Statistical Models for Frequency Distributions of Count Data with Applications to Scientometrics

Ruheyan Nuermaimaiti

School of Mathematics

University of Leeds

Submitted in accordance with the requirements for the degree of

Doctor of Philosophy

27th April 2023

Intellectual Property and Publication Statement

I confirm that the work submitted is my own, except where work which has formed part of jointly authored publications has been included. The contribution of myself and the other authors to this work has been explicitly indicated below. I confirm that appropriate credit has been given within the thesis where reference has been made to the work of others.

- Chapter 5 contains parts of the material published in a jointly authored preprint by Bogachev *et al.* (2023). The paper is concerned with the concept and derivation of the limit shape for the GIGP frequency distribution. In this work I performed most of the formal analysis, in particular with all mathematical proofs obtained by myself for the results included in the present thesis. I also carried out data analysis including goodness-of-fit assessment and interpretation, and wrote the first draft of the paper. Results of Bogachev *et al.* (2023) not included in the present thesis are concerned with (i) the uniform convergence to the limit shape (Theorem 4.1; §4.5, Lemma 4.5; and §§4.6–4.7) and (ii) finite-dimensional and sample path enhancements of the fluctuation results (Theorems 5.2, 5.3, 6.2 and 6.3). The latter results belong to Dr Bogachev and Dr Voss, who also contributed to the design of study and took part in preparation of the final version of the paper.

- Chapter 6 involves the material published in a jointly authored research-in-progress paper by Nuermaimaiti *et al.* (2021). The paper contains a sketch presentation of the Generalised Power Law model, which is treated in this chapter in full detail. Most of the calculations and all of the data collection and analysis were carried out by myself. I have presented the work at the ISSI 2021 conference (see bibliographic details in Nuermaimaiti *et al.* (2021)). I wrote up the first draft of the paper and also acted as the corresponding author in this publication. My co-authors, Dr Bogachev and Dr Voss, designed the study and assisted with some of the steps in the derivation of the limit shape. They also contributed to the interpretation and discussion of the results and participated in writing up the final version of the paper.

Copyright Statement

This thesis is dedicated to my mother, who loves me unconditionally, taught me to achieve my goal by doing every little thing steadfastly, and taught me to be strong. This thesis is also dedicated to people who love and support me.

Acknowledgements

I am deeply indebted to my parents for their never-ending encouragement and generous financial support throughout my PhD study, which has made this research possible.

I am particularly grateful to my supervisors Dr Leonid Bogachev and Dr Jochen Voss; without their abundant knowledge and patient guidance this work would not have been accomplished.

I gratefully acknowledge the constructive comments of the anonymous referees of my "research-in-progress" paper (Nuermaimaiti *et al.*, 2021) presented at the 18th International Society for Scientometrics and Informetrics Conference (ISSI 2021), in particular pointing out the GIGP model (Sichel, 1985) and emphasising the importance of a conceptual justification of models. I also wish to thank Dr Jens Peter Andersen and Dr Andrea Scharnhorst, who discussed my work and provided valuable advice at the Doctoral Forum of the ISSI 2021.

The author is grateful to Dr Yuri Yakubovich for valuable suggestions and discussions on integer partitions and to Dr Anna Senkevich for helpful insights concerning data retrieval from the web. Special thanks to Isabella Deutsch for useful discussions about the Hawkes processes. Thanks are also due to my progress review assessors Dr John Paul Gosling and Professor Charles Taylor, who gave me valuable advice in the early stage of this work.

I want to thank all my friends in the School of Mathematics for having lunch group, hiking group, badminton group, and cinema group, which made my PhD student life lively; special thanks are to the people who helped me through the challenging Covid lockdown period.

Abstract

This thesis investigates the statistical modelling of count data, common in diverse fields, including bibliometrics and scientometrics. The item production model approach reviews existing models, such as the power law and the generalised inverse Gaussian-Poisson model. Furthermore, a new generalised power law model is proposed to enhance the fitting of models to specific data types. The main focus of the study is on scientific production use cases, encompassing authors, papers, and citations. The research adopts probabilistic combinatorics and the theory of random integer partitions to gain valuable insights into the data structure, which examines Young diagrams and their scaling limits. These insights are then applied to model and estimate production metrics, including the h-index and g-index.

Contents

List of Figures

Summary of Basic Notation

$\mathbb{N}$	The set of natural numbers (positive integers), $\mathbb{N} = \{1, 2, \dots\}$
$\mathbb{N}_0$	The set of non-negative integers, $\mathbb{N}_0 = \{0, 1, 2, \dots\}$
$\mathbb{R}$	The set of real numbers, $\mathbb{R} = \{x \in (-\infty, \infty)\}$
$\mathbb{C}$	The set of complex numbers, $\mathbb{C} = \{z = x + \mathrm{i}y,\ x, y \in (-\infty, \infty)\}$
$a \sim b$	Asymptotic equivalence, $a/b \to 1$
$a \gg b$	Asymptotically bigger than, $a/b \to \infty$ (also written as $b = o(a)$)
$a \ll b$	Asymptotically smaller than, $a/b \to 0$ (also written as $a = o(b)$)
$\#B$	The number of elements in set B
$\mathbf{1}_D(x)$	Indicator function of set $D \subset \mathbb{R}$
I_A	Indicator of event A
$\mathsf{P}(A)$	Probability of event A
$\mathsf{E}(X)$	Expected value of random variable X
$\xrightarrow{\mathrm{p}}$	Convergence in probability
$\xrightarrow{\mathrm{d}}$	Convergence in distribution
$\doteq$	Approximately equal to
$\mathcal{N}(0, 1)$	Standard normal distribution (with zero mean and unit variance)
M	Total number of sources in the item production model
K	Number of batteries of sources in the composite item production model
X_i	Random number of items produced by source $i \in \{1, 2, \dots, M\}$
N	Total number of items produced, $N = X_1 + \cdots + X_M$
M_j	Multiplicity of the output j, defined as a random number of sources (out of M) that produced j items each $(j \in \mathbb{N}_0)$
$Y(x)$	Upper boundary of the Young diagram, $Y(x) = \sum_{j \ge x} M_j\ (x \ge 0)$
f_j	Theoretical frequency distribution, $f_j = \mathsf{P}(X_i = j)$ $(j \in \mathbb{N}_0)$
$F(x)$	Cumulative distribution function (CDF), $F(x) = \mathsf{P}(X_i < x) = \sum_{j<x} f_j\ \ (x \ge 0)$
$\bar{F}(x)$	Complementary cumulative distribution function (CCDF), $\bar{F}(x) = \mathsf{P}(X_i \ge x) = \sum_{j \ge x} f_j\ \ (x \ge 0)$
μ	Expected number of items produced by a source, $\mu = \mathsf{E}(X_i) = \sum_{j \in \mathbb{N}_0} j\, f_j$
$\hat{f}_j$	Sample frequency of j items, $\hat{f}_j = M_j/M$ (an estimate of f_j, $j \in \mathbb{N}_0$)
$\hat{\mu}$	Sample mean of items per source, $\hat{\mu} = N/M$ (an estimate of μ)

Chapter 1

Introduction

In this chapter, the focus of the research is established: the analysis of frequency distribution in count data with the application of scientometric data examples. The background and motivation for this study are also explained. In the end, an outline of the subsequent chapters is presented.

1.1 Background and Motivation

In many applied situations, one deals with *count data* in the form of sample frequencies of occurrence in one of the countably many categories ("boxes"). It is often appropriate to interpret occurrence in each box as the corresponding number of batched "items" produced by one out of the plurality of contributing "sources"; for example, falling into box with label zero is interpreted as no items produced by the source.

