第四章 分析結果與討論

JSWT$^+$變數選取流程為一個可以同時做係數壓縮與變數選取的方法，在第二章我們回顧了一些可以同時做係數壓縮與變數選取的方法，如 LASSO、Elastic Net 等，Zou 與 Hastie 在 2004 年提出 Elastic Net 並建立模擬與 LASSO 比較下，當解釋變數間具有高度相關時，LASSO 表現較差。由於我們引入線性迴歸，假設在資料中解釋變數間滿秩，所以本章將以 LASSO 與 JSWT$^+$去比較。其中第一、二節利用模擬比較，第三節比較 JSWT$^+$與 LASSO 估計量中參數選定差異，最後第四節以攝護腺癌資料做實證。

第一節 模擬分析(設定真實係數三個非零)

本節我們建立模擬研究。考慮以下線性模型：

$Y = X\beta + \sigma\varepsilon$, $\varepsilon \sim N_n(0, \sigma^2 I_n)$, $corr(x_i, x_j) = \rho^{|i-j|}$, $X \sim N_8(0\ ,\ corr.matrix)$ ，

設定 ρ=0.5，σ=3，八個解釋變數其真實係數數值為 $\beta = (3, 1.5, 0, 0, 2, 0, 0, 0)$，樣本數為 100，去生成模擬資料。

首先，我們將依照上述設定生成一組模擬資料，將此資料依照 JSWT$^+$變數選取流程，畫出不同 JSWT$^+$的參數 b 改變時，每個解釋變數其係數路徑圖及 *SURE* 函數圖，比較 JSWT$^+$變數選取流程(ii)中不同轉換矩陣之差異(圖 2)。在圖 2 的上列，我們可發現利用 JSWT$^+$變數選取流程(ii)之(1)狀況的轉換矩陣的結果，*SURE* 函數較平滑且每個係數壓縮比例很微小，但觀察其係數路徑圖 X_1、X_2、X_5 對應其係數與 0 的絕對距離最遠；在圖 2 的下列，在 JSWT$^+$變數選取流程(ii)之(2)狀況的轉換矩陣的結果，*SURE* 函數較不平滑且每個係數隨著參數 b 改變而逐漸壓縮，明顯看到每個係數其壓縮之效果，且有些係數路徑壓縮到零。觀察其變數選取結果，JSWT$^+$變數選取流程(ii)之(1)狀況，當 $b = 1.361$ 時，使得 *SURE* 函數有

最小值，挑選出全部 8 個變數與模擬設定不符；JSWT^{+}變數選取流程(ii)之(2)狀況，當 $b=0.697$ 時，使得 *SURE* 函數有最小值挑選出三個變數為 X_1、X_2、X_5，符合我們之前所設定的真實係數數值在第一、第二和第五位置非零(詳見表 1)。由於在 JSWT^{+}變數選取流程(ii)之(1)狀況中，每個係數壓縮比例很微小且與真實係數數值不符，所以之後我們僅考慮 JSWT^{+}變數選取流程(ii)之(2)狀況來當做轉換矩陣時的主要步驟。

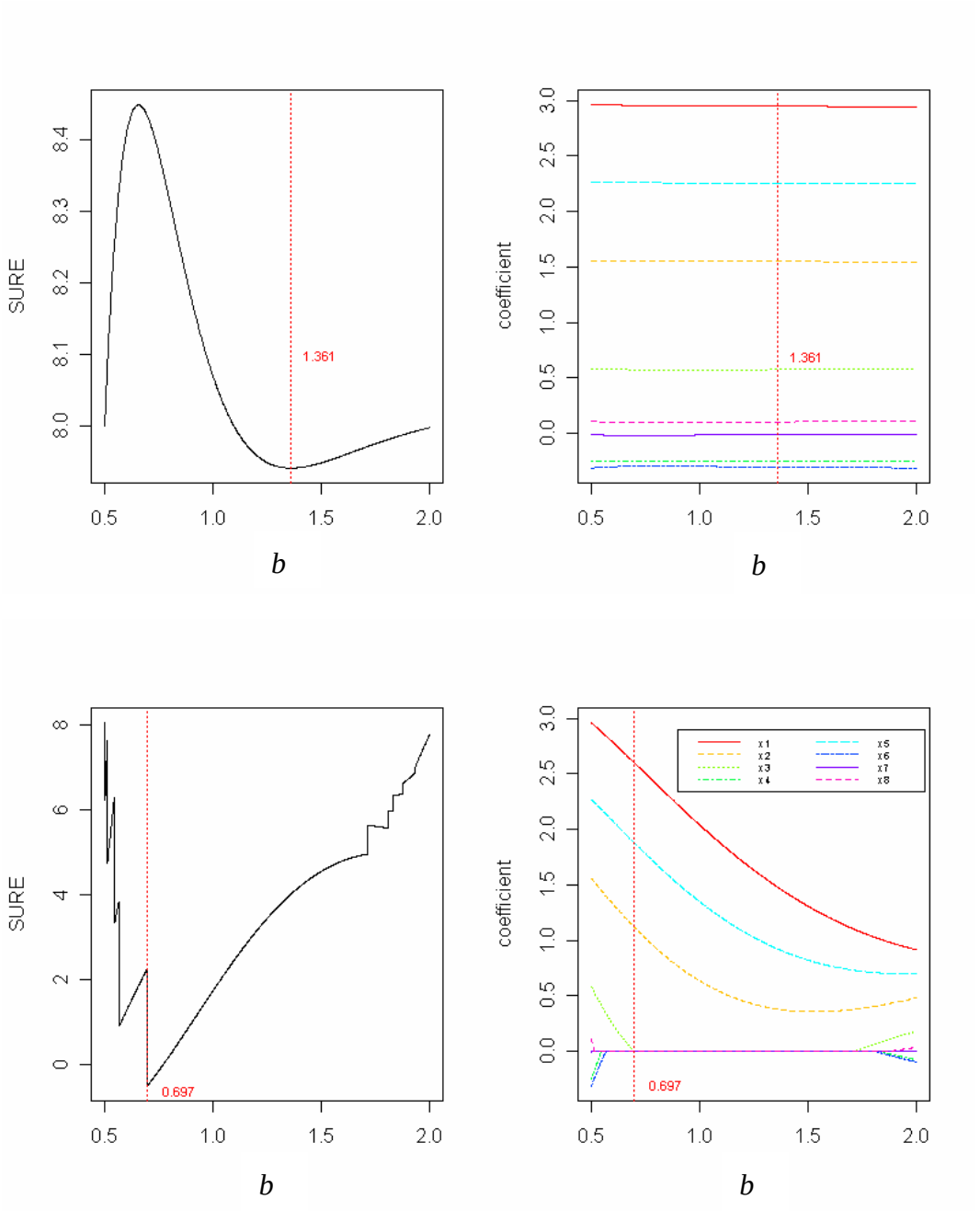

圖 2：JSWT^{+}在不同轉換矩陣之 *SURE* 函數及係數路徑圖(上列為(ii)之(1)；下列為(ii)之(2))

