Analysis of Citation Patterns in Society of Statistics

Minshi Peng
Advisor: Jiashun Jin(Carnegie Mellon University),
Pengsheng Ji(University of Georgia)

1 Introduction

In the past decades, the scientific community has grown substantially: we have way more researchers and annual publications than we ever had before. The network of citations to published papers has long been recognized to provide important information to researchers and librarians Gross and Gross (1927), but gained rapid popularity following the creation of the Science Citation Index in 1961 by the Institute for Scientific Information Smith (1981). This pioneering database was conceived and created by Eugene Garfield Garfield (1955), followed by the invention of journal impact factors to rank journals by average annual citations to papers published in the last 2 years Garfield et al. (1972); Garfield (2006).

In recent decades, the field of citation analysis has expanded rapidly. Due to the increasing availability of computational resources, it is possible to build large bibliographic data set and analyze them to an unprecedented level of accuracy. Redner Redner (2005) performed an analysis over all papers published in the 110 years long history of journals of the American Physical Society (APS), the citation data cover 353,268 papers and 3,110,839 citations. Another well-studied dataset *CiteSeer* contains over 800,000 research papers in computer science published by over 2 million authors. Some patent-citation data has also been collected and cleaned, like *cit-patents*Leskovec et al. (2005), which contains 3,774,768 papers and 16,518,948 citations.

Studies on these large bibliographic data sets cast light on the internal structure of the research communities, and provide us with a better understanding of the historical roots and development of those research fields. Though statisticians contributed a lot in studying those bibliographic data, however, till now few work has been done to study the structure of statistics society, that is the purpose of my project. Our data sets covers about $80,000$ research papers published in 36 journals in statistics from 1976 to the first half of 2015, consisting of titles, authors and affiliations, abstracts, MSC numbers and keywords, etc. The Phase I of the data, containing papers published in the top 4 journals in statistics in last ten years, has recently become publicJi and Jin (2017). The work has drawn unexpected attentions and was considered as a great step forward to provide the community with a first such data set for self-study. The Phase II data is a step further, with the increased breadth and width, opens many potential more interesting questions.

Citation networks are compact dynamic representations of the relationships between research products. they offer a fertile ground for studying research and collaboration patterns of scientific communities. One key question in citation analysis is what makes some papers highly cited while others remain uncited. In addition to the quality of the manuscript, other factors play a key role: more citations are received by longer papers with more authors, review articles, papers resulting from international collaborations, those published in journals with high-impact factors and papers that are relevant to multiple fields, as well as the "first-mover" effect, that the first papers in a field will, essentially regardless of content, receive citations at a rate enormously higher than papers published later Padial et al. (2010); Wuchty et al. (2007); Newman (2009). However, such analyses have not been conducted for the Statistics literature. With the available of the second-stage data, we can explore the diversity of the citation patterns, investigate the characteristics of the highly cited papers, then analyze to provide insights into trends over time, and into the key factors that result in higher citation counts. A better understanding of such citation features is useful in many perspectives: it may help administrators or

30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

funding agencies to prioritize research areas, and researchers to start a new topic or a publish a new paper.

2 Exploratory Data Analysis

Our data set is the second-stage data followed from Ji and Jin (2017), collected by Frof. Pengsheng Ji form University of Georgia, Prof. Jiashun Jin from Carnegie Mellon University and Frof. Tracy Ke from University of Chicago. The dataset covers papers from 36 journals that were published between 1976 and 2014 and half of 2015. The total number of papers is 80989. After removing the papers with missing information(those are letters to editors, book reviews and paper discussion) and the duplicated records, there are in total 71824 papers. From the reference list of the papers in terms of MR number, a link is formed between two papers in our data set (390239 citations).

2.1 Journal

The information of 36 journals covered in this data set is shown in table 1. These journals contains the major influential Statistical journals and some closely related important journals in other disciplines, for example, Journal of Machine Learning Research.

We can observe that not all the journals were established in 1976, instead, only 14 out of 36 journals existed in 1976, while the other journals came into appear between $1976 - 2014$. Figure 1 left show the number of journal at each year. Generally the journals that appeared later contribute fewer papers in our dataset, as is shown in Figure 1 right.

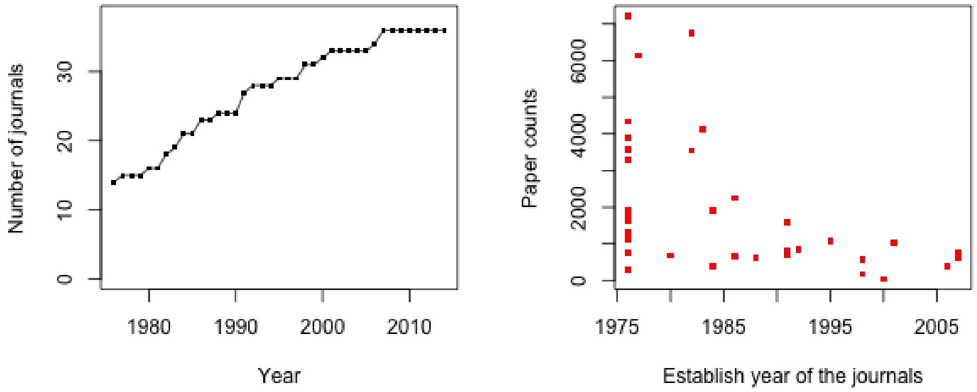

Figure 1: Left: The number of journals out of 36 total journals established before a given year. Right: the number of papers in the journal(within data set) versus the establish year of the journal (journals established before 1976 are shown on the left margin).

2.2 Author

Each paper is authored by one or co-authored by several people. The author name is cleaned at the stage of data collection so that each author has an unique authorID. The number of authors covered in the data set is 39616. Among them 21519 has only one paper and 32478 have less than 5 papers. The maximum count of publications belonging to a single author is 440, form HALL, PETER. We list the authors who have more than 100 papers in our data set in table 2.

Table 1: 36 journals covered in our data set(in shorhand)

Shortcut	Full name	issn	Year	# papers
Poincare	Ann. Inst. Henri Poincare Probab. Stat	0246-0203	1964	1163
AoAS	Annals of Applied Statistics	1932-6157, 1941-7330	2007	785
AoP	Annals of Probability	0091-1798, 2168-894X	1973	3278
AoS	Annals of Statistics	0090-5364, 2168-8966	1973	4341
AISM	Annals of the Institute of Statistical Mathematics	0020-3157,1572-9052	1949	1939
AuNZJS	Australian & New Zealand Journal of Statistics	1369-1473, 1467-842X	1998	564
BayN	Bayesian Analysis	1931-6690, 1936-0975	2006	390
Bern	Bernoulli	1350-7265, 1573-9759	1995	1073
Bmtrcs	Biometrics	0006-341X, 1541-0420	1945	1759
Bmtrk	Biometrika	0006-3444, 1464-3510	1901	3310
Biost	Biostatistics	1465-4644, 1468-4357	2000	45
CaJS	Canadian Journal of Statistics	0319-5724,(1708-945X)	1973	1133
ComSTM	Communications in Statistics-Theory and Methods	0361-0926,1532-415X	1976	7208
CSDA	Computational Statistics & Data Analysis	0167-9473, 1872-7352	1983	4132
EJS	Electronic Journal of Statistics	1935-7524	2007	627
Extrem	Extremes	1386-1999,1572-915X	1998-2015	177
ISRev	International Statistical Review	0306-7734,(1751-5823)	1972	288
JCGS	Journal of Computational and Graphical Statistics	1061-8600,1537-2715	1992	851
JMLR	Journal of Machine Learning Research	1532-4435, 1533-7928	2001	1036
JASA	Journal of the American Statistical Association	0162-1459, 1537-274X	1922	3899
JRSS-B	Journal of the Royal Statistical Society Series B-Statistical Methodology	1369-7412, 0035-9246, 1948 & 1988	1632	
JOAS	Journal of Applied Statistics	0266-4763,(1360-0532)	1984	1918
JClass	Journal of Classification	0176-4268,1432-1343	1984	377
JMVA	Journal of Multivariate Analysis	0047-259X	1971	3564
JRSS-A	Journal of the Royal Statistical Society Series A-Statistics in Society	0964-1998,(1467-985X)	1988	626
JRSS-C	Journal of the Royal Statistical Society Series C-Applied Statistics	0035-9254,1467-9876	1964	766
JSPI	Journal of Statistical Planning and Inference	0378-3758,(1873-1171)	1977	6129
JoTS	Journal of Time SeriesAnalysis	0143-9782, 1467-9892	1980	683
JoNpS	Journal of Nonparametric Statistics	1048-5252, 1026-7654	1991	827
PTRF	Probability Theory and Related Fields	0178-8051,1432-2064	1986	2255
StSci	Statistical Science	0883-4237, 2168-8745	1986	669
ScanJS	Scandinavian Journal of Statistics	0303-6898	1974	1328
Sinica	Statistica Sinica	1017-0405,1996-8507	1991	1574
StComp	Statistics and Computing	0960-3174, (1573-1375)	1991	679
SPLet	Statistics& Probability Letters	0167-7152	1982	6751
StMed	Statistics in Medicine	0277-6715, 1097-0258	1982	3549

Table 2: The authors that have more than 100 papers in the data set.

