2021 年美國職棒打者表現之評估

作者: 黃亮臻

一、 摘要

此作品挑選了 46 個打擊相關的變數，希望以全面而非單一的角度評估打者表現，並找出選手各自的優勢。由於得分為影響比賽勝負的關鍵，我針對打者「每打席的平均得分」建立迴歸模型，將影響得分的重要變數作為評估打者表現的指標。這裡採取向後選取法搭配主成分分析，以及 Sparse Group Lasso 兩種方法做變數選擇與維度縮減，再分別以階層式分群法，評估每一群打者的能力與狀況。

二、 研究動機

評估打者第一個想到的指標可能會是「打擊率」，但其實不然，打擊率不高但會選球、上壘率高或是速度快，甚至是具有長打能力的打者，若使用得當，可能為隊伍做出更多貢獻。憑藉單一指標，很難去評估一個打者的表現與價值，期許在本次研究中，可以以更全面的角度，評估打者的狀況與表現，並為球員做分群，找出不同的優勢，提供球團與球員作參考。

三、 資料說明

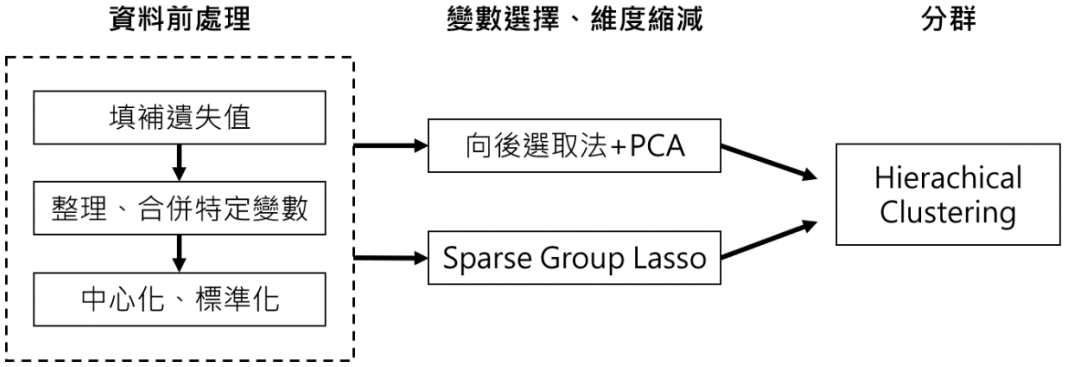
本研究資料來源為 Statcast (Statcast Custom Leaderboards | baseballsavant.com (mlb.com)) 。Statcast 官網中提供了 30 隻隊伍的 732 名打者資料，但僅僅關於打者的資料，就有超過 150 個變數。我針對了「打擊與跑壘表現」、自身對棒球的了解以及預期分析的內容，刪除了一些變數，最後採用 46 個變數，變數說明如下。

變數名稱	內容
Age	年齡
AB	打數
PA	打席
HR	全壘打數
K%	三振率
BB%	四壞率
AVG	打擊率
SLG	長打率

OBP	上壘率
RBI	打點
CS	盜壘失敗次數
SB	盜壘成功次數
GIDP	雙殺打次數
GITP	三殺打次數
HBP	觸身球次數
IBB	故意四壞球保送次數
SacBunt	犧牲觸擊次數
SacFly	犧牲高飛次數
R	得分
Sac	總犧牲數
Walkoff	再見安打數
AvgEV	平均擊球初速度 (MPH)
AvgLA	平均擊球角度 (°)
SweetSpot%	甜蜜點機率
Barrel%	擊球仰角與初速品質 (約仰角 45 度，見附錄一)
SolidContact%	擊球仰角與初速品質 (約仰角 30 度，見附錄一)
Flare/Burner%	擊球仰角與初速品質 (約仰角 15 度，見附錄一)
Under%	擊球仰角與初速品質 (約仰角 -15 度，見附錄一)
Topped%	擊球仰角與初速品質 (約仰角 0 度，見附錄一)
Poor/Weak%	擊球仰角與初速品質 (約仰角 -30 度，見附錄一)
HardHit%	重擊球，定義為以 95 英里/小時
ZoneSwing%	好球帶以內的球的揮擊比例
OutofZoneSwing%	好球帶以外的球的揮擊比例
OutofZoneContact%	揮擊所有好球帶以外的球的打中比例
InZoneContact%	揮擊所有好球帶以內的球的打中比例
Whiff%	揮空率
Swing%	面對來球揮棒的比例
Pull%	拉打率
StraightAway%	中間率
Oppo%	反向率
GB%	滾地比例

