# 第四章 雙維度模式建構與推論

## 4.1 雙維度分析之基本假設

在不失一般性的原則下，我們假設具有血緣關係的對象是姊妹，令 $\tilde{D}_j$ $(j=1,2)$ 代表個別是否帶有致病基因的指標函數。依據遺傳的定律，$(\tilde{D}_1, \tilde{D}_2)$ 具有相關性，我們可以用勝負比(odds ratio)來描述相關性強弱:

$$\gamma = \frac{\Pr(\tilde{D}_1=1,\tilde{D}_2=1)\Pr(\tilde{D}_1=0,\tilde{D}_2=0)}{\Pr(\tilde{D}_1=0,\tilde{D}_2=1)\Pr(\tilde{D}_1=1,\tilde{D}_2=0)} \text{ 。} \tag{4.1}$$

令 $X_j \mid \tilde{D}_j = 1$ 代表帶因者之潛在發病時間，$Y_j$ 代表來自死亡的競爭風險，以 $j=1,2$ 分別代表姊妹。對於帶因者$(\tilde{D}_j = 1)$，與單維度的情形一樣令 $X_j^* = X_j$ ；對於免疫者$(\tilde{D}_j = 0)$，則令 $X_j^* = \infty$ 。此時可定義 $D_j = I(X_j^* \le Y_j)$ $(j=1,2)$ ，代表在生前是否觀測到發病的指標函數。當觀察到發病$(D_j = 1)$，可知 $\tilde{D}_j = 1$ 且 $X_j \le Y_j$ ；然而若死亡前未觀察到發病$(D_j = 0)$，則有可能有兩種情形:有可能是免疫$(\tilde{D}_j = 0)$或是帶因$(\tilde{D}_j = 1)$但是 $X_j > Y_j$ 。

針對雙維度模式，我們所做以下假設：

**假設 一:** $X_j^* \perp Y_j$ $(j=1,2)$ (4.2)

**假設 二:** $Y_j \mid \tilde{D}_j = 1 \ =^d Y_j \mid \tilde{D}_j = 0$ $(j=1,2)$ (4.3)

**假設 三:** $\Pr(X_i > t \mid \tilde{D}_i = 1, \tilde{D}_j = 0) = \Pr(X_i > t \mid \tilde{D}_i = 1, \tilde{D}_j = 1) = \Pr(X_i > t \mid \tilde{D}_i = 1)$

$(i \ne j)$ 。 (4.4)

假設（一）（二）的理由和先前單維度的理由相同。第三個假設指一旦知道其中一個帶因$(\tilde{D}_i = 1)$，其發病時間$(X_i)$的分配和其姊妹是否帶因的資訊無關(由 $\tilde{D}_j$ 的值決定)。這個假設一方面是為了估計的簡化，另一方面也應屬合理，因為一旦有

了直接的證據，間接的證據便失去參考價值。在混合模式架構下，可以得到

$$\Pr(X_j^* > t) = \Pr(X_j^* > t \mid \tilde{D}_j = 1)\Pr(\tilde{D}_j = 1) + \Pr(X_j^* > t \mid \tilde{D}_j = 0)\Pr(\tilde{D}_j = 0)$$

$$= \Pr(X_j^* > t \mid \tilde{D}_j = 1)p_j + (1 - p_j)\ ,\qquad (4.5)$$

其中 $\Pr(\tilde{D}_j = 1) = p_j$ $(j = 1,2)$。值得一提的是我們可以模仿 Chatterjee & Shih 的作法假設 $p_1 = p_2 = p$，以減低未知參數的維度。但對我們的方法而言，這個簡化的假設並非必要的。

## 4.2 雙維度混合模式

### 4.2.1 雙維度資料

如前述之定義，我們令 $D_j$ 代表個體實際發病與否的指標函數，其對應的時間為 $T_j = X_j^* \wedge Y_j$。可將資料表示為 $\{(T_{i1}, T_{i2}, D_{i1}, D_{i2})\ (i = 1,2,...,n)\}$，其中 $T_{i1} = X_{i1}^* \wedge Y_{i1}$，$T_{i2} = X_{i2}^* \wedge Y_{i2}$，$D_{i1} = I(X_{i1}^* \le Y_{i1})$，$D_{i2} = I(X_{i2}^* \le Y_{i2})$。外生設限(censoring)的情形將在後面章節考慮，在此仍假設每個個體都可以追蹤到發病或是因其他疾病而死亡。

### 4.2.2 雙維度模式分配的假設

當免疫者不存在的情形下，以無母數的方法估計雙變量的存活函數 (nonparametric estimation of the bivariate survival function) 已是相當複雜與困難的問題。因此當存在免疫者時，無母數的方向更不可行，此時我們需要做一些分佈的假設以做為推論的依據。

(I). 發病時間之關聯模式：若姊妹皆可能致病 $(\tilde{D}_1 = 1, \tilde{D}_2 = 1)$，我們假設其發病的時間的聯合分配具有 copula model，可表示為

$$\Pr(X_1 > s, X_2 > t \mid \tilde{D}_1 = 1, \tilde{D}_2 = 1) = C_\alpha\{S_1(s), S_2(t)\}\ ,\qquad (4.6)$$

其中 $S_j(t) = \Pr(X_j > t \mid \tilde{D}_j = 1)$。參數 $\alpha$ 衡量發病時間關聯性的強弱，是我們最感興趣的參數之一，也是推論的主要目標。

(II). 競爭風險之關聯模式：假設 $(Y_1, Y_2)$ 亦具有 copula model，可表示為

$$\Pr(Y_1 > s, Y_2 > t) = D_\beta\{H_1(s), H_2(t)\}, \tag{4.7}$$

其中 $H_j(t) = \Pr(Y_j > t)$，參數 $\beta$ 衡量來自死亡的競爭風險關聯性的強弱。

### 4.2.3 符號定義

在討論估計方法之前，我們先定義符號，以簡化推導過程之表達。

(i). $C_\alpha^{00}(u,v) = C_\alpha(u,v)$，

(ii). $C_\alpha^{10}(u,v) = \dfrac{\partial C_\alpha^{00}(u,v)}{\partial u}$，

(iii). $C_\alpha^{01}(u,v) = \dfrac{\partial C_\alpha^{00}(u,v)}{\partial v}$，

(iv). $C_\alpha^{11}(u,v) = \dfrac{\partial C_\alpha^{00}(u,v)}{\partial u \partial v}$。

我們可以依類似的原則定義 $D_\beta^{kl}(s,t)$ $(k,l) = (0,0),(1,0),(0,1),(1,1)$。

## 4.3 雙維度混合模式的推論方法

### 4.3.1 概似函數

資料可表示為 $\{(T_{i1}, T_{i2}, D_{i1}, D_{i2})\ (i = 1,2,...,n)\}$。依據模式的架構，感興趣的參數有：

(i). 描述姊妹帶因與否的關聯性參數($\gamma = \dfrac{P_{11}P_{00}}{P_{01}P_{10}}$)；

(ii). 姊妹發病時間的關聯性參數（$\alpha$）；

(iii). 來自死亡競爭風險的關聯性參數（$\beta$）。

依據可觀測到的指標函數( $D_{i1}$, $D_{i2}$) 的值，我們可以將其對應到以下的幾種情形:

a. $(D_{i1}=0, D_{i2}=1)$:有兩種可能的情形

a1 $(\tilde{D}_{i1},\tilde{D}_{i2})=(1,1)$: 此時 $T_{i1}=X_{i1}<Y_{i1}$，$T_{i2}=X_{i2}<Y_{i2}$，

a2 $(\tilde{D}_{i1},\tilde{D}_{i2})=(0,1)$: 此時$T_{i1}=Y_{i1}<X_{i1}$，$T_{i2}=X_{i2}<Y_{i2}$；

b. $(D_{i1}=1, D_{i2}=0)$：有兩種可能的情形

b1 $(\tilde{D}_{i1},\tilde{D}_{i2})=(1,1)$:此時 $T_{i1}=X_{i1}<Y_{i1}$，$T_{i2}=X_{i2}<Y_{i2}$，

b2 $(\tilde{D}_{i1},\tilde{D}_{i2})=(1,0)$: 此時$T_{i1}=X_{i1}<Y_{i1}$， $T_{i2}=Y_{i2}<X_{i2}$；

c. $(D_{i1}=0, D_{i2}=0)$：有四種可能的情形

c1 $(\tilde{D}_{i1},\tilde{D}_{i2})=(1,1)$: 此時 $T_{i1}=X_{i1}<Y_{i1}$，$T_{i2}=X_{i2}<Y_{i2}$，

c2 $(\tilde{D}_{i1},\tilde{D}_{i2})=(1,0)$: 此時$T_{i1}=X_{i1}<Y_{i1}$， $T_{i2}=Y_{i2}<X_{i2}$，

c3 $(\tilde{D}_{i1},\tilde{D}_{i2})=(0,1)$: 此時$T_{i1}=Y_{i1}<X_{i1}$，$T_{i2}=X_{i2}<Y_{i2}$，

c4 $(\tilde{D}_{i1},\tilde{D}_{i2})=(0,0)$: 此時$T_{i1}=Y_{i1}<X_{i1}$，$T_{i2}=Y_{i2}<X_{i2}$；

d. $(D_{i1}=1, D_{i2}=1)$：只有一種情形

d1 $(\tilde{D}_{i1},\tilde{D}_{i2})=(1,1)$: 此時 $T_{i1}=X_{i1}<Y_{i1}$，$T_{i2}=X_{i2}<Y_{i2}$。

當邊際存活函數$S_j(t)=\Pr(X_j>t\mid\tilde{D}_j=1)$已知的情形下，$(\alpha,\beta,P_{11},P_{10},P_{01},P_{00})$的概似函數可表示為:

$$L=\prod_{i=1}^{n}(L_{1i})^{I(D_{i1}=1,D_{i2}=1)}(L_{2i}+L_{3i})^{I(D_{i1}=1,D_{i2}=0)}(L_{4i}+L_{5i})^{I(D_{i1}=0,D_{i2}=1)}(L_{6i}+L_{7i}+L_{8i}+L_{9i})^{I(D_{i1}=0,D_{i2}=0)} \tag{4.8}$$