表 1：JSWT$^+$在不同轉換矩陣挑選出的係數數值比較表(1 組模擬資料)

轉換矩陣	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
利用(ii)之(1)	2.95	1.55	0.575	-0.265	2.253	-0.305	-0.01	0.104
利用(ii)之(2)	2.604	1.121	0	0	1.881	0	0	0

在分析轉換矩陣不同造成係數路徑之差異後，我們依照上述設定再模擬一組資料去比較 JSWT$^+$與 LASSO 係數路徑圖之變化，在圖 3 中，JSWT$^+$估計量當 $b=0.772$ 時，挑選到 X_1 、X_2 、X_5 三個解釋變數；在圖 4 中，LASSO 估計量當其 1-norm 的限制式值 t 與 1-norm 最小平方估計總和的比值(bounds)為 0.697 時，挑選到 X_1 、X_2 、X_3 、X_5 四個變數。同時比較 X_1 、X_2 、X_5 三個解釋變數，JSWT$^+$估計量比 LASSO 估計量得到較小的係數值。挑選出係數數值見表 2。

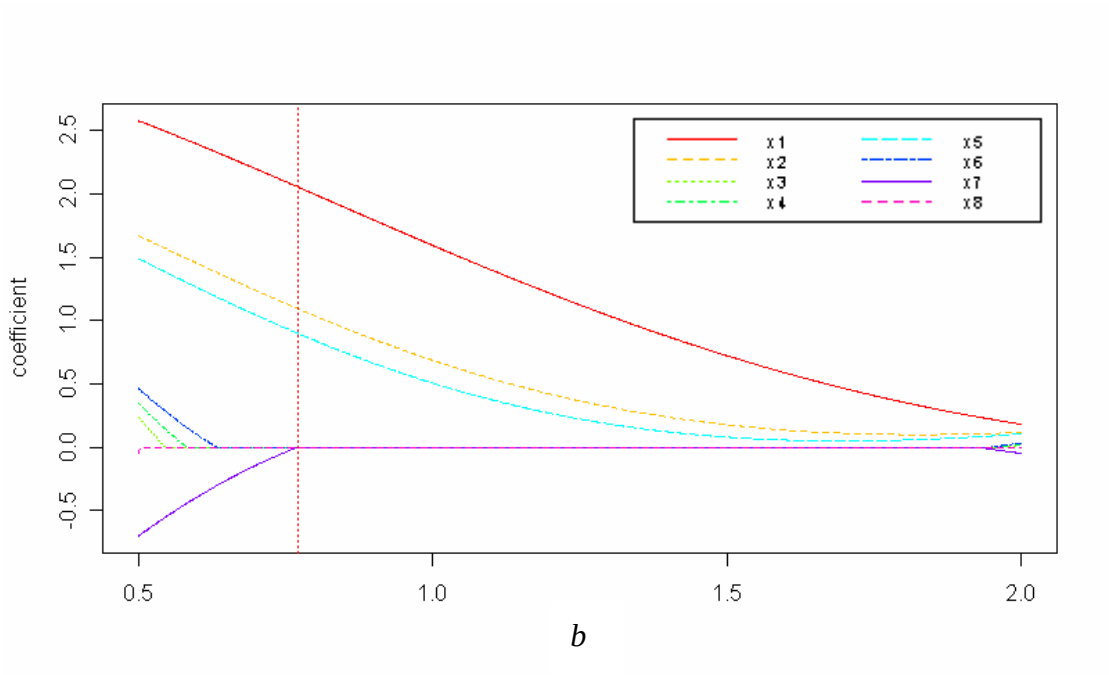

圖 3：一組模擬資料由 JSWT$^+$做變數選取之係數路徑圖(第一、二、五位置非零)

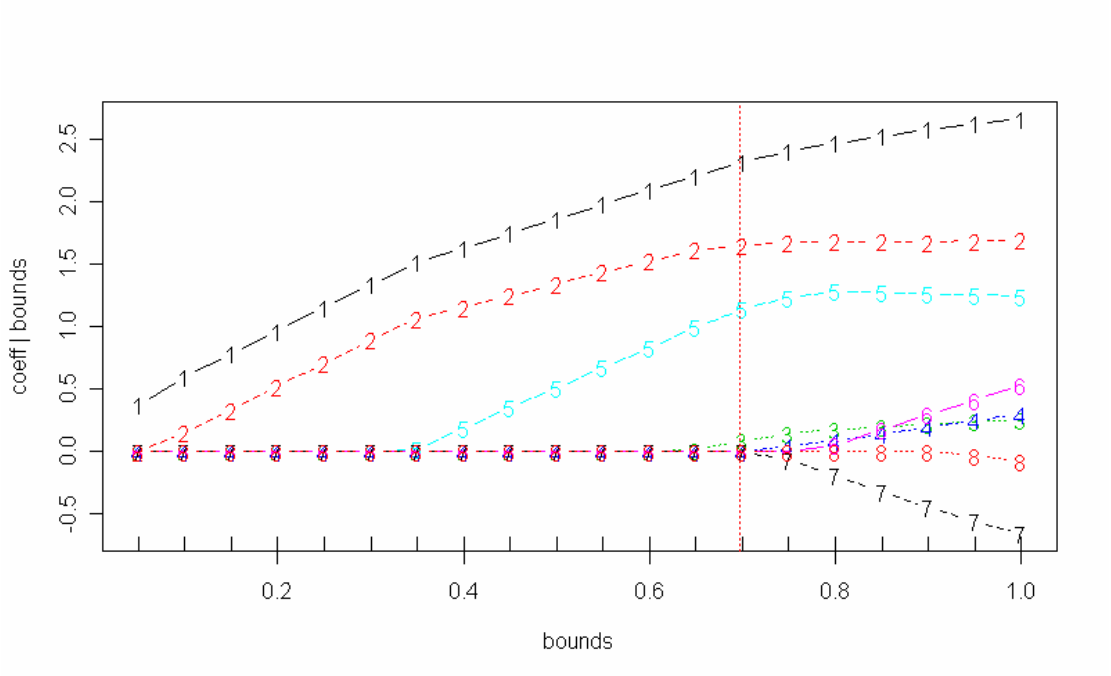

圖 4：一組模擬資料由 LASSO 做變數選取之係數路徑圖(第一、二、五位置非零)