FB%	高飛比例
LD%	平飛比例
Popup%	內野高飛比例
HPto1B	本壘到一壘秒數
SprintSpeed	壘間全力衝刺時的平均速度 :英尺/秒

四、 方法流程圖



五、 資料前處裡

(一) 填補遺失值

HPto1B (本壘到一壘秒數)、SprintSpeed (壘間全力衝刺時的平均速度) 有許多的遺失值，以及第 230 名打者 Adrian Sampson 的 AvgEV (平均擊球初速度)、AvgLA (平均擊球角度) 變數各有一個遺失值，分別做以下處理:

1. 本壘到一壘秒數、壘間全力衝刺時的平均速度，此兩個變數與打者跑步速度有關，且兩變數成反比，我將這兩欄的資料相乘後取平均。若某位打者缺少其中一個變數的資料，則使用所有球員相乘平均的值除以已知的變數資料，得到的結果用來填補遺失值。若兩個變數的值都遺失，則取此變數的平均值填補。
2. 查看後發現第 230 名打者 Adrian Sampson 皆使用短打，因此測不到擊球初速與角度，於是我自行補上短打對應的合理的值。

(二) 整理、合併特定變數

因部分變數單獨看提供的資訊量過少，因此將兩者合併，如 GIDP (雙殺打)、GITP (三殺打) 合併為 DPTP% (雙殺、三殺打率) ; CS (盜壘失敗)、SB (盜壘成功)

合併為 CSSB% (盜壘率)。

此外，因打者的上場次數會影響各指標的數量，例如，打擊次數多，安打數也會相對較多。僅從數量無法評估打者的表現，因此將數量的變數換成比率的型態，分別做以下處理。另外，考量到上場次數與球員能力有關聯，又打席與打數高度相關，因此僅保留打席變數。

原變數	處理後變數	內容	經過處理
HR	HR%	全壘打率	全壘打 HR / 打數 AB
RBI	RBI/PA	平均每打席打點	打點 RBI/ 打席 PA
CS、SB	CSSB%	盜壘率 (上壘時的盜壘機率)	(盜壘成功 SB+盜壘失敗 CS) / (上壘率 OBP* 打席 PA)
SB	SB%	盜壘成功率 (上壘時的盜壘成功機率)	(盜壘成功 SB) / (上壘率 OBP*打席 PA)
GIDP 、GITP	DPTP%	雙殺、三殺打率	(雙殺打 GIDP+三殺打 GITP) / 打數 AB
HBP	HBP%	觸身球率	觸身球 HBP / 打席 PA
IBB	IBB%	故意四壞球保送率	故意四壞球保送 IBB / 四壞球數 BB
SacBunt	SacBunt%	犧牲打中，觸擊的機率	觸擊數 SacBunt / 總犧牲數
SacFly	SacFly%	犧牲打中，高飛的機率	高飛球數 SacFly / 總犧牲數 Sac
R	R/PA	每打席會跑回幾分	得分 R / 打席 PA
Sac	Sac%	總犧牲率	總犧牲 Sac / 打席 PA

(三) 資料中心化與標準化

由於不希望變異大的變數掩蓋掉其他變異小，但可能有重要資訊的變數，因此將資料進行標準化；為使資料更漂亮方便觀察，也使用了中心化。

六、 變數選擇與維度縮減

由於採用的變數不一定皆與打者表現有關聯，因此我針對打者「每打席的平均得分」建立迴歸模型，並將影響得分的重要變數作為評估打者表現的指標。以下採用兩種不同的方法做特徵選取。

(一) 向後選取法 (backward) 搭配主成分分析 (PCA)

由於最初的目的是保留與打者表現有關聯的指標，雖然這裡以得分作為評斷依據，但不希望僅以與得分有顯著關係的變數作為評估的唯一標準。因此我將條件設得較寬鬆，採用向後選取法並以 p=0.2 作為決定水準。留下 24 個變數，見下表。