其中 $P_{ij}=\Pr(\tilde{D}_1=i,\tilde{D}_2=j)\ i,j=(0,1)$，

$$L_{1i}=\Pr(X_1=t_{i1},X_2=t_{i2}\mid\tilde{D}_{i1}=1,\tilde{D}_{i2}=1)\Pr(Y_1>t_{i1},Y_2>t_{i2})\Pr(\tilde{D}_{i1}=1,\tilde{D}_{i2}=1)$$

$$=C_{\alpha}^{11}(S_1(t_{i1}),S_2(t_{i2}))S_1(\Delta t_{i1})S_2(\Delta t_{i2})D_{\beta}^{00}(H_1(t_{i1}),H_2(t_{i2}))P_{11}\ ,$$

$$L_{2i} = \Pr(X_1 = t_{i1}, X_2 > t_{i2} \mid \tilde{D}_{i1} = 1, \tilde{D}_{i2} = 1) \Pr(Y_1 > t_{i1}, Y_2 = t_{i2}) \Pr(\tilde{D}_{i1} = 1, \tilde{D}_{i2} = 1)$$

$$= C_\alpha^{10}(S_1(t_{i1}), S_2(t_{i2})) S_1(\Delta t_{i1}) D_\beta^{01}(H_1(t_{i1}), H_2(t_{i2})) H_2(\Delta t_{i2}) P_{11} \text{ ，}$$

$$L_{3i} = \Pr(X_1 = t_{i1} \mid \tilde{D}_{i1} = 1, \tilde{D}_{i2} = 0) \times \Pr(Y_1 > t_{i1}, Y_2 = t_{i2}) \times \Pr(\tilde{D}_{i1} = 1, \tilde{D}_{i2} = 0)$$

$$= S_1(\Delta t_{i1}) D_\beta^{01}(H_1(t_{i1}), H_2(t_{i2})) H_2(\Delta t_{i2}) P_{10} \text{ ，}$$

$$L_{4i} = \Pr(X_1 > t_{i1}, X_2 = t_{i2} \mid \tilde{D}_{i1} = 1, \tilde{D}_{i2} = 1) \Pr(Y_1 = t_{i1}, Y_2 > t_{i2}) \Pr(\tilde{D}_{i1} = 1, \tilde{D}_{i2} = 1)$$

$$= C_\alpha^{10}(S_1(t_{i1}), S_2(t_{i2})) S_2(\Delta t_{i2}) D_\beta^{10}(H_1(t_{i1}), H_2(t_{i2})) H_1(\Delta t_{i1}) P_{11} \text{ ，}$$

$$L_{5i} = \Pr(X_2 = t_{i2} \mid \tilde{D}_{i1} = 0, \tilde{D}_{i2} = 1) \Pr(Y_1 = t_{i1}, Y_2 > t_{i2}) \times \Pr(\tilde{D}_{i1} = 0, \tilde{D}_{i2} = 1)$$

$$= S_2(\Delta t_{i2}) D_\beta^{10}(H_1(t_{i1}), H_2(t_{i2})) H_1(\Delta t_{i1}) P_{01} \text{ ，}$$

$$L_{6i} = \Pr(X_1 > t_{i1}, X_2 > t_{i2} \mid \tilde{D}_{i1} = 1, \tilde{D}_{i2} = 1) \times \Pr(Y_1 = t_{i1}, Y_2 = t_{i2}) \Pr(\tilde{D}_{i1} = 1, \tilde{D}_{i2} = 1)$$

$$= C_\alpha^{00}(S_1(t_{i1}), S_2(t_{i2})) D_\beta^{11}(H_1(t_{i1}), H_2(t_{i2})) H_1(\Delta t_{i1}) H_2(\Delta t_{i2}) P_{11} \text{ ，}$$

$$L_{7i} = \Pr(X_1 > t_{i1} \mid \tilde{D}_{i1} = 1, \tilde{D}_{i2} = 0) \Pr(Y_1 = t_{i1}, Y_2 = t_{i2}) \times \Pr(\tilde{D}_{i1} = 1, \tilde{D}_{i2} = 0)$$

$$= S_1(t_{i1}) D_\beta^{11}(H_1(t_{i1}), H_2(t_{i2})) H_1(\Delta t_{i1}) H_2(\Delta t_{i2}) P_{10} \text{ ，}$$

$$L_{8i} = \Pr(X_2 > t_{i2} \mid \tilde{D}_{i1} = 0, \tilde{D}_{i2} = 1) \Pr(Y_1 = t_{i1}, Y_2 = t_{i2}) \Pr(\tilde{D}_{i1} = 0, \tilde{D}_{i2} = 1)$$

$$= S_2(t_{i2}) D_\beta^{11}(H_1(t_{i1}), H_2(t_{i2})) H_1(\Delta t_{i1}) H_2(\Delta t_{i2}) P_{01} \text{ ，}$$

$$L_{9i} = \Pr(Y_1 = t_{i1}, Y_2 = t_{i2}) \Pr(\tilde{D}_{i1} = 0, \tilde{D}_{i2} = 0)$$

$$= D_\beta^{11}(H_1(t_{i1}), H_2(t_{i2})) H_1(\Delta t_{i1}) H_2(\Delta t_{i2}) P_{00} \text{ 。}$$

以下簡述以概似函數求估計量的難度：

（i） 概似函數（L）包含 $S_j(t) = \Pr(X_j > t \mid \tilde{D}_{ij} = 1)$ ，這部分是未知的，需要額外估計，這個問題已在第三章討論過。在模式中我們亦將比較無母數和假設母數分佈的最大概似法估計 $S_j(t_i)$ 。一般來說無母數的估計量較具

穩健性，但是有較大的變異性，在存在免疫者時，無母數方法依賴充分追蹤的假設。

（ii） 因為無法直接觀察到$(\tilde{D}_1,\tilde{D}_2)$，但是$(D_1,D_2)$提供的資訊包含數種可能性，使得概似函數（L）中有相加的機率項。在求極值的過程將對數概似函數(log-likelihood function)取微分時，score function 因此變得十分複雜，增加了估計的難度。即使以數值分析的方法求解，可是在計算二階微分以求 information 時，複雜度更為增加。為了簡化計算，我們將採用 EM 演算法的方法求解。

（iii） 在(4.8)的概似函數中，包含$S_1(\Delta t_{i1}), S_2(\Delta t_{i2}), H_1(\Delta t_{i1}), H_2(\Delta t_{i2})$這些衡量 point mass 的項目。對於一般 copula model 的推論問題，這些項目會在對參數微分時自動消失。然而我們的問題容許存在免疫的可能性，概似函數中連加的項目亦包含這些 mass 計算，無法以常數視之，使得我們必須處理多餘的估計問題。

4.3.2 $\beta$的概似函數與無母數推論方法

參數$\beta$描述了姊妹競爭風險發生時間的關聯性，如前所述$(Y_1,Y_2)$受到$(X_1^*,X_2^*)$的設限，這是典型分析二維 copula 模式的推論問題(不像估計$\alpha$牽涉到免疫與否的問題)，估計方法較為單純因此可以先行處理。前面將競爭風險$(Y_1,Y_2)$之關聯模式表示為:

$$\Pr(Y_1 > s, Y_2 > t) = D_\beta\{H_1(s), H_2(t)\}，$$

其中$H_j(t) = \Pr(Y_j > t)$ $(j = 1,2)$。我們以無母數方法估計$H_j(t) = \Pr(Y_j > t)$，

Kaplan-Meier 估計量可表示為:

$$\hat{H}_j(t) = \prod_{u \le t}\left\{1 - \frac{\sum_{i=1}^{n} I(T_{ij} = u, D_{ij} = 0)}{\sum_{i=1}^{n} I(T_{ij} \ge u)}\right\} \quad (j = 1,2)。$$

將 $\hat{H}_j(t)$ 代入原有的概似函數中，得到的擬概似函數(pseudo-likelihood function)可表示如下，

$$\tilde{L}_\beta \propto \prod_{i=1}^{n} \Big\{ \Big[D_\beta^{00}(\hat{H}_1(t_{i1}),\hat{H}_2(t_{i2}))\Big]^{I(D_{i1}=1,D_{i2}=1)} \Big[D_\beta^{01}(\hat{H}_1(t_{i1}),\hat{H}_2(t_{i2}))\hat{H}_2(\Delta t_{i2})\Big]^{I(D_{i1}=1,D_{i2}=0)} \times$$

$$\Big[D_\beta^{10}(\hat{H}_1(t_{i1}),\hat{H}_2(t_{i2}))\hat{H}_1(\Delta t_{i1})\Big]^{I(D_{i1}=0,D_{i2}=1)} \times$$

$$\Big[D_\beta^{11}(\hat{H}_1(t_{i1}),\hat{H}_2(t_{i2}))\hat{H}_1(\Delta t_{i1})\hat{H}_2(\Delta t_{i2})\Big]^{I(D_{i1}=0,D_{i2}=0)} \Big\} \qquad (4.9)$$

其中 $D_\beta^{kl}(s,t)$ 的定義如同(4.2.4)節的符號( $(k,l)=(0,0),(1,0),(0,1),(1,1)$ )。

對 $\tilde{L}_\beta$ 取 log 函數，可得

$$\log(\tilde{L}_\beta) \propto \sum_{i=1}^{n} \Big\{ I(D_{i1}=1,D_{i2}=1)\log\Big[D_\beta^{00}(\hat{H}_1(t_{i1}),\hat{H}_2(t_{i2}))\Big] +$$

$$I(D_{i1}=1,D_{i2}=0)\log\Big[D_\beta^{01}(\hat{H}_1(t_{i1}),\hat{H}_2(t_{i2}))\Big] +$$

$$I(D_{i1}=0,D_{i2}=1)\log\Big[D_\beta^{10}(\hat{H}_1(t_{i1}),\hat{H}_2(t_{i2}))\Big] +$$

$$I(D_{i1}=0,D_{i2}=0)\log\Big[D_\beta^{11}(\hat{H}_1(t_{i1}),\hat{H}_2(t_{i2}))\Big] \Big\}。 \qquad (4.10)$$

對上式 $E\big[\log(\tilde{L}_\beta)\,|\,Data\big]=\tilde{l}(\beta)$ 求極大值的解即為 $\beta$ 估計量。

我們利用 Newton-Raphson 方法求 $\tilde{l}(\beta)$ 的極大值發生的 $\beta$，做為 $\beta$ 的估計值。Newton-Raphson 方法的公式如下：

$$\hat{\beta}^k = \hat{\beta}^{k-1} - \frac{\dfrac{\partial \tilde{l}(\beta)}{\partial \beta}\Big|_{\beta=\hat{\beta}^{k-1}}}{\dfrac{\partial^2 \tilde{l}(\beta)}{\partial \beta^2}\Big|_{\beta=\hat{\beta}^{k-1}}} \qquad (4.11)$$

