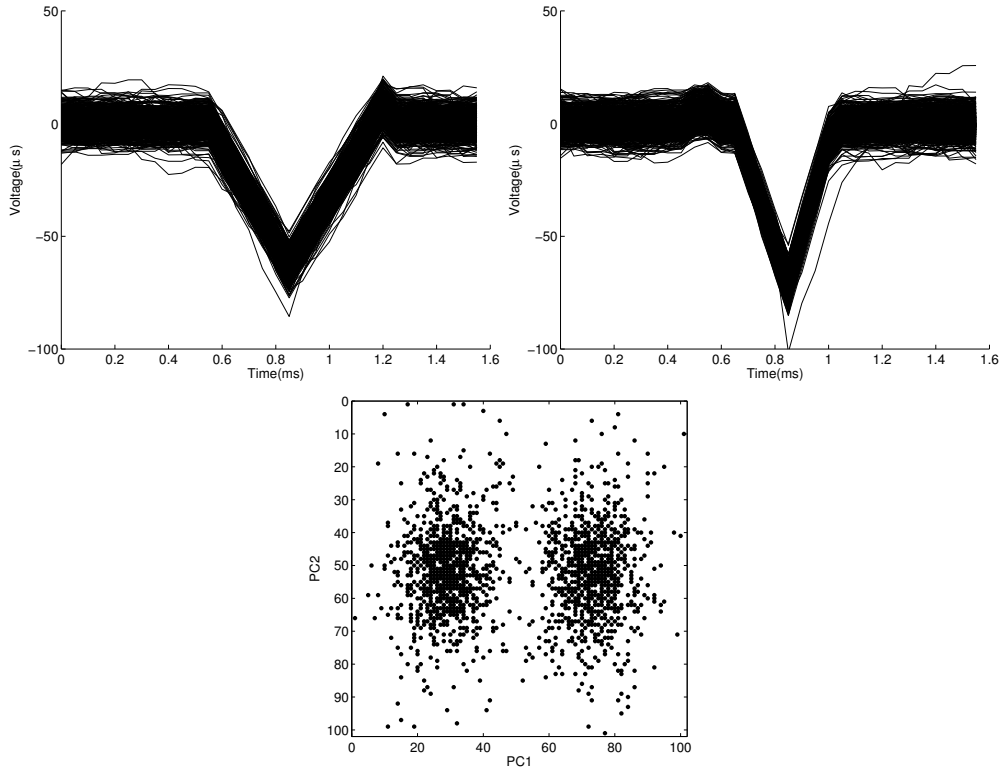
我們可以得到四個主成分：

$$\begin{cases} \boldsymbol{y}_1 = (-0.9997, -1.0509, 0.9997, 1.0509)^T \\ \boldsymbol{y}_2 = (-0.0354, 0.0354, -0.0354, 0.0354)^T \\ \boldsymbol{y}_3 = (0.025, -0.0238, -0.025, 0.0238)^T \\ \boldsymbol{y}_4 = (0, 0, 0, 0)^T \end{cases}$$

觀察 $\boldsymbol{y}_2$，我們可以發現 $\boldsymbol{y}_2(1) = \boldsymbol{y}_2(3)$，但是 $\boldsymbol{v}_1 \neq \boldsymbol{v}_3$。更進一步觀察，雖然 $\boldsymbol{v}_2$ 比 $\boldsymbol{v}_3$ 接近 $\boldsymbol{v}_1$，但是 $\boldsymbol{y}_2(1)$ 和 $\boldsymbol{y}_2(2)$ 卻恰恰是 $\boldsymbol{y}$ 中差距最遠的兩個數。不過觀察 $\boldsymbol{y}_1$，我們可以發現 $\boldsymbol{y}_1(1)$ 跟 $\boldsymbol{y}_1(2)$ 的差要比 $\boldsymbol{y}_1(1)$ 跟 $\boldsymbol{y}_1(3)$ 的差來得小。所以 $\boldsymbol{y}_1$ 正確的表現了資料間的差異。因此我們要是忽略 $\boldsymbol{y}_1$ 而直接由 $\boldsymbol{y}_2$ 來判斷向量的相似度，就會犯下兩種錯誤：一、將兩個差很遠的向量誤判成類似的向量，二、將兩個類似的向量誤判成差很遠的向量。所以沒考慮到第一主成分的分析很容易出錯*。

---

* 如果前面數個主成分對應的特徵值大小都差不多，甚至於相等，則這幾個主成分的重要性一樣，因此全都要考慮。

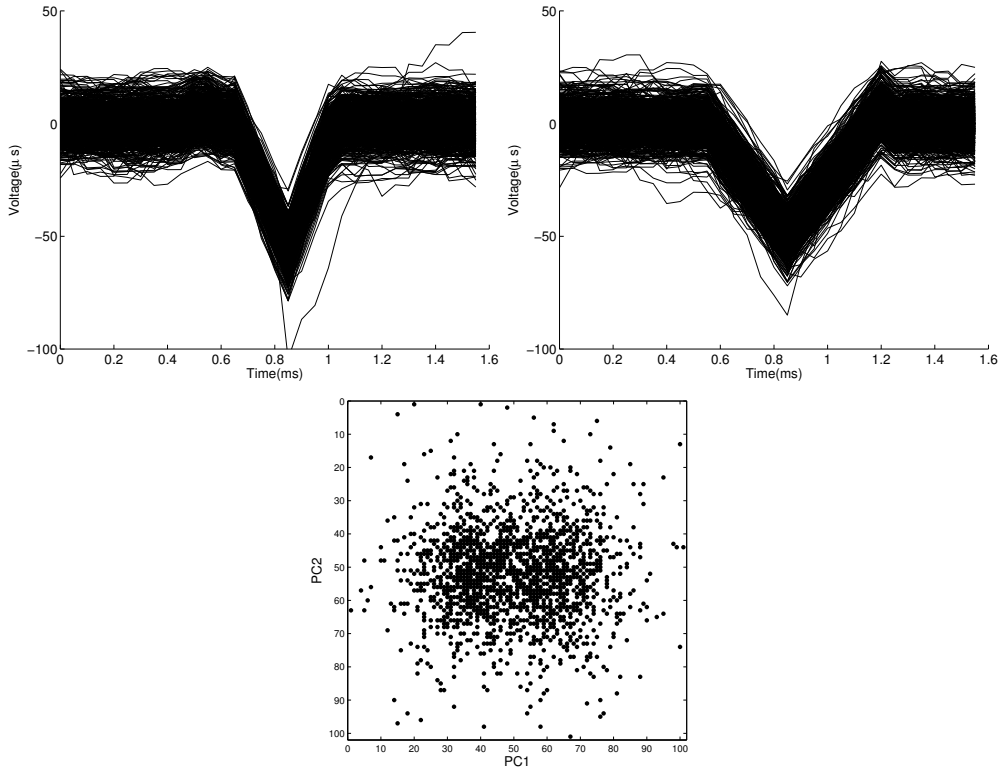


圖三

不過在實際應用時，我們卻不一定要考慮第一主成分。在上述的例子中，資料並沒有包含雜訊，換句話說，有意義的資料並沒有受到雜訊的干擾。如果資料受到雜訊的干擾，則主成分也會受到雜訊的干擾。如果雜訊的變異量太大，則第一主成分會受到嚴重的干擾，在這種情況下，第一主成分未必會帶來有用的資訊。

2.2 應用 PCA 分類動作電位

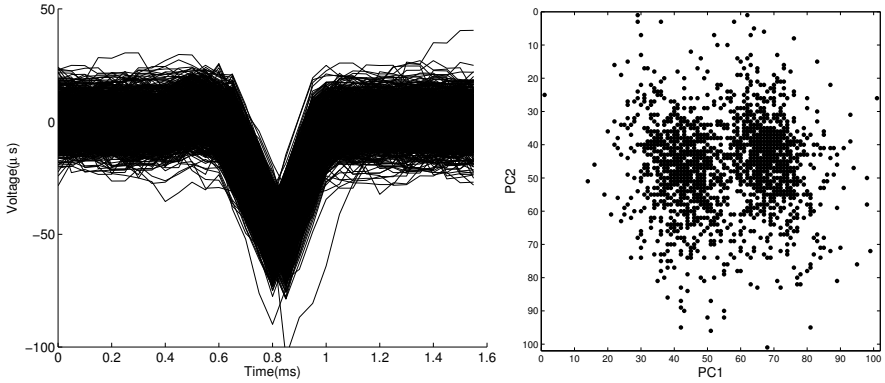
本論文雜訊來源訊號的取樣頻率為每秒 20000 個點，因此每個動作電位都用一個 32 維的向量表示。利用 PCA，我們可以得到 32 個主成分，然後用這些主成分來分類。一般來說，這 32 個主成分只有前三個—甚至於前兩個—被用來分類，其他的主成分通常無法用來分類。現在我們看看要如何使用這些主成分分類動作電位。圖三上面的兩張圖是兩種不同波形的動作電位加上雜訊後的結果，這兩類動作電位各有 1000 個，共有 2000 個動作電位，亦即，這裡有 2000 個 32 維的向量。計算出這些動作電位的主成分後，我們取出其中兩個—通常是第一跟第二主成分—用於分類。所以一個動作電位就表示成一個二維的向量，因此我們就有 2000 個二維向量，換個等價的說法，這 2000 的動作電位分別表示成平面上的 2000 的點。將這些點畫出來，得到的就是圖三的下圖。此圖的橫軸是第一主成分，縱軸是第二主成分。從這張圖可以看出 2000



