Wirtschaftsuniversität Wien May 24, 2013

The Ranking Lasso

Cristiano Varin
Università Ca' Foscari, Venice

Based on joint works with:
Guido Masarotto (Padova)
Manuela Cattelan (Padova) & David Firth (Warwick)

References

Masarotto, G. and Varin, C. (2012). The Ranking Lasso and its Application to Sport Tournaments. *The Annals of Applied Statistics* **6** (4), 1949–1970.

Varin, C., Cattelan, M., and Firth, D. (2013?). Paired Comparison Modelling of Citation Exchange Among Statistics Journals. In preparation.

College Ice Hockey

NCAA Men's Division I 2009-2010

- 58 teams partitioned into six conferences
- Regular season: 1 083 games
- Highly incomplete and unbalanced tournament design
  - 73.3% of the $\binom{58}{2}$ possible matches not played
  - 6.8% played just once
  - 10.5% played twice
  - 9.4% three or more times

  Number of matches per team ranges from 31 to 43
- Six automatic bids (division winners) go to the conference tournament champions
- The remaining 10 teams are selected upon ranking under the NCAA's system of pairwise comparisons
- Data available through R package BradleyTerry2

Turner and Firth (2012)

Notation

- Tournament with k teams
- Match between team i and team j played n_{ij} times ($n_{ij} \geq 0$)
- Total number of matches $n = \sum_{i<j} n_{ij}$
- Y_{ijr} outcome of the rth match between team i and team j

$$Y_{ijr} = \begin{cases} +1, & \text{if team } i \text{ defeats team } j, \\ 0, & \text{if teams } i \text{ and } j \text{ tied,} \\ -1, & \text{if team } i \text{ is defeated by team } j \end{cases} \qquad (1 < i < j < k)$$

for $r = 1, \ldots, n_{ij}$

Bradley-Terry with Ties and Home Advantage

Cumulative logit model with team abilities $\mu_1, \dots, \mu_n$

$$\log\left\{\frac{\text{pr}(Y_{ijr} \le m)}{\text{pr}(Y_{ijr} > m)}\right\} = \underbrace{\delta_{(m)}}_{\text{cutpoint}} + \underbrace{h_{ijr}\,\tau}_{\text{home effect}} + \underbrace{\mu_i - \mu_j}_{\text{ability difference}} \quad (m = -1, 0, 1)$$

Cutpoints $-\infty \le \delta_{(-1)} \le \delta_{(0)} \le \delta_{(1)} \equiv +\infty$

Home-field indicator

$$h_{ijr} = \begin{cases} +1, & \text{match played at home of team } i, \\ 0, & \text{match played at neutral field,} \\ -1, & \text{match played at home of team } j \end{cases}$$

Model identifiability:

- one constraint on ability vector $\sum_{i=1}^{n} \mu_i = 0$
- for every match played on a neutral field, the model must assure that $\text{pr}(Y_{ijr} = 1)$=$\text{pr}(Y_{jir} = -1)$, then $\delta_{(-1)} = -\delta_{(0)}$

Ranking Journals

Impact Factor?

Rank	Abbreviated Journal Title (linked to journal information)	ISSN	JCR Data					
			Total Cites	Impact Factor	5-Year Impact Factor	Immediacy Index	Articles	Cited Half-life
1	J STAT SOFTW	1548-7660	1795	4.010	4.791	1.537	95	4.3
2	J R STAT SOC B	1369-7412	12345	3.645	5.281	0.793	29	>10.0
3	STAT SCI	0883-4237	3054	3.035	4.205	0.259	27	>10.0
4	ANN STAT	0090-5364	11722	3.030	3.700	0.423	104	>10.0
5	ECONOMETRICA	0012-9682	19659	2.976	4.700	0.688	48	>10.0
6	STAT METHODS MED RES	0962-2802	1835	2.443	2.988	0.500	36	>10.0
7	STATA J	1536-867X	1250	2.222	3.063	0.147	34	6.4
8	BIOSTATISTICS	1465-4644	2225	2.145	3.162	0.519	54	7.3
9	J R STAT SOC A STAT	0964-1998	1685	2.110	2.275	0.327	49	>10.0
10	PHARM STAT	1539-1604	422	2.067	2.160	0.463	67	4.1
11	J AM STAT ASSOC	0162-1459	21348	1.992	3.310	0.240	121	>10.0
12	J CHEMOMETR	0886-9383	2422	1.952	1.976	0.273	66	8.9
13	CHEMOMETR INTELL LAB	0169-7439	4494	1.920	2.295	0.350	137	9.9
14	BIOMETRIKA	0006-3444	13222	1.912	2.575	0.179	78	>10.0
15	STAT MED	0277-6715	13901	1.877	2.582	0.395	258	9.7
16	BIOMETRICS	0006-341X	14212	1.827	2.249	0.251	171	>10.0
17	ANN PROBAB	0091-1798	3517	1.789	1.669	0.183	71	>10.0
18	J BUS ECON STAT	0735-0015	2919	1.779	2.442	0.583	48	>10.0
19	FUZZY SET SYST	0165-0114	9886	1.759	1.988	0.199	171	>10.0
20	BAYESIAN ANAL	1931-6690	502	1.650	3.077	0.258	31	5.2