Author name	Count	Author name	count
HALL, PETER	440	BALAKRISHNAN, N.	294
DETTE, HOLGER	184	CARROLL, RAYMOND J.	173
SEN, PRANAB KUMAR	145	ZHU, LIXING	135
IBRAHIM, JOSEPH G.	130	HORVATH, LAJOS	125
MUKERJEE, RAHUL	124	FAN, JIANQING	122
GHOSH, MALAY	115	MARRON, J. S.	113
DUNSON, DAVID B.	110	WALKER, STEPHEN G	108
MULLER, HANS-GEORG	106	MOLENBERGHS, GEERT	103

Table 3: The "top" highly cited authors.

Author name	Count	Author name	count
HALL, PETER	6428	FAN, JIANQING	5357
RUBIN, DONALD B.	4517	TIBSHIRANI, ROBERT	4032
EFRON, BRADLEY	3487	CARROLL, RAYMOND J.	3299
HASTIE, TREVOR	2765	JOHNSTONE, IAIN M.	2547
LAIRD, NAN M.	2287	MARRON, J. S.	2225
WEI, L. J.	2225	YING, ZHILIANG	2209
MULLER, HANS-GEORG	2087	DONOHO, DAVID	1989

Here the number of publication can be considered as the out-degree, while the number of citation from other papers received by the paper of the author is the in-degree. Table 3 presents the "top 10" highly cited authors. Both the out-degree and in-degree distribution are highly-skewed, where 10% of the authors account for nearly 60% of the publications and more than 80% of the citations, as is shown by the two Lorenz cure in Figure 2 left. Intuitively, the number of publication and the number of the citation are highly correlated. In fact from the scatter plot of two variable in logarithmic scale, it is demonstrated that two variables forms a strong linear relationship. The variation is larger when the value of two variables are small. This is indeed reasonable since if a author has published a large amount of papers, the exceptional performance of few papers will not cause large deviation from the mean, and the highly cited works always belongs to the productive authors.

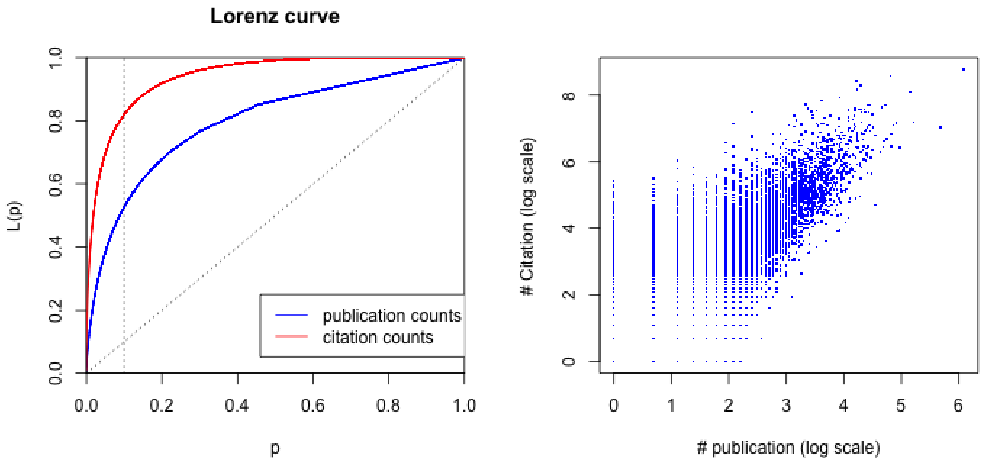

Figure 2: Left: The Lorenz curve of the distribution of publication counts (blue) and citation counts (red) of all 39616 authors. Right: the scatter plot the publication counts and citations in logarithmic scale.

2.3 Citation and reference count of papers

Since we have 40 years of data, the significance of earlier papers are essential indicated by the citations form the later papers. We are primarily interested in how the citation count of papers involve over the years.Figure 3 left presents the number of papers published in each year from 1976 to 2015. It is shown that the number kept increasing over the years, and increased most rapidly during 2000-2008. This is probably due to the growing interests in "Big Data". The abrupt decrease of the last bar is due to the half portion of data we have for 2015, thus the height of the bar is expected to be twice the amount, which is consistent with the trend of previous years. The middle figure shows the average number of citations received by papers published in each year, i.e. the y axis represents how many citations received by the papers published in this year till 2015, divided by the total number of papers published in this year. The general trend is that the number keeps going down with some fluctuations. This is reasonable since averagely older papers have more time to be discovered, explored and cited. The right figure is from dividing the citation count of each paper by the age of the paper, and reproduce the second plot. As the result the age effect is got rid off, and the average citation count per paper per year increases from 1976 to 2008. The drop followed is probably due to the edge effect, since there will be very few citations within a few year after publication.

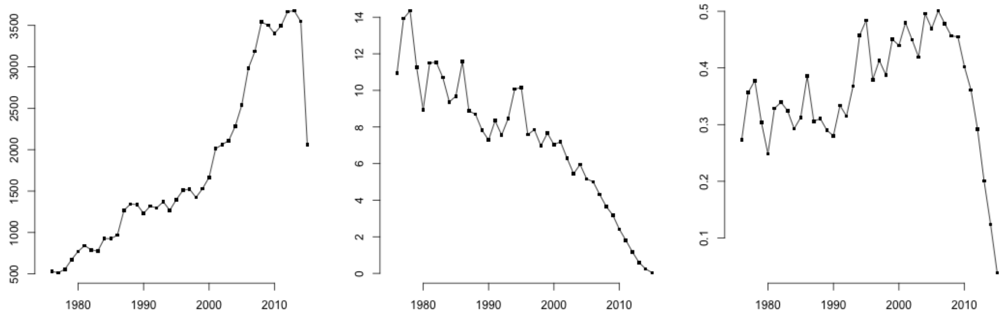

Figure 3: Left: number of papers published in each year from 1976 to 2015. Middle: The average number of citations for papers published in each year till 2015. Right: divide the total citation of each paper by the age and reproduce second figure.

From the view of network analysis, the references of each paper are the "out-edges" to the existed papers, while the citations received from later papers are the "in-edges". We want to investigate how the out-degree and in-degree are distributed over all papers in the dataset. First consider the out-degree, i.e. the reference counts. Figure 4 Left shows the average number of reference counts of papers in each two-year group. Since we only count the reference if the referred paper is in our dataset, the number of reference is restricted by the total number of papers at the year of publication. Therefore we adjust the number of reference by the publication year (divide the publication year minus 1975 and multiplied by 40). The adjusted reference count is shown in middle. The left panel shows the histogram of the adjusted reference count of all papers in logarithmic scale.

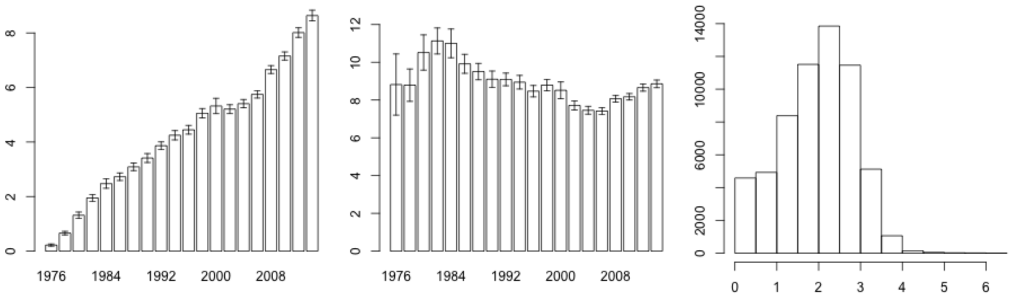

Figure 4: Left: The average number of references of papers published in each two-year group. Middle: the adjusted reference count by the year of publication. Right: the histogram of the adjusted reference count (in logarithmic scale).

Next consider the distribution of the in-degree (the number of citations received by each paper). For the 71824 papers in our data sets, the average citation per paper is 5.43. Among these papers (a) 25652 (36%) are not cited by any other paper in the data set, (b) 10699 (15%) do not cite any other paper in the data set, and (c) 5186 (7%) neither cite nor are cited by any other papers in the data set. Since we have a 40 years' long time span (from 1976 to 2015), and the citations of paper published in deferent years are not comparable, as is shown in Figure 2. Therefore we cut the 40 years into 8 five-year period and consider the citation distribution for papers published in each group, so as to remove the impact of age of a paper has on its citation count. Further, since the papers published between 2010 and 2015 don't have sufficient time to be cited, therefore we exclude them from this part of analysis.

The distribution of the citations in 7 groups show similar patterns. They are all highly skewed. For example, the top 5% highly cited papers receive about 45% of all citation counts, and the top 10% received nearly 60% of the citation counts. The mean of the Gini coefficient is 0.72 with the highest 0.76 and lowest 0.71, suggesting that the in=degree is highly dispersed.

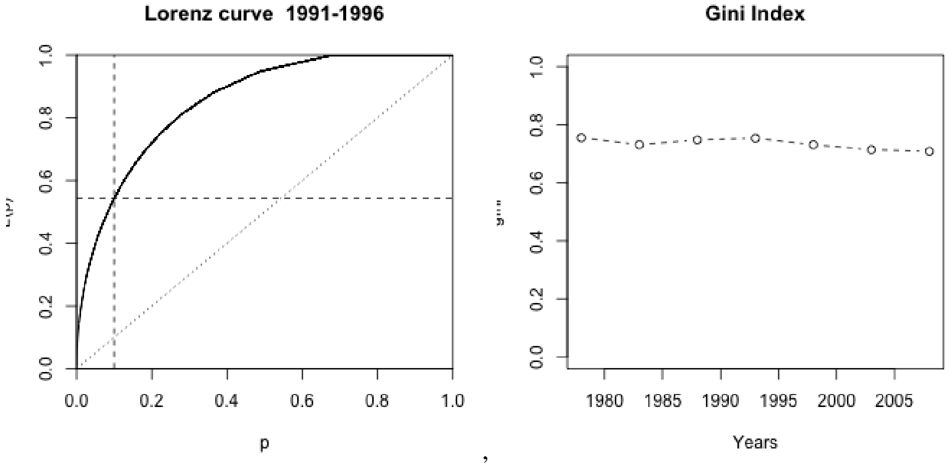

Figure 5: Left: The Lorenz curve for the number of citation received by each paper published between 1991-1996. Right: the Gini coefficient of the citation counts for each of the 7 groups.