其中 $\hat{\beta}^k$ 代表第 k 次疊代之估計量，

$$\frac{\partial \tilde{l}(\beta)}{\partial \beta} = \sum_{i=1}^{n} \Big\{ I(D_{i1}=1,D_{i2}=1)D_\beta^{100}(\hat{H}_1(t_{i1}),\hat{H}_2(t_{i2})) +$$

$$I(D_{i1}=1,D_{i2}=0)D_\beta^{101}(\hat{H}_1(t_{i1}),\hat{H}_2(t_{i2})) +$$

$$I(D_{i1}=0,D_{i2}=1)D_\beta^{110}(\hat{H}_1(t_{i1}),\hat{H}_2(t_{i2})) +$$

$$I(D_{i1}=0,D_{i2}=0)D_{\beta}^{111}(\hat{H}_1(t_{i1}),\hat{H}_2(t_{i2}))\Big\}\ ; \tag{4.12}$$

$$\frac{\partial^2\tilde{l}(\beta)}{\partial\beta^2}=\sum_{i=1}^{n}\Big\{\ I(D_{i1}=1,D_{i2}=1)D_{\beta}^{200}(\hat{H}_1(t_{i1}),\hat{H}_2(t_{i2}))+$$

$$I(D_{i1}=1,D_{i2}=0)D_{\beta}^{201}(\hat{H}_1(t_{i1}),\hat{H}_2(t_{i2}))+$$

$$I(D_{i1}=0,D_{i2}=1)D_{\beta}^{210}(\hat{H}_1(t_{i1}),\hat{H}_2(t_{i2}))+$$

$$I(D_{i1}=0,D_{i2}=0)D_{\beta}^{211}(\hat{H}_1(t_{i1}),\hat{H}_2(t_{i2}))\ \Big\}\ , \tag{4.13}$$

其中 $D_{\beta}^{1kl}(s,t)=\dfrac{\partial\log\left[D_{\beta}^{kl}(s,t)\right]}{\partial\beta}$ ， $D_{\beta}^{2kl}(s,t)=\dfrac{\partial^2\log\left[D_{\beta}^{kl}(s,t)\right]}{\partial\beta^2}$ $(k,l)=(0,0),(1,0)\ \ ,(0,1),(1,1)$ 。

當$\beta$的估計值求出後，我們可將$\hat{\beta}$代入(4.8)的概似函數中。在 4.3.1 節與 4.3.2 節我們求得了邊際函數與$\beta$的估計量，可視為第一階段的估計。在下一節中，我們討論利用先前獲得之估計量，以“代入”(plug-in)的原則，用以估計最感興趣的參數$\alpha$與$r$。

4.3.3 $\alpha, P_{ij}\ (i,j=0,1)$ 的概似函數與估計

再複習二階段的推論架構**:**

**二階段估計法:**

階段一: 先估計邊際分配 $S_j(.)\ (j=1,2)$ 與 $\beta$

第三章已分別討論以有母數和無母數方法估計 $S_j(.)\ (j=1,2)$。在 4.3.2 節探討參數 $\beta$ 的估計,因為未牽涉到免疫問題且是傳統右設限的資料,可以用針對一般 copula 模式擬概似函數估計法。將第一階段所得估計量 $\hat{S}_j(.)$ 和 $\hat{\beta}$ 代入概似函數中，此時概似函數只剩下未知參數 $(\alpha, P_{00}, P_{10}, P_{01}, P_{11})$。

階段二: 估計 $(\alpha, P_{00}, P_{10}, P_{01}, P_{11})$

原有參數 $P_{ij}\ (i,j)=(1,0)$ 中，自由度為 3，在第一階段中估了邊際機率，等於只剩下一個維度的未知參數。先前提及的概似函數（4.8）是根據不完整資料 $\{(T_{i1}, T_{i2}, D_{i1}, D_{i2})\ \ (i=1,2,...,n)\}$ 而建構的，求取極大值時，若未知參數的危度大會造成分析的困難。我們利用 E-M 演算法的想法，先建構完整資料 $\{(T_{i1}, T_{i2}, \tilde{D}_{i1}, \tilde{D}_{i2}, D_{i1}, D_{i2})\ \ (i=1,2,...,n)\}$ 的概似函數，其中會牽涉部分未知的 $\tilde{D}_{ij}$，我們對此概似函數根據已知資料先求條件期望值(E-step),再對期望概似函數取極大值(M-step)。

**E-M 演算法:** 根據完整資料建立的概似函數可表示為

$$\tilde{L}(\alpha, P_{11}, P_{10}, P_{01}, P_{00})$$

$$= \prod_{i=1}^{n} \{ L_{1i}^{I(D_{i1}=1, D_{i2}=1)} L_{2i}^{I(D_{i1}=1, D_{i2}=0, \tilde{D}_{i1}=1, \tilde{D}_{i2}=1)} L_{3i}^{I(D_{i1}=1, D_{i2}=0, \tilde{D}_{i1}=1, \tilde{D}_{i2}=0)}$$

$$L_{4i}^{I(D_{i1}=0, D_{i2}=1, \tilde{D}_{i1}=1, \tilde{D}_{i2}=1)} L_{5i}^{I(D_{i1}=0, D_{i2}=1, \tilde{D}_{i1}=0, \tilde{D}_{i2}=1)}$$

$$L_{6i}^{I(D_{i1}=0, D_{i2}=0, \tilde{D}_{i1}=1, \tilde{D}_{i2}=1)} L_{7i}^{I(D_{i1}=0, D_{i2}=0, \tilde{D}_{i1}=1, \tilde{D}_{i2}=0)}$$

$$L_{8i}^{I(D_{i1}=0,D_{i2}=0,\tilde{D}_{i1}=0,\tilde{D}_{i2}=1)} L_{9i}^{I(D_{i1}=0,D_{i2}=0,\tilde{D}_{i1}=0,\tilde{D}_{i2}=0)} \Big\},$$

(4.14)

其中，$\tilde{L}(\alpha, P_{11}, P_{10}, P_{01}, P_{00})$ 中各項 $L_{ki}$ $(k = 1,2,...,9)$ 的定義如同（4.8）概似函數 L 的定義。對 $\tilde{L}(\alpha, P_{11}, P_{10}, P_{01}, P_{00})$ 取 log 函數，可得：

$$\tilde{l}(\alpha, P_{11}, P_{10}, P_{01}, P_{00})$$

$$= \sum_{i=1}^{n} \{ I(D_{i1}=1, D_{i2}=1)\log L_{1i} + I(D_{i1}=1, D_{i2}=0, \tilde{D}_{i1}=1, \tilde{D}_{i2}=1)\log L_{2i} +$$

$$I(D_{i1}=1, D_{i2}=0, \tilde{D}_{i1}=1, \tilde{D}_{i2}=0)\log L_{3i} + I(D_{i1}=0, D_{i2}=1, \tilde{D}_{i1}=1, \tilde{D}_{i2}=1)\log L_{4i} +$$

$$I(D_{i1}=0, D_{i2}=1, \tilde{D}_{i1}=0, \tilde{D}_{i2}=1)\log L_{5i} + I(D_{i1}=0, D_{i2}=0, \tilde{D}_{i1}=1, \tilde{D}_{i2}=1)\log L_{6i} +$$

$$I(D_{i1}=0, D_{i2}=0, \tilde{D}_{i1}=1, \tilde{D}_{i2}=0)\log L_{7i} + I(D_{i1}=0, D_{i2}=0, \tilde{D}_{i1}=0, \tilde{D}_{i2}=1)\log L_{8i} +$$

$$I(D_{i1}=0, D_{i2}=0, \tilde{D}_{i1}=0, \tilde{D}_{i2}=0)\log L_{9i} \},$$ (4.15)

其中 $\log L_{ki}$ 為對各項 $L_k$ $(k = 1,2,...,9)$ 取 log 轉換。原有 $L_{ki}$ 內的每一項項數都是相乘的關係，經由 log 的轉換後，$\log L_k$ 內的每一項就可寫成相加的項，因此線性的關係簡化了期望值的操作。我們可將 $\log L_k$ 改寫成與參數 $\alpha, \beta$ 和 $(P_{11}, P_{10}, P_{01}, P_{00})$ 有關函數的線性組合，令

$$\tilde{l}(\alpha, P_{11}, P_{10}, P_{01}, P_{00}) = l_\alpha + l_\beta + l_P + l_C \ ,$$ (4.16)

其中

$$l_\alpha = \sum_{i=1}^{n} \{ I(D_{i1}=1, D_{i2}=1)\log C_\alpha^{11}(S_1(t_{i1}), S_2(t_{i2})) +$$

$$I(D_{i1}=1, D_{i2}=0, \tilde{D}_{i1}=1, \tilde{D}_{i2}=1)\log C_\alpha^{10}(S_1(t_{i1}), S_2(t_{i2})) +$$

$$I(D_{i1}=0, D_{i2}=1, \tilde{D}_{i1}=1, \tilde{D}_{i2}=1)\log C_\alpha^{01}(S_1(t_{i1}), S_2(t_{i2})) +$$

$$I(D_{i1}=0, D_{i2}=0, \tilde{D}_{i1}=1, \tilde{D}_{i2}=1)\log C_\alpha^{00}(S_1(t_{i1}), S_2(t_{i2})) \},$$

(4.17)

$$l_\beta = \sum_{i=1}^{n} \Big\{ I(D_{i1}=1, D_{i2}=1)\log D_\beta^{00}(H_1(t_{i1}), H_2(t_{i2})) +$$

$$I(D_{i1}=1, D_{i2}=0)\log D_\beta^{01}(H_1(t_{i2}), H_2(t_{i2})) +$$

$$I(D_{i1}=0, D_{i2}=1)\log D_\beta^{10}(H_1(t_{i2}), H_2(t_{i2})) +$$

$$I(D_{i1}=0, D_{i2}=0)\log D_\beta^{11}(H_1(t_{i1}), H_2(t_{i2})) \Big\}\,, \tag{4.18}$$

$$l_P = \sum_{i=1}^{n} \Big\{ \Big[ I(D_{i1}=1, D_{i2}=1) + I(D_{i1}=1, D_{i2}=0, \tilde{D}_{i1}=1, \tilde{D}_{i2}=1) +$$

$$I(D_{i1}=0, D_{i2}=1, \tilde{D}_{i1}=1, \tilde{D}_{i2}=1) +$$

$$I(D_{i1}=0, D_{i2}=0, \tilde{D}_{i1}=1, \tilde{D}_{i2}=1) \Big] \log \Pr(\tilde{D}_{i1}=1, \tilde{D}_{i2}=1) +$$

$$\Big[ I(D_{i1}=1, D_{i2}=0, \tilde{D}_{i1}=1, \tilde{D}_{i2}=0) +$$

$$I(D_{i1}=0, D_{i2}=0, \tilde{D}_{i1}=1, \tilde{D}_{i2}=0) \Big] \log \Pr(\tilde{D}_{i1}=1, \tilde{D}_{i2}=0) +$$

$$\Big[ I(D_{i1}=0, D_{i2}=1, \tilde{D}_{i1}=0, \tilde{D}_{i2}=1) +$$

$$I(D_{i1}=0, D_{i2}=0, \tilde{D}_{i1}=0, \tilde{D}_{i2}=1) \Big] \log \Pr(\tilde{D}_{i1}=0, \tilde{D}_{i2}=1) +$$

$$\Big[ I(D_{i1}=0, D_{i2}=0, \tilde{D}_{i1}=0, \tilde{D}_{i2}=0) \Big] \log \Pr(\tilde{D}_{i1}=0, \tilde{D}_{i2}=0) \Big\}\,, \tag{4.19}$$

$$l_C = \sum_{i=1}^{n} \Big\{ I(D_{i1}=1, D_{i2}=0, \tilde{D}_{i1}=1, \tilde{D}_{i2}=0)\log \Pr(X_1 = t_{i1} \mid \tilde{D}_{i1}=1) +$$

$$I(D_{i1}=0, D_{i2}=1, \tilde{D}_{i1}=0, \tilde{D}_{i2}=1)\log \Pr(X_2 = t_{i2} \mid \tilde{D}_{i2}=1) +$$

$$I(D_{i1}=0, D_{i2}=0, \tilde{D}_{i1}=1, \tilde{D}_{i2}=0)\log \Pr(X_1 > t_{i1} \mid \tilde{D}_{i1}=1) +$$

$$I(D_{i1}=0, D_{i2}=0, \tilde{D}_{i1}=0, \tilde{D}_{i2}=1)\log \Pr(X_2 > t_{i2} \mid \tilde{D}_{i2}=1) \Big\}\,。 \tag{4.20}$$

