Analysis of Paper Citations in Statisticians Network

Minshi Peng

Collaborators: Jiashun Jin (CMU)
Pengsheng Ji (Univ. Georgia)
Tracy Ke (Univ. Chicago)

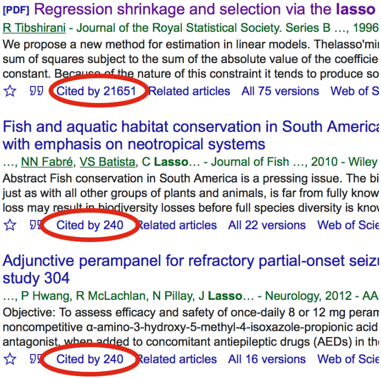
Dec 8, 2017

Data

All published papers in 36 journals in Statistics and related fields, spanning 40 years (1976-2015)

- 70824 papers, 39616 authors
- For each paper, all informations except main content (bibtex, MathSciNet)
- Citation relationships (Web of Science)

Why citation analysis



[PDF] Regression shrinkage and selection via the **lasso**
R Tibshirani - Journal of the Royal Statistical Society. Series B ..., 1996
We propose a new method for estimation in linear models. Thelasso'mi
sum of squares subject to the sum of the absolute value of the coefficie
constant. Because of the nature of this constraint it tends to produce so
☆ 99 Cited by 21651 Related articles All 75 versions Web of S

Fish and aquatic habitat conservation in South America
with emphasis on neotropical systems
..., NN Fabré, VS Batista, C **Lasso**... - Journal of Fish ..., 2010 - Wiley
Abstract Fish conservation in South America is a pressing issue. The bi
just as with all other groups of plants and animals, is far from fully know
loss may result in biodiversity losses before full species diversity is knov
☆ 99 Cited by 240 Related articles All 22 versions Web of Scie

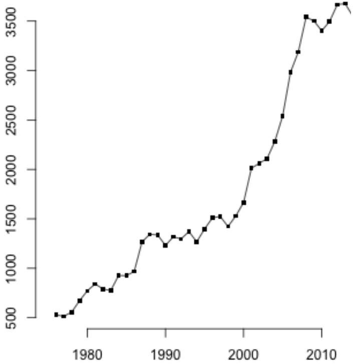
Adjunctive perampanel for refractory partial-onset seizu
study 304
..., P Hwang, R McLachlan, N Pillay, J **Lasso**... - Neurology, 2012 - AA
Objective: To assess efficacy and safety of once-daily 8 or 12 mg peran
noncompetitive α-amino-3-hydroxy-5-methyl-4-isoxazole-propionic acid
antagonist, when added to concomitant antiepileptic drugs (AEDs) in th
☆ 99 Cited by 240 Related articles All 16 versions Web of Scie

For evaluation:

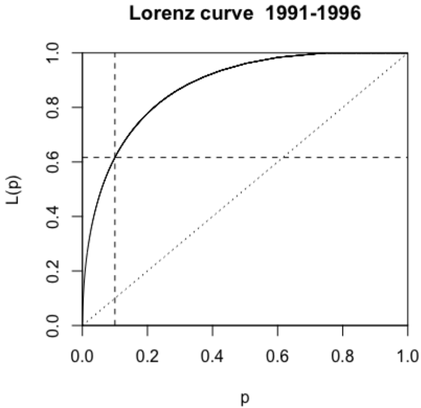
- Journal impact factor
- H index

For understanding the research comminity

Data overview

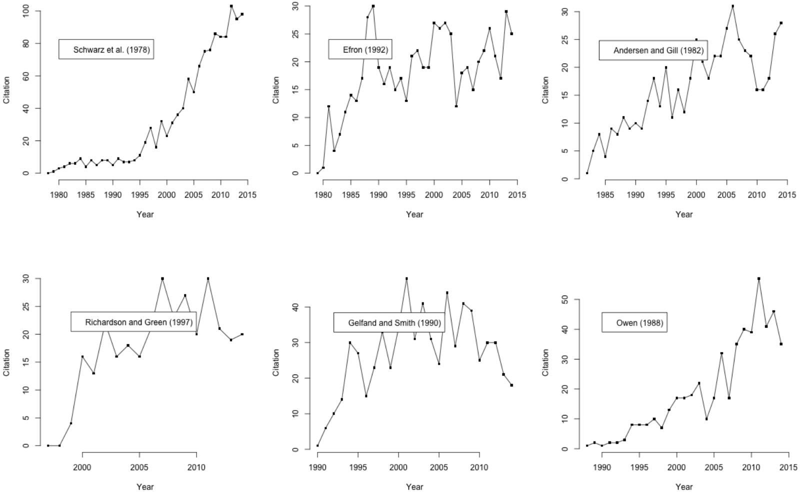


Number of papers published each year



Average citation: 5.43.