Table 4 presents the top three high-cited papers in each five-year group, as well there citation patterns. add more explanation on this table.

Table 4: The top there high-cited papers in five-year group

Paper	keywords	Journal
Dempster et al. (1977)	Maximum likelihood, EM	JRSS-B
Schwarz et al. (1978)	Dimension, (AIC)	AOS
Efron (1992)	Bootstrap	
Andersen and Gill (1982)	(Censoring; survival analysis)	AOS
Louis (1982)	EM	JRSS-B
Rosenbaum and Rubin (1983)	(Covariance adjustment, Discriminant matching)	Bmtrk
Liang and Zeger (1986)	Longitudinal data, Generalized linear model	Bmtrk
Gelfand and Smith (1990)	Sampling, (Conditional probability structure)	JASA
Owen (1988)	Empirical likelihood ratio. Confidence interval	Bmtrk
Benjamini and Hochberg (1995)	FDR, Multiple testing	JRSS-B
Green (1995)	MCMC, Bayesian model	Bmtrk
Donoho and Johnstone (1994)	Wavelet (Minimax)	Bmtrk
Tibshirani (1996)	Lasso	JRSS-B
Eilers and Marx (1996)	GLM, smoothing, splines	StSci
Richardson and Green (1997)	(Birth and death process), Mixtures, Bayesian	JRSS-B
Fan and Li (2001)	Variable selection, Penalize likelihood	JASA
Spiegelhalter et al. (2002)	Bayesian model comparision	JRSS-B
Efron et al. (2004)	(Lasso, boosting), linear regression	AOS
Zou (2006)	Adaptive Lasso	JASA
Meinshausen and Bühlmann (2006)	High-dim, (Lasso, graphical models)	AOS
Yuan and Lin (2006)	Grouped Lasso	JRSS-B

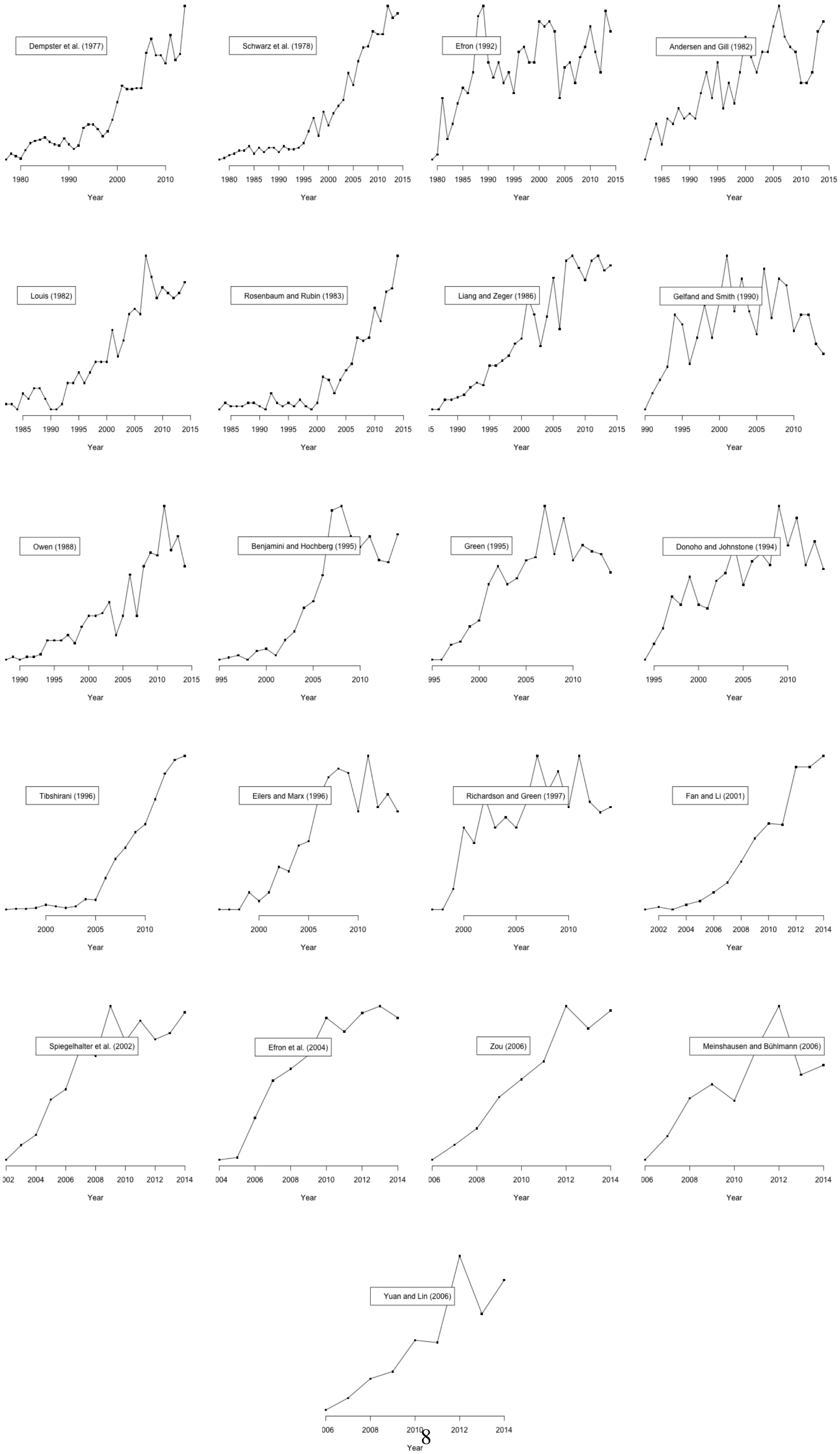

Figure 6: The citation pattern of papers in table 4

2.4 How variables correlated with citations

2.4.1 Keywords

We want to investigate whether the topics of the papers are associated with the number of citations received, and how dose the relation change over time. We measure the association by taking correlation of the existence of keywords in the title with the citation counts the papers receive in each period. This is because in univariate regression, the correlation is equivalent to the slope. We consider the citations appeared in a sequence of two-year period from 1980 to 2013, so that we could measure how the correlations evolved over time.

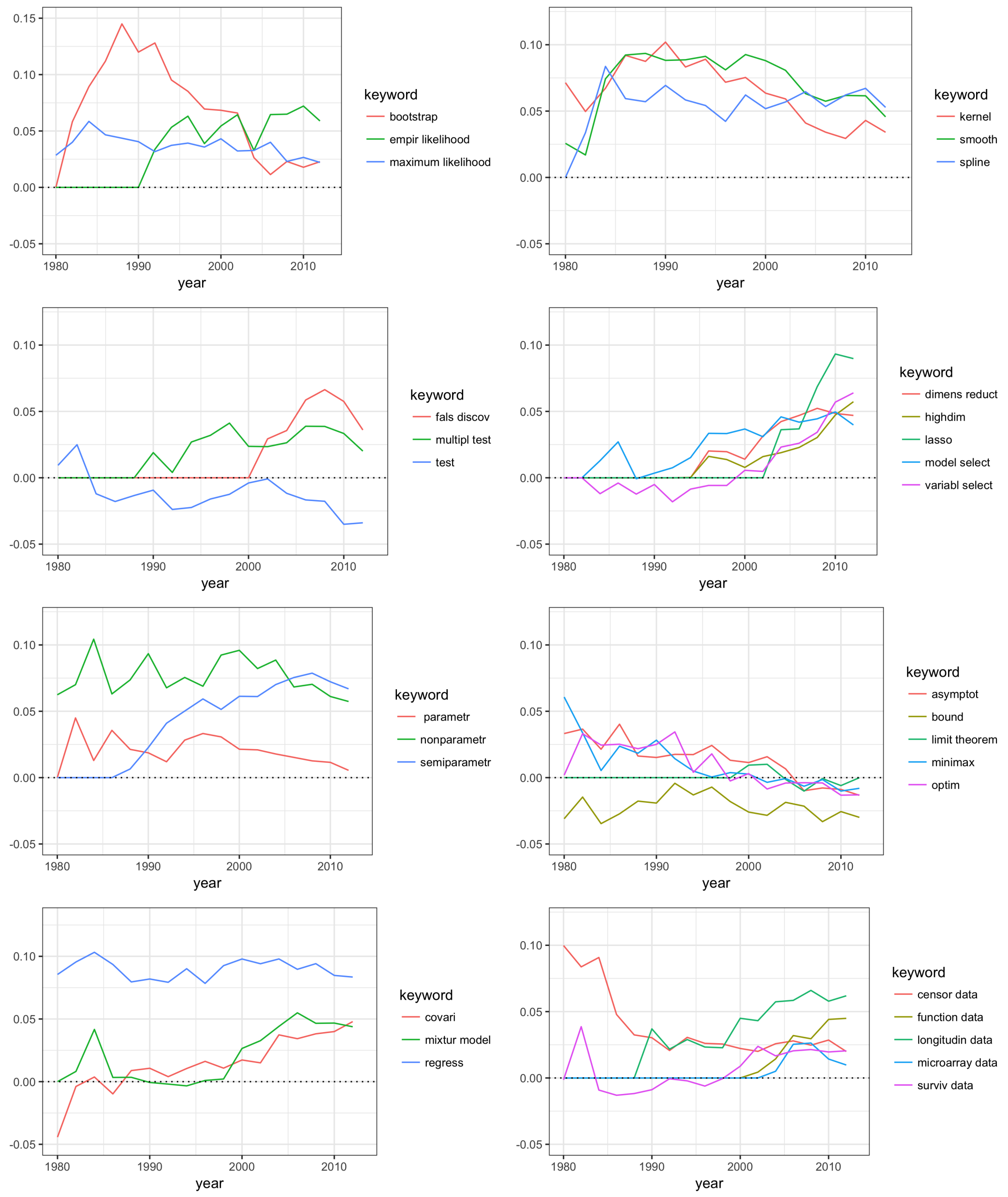

Figure 7: The correlation of existence of a keyword in the title with the number of citations the papers received in each two-year period, i.e., how much popularity of paper that can be explained by the topics.