圖四

個點明白地分成兩群，左半邊的點集跟右半邊的點集分別代表了這兩類動作電位。也就是說波形較靠近的動作電位，在這張 PCA 的圖上也會比較靠近。因此我們可以從這張圖清楚的將動作電位分成兩類。這個例子還不代表所有的實驗。有些動作電位是無法清楚的從 PCA 的圖形分類，圖四與第四章的圖六、圖七和圖八都是這樣的例子。這種情況和雜訊的大小有關，當雜訊較大時，前面數個主成分也會受到較大的雜訊干擾，分類也會較不順利。既然有雜訊的干擾，第一主成分就不再是非考慮不可的主成分。所以如果在第一主成分跟第二主成分的圖上無法分類動作電位，我們也可以試著從第一主成分跟第三主成分或是第二主成分跟第三主成分來分類動作電位。不過根據蔡孟利教授的經驗，如果在前三個主成分無法分類動作電位，後面的主成分也無法分類動作電位。這裡無法分類的情況有兩種，一種是兩類不同的動作電位在 PCA 的圖上聚在一起，我們或許能看出有兩類以上的動作電位，但很難從圖上分類動作電位。第二種情況稍微好一點，我們可以看出有幾類動作電位，但這些動作電位在圖上有一部份會混在一起，這個部份是不容易分開的。圖四的例子就是屬於第二種情況。

圖四的例子的雜訊稍微大一點。下圖是由第一主成分跟第二主成分構成的圖。雖然我們可以從圖看出有兩類動作電位，但這兩類動作電位有一部份是重疊的。碰到這種



圖五

情況，我們會這麼做：在圖上畫一條線，將圖分成兩邊，這兩邊就分別是兩類不相同的動作電位。這個操作並不客觀，分類結果會隨著操作者的不同而不同。在這個例子中，這兩類動作電位有大部分是分開的，只有少部份重疊，所以不同操作者的分類結果不會有太大的差異。但如果雜訊再大一點，則重疊的比率就會提高，這時不同操作者的分類結果就會有很大的差異。所以這些分類結果就不像圖三那麼可靠。本論文試圖解決這個問題。事實上，本論文的方法依然無法準確分類這兩群動作電位，但是此方法會客觀的提供一個估計正確率，讓使用者判斷分類結果是否可信賴。

2.3 其他的分類問題

有兩個問題在分類動作電位時會碰到，而且是 PCA 方法無法解決的。雖然這兩個問題不是本論文想解決的問題，但為了完整性，我們還是敘述一下。第一個問題是動作電位的重疊。當兩個大動作電位在接近的時間發生時，動作電位會因為彼此的疊合而導致那段時間波形發生很大的變化。由於這些動作電位跟一般神經元發出的動作電位有較大的差異，因此我們無法經由 PCA 方法分類這些動作電位。這些重疊的波形在 PCA 的圖上常常會形成孤立點，不靠近任何群聚的點集。我們得將這些波形拆成兩個動作電位才有辦法分類，但這件事無法由 PCA 完成，我們必須使用其他的方法分類這些波形。

另一個分類問題是對齊。我們希望動作電位是對齊的，換句話說，就是同一個神經元的動作電位發生的時間點是對齊的。例如我們用 32 維的向量記錄某個神經元的動作電位，我們會希望這些向量剛好都是在第 8 個元素開始不為 0。圖五的例子可以證明動作電位的對齊會對分類帶來嚴重的影響。圖五的左圖有 2000 個動作電位，是由同一種波形加上雜訊造成的。其中有 1000 個動作電位在橫軸平移了一個點。而圖五

的右圖是它們在第一主成分跟第二主成分下的點圖，如果從右圖來看，它們應該是兩類不同的動作電位。但實際上，它們是同一種動作電位，只是在時間點上有一點平移。這表示 PCA 方法對於時間平移是很敏感的，只要動作電位沒有對齊，分類結果就會受到嚴重的影響。對於時間敏感的方法，動作電位的對齊與否是非常重要的，一點誤差就可能讓我們將同一類動作電位誤分成兩類動作電位。一個簡單的對齊方法是最低點對齊，細節我們留待下節說明。

在本論文中，我們不考慮上述的兩個問題。因為我們採用的是人工訊號，所以我們可以控制這兩個問題不會發生。

2.4 真實訊號的處理

雖然本論文的實驗全都使是用人工訊號，但是我們還是在這裡交代一下在碰到真實訊號時，我們的處理流程。

當取得一筆神經訊號後，要分類動作電位之前，我們要先找出動作電位。這時使用上述的 threshholding 探測，設定一個門檻，將超過門檻的波形取出來。在這一步我們不檢查這些波形是否是真的動作電位，因為不是動作電位的波形會在分類時被挑出來。

接著我們將動作電位以向量的形式存下，我們需要取出 1.6ms 的訊號，因此向量長度會隨取樣頻率的不同而不同。本論文訊號的取樣頻率 (雜訊來源) 為每秒 20000 個點，因此動作電位就以 32 維的向量儲存。取出動作電位前，我們需要對齊，這裡提供一個簡單的對齊方法。當碰到一個超過門檻的波形後，找到波形最低點的位置，取出這個位置的前 15 個點以及後 16 點和那個位置本身，這就形成了一個 32 維的向量。這個方法找到的動作電位將會全部在波形的最低點對齊。

接著我們用 PCA 分類這些動作電位，找出本論文能處理的動作電位，當然這些動作電位得是 PCA 不能分類的。本論文的方法還不完全，只能針對某些動作電位做分類。 能處理的情況就類似圖四，有兩類動作電位的情況。所以我們先從 PCA 的圖中找出看起來有兩類動作電位，但是無法清楚分開的點集，或是無法分辨由一類或兩類動作電位所構成的點集。然後用三、四章描述的方法對這些點集所代表的動作電位做分類。

最後，有一個步驟是採取本論文方法需要做的，就是取出雜訊。正確地說，我們需要的是雜訊的幾個統計量。細節現在先不提，取出雜訊的方法在前面有提過，將所有

的動作電位拿掉，剩下的就是雜訊。根據 [2]，我們常常假設雜訊的統計量是固定的。以下是 [2] 的 §7.3 的翻譯：

> 一個可能不成立的假設就是雜訊的統計量。如果背景雜訊的統計量保持不變，分類結果就會一致。反之，受到較大的背景雜訊干擾的動作電位的分類就會有較多的錯誤。如果背景雜訊的統計量在一次實驗中發生多次變化，那麼準確的分類方法也會變得不準確。雖然我們應該根據雜訊模型的變化調整我們的分類方法，但由於這個作法實在過於複雜，所以我們只能假設雜訊的統計量不會發生太大的變化。

本論文假設雜訊的分佈成正規分佈 [3]。在第三章提出兩群點的分類法，在第四章應用這個分類法分類動作電位。

第三章、兩群點的分類

3.1 分類方法

令 $\boldsymbol{N_1}$ 和 $\boldsymbol{N_2}$ 是兩筆正規分佈 (normal distribution) 的訊號，平均值都是 0，標準差都是 σ。給定兩個相異實數 a 和 b，我們這樣定義訊號 $\boldsymbol{A}$ 和 $\boldsymbol{B}$：

$$\boldsymbol{A} = a + \boldsymbol{N_1}, \quad \boldsymbol{B} = b + \boldsymbol{N_2}$$

因此 $\boldsymbol{A}$ 和 $\boldsymbol{B}$ 都是標準差 σ 的正規分佈訊號，$\boldsymbol{A}$ 的平均值是 a，$\boldsymbol{B}$ 的平均值是 b。因為 $\boldsymbol{N_1}$ 和 $\boldsymbol{N_2}$ 是模擬雜訊，所以我們可以假設 $\boldsymbol{A}$ 和 $\boldsymbol{B}$ 中沒有相同的點值。接著我們將 $\boldsymbol{B}$ 接在 $\boldsymbol{A}$ 的後面，產生一個更長的訊號，然後隨意重排這個訊號，將最後得到的訊號命名為 $\boldsymbol{C}$。因為 $\boldsymbol{A}$ 跟 $\boldsymbol{B}$ 沒有相同的點值，所以 $\boldsymbol{C}$ 剛好可以分成 $\boldsymbol{A}$ 跟 $\boldsymbol{B}$ 兩部份。假設定義 $\boldsymbol{A}$ 在 $\boldsymbol{C}$ 中所佔的比率是 α，$\boldsymbol{B}$ 所佔的比率就是 $1-\alpha$。我們的問題是在 $\boldsymbol{C}$ 跟 σ 是已知，而 a、b 以及 α 都是未知的情況下，我們要如何將 $\boldsymbol{C}$ 分類為 $\boldsymbol{A}$ 和 $\boldsymbol{B}$？

理論上，$\boldsymbol{A}$ 跟 $\boldsymbol{B}$ 的分佈範圍是整個實數域，所以實數上的任何一個點都有可能屬於 $\boldsymbol{A}$ 跟 $\boldsymbol{B}$。因此我們無法找到一個方法能絕對正確地將 $\boldsymbol{C}$ 分成 $\boldsymbol{A}$ 和 $\boldsymbol{B}$。如果我們用某種方法將 $\boldsymbol{C}$ 中的某一個點分類到 $\boldsymbol{A}$ 或 $\boldsymbol{B}$ 時，我們應該考慮分類的正確率。在應用上，如果 $\boldsymbol{A}$ 跟 $\boldsymbol{B}$ 是真實訊號，那麼 $\boldsymbol{A}$ 跟 $\boldsymbol{B}$ 的分佈範圍通常會分別落在兩個有限區間。只要 a 和 b 距離夠遠，我們就能夠幾乎正確地將 $\boldsymbol{C}$ 分類成 $\boldsymbol{A}$ 跟 $\boldsymbol{B}$；反之，如果 a 和 b 的距離太靠近使得 $\boldsymbol{A}$ 跟 $\boldsymbol{B}$ 的分佈範圍有重疊，那麼我們就無法在不發生錯誤的情況下將 $\boldsymbol{C}$ 分類成 $\boldsymbol{A}$ 和 $\boldsymbol{B}$—至少落在重疊部分的點會分錯。因為 $\boldsymbol{A}$ 跟 $\boldsymbol{B}$ 是正規分佈，所以 $\frac{|a-b|}{\sigma}$ 暗示了分類正確率的極限。如果 $\frac{|a-b|}{\sigma} > 6$，那麼分類的正確率會超過 99%；但要是 $\frac{|a-b|}{\sigma} < 1$，分類的正確率就會接近 50%—也就是分類正確率的下界。以下提出一個簡單的想法，透過這個想法幫助我們了解一些事實。根據這些事實，我們再找尋新的方法。