表 2：JSWT$^+$與 LASSO 估計量係數比較表(1 組模擬資料，第一、二、五位置非零)

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
JSWT$^+$	2.053	1.091	0	0	0.896	0	0	0
LASSO	2.321	1.577	0.088	0	1.282	0	0	0

接下來，我們模擬 2500 組資料，比較 JSWT$^+$與 LASSO 變數被選取個數的分布。見表 3，在 JSWT$^+$中，2500 組被挑選出三個解釋變數的組數是最多的，總共有 1919 組，其次為解釋變數被挑選個數為 2 個及 4 個，且至少會挑到一個；在 LASSO 中，2500 組被挑選出五個解釋變數的組數是最多的，總共有 582 組，其次為 4 個及 6 個，至少會挑到三個。計算其 2500 次平均被挑選 JSWT$^+$為 3.086 個，而 LASSO 為 5.362 個，JSWT$^+$較符合我們期待挑選三個變數。

表 3：2500 組模擬資料中被挑選出解釋變數個數對應之組數表

解釋變數被選取個數(n)	2500 組被挑選出 n 的組數	
	JSWT$^+$	LASSO
8	0	291
7	1	356
6	16	452
5	67	582
4	273	507
3	**1919**	**312**
2	205	0
1	19	0
0	0	0
平均被挑選個數	3.086	5.362

我們試著比較每個解釋變數在 2500 組模擬資料被挑選的次數與比例，見表 5 與表 6，在 JSWT$^+$中，X_1 、X_2 、X_5 三個解釋變數被 JSWT$^+$挑選的次數皆大於 2000 次，遠大於其他解釋變數被挑選的次數(60~110 次)；在 LASSO 中，X_1 、X_2 、X_5 三個解釋變數被 LASSO 挑選的次數皆為 2500 次，但其他解釋變數被挑選次數卻超過了 1000 次。由挑選次數比例來看，模擬設定不為零的解釋變數被挑選比例兩方法皆接近於 100%，但在 JSWT$^+$中，模擬設定為零的解釋變數被挑選比例皆小於 5%，LASSO 在模擬設定為零的解釋變數被挑選比例皆大於 40%。詳見表 5 與表 6。而在 2500 次模擬資料中，正確挑選出 X_1 、X_2 、X_5 三個解釋變數次數，JSWT$^+$為 1905 次，佔 2500 組的 76%，而 LASSO 只有 312 次，佔 2500 組的 12.5%。(見表 4)

表 4：2500 組挑選正確變數(3 個)之組數與比例比較表

	2500 組挑選正確變數之組數	2500 組挑選正確變數之比例
JSWT$^+$	1905	76%
LASSO	312	12.5%

表 5：每個解釋變數在 2500 組模擬資料被挑選的個數比較表

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
JSWT$^+$	2500	2274	109	92	2469	106	104	62
LASSO	2500	2500	1238	1216	2500	1183	1079	1190

表 6：每個解釋變數在 2500 組模擬資料被挑選的比例比較表

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
JSWT$^+$	1	0.91	0.044	0.037	0.988	0.042	0.042	0.025
LASSO	1	1	0.495	0.486	1	0.473	0.432	0.476

表 7 與表 8 顯示在 2500 組模擬資料每個解釋變數其係數值的平均與標準差，而 X_1 、X_2 、X_5 三個解釋變數其係數平均值，JSWT$^+$比 LASSO 更小，標準差則 JSWT$^+$比 LASSO 更大。除了 X_1 、X_2 、X_5 三個解釋變數外，JSWT$^+$其係數平均值的絕對值皆小於 0.008 且係數標準差皆小於 0.14；LASSO 其係數平均值的絕對值介於 0.06 與 0.006 之間且係數標準差皆大於 0.21。

表 7：JSWT$^+$在 2500 組模擬資料每個解釋變數其係數值的平均值與標準差表

JSWT$^+$	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
平均值	2.629	1.078	-0.007	-0.002	1.605	0.001	0.002	-0.002
標準差	0.412	0.523	0.129	0.129	0.51	0.135	0.128	0.099

表 8：LASSO 在 2500 組模擬資料每個解釋變數其係數值的平均值與標準差表

LASSO	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
平均	2.844	1.342	0.048	0.06	1.73	0.043	0.006	0.006
標準差	0.363	0.385	0.245	0.244	0.376	0.252	0.218	0.213

見圖 5 與圖 6，在 2500 組模擬資料每個解釋變數其係數值的盒鬚圖，JSWT$^+$

和 LASSO 兩者在 X_1 、X_2 、X_5 三個解釋變數其係數值分布較分散，其他係數都在 0 附近。

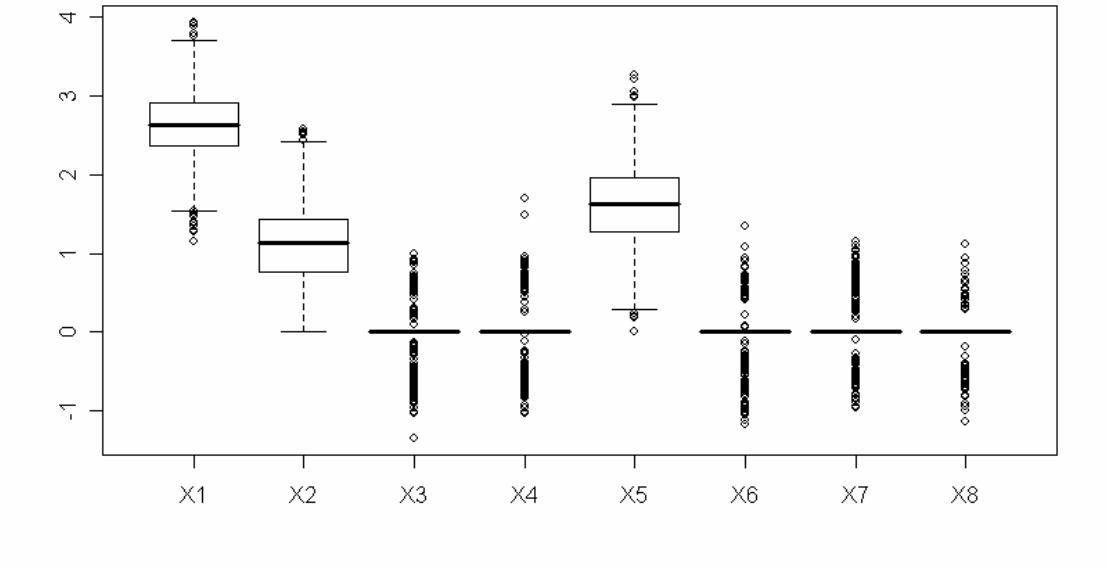

圖 5：利用 JSWT^{+}做變數挑選對應其係數值之盒鬚圖(2500 組模擬資料)