Figure 7 presents some representative topics that had non-trivial correlations with the citation counts. The keywords in the first four figures covers some hot topics in Statistics. We can clearly find some trend in these patterns, and this, to some extent, concords with the popularity of the these areas in Statistics over the years. For example, the pattern of keyword 'bootstrap' in the first figure shows that in 1990 papers about 'bootstrap' were averagely receiving higher citations, which agreed with the mainstream research interest at that time. The second figure shows that the 'smoothing'-type of topic was most popular and highly cited in 1990s and the positive impact continued to recent years, though slightly weakened. Also , the pattern for these three keywords are highly-similar, which might due the high relevance of the content. The third figure shows the popularity pattern of multiple testing, which were highly-cited around 2000s, promoted with tide of FDR. In contrast, the general testing becomes less popular as time goes on. The forth figure shows some of the most popular topics for the last ten years, like 'high dimensional', 'lasso', 'model selection' and 'variable selection'. The fifth figure shows the impact on citation by which class of the papers belongs to, parametric, nonparametric or semi-parametric. Averagely nonparametric papers were associated with higher citations than the parametric papers, and the semi-parametric papers seems to be a rising star, which was becoming popular. We can possibly guess this topic would becomes more popular in the near future. The sixth figure shows that some theoretical topics resulted in comparatively lower citations. This seems to be counterintuitive at first, since most of these papers were written by renowned statisticians. But it also reasonable, since the theoretic papers always serve as 'topic killer', because it always marks the end of this type of problem, with little opening for future work. Plot 7 also shows some popular problems, for example, those related to 'covariance' and 'mixture model', which have the potential of getting more and more citations. The last plot are the trend of interest on different type of data. We can observe that the censor data were a hot choice in the 1980s, while the functional, longitudinal and genetic data are becoming popular recently. By analyze the correlation of citations and the keyword, we can gain some insight in how the research interests in statistics society changed over the years.

2.4.2 Number of authors

We can observe that there are many papers that have more than one authors. Indeed, in academia the collaboration is very common since each scholar has his/her strength and weakness. Therefore the number of authors of the paper serves as an potential factor that determine the citations of the paper. So we are interested in what proportion of the papers having one, two, three or four and more authors, and are there any difference between highly-cited papers and averagely-cited papers. Since the co-authorship might have been changing over time, we cut the 40-years period from 1976 to 2015 as 10 equal time segments. So papers published in each time segments can be view as appearing at the same time, and therefore their citation counts are comparable. First we calculate the proportion of different co-authorship for papers published in each time period. The result is shown in left panel of figure 8. We can observe that over the years, the proportion of single-authored papers kept decreasing, while the proportion of papers having two, three or four and more authors were all increasing, and among them the proportion of three-author papers increased comparatively faster.

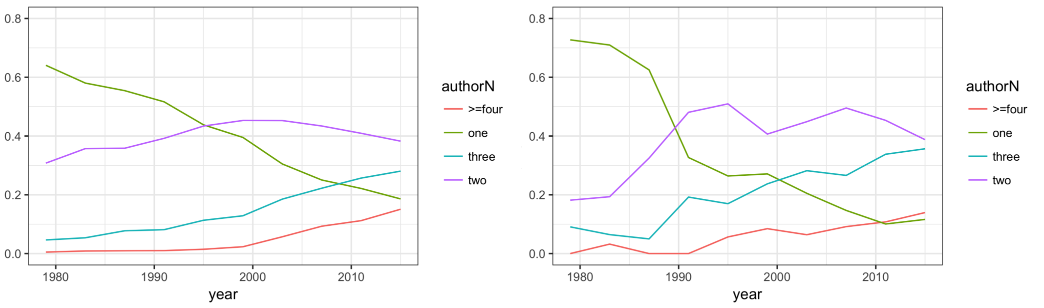

Figure 8: Left: the proportion of all papers having one, two, three or four and more authors, for each four-year group from 1976 to 2015. Right: the proportion of each number of authors for the top 1% highly-cited papers.

1976-1979	1980-1983	1984-1987	1988-1991	1992-1995
22	30	40	52	53
1996-1999	2000-2003	2004-2007	2008-2011	2012-2015
59	78	109	139	129

Table 5: The number of papers having top 1% citation counts among papers in each four-year group.

Then we select the top 1% highly-cited papers during each time period, and summarize the distribution of the number of authors as proportion.Table 5 shows the number of papers selected for each four-year period. Figure 8 shows the proportion of those papers having different number of authors. The general trend over the years follows the similar pattern, that the number of single-authored papers decreased sharply while the multiple-authored papers were gaining increasing popularity. However the difference is also evident between two patterns. The most prominent difference is the proportion of papers having four or more authors, which eventually became dominated among the top-cited papers, while it remained modest among the average papers. It seems the highly-cited papers prefer more authors, and this makes sense because creativity is often sparked by the collision of different ideas. We can probably conjecture that the proportion of multiple-authored papers would continue to increase, especially among the highly-cited papers.

2.4.3 Journal

Our data set covers papers in 36 journals in Statistics and some closely related fields. It is conceivable that papers published in more famous journals would probably receive more citations. However, though there exists various evaluation of different journals, there is no universally acknowledged ranking among them. In addition, the popularity of the journals keeps changing over the years. Therefore we want to investigate how the citation behavior of the papers is associated with the journals where they were published, and how it changes over time.

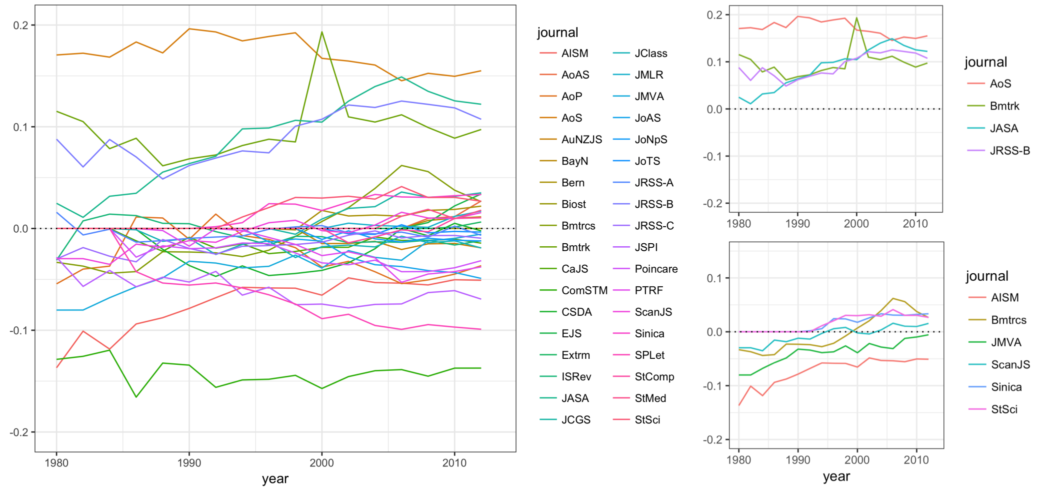

Figure 9: Left: The correlation of the citation counts with the 36 journals in each two-year period. The journal with higher correlation indicates the averagely more citations received by papers. Right upper: the "top" four journals in terms of association. Right bottom: examples of the journals with increasing correlation with citation counts.

Here again we cut the 40-years into 20 segments, and consider the citation counts received by the existing papers in each 2-year time segment separately. For each time segment, we calculate the correlation of the journal(binary variable) with the citation counts (in logarithmic scale) appeared in the two-years for the existing papers (papers published before this time segment). Each line in Figure 9 left represents the correlation pattern of a journal over 20 time periods. The journal with higher correlation indicate that the paper in this journal receive averagely higher citation counts in each two-year period. We can observe that the correlation pattern of most journals are around $y = 0$

line, and this is partly due to that some journals were established in latter years. The top right figure shows the correlation patterns of the four journals having apparent higher correlations than the rest of the journals. Among them the *Annals of Statistics* has an evident advantage in the early years, while *Journal of the American Statistical Association* has the fast-growing correlations over the years. The bottom right figure shows some example journals having the growing pattern.

3 Interesting problems

In recent decades, the field of citation analysis has expanded rapidly. Citations provide a key method of sorting through the enormous volume of scientific papers produced every year, to pinpoint the most important papers in a given topic. For example, in conducting a literature search for this paper, we used citation databases in the following ways: (i) identifying classic papers with the most citations, (ii) following the trail of papers cited by the classic papers to find the oldest papers in the field, (iii) reading the most recent published papers that cited classic papers and (iv) examining the references lists in all of these identified papers to select a further round of relevant papers to read. In this manner, we quickly identified key papers in the field of citation analysis.