變數	參數估計值	標準誤差	t 值	Pr > |t|
打席	-1.55	0.57	-2.73	0.007
四壞率	-0.62	0.18	-3.53	0.000
打擊率	-1.48	0.34	-4.40	<.0001
長打率	0.88	0.09	10.26	<.0001
上壘率	1.77	0.43	4.11	<.0001
盜壘率	-0.11	0.05	-2.30	0.022
盜壘成功率	0.23	0.05	4.76	<.0001
觸身球率	-0.19	0.06	-2.97	0.003
犧牲打中，觸擊的機率	-0.07	0.03	-2.10	0.036
犧牲打中，高飛的機率	-0.06	0.03	-1.74	0.083
甜蜜點機率	0.12	0.04	3.33	0.001
Barrel%	-0.33	0.07	-4.53	<.0001
SolidContact%	-0.08	0.05	-1.61	0.108
Flare/Burner%	-0.24	0.09	-2.70	0.007
Under%	-0.18	0.11	-1.59	0.112
Topped%	-0.24	0.12	-1.98	0.049
Poor/Weak%	-0.25	0.14	-1.83	0.068
重擊球，定義為以 95 英里/小時	0.05	0.04	1.35	0.177
揮擊好球帶以外的球的打中比例	-0.09	0.04	-2.17	0.030

變數	參數估計值	標準誤差	t 值	Pr > |t|
揮擊好球帶以內的球的打中比例	-0.11	0.05	-2.21	0.027
揮空率	-0.19	0.06	-3.35	0.001
中間率	0.05	0.02	2.33	0.020
內野高飛比例	-0.05	0.03	-2.06	0.039
壘間全力衝刺時的平均速度	0.09	0.03	3.36	0.001

由迴歸分析結果知，在 95%的信心水準下，有 19 個指標與得分有顯著的關聯，即為影響得分的重要因子。

我將向後選取法留下的 24 個變數再做主成分分析，取前 5 個主成分作為特徵 (主成分與特徵值陡坡圖詳見附錄二)，可解釋 65%的變異。五個主成分的特徵分別為 (詳見附錄三):

第一主成分	打擊率低、長打率低、上壘率低、擊球仰角集中於-30 度、重擊球少
第二主成分	好/壞球擊中率高、揮空率低
第三主成分	盜壘率及盜壘成功率高、壘間衝刺速度快
第四主成分	擊球仰角集中於 -15 度 (hit under，多為弱滾地)、內野高飛球多
第五主成分	犧牲高飛少、甜蜜點多、擊球仰角集中於 15 度左右 (Flare/Burner %)

(二) 將變數分組並使用 Sparse Group Lasso 做變數選擇

考量到有些特徵可能需要放在同一組裡才有一定的意義，我自行將一些變數合併成組，合併的變數有:

1. 犧牲觸擊率、犧牲高飛率
2. Weak%~Barrel%　6 個擊球仰角與初速的品質
3. 好球帶內、外的球的揮擊比例
4. 揮擊好球帶內、外的球的打中比例
5. 拉打率、中間率、推打率

6. 滾地、高飛、平飛、內野高飛比例

此六組與其餘變數單獨一組，總共 31 組。

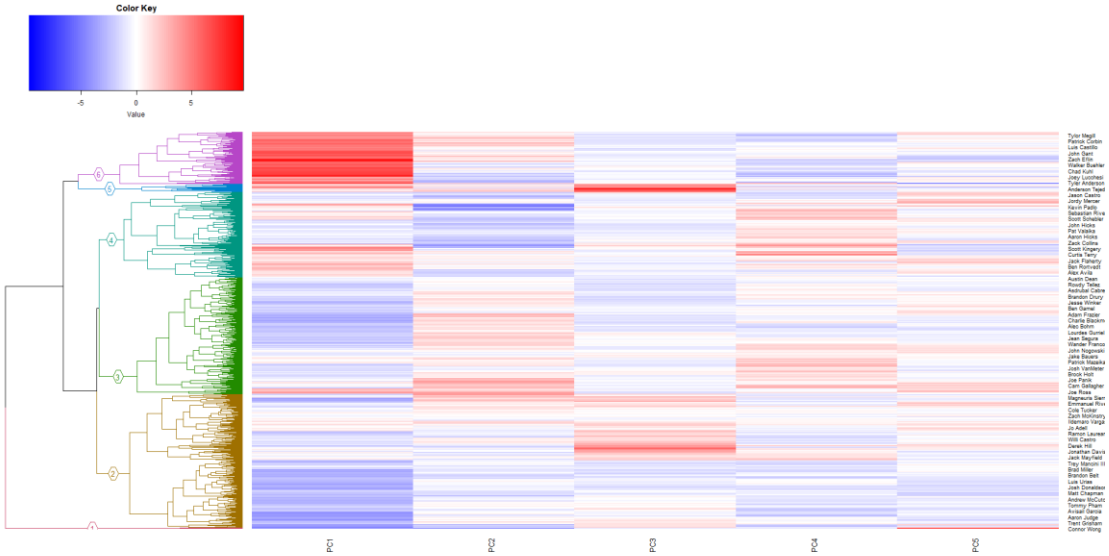
為了使組別與單一特徵皆有稀疏性，我選擇採用 Sparse Group Lasso 做變數選擇，並希望將變數控制在 10 個以內，因此將懲罰係數設為 0.003 (lambda 與組係數收縮關係圖見附錄四)。大部分的變數被收縮至 0，留下下表 8 個對得分影響最大的變數，其中又以長打率最為重要。

