Hockey players analysis using shots and goals data sets

In the paper 'The Gamma Lasso'[1], the author evaluates the performance of hockey players by using the logistic regression. The model he used is a 'regression plus-minus' model for player contribution, which is

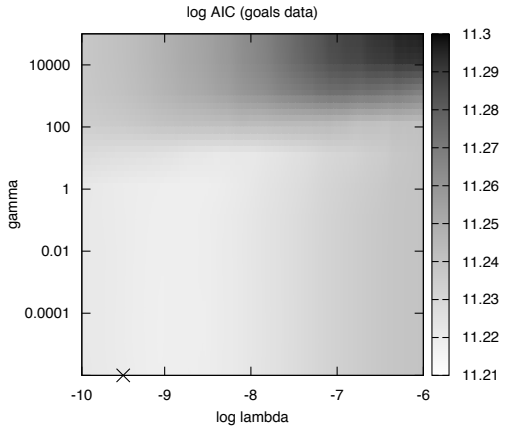
$$logit[P(\text{home team scored goal } i)] = \alpha + \mathbf{u}'\phi + \mathbf{x}'\beta,$$

where $\mathbf{u}$ is a length-7 vector indicating the special-teams scenario and $\mathbf{x}$ is a vector of player effects. For goal/shot i, $x_{ij} = 1$ if player j was on the home team, $x_{ij} = -1$ if he was on the away team and $x_{ij} = 0$ if no one was on ice.

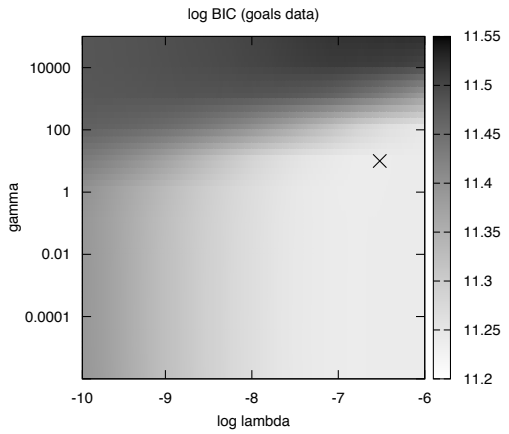
The data of goals applied by the author in 'The Gamma Lasso' paper includes the players on the ice for each goal in the National Hockey League (NHL) from 2002-2003 season. The data of gaols, where there are 64448 goals and 2302 players, can be obtained in the 'gamlr' package for R. More generally, we apply the data of shots, which record each shot information instead of goal (goal is just a special case of shot). The data of shots, including 726352 shots and 2412 players, are in the 'Mataddy/hockey' Github. The aim of our work is to compare the results using the two different data sets.

The gamma lasso regression, which can also be found in the 'gamlr' package for R, is applied. We estimate the gamma lasso paths for β in the model with α and ϕ left unpenalized. The penalties are not standardized since this would penalizing less the players with small standard deviation, i.e. this would have favored players with little ice time. The 100 paths are from the max-abs-gradient $\lambda^1 \approx exp(-6)$ down to $\lambda^{100} = 0.01\lambda^1$. To find out under what value of γ, AIC or BIC attains the minimum, we run the algorithm for $log_{10}\gamma = -5..5$ and $\gamma = 0$ lasso. As we can see in figure 1, in both shots and goals cases, BIC selects higher (or the same) λ and γ than AIC. Note here we have slightly different results as presented in the paper. In figure 1(b), the minimum BIC, that is 11.23818, is attained when $\gamma = 10$ and $log(\lambda) = -6.5$. However, in the paper, the minimum BIC is shown to be obtained as 11.23837 when $\gamma = 1$ and $log(\lambda) = -6.9$. This needs to be figured out later.