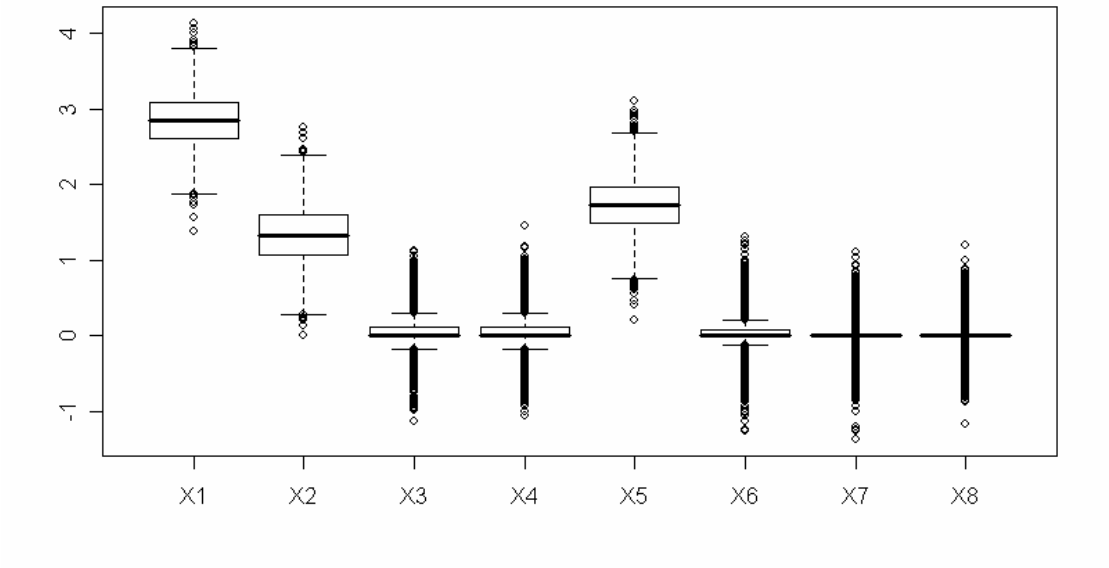

圖 6：利用 LASSO 做變數挑選對應其係數值之盒鬚圖(2500 組模擬資料)

第二節 模擬分析(設定真實係數僅一個非零)

本節我們依照第一節一開始所提的模擬假設，設定 ρ=0.5，σ=3，除了真實係數數值改為僅有第一位置非零與第一節設定不同外，其他設定皆不變。其真實係數數值為 $\beta=(5,0,0,0,0,0,0,0)$ ，樣本數為 100，去生成模擬資料。

我們依照上述設定模擬一組資料去比較 JSWT^{+}與 LASSO 係數路徑圖之變化，見圖 7，JSWT^{+}估計量當 $b=0.568$ 時，挑選到 X_1 解釋變數；見圖 8，LASSO 估計量當其 1-norm 的限制式其門檻值 t 與 1-norm 最小平方估計總和的比值 (bounds)為 0.747 時，也挑選到 X_1 解釋變數。見表 9，同時比較 X_1 其係數值，JSWT^{+}估計量比 LASSO 估計量得到較大的係數值，與第一節的結果不同。

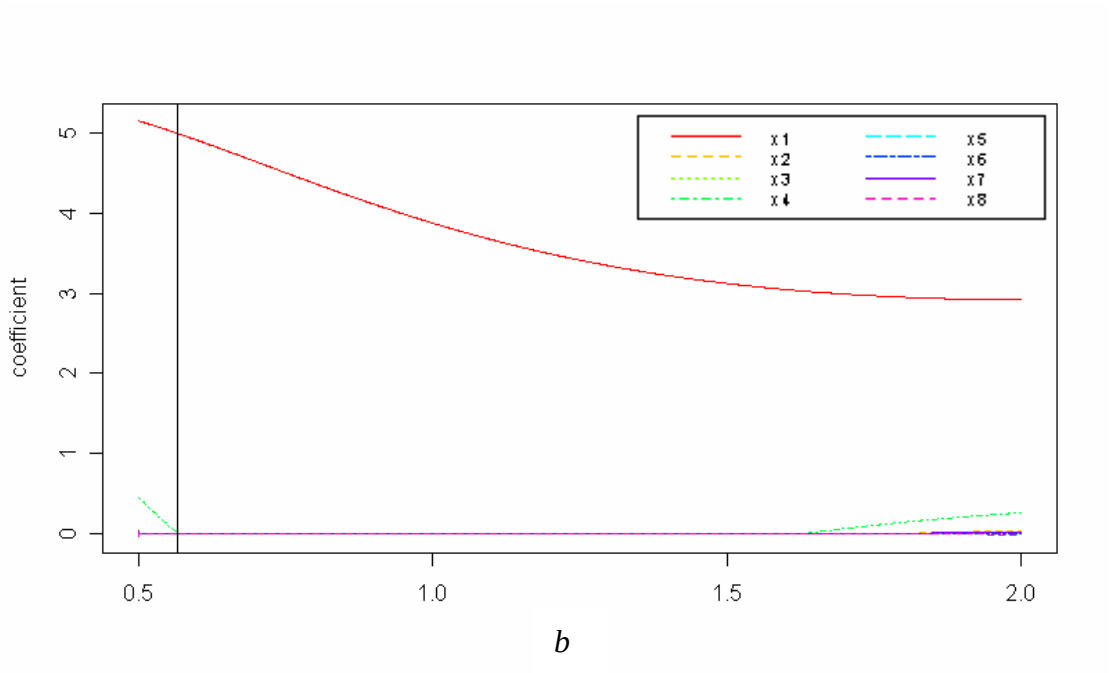

圖 7：一組模擬資料由 JSWT^{+}做變數選取之係數路徑圖(第一位置非零)