變數	收縮後係數
全壘打率	0.004
長打率	0.469
上壘率	0.184
盜壘成功率	0.073
總犧牲率	-0.020
甜蜜點機率	0.017
本壘到一壘秒數	-0.002
壘間全力衝刺時的平均速度 :英尺/秒	0.061

七、 分群結果

經過一些嘗試後，決定對於兩種特徵選取方法，皆採用 complete linkage 階層式分群法，以此觀察各群球員的表現。

(一) 採用向後選取法搭配主成分分析 (PCA) 之分群結果



群	1	2	3	4	5	6
人數	2	247	215	157	15	96

- 第一群: 離群樣本 (PC1 低、PC2 低、PC5 高)
  綜合特徵: 打擊率、長打率、上壘率、重擊球率皆高，但揮空率也較高。此群的特徵以白話文解釋是: 很常揮空，但一擊到球就是安打。以常理判斷並不太合理，我抓出第一群僅有的兩個球員: Connor Wong 與 Seth Beer 檢查，發現 Connor Wong 的特徵皆符合上述，但 Seth Beer 的揮空率其實大約位於平均值，並不算特別高。另外，他們的打席數分別為 14 與 10 個，皆遠遠低於平均 247 個，因此我判斷這一群的資料為上場機會較少導致的離群結果。

- 第二群: 狀況/打擊優秀的球員 (PC1 偏低)
  特徵: 打擊率、長打率、上壘率、重擊球率有一定的水準
  代表球員: Paul Goldschmidt、Matt Olson、Nick Castellanos

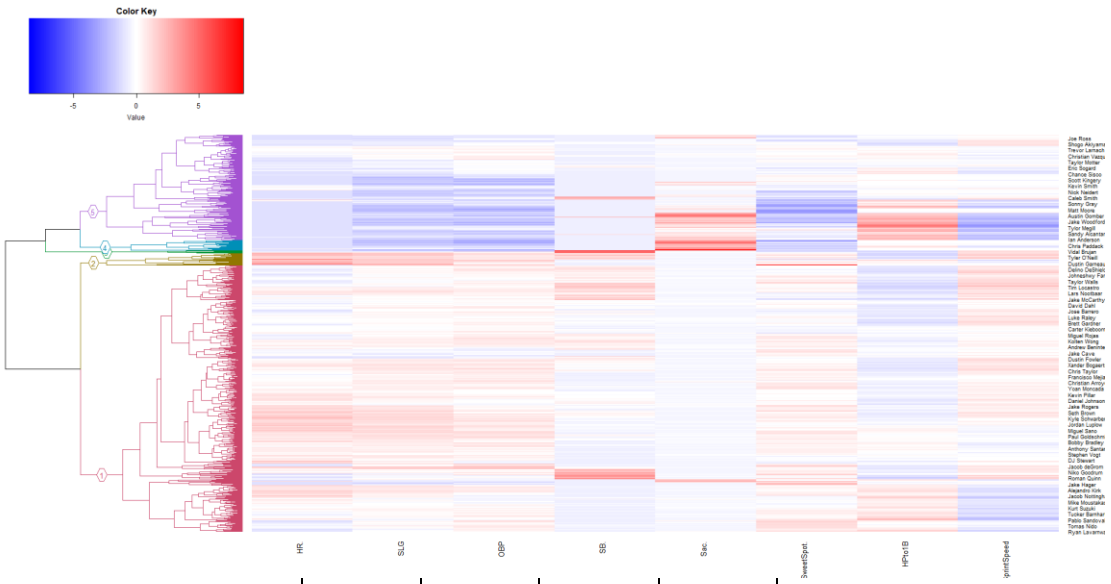
- 第三群: 具技巧性且打擊優秀的球員 (PC1 偏低、PC2 偏高)
  特徵: 打擊率、長打率、上壘率、重擊球率有一定水準、揮空率低
  打者打擊表現良好，且由揮空率低推測此群的打者在技巧上也非常突出，適合技巧性執行進壘、打帶跑等戰術。
  代表球員: DJ LeMahieu、Alex Bregman、Josh Rojas