在此有兩項$l_\beta$ & $l_C$與參數$(\alpha, P_{11}, P_{10}, P_{01}, P_{00})$無關，所以這兩項可視為

$(\alpha, P_{11}, P_{10}, P_{01}, P_{00})$ 概似函數的常數項。因為隨機變數 $\tilde{D}_j$ $(j=1,2)$ 無法觀測到屬缺失值（missing data），所以我們提出用 EM 演算法計算 $\tilde{l}(\alpha, P_{11}, P_{10}, P_{01}, P_{00})$ 的極值。

**E-step:** 針對 For $l,k,p,q = 0 \text{ or } 1$

$$E\left[I(D_{i1}=l, D_{i2}=k, \tilde{D}_{i1}=p, \tilde{D}_{i2}=q) \mid D_{i1}=l, D_{i2}=k, T_1, T_2\right]$$

$$= I(D_{i1}=l, D_{i2}=k) \times \Pr\left[\tilde{D}_{i1}=p, \tilde{D}_{i2}=q \mid D_{i1}=l, D_{i2}=k, T_1=t_{i1}, T_2=t_{i2}\right]$$

$$= I(D_{i1}=l, D_{i2}=k) \times \frac{\Pr\left[\tilde{D}_{i1}=p, \tilde{D}_{i2}=q, D_{i1}=l, D_{i2}=k, T_1=t_{i1}, T_2=t_{i2}\right]}{\Pr\left[D_{i1}=l, D_{i2}=k, T_1=t_{i1}, T_2=t_{i2}\right]}。$$

(4.21)

**M-step:** 對 $\tilde{l}(\alpha, P_{11}, P_{10}, P_{01}, P_{00})$ 求極值

當對 $\tilde{l}(\alpha, P_{11}, P_{10}, P_{01}, P_{00})$ 微分時，因為函數的複雜度使我們利用數值方法 Newton-Raphson 法來求解。另外，與 $\alpha$ 有關的概似函數項 $l_\alpha$ 和與 $(P_{11}, P_{10}, P_{01}, P_{00})$ 有關的概似函數項 $l_P$ 是分開的，所以我們可以個別對 $\alpha$ 和 $(P_{11}, P_{10}, P_{01}, P_{00})$ 分別做估計。

（i） $\alpha$ 的估計：

$\alpha$ 對 $l(\alpha, P_{11}, P_{10}, P_{01}, P_{00})$ 微分相當於只需對 $l_\alpha$ 取微分，因為只有 $l_\alpha$ 與 $\alpha$ 有關，而其他項對 $\alpha$ 而言是常數項。故

$$\frac{\partial \tilde{l}(\alpha, P_{11}, P_{10}, P_{01}, P_{00})}{\partial \alpha}$$

$$= \frac{\partial l_\alpha}{\partial \alpha}$$

$$= \sum_{i=1}^{n} \Big\{ I(D_{i1}=1, D_{i2}=1) C_\alpha^{111}(S_1(t_{i1}), S_2(t_{i2})) +$$

$$I(D_{i1}=1, D_{i2}=0, \tilde{D}_{i1}=1, \tilde{D}_{i2}=1) C_\alpha^{110}(S_1(t_{i1}), S_2(t_{i2})) +$$

$$I(D_{i1}=0, D_{i2}=1, \tilde{D}_{i1}=1, \tilde{D}_{i2}=1) C_\alpha^{101}(S_1(t_{i1}), S_2(t_{i2})) +$$

$$I(D_{i1}=0,D_{i2}=0,\tilde{D}_{i1}=1,\tilde{D}_{i2}=1)C_{\alpha}^{100}(S_1(t_{i1}),S_2(t_{i2}))\Big\}\ , \qquad (4.22)$$

其中 $C_{\alpha}^{1kl}(s,t)=\dfrac{\partial \log\left[C_{\alpha}^{kl}(s,t)\right]}{\partial \alpha}$ $(k,l)=(0,0),(1,0),(0,1),(1,1)$

以牛頓法求$\alpha$的數值解，需要做以下二階導數的計算：

$$\frac{\partial^2 \tilde{l}(\alpha,P_{11},P_{10},P_{01},P_{00})}{\partial \alpha^2}$$

$$=\frac{\partial^2 l_{\alpha}}{\partial \alpha^2}$$

$$=\sum_{i=1}^{n}\Big\{I(D_{i1}=1,D_{i2}=1)C_{\alpha}^{211}(S_1(t_{i1}),S_2(t_{i2}))+$$

$$I(D_{i1}=1,D_{i2}=0,\tilde{D}_{i1}=1,\tilde{D}_{i2}=1)C_{\alpha}^{210}(S_1(t_{i1}),S_2(t_{i2}))+$$

$$I(D_{i1}=0,D_{i2}=1,\tilde{D}_{i1}=1,\tilde{D}_{i2}=1)C_{\alpha}^{201}(S_1(t_{i1}),S_2(t_{i2}))+$$

$$I(D_{i1}=0,D_{i2}=0,\tilde{D}_{i1}=1,\tilde{D}_{i2}=1)C_{\alpha}^{200}(S_1(t_{i1}),S_2(t_{i2}))\Big\}\ , \qquad (4.23)$$

其中 $C_{\alpha}^{2kl}(s,t)=\dfrac{\partial^2 \log\left[C_{\alpha}^{kl}(s,t)\right]}{\partial \alpha^2}$ $(k,l)=(0,0),(1,0),(0,1),(1,1)$。值得一提的是在這些式子中，因為 $\tilde{D}_j$ $(j=1,2)$ 是缺失值，所以對於 $l,k,p,q=0$ 或1 ，

$I(D_{i1}=l,D_{i2}=k,\tilde{D}_{i1}=p,\tilde{D}_{i2}=q)$ 的值要利用 E-step 中求得的

$E\left[I(D_{i1}=l,D_{i2}=k,\tilde{D}_{i1}=p,\tilde{D}_{i2}=q)\mid D_{i1}=l,D_{i2}=k,T_1,T_2\right]$來取代之。以下是 E-step 中八種條件期望值，

$$E\left[I(D_{i1}=1,D_{i2}=0,\tilde{D}_{i1}=1,\tilde{D}_{i2}=1)\mid D_{i1}=1,D_{i2}=0,T_1,T_2\right]$$

$$=I(D_{i1}=1,D_{i2}=0)w_{1i}\ ,$$

其中 $w_{1i}=\dfrac{C_{\alpha}^{10}(S_1(t_{i1}),S_2(t_{i2}))\times P_{11}}{C_{\alpha}^{10}(S_1(t_{i1}),S_2(t_{i2}))\times P_{11}+S_1(\Delta t_{i1})\times P_{10}}$ ；

$$E\left[I(D_{i1}=1,D_{i2}=0,\tilde{D}_{i1}=1,\tilde{D}_{i2}=0)\mid D_{i1}=1,D_{i2}=0,T_1,T_2\right]$$

$$=I(D_{i1}=1,D_{i2}=0)w_{2i}\ ,$$

其中 $w_{2i} = \dfrac{S_1(\Delta t_{i1}) \times P_{10}}{C_\alpha^{10}(S_1(t_{i1}), S_2(t_{i2})) \times P_{11} + S_1(\Delta t_{i1}) \times P_{10}}$ ；

$$E\left[I(D_{i1}=0, D_{i2}=1, \tilde{D}_{i1}=1, \tilde{D}_{i2}=1) \mid D_{i1}=0, D_{i2}=1, T_1, T_2\right]$$

$$= I(D_{i1}=0, D_{i2}=1) w_{3i} \text{ ，}$$

其中 $w_{3i} = \dfrac{C_\alpha^{01}(S_1(t_{i1}), S_2(t_{i2})) \times P_{11}}{C_\alpha^{01}(S_1(t_{i1}), S_2(t_{i2})) \times P_{11} + S_2(\Delta t_{i2}) \times P_{01}}$ ；

$$E\left[I(D_{i1}=0, D_{i2}=1, \tilde{D}_{i1}=0, \tilde{D}_{i2}=1) \mid D_{i1}=0, D_{i2}=1, T_1, T_2\right]$$

$$= I(D_{i1}=0, D_{i2}=1) w_{4i} \text{ ，}$$

其中 $w_{4i} = \dfrac{S_2(\Delta t_{i2}) \times P_{01}}{C_\alpha^{01}(S_1(t_{i1}), S_2(t_{i2})) \times P_{11} + S_2(\Delta t_{i2}) \times P_{01}}$ ；

$$E\left[I(D_{i1}=0, D_{i2}=0, \tilde{D}_{i1}=1, \tilde{D}_{i2}=1) \mid D_{i1}=0, D_{i2}=0, T_1, T_2\right]$$

$$= I(D_{i1}=0, D_{i2}=0) w_{5i} \text{ ，}$$

其中 $w_{5i} = \dfrac{C_\alpha^{00}(S_1(t_{i1}), S_2(t_{i2})) \times P_{11}}{\left[C_\alpha^{00}(S_1(t_{i1}), S_2(t_{i2})) \times P_{11} + S_1(t_{i1}) \times P_{10} + S_2(t_{i2}) \times P_{01} + P_{00}\right]}$ ；

$$E\left[I(D_{i1}=0, D_{i2}=0, \tilde{D}_{i1}=1, \tilde{D}_{i2}=0) \mid D_{i1}=0, D_{i2}=0, T_1, T_2\right]$$