Diverse real-life examples of such a scenario include: abundance data of various species such as butterflies, with different species treated as sources and the observed counts as items; the number of followers (items) of different accounts in Twitter (sources); repeat-buying data, with the number of units bought (items) by households (sources); the number of papers (items) produced by authors (sources); etc. For more examples and further references, see, e.g., a monograph by Egghe (2005) or a review paper by Clauset *et al.* (2009). In the latter example, papers may themselves play the role of sources, with citations as items. Altogether, this forms an interesting triangle of relationships, *authors–papers–citations (APC)*, which is one of the main subjects of investigation in *scientometrics* (Egghe, 2005).

A natural objective with such types of data is to explain the observed (relative) frequencies by fitting a suitable distributional model, preferably possessing some conceptual foundation and applicable to a variety of use cases. Of course, in any real-life data set the numbers of sources will reduce to zero for the value of items big enough, but this is reconciled with the modelling prediction simply by the fact that the theoretical frequencies tend to zero as the number of items goes to infinity. However, adequate modelling of the long-tail frequencies is of importance in relation to understanding the behaviour of extreme values in the count data (e.g., untypically high citations).

An important principle of statistical modelling is that the choice of a suitable model for the data is not driven only by goodness-of-fit, but should be intuitively justifiable and interpretable. Quoting from Rousseau (2002, p. 320), the distribution chosen to fit scientometric data "should not be the result of a purely statistical fitting exercise, but should be explained... One should first make a model based on reasonable and acceptable assumptions... A best fitting distribution should then be derived from this model." On the other hand, the observed goodness-of-fit of a chosen model to real data may shed light on the proposed mechanism by providing evidence in favour or against it. Specifically, a reasonable fit would support the validity of the mechanism in question whereas a "bad" fit might signal the need to review the model and either improve or discard it.

A celebrated example of a theoretical frequency model is the *power law*, first proposed by Lotka (1926) to describe the publication statistics in chemistry and physics, and based on the empirical observation that the sample frequencies approximately follow a power law distribution (Coile, 1977; Egghe, 2005). Price (1965) discovered an important connection with networks, whereby citations were interpreted as nodes' degrees. Examples of fitting the power law to the citation data can be found in Coile (1977), Redner (1998), and Clauset *et al.* (2009).

An evident heuristic tool to fit a power law model to the count data is by looking at the frequency plots (e.g., histograms) with logarithmic scales on both axes, whereby one seeks a straight-line fit (Nicholls, 1987). An alternative approach

(Clauset *et al.*, 2009), which provides the helpful smoothing of the discrete data, is via the complementary cumulative frequencies where, using again the log-log plots, a good fit approximately corresponds to a straight line. More formally, the model can be fitted using standard statistical methods such as the maximum likelihood or ordinary least squares estimation (Nicholls, 1987).

The conventional explanation of universality of the power law is based on the principle of *cumulative advantage*, also expressed as the catchphrase "success breeds success", originally coined in the context of scientific productivity (Price (1965); Price (1976); Egghe & Rousseau (1995)); see also a more recent review by Huber (2002) with a critique of cumulative advantage. Unfortunately, the utility of Lotka's power law for real data modelling is often limited by fitting to the data well only on a reduced range of count values, requiring a truncation of lower values (see an extensive discussion in Clauset *et al.* (2009)) or of higher (long-tail) values better described by a *stretched-exponential law* (Laherrere & Sornette, 1998).

Numerous other attempts to fit theoretical distributions to a variety of count data sets included using the negative binomial distribution, the modified geometric distribution, the beta binomial distribution, and many more (Johnson *et al.*, 2005) (see the discussion and further references in Huber, 2002; Sichel, 1985), however, none of these distributional families proved to be sufficiently "universal" in explaining diverse count data sets, often failing to capture some characteristic features such as modality and long-tail behaviour.

In a series of papers, Sichel (1971, 1973, 1974, 1975, 1982, 1985) introduced and developed the so-called *generalised inverse Gaussian-Poisson (GIGP)* model, proposed in an attempt to grasp a plausible production of items by respecting statistical differences in the individual productivity of sources (e.g., papers and authors, respectively). More precisely, a source is assumed to produce items according to a Poisson law whose rate is itself random with a specific choice of the *generalised inverse Gaussian (GIG)* density (Johnson *et al.*, 1994; Sichel, 1971). In other words, the GIGP distribution is a mixed Poisson distribution under the GIG mixing density (see Gupta & Ong, 2005, for a survey of Poisson mixture models for long-tailed count data). Sichel applied his GIGP model to a great variety of use cases and multiple data sets, from sentence-lengths and word frequencies in written prose (Sichel, 1974, 1975) to number of stones found in diamondiferous deposits (Sichel, 1973) and scientific production (papers and/or citations) (Sichel, 1985). These examples have demonstrated a remarkable flexibility and versatility of the GIGP distribution family.

In a more recent development, Yong (2014) proposed to use combinatorial models of random integer partitions to mimic citation count data, where the constituent parts of the integer partition represent the author's papers with the corresponding numbers of citations, respectively. One drawback of this approach is that papers with zero citations are not represented in the partition, but there are ways in which the concept of a partition could be modified to include zeros. The main perceived advantage of this approach was to leverage the knowledge of so-called *limit shape* for suitably scaled *Young diagrams* visualising parts in the (random) integer partition, which would then enable one to estimate statistically some citation metrics such as the *h-index*. Specifically, noting that the h-index corresponds geometrically to the location with equal coordinates at the upper boundary of the Young diagram, and using an explicit equation for the limit shape under the scaling $\sqrt{N}$ along both axes, where $N \gg 1$ is the total number of citations (Pittel, 1997; Vershik, 1996), Yong came up with a simple estimate of the h-index, $h \approx 0.54\sqrt{N}$, which he then tested using several data sets of mathematical citations (Yong, 2014).

When it comes to data analysis of real-life citations, the aforementioned models can be fitted using standard statistical methods such as the maximum likelihood or ordinary least squares estimation. Unfortunately, neither of these models appear to provide a good match, at least not in the entire spectrum of the citation values. Indeed, it has been documented across many use cases (Clauset *et al.*, 2009) that the power law usually fits quite well but only in the tail region of the frequency range, which motivates the use of truncated models by excluding the lower values. In comparison, partition models demonstrate a reasonable fit only over the initial range, but perform poorly at the tail. A simple pragmatic idea to sew both models in order to cover the entire frequency range may not work because there is usually a gap between the fitted domains. We have encountered this difficulty in our work trying to model the h-index.

To overcome these deficiencies and shortcomings, we proposed a *generalised power law (GPL)* model by modifying the power law setting (Nuermaimaiti *et al.*, 2021). This model interpolates between slow (almost flat) decay of the citation frequencies at the lower end of the citation spectrum and then displaying the power law behaviour at the tail of the frequency distribution. As we discovered after the paper was published, a similar frequency model has been known in the earlier literature as the *hooked power law* (see, e.g., Pennock *et al.* (2002); Thelwall & Wilson (2014); Shahmandi *et al.* (2020)), where the word "hooked" refers to the behaviour of the complementary cumulative distribution function for small counts in log-log coordinates.

The conceptual justification of the GPL model is also based on the mixing idea as in Sichel (1974), but under the different choices of the source production law (geometric instead of Poisson) and the mixing density (a beta distribution instead of the GIG one). As we have demonstrated using a variety of real data sets (see Section 6.4), the GPL model provides a very good fit across the entire citation spectrum. In addition, the GPL model possesses a limit shape, which can be used, for example, to make meaningful estimation of the h-index. In particular, the estimation of the h-index based on the GPL model appears to be significantly more accurate as compared to the partition model.

1.2 Thesis Outline

The rest of the thesis is organised as follows.

Chapter 2 draws forth the research interests of this thesis by presenting informetric data examples. Moreover, this chapter delves into the distinctions and connections among informetrics, scientometrics, and bibliometrics. Furthermore, it provides a comparative analysis of prominent scientometric data platforms.