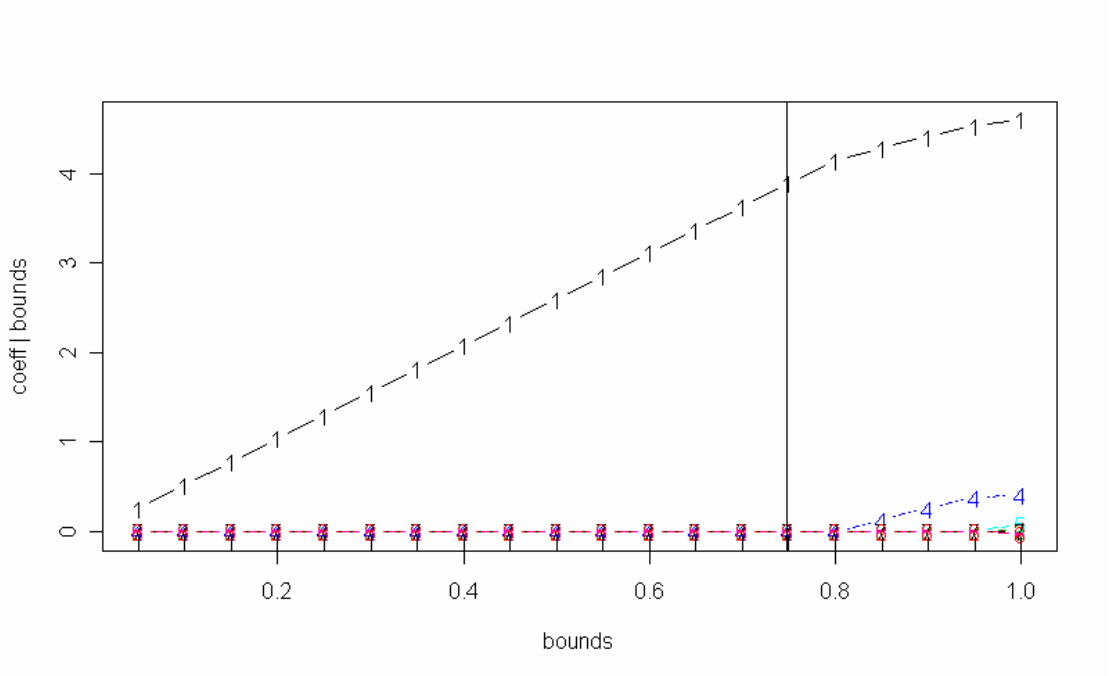

圖 8：一組模擬資料由 LASSO 做變數選取之係數路徑圖(第一位置非零)

表 9：JSWT$^+$與 LASSO 估計量係數比較表(1 組模擬資料，第一位置非零)

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
JSWT$^+$	4.997	0	0	0	0	0	0	0
LASSO	4.338	0	0	0	0	0	0	0

接下來，我們模擬 2500 組資料，比較 JSWT$^+$與 LASSO 變數被選取個數的分布。見表 10，在 JSWT$^+$中，2500 組被挑選出一個解釋變數的組數是最多的，總共有 2209 組，其次為解釋變數被挑選個數為兩個及三個，至少會挑到一個且最多挑選到六個的僅有 1 組；在 LASSO 中，2500 組被挑選出一個解釋變數的組數也是最多的，總共有 660 組，至少會挑到一個，然而在此模擬設定狀況下，挑選到六個的有 137 組。計算其 2500 組平均被挑選個數 JSWT$^+$為 1.154 個，接近於挑選到一個解釋變數，；而 LASSO 為 2.968 個，接近於挑選到三個解釋變數，所以由此分析 JSWT$^+$較符合我們期待挑選一個解釋變數。

表 10：2500 組模擬資料中被挑選出解釋變數個數對應之組數表(第一位置非零)

解釋變數被選取個數(n)	2500 組被挑選出 n 的組數	
	JSWT$^+$	LASSO
8	0	54
7	0	89
6	1	137
5	4	219
4	8	310
3	62	485
2	216	546
1	**2209**	**660**
0	0	0
平均被挑選個數	1.154	2.968

而解釋變數在 2500 組模擬資料被挑選的次數與比例，見表 12，在 JSWT$^+$ 中，X_1 解釋變數被 JSWT$^+$挑選的次數即 2500 次，遠大於其他七個解釋變數被挑選的次數(30~75 次)；在 LASSO 中，X_1 解釋變數被 LASSO 挑選的次數也為 2500 次，但其他解釋七個解釋變數被挑選次數皆超過了 640 次。見表 13，由挑選次數比例來看，模擬設定不為零的解釋變數被挑選比例兩方法皆為 100%，但在 JSWT$^+$中，模擬設定為零的解釋變數被 JSWT$^+$挑選比例皆小於 3%，LASSO 在模擬設定為零的解釋變數被挑選比例皆大於 25%。而在 2500 次模擬資料中(見表 11)，正確挑選出 X_1 解釋變數次數，JSWT$^+$為 2209 次，即 2500 次模擬資料中只要挑選到一個解釋變數就是挑中 X_1 解釋變數，佔 2500 組的 88.4%，而 LASSO 只有 660 次，與 JSWT$^+$相同只要挑選到一個解釋變數也是必挑到 X_1 解釋變數，佔 2500 組的 26.4%。

表 11：2500 組挑選正確變數(1 個)之組數與比例比較表

	2500 組挑選正確變數之組數	2500 組挑選正確變數之比例
JSWT$^+$	2209	88.4%
LASSO	660	26.4%

表 12：每個解釋變數在 2500 組模擬資料被挑選的個數比較表(第一位置非零)

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
JSWT$^+$	2500	73	58	47	63	49	62	33
LASSO	2500	796	742	650	699	646	662	724

表 13：每個解釋變數在 2500 組模擬資料被挑選的比例比較表(第一位置非零)

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
JSWT$^+$	1	0.029	0.023	0.019	0.025	0.02	0.025	0.013
LASSO	1	0.308	0.297	0.26	0.28	0.258	0.265	0.29

表 14 與表 15 分別顯示在 2500 組模擬資料 JSWT$^+$與 LASSO 每個解釋變數其係數值的平均值與標準差，而 X_1 解釋變數其係數平均值，LASSO 比 JSWT$^+$更小，此與第一節結果相反，而係數標準差 JSWT$^+$依舊比 LASSO 更大。除了 X_1 解釋變數外，JSWT$^+$其係數平均值的絕對值介於 0 與 0.003 之間且係數標準差皆小於 0.11；LASSO 其係數平均值的絕對值介於 0.001 與 0.046 之間且係數標準差皆大於 0.15。

表 14：JSWT$^+$在 2500 組模擬資料每個解釋變數其係數值的平均值與標準差表(第一位置非零)

