Empirical Bayes approaches to PageRank type algorithms for rating scientific journals

Jean-Louis Foulley, Gilles Celeux, Julie Josse

INRIA, École Polytechnique

May 29, 2017

Overview

Motivation

Assessing and ranking journals

- **Impact Factor** (Garfield, 1972): average number of annual citations received / published article in the last 2 years. Impacts scientific life.

  Highly criticized: + or - assessment of citations; depends on the field; citation window too narrow; self-citations influence; equal weight to each quotation whatever its origin, etc.

Assessing and ranking journals

- **Impact Factor** (Garfield, 1972): average number of annual citations received / published article in the last 2 years. Impacts scientific life.

  Highly criticized: + or - assessment of citations; depends on the field; citation window too narrow; self-citations influence; equal weight to each quotation whatever its origin, etc.

$\Rightarrow$ Alternatives: increase the citation window, standardize by field, etc...
Take into account the importance of citing sources:

- Group lasso, stochastic bloc models, etc. (Varin et al., 2016)
- **Google PageRank algorithms** (Waltman et van Eck, 2010) :
  - Prestige Scimago Journal Rank (PSJR)
  - EigenFactor (EIFA): excludes self-citations

  $\Rightarrow$ Widely use: ease of computation.
  $\Rightarrow$ Lacks a probabilistic model.

PageRank influence scores

Table: $C \in \mathbb{N}^{N \times N}$ cross-citation: citing (issuing ref) in rows and cited (receiving cit) in cols. c_{ij}: number of times i, in a given year, quotes articles published by j over a previous period of time (5 years).

		AmS	AISM	AoS	ANZS	Bern
1	AmS	43	1	2	0	0
2	AISM	0	18	3	3	5
3	AoS	9	24	291	4	53
4	ANZS	0	5	2	5	0
5	Bern	1	7	27	0	22
6	BioJ	0	0	3	2	0

Adjacency matrix P $p_{ij} := (\frac{c_{ij}}{c_{i+}})$: weighted oriented network citing→cited.
PageRank: "random surfer" of Google, teleportation matrix $\pi = \frac{1}{N}\mathbf{1}_N$

$$G = \alpha P + (1 - \alpha)\mathbf{1}_N \pi^\top. \tag{1}$$

Discrete-time, irreducible & aperiodic Markov chain between the N nodes. The parameter α damping factor a 0.85.

PageRank influence scores

PageRank (Brin and Page, 1998)

$$r_j^{\ell+1} = \sum_{i=1}^{N} g_{ij} r_i^{\ell} \ , j = 1, \ldots, N. \tag{2}$$

$\Rightarrow$ A quotation from JRRS-B does not have the same weight than others.
Solution: eigenvector r with unit norm, $r^\top = r^\top G$.

PageRank influence scores

PageRank (Brin and Page, 1998)

$$r_j^{\ell+1} = \sum_{i=1}^{N} g_{ij} r_i^{\ell} \ , j = 1, \ldots, N. \tag{2}$$

$\Rightarrow$ A quotation from JRRS-B does not have the same weight than others.
Solution: eigenvector r with unit norm, $r^\top = r^\top G$.

EIFA Bergstrom (2007): self-citations excluded ($c_{ii} = 0, p_{ii} = 0, \forall i$);
$\pi = \left(\frac{a_i}{a_+}\right)$: freq of references published by each journal in 5 years.
r eigenvector of G. $\tilde{r} = P^\top r$ and $r^\star = \frac{\tilde{r}}{\mathbf{1}_N^\top \tilde{r}}$. This is equivalent to

$$r^\star = \frac{r - (1-\alpha)\pi}{\alpha}. \tag{3}$$

$\Rightarrow$ EIFA introduces the "teleportation" part and then attenuate its effect.

A Bayesian Dirichlet-multinomial model without diagonal

- Model
- Empirical Bayes

A Bayesian Dirichlet-multinomial model without diagonal

Let $\underline{C}_i^\top := (c_{ij})$: the row i of C without its diagonal elements. $\underline{C}_{(N\times N-1)}$

1. Multinomial sampling of the elements of $\underline{C}_i^\top$

$$\underline{C}_i^\top | \underline{\theta}_i^\top \sim \mathcal{M}(n_i, \underline{\theta}_i^\top)$$

$n_i = \sum_{j\neq i} c_{ij}$ and $\underline{\theta}_i = (\theta_{i1}, \ldots, \theta_{i,i-1}, \theta_{i,i+1}, \ldots, \theta_{i,N})^\top$.

2. Dirichlet prior distributions

$$\underline{\theta}_i^\top | \gamma_{\setminus i} \sim \mathcal{D}(\gamma_{\setminus i}^\top)$$

where $\gamma_{\setminus i}^\top = (\gamma_1, \gamma_2, \ldots, \gamma_{i-1}, \gamma_{i+1}, \ldots, \gamma_N)$.