Different citation patterns



Analysis objective

Two related problems about paper citations:

- **Why people cite:** how to select predictive features and use them to predict # of future citations.
- **Unsupervised citation pattern learning:** characterize citation patterns of individual papers, and so to cluster into a few groups of papers.

Problem I: Why people cite

Goal: build a prediction model, where we predict future citations using a few predictive features.

- Y variable: divide papers into two classes: "highly cited" and "moderately cited".
- X variables: predictive features.
- Models and methods

Challenges:

- Total citations are not comparable for papers published in different years
- Have limited predictors, need to extract hidden information

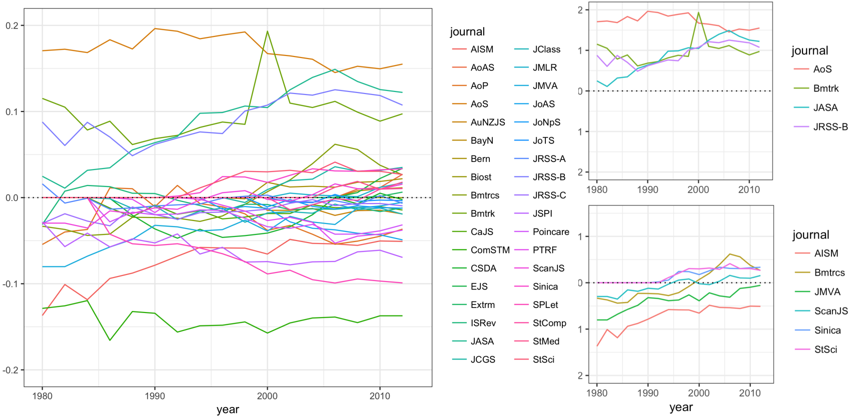
Categorize citation counts (Y variable)

- Related citation counts for papers published in each 5 year periods.
- top 10% → "highly cited" (1226)
- The rest → "moderately cited" (11061)

Select predictive features (X variables)

- Journal, year, title
- Author-related features
- Collaborator-related features
- Reference-related features

Regress the citation each year on the **Journal**



Methods

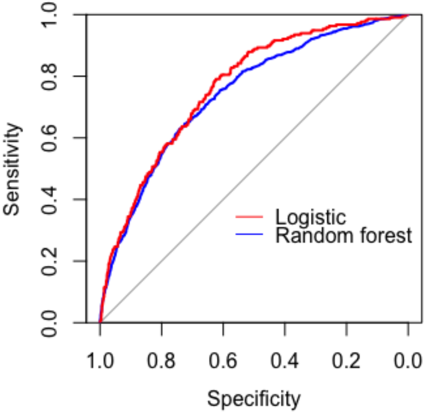
17 features, 12287 observations and 2 classes.

↙ ↘

Weighted logistic regression
Stepwise selection

Random forest

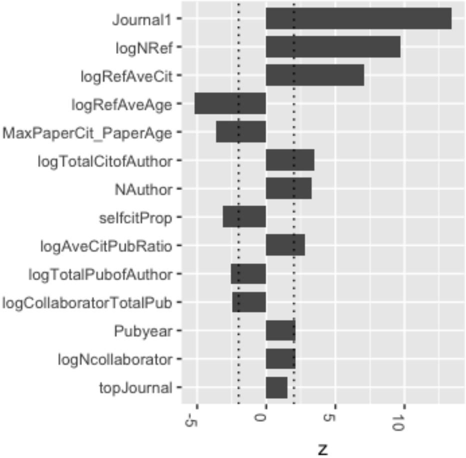
80% training
20% testing



Logistic regression: fitted results

- Accuracy: 67.9%
- Balanced accuracy 69.5%
- Confusion matrix

Prediction vs reference	moderately cited	highly cited
mod cited	1492	69
highly cited	719	175



Interpretations

- Have accuracy approximately 70%, far better than flipping coins.
- The reference are play essential roles
- Problem: rank may change when measured at different time.

Problem II: Citation patterns clustering

Goal: find characteristic citation patterns by clustering citation patterns.

Challenge: model citation patterns

Problem II: Citation patterns clustering

Goal: find characteristic citation patterns by clustering citation patterns.