Increasingly, citation rates are also being used to rank papers and scientists. The basic theory of citation analysis holds that citations give credit to colleagues for influence on their own work, and can therefore be used to measure the impact of a particular scientist Bornmann and Daniel (2008). Under this theory, scientists can be compared within a field through metrics such as the total citations to their work, the number of papers with more than 10 citations each (e.g. Google Scholar), or the increasingly popular h-index: where h of their papers have at least h citations each and the other papers have $\leq h$ citations Hirsch (2005). However, an alternative perspective arises suggesting that citations are not unalloyed measures of another scientist's worth, but a highly flawed measure reflecting complex motives for citing particular papers and not others. For instance, citations may be used to defend a body of research, promote your own interests and convince others that you are right, with the ultimate aim to gain a dominant position in your field Bornmann and Daniel (2008).

Regardless of the different motivation, there is certainly arbitrariness and bias in the choice of which references get cited. To list just a few factors, citation rates are influenced by a paper's accessibility, age, language, country, the size of the specialty and the journal in which it was published Smith (1981). From Figure 2 we have seen the intense unbalance of citations among all the papers, and this motivates us to analyze why some papers can receive far more citation than the others, what features the top-cited papers have in common, and what features are playing an essential role. Moreover, we observe that the distribution of cites for each paper over the year, i.e. the citation pattern, vary a lot from papers to papers. Even among the top-cited papers, some performed like 'the sleeping beauty', popular after years of being unnoticed, while some got large amount of citations soon after published. The existence of the diversity inspires us to study the mechanism of the citations and look into the potential driving factors.

3.1 What type of papers can get highly cited?

3.1.1 Label the highly cited papers

We are interested in how the features that are available at the time of publication are associated with the amount of citations received by the paper in later years, i.e. whether they would get highly-cited or not. As is can be seem from Figure 3, papers from different years are not comparable. To account for this, in addition to presenting the top all-time references, we split the 40 years into 8 five-year segments and compare the citation count among papers published in the same time segment. To characterize the "highly cited", for each five year period, we select out the papers with citation counts ranked the top 10% among all the papers published in the same period. Due to the highly skewed distribution and the average of 5 citation counts per paper, we exclude the papers that have total citations less or equal to 5, which consist of 56163 papers. In this way, we have labeled our remaining 15661 papers into two class, 1519 are "highly cited" and 14142 are "moderately cited".

3.1.2 Features

Length of title we consider whether we can find any evidence that the style in which a paper is written may relate to its success. Specifically, we consider the length of the article title chosen by the authors and investigate whether the length bears any relation to the number of citations. Some studies have investigate into the relationship have reported some conflicting results, with one study suggesting that papers with longer titles might receive more citationsJacques and Sebire (2010) and another finding no evidence of a relationship. Here we measure the length of the title as the number of words including the numbers.

Publication year Due to the digitization of publishing and the marvelous improvements to search and relevance ranking, referring to the old papers become far easier than before. Also, the expansion of the volume of papers published per year also changed the way people cite the the previous work. Conceptually the publication year can explain some of the variance in citations, and therefore we take it into consideration.

Number of authors Figure 8 shows that the highly cited papers have less proportion of single-authored papers, instead they usually have multiple authors. Some research result shows that studies that have more authors tend to draw on a greater diversity of expertise, and thus present a greater diversity of ideas and/or data types, especially when collaborations are interdisciplinary

Number of reference The increasing number of references in scientific journal articles suggests editors may prefer articles with many references. Naturally we want analyze whether the total number of references will affect citations. Since we only consider the citation action within the data set, papers published in earlier years have fewer reference within our scope, as is shown in Figure 4. Therefore we fix the year effect by by dividing the number of reference by the length of publication year from 1975. The resulting distribution of reference count are similar for each time period.

Proportion of self citation in the reference There are many reason people would cite there own previous work. Some people perceives the lack of self-citation as a sign that the author has little or no background on the subject in question. On the other hand, they perceive over self-citation as a sign of the author's narcissism and of ignoring the research of their co-workers. Thus we also want exploit our data to investigate wether the self-citation proportion has a relationship with the citations Fox et al. (2016). Therefore we include the number of authors as a quantitive feature to investigate whether the increase in author number would promote the citations of papers.

Information of reference papers The literature review always starts form the most classic papers in the field, and then trace back or forward to find the order or more recent related papers. As a result, chances are that the more influential the reference papers are, the more citations a paper would receive. So here we want include the past performance of the reference papers to analyze there association with the most recent paper.

- The average citations received by the reference papers
- The average age of the reference papers

Priori author-based features A number of studies have provided evidence that the long-term success of scientists depends on their early publications van Dijk et al. (2014); Laurance et al. (2013). analyses have indicated that a paper's success can be partially predicted by its early success Acuna et al. (2012); Wang et al. (2013) as well as the reputation of the authors. Here we try to capture how well previous papers from the same authors have performed in the past, where the past publication history of the authros can be expressed in terms of: (the average is taken over all the authors of the new paper)

- The average number of citations received.
- The average number of papers authored.
- The average citation/publication ratio. Here we take into consideration the possible scenario that the author published a lot of papers, and the total past citation is high in terms of absolute value, however, the citation per paper is comparatively moderate. We are interested in whether this situation would have an effect on the citation of the author's future work.
- Citation of previous "best" papers. Among all the previous papers of the authors, record the citation counts and age of the paper that has the maximum number of citations up to the year the new paper published.

Collaborators We want to investigate whether highly collaborative authors tend to be cited differently from authors in their field who work alone or with very few people? Wallace et al. (2012) have shown that an increasing share of citations received is comes from collaborators, as well as collaborators of collaborators. Similarly, Ajiferuke et al. (2011) provided evidence that collaborators cite each other's papers more than non-collaborators. Therefore we introduce several variables to characterize the collaboration level of the authors to get an insight in how this is associated with the citations the paper recieves.

- The number of all the previous collaborators of the authors. Here the collaborator is defined as co-authoring at least one paper in the past.
- The average number of previous citations of all the collaborators. This measures the early success of the collaborators.
- The average number of publications of all the collaborators.

Journal As is shown in Figure 9, different journals have different association with citations of their papers. It seems "good journals" always attract the best work, and in turn the excellent papers contribute to the higher rank of the journal. Also, Larivière and Gingras (2010) shows that the same papers when published in an already high impact journal will accrue more citations than when published in a less high impact journal. This is probably due to that scientists often cite material to which they have been readily exposed, often from high impact journals. Here we define a factorial variable with two level, one refers to the four "top journals" identified from Figure 9: AOS, JASA, Bmtrk, and JRSS-B, the other level represents the rest of the journals.

3.1.3 Collinearity

Before fitting the model, we analyze whether there is strong collinearity existing in the covariates. Since highly correlated covariates might result in signal canceling in parameter estimation. We have altogether 16 variables from the above feature construction process, while we have 15661 observations, which is about 100 times of the predictor size. Therefore a slight difference in the predictors might explain a relative amount of difference in observations. Figure 10 shows the the variables pairs that having correlation greater than 0.85. Though strongly correlated, we can still observe some variance away from the dense central line. Considering that we have quite an unbalanced classes, where the "highly cited" class is only ninth the size of the ”moderately cited" class, therefore we incline no to desert many of the variables. Here we only exclude the total number of collaborators' publication from the model due to the extremely high correlation. Ideally, the redundancy can be filtered through variable selection.

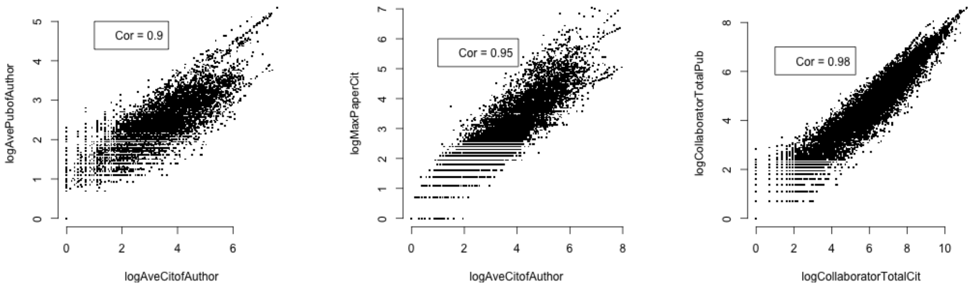

Figure 10: The scatter plots of variables pairs having correlation greater than 0.85.

Since the data set covers 40 years of papers, due to the expansion of the volume of papers published per year and the digitization of publishing and the marvelous improvements to search and relevance ranking, the way people cite the old reference might also has changed over the time. Also, due the limitation of papers within the data set, we only have limited information of the authors of earlier papers. Therefore, we introduce the interaction of publication year with some of the variables that potentially subject to time effect, including number of authors, the average citation of reference, the average age of reference papers, the average number of previous citation and publication of authors as well as the citation counts of the best paper of the author, and the number of citation and publication of the collaborators.

3.1.4 Fit weighted logistic regression

Together with the interaction, we have in total 22 features, 15661 observations and two classes, including the "highly cited"(1519) and "moderately cited"(14142). Therefore it is binary classification problem with unbalanced classes. We fit the weighted logistic regression with the weights chosen as the inverse of class size. Variables are ranked according to the absolute z-value in Figure 11.