$$= I(D_{i1}=0, D_{i2}=0) w_{6i} \text{ ，}$$

其中 $w_{6i} = \dfrac{S_1(t_{i1}) \times P_{10}}{\left[C_\alpha^{00}(S_1(t_{i1}), S_2(t_{i2})) \times P_{11} + S_1(t_{i1}) \times P_{10} + S_2(t_{i2}) \times P_{01} + P_{00}\right]}$ ；

$$E\left[I(D_{i1}=0, D_{i2}=0, \tilde{D}_{i1}=0, \tilde{D}_{i2}=1) \mid D_{i1}=0, D_{i2}=0, T_1, T_2\right]$$

$$= I(D_{i1}=0, D_{i2}=0) w_{7i} \text{ ，}$$

其中 $w_{7i} = \dfrac{S_2(t_{i2}) \times P_{01}}{\left[C_\alpha^{00}(S_1(t_{i1}), S_2(t_{i2})) \times P_{11} + S_1(t_{i1}) \times P_{10} + S_2(t_{i2}) \times P_{01} + P_{00}\right]}$ ；

$$E\left[I(D_{i1}=0, D_{i2}=0, \tilde{D}_{i1}=0, \tilde{D}_{i2}=0) \mid D_{i1}=0, D_{i2}=0, T_1, T_2\right]$$

$$= I(D_{i1}=0, D_{i2}=0) w_{8i} \text{ ，}$$

其中 $w_{8i} = \dfrac{P_{00}}{\left[C_\alpha^{00}(S_1(t_{i1}), S_2(t_{i2})) \times P_{11} + S_1(t_{i1}) \times P_{10} + S_2(t_{i2}) \times P_{01} + P_{00}\right]}$ 。

利用上述 8 種情況的條件期望值估計缺失值。另外，(4.23)與(4.24)是概似函數 $l_\alpha$ 對 $\alpha$ 的一階和二皆微分式。所以，估計 $\alpha$ 的牛頓法公式如下：

$$\alpha^{(k+1)} = \alpha^{(k)} - \frac{\frac{\partial}{\partial \alpha} l_\alpha(\alpha, P_{11}^{(k)}, P_{10}^{(k)}, P_{01}^{(k)})|_{\alpha=\alpha^{(k)}}}{\frac{\partial^2}{\partial \alpha^2} l_\alpha(\alpha, P_{11}^{(k)}, P_{10}^{(k)}, P_{01}^{(k)})|_{\alpha=\alpha^{(k)}}} \text{。}$$

（ii） $P_{11}, P_{10}, P_{01}$ 的估計：

因為 $P_{00} = 1 - P_{11} - P_{10} - P_{01}$ 為 $P_{11}, P_{10}, P_{01}$ 的線性組合，故只需要估計出 $P_{11}, P_{10}, P_{01}$ 就能得到 $P_{00}$ 的估計值，若做了可交換性的假設，則參數的個數會減少。同理 $P_{11}, P_{10}, P_{01}$ 個別對 $\tilde{l}(\alpha, P_{11}, P_{10}, P_{01}, P_{00})$ 微分相當於只需對 $l_P$ 取微分，故

$$\frac{\partial l_P}{\partial P_{11}} = \sum_{i=1}^{n} \{ [ I(D_{i1}=1, D_{i2}=1) + I(D_{i1}=1, D_{i2}=0, \tilde{D}_{i1}=1, \tilde{D}_{i2}=1) + I(D_{i1}=0, D_{i2}=1, \tilde{D}_{i1}=1, \tilde{D}_{i2}=1) + I(D_{i1}=0, D_{i2}=0, \tilde{D}_{i1}=1, \tilde{D}_{i2}=1) ] \frac{1}{P_{11}} - [ I(D_{i1}=0, D_{i2}=0, \tilde{D}_{i1}=0, \tilde{D}_{i2}=0) ] \frac{1}{1 - P_{11} - 2P_{01}} \} , \quad (4.24)$$

$$\frac{\partial l_P}{\partial P_{10}} = \sum_{i=1}^{n} \{ [ I(D_{i1}=1, D_{i2}=0, \tilde{D}_{i1}=1, \tilde{D}_{i2}=0) + I(D_{i1}=0, D_{i2}=1, \tilde{D}_{i1}=0, \tilde{D}_{i2}=1) + I(D_{i1}=0, D_{i2}=0, \tilde{D}_{i1}=0, \tilde{D}_{i2}=1) + I(D_{i1}=0, D_{i2}=0, \tilde{D}_{i1}=1, \tilde{D}_{i2}=0) ] \frac{1}{P_{10}} - [ I(D_{i1}=0, D_{i2}=0, \tilde{D}_{i1}=0, \tilde{D}_{i2}=0) ] \frac{2}{(1 - P_{11} - 2P_{01})} \} , \quad (4.25)$$

同理 $\tilde{D}_j$ $(j = 1,2)$ 是缺失值，所以 $I(D_{i1}=l, D_{i2}=k, \tilde{D}_{i1}=p, \tilde{D}_{i2}=q)$ $(l,k,p,q = 0 \text{ or } 1)$ 的值利用 E-step 中求得的 $E[I(D_{i1}=l, D_{i2}=k, \tilde{D}_{i1}=p, \tilde{D}_{i2}=q) \mid D_{i1}=l, D_{i2}=k, , T_1, T_2]$ 來取代之。同樣利用牛頓法，解出 $P_{11}, P_{10}=P_{01}$ 的數值解，公式如下：

$$\begin{bmatrix} P_{11}^{(k+1)} \\ P_{10}^{(k+1)} \end{bmatrix} = \begin{bmatrix} P_{11}^{(k)} \\ P_{10}^{(k)} \end{bmatrix} - \left[ \begin{matrix} \frac{\partial^2 l_P}{\partial P_{11}^2} & \frac{\partial^2 l_P}{\partial P_{11}\partial P_{10}} \\ \frac{\partial^2 l_P}{\partial P_{11}\partial P_{10}} & \frac{\partial^2 l_P}{\partial P_{10}^2} \end{matrix} \right]^{-1} \left[ \begin{matrix} \frac{\partial\, l_P}{\partial P_{11}} \\ \frac{\partial\, l_P}{\partial P_{10}} \end{matrix} \right] \Bigg|_{\begin{bmatrix} P_{11} \\ P_{10} \end{bmatrix} = \begin{bmatrix} P_{11}^{(k)} \\ P_{10}^{(k)} \end{bmatrix}},$$

(4.26)

其中

$$\frac{\partial^2 l_P}{\partial P_{11}^{\ 2}} = \sum_{i=1}^{n} \{ \left[ I(D_{i1}=1, D_{i2}=1) + I(D_{i1}=1, D_{i2}=0, \tilde{D}_{i1}=1, \tilde{D}_{i2}=1) + \right.$$

$$I(D_{i1}=0, D_{i2}=1, \tilde{D}_{i1}=1, \tilde{D}_{i2}=1) + I(D_{i1}=0, D_{i2}=0, \tilde{D}_{i1}=1, \tilde{D}_{i2}=1) \left. \right] \times$$

$$\frac{-1}{P_{11}^{\ 2}} - \left[ I(D_{i1}=0, D_{i2}=0, \tilde{D}_{i1}=0, \tilde{D}_{i2}=0) \right] \frac{1}{(1-P_{11}-2P_{01})^2} \Bigg\} \qquad (4.27)$$

$$\frac{\partial^2 l_P}{\partial P_{10}^{\ 2}} = \sum_{i=1}^{n} \{ \left[ I(D_{i1}=1, D_{i2}=0, \tilde{D}_{i1}=1, \tilde{D}_{i2}=0) + I(D_{i1}=0, D_{i2}=1, \tilde{D}_{i1}=0, \tilde{D}_{i2}=1) + \right.$$

$$I(D_{i1}=0, D_{i2}=0, \tilde{D}_{i1}=0, \tilde{D}_{i2}=1) + I(D_{i1}=0, D_{i2}=0, \tilde{D}_{i1}=1, \tilde{D}_{i2}=0) \left. \right] \times$$

$$\frac{-1}{P_{10}^2} - \left[ I(D_{i1}=0, D_{i2}=0, \tilde{D}_{i1}=0, \tilde{D}_{i2}=0) \right] \frac{4}{(1-P_{11}-2P_{01})^2} \Bigg\}, \qquad (4.28)$$

$$\frac{\partial^2 l_P}{\partial P_{11}\partial P_{10}} = \frac{\partial^2 l_P}{\partial P_{01}\partial P_{11}} = \sum_{i=1}^{n} \left\{ - I(D_{i1}=0, D_{i2}=0, \tilde{D}_{i1}=0, \tilde{D}_{i2}=0) \frac{2}{(1-P_{11}-2P_{01})^2} \right\}, \qquad (4.29)$$

為了運算的方便，我們計算了訊息矩陣的反函數如下：

$$\left[ \begin{matrix} \frac{\partial^2 l_P}{\partial P_{11}^2} & \frac{\partial^2 l_P}{\partial P_{11}\partial P_{10}} \\ \frac{\partial^2 l_P}{\partial P_{11}\partial P_{10}} & \frac{\partial^2 l_P}{\partial P_{10}^2} \end{matrix} \right]^{-1} = \begin{bmatrix} a & c \\ c & b \end{bmatrix}^{-1} = \frac{1}{ab-c^2} \begin{bmatrix} b & -c \\ -c & a \end{bmatrix} \qquad (4.30)$$

其中 $a = \frac{\partial^2 l_P}{\partial P_{11}^2}$ ， $b = \frac{\partial^2 l_P}{\partial P_{10}^2}$ 和 $c = \frac{\partial^2 l_P}{\partial P_{11}\partial P_{10}}$ 。

# 第五章 模擬實驗

## 5.1 模擬資料生成

在本章我們利用電腦模擬以檢驗先前提出的推論方法在有限樣本的表現。 在先前的討論中，我們假設發病時間的聯合分配和競爭風險發生時間的聯合分配均呈現 copula 模式分配。在模擬程式中，我們選擇 Clayton 分配，其聯合分配函數可表示為

$$C_\alpha(u,v)=\begin{cases}(u^{1-\alpha}+v^{1-\alpha}-1)^{\frac{1}{1-\alpha}} & if\ \alpha>1\\ uv & if\ \alpha=1\end{cases}\quad 。$$

其中 $\tau_\alpha=\dfrac{\alpha-1}{\alpha+1}$。模擬實驗資料的生成過程整理如下：

第一步:生成具有關聯性的離散型變數$(\tilde{D}_1,\tilde{D}_2)$：

令 $\Pr(\tilde{D}_1=i,\tilde{D}_2=j)=P_{ij}\ \ (i,j=0,1)$ ，$\gamma=\dfrac{P_{11}P_{00}}{P_{10}P_{01}}$ 。