Moving forward, Chapter 3 commences by presenting a mathematical setup for describing these informetric data sets introduced in Chapter 3 within the "sources-items" system. Additionally, this chapter introduces the Young diagram as a graphical representation of the data. The notion of limit shape is also expounded upon, and an example of the limit shape of random integer partitions is provided. The chapter also introduce the composite item production model, and provide production metrics and model-based estimators for these metrics.

In Chapter 4, one of the classical models, namely power law, is applied to the scientometric data to explore its fitting properties. However, it is discovered that the standard power law model is inadequate for the data set. Thus, the power law model with truncation is explored in addition. Furthermore, by comparing the results obtained from fitting the integer partitions model to the data, a gap in the data domain is discovered.

In Chapter 5, the GIGP model, another classical model, is explored. A mixed Poisson model is employed to explicate the GIGP distribution. This chapter also provides a detailed account of the universal limit shape of the GIGP model, as well as the fluctuations and convergence of this limit shape. Furthermore, computer simulations and data examples are furnished to support the analysis.

Chapter 6 introduces a novel model, the GPL model, to address the shortcomings of the power law model and the integer partitions model previously discussed in Chapter 4. Building upon the conceptualisation of the GIGP model presented in Chapter 5, the GPL model is explained via a mixed geometric model. The limit shape of the GPL model is derived, and data examples of fitting the GPL model are presented.

In Chapter 7, a breakthrough is made by considering the time evolution of citations instead of just analysing a snapshot of scientometric data. This chapter employs exploratory data analysis to gain insights from dynamic citation data. Then the content of the study focuses on the survival analysis of the first citation of publications with considerations given to the number of pages and co-authors of a paper as covariates. At the end of this chapter, the dynamic citations are researched using the point processes.

The thesis concludes with Chapter 8, which provides a summary of the main findings and contributions of the research. Additionally, this chapter outlines future directions for further work.

In addition to the main chapters, the Appendix provides detailed computations to support the research conducted in this thesis. Appendix A summarises the method used for scraping data from Google Scholar web pages. Appendix

B lists asymptotic formulas for the Bessel function, which are helpful for the calculations in Chapter 5.

Chapter 2

Count Data: Examples

The present chapter serves as an introduction to the various types of informetric data that are relevant to the research presented in this thesis. Additionally, it outlines the five data sets that have been utilised in this study. Subsequently, the chapter proceeds to offer a comparative analysis of three mainstream platforms that are commonly used for collecting scientometric data.

2.1 Types of Informetric Data

The aim of this section is twofold: firstly, to introduce the data that will be used for the remainder of the thesis, and secondly, to compare various platforms that offer scientometric data.

Informetrics is a general field of study that focuses on the quantitative aspects of information. Informetrics encompasses more specialised domains, such as bibliometrics and scientometrics.

Bibliometrics refers to statistical analyses of publications such as articles and books, used in the fields of libraries and information science. Pritchard (1969) first used bibliometrics in English in 1969 and defined it as the mathematical and statistical methods for books and media of communication. In many research fields, bibliometrics is used to analyse the impact of fields, researchers and papers. Also, bibliometrics can be applied to descriptive linguistics, the assessment of the use of the reader and evaluating budgets through the analysis of academic literature. Citation analysis is a general method for researching bibliometrics.

Citation indices and citation graphs (or citation networks) are the most commonly used methods in the field of citation analysis. The h-index is one of the science citation indices.

Scientometrics is a sub-field of informetrics, which centres on the analysis and quantification of scientific publications. Scientometrics study is based on the work of Price (1965) (who is credited as the father of scientometrics) and Garfield (1955) (who created the science citation index). Hirsch (2005) proposed a citation index referred to as *h-index* and defined as the maximum number of an author's papers, h, each one cited at least h times (Hirsch, 2005). After the h-index was introduced, to remedy the censoring of the larger citation the g-index was proposed by Egghe (2006). Yong (2014) connected citations and integer partitions, and estimated the h-index using the limit shape of random partitions.

There is a significant overlap between bibliometrics and scientometrics. Figure 2.1 presents a Venn diagram that illustrates the relationship among informetrics, bibliometrics, and scientometrics.

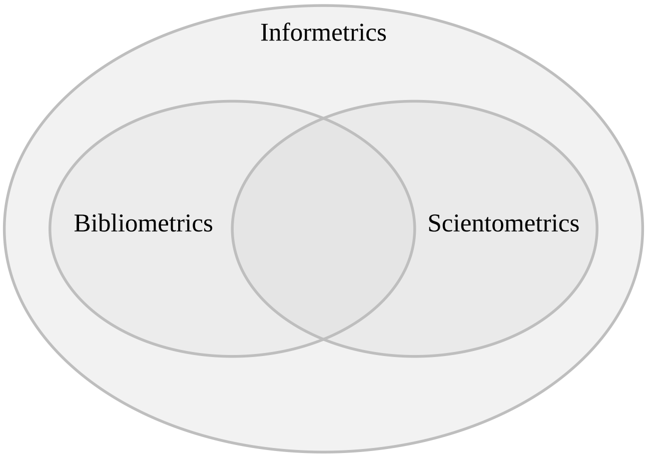


Figure 2.1: Venn Diagram: relationships among the informetrics, bibliometrics and scientometrics.

The present thesis encompasses methodology applicable to diverse informetric data sets, but its primary focus is on the statistical analysis of scientometric data.

2.2 Examples of Informetric Data Sets

This section provides an overview of the data sets used throughout this thesis. Two of these data sets (EJP and AMS) were collected by the author of the present thesis, while other data sets were retrieved from the literature and web databases.

We describe the nature and sources of the data and illustrate the resulting data sets using two types of empirical plots — the frequency plots showing the observed multiplicities per each count value, and complementary cumulative plots based on relative frequencies, depicting the corresponding distributional tails. Noting that the frequencies are typically getting very small for larger counts, in order to visualise details of the tail behaviour it is often useful to plot the data in logarithmically transformed coordinates (referred to as log-log coordinates).

A: Lotka's Data

This data set is from a seminal article by Lotka (1926). It presents the number of papers featured in *Chemical Abstracts* in 1907–1916, restricted to authors whose surnames begin with A and B. The data set comprises 6,891 authors who produced 22,939 papers. Among these authors, 3,991 had authored only one paper each, while 1,059 had authored two papers, and so on. The complete data set can be found in Lotka (1926, page 318, table 1). Notably, the names of authors were not included in the table. The full index of *Chemical Abstracts* for volumes 1 to 10 (1907–1916) is available online, https://babel.hathitrust.org/cgi/pt?id=mdp.39015023498507&view=1up&seq=19&skin=2021.

Lotka's data set is graphically illustrated in Figure 2.2, showing the frequency and complementary cumulative plots, both in the original and log-log coordinates.