Stigler Model

Stigler (1994):

- Journal importance given by the ability to "export intellectual influence"
- The export of influence is measured by the citations received by the journal
- Bradley-Terry model

$$\text{log-odds}\,(\text{ journal } i \text{ is cited by journal } j\,) = \mu_i - \mu_j$$

  where μ_i is the export score of journal i
- The larger the export score, the greater the propensity to export intellectual influence

Maximum Likelihood?

- Likelihood function of the Bradley-Terry simple...
- ...but is maximum likelihood estimation appropriate here?
- Shrinkage estimation outperforms maximum likelihood for simultaneous inference on a vector of mean effects
- The Bradley-Terry model is identified through pairwise differences $\mu_i - \mu_j$
- Plan: fit Bradley-Terry model with penalty on each pairwise difference $\mu_i - \mu_j$

The Ranking Lasso

- Lasso estimation of the Bradley-Terry model

$$\hat{\boldsymbol{\mu}}_s = \arg\max \ell(\boldsymbol{\mu}) \text{ subject to } \sum_{i<j}^{k} w_{ij}|\mu_i - \mu_j| \leq s$$

  where w_{ij} are pair-specific weights
  [likelihood for ice hockey data also contains the home effect and a cut point parameter: $\ell(\boldsymbol{\mu}, \tau, \delta)$]
- Standard maximum likelihood for a sufficiently large value of the bound s
- Fitting penalized as s decreases to zero
- Ranking in groups: L1 penalty induces groups of team ability parameters estimated to the same value

Generalized Fused Lasso

- Fused lasso designed for problems where parameters have natural order — Tibshirani et al. (2005)
- L1 penalty on pairwise differences of successive coefficients

$$\hat{\boldsymbol{\mu}}_s = \arg\max \ell(\boldsymbol{\mu}) \text{ subject to } \sum_{i=1}^{k-1} w_i |\mu_i - \mu_{i+1}| \leq s$$

- Ranking lasso as generalized fused lasso with penalty on all possible pairs $\mu_i - \mu_j$
- Lack of order in the ranking lasso implies <u>substantial</u> computational complications
- Difficulty from the one-to-many relationship between μ_i and penalized parameters $\theta_{ij} = \mu_i - \mu_j$, $i < j$

Ranking lasso equivalent to the penalized minimization problem

$$\hat{\boldsymbol{\mu}}_\lambda = \arg\min \left\{ -\ell(\boldsymbol{\mu}) + \lambda \sum_{i<j}^{k} w_{ij} \left| \mu_i - \mu_j \right| \right\}$$

Helpful to re-express as a constrained ordinary lasso problem

$$\left(\hat{\boldsymbol{\mu}}_\lambda, \hat{\boldsymbol{\theta}}_\lambda \right) = \arg\min \left\{ -\ell(\boldsymbol{\mu}) + \lambda \sum_{i<j}^{k} w_{ij} |\theta_{ij}| \right\}$$

$$\text{subject to} \quad \theta_{ij} = \mu_i - \mu_j, \quad 1 < i < j < k$$

Lagragian Form of the Ranking Lasso

- Lagrangian form: minimize

$$-\ell(\boldsymbol{\mu}) + \lambda \sum_{i<j}^{k} w_{ij}|\theta_{ij}| + \sum_{i<j}^{k} u_{ij}\left(\theta_{ij} - \mu_i + \mu_j\right)$$

- Computation of Lagrangian multipliers u_{ij} is ill-posed problem
- Simpler solution: replace the Lagrangian term with

$$\frac{v}{2}\sum_{i<j}^{k}(\theta_{ij} - \mu_i + \mu_j)^2$$

- Quadratic penalty form converges to the solution of the ranking lasso as v diverges
- Numerical analysis literature discourages quadratic penalty, because of instabilities for large values of v