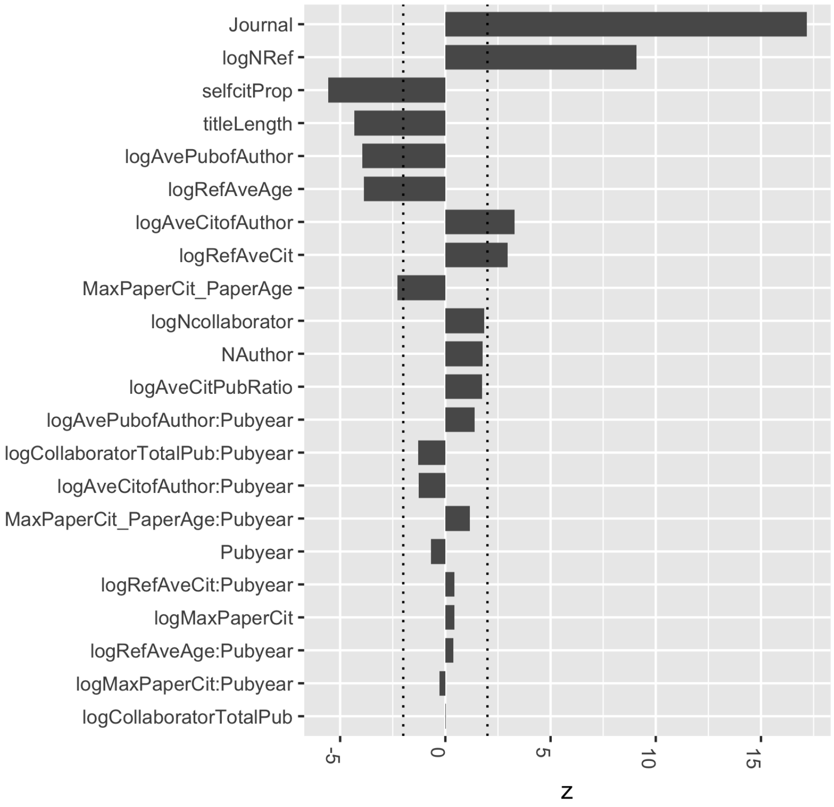

Figure 11: Variable ranking by z-value of weighted logistic regression model. Two vertical lines shows the significance threshold.

3.1.5 Variable selection

In order to avoid double fitting of the data, We split it as 80% for training and 20% for testing. Then perform the model selection on the training data, leaving the test data untouched.

Backward stepwise selection We apply logistic regression with backward stepwise selection. Again we use only the training data to fit the model. As the result, it selects 12 variables out of 22 total variables, Including in interaction term.

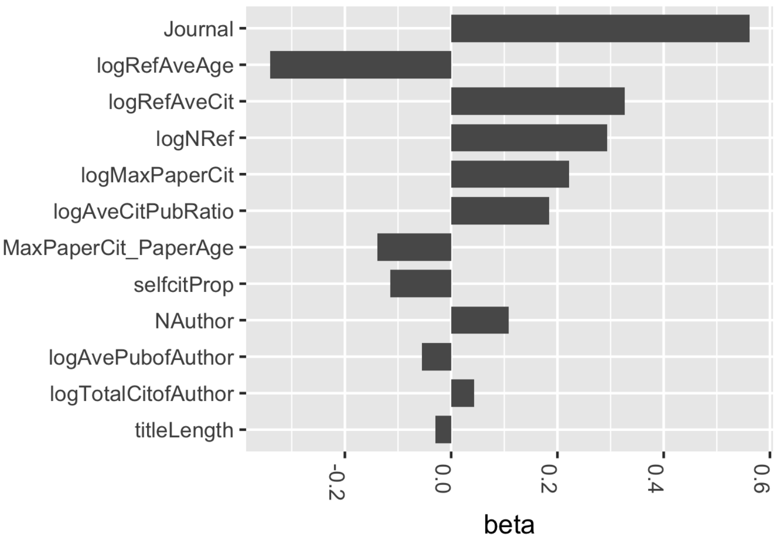

Figure 12: Variables selected by stepwise selection.

3.1.6 Variable ranking

Variable ranking by z-value One approach is to rank predictors by the z-statistics of the Wald chi?square test, $H_0 : \beta_i = 0$; the null hypothesis is that there is no association between the predictor i and the outcome after taking into account the other predictors in the model. Small p?values indicate that the null hypothesis should be rejected, meaning that there is evidence of a non?zero association. This metric only indicates the strength of evidence that there is some association, not the magnitude of the association. Thus the ranking is interpreted as a ranking in terms of strength of evidence of non?zero association.

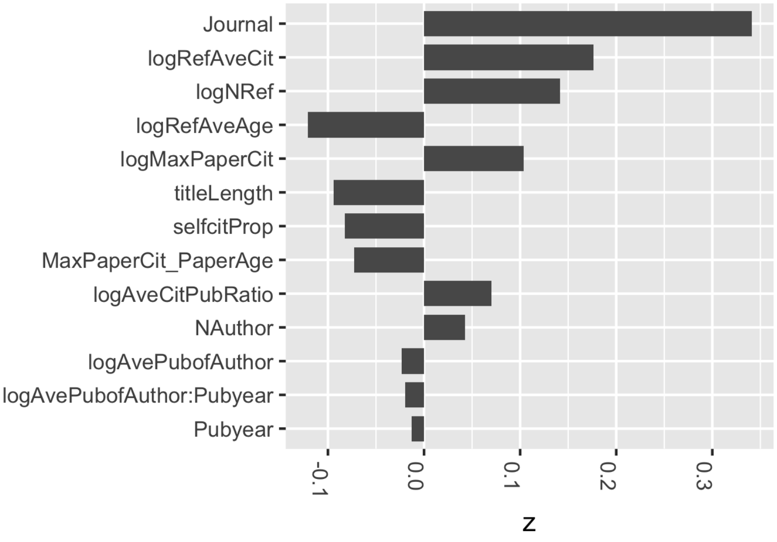

Figure 13: Variable ranking by z-statistics of Wald chi-square test

Variable ranking by marginal R^2 In linear regression, partial correlation is the marginal contribution of a single predictor to explain the variance in the outcome. Alternatively, we can do a separate linear regression on each predictors and check the percent of variation explained by R^2. For logistic regression the procedure is a little bit different, we have to look at pseudo-R^2s instead of R^2. By doing this we can rank the predictor according to how much each of them explains the response variable. The ranking result in shown in Figure 14

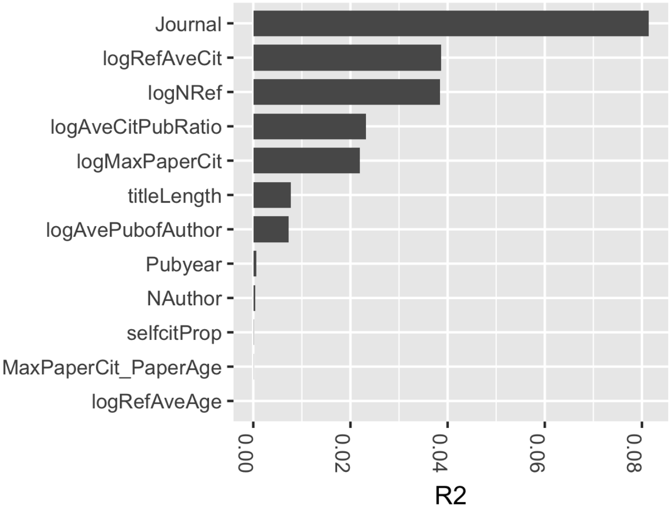

Figure 14: Variable ranking by marginal R^2

Compared with the ranking from Wald chi-square test statistics, we observer that among the top 6 variables, 5 of them are the same except for the difference of average age of reference, which ranked the fourth in ranking by z while ranked the last by marginal R^2. The reason is not yet known.

3.1.7 Prediction

Using the selected variable, we can use the fitted weighted logistic regression model to predict on the test data. The accuracy of prediction is 67.7%, the balanced accuracy is 69.8% and the confusion matrix is shown in table 6. Though the classes are unbalance, the weighted logistic regression performs well in producing balanced accuracy for each class. The result seems promising considering the limited given information of the papers and authors.

Prediction vs Reference	moderately-cited	highly-cited
moderately-cited	1902	84
highly-cited	926	219

Table 6: Confusion matrix of prediction result on the testing data set.

The prediction result is good given the limited information we have utilized for fitting the model. We do not having any information regarding the content of the articles, and we do not incorporate any objective evaluation of the research contributions. The features of the papers are all available and easy to obtain at the time the paper went out. Therefore this prediction model provide us with a handy way to make a rough prediction on the future success of the published article in term of number of citations.

3.2 Why different citation patterns?

In the first part, we considered the characters of highly-cited papers versus the moderately-cited papers and got some interesting results. Next, we want to dig further into the behaviors of the highly-cited papers. Especially we are curious about how the citation counts accumulated over the years (citation patterns). In order to allow for a long enough time span to form a clear pattern, here we use the earlier part of the data, papers published from 1976 to 2001, to fit the model and perform analysis. We may use the later part for prediction, for instance, how their citation patterns will involve in the future. We still consider the papers ranking top 10% for each five-year publication period in terms of the total citation counts, and before that we remove the papers having less that 5 citations. 922 papers are selected out for analysis. Among them, the highest citation count is 1686 and lowest is 41. Since the papers are all top-cited, we are interested in the relative ups and downs over the years than the absolute number of cites per year, therefore we first did some pre-processing to get a clear citation pattern for each paper:

- First calculate the citation count per year from the year of publication to 2014, which is a vector of length $(2014 - year_{pub})$.
- Then Smooth the counts for each year by taking average of five consecutive years (two before and two afterwards).
- Normalize the vector to make the sum equals 1.