B: Chen's Data

Chen's data is from Chen (1972), which comprised counts of the use of physics journals in M.I.T. Science Library in 1971, recorded per each volume taken from the shelves for reading or photocopying. The total number of volumes ever requested was 138, and the total number of requests was 4,292. The frequency plots

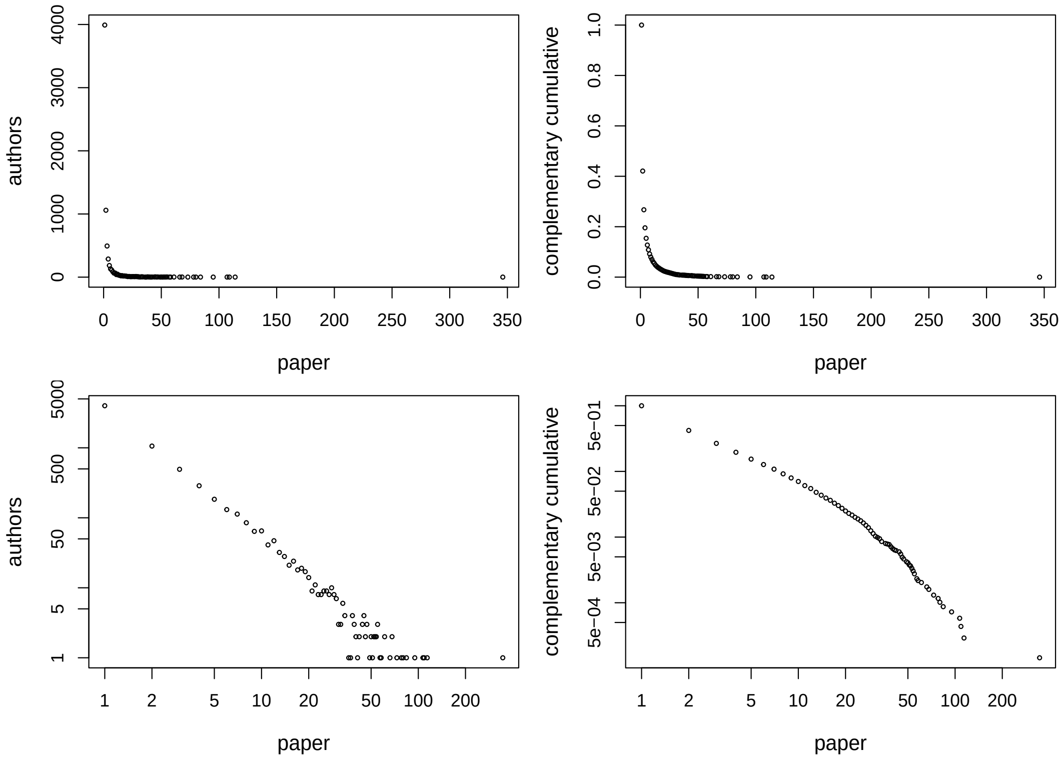


Figure 2.2: Sample frequency (upper left) and complementary cumulative frequencies (upper right) plots of Lotka's data in original scale. Logarithmic scaled sample frequency (lower left) and logarithmic scaled complementary cumulative (lower right) plots.

graphically representing Chen's data, both in the original and log-log coordinates, are given in Figure 2.3.

C: Moby Dick Data

This data set, consisting of word frequencies in the novel "Moby Dick: The Whale" by American writer Herman Melville, is a classic example for text analysis. The Moby Dick data set used in this thesis can be accessed directly in R under the name "moby" within the `poweRlaw` package (Gillespie, 2015).

The data set includes unique words as sources and their time of occurrence as items. The novel consists of 18,855 unique words, with a total of 245,567 occurrences. The data set is provided in the form of occurrences of each word,

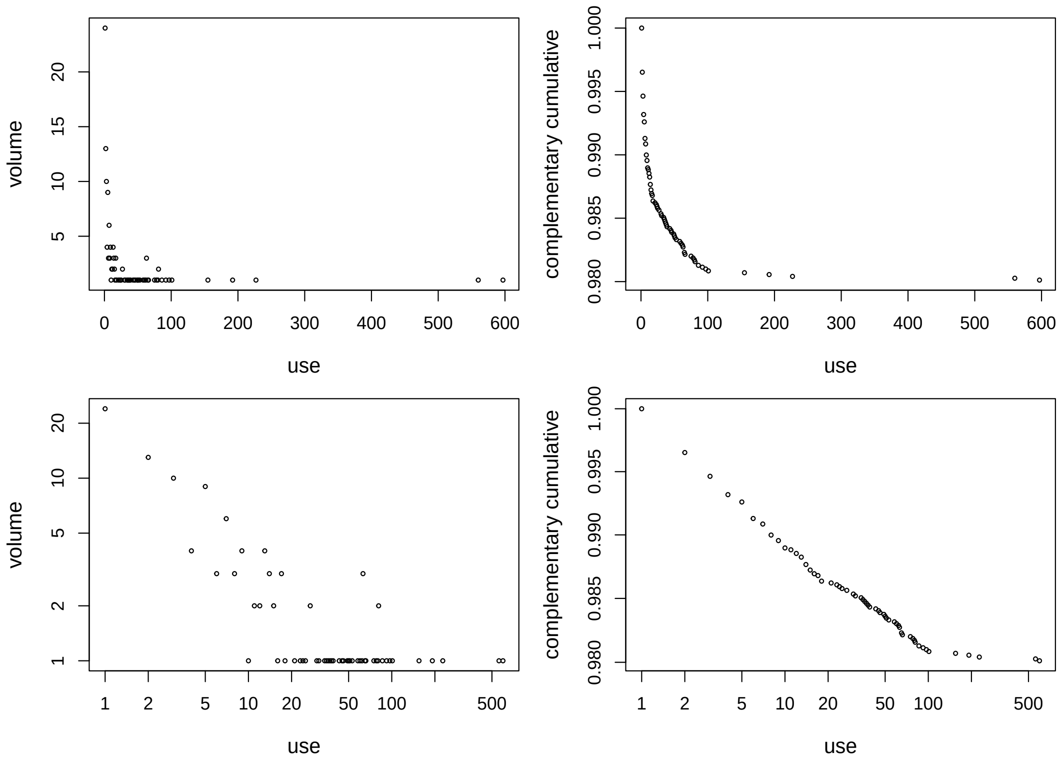


Figure 2.3: Chen's data: frequency plots (left column) and complementary cumulative frequency plots (right column) in the original (top row) and log-log (bottom row) coordinates.

but one may apply `table(moby)` in R to obtain the data shown in an aggregated way. Of these words, 9,161 words occurred only once, and 3,085 occurred twice. The most used word occurred 14,086 times. Note that the study in this thesis focuses on the frequency distribution of the count data, thus not accounting for specific counts of particular words. If relevant, the latter information is available through an online tutorial by Bonnell & Ogihara (2023).

The frequencies and complementary cumulative plots of the Moby Dick data, both in the original and the log-log coordinates, are depicted in Figure 2.4.

D: EJP Data

This data was collected by the author of this thesis on 27th January, 2020. It comprises citations of each publication of 113 authors who pub-

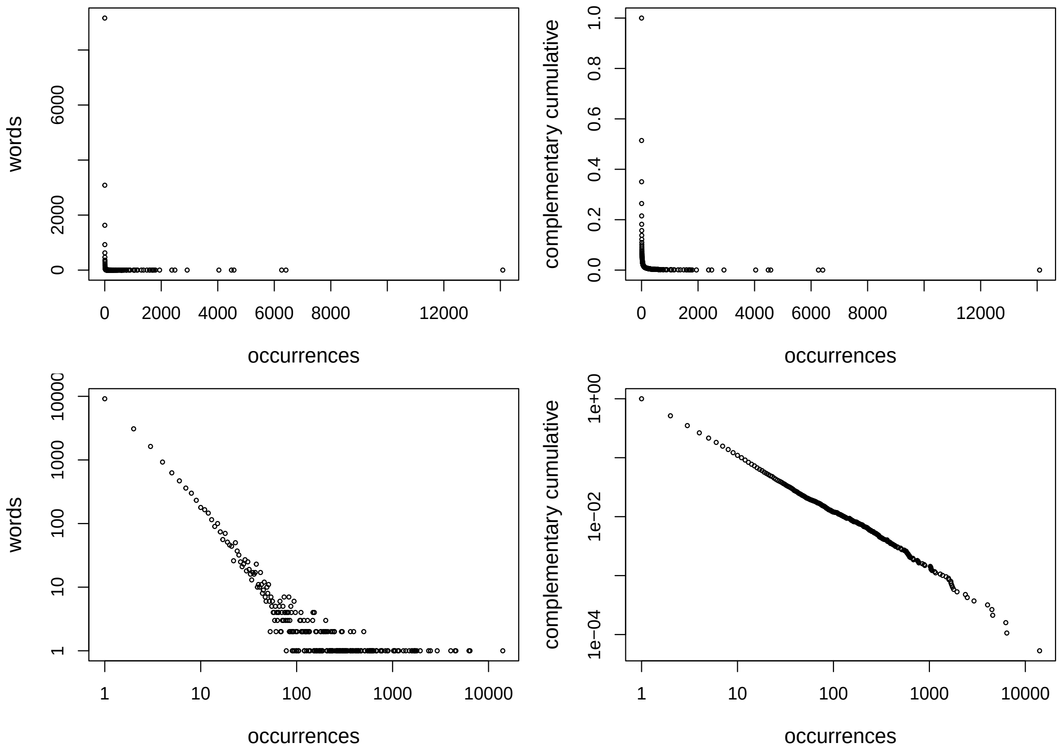


Figure 2.4: Moby Dick data: frequency plots (left column) and complementary cumulative frequency plots (right column) in the original (top row) and log-log (bottom row) coordinates.