JSWT$^+$	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
平均值	4.639	-0.001	0	-0.003	0.003	-0.002	-0.002	0.002
標準差	0.435	0.107	0.085	0.069	0.089	0.074	0.088	0.061

表 15:LASSO 在 2500 組模擬資料每個解釋變數其係數值的平均值與標準差表(第一位置非零)

LASSO	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
平均值	4.592	0.046	0.01	0.004	0.004	-0.007	0.001	0.001
標準差	0.368	0.169	0.173	0.158	0.165	0.167	0.156	0.154

見圖 9 與圖 10，在 2500 組模擬資料每個解釋變數其係數值的盒鬚圖，JSWT^{+} 和 LASSO 兩者在 X_1 解釋變數其係數值分布較分散，其他係數值都在 0 附近。

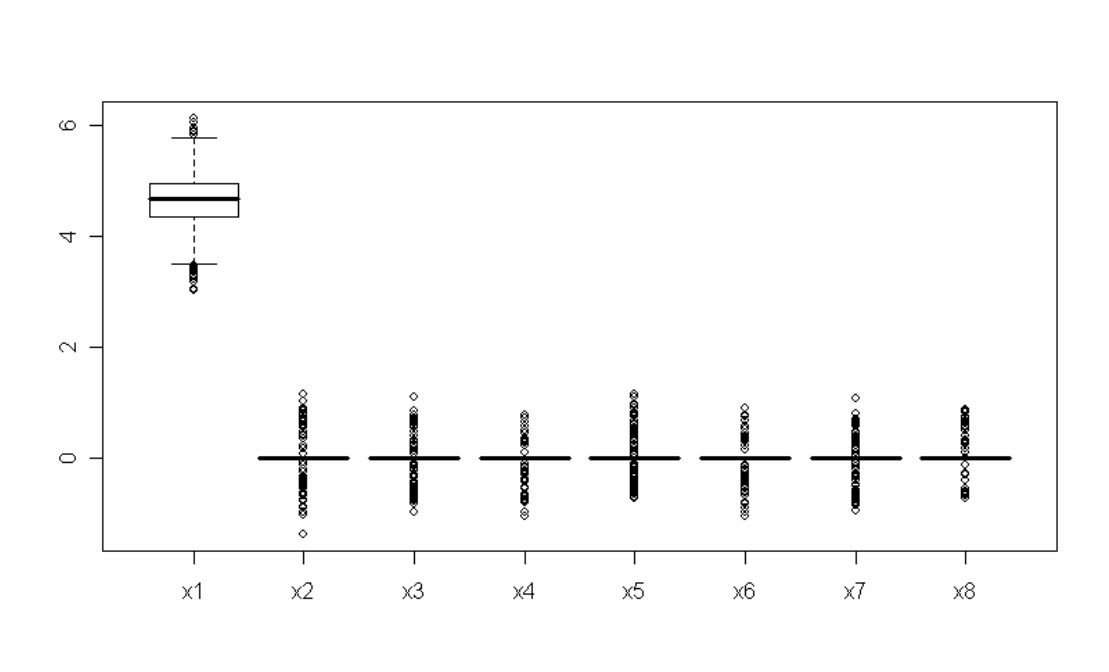

圖 9：利用 JSWT^{+}做變數挑選對應其係數值之盒鬚圖(2500 組模擬資料，第一位置非零)

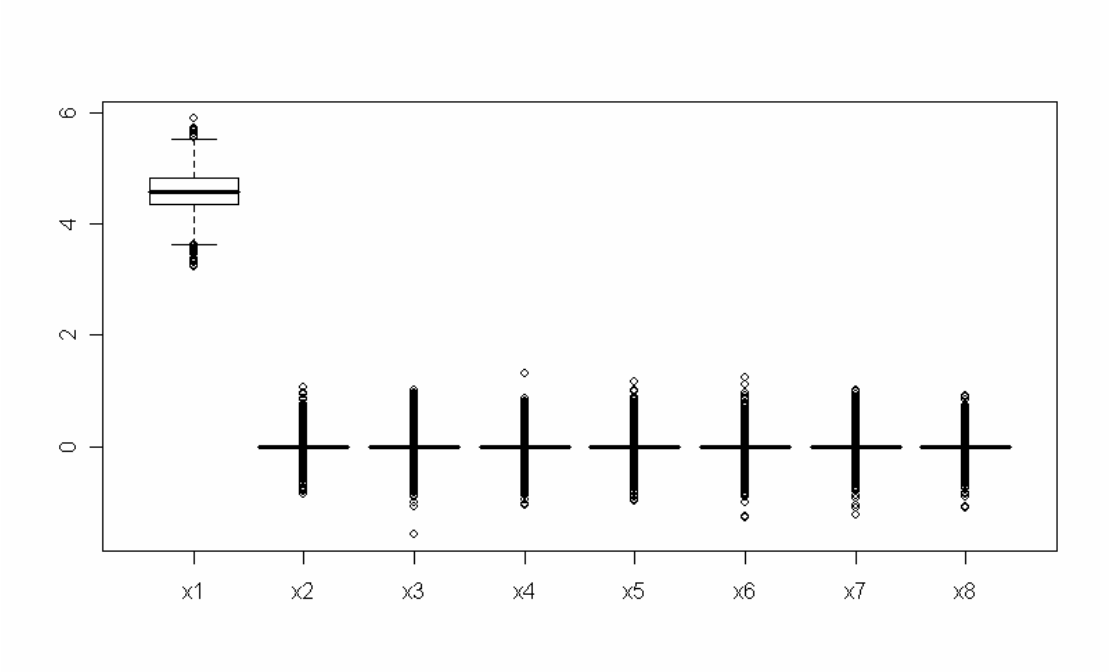

圖 10：利用 LASSO 做變數挑選對應其係數值之盒鬚圖(2500 組模擬資料，第一位置非零)

第三節 JSWT^{+}與 LASSO 參數挑選

JSWT^{+}與 LASSO 兩者都是可同時做到係數壓縮及變數選取的方法，而 JSWT^{+}估計量其中只有參數 b 需要被估計，經由 *SURE* 函數可以找到唯一參數 b 決定唯一的壓縮係數值，但 LASSO 限制式的門檻值 t 必須人工自訂其數值，或者利用交叉驗證(cross-validation)去找出合適的門檻值 t，而每次交叉驗證所到的 t 亦不相同。現在我們考慮迴歸矩陣為正交集時，JSWT^{+}估計量與 LASSO 估計量的壓縮函數即：

$$Shrink_i^{JSWT^+}(\hat{\beta}_i^{OLS}) = \left(1 - a_b D_b^{-1} \left|\hat{\beta}_i^{OLS}\right|^{b-2}\right)_+ \hat{\beta}_i^{OLS} \text{ ，其中 } D_b = \sum_{i=1}^{d} \left|\hat{\beta}_i^{OLS}\right|^b$$

$$Shrink_i^{LASSO}(\hat{\beta}_i^{OLS}) = sign(\hat{\beta}_i^{OLS}) \times (\left|\hat{\beta}_i^{OLS}\right| - \lambda)_+ \text{ ，其中 } \lambda \text{ 由限制式的門檻值決定}$$