Augmented Lagrangian Method

- First developed in late 60's, then loss of attention in favor of sequential quadratic programming and interior point methods
- Recently, revived for total-variation denoising and compressed sensing Nocedal and Wright (2006)
- Augmented objective function

$$F_{\lambda,v}(\boldsymbol{\mu}, \boldsymbol{\theta}, \mathbf{u}) = -\ell(\boldsymbol{\mu}) + \lambda \sum_{i<j}^{k} w_{ij}|\theta_{ij}| + \underbrace{\sum_{i<j}^{k} u_{ij}\,(\theta_{ij} - \mu_i + \mu_j)}_{\text{Lagrangian term}} + \underbrace{\frac{v}{2}\sum_{i<j}^{k}(\theta_{ij} - \mu_i + \mu_j)^2}_{\text{quadratic penalty}}$$

- Augmented Lagrangian method iterates through
  (1) Given $(\mathbf{u}, v)$, minimize $F_{\lambda,v}(\boldsymbol{\mu}, \boldsymbol{\theta}, \mathbf{u})$ with respect to $(\boldsymbol{\mu}, \boldsymbol{\theta})$
  (2) Given $(\boldsymbol{\mu}, \boldsymbol{\theta})$, update tuning coefficients $\mathbf{u}$ and v

Minimization Step

Cycle between

- Minimization with respect to $\boldsymbol{\mu}$ given $\boldsymbol{\theta}$
  - Approximated by Bradley-Terry regression with ridge penalty

$$\hat{\boldsymbol{\mu}} = \arg\min \left\{ -\ell(\boldsymbol{\mu}) + \frac{v}{2} \sum_{i<j}^{k} (\theta_{ij} - \mu_i + \mu_j)^2 \right\}$$

- Minimization with respect to $\boldsymbol{\theta}$ given $\boldsymbol{\mu}$
  - Equivalent to ordinary lasso problem with an orthogonal design
  - Solution provided by soft-thresholder operator

$$\hat{\theta}_{ij} = \text{sign}(\tilde{\theta}_{ij}) \left(|\tilde{\theta}_{ij}| - \frac{\lambda w_{ij}}{v} \right)_+, \quad 1 < i < j < k$$

where $\tilde{\theta}_{ij} = \hat{\mu}_i - \hat{\mu}_j - u_{ij}/v$

Updating Lagrangian Multipliers

- The Augmented Lagrangian function provides a direct recursion for updating Lagrangian multipliers
- Rearranging terms, we have

$$F_{\lambda,v}(\boldsymbol{\mu},\boldsymbol{\theta},\mathbf{u}) = -\ell(\boldsymbol{\mu}) + \lambda \sum_{i<j}^{k} w_{ij}|\theta_{ij}| +$$

$$+ \sum_{i<j}^{k} \underbrace{\left\{ u_{ij} + \frac{v}{2}\left(\theta_{ij} - \mu_i + \mu_j\right) \right\}}_{\text{new } u_{ij}} (\theta_{ij} - \mu_i + \mu_j)$$

- Thus, suggesting the recursion

$$\hat{u}_{ij}^{(\text{new})} = \hat{u}_{ij}^{(\text{old})} + \frac{v}{2}(\hat{\theta}_{ij} - \hat{\mu}_i + \hat{\mu}_j)$$

- Set $\hat{v} = \max\{\hat{u}_{ij}^2\}$

Adaptive Ranking Lasso

- Lasso can yield inconsistent estimation of the nonzero effects because the shrinkage produced by the $L1$ penalty is too severe
- Solutions:
  - substitute $L1$ penalty with another penalty that penalizes large effects less severely, *e.g.* SCAD — Fan and Li (2001)
  - adaptive lasso: give more weight to terms of the $L1$ penalty as the size of the effect decreases — Zou (2006)
- Adaptive ranking lasso

$$\hat{\boldsymbol{\mu}}_\lambda = \arg\min \left\{ -\ell(\boldsymbol{\mu}) + \lambda \sum_{i<j}^{k} w_{ij} \, |\mu_i - \mu_j| \right\}$$

with weights inversely proportional to a consistent estimator of the ability difference

$$w_{ij} = |\hat{\mu}_i^{(\mathrm{mle})} - \hat{\mu}_j^{(\mathrm{mle})}|^{-1}$$

- Maximum likelihood estimates $\hat{\mu}_i^{(\text{mle})}$ diverge when team i wins or loses all its matches
- Compute weights

$$w_{ij} = |\tilde{\mu}_i - \tilde{\mu}_j|^{-1}$$

with $\tilde{\mu}_i$ modified maximum likelihood estimator constructed so to guarantee finiteness, for example