lished at least one paper in the first 10 issues of the *Electronic Journal of Probability* (EJP), volume 24 (2019) (`https://projecteuclid.org/journals/electronic-journal-of-probability/volume-24/issue-none`) and who are also featured on Google Scholar (`https://scholar.google.com/`). This data set contains 245,567 citations and 15,400 publications of 113 authors in total.

The citation score of each publication was obtained by using an R command `get_publications()$cites` in the package `scholar`, after collecting the Google Scholar IDs of authors; here, the Google Scholar ID should be entered inside the parentheses in the template command (Yu *et al.*, 2016). In turn, Google Scholar IDs were identified using *web scraping* techniques on Google Scholar web pages. For more details about web scraping, see Appendix A.

The complete EJP data set, with citation numbers per paper per individual author, is available online from the GitHub (see `https://github.com/`

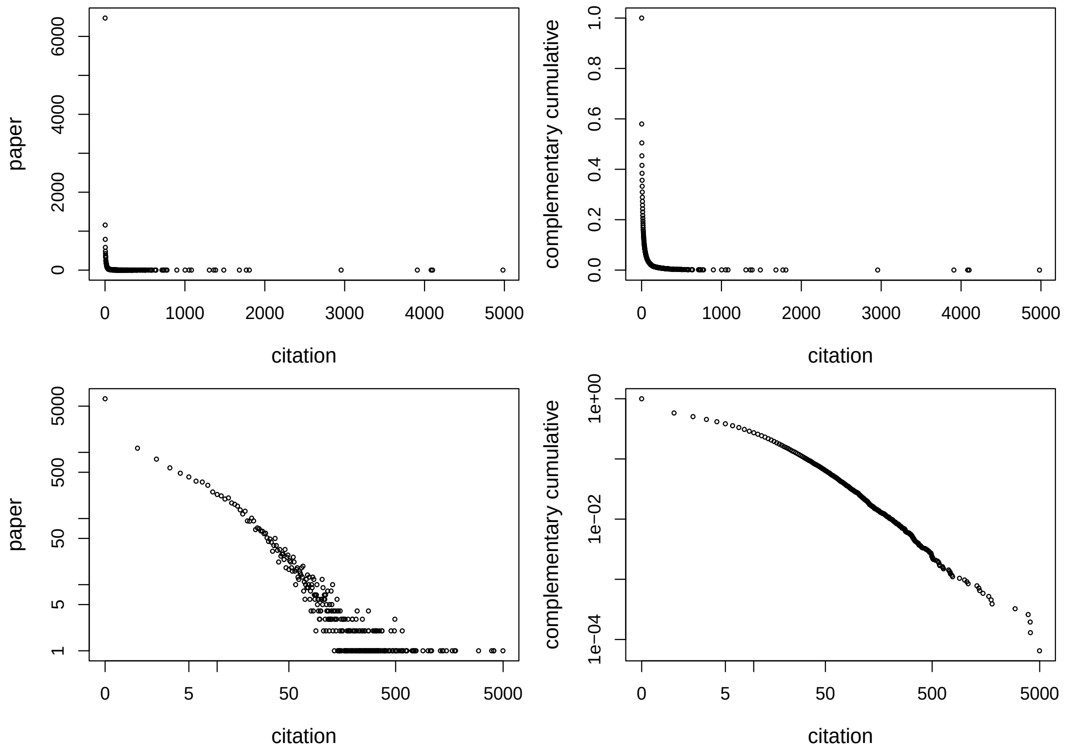


Figure 2.5: EJP data: frequency plots (left column) and complementary cumulative frequency plots (right column) in the original (top row) and log-log (bottom row) coordinates.

Ruheyan/Citation-data/blob/main/data/cejop.csv), listing authors and the citation numbers of each of their papers. Authors and papers are anonymised, as shown in Table 2.1. For example, the first paper of author #1 is cited 1,486 times, the second paper of the same author is cited 506 times, etc. The last paper of author #113 has no citations. The aggregated EJP data, listing the citation counts of the pooled population of all papers (Table 2.2), is available from https://github.com/Ruheyan/Citation-data/blob/main/data/ejop_citations_frequency.csv.

The frequency and complementary cumulative frequency plots for the EJP data are shown in Figure 2.5, both in the original and log-log coordinates.

Table 2.1: The structure of the EJP data set, listing all authors with citations of each of their paper.

Author	Paper	Citations
1	1	1,486
1	2	506
1	3	475
1	4	241
⋮	⋮	⋮
113	110	0
113	111	0

Table 2.2: Structure of the aggregated EJP data showing the numbers of papers per numbers of citations.

Citations	Papers
0	6,472
1	1,157
2	790
3	583
⋮	⋮
4,100	1
4,981	1

D′: Extended EJP Data (with covariates)

The extended EJP data consists of yearly citation data for authors included in the EJP data. The data was retrieved from the Web of Science (WoS) by the author of this thesis in January 2020. The data set includes a total of 3,588 publications from 111 authors. The identities of authors are consistent with those appearing in the EJP data, except for two authors who lack a Web of Science page. It is worth noting that the data set excludes all books and prefaces. The WoS archives citation data from 1900 until the date of collection (18th September, 2022). Additionally, the data set includes the publication year, total citations, number of pages, and the number of authors for each publication. The titles of the papers are labelled by numbers. The data set is available online at https://github.com/Ruheyan/Citation-data/blob/main/data/Extended_EJP_Data.csv.

E: AMS Data

This data was collected by the author of the present thesis in March 2021 with the aid of the Google Scholar, using the same techniques as with the EJP data. The authors in this data set are the academic members (excluding PhD students) of the American Mathematical Society (AMS), https://www.ams.org/. The initial list of the current members was provided by Dr Leonid Bogachev via accessing the AMS online database through his AMS membership. Overall, this data set comprises 3,089 authors with the total of 316,361 papers and 12,351,608 citations. The complete data set per individual authors is available online at https://github.com/Ruheyan/Citation-data/blob/main/data/cAMS.csv. An aggregated version, with paper counts versus citations, is stored online at https://github.com/Ruheyan/Citation-data/blob/main/data/AMS_citations_frequency.csv.

Frequency plots of citations of papers in the AMS data are depicted in Figure 2.6, both in the original and log-log coordinates.