（i） 先給定 $P_{11}$ 和 $\gamma$ ；

（ii） 利用下列兩項條件：

(1) $\gamma=\dfrac{P_{11}P_{00}}{P_{10}^2}$ ，

(2) $P_{11}+P_{00}+2P_{10}=1\Rightarrow P_{00}=1-P_{11}-2P_{10}$ ，

可求出 $P_{10}=\dfrac{-a+\sqrt{a^2-4b}}{2}$ 和 $P_{00}=1-P_{11}-2P_{10}$ ，

其中 $a=\dfrac{2P_{11}}{\gamma},\quad b=\dfrac{P_{11}^{\ 2}}{\gamma}-\dfrac{P_{11}}{\gamma}=\dfrac{P_{11}}{\gamma}(P_{11}-1)$ 。

（iii） 生成變數$U_i\ \ (i=1,2,...,n)$ ，其中$U_i\sim U(0,1)$ 且$U_i$ 與$U_j$ 獨立。

（iv） if $\quad U_i<P_{11}\Rightarrow\tilde{D}_{1i}=1,\tilde{D}_{2i}=1$

else if $\quad U_i<P_{11}+P_{10}\Rightarrow\tilde{D}_{1i}=1,\tilde{D}_{2i}=0$

else if $\ U_i<P_{11}+P_{10}+P_{01}\Rightarrow\tilde{D}_{1i}=0,\tilde{D}_{2i}=1$

else $\quad \tilde{D}_{1i}=0, \tilde{D}_{2i}=0$

因此可得$(\tilde{D}_{1i}, \tilde{D}_{2i})$ $(i=1,2,...,n)$。

第二步: 生成 $X_1, X_2 \mid \tilde{D}_1=1, \tilde{D}_2=1 \sim Clayton\,(\tau_\alpha)$：

(i) 先計算 $cc=\dfrac{1-\tau_\alpha}{1-\tau_\alpha}$；

(ii) 再生成彼此獨立的變數$(U_{1i}, U_{2i})$ $(i=1,2,...,n)$，其中

$U_{ji} \sim U(0,1)$ $(j=1,2)$；

(iii) 令 $X_j \mid \tilde{D}_j=1$ $(j=1,2)$ 的邊際分配為 exp(1)，則根據 Prentice and Cai (1993)的程式，生成$(X_{1i}, X_{2i})$如下：

$a_i=(1-U_{2i})^{-\frac{1}{cc}}$，

$X_{2i}=-\log(1-U_{2i})$，

$X_{1i}=cc\times\log[1-a_i+a_i\times(1-U_{1i})^{-\frac{1}{1+cc}}]$；

(iv) if$(\tilde{D}_{1i}=1, \tilde{D}_{2i}=1)$，set$(X_{1i}^*, X_{2i}^*)=(X_{1i}, X_{2i})$

else if $(\tilde{D}_{1i}=1, \tilde{D}_{2i}=0)$，set$(X_{1i}^*, X_{2i}^*)=(X_{1i}, \infty)$

else if $(\tilde{D}_{1i}=0, \tilde{D}_{2i}=1)$，set$(X_{1i}^*, X_{2i}^*)=(\infty, X_{2i})$

else $(\tilde{D}_{1i}=0, \tilde{D}_{2i}=0)$，set$(X_{1i}^*, X_{2i}^*)=(\infty, \infty)$。

重複以上步驟$n$次可得$(\tilde{D}_{1i}, \tilde{D}_{2i}, X_{1i}^*, X_{2i}^*)$ $(i=1,2,...,n)$。

第三步: 生成$(Y_1, Y_2) \sim Clayton\ \ distribution(\tau_\beta)$：

生成方式如同第二步驟中的(i)~(iii)，重複$n$次可得$(Y_{1i}, Y_{2i})$ $(i=1,2,...,n)$。

第四步: 製造觀察到的時間$(T_1, T_2)$和生前經歷發病與否的指標$(D_1, D_2)$：

If $X_{1i}^* \le Y_{1i}, X_{2i}^* \le Y_{2i}$，set$(T_{1i}, T_{2i})=(X_{1i}^*, X_{2i}^*)$和$(D_{1i}, D_{2i})=(1,1)$

else if $X_{1i}^{*} \le Y_{1i}, X_{2i}^{*} > Y_{2i}$ ，$\text{set}(T_{1i}, T_{2i}) = (X_{1i}^{*}, Y_{2i})$ 和 $(D_{1i}, D_{2i}) = (1,0)$

else if $X_{1i}^{*} > Y_{1i}, X_{2i}^{*} \le Y_{2i}$ ，$\text{set}(T_{1i}, T_{2i}) = (Y_{1i}, X_{2i}^{*})$ 和 $(D_{1i}, D_{2i}) = (0,1)$

else $X_{1i}^{*} > Y_{1i}, X_{2i}^{*} > Y_{2i}$ ，$\text{set}(T_{1i}, T_{2i}) = (Y_{1i}, Y_{2i})$ 和 $(D_{1i}, D_{2i}) = (0,0)$ 。

可得資料 $(T_{1i}, T_{2i}, D_{1i}, D_{2i})$ $(i = 1,2,...,n)$ 。

## 5.2 模擬結果

我們建立模擬實驗來評估論文中提出的二階段估計法。個別樣本大小（sample size）為 $n = 500$ 和 $n = 1000$ 的成對資料，而模擬實驗的重複次數則是樣本大小的 2 倍，即個別重複 1000 次和 2000 次，再求平均誤差與標準差。我們選用 Clayton 模式來描述姊妹倆發病時間的聯合分配和競爭風險發生時間的聯合分配，而描述關聯性的參數之設定方式則是令 $\tau_\alpha = 0.3, 0.5, 0.7$ 和 $\tau_\beta = 0.3, 0.5$ 配對成六種情形，由此轉換為 $\alpha$ 和 $\beta$ 值。另一方面我們固定姊妹倆帶因與否的關聯性為 $\gamma = 0.8$ ，而姊妹皆帶因的機率 $(\Pr(\tilde{D}_1 = 1, \tilde{D}_2 = 1))$ 設定為 0.2，經由 5.1 節的第一步之（ii）的計算，可得其他三種帶因與否的參數值各為 $\Pr(\tilde{D}_1 = 1, \tilde{D}_2 = 0) = \Pr(\tilde{D}_1 = 0, \tilde{D}_2 = 1) = 0.262348$ 及 $\Pr(\tilde{D}_1 = 0, \tilde{D}_2 = 0) = 0.275305$ 。

圖表（一）是當 $\tau_\beta = 0.3$ 和 $\tau_\alpha = 0.3$(弱相關)，0.5（中度相關），0.7（高度相關）時，分別利用有母數和無母數方法估計邊際分配和單維度個體可致病機率的模擬結果。我們令 $\hat{P}_1$ 和 $\hat{P}_2$ 為 $\Pr(\tilde{D}_1 = 1)$ 和 $\Pr(\tilde{D}_2 = 1)$ 的估計值（而真值皆為 0.462348）；另外在有母數的估計方法中，我們假設姊姊/妹妹的邊際分佈為指數分配（exponential distribution），因此 $\lambda_1$ 和 $\lambda_2$ 為指數分配母數（真值為 $\lambda_j = 1$ $(j = 1,2)$ ），而 $\hat{\lambda}_j$ $(j = 1,2)$ 為母數的估計值。由圖表（一）的結果發現不論是用有母數法或是無母數法得之 $\hat{P}_j$ $(j = 1,2)$ 估計量偏差均理想。當 $\tau_\beta = 0.3$ 和 $\tau_\alpha = 0.3, 0.5, 0.7$ 下，母數

方法所估計 $\lambda_j = 1 (j = 1,2)$ 的估計量的表現並沒有很好，因為估計量的準確度只到小數第二位，而標準差也很大。當樣本數增加時（樣本大小由 500 增加一倍到 1000），其偏差值和標準差會略微降低。儘管如此在相同的條件下利用有母數的估計方法所估計出來的 $\hat{P}_j$ $(j = 1,2)$ 值仍具有較小的標準差。

圖表（二）、圖表（三）和圖表（四）則是令真值 $\tau_\alpha = 0.3, 0.5, 0.7$ 與 $\tau_\beta = 0.3, 0.5$ 配對成六種情形和樣本大小各為 $n = 500$ 和 $n = 1000$ 之下，對 $\tau_\alpha$、$\tau_\beta$、$P_{11}$ 和 $P_{10}$ 估計的模擬結果。圖表（二）中顯示，不論 $\tau_\alpha$ 和 $\tau_\beta$ 的真值如何的變化，對 $P_{11}$ 和 $P_{10}$ 所做的有母數和無母數估計都具有小的偏差值，但是用無母數估計方法會得到較大的標準差。與圖表（一）的表現相同，當樣本數增加時（樣本大小由 500 增加一倍到 1000），其偏差值和標準差會略微降低。

對於競爭風險發生時間之關聯性（$\beta$）的估計，我們只採用無母數（Kaplan-Meier estimator）估計競爭風險發生時間的邊際分配。理由之一是 $(Y_1, Y_2)$ 呈現傳統右設限資料，沒有所為免疫的問題。其次是 $Y_j (j = 1,2)$ 非感興趣的變數，所以不希望給予強的模式假設。另外由於模擬的結果顯示死亡的次數較發病的次數多，因此關聯性 $\tau_\beta$ 的估計值較為 efficient。這樣的模擬設定應該是與實際現象相符合，以研究罕見疾病為例，只會有少數人會發生該疾病；即使有帶因的話，也有可能被死亡所設限。我們發現對於帶因者發病時間之關聯性（$\alpha$）的估計，即使是用有母數估計法來估計邊際機率，但是 $\hat{\tau}_\alpha$ 的標準差仍比以無母數方法做為邊際估計量之 $\hat{\tau}_\beta$ 的標準差大。至於用無母數估計法估計 $\tau_\alpha$，在偏差值和標準差的表現就顯得更差。若要進一步改善 $\tau_\alpha$ 估計在偏差值和標準差的表現的話，就要增加樣本大小。在圖表（五）是令樣本大小為 5000，實驗的次數降為 200 次的結果。以圖表（五）與圖表（三）和圖表（四）的數據作比較，發現樣本數增加對 $\tau_\alpha$ 估