  - add ϵ-ridge penalty

$$\tilde{\boldsymbol{\mu}} = \arg\min \left\{ -\ell(\boldsymbol{\mu}) + \epsilon \sum_{i<j} (\mu_i - \mu_j)^2 \right\}$$

    for a small $\epsilon \approx 0.0001$
  - Firth's bias correction Firth (1993)

$$\tilde{\boldsymbol{\mu}} = \arg\min \left\{ -\ell(\boldsymbol{\mu}) - \frac{1}{2} \log |\mathbf{I}(\boldsymbol{\mu})| \right\}$$

    with $\mathbf{I}(\boldsymbol{\mu})$ Fisher information [Jeffreys prior]

Selection of the Ranking Lasso Penalty

- Compute ranking lasso solution for a range of values of λ
- Efficient implementation: increase (decrease) λ smoothly and use estimates at previous step as warm starts for the successive step
- Selection of λ through information criteria

$$\text{AIC}(\lambda) = -2\,\ell(\hat{\boldsymbol{\mu}}_\lambda) + 2\,\text{enp}(\lambda)$$
$$\text{BIC}(\lambda) = -2\,\ell(\hat{\boldsymbol{\mu}}_\lambda) + \log(n)\,\text{enp}(\lambda)$$

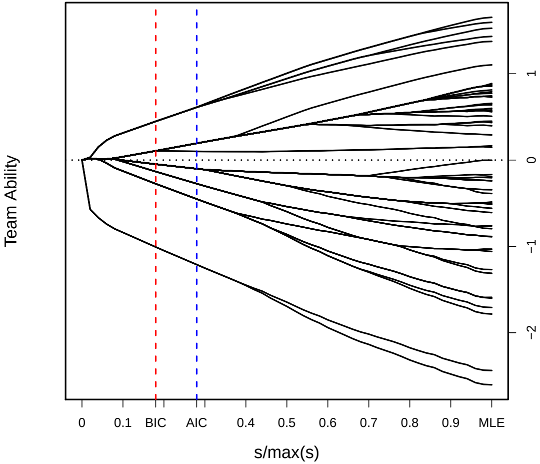
with

- effective number of parameters (enp) estimated as the number of distinct groups formed with a certain λ
- $\hat{\boldsymbol{\mu}}_\lambda$ hybrid adaptive ranking lasso estimate

Chen and Chen (2008)


HERE'S THE FACE-OFF FOR THE START OF THE BIG HOCKEY GAME..

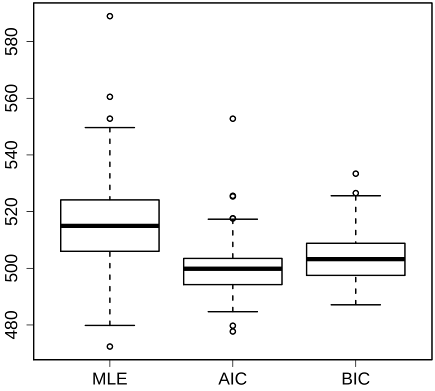
Ranking Lasso Path



AIC: 7 groups BIC: 6 groups

Cross-validation exercise

(1) training/validation: half of the matches randomly sampled
(2) fit model by adaptive ranking lasso on training set
(3) compute log-likelihood on the validation set