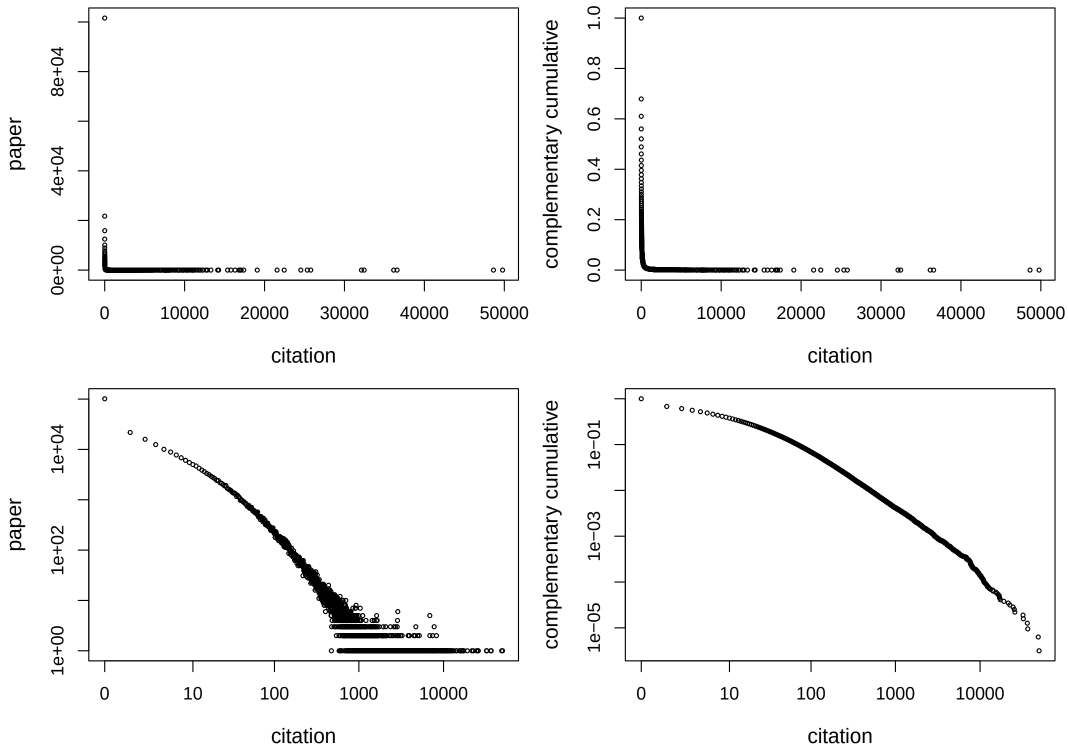


Figure 2.6: AMS data: frequency plots (left column) and complementary cumulative frequency plots (right column) in the original (top row) and log-log (bottom row) coordinates.

2.3 Scientometric Databases

This subsection compares three mainstream platforms for obtaining scientometric data, and explains the reasons for using the Google Scholar for collecting the EJP and the AMS data. Furthermore, other databases are provided at the end of this subsection.

The three most popular online databases for bibliometric data are as follows:

- Google Scholar (https://scholar.google.com/)
- Scopus (https://www.scopus.com/freelookup/form/author.uri)
- Web of Science (https://www.webofscience.com/wos/author/search)

These three databases display scientometric indicators, such as publications, citations, the h-index etc. For the definition of the h-index and more detail, see Section 2.4.1.

However, various websites include different data on the publication profile of the same authors. Table 2.3 illustrates the differences among websites using Stephen Hawking and Alan Turing as examples. It should be noted that Google Scholar does not directly list the total number of papers of an author. Instead, users can click on 'show more' to view the entire publication list and find the publication count at the centre of the bottom.

Table 2.3: Scientometric data of Stephen Hawking and Alan Turing featured in different data platforms (snapshot taken on 13 December 2022).

Author	**Stephen Hawking**			**Alan Turing**		
	citations	papers	h-index	citations	papers	h-index
Google Scholar	141,235	1,024	130	58,233	301	43
Scopus	50,161	160	77	704	18	5
Web of Science	51,031	160	83	12,557	38	10

Different collection methods reason for the differences among these databases. Google Scholar uses an automated approach to collect citation data, which can cover all the related data of a publication without many restrictions on publication type and published time, and it updates fast. However, it may cause some technical errors, such as duplication. Scopus and Web of Science have their experts using their selective criteria. For including a paper, the journal where the paper is published needs to be included in the databases first, so some citations are not counted or take a longer time to be included.

In terms of citation coverage, according to Martín-Martín *et al.* (2018), the citation coverage of Google Scholar is broader than Scopus and Web of Science in all disciplines. Citations on Google Scholar are not only limited to journal articles but also include theses, preprints, books etc., while Scopus and Web of Science are journal based. Furthermore, Google Scholar is more sensitive than Scopus and Web of Science on non-English publications (Martín-Martín *et al.*, 2018).

Google Scholar has no restrictions on the publication time or items that can be viewed, but it does not provide download access. Scopus only allows ten documents to be viewed without signing in, and after signing in, only citation information for 2,000 documents can be downloaded at once. Scopus provides yearly citation data, which can be downloaded for 200 documents at once, and also provides graphical outputs of author analysis and citation overview. Web of Science also provides citation reports and yearly citation data, and it allows up to 1,000 records to be downloaded at once, which is better than Scopus. Additionally, citation data from Scopus starts from 1970, while that of Web of Science starts from 1900. The yearly citation data is useful for Chapter 7, where the time evolution of citations is researched, and data is downloaded from the Web of Science.

After comparing these three databases (as summarised in Table 2.4), the data from Google Scholar was selected for use. In both time coverage and publication types coverage, Google Scholar contains the most complete and informative citation data for each author. Additionally, the software `Publish or Perish` for analysing citation also uses Google Scholar data, which gives more confidence in choosing Google Scholar as our database for this thesis. Although collecting citation data from Google Scholar is more complicated than from the other two databases, we decided to face the challenge.

Table 2.4: Comparison of three data platforms on citation data: Google Scholar, Scopus, and Web of Science.

Platform	**Advantages**	**Disadvantages**
Google Scholar	• Broad coverage • No time restriction • Fast updates • Non-English publications included	• Downloading not available • Technical errors
Scopus	• Downloading available (registration required) • Annualised data available	• Limited coverage • Data only after 1970
Web of Science	• Downloading available • Annualised data available	• Limited coverage • Data only after 1900

2.4 Some Metrics of Scientific Production

This section introduces two most frequently used citation metrics in the citations analysis, the h-index and the g-index.

2.4.1 The h-index

The h-index introduced by Hirsch (2005) is defined as the maximum number h of an author's papers, each one cited at least h times. For illustration, Figure 2.7 shows the h-indexes of the authors from the EJP and the AMS data sets against their citation scores, N. The dependence appears to be of a square-root type, $h = a\sqrt{N} + b$; to check this out, a linear regression model was used, giving the estimates $\hat{a} \doteq 0.4146$, $\hat{b} \doteq 3.0576$ (EJP) and $\hat{a} \doteq 0.4146$, $\hat{b} \doteq 2.4116$ (AMS); see the solid red lines in Figure 2.7. This empirical observation will be further discussed in Section 3.6.1 in the general context of the item production model and in Section 4.5.2 with regard to fitting the integer partition model to citation data.