計的偏差值和標準差確實有所改進。其實不僅僅是$\tau_\alpha$有所改善，其他的參數在偏差值和標準差的表現也改進許多。另外比較圖表（三）、圖表（四）和圖表（五）會發現當$\tau_\alpha$增加，所有估計量都表現得更好。在圖表（六）是在不同的樣本數下，帶因者發病時間的存活函數，當樣本數越大時，Kaplan-Meier 曲線就與真值 exponential(1)的曲線越接近。

| | | | N=500 Rep=1000 | N=1000 Rep=2000 |
|---|---|---|---|---|
| | | | $\tau_\beta = 0.3$ | $\tau_\beta = 0.3$ |
| | | | Bias$\times 10^{-2}$ | Bias$\times 10^{-2}$ |
| | | | (St.error$\times 10^{-2}$) | (St.error$\times 10^{-2}$) |
| $\tau_\alpha = 0.3$ | $\hat{P}_1$ | P | 0.33(5.46) | 0.10(3.65) |
| | | NP | -0.09(6.15) | -0.04(4.63) |
| | $\hat{P}_2$ | P | 0.40(5.23) | 0.21(3.69) |
| | | NP | 0.50(5.93) | -0.14(4.67) |
| | $\hat{\lambda}_1$ | P | 1.42(18.52) | 0.68(12.53) |
| | $\hat{\lambda}_2$ | P | 2.04(18.33) | 0.78(12.73) |
| $\tau_\alpha = 0.5$ | $\hat{P}_1$ | P | 0.26(5.21) | 0.27(3.72) |
| | | NP | -0.18(5.98) | -0.18 (4.76) |
| | $\hat{P}_2$ | P | 0.24(5.24) | 0.20(3.77) |
| | | NP | 0.33(6.78) | -0.18(4.61) |
| | $\hat{\lambda}_1$ | P | 1.97(18.41) | 0.89(12.93) |
| | $\hat{\lambda}_2$ | P | 1.21(18.59) | 0.89(12.79) |
| $\tau_\alpha = 0.7$ | $\hat{P}_1$ | P | 0.00(5.45) | 0.13(3.74) |
| | | NP | -0.64(5.95) | -0.16(4.72) |
| | $\hat{P}_2$ | P | 0.29(5.46) | 0.07(3.75) |
| | | NP | -0.51(5.97) | -0.20(4.60) |
| | $\hat{\lambda}_1$ | P | 1.39(18.88) | 0.94(12.91) |
| | $\hat{\lambda}_2$ | P | 1.37(18.96) | 0.71 (12.96) |

***圖表(一):在樣本大小為 500 和 1000 下,利用有母數估計法("P")和無母數估計法("NP")估計邊際分配和個體帶因與否的機率之估計值的比較。***

| | | | N=500 Rep=1000 | N=1000 Rep=2000 |
|---|---|---|---|---|
| | | | $\tau_\beta = 0.3$ | $\tau_\beta = 0.3$ |
| | | | Bias$\times 10^{-2}$ (St.error$\times 10^{-2}$) | Bias$\times 10^{-2}$ (St.error$\times 10^{-2}$) |
| $\tau_\alpha = 0.3$ | $\hat{\tau}_\alpha$ | P | 1.88(12.37) | -0.09(9.78) |
| | | NP | -9.31(14.32) | -8.39(11.60) |
| | $\hat{\tau}_\beta$ | NP | -0.38(3.04) | -0.31(2.19) |
| | $P_{11}$ | P | 1.56(3.62) | 0.70(2.63) |
| | | NP | 0.22(4.43) | 0.45(3.50) |
| | $P_{10}$ | P | 0.26(2.78) | 0.17(2.10) |
| | | NP | -1.39(3.66) | -0.92(2.52) |

***圖表(二):在樣本大小為 500 和 1000 下,給定$\tau_\alpha = 0.3$和$\tau_\beta = 0.3$利用有母數估計法("P")和無母數估估計法("NP")估計參數$\tau_\alpha$、$\tau_\beta$、$P_{11}$和$P_{10}$。***

| | | | N=500 Rep=1000 | | N=1000 Rep=2000 | |
|---|---|---|---|---|---|---|
| | | | $\tau_\beta = 0.3$ | $\tau_\beta = 0.5$ | $\tau_\beta = 0.3$ | $\tau_\beta = 0.5$ |
| | | | Bias$\times 10^{-2}$ (St.error$\times 10^{-2}$) | Bias$\times 10^{-2}$ (St.error$\times 10^{-2}$) | Bias$\times 10^{-2}$ (St.error$\times 10^{-2}$) | Bias$\times 10^{-2}$ (St.error$\times 10^{-2}$) |
| $\tau_\alpha = 0.5$ | $\hat{\tau}_\alpha$ | P | 0.46(8.99) | 0.04(8.68) | 0.16(6.28) | 0.40(5.84) |
| | | NP | -12.68(16.59) | -16.9815(17.54) | -10.27(12.45) | -12.23(12.49) |
| | $\hat{\tau}_\beta$ | NP | -0.41(3.12) | -0.80(2.55) | -0.36(2.17) | -0.38(1.88) |
| | $\hat{P}_{11}$ | P | 1.35(3.13) | 1.46(3.21) | 0.78(2.24) | 0.82(2.20) |
| | | NP | 0.82(4.02) | 0.93(4.28) | 0.56(3.24) | 0.71(3.37) |
| | $\hat{P}_{10}$ | P | 0.59(2.44) | 0.34(2.33) | 0.37(1.67) | 0.35(1.62) |
| | | NP | -1.05(2.93) | -1.53(2.97) | -0.88(2.18) | -1.03(2.07) |

***圖表（三）：在樣本大小為 500 和 1000，令$\tau_\alpha = 0.5$和$\tau_\beta = 0.3,0.5$，利用有母數估計法和無母數估估計法估計參數$\tau_\alpha$、$\tau_\beta$、$P_{11}$和$P_{10}$。***

| | | | N=500 Rep=1000 | | N=1000 Rep=2000 | |
|---|---|---|---|---|---|---|
| | | | $\tau_\beta = 0.3$ | $\tau_\beta = 0.5$ | $\tau_\beta = 0.3$ | $\tau_\beta = 0.5$ |
| | | | Bias$\times 10^{-2}$ (St.error$\times 10^{-2}$) | Bias$\times 10^{-2}$ (St.error$\times 10^{-2}$) | Bias$\times 10^{-2}$ (St.error$\times 10^{-2}$) | Bias$\times 10^{-2}$ (St.error$\times 10^{-2}$) |
| $\tau_\alpha = 0.7$ | $\hat{\tau}_\alpha$ | P | 0.29 (4.6) | 0.45 (4.26) | 0.19(3.26) | 0.15(3.13) |
| | | NP | -15.30(17.53) | -17.34(18.21) | -9.24(10.07) | -10.67(10.80) |
| | $\hat{\tau}_\beta$ | NP | -0.43(3.08) | -0.58(2.37) | -0.32(2.16) | -0.32(1.82) |
| | $\hat{P}_{11}$ | P | 0.97(2.93) | 1.36(3.05) | 0.63(2.01) | 0.66(2.06) |
| | | NP | 0.64(3.72) | 1.20(4.03) | 0.48 (2.99) | 0.57(2.64) |
| | $\hat{P}_{10}$ | P | 0.79(2.11) | 0.73(2.17) | 0.41(1.55) | 0.34(1.47) |
| | | NP | -1.44(2.76) | -1.48(2.83) | -0.74(2.02) | -0.90(2.04) |

***圖表（四）：在樣本大小為 500 和 1000，令$\tau_\alpha = 0.7$和$\tau_\beta = 0.3,0.5$，利用有母數估計法和無母數估估計法估計參數$\tau_\alpha$、$\tau_\beta$、$P_{11}$和$P_{10}$。***

N=5000　Rep=200　$\tau_\alpha = 0.5$　$\tau_\beta = 0.3$

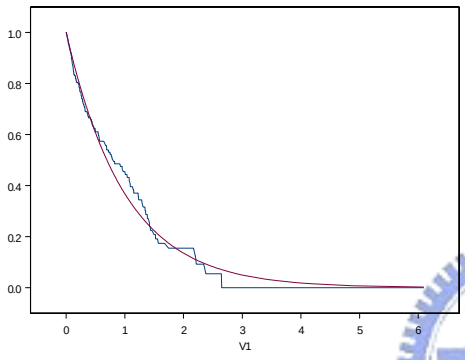
**單維度的估計：**

| | **有母數估計法：** | **無母數估計法：** |
|---|---|---|
| | $Bias \times 10^{-2}$ $(St.error \times 10^{-2})$ | $Bias \times 10^{-2}$ $(St.error \times 10^{-2})$ |
| $\hat{P}_1$ | 0.06(1.71) | 0.28(3.57) |
| $\hat{P}_2$ | -0.08(1.74) | 0.08(2.95) |
| $\hat{\lambda}_1$ | 0.59(5.87) | --- |
| $\hat{\lambda}_2$ | -0.13(5.78) | --- |

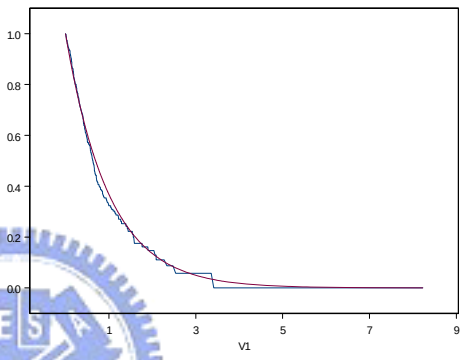
**雙維度的估計：**

| | | |
|---|---|---|
| $\hat{\tau}_\alpha$ | 0.09(2.76) | -2.57(4.05) |
| $\hat{\tau}_\beta$ | --- | -0.13(0.91) |
| $\hat{P}_{11}$ | 0.14(0.94) | 0.14(1.49) |
| $\hat{P}_{10}$ | 0.16(0.68) | -0.08(1.03) |

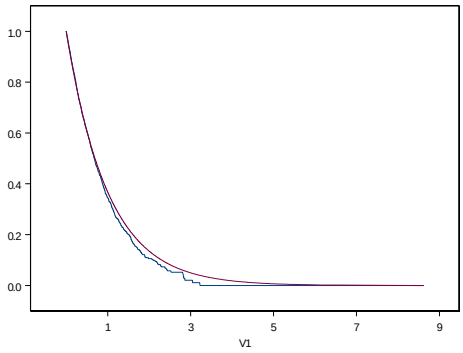
***圖表（四）：在樣本大小為 5000，令$\tau_\alpha = 0.5$和$\tau_\beta = 0.3$，利用有母數估計法和無母數估估計法對各種參數做估計。***




N=500




N=1000




N=5000

***圖表（五）：樣本大小（n）不相同時，Kaplan-Meier 曲線與 exponential(1) 曲線。***

# 第六章　結論

本論文將二維存活分析（bivariate survival analysis）和治癒模式（cure model）結合，以探討遺傳現象展現於發病（incidence)和發病年齡間(age of onset)的關聯性。我們的論文和 Chatterjee & Shih（2001）所探討的問題相同，最大的差別在於我們的架構中直接考慮了“死亡”這個競爭風險，而 Chatterjee & Shih 似乎把死亡當成是設限的原因之一，換言之她們的架構未將外生設限和因死亡競爭風險的設限予以分開。然而 Chatterjee & Shih 意識到死亡對發病是無法避免的競爭風險，並在其文章最後討論的部份提及了解釋力（interpretability）和模式的辨識性（identifiability）的問題。該文強調所分析的 WAS 乳癌資料，最晚發病的乳癌患者發病年齡是 91 歲，但是仍有一些活得更久的人未見發病，作者認為在此資料中應有充份證據顯示 91 歲已是乳癌發病年齡的上限。