$\Rightarrow$ Posterior distributions are Dirichlet:

$$\underline{\theta}_i^\top | \gamma_{\setminus i}, \underline{C}_i^\top \sim \mathcal{D}(\underline{C}_i^\top + \gamma_{\setminus i}^\top), \ i = 1, \ldots, N.$$

A Bayesian Dirichlet-multinomial model without diagonal

$$
\begin{aligned}
E_{\text{post}}(\underline{\theta_{ij}}) &= \frac{c_{ij} + \gamma_j}{\sum_{j \neq i} c_{ij} + \sum_{j \neq i} \gamma_j}, \text{ for } j \neq i, i = 1, \ldots, N. \\
G_i^{\star\top} &= \alpha_i p_i^\top + (1 - \alpha_i)(\pi_i^\star)^\top
\end{aligned}
$$

Parametrization with prior expectation $E(\underline{\theta}_i^\top) = (\pi_i^\star)^\top$: prior proba that i cites any other journals. $\pi_i^\star = \frac{\gamma_{\setminus i}^\top}{K_{\setminus i}}$ with $K_{\setminus i} = K - \gamma_i$, $K = \sum_{i=1}^N \gamma_i$.
$K_{\setminus i}$, concentration: total number of "fictive" quotes given by i to others.

$\Rightarrow$ Same form as EIFA but with shrinkage factor no longer fixed:

$$
\alpha_i = \frac{n_i}{n_i + K_{\setminus i}},
$$

$0 \leq \alpha_i \leq 1$: large when n_i large & $K_{\setminus i}$ small (large prior variance).

Empirical Bayes approach

$$\begin{aligned}\mathcal{L}(\underline{C}|\gamma) &= \Pi_{i=1}^{N}\mathcal{L}(\underline{C}_i^\top|\gamma_{\setminus i})\\ \mathcal{L}(\underline{C}_i^\top|\gamma_{\setminus i}) &= \int p(\underline{C}_i^\top|\underline{\theta}_i^\top)p(\underline{\theta}_i^\top|\gamma_{\setminus i})d\underline{\theta}_i^\top\end{aligned}$$

Multinomial component: $p(\underline{C}_i^\top|\underline{\theta}_i^\top) = \frac{n_i!}{\Pi_{j\neq i}c_{ij}!}\Pi_{j\neq i}\theta_{ij}^{c_{ij}}$

Dirichlet component: $p(\underline{\theta}_i^\top|\gamma_{\setminus i}) = \frac{\Gamma\left(\sum_{j\neq i}\gamma_j\right)}{\Pi_{j\neq i}\Gamma(\gamma_j)}\Pi_{j\neq i}\theta_{ij}^{\gamma_j-1}$

$$\mathcal{L}_i(\underline{C}_i^\top|\gamma_{\setminus i}) = \frac{n_i!\Gamma\left(\sum_{j\neq i}\gamma_j\right)}{\Pi_{j\neq i}c_{ij}!\Gamma\left(\sum_{j\neq i}(c_{ij}+\gamma_j)\right)}\Pi_{j\neq i}\frac{\Gamma(c_{ij}+\gamma_j)}{\Gamma(\gamma_j)}.$$

$\Rightarrow$ Clear distinction between structural and sampling zeroes.
Multinomial part: no impact, it multiplies the likelihood by 1.
Dirichlet part: impact, $\underline{\theta}_i$ of size $(N-1)$ instead of N.

Fixed point algorithm

The log-likelihood can be written as:

$$L_i(\gamma_{\setminus i}) = \log\Gamma(\mathrm{K}_{\setminus \mathrm{i}}) - \log\Gamma(\mathrm{n_i} + \mathrm{K}_{\setminus \mathrm{i}}) + \sum_{\mathrm{j} \neq \mathrm{i}} \left[\log\Gamma(\mathrm{c_{ij}} + \gamma_\mathrm{j}) - \log\Gamma(\gamma_\mathrm{j})\right], \text{and}$$

its gradient can be written as:

$$\frac{d\mathcal{L}_i(\underline{C}_i^\top | \gamma_{\setminus i})}{d\gamma_j} = \psi(K_{\setminus i}) + \psi(n_i + K_{\setminus i}) + \psi(c_{ij} + \gamma_j) - \psi(\gamma_j), \text{ for all } i \neq j$$

- First order algorithms: Majorize-Minimize algorithm.
  $\Rightarrow$ Fixed point iteration algorithm (Minka, 2012)

$$\gamma_j^{\ell+1} = \gamma_j^\ell \frac{\sum_{i \neq j} \psi(c_{ij} + \gamma_j^\ell) - (N-1)\psi(\gamma_j^\ell)}{\sum_{i \neq j} \psi(n_i + K_{\setminus i}^\ell) - \psi(K_{\setminus i}^\ell)}.$$

- Second order algorithms: Levenberg-Marquardt, EM variant.