在不失去一般性的情況下，我們假設 $a > b$。既然 $\boldsymbol{A}$ 和 $\boldsymbol{B}$ 是由 a 和 b 加上正規分佈的雜訊造成的，那 $\boldsymbol{A}$ 會有超過 99% 的點落在 $[a-3\sigma, a+3\sigma]$，$\boldsymbol{B}$ 則有超過 99% 的點落在 $[b-3\sigma, b+3\sigma]$。我們忽略那不到 1% 的誤差，假設 $\boldsymbol{A}$ 落在 $[a-3\sigma, a+3\sigma]$，$\boldsymbol{B}$ 落在 $[b-3\sigma, b+3\sigma]$。依據這個假設，我們採取消去法，將所有小於 $a-3\sigma$ 的點分類到 $\boldsymbol{B}$；所有大於 $b+3\sigma$ 的點分類到 $\boldsymbol{A}$。這的流程只是一個想法，無法實作，因為我們不知道 a 跟 b；但這個想法帶來一些對我們有幫助的事實。

首先，這個想法有一個缺點，就是有一部份的點不會被分類，這些點就是落在 $P=(a-3\sigma,b+3\sigma)$ 內的點。直觀來看，不在區間 P 的點就是屬於「好」分類的點，因為如果照上一段的想法分類，這些點的分類正確率會高達 99%。所以光是看「可分類」的點的分類正確率，這個分類法似乎是很優秀；但如果 $\boldsymbol{C}$ 有超過一半的點屬於 P，那麼我們就等於有超過一半的點沒有分類，因此分類正確率不到 50%—甚至於比理論下界還低。這裡我們看到兩種分類錯誤，一種是有分類，但分錯了；另一種則是沒分類。在此重複分類的錯誤不會發生，因為在此分法下，重複分類的錯誤跟未分類的錯誤不可能同時發生。為了描述的方便，在以下的討論中，當我們提到正確率時，我們是指分類正確的點在能分類的點中所佔的比率。而未分類的點所佔的比率就稱作未分類比率。

現在看看未分類比率跟分類正確率的關係。當 $\frac{|a-b|}{\sigma}>6$，P 是空集合，所以沒有未分類的錯誤，而分類正確率會超過 99%，因此分類結果幾乎正確。如果 $\frac{|a-b|}{\sigma}<6$，$\boldsymbol{C}$ 中會有部分的點不能分類，比值越小，未分類比率就越大。如果要降低未分類比率，我們這麼修改：將所有小於 $a-2\sigma$ 的點分類到 $\boldsymbol{B}$；所有大於 $b+2\sigma$ 的點分類到 $\boldsymbol{A}$。這時只有落在 $\hat{P}=(a-2\sigma,b+2\sigma)$ 的點沒有分類，和原來的作法相比，這個作法的未分類比率會下降。但是這會付出代價—分類正確率降到 97%。我們可以這樣想：不落在 $\hat{P}$ 但落在 P 的點比不落在 P 的點「差」，這些點被分類後只能保證有 97% 的正確率。所以我們降低未分類比率時，分類正確率也會下降。反過來說也是對的：當我們提高分類正確率，未分類比率也會上升。所以我們只能在高分類正確率跟低未分類比率中做一個選擇，想要提高分類正確率又不降低未分類比率是不可能的。

雖然這個想法沒有實用價值，但給了我們一個方向。新的演算法有兩個目標，第一個目標是可以控制，由於分類正確率跟未分類比率是互斥的，所以我們想找到一個能依需要而調整它們的方法。而第二個目標是能估計分類正確率。既然我們無法正確分類，那麼就應該有一個分類正確率告訴使用者分類結果有多可靠。不可分類的點可以直接從分類結果得到，所以我們不需要估計未分類比率。但是由於我們事先並不知道正確的分類結果，因此分類正確率只能用估計的，無法精確得到。但只要估計的正確率接近真正的分類正確率，那麼就能告訴使用者分類結果是否可靠。

現在令 $\overline{\mu}$ 是 $\boldsymbol{C}$ 的平均值，$\overline{\sigma}$ 是 $\boldsymbol{C}$ 的標準差。再令

$$f_a(x)=\frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left(\frac{x-a}{\sigma}\right)^2\right]$$

是 $\boldsymbol{A}$ 的機率分佈函數，

$$f_b(x) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left(\frac{x-b}{\sigma}\right)^2\right]$$

是 $\boldsymbol{B}$ 的機率分佈函數。現在假設 $\boldsymbol{C}$ 的主要分佈範圍在 $[\overline{\mu}-k\overline{\sigma},\overline{\mu}+k\overline{\sigma}]$，$\boldsymbol{A}$ 的主要分佈範圍在 $[a-\ell\sigma,a+\ell\sigma]$，$\boldsymbol{B}$ 的主要分佈範圍在 $[b-\ell\sigma,b+\ell\sigma]$。由於 $\boldsymbol{C}$ 的分佈上界接近 $\boldsymbol{A}$ 的分佈上界，$\boldsymbol{C}$ 的分佈下界接近 $\boldsymbol{B}$ 的下界；所以推得 $\boldsymbol{A}$ 分佈於 $[\overline{\mu}+k\overline{\sigma}-2\ell\sigma,\overline{\mu}+k\overline{\sigma}]$，而 $\boldsymbol{B}$ 分佈於 $[\overline{\mu}-k\overline{\sigma},\overline{\mu}-k\overline{\sigma}+2\ell\sigma]$。在此 k 和 ℓ 尚未決定，因為我們要讓它們可以調整。由於雜訊呈正規分佈，所以 k 和 ℓ 設為 1–3 都在可接受範圍。現在我們要開始分類。

給定兩個正數 s_1 和 s_2，如果 $\boldsymbol{A}$ 分佈於 $[\overline{\mu}+k\overline{\sigma}-s_1,\overline{\mu}+k\overline{\sigma}]$，$\boldsymbol{B}$ 分佈於 $[\overline{\mu}-k\overline{\sigma},\overline{\mu}-k\overline{\sigma}+s_2]$。我們就可將所有小於 $\overline{\mu}+k\overline{\sigma}-s_1$ 的點分到 $\boldsymbol{B}$，所有大於 $\overline{\mu}-k\overline{\sigma}+s_2$ 的點分到 $\boldsymbol{A}$。通常 $\boldsymbol{A}$ 不會恰好只落在 $[\overline{\mu}+k\overline{\sigma}-s_1,\overline{\mu}+k\overline{\sigma}]$，但會有一部分落在這個區間。如果我們知道屬於 $\boldsymbol{A}$ 但不落在這個區間的點的比率，或反過來說，落在這個區間的比率，我們就可以知道不屬於 $\boldsymbol{B}$ 卻被分到 $\boldsymbol{B}$ 的比率。同樣地，$\boldsymbol{B}$ 也不會恰好只落在 $[\overline{\mu}-k\overline{\sigma},\overline{\mu}-k\overline{\sigma}+s_2]$，但也會有一部分落在這個區間。要是我們也知道落在這個區間比率，那我們就可以知道不屬於 $\boldsymbol{A}$ 卻被分到 $\boldsymbol{A}$ 的比率。這兩個比率加起來就是有分類，但卻分類錯誤的比率。當 s_1 和 s_2 改變時，未分類比率及分類正確率也會改變，因此我們有必要了解 s_1 和 s_2 會如何影響未分類比率及分類正確率。為了描述的方便，我們假設 α 和 $a-b$ 是已知的。在 §3.2 會說明如何估計這兩個值。

已知 s_1 和 s_2，未分類比率差不多等於

$$\int_{\overline{\mu}+k\overline{\sigma}-s_1}^{\overline{\mu}-k\overline{\sigma}+s_2} (\alpha f_a(x) + (1-\alpha) f_b(x))\, dx$$

這個積分只有在 $\overline{\mu}-k\overline{\sigma}+s_2 > \overline{\mu}+k\overline{\sigma}-s_1$ 才是正值，如果這個值小於 0，不可分類比率就是 0。從這個積分式來看，要降低未分類比率就要降低 s_1+s_2，不過這會導致分類正確率下降。當 $s_1+s_2=2k\overline{\sigma}$ 時，未分類比率就是 0。如果繼續降降低 s_1+s_2，未分類比率不會再下降；但分類正確率卻會繼續下降，而且也會開始產生重複分類的錯誤，所以我們要求 $s_1+s_2 \geq 2k\overline{\sigma}$。

再是分類正確率。由於知道分類正確率就等於知道分類錯誤率，因此我們估計的是分類錯誤率。有兩種分類錯誤：一是將該分到 $\boldsymbol{A}$ 的點分到 $\boldsymbol{B}$，二是將該分到 $\boldsymbol{B}$ 的點

分到 $\boldsymbol{A}$。該分到 $\boldsymbol{A}$ 但被分到 $\boldsymbol{B}$ 的點的比率為

$$\alpha \int_{a-\ell\sigma}^{\overline{\mu}+k\overline{\sigma}-s_1} f_a(x)\,dx \tag{1}$$