Challenge: model citation patterns

Methods

- Model the individual citation pattern.
- Clustering based on some distance measure.
- Group-based trajectory modeling (mixture model)

model: Group Based Trajectory of Development

- Input: $Y_i = (Y_i^{(1)}, \cdots, Y_i^{(t)}, \cdots, Y_i^{(T)})$ are observations for sample i. Assume K classes

$$P(Y_i) = \sum_k^K P^k(Y_i)\pi_k, \; P^k(Y_i) = \prod^T P^k(Y_i^{(t)})$$

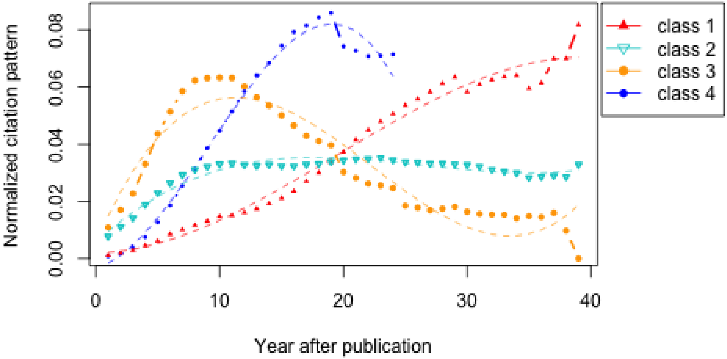
  where $\mathbb{E}[Y_i^{(t)}|\text{class}_k] = f_k(t)(\text{ polynomial})$

- Output:
  posterior probabilities $P(\text{class}_k|Y_i)$

Results:

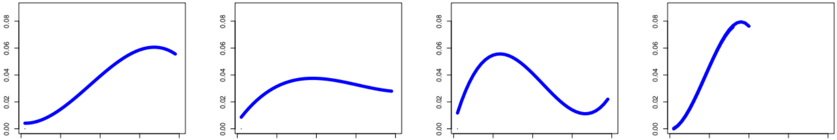
Preprocessing: smoothing and normalization.

Fitted result: K is chosen by BIC



class 1 (15.7%), class 2 (29.8%), class 3 (25.4%), class 4 (29.2%)

Distinct keywords for each group



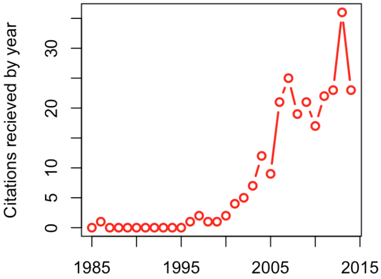
Group 1	Group 2	Group 3	Group 4
time data(.72)	spline smooth(.88)	bay factor(.71)	miss data(.72)
failur time(.56)	kernel densiti(.72)	wavelet(.68)	semiparametr(. 7)
multipl regress(.5)	kernel method(.72)	empir process(.57)	empir likelihood(.69)
procedu(.5)	censor data(.63)	bandwidth(.54)	longitudin data(.67)
class(.45)	repeat measur(.6)	densiti estim(.52)	model select(.67)
regress analysi(.43)	least squar(.57)	bay(.5)	p value(.67)
theori(.38)	CLT(.55)	bayesian analysi(0.44)	random effect(.6)
surviv data(.33)	logist regress(.5)	function estim (.44)	mix model(.5)
hazard model(.33)	nonparametr estim(.45)	markov chain (.4)	algorithm(.5)

”Sleeping beauty”

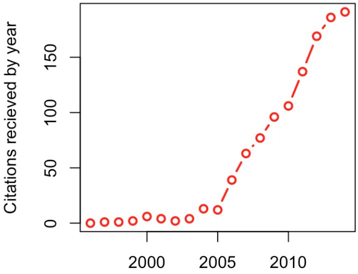
sleeping score 1: $\dfrac{\sum_{t=11}^{15} c(t)}{\sum_{t=1}^{10} c(t)}$

sleeping score 2: $\dfrac{\sum_{t=16}^{20} c(t)}{\sum_{t=1}^{15} c(t)}$

Azzalini A. A class of distributions which includes the normal ones[J].



Tibshirani R. Regression shrinkage and selection via the lasso[J].



Summary:

- Identified citation pattern groups.
  Analyzed the features of these groups.
- Investigated some special patterns like ”sleeping beauties”
- Built a prediction model. Identified some predictive variables.

Future work:

- Clean the abstract data.
- Analyze how the interests in different statistical topics change over time.
- Incorporate keywords information in prediction model.

Thank you !