After these three steps, we get a smooth citation pattern for each paper when plot the resulting vector over the years. The patterns are of different length for papers published in different years. After some rough clustering analysis, we can observe that the patterns fall into several categories, as is shown in Figure 15.

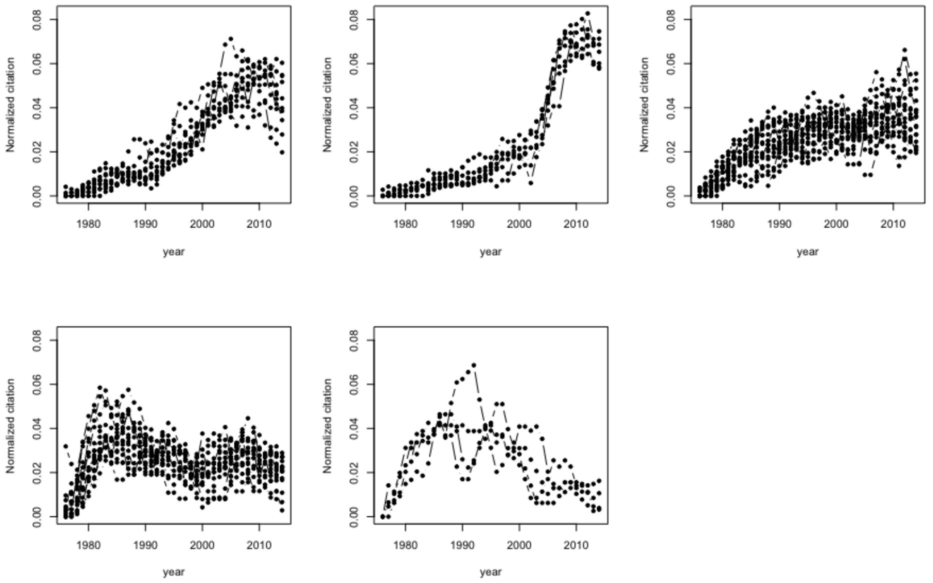

Figure 15: Clusters from hierarchical clustering of highly-cited papers published between 1976 and 1981, where the distance is measured by the correlation of two pattern vectors.

The different patterns depict different process of how the papers are acknowledged by scholars and popularized. For example, Some papers rose to famous only few years after the article went out, while other papers might have waited years before the they began to become famous. In contrast, some other papers may have a stable performance over the years, receiving even amount of citation each years. An interesting and straightforward problem arises here that why the evolution of citations show distinct patterns, and what features are associated with each patterns.

3.2.1 Label the patterns

At first sight, the results from hierarchical clustering can serve as a good guidance for labeling. However there are several problems come with this. Since the pattern vectors are of different length, we need to produce the clustering result for papers for each five-year period. Thus problem arises that the structure of the dendrogram are different and the patterns shown in each result are slightly different. It is hard to form an unified labeling criterion.

Here we adopt the Group-Based Modeling of Development(GBMD) model from Nagin (2005), which is designed for longitudinal data (with a time-based dimension) and can divide the population into subgroups according to the developmental trajectories. Group-based trajectory models are a specialized application of finite mixture models. While the conceptual aim of the analysis is to identify clusters of individuals with similar trajectories, the model?s estimated parameters are not the result of a cluster analysis. Rather they are the product of maximum likelihood estimation. As such, they share the many desirable characteristics of maximum likelihood parameter estimates?they are consistent and asymptotically normally distributed. Let $Y_i = (y_{i1}, y_{i,1}, \dots, y_{iT})$ denote a longitudinal sequence of measurements on individual i over T periods. Let $P(Y_i)$ denote the probability of Y_i. The group-based method assumes that individual differences in trajectories can be summarized by a finite set of different polynomial functions of age or time. Each such set corresponds to a trajectory group which is hereafter indexed by j. Let $P^j(Y_i)$ denote the probability of Y_i given membership in group j , and π_j denote the probability of a randomly chosen population member belonging to group j. Since the group membership is not observed, the construction of the likelihood function requires the aggregation of the J conditional likelihood functions, $P^j(Y_i)$, to form the unconditional probability of the data, Y_i:

$$P(Y_i) = \sum_{j}^{J} \pi_j P^j(Y_i)$$

where $P(Y_i)$ is the unconditional probability of observing individual i's longitudinal sequence of behavioral measurements, Y_i . It equals the sum across the J groups of the probability of Y_i given i's membership in group j weighted by the probability of membership in group j. For given j, conditional independence is assumed for the sequential realizations of the elements of Y_i, y_{it}, over the T periods of measurement. Thus $P^j_{Y_i} = \Pi^T p^j(y_{it})$, where $p^j(y_{it})$ is the probability distribution function of y_{it} given membership in group j. The expected developmental trajectories are modeled as the polynomial of time t, which accepted input of different length of vectors. The output of th model are the posterior probability $P(\text{class k}|Y_i)$ of each individual belong to each subgroups and the assigned label of each observation into the subgroups which correspond to the highest probability.

We fit the GBMD model with our 922 pattern vectors, where the number of groups is set to be 4 and the expected trajectory for each group is restricted to third order polynomials. The four trajectories of four subgroups and the percentage of each group is shown in Figure 16.

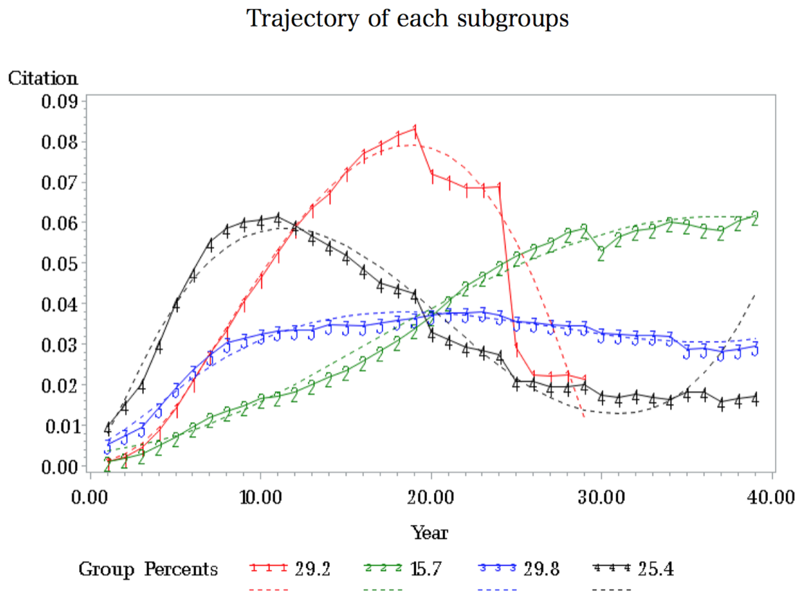

Figure 16: The output trajectories for the 4 groups from GBMD model

The trajectories captures the characteristic patterns we observe in clustering analysis. The first and fourth group include those papers that captures most of their citations in a comparatively short period after the paper is published and afterward it popularity decreases; the papers in the second group have stably increasing citation per year over the years; papers in the third group have rough same citation each year and haven't changed for a long time. Since the first and fourth group have essentially the same trend, that is, increasing at first followed by decrease. Therefore we combine the two groups together as group one. As a result, all the 922 are labeled into three classes.

3.2.2 Features

We adopt the same features as in the previous analysis. There are in total 22 features and 922 observations, including the interactions with the publish year.

3.2.3 Method and results

There are 922 observations 22 features and 3 classes. We split the data as 80% for training and 20% for testing. Fit the training data with the multinomial logistic regression model with the lasso regularization. The parameter lambda is chosen by cross-validation.

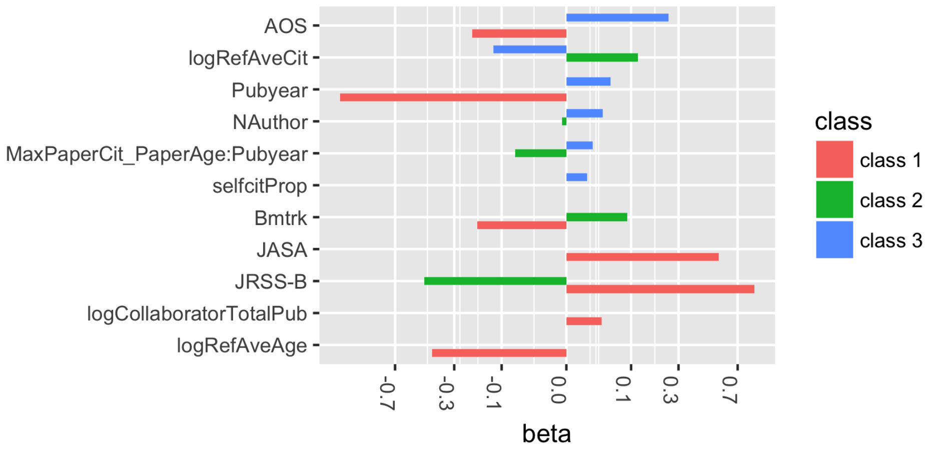

Figure 17: The parameter β of the fitted multiclass weighted logistic regression. Three colors show the parameter of three classes.

Table 7: Confusion matrix of prediction result on the testing data set.