這個積分只有在 $\overline{\mu}+k\overline{\sigma}-s_1 > a-\ell\sigma$ 時才是正值，代入 $\overline{\mu}=\alpha a+(1-\alpha)b$ 會得到

$$s_1 < k\overline{\sigma}+\ell\sigma-(1-\alpha)(a-b) \tag{2}$$

所以只有在 (2) 成立時 (1) 才是正值。如果 (1) 不是正值，則幾乎不會有將該分到 $\boldsymbol{A}$ 的點分到 $\boldsymbol{B}$ 的錯誤。類似地，將該分到 $\boldsymbol{B}$ 的點分到 $\boldsymbol{A}$ 的錯誤比率為

$$(1-\alpha)\int_{\overline{\mu}-k\overline{\sigma}+s_2}^{b+\ell\sigma} f_b(x)\,dx \tag{3}$$

這個積分只有在 $b+\ell\sigma > \overline{\mu}-k\overline{\sigma}+s_2$ 時才是正值。代入 $\overline{\mu}=\alpha a+(1-\alpha)b$ 會得到

$$s_2 < k\overline{\sigma}+\ell\sigma-\alpha(a-b) \tag{4}$$

所以只有在 (4) 成立時 (3) 才是正值。如果 (3) 不是正值，則幾乎不會有將該分到 $\boldsymbol{B}$ 的點分到 $\boldsymbol{A}$ 的錯誤。從 (1)、(3) 式來看，要提高分類正確率就要提高 s_1 和 s_2，不過這也會提高未分類比率。在 $s1=k\overline{\sigma}+\ell\sigma-(1-\alpha)(a-b)$ 且 $s_2=k\overline{\sigma}+\ell\sigma-\alpha(a-b)$ 時，分類正確率幾乎是 100%。如果繼續提高 s_1 和 s_2，不僅不會提高分類正確率，反而會提高未分類比率。所以我們要求 $s1 \le k\overline{\sigma}+\ell\sigma-(1-\alpha)(a-b)$ 且 $s_2 \le k\overline{\sigma}+\ell\sigma-\alpha(a-b)$。加上前一段的結論，我們可以知道 s_1 和 s_2 的上下界：

$$s1+s2 \ge 2k\overline{\sigma}$$

$$s1 \le k\overline{\sigma}+\ell\sigma-(1-\alpha)(a-b)$$

$$s2 \le k\overline{\sigma}+\ell\sigma-\alpha(a-b)$$

我們讓 s_1 跟 s_2 在這個範圍內變動。接著我們要估計分類的正確率（錯誤率），只要能知道這個比率，我們就能夠選擇一個可接受的分類結果。

雖然分類錯誤率是 (1) 加 (3)，但 a 和 b 未知，所以不能用 (1) 和 (3) 估計分類錯誤率。我們需要用新的公式估計分類錯誤率。根據之前的定義，我們知道 (1) 等於

$$\alpha \int_{-\ell\sigma}^{k\overline{\sigma}-(1-\alpha)(a-b)-s_1} f_0(x)\,dx \tag{5}$$

其中

$$f_0(x) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left(\frac{x-0}{\sigma}\right)^2\right]$$

(2) 等於

$$(1-\alpha)\int_{-k\overline{\sigma}+\alpha(a-b)+s_1}^{\ell\sigma} f_0(x)\,dx \tag{6}$$

又 α 和 $a-b$ 是能估計的，所以 (5) 和 (6) 就能計算。而分類錯誤率就是 (5) 加 (6)。

這個錯誤率只是一個估計值，不等於真正的錯誤率。有兩個因素會影響估計值的準確性，一個是 σ 跟 $a-b$ 的比率。在 §3.2 會介紹估計 α 跟 $a-b$ 的方法，這兩個估計值也不等於真正的 α 以及 $a-b$。它們的誤差會隨著 $\frac{|a-b|}{\sigma}$ 的縮小而變大，所以估計錯誤率的誤差也會變大。另一個會影響估計錯誤率的因素是 $\boldsymbol{N_1}$ 跟 $\boldsymbol{N_2}$ 的分佈及長度。上述的的推論要求 $\boldsymbol{N_1}$ 及 $\boldsymbol{N_2}$ 是正規分佈，但實際的問題不一定是正規分佈，$\boldsymbol{N_1}$ 及 $\boldsymbol{N_2}$ 離正規分佈越遠，估計錯誤率的誤差就越大。幸運的是我們的推論以積分為主，所以只要 $\boldsymbol{N_1}$ 跟 $\boldsymbol{N_2}$ 不要離正規分佈太遠，誤差就不會太大。$\boldsymbol{N_1}$ 跟 $\boldsymbol{N_2}$ 的長度也會對這個方法造成影響。換個說法，$\boldsymbol{A}$ 和 $\boldsymbol{B}$ 的點數會對這個方法造成影響。理由是上述推論 (以及下一節的推論) 中最重要的兩個數值就是 $\overline{\mu}$ 和 $\overline{\sigma}$，當 $\boldsymbol{N_1}$ 或 $\boldsymbol{N_2}$ 太短時，這兩個數就無法採用，我們的分類就會受到影響。所以要使用這個方法，$\boldsymbol{N_1}$ 和 $\boldsymbol{N_2}$ 的點數就不能太少。

現在我們回頭看看 k 和 ℓ。在上述推導中，分類的的界線為 $\overline{\mu}+k\overline{\sigma}-s_1$ 以及 $\overline{\mu}-k\overline{\sigma}+s_2$。為了方便，我令 r_1 為前一個數，r_2 為後一個數。配合 s_1 和 s_2 上界的條件，r_1 要大於 $\overline{\mu}-\ell\sigma+(1-\alpha)(a-b)$，$r_2$ 要小於 $\overline{\mu}+\ell\sigma-\alpha(a-b)$。這兩個數都與 k 無關。當 $s_1+s_2=2k\overline{\sigma}$，$r_1-r_2=0$。也就是說，只要 $r_1=r_2$，就不會有未分類的點。因此雖然表面上 k 會對分類的上下界—s_1 和 s_2—造成影響，但實際上真正用來分類的兩個數—r_1 和 r_2—的上下界卻與 k 無關。從這點來看，我們似乎可以隨意決定 k。但是我們還是讓 $k=3$比較好，因為在估計分類錯誤率時需要用到 k，所以 k 不能太小，否則錯誤率就會估錯。令 $k=3$，對錯誤率的估計值不會有太大的影響。而 ℓ 會影響到 r_1 和 r_2 的上下界，所以不能隨便決定。先前我們假設 $\boldsymbol{A}$ 分佈於 $[a-\ell\sigma, a+\ell\sigma]$，$\boldsymbol{B}$ 分佈於 $[b-\ell\sigma, b+\ell\sigma]$。理論上，如果 $\ell\neq\infty$，估計的分類錯誤率就會有誤差。當 $\ell=2$ 時，增加的誤差約為 5%，$\ell=3$ 時，增加的誤差約為 1%。我們不能真的讓 $\ell=\infty$，因為這樣 r_1 和 r_2 就會跑到正負無窮大。通常 $\ell=2$ 帶來的 5% 誤差是可以接受的，$\ell=3$ 稍嫌太大，會帶來一些負面效應。ℓ 越大，估計分類錯誤率

會越大，亦即錯誤率會被高估。舉例來說，當 $\ell = 2$ 時，我們從 (5) 和 (6) 得到錯誤率為 5%；但 $\ell = 3$ 時，這個錯誤率可能會提高到 10%。考慮到估計錯誤率的誤差，在 $\ell = 2$ 時，我們只能說真實的錯誤率大約就是估計的錯誤率，可能較大，也可能較小。但在 $\ell = 3$ 時，我們可以說真實的錯誤率不會超過估計的錯誤率，因為在 $\ell = 3$ 時，錯誤率往往會被高估。這是優點，也是缺點。優點是當估計錯誤率很小時，我們可以相信真實的錯誤率也會很小。缺點是有時會發生真實錯誤率很小，估計錯誤率很大的現象，使得本來可以信賴的分類結果變得不可信賴。

3.2 α 及 $a-b$ 的估計

我們要估計 α 和 $a-b$。根據平均值的定義，我們知道

$$\overline{\mu} = \alpha a + (1-\alpha)b \tag{7}$$

從標準差的定義，我們知道

$$\overline{\sigma}^2 = \int_{-\infty}^{\infty} (\alpha f_a(x) + (1-\alpha) f_b(x))(x-\overline{\mu})^2 \, dx \tag{8}$$

但是光靠 (7) 和 (8) 無法估計 α 和 $a-b$，所以我們需要其他的條件。令 N 是 $\boldsymbol{C}$ 的長度，定義

$$r = \frac{1}{N} \sum (x-\overline{\mu})^3$$

已知 $\boldsymbol{C}$ 是 $\boldsymbol{A}$ 跟 $\boldsymbol{B}$ 的混合，且 $f_a(x)$ 是 $\boldsymbol{A}$ 的機率分佈函數，$f_b(x)$ 是 $\boldsymbol{B}$ 的機率分佈函數，我們可得

$$r = \int_{-\infty}^{\infty} (\alpha f_a(x) + (1-\alpha) f_b(x))(x-\overline{\mu})^3 \, dx \tag{9}$$

有 (7)、(8)、(9) 三式，我們就能估計 α 和 $a-b$。

首先把 (7) 代入 (8)，然後做等號右邊的積分*得到

$$\begin{aligned}
\overline{\sigma}^2 &= 2\alpha^2 ab - \alpha^2 b^2 + \alpha b^2 + \alpha a^2 - \alpha^2 a^2 + \sigma^2 - 2\alpha ab \\
&= \sigma^2 + \alpha(a-b)^2 - \alpha^2 (a-b)^2 \\
&= \sigma^2 + \alpha(1-\alpha)(a-b)^2
\end{aligned}$$