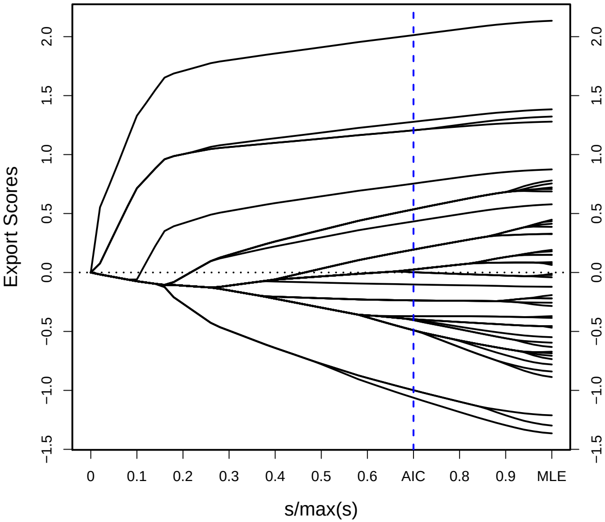


Boxplot of 100 cross-validated *negative* log-likelihoods

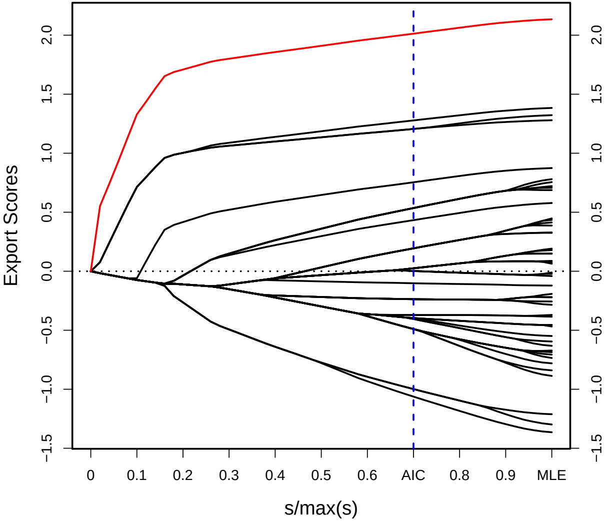
Ranking Journals

- Data from Thomson Reuters Journal Citation Reports edition 2011
- Statistics and Probability category: 106 journals in Statistics, Probability, Econometrics, Chemiometrics, ...
- Most journals within the category exchange very few citations
- Analysis using a selection of 51 journals in Statistics (no Probability, no Econometrics, no ...)
- Adaptive ranking lasso fit:
  - AIC identifies 16 groups
  - BIC identifies 14 groups

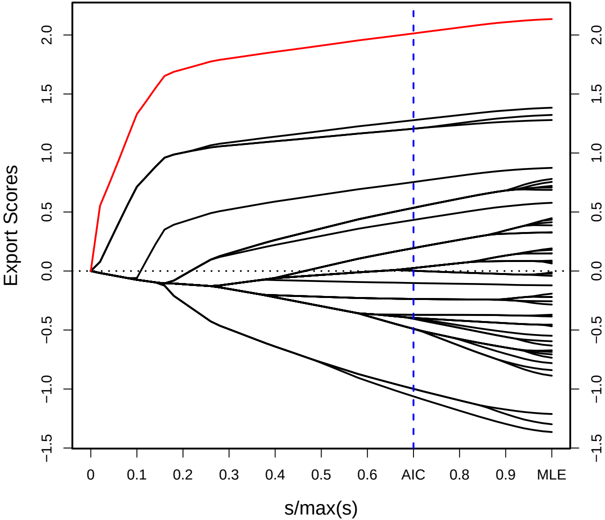
Ranking Lasso Path



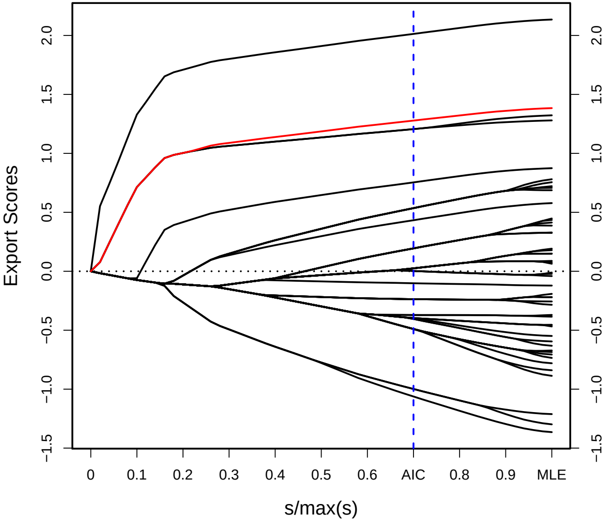
Ranking Lasso Path



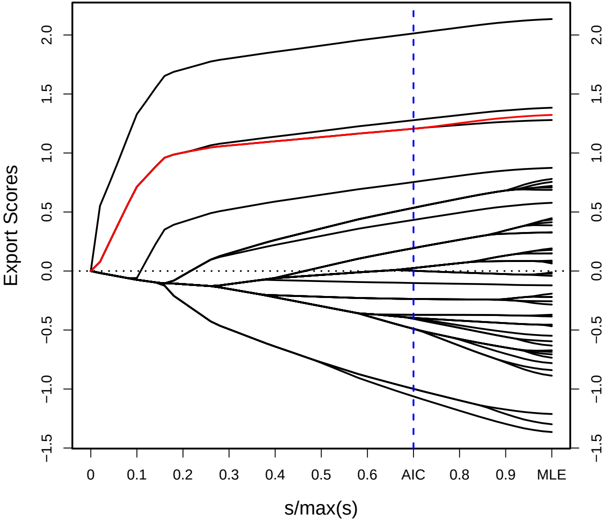
JRSS-B


Export Scores
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
0
0.1
0.2
0.3
0.4
0.5
0.6
AIC
0.8
0.9
MLE
s/max(s)