Prediction vs Reference	class 1	class 2	class 3
class 1	84	2	8
class 2	13	14	19
class 3	4	13	27

The accuracy of prediction on the test data is 68% and the confusion matrix is shown in table 7. In Figure 12 right we visualize the model parameters for different classes. From the confusion matrix we can observe that the first class is decently well-distinguished from the other two classes. However the second and the third classes are less separable from each other. This can probably be explained by the similar trajectories of the two classes considering the overall development trend of the discipline. Due to the growing interests in the Statistics in recent years that have been boosting the amount of published articles and citations each year. The articles assigned to the increasing pattern in the second class might belong to the flat group since the increasing popularity merely comes from the increasing amount of citations generated each year, but not due to the inherent difference of the papers.

3.2.4 Discussion on the different citation patterns

Apart from providing the underling class structure for predicting the pattern membership, the result output form the Group-based Trajectory Development model is of interest in many other ways. Therefore in this section we will devote few more lines to giving a brief discussion on the identified citation patterns. Figure 18 shows the the fitted trajectories and example article for each of the four groups. We can observe that the individual patterns follow the general trajectory of the assigned group with some perturbations. One interesting result that worth noticing is the example from group 2, which was the first paper in EM Algorithm Dempster et al. (1977). It is shown from the citation pattern that this paper has been receiving more and more citation each year despite it was published more than 30 years ago. Though this is reasonable considering that the papers marks the appearance of the a completely new field. So whenever people exploit or improve upon EM algorithm, they would always cite the original paper.

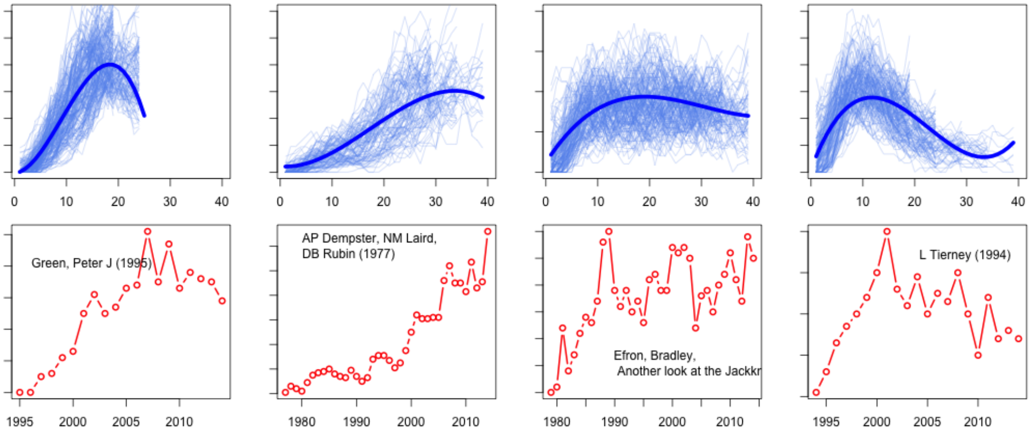

Figure 18: Upper: The fitted trajectory for each four groups and all the trajectories that are clustered into the group are shown in the background. Lower: examples of the papers belong to each four groups.

In order to gain further insight into the composition of each group, we performed topic analysis from the given title information. The topics that are distinct for each group can be explained by the keywords that exist in abundant in one group while less frequent in other groups. We measure the 'uniqueness' by defining a 'unique score' for each keyword, where

$$\text{Unique score} = \frac{\#\text{exists in group } i}{\#\text{exists in all roups}}$$

Here we take into consideration both the unigram words and bigram keywords so as to capture the statistically meaningful phrases. Such that a higher unique score is an indicator of distinct topic of a group. In Table 8 we list the keywords that show a relatively high uniqueness for each group.

Table 8: The unique keywords for each group.

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

We can observe that among the unique keywords for the first group, some keywords are extremely hot in most recent years, for instance, "model selection" and "semiparametric". For the third group, there are a couple of nonparametric-analysis-related keywords including "spline smooth", "kernel density" and "nonparametric estimation", etc. For the fourth group, some bayesian-analysis-related keywords like "bayes factor", "bayesian analysis" and "markov chain" coms into our attention.

4 Summary and future work.

In this analysis, we investigated the association between citations and various features of papers including the journal, author, keywords by marginal regression analysis. Then we built a classification model with the features of paper available at the time the paper was published by either weighted logistic regression and random forest, which can achieve around 68% accuracy in predicting whether the paper are ranked among top 10% in terms of total citation. By variable selection and variable ranking we pointed out the predictive features.

For analyzing the citation patterns, we fitted the normalized citation patterns of the highly-cited papers with Group-based Trajectory Development model, and from which we identified 4 characteristic trajectories. We investigated the features of different groups and analyzed the unique topics for each group.

For future work, we would proceed to clean the abstract data, which was available just recently. Since in the previous analysis we only utilized the keywords in title, which only provide a weak signal. Therefore with the additional keywords information we can perform more detailed topic analysis. For example, we can analyze what are the hot topics in statistical discipline for each period, and how the popularity of different topics are developing over the years, which can provide us with further insight into how the society of Statistics have been developing.

References

Acuna, D. E., Allesina, S., and Kording, K. P. (2012). Future impact: Predicting scientific success. *Nature*, 489(7415):201–202.

Ajiferuke, I., Lu, K., and Wolfram, D. (2011). Who are the research disciples of an author? examining publication recitation and oeuvre citation exhaustivity. *Journal of Informetrics*, 5(2):292–302.

Andersen, P. K. and Gill, R. D. (1982). Cox's regression model for counting processes: a large sample study. *The annals of statistics*, pages 1100–1120.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)*, pages 289–300.

Bornmann, L. and Daniel, H.-D. (2008). What do citation counts measure? a review of studies on citing behavior. *Journal of documentation*, 64(1):45–80.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38.

Donoho, D. L. and Johnstone, J. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *biometrika*, 81(3):425–455.

Efron, B. (1992). Bootstrap methods: another look at the jackknife. In *Breakthroughs in statistics*, pages 569–593. Springer.

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., et al. (2004). Least angle regression. *The Annals of statistics*, 32(2):407–499.

Eilers, P. H. and Marx, B. D. (1996). Flexible smoothing with b-splines and penalties. *Statistical science*, pages 89–102.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360.

Fox, C. W., Paine, C., and Sauterey, B. (2016). Citations increase with manuscript length, author number, and references cited in ecology journals. *Ecology and Evolution*, 6(21):7717–7726.

Garfield, E. (1955). Citation indexes for science. *Science*, 122:108–111.

Garfield, E. (2006). The history and meaning of the journal impact factor. *Jama*, 295(1):90–93.

Garfield, E. et al. (1972). Citation analysis as a tool in journal evaluation. American Association for the Advancement of Science.

Gelfand, A. E. and Smith, A. F. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American statistical association*, 85(410):398–409.

Green, P. J. (1995). Reversible jump markov chain monte carlo computation and bayesian model determination. *Biometrika*, 82(4):711–732.

Gross, P. L. and Gross, E. M. (1927). College libraries and chemical education. *science*, 66(1713):385–389.

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National academy of Sciences of the United States of America*, 102(46):16569.

Jacques, T. S. and Sebire, N. J. (2010). The impact of article titles on citation hits: an analysis of general and specialist medical journals. *JRSM short reports*, 1(1):1–5.

Ji, P. and Jin, J. (2017). Coauthorship and citation networks for statisticians. *The Annals of Applied Statistics*, 10(4):1779–1812.

Larivière, V. and Gingras, Y. (2010). The impact factor's matthew effect: A natural experiment in bibliometrics. *Journal of the Association for Information Science and Technology*, 61(2):424–427.

Laurance, W. F., Useche, D. C., Laurance, S. G., and Bradshaw, C. J. (2013). Predicting publication success for biologists. *BioScience*, 63(10):817–823.

Leskovec, J., Kleinberg, J., and Faloutsos, C. (2005). Graphs over time: densification laws, shrinking diameters and possible explanations. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 177–187. ACM.

Liang, K.-Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22.

Louis, T. A. (1982). Finding the observed information matrix when using the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 226–233.

Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *The annals of statistics*, pages 1436–1462.

Nagin, D. (2005). *Group-based modeling of development*. Harvard University Press.

Newman, M. E. (2009). The first-mover advantage in scientific publication. *EPL (Europhysics Letters)*, 86(6):68001.

Owen, A. B. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, 75(2):237–249.

Padial, A. A., Nabout, J. C., Siqueira, T., Bini, L. M., and Diniz-Filho, J. A. F. (2010). Weak evidence for determinants of citation frequency in ecological articles. *Scientometrics*, 85(1):1–12.

Redner, S. (2005). Citation statistics from 110 years of physical review. *Physics today*, 58(6):49–54.

Richardson, S. and Green, P. J. (1997). On bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society: series B (statistical methodology)*, 59(4):731–792.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.

Schwarz, G. et al. (1978). Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464.

Smith, L. C. (1981). Citation analysis.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):583–639.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288.

van Dijk, D., Manor, O., and Carey, L. B. (2014). Publication metrics and success on the academic job market. *Current Biology*, 24(11):R516–R517.

Wallace, M. L., Larivière, V., and Gingras, Y. (2012). A small world of citations? the influence of collaboration networks on citation practices. *PloS one*, 7(3):e33339.

Wang, D., Song, C., and Barabási, A.-L. (2013). Quantifying long-term scientific impact. *Science*, 342(6154):127–132.

Wuchty, S., Jones, B. F., and Uzzi, B. (2007). The increasing dominance of teams in production of knowledge. *Science*, 316(5827):1036–1039.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429.