* 這是 Maple 計算的結果。

把 (7) 代入 (9) 後做等號右邊的積分得到

$$\begin{aligned} r =& \alpha(2\alpha^2a^3 + a^3 + 3ab^2 - 3a^2b + 9\alpha a^2b + 3\alpha b^3 \\ & - 9\alpha ab^2 - 3\alpha a^3 - b^3 - 2\alpha^2b^3 + 6\alpha^2ab^2 - 6\alpha^2a^2b) \\ =& \alpha(2\alpha^2(a^3 - 3a^2b + 3ab^2 - b^3) - 3\alpha(a^3 - 3a^2b + 3ab^2 - b^3) + (a^3 - 3a^2b + 3ab^2 - b^3)) \\ =& \alpha(2\alpha^2(a-b)^3 - 3\alpha(a-b)^3 + (a-b)^3) \\ =& \alpha(2\alpha - 1)(\alpha - 1)(a-b)^3 \end{aligned}$$

所以我們有

$$\overline{\sigma}^2 - \sigma^2 = \alpha(1-\alpha)(a-b)^2$$

和

$$r = \alpha(2\alpha - 1)(\alpha - 1)(a-b)^3$$

兩式。令 $d = a - b$，$q = \overline{\sigma}^2 - \sigma^2$，得到

$$q = \alpha(1-\alpha)d^2 \tag{10}$$

和

$$r = \alpha(2\alpha - 1)(\alpha - 1)d^3 \tag{11}$$

因為前面假設 $a > b$，所以 $d > 0$，由 (10) 可知 $q > 0$。 又，α 跟 $(1-\alpha)$ 也大於 0，所以從 (10) 可得

$$d = \sqrt{\frac{q}{\alpha(1-\alpha)}} \tag{12}$$

將 (12) 代入 (11) 的平方得到

$$r^2 = \alpha^2(2\alpha-1)^2(\alpha-1)^2\frac{q^3}{\alpha^3(1-\alpha)^3} = \frac{(2\alpha-1)^2q^3}{\alpha(1-\alpha)}$$

因為 $\alpha(1-\alpha) \neq 0$，所以

$$\alpha r^2 - \alpha^2 r^2 = 4\alpha^2q^3 - 4\alpha q^3 + q^3$$

然後

$$(4q^3 + r^2)\alpha^2 - (4q^3 + r^2)\alpha + q^3 = 0$$

用公式求根

$$\begin{aligned}\alpha =& \frac{(4q^3+r^2) \pm \sqrt{(4q^3+r^2)^2 - 4(4q^3+r^2)q^3}}{2(4q^3+r^2)} \\ =& \frac{(4q^3+r^2) \pm r\sqrt{4q^3+r^2}}{2(4q^3+r^2)} \\ =& \frac{1}{2}(1 \pm \frac{r}{\sqrt{4q^3+r^2}})\end{aligned}$$

由於 $q > 0$ 且 $r^2 > 0$，所以分母不為 0。解出來的 α 有兩個值，都是正數且相加為 1。α 是 $\boldsymbol{A}$ 在 $\boldsymbol{C}$ 的比例，但是這個比例是 α 或是 $1-\alpha$ 都不會影響到 (10) 和 (12) 式。換句話說，我們不能從這裡知道 α 的解是哪一個。這個問題留到最後解決，因為這不影響 d 的解。令

$$\alpha = \frac{1}{2}(1 + \frac{r}{\sqrt{4q^3+r^2}}), \quad 1-\alpha = \frac{1}{2}(1 - \frac{r}{\sqrt{4q^3+r^2}})$$

或令

$$\alpha = \frac{1}{2}(1 - \frac{r}{\sqrt{4q^3+r^2}}), \quad 1-\alpha = \frac{1}{2}(1 + \frac{r}{\sqrt{4q^3+r^2}})$$

也可以。將 α 和 $1-\alpha$ 代入 (10) 會得到

$$q = \frac{1}{4}(1 - \frac{r^2}{4q^3+r^2})d^2 = \frac{q^3}{4q^3+r^2}d^2$$

然後可解出

$$d = \pm\frac{\sqrt{4q^3+r^2}}{q}$$

d 是正數，所以

$$d = \frac{\sqrt{4q^3+r^2}}{q}$$

因此 d 的解就求出來了。

現在回頭處理 α 的問題。從上述的推論可知如果要求出 α，我們必須知道 a 和 b。但我們不知道 a 跟 b，所以無法從正規的數學計算得知 α。但是我們可以利用一些特殊的方式求得 α。令

$$U = \sum_{x \in [\overline{\mu}, \overline{\mu}+\overline{\sigma}]} x, \quad L = \sum_{x \in [\overline{\mu}-\overline{\sigma}, \overline{\mu}]} x$$

由於 $\boldsymbol{A}$ 跟 $\boldsymbol{B}$ 是 (或接近) 正規分佈，如果 $\boldsymbol{A}$ 所佔的比例比較大時，$\overline{\mu}$ 會接近 a，這時 U 會大於 L。類似地，如果 $\boldsymbol{B}$ 所佔的比例比較大時，$\overline{\mu}$ 會比較接近 b，這時 U 會小於 L。因此當 $U > L$ 時

$$\alpha = \frac{1}{2}(1 + \frac{r}{\sqrt{4q^3 + r^2}})$$

當 $U < L$ 時

$$\alpha = \frac{1}{2}(1 - \frac{r}{\sqrt{4q^3 + r^2}})$$

3.3 實驗

實驗目的是檢查當雜訊是正規分佈時，前述的推論是否成立。有兩件事是我們想確認的，一個是 s_1 和 s_2 變大或變小時，分類正確率的變化是否如 §3.1 的描述。另一件事則是估計錯誤率是否可靠。所以我們選擇五組不同的 s_1 和 s_2 實驗，其中包含了上界與下界。以下用兩個點 a 和 b 製作點集 $\boldsymbol{A}$ 和 $\boldsymbol{B}$，$\boldsymbol{A}$ 和 $\boldsymbol{B}$ 是由 a 和 b 分別加上正規分佈的雜訊造成的。我們令 $\boldsymbol{A}$ 有 4000 個點，$\boldsymbol{B}$ 有 1000 個點，雜訊的平均值為 0，標準差為 1。我們想確認 a 和 b 的距離會如何影響這個方法的效率，因此我們用三組不同的 a 和 b 來實驗。

a 和 b 有三組，分別是 $a = 4, b = 3$、$a = 6, b = 4$、 和 $a = 3, b = 0$；s_1 和 s_2 有五組，包含上下界。由於前面關於下界的推導只有 $s_1 + s_2$ 的下界，所以在這個實驗選擇的下界只是令未分類比率達到 0%，而並沒有其他的根據說明這個選擇會是最好。但是 $\boldsymbol{A}$ 的點數比 $\boldsymbol{B}$ 多，所以當 $s_1 + s_2$ 改變時，我們會讓 s_1 改變的幅度大於 s_2 改變的幅度。我們將結果分成 3 個 3×5 的表格。表格的每列代表一組 a 和 b 的結果，每行代表一組 s_1 和 s_2 的結果。由於 s_1 和 s_2 的數據會隨著實驗的不同而不同，所以在此不列出數據，而是按照大小放入表格內；表格第一行是 s_1 和 s_2 在上界的數據，最後一行則是 s_1 和 s_2 在下界的數據。第一個表格的欄位記錄的是未分類的點數，這個數目除以 5000 就是未分類比率。

	上界				下界
$a = 4, b = 3$	4913	4837	3868	2642	0
$a = 6, b = 4$	4145	3898	2490	1541	0
$a = 3, b = 0$	2601	2343	1251	739	0

第二個表格的欄位有三個數字，用 / 隔開，第一個數字是被分到 $\boldsymbol{A}$ 的點數，第二個數字是不屬於 $\boldsymbol{A}$ 卻被分到 $\boldsymbol{A}$ 的點數，而第三個數字則是所估計不屬於 $\boldsymbol{A}$ 但被分到

$\boldsymbol{A}$ 的點數。估計錯誤率的分母是 $\boldsymbol{A}$ 和 $\boldsymbol{B}$ 的總點數，所以用 5000 乘上這個比率就是估計分錯的點數，也就是欄位內的第三個數字。

	上界				下界
$a=4,b=3$	53/1/0	98/2/20	648/24/211	1325/72/399	2678/229/803
$a=6,b=4$	680/1/0	860/2/8	1897/19/68	2478/39/121	3289/105/238
$a=3,b=0$	1910/2/0	2096/2/6	2892/10/45	3228/15/77	3590/40/143

第三個表格類似第二個表格，只是 $\boldsymbol{A}$ 變成 $\boldsymbol{B}$。

	上界				下界
$a=4,b=3$	34/8/0	65/16/45	484/224/483	1033/576/917	2322/1551/1840
$a=6,b=4$	175/5/0	242/9/43	613/92/412	981/270/761	1711/816/1527
$a=3,b=0$	489/9/0	561/15/33	857/69/276	1033/154/489	1410/450/956

經由這個實驗，我們知道四個事實。第一、當 s_1 和 s_2 非常靠近上界時，估計錯誤率會小於真實的錯誤率。因為這個實驗令 $\ell=3$，因此錯誤率估計有約 0.005 的誤差。這個誤差只有在錯誤率極小，也就是 s_1 和 s_2 極靠近上界時才會造成負面影響。錯誤率估計不會影響分類的結果，但如果低估錯誤率，可能會讓人誤判分類結果的可靠性。為了避開這個問題，我們可以在估計錯誤率低於 0.01 時，就判定估計錯誤率為 0.01。因為當錯誤率高於 0.01，0.005 誤差的影響不會太嚴重。當 ℓ 發生變化，我們就用不同的界限修正估計錯誤率，因此我們可以保證估計錯誤率不會低於真實錯誤率。

第二、a 和 b 越靠近，分類就越困難。看看 $a=4,b=3$ 的分類結果，此時 $|a-b|$ 剛好等於雜訊標準差，不管我們再怎麼調整 s_1 和 s_2，未分類的比率及分類的錯誤率還是偏高。因此只有在 a 和 b 距離超過一個標準差時，我們才有辦法用夠高的準確率分開 $\boldsymbol{A}$ 和 $\boldsymbol{B}$。