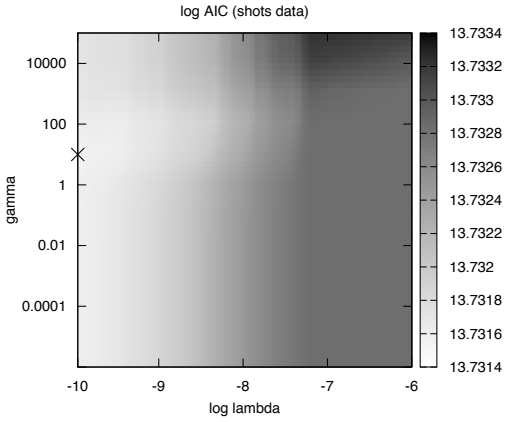
In figure 2, we can observe that estimates jump from zero. The BIC selected model is very sparse. As we can see from figure 3, there are only 7 players get nonzero coefficients $\hat{\beta}_j$ when the data of goals are applied. All the effects are positive. But this seven-player optimal model seems to be unrealistic since there can not only be 7 players that are above or below the average level of NHL. The number of nonzero player effects increases to 34 when we apply the data of shots. At this time, we can see negative player effects. All the seven players shown in graph 3(a) do not appear in graph 3(b). This means that the influential players in making goals are at the average level of making shots. Since the coefficients are not allowed to change from season to season, these can be regarded as career effects.