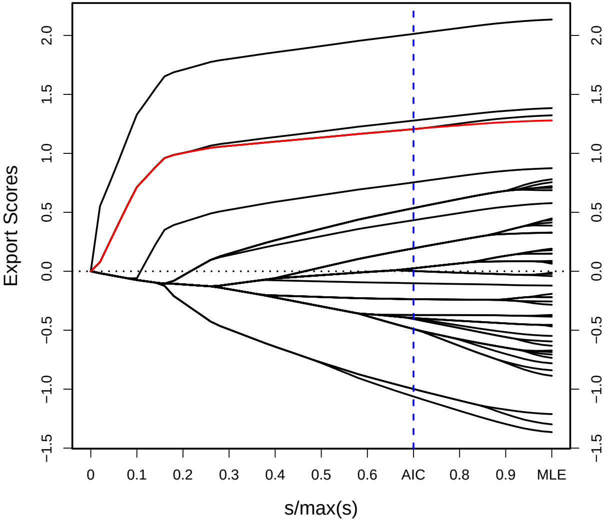
Annals of Statistics



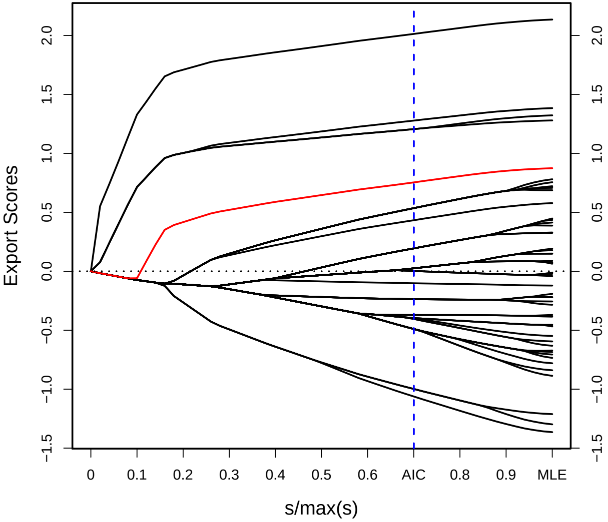
Biometrika


Export Scores
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
0
0.1
0.2
0.3
0.4
0.5
0.6
AIC
0.8
0.9
MLE
s/max(s)

JASA


Export Scores
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
0
0.1
0.2
0.3
0.4
0.5
0.6
AIC
0.8
0.9
MLE
s/max(s)

Biometrics


Export Scores
s/max(s)
0
0.1
0.2
0.3
0.4
0.5
0.6
AIC
0.8
0.9
MLE
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0

Top ten journals according to Stigler Model

			Lasso	
	Journal	MLE	AIC	BIC
1	JRSS B	2.13	2.00	1.97
2	Annals	1.38	1.27	1.24
3	Biometrika	1.32	1.20	1.20
4	JASA	1.28	1.20	1.20
5	Biometrics	0.87	0.74	0.71
6	Bernoulli	0.78	0.52	0.47
7	JRSS A	0.76	0.52	0.47
8	JCGS	0.72	0.52	0.47
9	Scandinavian J	0.71	0.52	0.47
10	Biostatistics	0.69	0.52	0.47

Final Remarks

- Uncertainty quantification
  - Uncertainty quantification of adaptive lasso estimators can be performed via parametric bootstrap

    Chatterjee and Lahiri (2011)
  - By construction, adaptive ranking lasso estimators are biased, then sensible to adjust bootstrap confidence intervals for bias

    Efron (1987)
- Future extensions to deal with dynamic evolution of team/player/journal abilities during several seasons/years
- Augmented Lagrangian method does not scale enough for large ranking lasso applications
  - needs $\mathcal{O}(k^2)$ pairwise difference parameters θ_{ij} for estimation of $\mathcal{O}(k)$ ability parameters
  - looks for more efficient alternatives for large scale problems

Many thanks for your attention!