第三、當要分類的兩個點集的數目差很多時，點數較少的集合通常較不易分類。以此實驗為例，$\boldsymbol{B}$ 的點數只有 $\boldsymbol{A}$ 的四分之一，所以 $\boldsymbol{B}$ 的分類結果都很糟糕。直觀來看，如果將 $\boldsymbol{B}$ 視為有意義的訊號，將 $\boldsymbol{A}$ 當成干擾 $\boldsymbol{B}$ 的雜訊，那麼 $\boldsymbol{B}$ 就等於受到強烈的雜訊干擾，要將 $\boldsymbol{B}$ 分出來就變得非常困難。反過來說，對 $\boldsymbol{A}$ 而言，$\boldsymbol{B}$ 所造成的干擾並不嚴重，因此要將 $\boldsymbol{A}$ 分類出來就容易得多。雖然我們沒有關於 α 的估計結果，但是從估計錯誤率來看，我們相信 α 的估計是可信賴的 (在此實驗中，$1-\alpha$ 的估計值大約介於 0.20 和 0.21 之間)。因此我們可以事前就知道我們的分類是否會碰到這類的問題。

第四、估計的錯誤率高於真實的錯誤率。從實驗可以看出，除了 s_1 和 s_2 靠近上界的情況外，估計錯誤率遠遠高出真實錯誤率。雖然過度高估不是一件好事，但至少當估計錯誤率低時，分類結果是可信賴的。

3.4 完全分類

有些時候，我們會希望所有的點都被分類。我們只要讓 $s_1 + s_2 = 2k\overline{\sigma}$，幾乎所有的點都會被分類。延用上面的符號，也就是說當 r_1 等於 r_2 時，所有的點都會被分類。令 $r = r_1 = r_2$，我們將大於 r 的點分到 $\boldsymbol{A}$，小於 r 的點分到 $\boldsymbol{B}$。到底 r 要如何決定，分類的結果會達到最好？我們並沒有公式決定 r 的大小要多大才是最好，但是我們有一個方法決定 r，這個方法得到的 r 雖然不是最佳選擇，但是會靠近最佳選擇。現在令 $h_1 = \overline{\mu} - \ell\sigma + (1+\alpha)(a-b)$ 是 r_1 的下界，$h_2 = \overline{\mu} + \ell\sigma - \alpha(a-b)$ 是 r_2 的上界。r 的範圍應該介於 $[h_1, h_2]$。由於 $\boldsymbol{A}$ 的數目比 $\boldsymbol{B}$ 的數目為 $\alpha : (1-\alpha)$，因此我們猜測如果 $r - h_1 : h_2 - r = \alpha : 1-\alpha$ 分類的結果應該會很好。換句話說，如果 $\boldsymbol{A}$ 的數目較多，那麼就應該分較多的點到 $\boldsymbol{A}$，這樣分類的錯誤應該會比較少。但我們無法直接證明這個猜想，所以用兩個實驗當做佐証。

這兩個實驗都是 $a = 6$、$b = 3$。第一個實驗 $\boldsymbol{A}$ 與 $\boldsymbol{B}$ 點數比為 $4:1$。結果如下表：

	1	2	3	4	5	6	7	8	9
估計錯誤率 (%)	44	36	28	22	17	14	12	12	12
實際錯誤率 (%)	28	21	14	9	7	5	6	7	8

這張表格將 $[h_1, h_2]$ 切成十等分，1 表示 $r = h_2 - (h_2 - h_1)/10$ 的分類錯誤率， 2 表示 $r = h_2 - 2 \times (h_2 - h_1)/10$ 的分類錯誤率，依此類推。由於這裡不再有未分類比率，因此我們直接觀察分類的總錯誤率就夠了。剛提到的方法會令 $r = h_2 - 8 \times (h_2 - h_1)/10$，這個選擇並不是最佳的，但結果不會比最佳選擇差太多。

第二個實驗 $\boldsymbol{A}$ 與 $\boldsymbol{B}$ 點數比為 $1:1$。結果如下表：

	1	2	3	4	5	6	7	8	9
估計錯誤率 (%)	28	24	21	19	18	19	21	25	30
實際錯誤率 (%)	20	15	11	8	7	7	9	13	19

以上述方法得到的 r 會等於 $h_2 - 5 \times (h_2 - h_1)/10$，為這張表格中的最佳選擇。這兩個實驗顯示這個演算法是可行的。因此當我們有需要將所有的點都分類時，我們用這個方法決定 r。

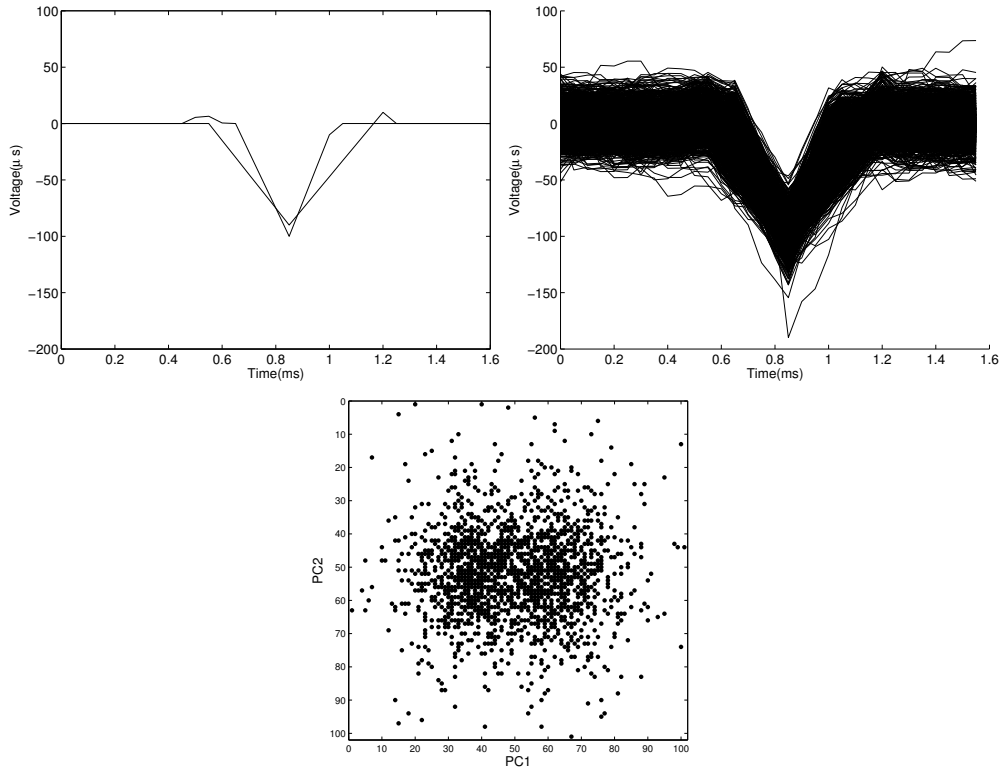
第四章、動作電位分類

4.1 基本的分類方法

現在我們將第三章的方法應用到動作電位分類。假設我們現在有兩類動作電位，$\boldsymbol{V}$ 和 $\boldsymbol{U}$。$\boldsymbol{V}$ 中的動作電位是由同一個波形 $\boldsymbol{v}$ 加上雜訊造成的，$\boldsymbol{U}$ 中的動作電位則是由另一個波形 $\boldsymbol{u}$ 加上雜訊造成的。流程請看 §1.2。根據 [3] 的說明，我們假設雜訊分佈接近正規分佈，因此我們利用第三章的方法來分類動作電位。假設我們得到的只有 $\boldsymbol{V}$ 和 $\boldsymbol{U}$ 混在一起的結果—命名為 $\boldsymbol{W}$，我們試圖以第三章的方法將 $\boldsymbol{V}$ 和 $\boldsymbol{U}$ 從 $\boldsymbol{W}$ 中分開來。第三章的方法只用在一維實數，而我們要分類的是 32 維向量，所以我們要從 32 個位置中選一個位置，對這個位置分類。由於不論是哪一個位置，雜訊標準差都不會改變，因此分類效果最好的位置應該是 $\boldsymbol{v}-\boldsymbol{u}$ 取絕對值後最大的位置。但我們不知道 $\boldsymbol{v}$ 和 $\boldsymbol{u}$，所以我們必須只能用其他的方法決定位置。有兩種方法可以選擇，第一種是利用 §3.2 的公式估計這兩個波形在每個位置上的距離，然後選最大的位置。而第二種是個別計算 $\boldsymbol{W}$ 中每個位置的標準差，然後選擇最大的位置。因為在波形差距較大的位置上，$\boldsymbol{W}$ 的標準差通常會比較大。第一種處理方法有一個問題，就是如果在某個位置上波形的距離小於雜訊標準差，則 §3.2 的方法會變得不穩定，因此估計的結果就會不準。因此我們採用第二種方法選擇分類位置。

我們有三個實驗，每個實驗會有兩個人工的動作電位波形，利用這兩個波形分別產生 1000 個動作電位，所以一共是 2000 個動作電位。實驗所使用的雜訊都是從同一個真實訊號中取出，標準差約為 11μV。

第一個實驗波形是圖六的左圖，而加上雜訊的結果則是右圖。 下圖是它們在第一主成分及第二主成分下的點圖。以下的表格是以 §3.1 的方法分類後的結果。這個實驗以六組不同的 s_1 和 s_2 分類，放在第二列的是 s_1 和 s_2 在上界時的分類結果，依序遞減，第六列是 s_1 和 s_2 在下界時的分類結果。C1 行是未分類點數，C2 行是分類到 $\boldsymbol{V}$ 的點數， C3 行是屬於 $\boldsymbol{U}$，但分到 $\boldsymbol{V}$ 的點數， C4 行是用 §3.1 公式估計的分錯點