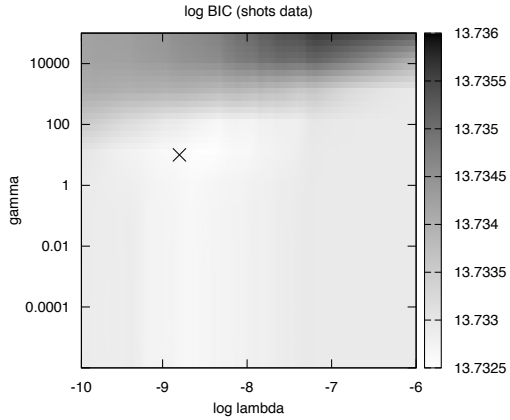
(a) AIC for data of goals



(b) BIC for data of goals

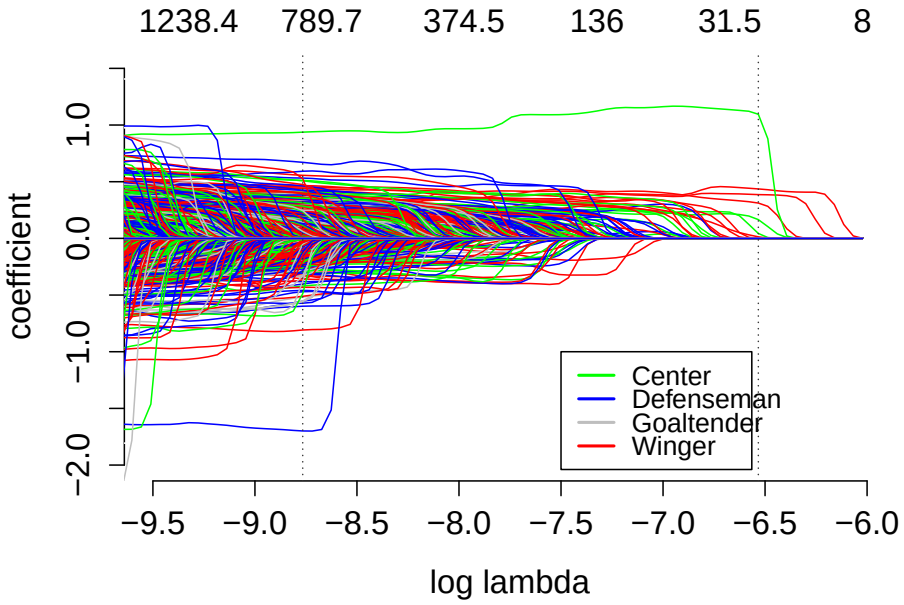


(c) AIC for data of shots

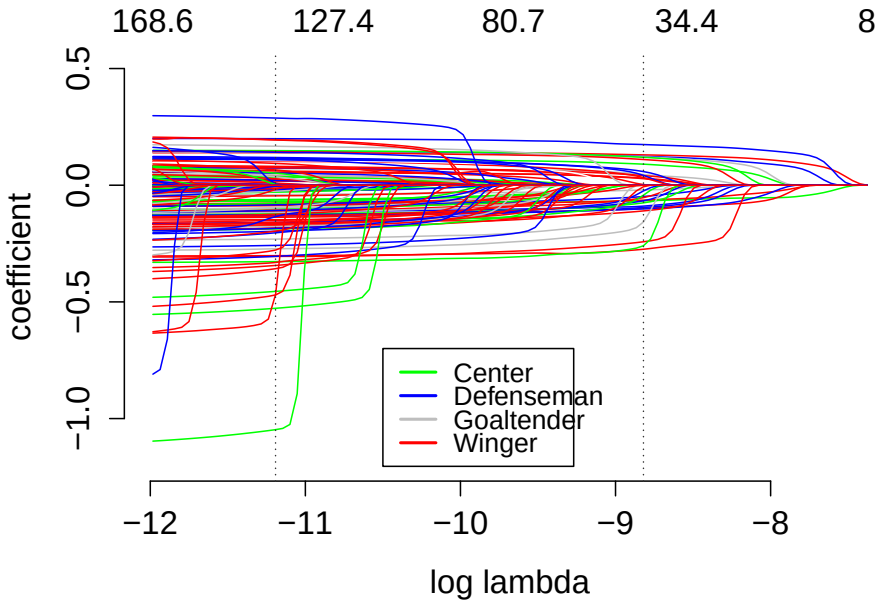


(d) BIC for data of shots

Figure 1: The figure describes the AIC and BIC paths for the regressions using data of goals and data of shots. The 100 $log(\lambda)$ ranges from -6 to -10.6 in models with $log_{10}\gamma = -5..5$ and with lasso $\gamma = 0$ on the bottom edge. In the graphs, 'X' marks the minima.



(a) Regularization path (data of goals)



(b) Regularization path (data of shots)

Figure 2: This figure are the regularization paths for $\gamma = 10$ by using goals and shots data sets. The minimum AIC and BIC are marked on each graph with dashed vertical lines. In both graphs, the left dashed lines indicate minimum AIC. The curves in graphs are colored by different positions of the players.

	name	position	plus.minus	coefficent.beta
599	PETER_FORSBERG	C	255	1.09518866
418	PAVEL_DATSYUK	C	584	0.43624989
851	MARIAN_HOSSA	R	422	0.38606779
1854	HENRIK_SEDIN	C	535	0.31165957
308	ZDENO_CHARA	D	288	0.16549786
210	DAN_BOYLE	D	269	0.04337137
2075	JOE_THORNTON	C	476	0.01379417

(a) Non-zero player effects (data of goals)

	name	position	plus.minus	coefficent.beta
69	NICKLAS_BACKSTROM	C	1864	0.174292201
16	DANIEL_ALFREDSSON	R	2322	0.130011294
135	PATRICE_BERGERON	C	1389	0.123064046
65	DAVID_BACKES	R	894	0.121282246
59	SEAN_AVERY	L	812	0.119601566
41	TYLER_ARNASON	C	1131	0.110099132
43	JASON_ARNOTT	C	2018	0.090156873
133	SEAN_BERGENHEIM	L	141	0.058386150
164	JASON_BLAKE	L	917	0.045698837
25	TONY_AMONTE	R	644	0.042319965
134	MARC-ANDRE_BERGERON	D	1494	0.041279004
191	DAVID_BOOTH	L	689	0.032699819
145	TODD_BERTUZZI	R	1828	0.030175934
165	ROB_BLAKE	D	735	0.027907976
139	PATRIK_BERGLUND	C	689	0.002622703
30	DAVE_ANDREYCHUK	L	-207	-0.274424815
136	MARC_BERGEVIN	D	-473	-0.259324141
185	BRAD_BOMBARDIR	D	-455	-0.242422204
51	SERGE_AUBIN	L	-548	-0.157449485
147	BLAIR_BETTS	C	-1306	-0.110138937
113	ERIC_BELANGER	C	-1080	-0.101538332
76	KEITH_BALLARD	D	-1098	-0.099650151
14	ANDREW_ALBERTS	D	-1047	-0.095428860
38	COLBY_ARMSTRONG	R	-737	-0.089536458
110	STEVE_BEGIN	L	-936	-0.075541641
73	JOSH_BAILEY	C	-91	-0.066803293
70	NIKLAS_BACKSTROM	G	-916	-0.066658763
26	CRAIG_ANDERSON	G	-974	-0.063932796
87	STU_BARNES	C	-678	-0.053162515
19	BRYAN_ALLEN	D	-1360	-0.048499563
98	SHAWN_BATES	C	-419	-0.042569951
176	MIKKEL_BOEDKER	L	-122	-0.024467235
155	MARTIN_BIRON	G	-565	-0.020498559
115	MATT_BELESKEY	L	-241	-0.006659799

(b) Non-zero player effects (data of shots)

Figure 3: Non-zero player effects in the minimum BIC model.

References

[1] M. Taddy (2013): *The Gamma Lasso*.