Ranking statistical journals

Statistical journals

47 statistical journals (Varin, 2016), citations published in 2010 related to articles published from 2001 to 2010.

- EBEF $\alpha_i = 0.95$ for CSDA, STMED $\alpha_i = 0.39$ STATAJ (mean 0.77).
  $\Rightarrow$ Teleportation decreasing with the number of references emitted.
- EIFA

- 2 scores with self-citations:
  - EBPR (Empirical Bayes Page Rank): Multinomial-Dirichlet with diag
  - Prestige Scimago Journal Rank (PSJR) (Scimago Lab, Scopus)

$$G_2 = \alpha_2 P + (1 - \alpha_2 - \beta)\mathbf{1}\pi^\top + \beta\frac{\mathbf{1}\mathbf{1}^\top}{N},$$

$\pi := a_i/a_+$, $\alpha_2 = 0.90$, $\beta = 10^{-4}$. First eigenvector: $G_2^\top r_2 = r_2$.
self-citations restricted to 33%

Total scores & rankings

	Journal	PSJR	EBPR	EIFA	EBEF
1	JASA	117.84	132.37	127.10	127.28
2	AOS	103.01	116.12	96.37	97.17
3	JRSS-B	70.79	79.49	78.91	79.91
4	BKA	61.37	68.47	71.98	72.97
5	BCS	63.58	66.49	67.23	63.81
6	STMED	70.04	64.75	58.71	51.27
7	JSPI	41.23	40.59	41.58	42.39
8	CSDA	43.29	42.33	39.11	38.44
9	STSIN	31.51	29.69	33.05	34.45
10	JMA	27.75	29.22	29.55	30.27
11	BIOST	27.26	24.83	27.58	26.65
12	JCGS	21.69	23.22	25.19	25.10
13	SPL	27.40	23.91	23.17	23.75
14	SJS	17.89	19.73	22.83	23.47
15	STSCI	18.57	20.85	23.03	23.26
16	BERN	13.75	14.52	15.98	16.31
17	CJS	10.73	11.84	13.17	13.79
18	STCMP	11.67	11.79	12.82	13.42
19	BIOJ	11.84	11.03	12.65	12.14
20	TECH	11.06	11.57	11.37	11.80
21	CSTM	12.67	10.75	12.18	11.61
22	JRSS-C	9.90	9.53	10.74	11.08
23	TEST	7.54	8.48	9.62	10.27
24	JRSS-A	10.55	10.22	9.83	9.81
25	AISM	8.72	8.64	9.61	9.79
26	AMS	9.92	9.73	9.50	9.62
27	JNS	9.50	7.23	8.06	8.54
28	LTA	7.43	7.56	8.71	8.52
29	JSCS	7.02	6.45	6.86	7.31
30	ENVR	7.50	6.80	7.30	7.26
31	SMMR	12.39	5.81	7.02	6.54

	EBPR	PSJR	EIFA	EBEF
1	JASA	JASA	JASA	JASA
2	AOS	AOS	AOS	AOS
3	JRSS-B	JRSS-B	JRSS-B	JRSS-B
4	STMED	BKA	BKA	BKA
5	BCS	BCS	BCS	BCS
6	BKA	STMED	STMED	STMED
7	CSDA	CSDA	JSPI	JSPI
8	JSPI	JSPI	CSDA	CSDA
9	STSIN	STSIN	STSIN	STSIN
10	JMA	JMA	JMA	JMA
11	SPL	BIOST	BIOST	BIOST
12	BIOST	SPL	JCGS	JCGS
13	JCGS	JCGS	SPL	SPL
14	STSCI	STSCI	STSCI	SJS
15	SJS	SJS	SJS	STSCI
16	JSS	BERN	BERN	BERN
17	BERN	CJS	CJS	CJS
18	CSTM	STCMP	STCMP	STCMP
19	SMMR	TECH	BIOJ	BIOJ
20	BIOJ	BIOJ	CSTM	TECH
21	STCMP	CSTM	TECH	CSTM
22	TECH	JRSS-A	JRSS-C	JRSS-C
23	EES	AMS	JRSS-A	TEST
24	CJS	JRSS-C	TEST	JRSS-A
25	JRSS-A	AISM	AISM	AISM
26	JBS	TEST	AMS	AMS
27	AMS	LTA	LTA	JNS
28	JRSS-C	JNS	JNS	LTA
29	JNS	ENVR	ENVR	JSCS
30	AISM	JSCS	SMMR	ENVR
31	TEST	JSS	JSCS	SMMR