我們可以發現兩者皆為 $\hat{\beta}_i^{OLS}$ 的函數形式，見圖 11，假設 $\hat{\beta}_i^{OLS} \in (-4, 4)$，在 $(-4, 4)$ 生成等距 20 點，利用 JSWT^{+}變數選取流程可以挑選出唯一參數 b 估計為 0.553，代入 JSWT^{+}其壓縮函數可得唯一曲線(綠色)；假設給定合適的門檻值 t 使得 $\lambda = 1, 1.5, 2$，LASSO 其壓縮函數則得到不同曲線(藍色)。

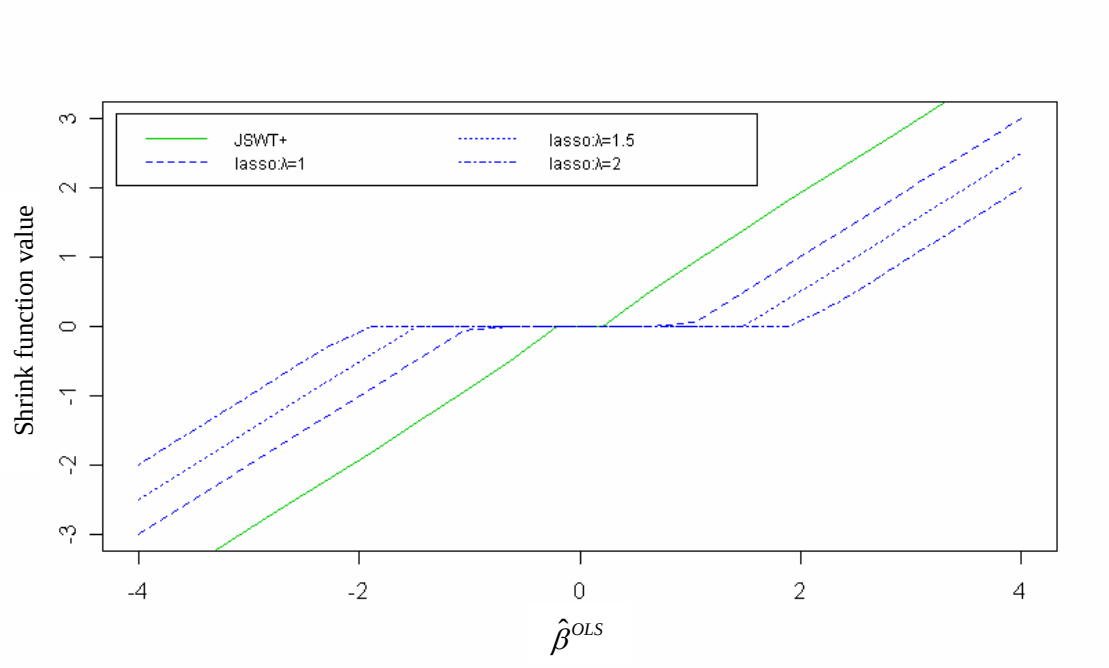

圖 11：JSWT^{+}與 LASSO 其壓縮函數比較圖

第四節 攝護腺癌資料分析

最後，我們利用Tibshirani (1996)的文章中攝護腺癌資料做實證，此資料源自於Stamey *et* al. (1989)，為了檢查攝護腺的特定抗原(prostate specific antigen)與一些臨床測量是否有關係，主要有以下8個解釋變數：log(cancer volume) (X_1)、log(prostate weight) (X_2)、 age (X_3)、log(benign prostatic hyperplasia amount) (X_4)、seminal vesicle invasion (X_5)、log(capsu1arpenetration) (X_6)、Gleason score (X_7) 、percentage Gleason scores 4 or 5(X_8)，其反應變數為log(prostate specific antigen) (Y)，樣本數為97筆。依照我們所提出的JSWT$^+$變數選取流程與LASSO所做的結果做比較，見圖12，結果顯示當$b=0.617$ 時，使得*SURE*函數有最小值，且*SURE*函數值在$b=0.617$之後呈一直線，所有解釋變數其係數被壓縮到零，依照我們所提出JSWT$^+$變數選取流程並沒有挑出解釋變數。