The online platforms such as Google Scholar, Scopus and Web of Science (see Section 2.3) provide information about an author's h-index. This metric can serve

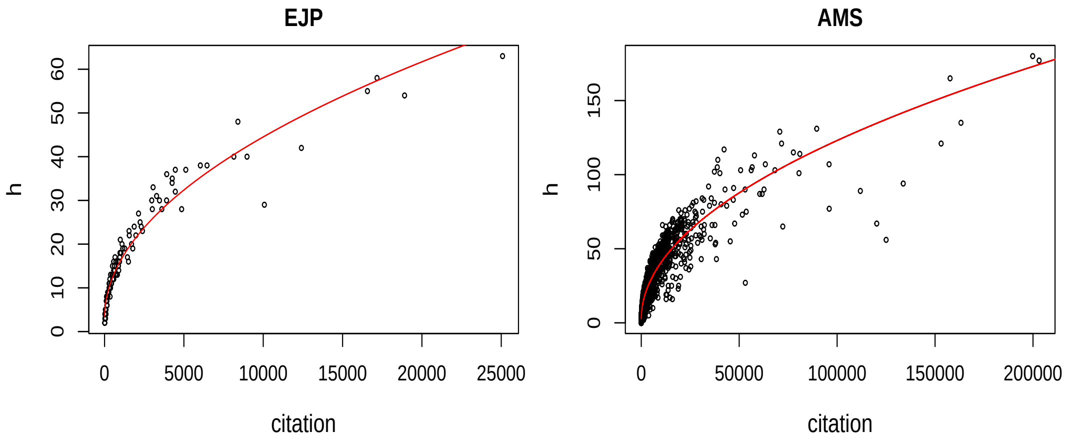


Figure 2.7: The h-index versus total citations: EJP data (left) and AMS data (right). The red lines are fitted via regression of the form $h = a\sqrt{N} + b$, where N is the total number of citations. Specific fits are $h \approx 0.4146\sqrt{N} + 3.0576$ (EJP) and $h \approx 0.3815\sqrt{N} + 2.4116$ (AMS).

as a benchmark for evaluating faculty recruitment, promotion, and scholarship. Its potential applications in research make it a valuable tool in the fields of discrete probability and statistics in the social sciences. Furthermore, it involves only a simple calculation that can be carried out by knowing the number of papers and their respective citation counts, resulting in a single numerical value.

In a subsequent development, Hirsch (2007) demonstrated the superior performance of the h-index in predicting scientific achievement among physicists, compared to other measures such as the total number of published papers, total citations garnered, and mean number of citations per year. This research placed a greater emphasis on predicting future scientific achievement, with less consideration given to future citations of previous papers.

However, despite the h-index has become one of the commonly accepted metrics of scientific productivity, disputes and discussions about the utility and faithfulness of the h-index are still prevalent. The following issues have been noted in the literature with regard to the use of the h-index.

- **Disciplines**

  Hirsch (2005) initially focused his study of the h-index in the physical sciences, commenting that it may be necessary to analyse different disciplines separately. Since then, numerous scholars have conducted related research. However, some objections have been raised. Radicchi *et al.* (2008) presented a universal curve that allows for the rescaling of h-indices across different scientific disciplines. As a result, there is a question as to whether the classification of disciplines should be taken into account when studying the h-index.

- **Co-authorship**

  The term “co-authors” refers to individuals who collaborated in writing a paper together. According to Yong (2014), co-authors impact the accuracy of h-index estimations since it does not consider authorship order. In response to this issue, Hirsch (2019) introduced a new index, h_α, which aims to measure leadership among co-authors.

- **Self-citations**

  The h-index is susceptible to manipulation through self-citations. Although self-citations can clearly boost the h-index, their impact is relatively small when compared to the total h value of a scientist (Hirsch, 2005). Ideally, scientists should eliminate self-citations when calculating their citation metrics.

- **Book citations**

  Citations of books (especially textbooks) may also be worth removing when calculating or estimating the h-index (Yong, 2014). The rationale here is that books may be cited for reasons different from citing research papers, especially with regard to well-known books of famous academics. Using an estimation method based on the limit shape of integer partitions, Yong (2014) argued that the estimation often becomes more accurate if textbook citations are removed. For example, for a prominent combinatorialist R. P. Stanley, with 6,510 citations (by 2014), the estimated h-index of 43.6 has a 20% error as compared to the actual value of 35; however, after subtracting 3,362 citations from textbooks, a revised h-index is 32, while the adjusted estimate is 30.3. In contrast, when dealing with more homogeneous subsets of researchers, the results are often less sensitive to books;

for example, for mathematicians in the National Academy of Sciences of the USA, Yong (2014) reports that the correlation coefficient between the sample and estimated h-indexes with and without books only improves from 0.94 to 0.95.

The general preference in the community is to exclude books; for example, according to the web guidance by the University of Waterloo (see `https://subjectguides.uwaterloo.ca/calculate-academic-footprint/YourHIndex`), citations of books and conference proceedings should be removed when calculating the h-index, since these records are not well represented in scientometric data platforms.

- **Conference proceedings**

  In computer science, conference proceedings play a major role in updating cutting-edge algorithms and techniques, while other subjects, such as mathematics, may not consider conference proceedings and other book collections as valuable as journal publications.

Hirsch (2005) mentioned that the value of the h-index increases with the academic age A of scientific research, that is, $h \approx \beta A$. In his opinion, we can measure the achievement of scientists by the value of β. When $\beta \approx 1$, one is a successful scientist, when $\beta \approx 2$, he/she is an outstanding scientist, and when $\beta \approx 3$, the person is a truly unique individual. Hirsch collected data from 1985 to 2005 of physicists who obtained the Nobel Prize, and the average h-index of them is 41, and the average value of β is 1.14 (Hirsch, 2005).

Let T be the time elapsed after an author's first paper is published; this time may be interpreted as the individual's *academic age*. Following Hirsch (2005), Figure 2.8 illustrates the relationship between an author's academic age T and their h-index. The time T is measured in years, with an author's first publication assigned a value of $T = 0$, and subsequent years incremented accordingly. The data presented in Figure 2.8 corresponds to the h-index of a randomly selected author from the EJP data set (Section 2.2, D). The increasing h-index of this randomly chosen author over time is evident, with a slope of approximately $\beta = 1.257 \approx 1$, indicating a successful scientific career according to Hirsch's definition.

Note that the dynamic data was collected by the author of this thesis from the Web of Science, since Google Scholar does not provide annualised data.

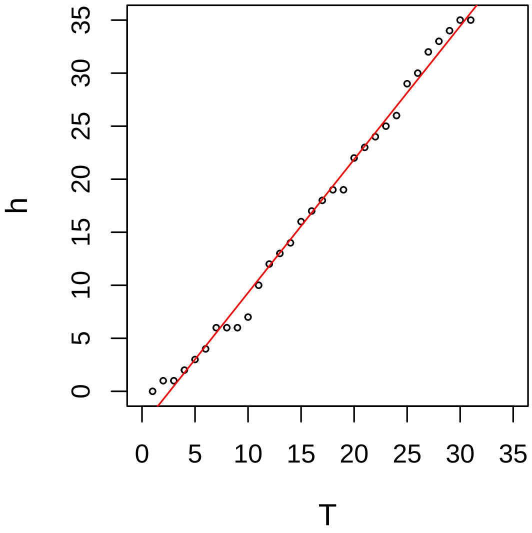


Figure 2.8: Growth of the h-index with the academic age (defined as the number of years in research) for authors in the EJP data set. The slope of the straight line (fitted via linear regression) is 1.257.

2.4.2 The g-index

The h-index focuses on h most cited papers with at least h citations each, but the citation counts larger than h are censored. To address this limitation, Egghe (2006) proposed the *g-index*, calculated by arranging an author's papers in decreasing order of generated citations and determining the largest value of g such that the sum of citations of the top g papers is at least g^2. Figure 2.9 using the EJP and the AMS data sets, illustrating the scatter plot of the g-index as a function of citations N, and also its relation with the h-index.

The g-index and corresponding citation counts tend to a square-root relation, i.e., $g = a\sqrt{N} + b$; to check this out, a linear regression was used, giving the esimates $\hat{a} \doteq 0.92417$, $\hat{b} \doteq 0.03961$ (EJP) and $\hat{a} \doteq 0.9235$, $\hat{b} \doteq -0.7075$ (AMS); these are depicted in solid red lines in Figure 2.9. Same as the h-index, this empirical observation will be further discussed in both in the general context of the item production model (see Section 3.6.2) and with regard to fitting the integer partiton model to citation data (see Section 4.5.2).

The h-index and the g-index tend to have a linear relation, i.e., $g = \alpha h + \beta$, checking through linear regression, we obtain $\hat{\alpha} \doteq 2.151$, $\hat{\beta} \doteq -5.409$ (EJP) and $\hat{\alpha} \doteq 2.240$, $\hat{\beta} \doteq -3.166$ (AMS). These are illustrated in Figure 2.9 in solid red lines.