- 第四群: 技巧較差或狀況不佳的球員 (PC2 偏低、PC4 偏高)
  特徵: 揮空率偏高、常擊出弱滾地與內野高飛球
  代表球員: Brandon Marsh、Pablo Sandoval

- 第五群: 飛毛腿球員 (PC1 偏高、PC3 超高)
  特徵: 打擊率、長打率、上壘率、重擊球率偏低，但盜壘率非常高、壘間衝刺速度快
  代表球員: Cameron Maybin、Adalberto Mondesi、Tyler Wade

- 第六群: PC1 超高 PC4 偏低 (狀況不佳的球員)
  特徵: 打擊率、長打率、上壘率、重擊球率很低，弱滾地、內野高飛球不多
  代表球員: Sammy Long、Walker Buehler、Carlos Carrasco

(二) 將變數分組並使用 Sparse Group Lasso 變數選擇之分群結果



群	1	2	3	4	5
人數	491	23	5	19	194

- 第一群:　狀況佳的球員
  特徵: 長打率、上壘率、甜蜜點率不錯
  代表球員: Buster Posey III、Paul Goldschmidt、Mike Trout

- 第二群:　具有長打能力的球員
  特徵: 全壘打率與長打率很高、上壘率不錯、本壘到一壘秒數偏長、壘間全力衝刺時的平均速度偏慢
  代表球員: Javier Baez、Fernando Tatis Jr.

- 第三群: 飛毛腿球員
  特徵: 盜壘成功率很高、本壘到一壘秒數短、壘間全力衝刺時的平均速度快，全壘打率、長打率、上壘率皆中偏低
  代表球員: Adalberto Mondesi、Tyler Wade

- 第四群: 推進型球員
  特徵: 全壘打率、長打率與上壘率皆偏低，總犧牲率高、甜蜜點偏少
  代表球員: David Price、Eric Lauer

- 第五群: 狀況不佳且較無速度的球員
  特徵: 全壘打率、長打率與上壘率皆低、甜蜜點偏少、本壘到一壘秒數長、壘間全力衝刺時的平均速度慢
  代表球員: Austin Romine、Austin Nola、Yoshi Tsutsugo

由以上兩種特徵選取分別的分群結果，很明顯看出經過 PCA 的特徵會較明顯，但我們在對主成分做解釋時，會以「與此主成分最有關聯的變數」作為解釋主成分的依據，不代表此主成分與其他變數沒有關聯，所以即使看似主成分在群中的特徵明顯，在解釋上仍很難非常全面，導致部分球員無法完全符合此群的特徵。

另外，兩種方法的分群結果皆有分出狀況不佳的球員，經過檢查後，發現此群有許多國聯王牌等級的投手，例如: Yu Darvish (達比修有)、Zack Wheeler、Aaron Nola。國聯球隊直到 2022 年前，投手也要上場打擊 (無 DH 制)，推測因出場投球次數多，打席數多加上打擊表現不佳，而被歸類成狀況不佳的球員。

八、 結論

兩種特徵選取方法選出的特徵稍有不同，但最後的分群結果皆依打擊率、上壘率等綜合指標，分出了打擊優秀與狀況不佳的球員，也依跑步速度相關的變數找出飛毛腿球員。未來評估球員表現時，也不妨同時參考兩種方法選出的特徵，球隊可依此做棒次的調度與戰術構想，提升球隊總戰力，球員也可以針對弱項做進一步加強。

九、 參考文獻

[1] J. Friedman, T. Hastie & R. Tibshirani. (2010). A note on the group lasso and a sparse group lasso