我們的模式直接把死亡的影響納入了分析的架構中，在單維度的分析裏，我們以無母數方法分析競爭風險的影響，在雙維度的分析中，我們以 copula 模式描述親屬競爭風險間的關聯性。我們提出的推論方法在假設 $X^* \perp Y$ 的情形下，會和 Chatterjee & Shih 得到一致的結論，然而當兩者非獨立時，我們的架構容許做適度的修正。例如可以 frailty 角度假設兩者具有條件獨立的關係：$X \perp Y \mid \xi$，其中 $\xi$ 代表解釋兩者關聯性的變數（如環境因子），此時要建構 $(X,Y)$ 的聯合分配只需要對 $\xi$ 的分配積分即可。此外，我們的方法可利用以下關係式以推估帶因者在生前未發病的比例：

$$\Pr(X > Y \mid \tilde{D} = 1) = E_Y[S(Y)] = -\int_y S(y)H(dy) \text{。}$$

將死亡的競爭風險納入分析中的好處之一是透過比較 $\tau_S, \tau_Y, \tau_C$ 的大小可以判斷 sufficient follow-up 的條件是否成立，將問題的本質看得更為清楚。如果研究者手邊的資料並非像 Chatterjee & Shih 的乳癌資料品質這麼好（樣本大、追蹤時間夠久），我們的方法亦可以清楚的知道問題出在哪裏。

我們提出的推論方法利用 EM 演算法以簡化概似函數的估計問題，Chatterjee & Shih 則是假設 pair 間的可交換性以減低參數個數，再直接求概似函數的極值。我們的方法較不受參數個數的影響，但是 EM 以疊代方式求解，欲推導所得之估計量變異數可能會遇到困難。

在 Chatterjee & Shih 的模擬中樣本數令為 5000，猜測可能是在小樣本無法得到理想結果，這個情形和我們的實驗是一致的。我們同時考慮母數與無母數兩種方法，是發現當免疫者存在的情形下，藉著無母數估計量的尾端來估計 $p$，進而以 $\frac{\hat{\Pr}(X^* > t) - (1-\hat{p})}{\hat{p}}$ 求 $\Pr(X > t \mid \tilde{D} = 1)$ 的估計量，不但深深受限於“充分追蹤時間”的假設，而且 $\hat{p}$ 置於分母亦帶來不穩定的表現，因此我們建議以母數的方法估計邊際分配。至於由無母數 Kaplan-Meier 所得之曲線可以將其和母數方法所得之 $\Pr(X^* > t)$ 估計量予以比較，做為檢驗模式假設合理性的方法。

後續的研究希望整理論文所提出之方法的大樣本性質，並透過模擬程式和 Chatterjee & Shih 的方法比較。我們已獲得 Johns Hopkins 大學有關 dementia 家族資料的使用權，希望透過分析實際資料對模式和分析方法有更清楚的了解。此外發展方法以處理競爭風險非獨立的情形，以及發展考慮解釋變數的迴歸模式亦是未來的具體目標。

# 參考文獻


Bishop, J.E. and Waldholz, M., 【基因聖戰--擺脫遺傳的命運】楊玉齡 譯，天下文化出版。

R. A. Weinberg, 【追獵癌症—癌症病因研究之路】許英昌、陳雅茜譯，天下文化出版。

Chatterjee, N. & Shih, J. (2001). "A Bivariate Cure-Mixture Approach for Modeling Familial Association in Disease". *Biometrics*, **57**, 779-786.

Li, C.S and Taylor, J.M.G. (2001). "Identifiability of cure models". *Judy P. Sy. Stat. & Prob. Letters*, **54**, 389-395.

Dabrowska,D.M. (1988). "Kaplan-Meier estimate on the plane". *Ann. Statist*., **16**, 1475-1489.

Eston et al. (1997). "Cancer Risk in Two Large Breast Cancer Families Linked to BRCA2 on Chromosome 13q12-13". *Am. J. Hum. Genet*., **61**, 120-128.

Farewell, V. T. (1982). "The use of mixture models for the analysis of survival data with long-term survivors". *Biometrics*, **38**, 1041-1046.

Frank. M. J. (1979). "On the simultaneous associativity of $F(x, y)$ and $x + y - F(x, y)$". *Aequationes Mathematicae*, **19**, 194-226.

Fort et al. (1998). "Genetic Heterogeneity and Penetrance Analysis of the BRCA1 and BRCA2 Genes in Breast Cancer Families." *Am. J. Hum. Genet*., **62**, 676-689.

Genest, C., Ghoudi, K. & Rivest, L.-P. (1995). "A semiparametric estimation procedure of dependence parameters in multivariate families of distribution". *Biometrika*, **82**, 543-552.

Hougaard, P. (1986). "A class of multivariate failure time distributions". *Biometrika*, **73**, 671-678.

Hsu, L. and Zhao, L. P. (1996). "Assessing Familial Aggregation

of Age at Onset, By Using Estimating Equation, with Application to Breast Cancer". *Am. J. Hum. Genet*., **58**, 1057-1071.

Jeffery et al. (1997)."The Risk Of Cancer Associated With Specific Mutations Of BRCA1 and BRCA2 Among Ashkenazi Jews". *The New England Journal od Medicine*, **336**, 1401-1408.

Kuk, A. Y. C. and Chen, C. (1992). " A mixture model combining logistic regression with proportional hazards regressions". *Biometrika*, **79**, 531-541.

Larson, M.G. and Dinse, G.E. (1985). "A Mixture Model for the Regression Analysis of Competing Risks Data". *Appl. Statistist.*, **34**, 201-211.

Li, C.-S., Taylor, J. M. G. and Sy, J. P. (2001). " Identifiability of cure models". *Statistics & Probability Letters*, **54**, 389-395.

Lin, D. Y. & Ying, Z.(1993). "A simple nonparametric estimator of the bivariate survival function under univariate censoring". *Biometrika*, **80**, 573-581.

Peng, Y. and Dear, K. B. G. (2000). "A nonparametric mixture model for cure rate estimation". *Biometrics*, **56**, 237-243.

Prentice, R. L. & Cai, J. (1992). "Covariance and survival function estimation estimation using censoried multivariate failure time data". *Biometrika*, **79**, 495-512.

Shih, J. H. & Louis, T. A. (1995). "Inference on the association parameter in copula models for bivariate survival data". *Biometrics*, **51**, 1584-1399.

Sy, J. P. and Taylor, J. M. G. (2000). "Estimation in a Cox Proportional Hazards cure model". *Biometrics*, **56**, 227-236.

lversen et al (2000)."Genetic Susceptibility and Survival:

Application to Breast Cancer". *Journal of American Statistical Association*, **95**, 28-42.

Taylor, J. M. G. (1995)."Semiparametric Estimation in Failure Time Mixture Models". *Biometrics*, **51**,899-907.

Wang, W. & Wells, M. T. (1997). "Nonparametric estimators of the bivariate survival function under simplified censoring conditions". *Biometrika*, **84**,863-880.

Wang, W. & Ding, A. A. (2000). "On assessing the association for bivariate current status data". *Biometrika*, **87**, 879-893.

Wang, W. (2003). "Nonparametric Estimation of the Sojourn Time Distributions for a Multi-Path Model". *Journal of the Royal Statistical Society, Series B*. **65**, 921-936.

# 附錄

附錄一：

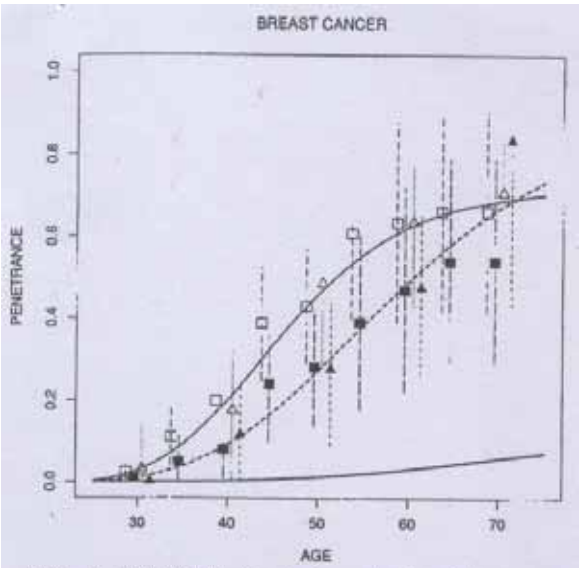



*Figure 1. Estimated Penetrance Curves for Female Breast Cancer Among Noncarriers (Solid Line), Carriers of a Mutation at BRCA1 (Dashed Line) and Carriers of a Mutation at BRCA2 (Dotted Line). Ford (triangular plotting symbols) and Struewing incidence data are included (square plotting symbols) for both BRCA1 (empty symbol) and BRCA2 (solid symbol). Plotting symbols are offset from the associated age by a small amount to the right for the Ford data and to the left for the Stuewing data. Confidence intervals are plotted as vertical lines through the point estimates.*

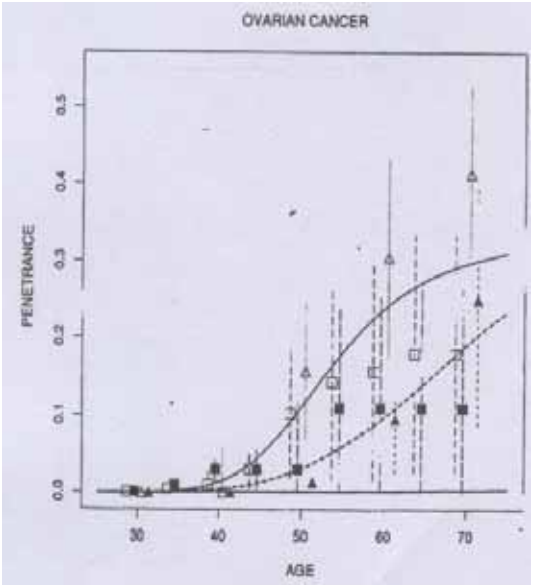



*Figure 2. Estimated Penetrance Curves for Ovarian Cancer. Plotting symbols and line types have the same interpretation as those in Figure 1.*

圖形截自論文“Genetic Susceptibility and Survival: Application to Breast Cancer”.lversen et al. JASS, March 2000.