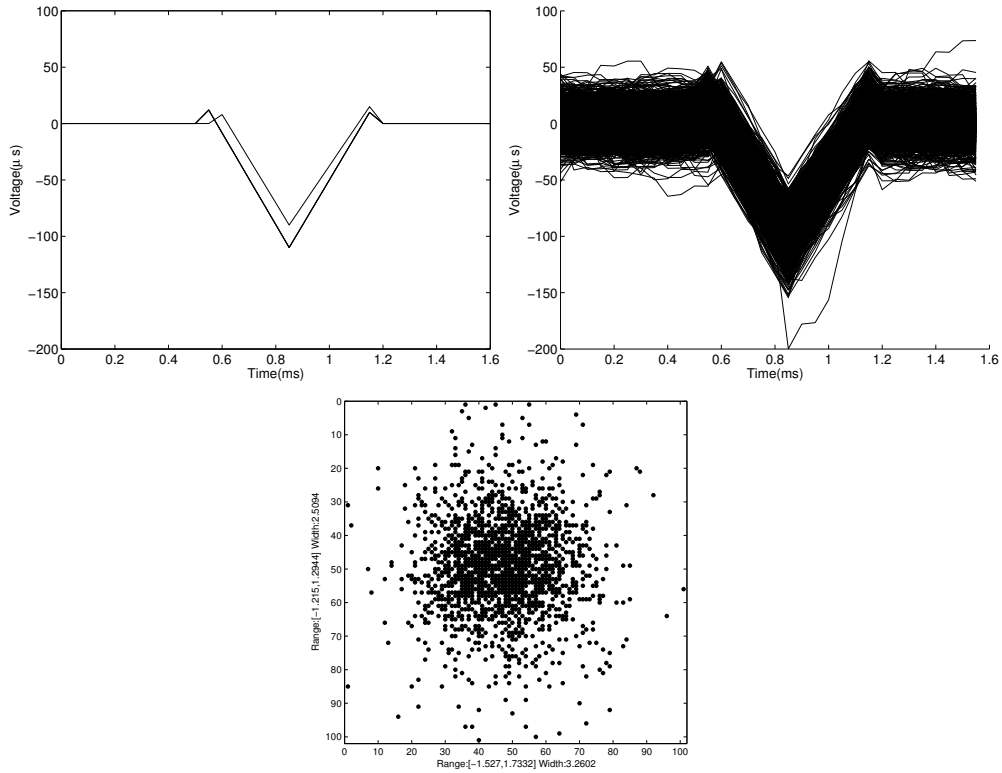
圖六

數， C5、C6、C7 的意義跟 C2、C3、C4 的意義類似，只是 $\boldsymbol{V}$ 跟 $\boldsymbol{U}$ 對調。

	C1	C2	C3	C4	C5	C6	C7
上界	692	630	4		678	7	
	491	742	7	20	767	13	22
	327	825	13	31	848	22	35
	211	884	21	56	905	30	63
	96	942	36	89	962	42	102
下界	0	991	51	133	1009	60	151

在 s_1 和 s_2 接近上界時，沒有估計的錯誤點數，這是因為在上界的估計值不可信賴，因此不列入表格內在此例子中我們也能從 PCA 的圖將這兩類動作電位分開，但我們沒有數據顯示這個結果可不可靠。如果用 §3.1 的方法，我們至少能知道最多有幾個點分錯。

第二個實驗的動作電位放在圖七上圖，而第一主成分及第二主成分的點圖則放在下



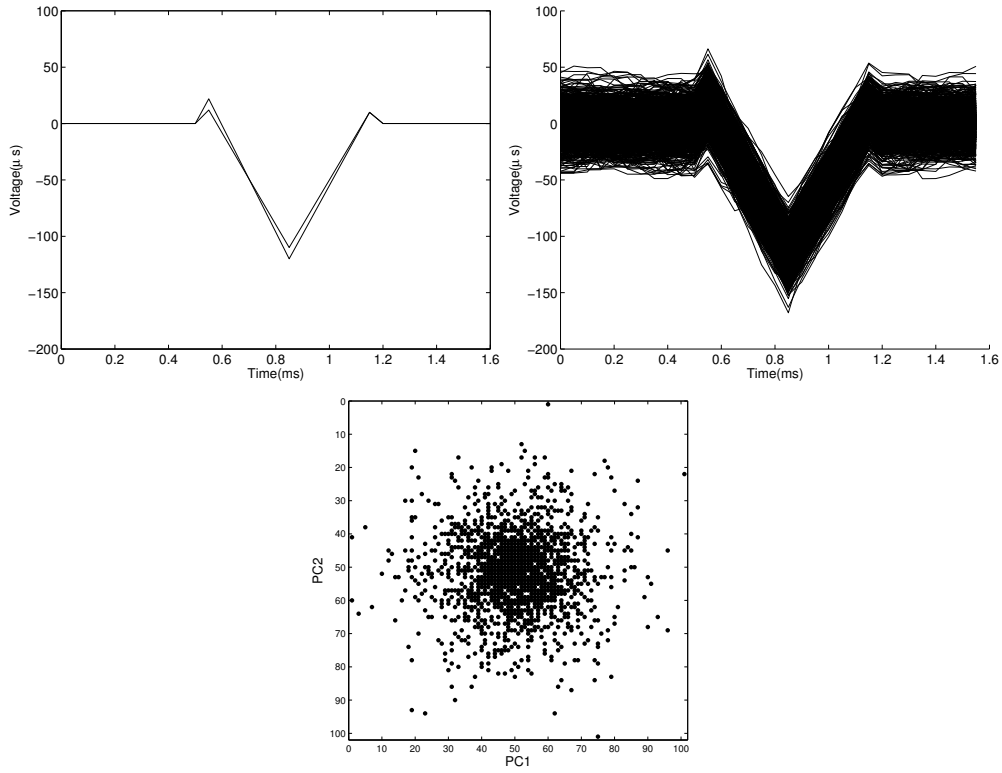
圖七

圖。從 PCA 的圖形來看，這群 spikes 並不容易分類，以§3.1 的方法分類，結果如下

	C1	C2	C3	C4	C5	C6	C7
上界	1782	78	5		190	8	
	1553	142	8	20	305	27	33
	1231	253	15	38	516	46	90
	849	415	27	74	738	86	179
	439	647	62	126	914	138	309
下界	0	868	119	194	1132	251	479

這個實驗除了 s_1 及 s_2 達下界外，未分類比率都偏高。而且分類錯誤率都大於第一個實驗的分類錯誤率。從實際數據來看，在完全分類時分類的正確率約有 80%；但考慮的真實情況，我們只能保證分類正確率會高於 65%，因為在真正的實驗中我們只能相信估計錯誤率。

圖八就是第三個實驗。從左圖來看，$\boldsymbol{v}$ 和 $\boldsymbol{u}$ 只有些微的差異 (相對雜訊而言)。觀察它們第一主成分及第二主成分的點圖，我們也不會認為它們有兩群。事實上，$\boldsymbol{v}$ 和



圖八

$\boldsymbol{u}$ 最大差距只有 0.1，比雜訊的標準差還小。以下是用 §3.1 方法分類的結果。

	C1	C2	C3	C4	C5	C6	C7
上界	1912	30	3		58	17	
	1805	67	12	20	128	32	42
	1620	136	20	47	244	57	120
	1230	303	65	97	467	103	250
	677	553	131	168	779	196	437
下界	0	907	255	261	1093	348	676

如果是以估計錯誤率來看，分類的結果非常糟，分類正確率約 50%。在這類的分類中，分類正確率最差就是 50%，因為我們沒有限定 $\boldsymbol{V}$ 和 $\boldsymbol{U}$ 的順序，所以當分類正確率低於 50% 時，我們只要將分類結果反過來，分類正確率就會高於 50%。所以這個分類是失敗的，因為就算我們隨便亂分，分類正確率最低也不過 50%。所以在碰到這種現象時，我們應該認定這裡只有一類動作電位。這個實驗說明了 §3.1 的方法不會誤將一類動作電位分成兩類。

4.2 應用

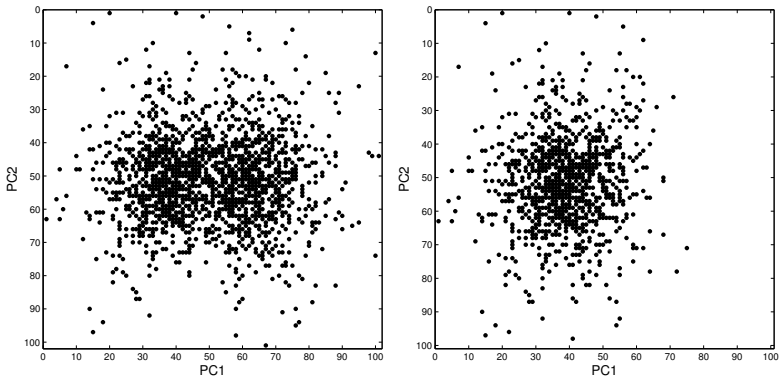
由於動作電位會用超過一維的向量來表示，因此我們可以使用多個位置分類。不過這麼做可能會有兩個缺點，第一個缺點是分類正確率會派不上用場，因為我們現在採用

多個位置的分類，所以分類正確率就變得難以估計。第二個缺點是會發生重複分類。舉例來說，假設我們選擇向量的第八個位置和第十個位置分類動作電位，我們可能找到一個動作電位在第八個位置分類到 $\boldsymbol{V}$，在第十個位置卻被分類到 $\boldsymbol{U}$。為了解決這個問題，我們決定採取投票。首先我們應該選擇奇數個位置分類動作電位，然後我們讓每個位置都是完全分類，亦即每個位置都不會有未分類的動作電位。如果有超過半數的位置將一個動作電位分到 $\boldsymbol{U}$，那我們就認定這個動作電位屬於 $\boldsymbol{U}$。由於我們要求奇數個位置以及完全分類，因此所有動作電位都會被分類。