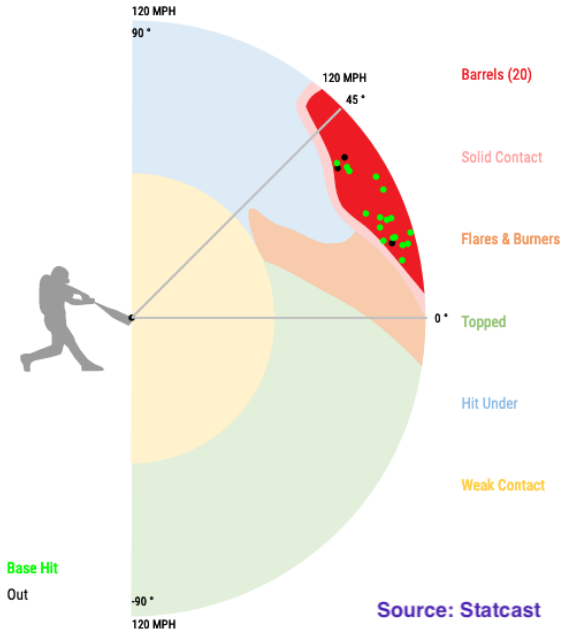
[2] 陳邦瑜 (2012)。資料採礦在中華職棒球員價值之應用

[3] 佚名 (2019)。反其道而行之：讓 Carlos Santana 進化的滾地球革命論。https://medcytw.com/2019/09/05/carlos-santana/

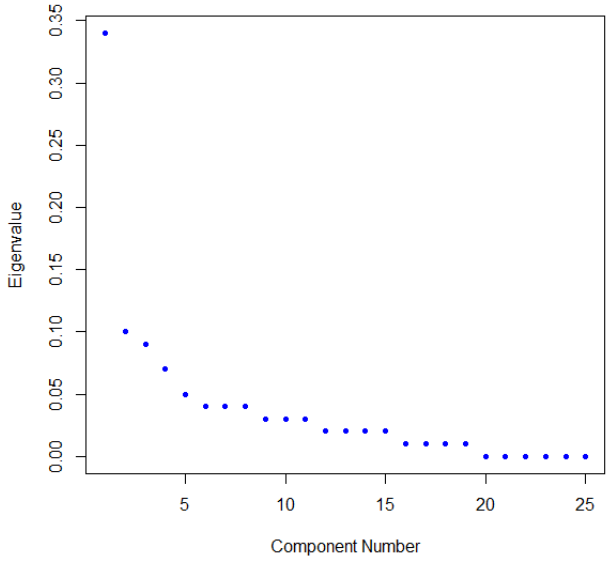
[4] Jonathan Metzelaar. (2020). Beyond the Barrel: An Introduction to Ideal Contact Rate. https://www.pitcherlist.com/beyond-the-barrel-an-introduction-to-ideal-contact-rate/

十、 附錄

附錄一: 擊球仰角與初速品質示意圖



附錄二: 主成分與特徵值之陡坡圖



附錄三: 主成分與變數之關係 (螢光筆為將其列入解釋之變數)

變數	PC1	PC2	PC3	PC4	PC5
打席	-0.25	0.11	-0.07	-0.25	-0.32
四壞率	-0.17	-0.12	-0.04	-0.01	-0.02
打擊率	-0.3	0.13	-0.02	-0.07	0.21
長打率	-0.3	-0.05	-0.03	-0.11	0.06
上壘率	-0.31	0.04	-0.01	-0.04	0.14

盜壘率	-0.06	0.06	0.63	-0.07	-0.02
盜壘成功率	-0.06	0.06	0.62	-0.08	-0.04
觸身球率	-0.06	-0.11	0.11	0.13	-0.07
故意四壞保送率	0.21	0.12	0.03	-0.08	0.17
犧牲打中，觸擊的機率	-0.23	0.04	-0.11	-0.13	-0.32
甜蜜點機率	-0.23	-0.15	0	0.03	0.4
Barrel%	-0.2	-0.31	-0.07	-0.2	-0.14
SolidContact%	-0.16	-0.24	-0.01	-0.04	0.18
Flare/Burner%	-0.2	0.14	-0.03	0.03	0.45
Under%	-0.14	-0.22	0.06	0.5	-0.17
Topped%	0.14	0.3	0.02	-0.21	-0.2
Poor/Weak%	0.26	0.04	-0.01	-0.13	0.04
重擊球，定義為以 95 英里/小時	-0.24	-0.2	-0.06	-0.21	0.01
好球帶以內的球的揮擊比例	-0.19	0.37	-0.07	0.19	0.05
好球帶以外的球的揮擊比例	-0.22	0.32	-0.01	0.26	-0.05
揮擊所有好球帶以外的球的打中比例	0.15	-0.48	0.06	-0.23	0.01
揮擊所有好球帶以內的球的打中比例	0.05	0.19	-0.03	-0.19	0.03
揮空率	-0.06	-0.15	0.03	0.47	-0.31
面對來球揮棒的比例	-0.15	-0.04	0.4	-0.04	-0.01

附錄四: lambda 與組係數收縮關係圖