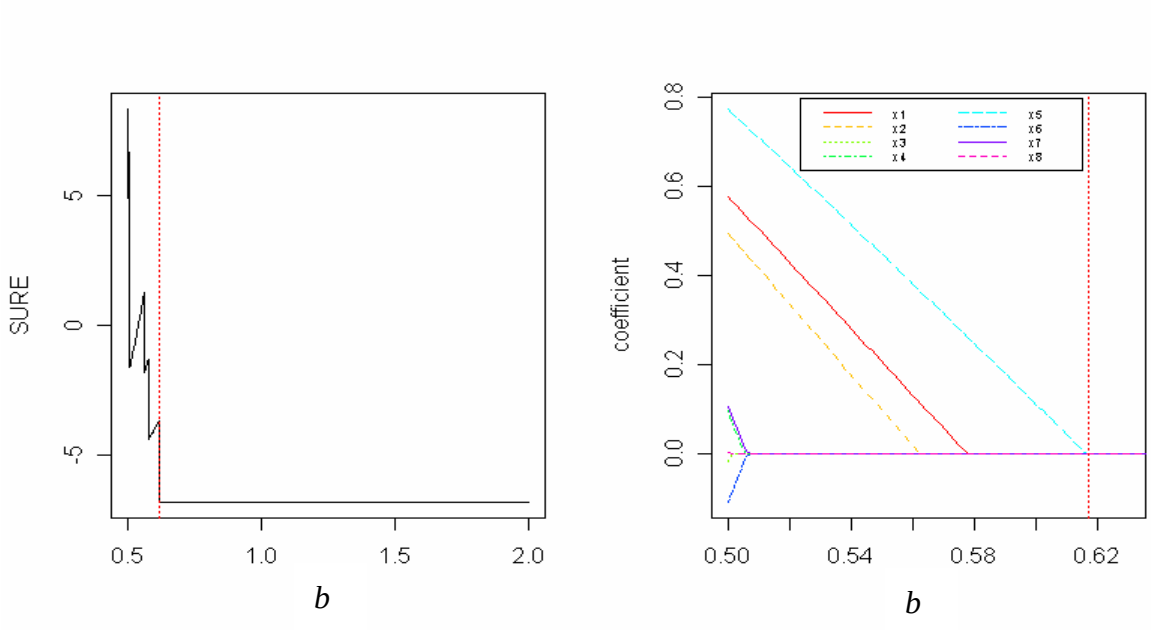

圖 12：攝護腺癌資料由 JSWT$^+$做變數選取之 *SURE* 函數圖與係數路徑圖

圖 13 與圖 14，分別為 JSWT$^+$與 LASSO 在攝護腺癌資料做變數選取時每個係數路徑變化，JSWT$^+$並沒有挑選出變數；LASSO 由交叉驗證選出當 bounds=0.44 時，挑選出 log(cancer volume) (X_1)、log(prostate weight) (X_2)、seminal vesicle

invasion (X_5)三個變數。再去觀察圖 13，JSWT$^+$係數路徑圖中，X_1、X_2、X_5三個變數為最後壓縮到零的變數，如果我們將 b 值從 0.617 向左移動至 0.54，依序會挑選到 X_5、X_1、X_2；而見圖 13，LASSO 係數路徑圖顯示變數被挑選之順序依序為 X_1、X_5、X_2。

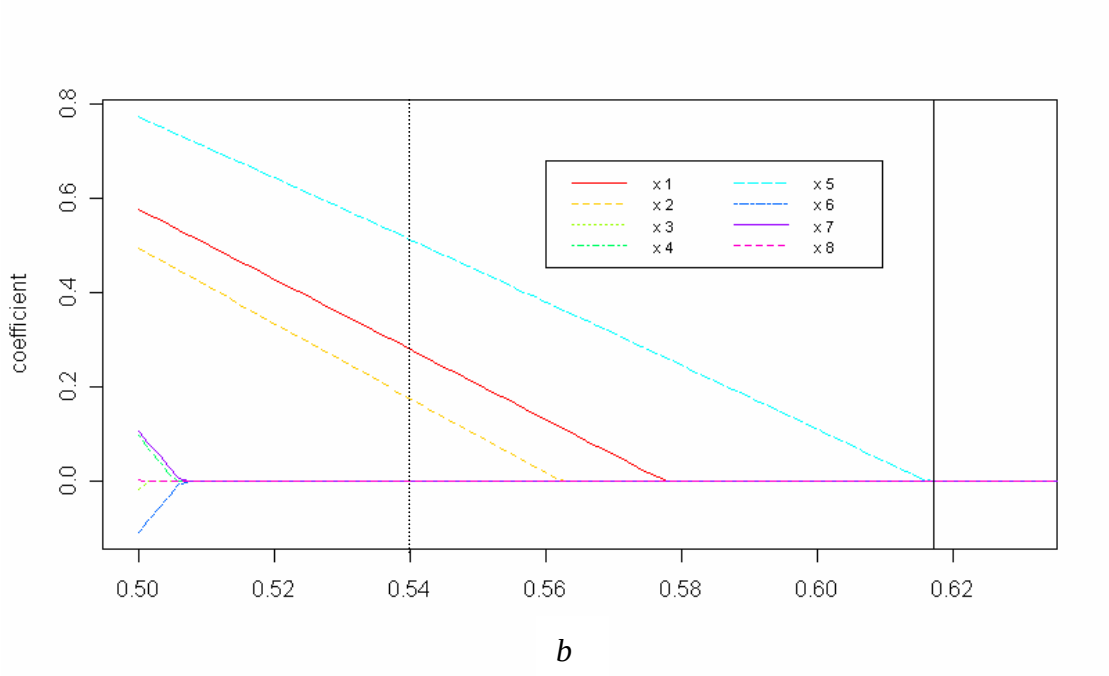

圖 13：攝護腺癌資料由 JSWT$^+$做變數選取之係數路徑圖

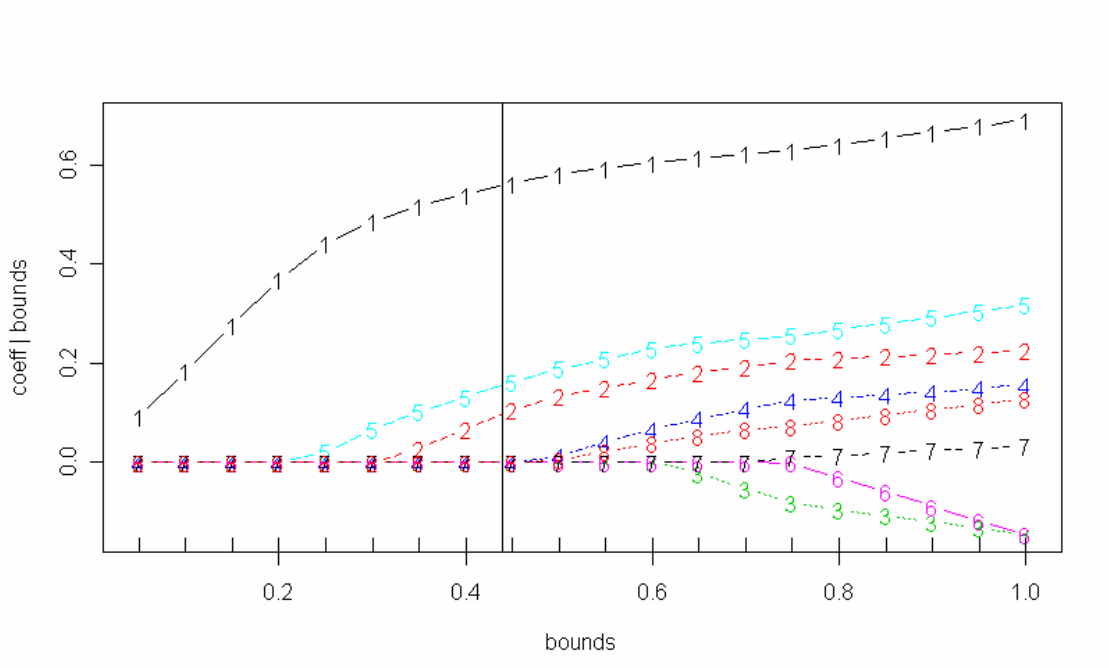

圖 14：攝護腺癌資料由 LASSO 做變數選取之係數路徑圖