Articles scores & rankings

	Journal	PSJR	EBPR	EIFA	EBEF
1	JRSS-B	7.77	8.72	8.66	8.77
2	STSCI	5.05	5.67	6.26	6.32
3	JASA	4.98	5.60	5.38	5.38
4	AOS	5.30	5.97	4.96	5
5	BKA	3.85	4.30	4.52	4.58
6	SJS	2.01	2.22	2.57	2.64
7	JCGS	2.23	2.39	2.59	2.58
8	BCS	2.52	2.63	2.66	2.53
9	TEST	1.61	1.81	2.06	2.19
10	CJS	1.64	1.80	2.01	2.10
11	BERN	1.44	1.52	1.67	1.71
12	TECH	1.46	1.52	1.50	1.55
13	LTA	1.34	1.36	1.57	1.53
14	JRSS-A	1.46	1.42	1.36	1.36
15	JMA	1.21	1.27	1.29	1.32
16	STMOD	0.93	0.96	1.10	1.14
17	AMS	1.07	1.05	1.03	1.04
18	ISR	1	0.99	0.95	1.03
19	AISM	0.85	0.85	0.94	0.96
20	STMED	1.26	1.16	1.05	0.92
21	STCMP	0.77	0.78	0.85	0.89
22	CSDA	0.92	0.90	0.83	0.81
23	JSPI	0.77	0.76	0.78	0.79
24	ANZS	0.60	0.61	0.67	0.73
25	BIOJ	0.71	0.66	0.75	0.72
26	JRSS-C	0.61	0.58	0.66	0.68
27	STNEE	0.54	0.56	0.59	0.67
28	STATS	0.55	0.54	0.61	0.66
29	BIOST	0.66	0.60	0.67	0.65
30	JTSA	0.68	0.61	0.62	0.63
31	ENVR	0.65	0.59	0.63	0.63

	EBPR	PSJR	EIFA	EBEF
1	JRSS-B	JRSS-B	JRSS-B	JRSS-B
2	AOS	AOS	STSCI	STSCI
3	STSCI	STSCI	JASA	JASA
4	JASA	JASA	AOS	AOS
5	BKA	BKA	BKA	BKA
6	BCS	BCS	BCS	SJS
7	JCGS	JCGS	JCGS	JCGS
8	SJS	SJS	SJS	BCS
9	CJS	TEST	TEST	TEST
10	TEST	CJS	CJS	CJS
11	JRSS-A	TECH	BERN	BERN
12	TECH	BERN	LTA	TECH
13	BERN	JRSS-A	TECH	LTA
14	LTA	LTA	JRSS-A	JRSS-A
15	STMED	JMA	JMA	JMA
16	JMA	STMED	STMOD	STMOD
17	AMS	AMS	STMED	AMS
18	ISR	ISR	AMS	ISR
19	STMOD	STMOD	ISR	AISM
20	CSDA	CSDA	AISM	STMED
21	AISM	AISM	STCMP	STCMP
22	STCMP	STCMP	CSDA	CSDA
23	JSPI	JSPI	JSPI	JSPI
24	BIOJ	BIOJ	BIOJ	ANZS
25	JTSA	ANZS	BIOST	BIOJ
26	BIOST	JTSA	ANZS	JRSS-C
27	ENVR	BIOST	JRSS-C	STNEE
28	JRSS-C	ENVR	ENVR	STATS
29	ANZS	JRSS-C	JTSA	BIOST
30	STATAJ	STNEE	STATS	JTSA
31	STATS	STATS	STNEE	ENVR

Discussion

- EBEF for EigenFactor with Dirichlet-multinomial model (cor 0.90)
- Teleportation varies from one journal to another

$\Rightarrow$ Simplicity - Quick (6s/ 35s second order algo)
$\Rightarrow$ Taking into account the zeros: exclusion of a specific field.

Discussion

- EBEF for EigenFactor with Dirichlet-multinomial model (cor 0.90)
- Teleportation varies from one journal to another

$\Rightarrow$ Simplicity - Quick (6s/ 35s second order algo)
$\Rightarrow$ Taking into account the zeros: exclusion of a specific field.

$\Rightarrow$ Underweight self-citations in a data-driven way? (PSJR 33%).

total received (R) / total made (M): $S_i = \frac{c_{+i}}{c_{i+}} = \frac{c_{ii}+R_i}{c_{ii}+M_i}$. Let $\kappa \in [0,1]$

$$S_i(\kappa) = \frac{\kappa c_{ii} + R_i}{\kappa c_{ii} + M_i}.$$

- if $S_i(0) < 1$, $S_i(\kappa)$ increasing function of κ upper bounded by 1.
- if $S_i(0) = 1$, $S_i(\kappa) = 1$ for any κ.
- if $S_i(0) > 1$, $S_i(\kappa)$ decreasing function of κ lower bounded by 1.

$\Rightarrow$ Good journals no interest in self-citations: $\kappa_i = \min\left(\frac{\min(R_i,M_i)}{c_{ii}}, 1\right)$