從 §3.3 的結果來看，兩個點之間的距離超過兩倍標準差的分類是比較可信的，因此我們決定只用波形差距超過兩倍標準差的位置分類動作電位。因此我們會用到 §3.2 的式子。由於當兩個點間的距離太小時，§3.2 的估計式會失效，所以我們要再多加一個限制，就是我們只採用 $\boldsymbol{W}$ 標準差大於 1.5 倍雜訊標準差的位置。因為當某個位置的標準差超過 1.5 倍雜訊標準差，則 $\boldsymbol{u}$ 和 $\boldsymbol{v}$ 在那個位置的差距就不會太小，§3.2 估計式就能正確運作。現在以圖六的實驗為例，將 32 個位置上的標準差、波形距離和估計距離列出來。

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
標準差	0.12	0.11	0.11	0.12	0.12	0.12	0.12	0.12
波形距離	0	0	0	0	0	0	0	0
估計距離	0.4	1.2	0.6	0.2	0.19	0.14	0.06	0.08
	9	**10**	**11**	**12**	**13**	**14**	**15**	**16**
標準差	0.12	0.11	0.12	0.12	0.14	0.19	0.15	0.13
波形距離	0	0	0.06	0.07	0.16	0.3	0.2	0.1
估計距離	0.12	0.1	0.08	0.09	0.17	0.31	0.21	0.12
	17	**18**	**18**	**20**	**21**	**22**	**23**	**24**
標準差	0.12	0.13	0.12	0.16	0.22	0.20	0.15	0.12
波形距離	0	0.1	0.06	0.21	0.37	0.33	0.19	0.04
估計距離	0.19	0.22	0.15	0.23	0.38	0.34	0.20	0.11
	25	**26**	**27**	**28**	**29**	**30**	**31**	**32**
標準差	0.13	0.13	0.12	0.12	0.12	0.12	0.12	0.12
波形距離	0.1	0	0	0	0	0	0	0
估計距離	0.13	0.06	0.08	0.25	0.20	0.18	0.14	0.08

雜訊的標準差約為 11μV，從表格來看，波形距離為 0 的位置上，標準差都接近 0.11。波形距離越靠近 0，標準差就越小，估計距離誤差越大。因此我們只能相信在標準差夠大的位置上的估計距離。我們基本上只使用估計距離大於兩倍雜訊標準差的位置分類。如果說這些位置有偶數個，那麼我們會再多取一個位置，這個位置的估計距離比其他沒取到的位置的估計距離大。我們一定會用奇數個位置分類動作電位。如



圖九

果說沒有一個位置的標準差和估計距離符合上述兩個條件，那麼我們就選擇標準差最大的位置分類動作電位。在這種情況下，標準差會比估計距離可靠，所以我們採用標準差而不採用估計距離。

要用多個位置分類動作電位，有一個該解決的問題是該如何將不同位置的分類結果結合。舉例來說，假設我們採用三個位置分類動作電位，這三個位置會得到三種分類結果。第一個位置的結果為 $\boldsymbol{A_1}$ 和 $\boldsymbol{B_1}$，第二個位置的結果為 $\boldsymbol{A_2}$ 和 $\boldsymbol{B_2}$，第三個位置的結果為 $\boldsymbol{A_3}$ 和 $\boldsymbol{B_3}$。我們知道 $\boldsymbol{A_1}$ 和 $\boldsymbol{B_1}$ 是兩類不同的動作電位，但我們不知道 $\boldsymbol{A_1}$ 和 $\boldsymbol{A_2}$ 會不會是同一類動作電位。這裡我們再次應用 PCA 方法來判斷，圖九左圖是 $\boldsymbol{A_1}$ 和 $\boldsymbol{B_2}$ 在第一及第二主成分上的點圖，從圖來看，$\boldsymbol{A_1}$ 和 $\boldsymbol{B_2}$ 是不同的兩類動作電位。因此我們判斷出 $\boldsymbol{A_1}$ 和 $\boldsymbol{A_2}$ 是同一類動作電位。為了保險起見，我們還是畫出 $\boldsymbol{A_1}$ 和 $\boldsymbol{A_2}$ 的點圖，也就是圖九右圖。從這兩張圖，我們可以斷定 $\boldsymbol{A_1}$ 和 $\boldsymbol{A_2}$ 是同一類動作電位，因此 $\boldsymbol{B_1}$ 和 $\boldsymbol{B_2}$ 就是另一類動作電位。而 $\boldsymbol{A_3}$ 和 $\boldsymbol{B_3}$ 也由同樣的流程與 $\boldsymbol{A_1}$ 和 $\boldsymbol{B_1}$ 結合。

結合分類結果後，我們用投票機制分類動作電位。圖九的動作電位來自 §4.1 的實驗一，這三個位置的分類結果是這樣的：$\boldsymbol{A_1}$、$\boldsymbol{A_2}$ 和 $\boldsymbol{A_3}$ 是同一類動作電位，$\boldsymbol{B_1}$、$\boldsymbol{B_2}$ 和 $\boldsymbol{B_3}$ 是另一類動作電位。我們現在令 $\boldsymbol{A}$ 是第一類動作電位，$\boldsymbol{B}$ 是第二類動作電位。如果一個動作電位在 $\boldsymbol{A_1}$、$\boldsymbol{A_2}$ 和 $\boldsymbol{A_3}$ 出現兩次—例如這個動作電位屬於 $\boldsymbol{A_1}$ 和 $\boldsymbol{A_2}$，但不屬於 $\boldsymbol{A_3}$—就將這個動作電位分到 $\boldsymbol{A}$，反之則分到 $\boldsymbol{B}$。所有的動作電位都以這個方式分類，最後 $\boldsymbol{W}$ 會被分成 $\boldsymbol{A}$ 和 $\boldsymbol{B}$，也就是我們想得到的 $\boldsymbol{U}$ 和 $\boldsymbol{V}$，分類結束。

最後的分類結果，有 995 個動作電位分到 $\boldsymbol{V}$，51 個分錯；1005 個動作電位分到 $\boldsymbol{U}$，56 個分錯。結果比 §4.1 方法好一點，但沒有估計錯誤率。實驗二沒有一個位置通

過篩選條件，因此只能使用 §4.1 的方法分類。

4.3 結論及發展

本論文提供的方法有兩個用途，一個是分類兩群的動作電位。另一個用途則是檢查一群動作電位是否該再分類。第一個用途是此方法最初的目的，如果我們採用 §4.1 的方法分類，則我們可以得知分類結果準不準確。而採用 §4.2 方法會失去這個好處，但結果會比 §4.1 好上一點。但不是所有的訊號都適合使用 §4.2 的方法，因為不是每個位置的分類都是可信賴的。因此 §4.2 要求我們只採用可信賴的位置來分類動作電位，如果說訊號中不存在這樣的的位置，我們就只能採用 §4.1 的方法。

而另一個用途的實例就是 §4.1 的實驗三，這個實驗的動作電位雖然有兩類，但在雜訊的干擾下，這兩類動作電位已經無法分開，所以我們的分類結果一踏糊塗。這顯示如果一群動作電位分類後的估計分類錯誤率高達 50% 時，我們最好認為這群動作電位屬於同一類動作電位。

本論文的方法有一個很大的缺點，就是只能對兩群動作電位分類。此方法的目標是解決 PCA 方法所無法解決的問題—當有幾類動作電位在 PCA 的點圖上的分佈範圍有重疊時，PCA 不能幫助我們客觀地分類動作電位。所以我們的預處理的最後一步是用 PCA 挑出兩類無法分開的動作電位。在一般的實驗中，動作電位常常有好幾類混在一起，只能分開兩類動作電位只能解決一小部分的問題，離實用還有一段距離。

要解決三類以上的問題，我們的第一個想法就是仿照 §3.1 的流程，先處理三個點的分類問題。但是這個想法最後失敗了，原因是三個點的推導公式過於複雜，我們無法順利的推出如第三章的簡單結果。而第二個想法則是希望將三類動作電位分成兩邊，一邊一類，另一邊有兩類。例如我們用 §4.1 的方法從某個位置將動作電位分成兩類，再從另一個位置將這兩類動作電位再分類。但我們也尚未找到一個方法告訴我們「某個位置」該如何決定。

參考書目

[1] K. V. Mardia, J. T. Kent and J. M. Bibby, "Multivariate Analysis," Academic Press, 1979.

[2] Michal S. Lewicki, "A review of methods for spike sorting:the detection and classification of neural action potentials," Network:Comput. Neural Syst. 9 (1998) R53-R78.

[3] Andre Diedrich, Warakorn Charoensuk, Robert J. Brycyta, Andrew C. Ertl, and Richard Shiavi, "Analysis of Raw Microneurographic Recordings Based on Wavelet De-Noising Technique and Classification Algorithm:Wavelet Analysis in Microneurography" IEEE Transactions on Biomedical Engineering, vol. 50, no. 1, January 2003.

[4] Kenneth D. Harris, Darrell A. Henze, Jozsef Csicsvari, Hajime Hirase, and György Buzsáki, "Accuracy of Tetrode Spike Separation as Determined by Simultaneous Intracellular and Extracellular Measurements," Journal of Neurophysiology 84:401-414, 2000.

[5] 吳慶餘, 林金盾, "基礎生命科學," 藝軒圖書, 2003, pp. 386–390.

[6] 黃世傑, 王瑋龍, 陳森香, "生物學," 偉華, 2003, pp. 386–390.

[7] George B. Johnson, "生物學," 學銘, 2004, p. 460.

[8] Sylvia S. Mader, "生物學：探索生命," 偉明, 2004, pp. 368–370.