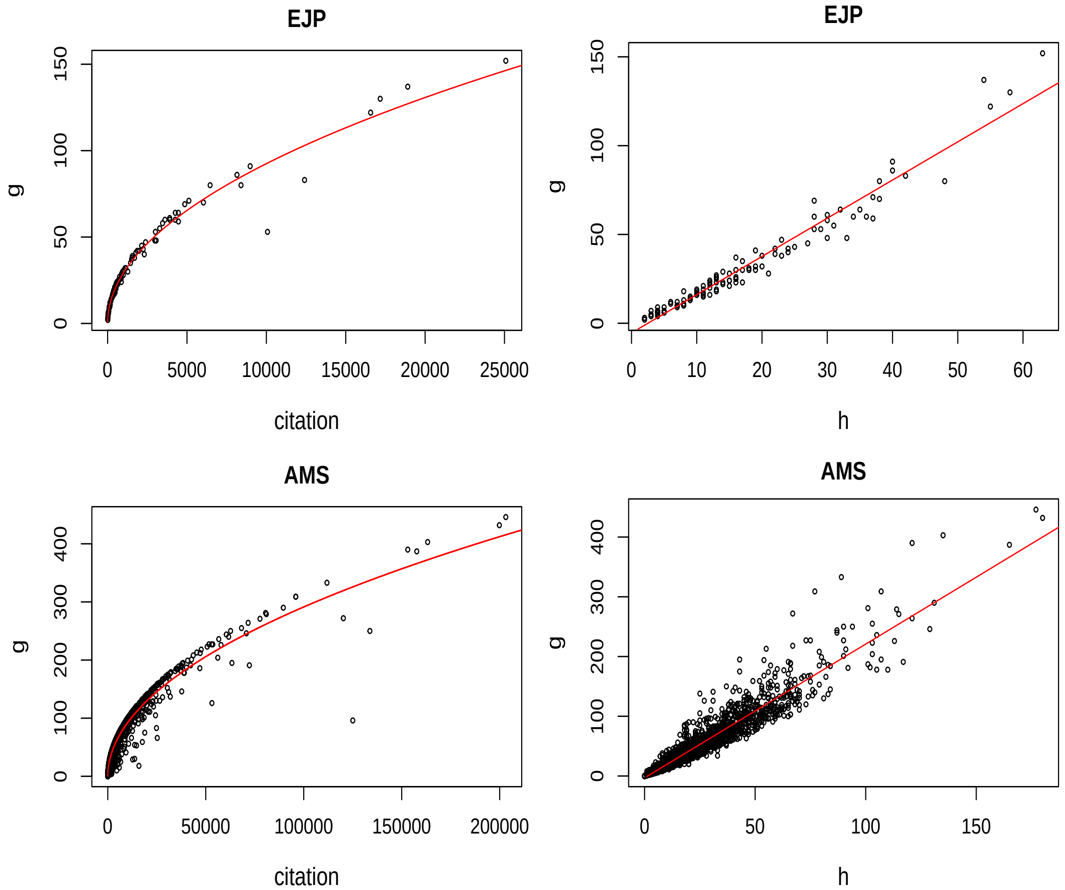


Figure 2.9: Illustartion of g-index using EJP data (top row) and AMS data (bottom row). Left panels show scatter plots of the g-index versus the total number of citation N. The red lines show fitted lines via regression $g = a\sqrt{N} + b$, yielding $g \approx 0.92417\sqrt{N} + 0.03961$ (EJP) and $g \approx 0.9235\sqrt{N} - 0.7075$ (AMS). Right panels display scatter plots of g versus the h-index, with red lines depicting linear regression fits, $g \approx 2.151\,h - 5.409$ (EJP) and $g \approx 2.240\,h - 3.166$ (AMS).

2.5 Possible Ethical Issues

In the context of research involving personal details, such as gender, age, and personal address, ethical and data protection issues are crucial. This also applies in scientometric research, where personal details may be included in the data. However, this thesis does not require an ethical approval review since publications and citation data are publicly available. Nonetheless, according to the course on Introduction to Research Ethics attended by the author on September 19, 2019, personal data that are not publicly available must be either anonymous or confidential. Anonymity refers to the inability to identify a person from the information provided, while confidentiality ensures that only a limited number of individuals can access data containing personal information. These ethical considerations become even more critical when research involves children's data.

In most of our studies, personal names were omitted from the scientometric data analysis, to make sure there are no ethical issues arising. However, it is still possible and legitimate to retrieve the author's identity from the data in cases of research interest, such as outliers in a data set. For instance, the data set presented in Lotka (1926) did not include authors' names, but the data source was known (i.e., the index of *Chemical Abstracts*), which would enable a researcher to find the relevant information if required, by referring to the original source. For example, an outlier in Lotka's data identified due to the GIGP fitting is discussed in Section 5.5.1. As another example, a specific paper from the EJP data set, with over 2,300 authors (!), is discussed in Section 7.2.5.

Chapter 3

Item Production Model: Mathematical Setup

This chapter aims to establish the foundation for the present thesis by setting up the basic model for the data. The model enables later chapters to investigate under different assumptions.

3.1 Item Production Model

3.1.1 Context and motivation

In informetrics, it is customary to use the terms “sources” and “items” to describe the count data (Egghe, 2005). Informally, “sources” produce “items”; for example, in Lotka’s data set authors produce papers. Likewise, in Chen’s data set, journal volumes are sources and their individual uses are items, while in the Moby Dick data set, different words are sources and their occurrences are items.

In the context of scientometrics, there are three pillars of the data depending on the focus of study — authors producing papers which, in turn, are producing citations. All of these features are present in our EJP and AMS data sets (see Section 2.2). The three “sources-items” relations arising here are as follows. Firstly, authors (sources) produce papers (items); on the other hand, papers (sources) produce citations (items). Finally, authors may be interpreted as sources directly producing citations as items. These relations are symbolically depicted

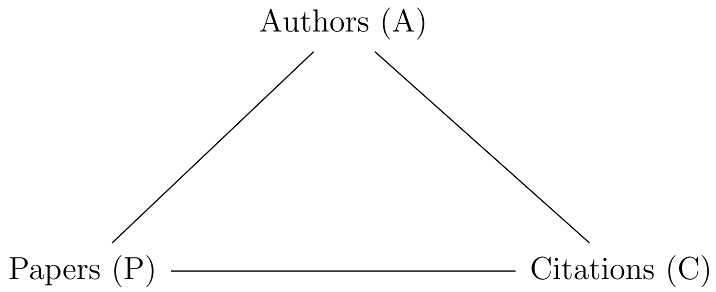


Figure 3.1: Authors-papers-citations (APC) triangle.

in Figure 3.1. In the present thesis, we are mostly interested in the relation "papers–citations", in particular due to the relevance of the h-index.

Table 3.1 displays the summary of the data sets A–E described in Section 2.2 by highlighting the suitable interpretation of sources and items.

Table 3.1: Summary of the data sets A to E introduced in Section 2.2

Data set	sources – items	sources	items	items per source
A (Lotka's)	authors – papers	6,891	22,939	3.328835
B (Chen's)	volumes – uses	138	4,292	31.10145
C (Moby Dick)	words – occurrences	18,855	209,994	11.13731
D (EJP)	papers – citations	15,400	245,567	15.94591
E (AMS)	papers – citations	316,361	12,351,608	39.04276

3.1.2 Outputs and multiplicities

Suppose there are M sources, each one producing a batch of items, and let X_i denote the random size of the batch produced by the i-th source ($i = 1, \dots, M$). The range of the output size can be $j \in \mathbb{N}_0$ if empty output is allowed (e.g., citations of a paper), or it can be zero truncated, with $j \in \mathbb{N}$ (e.g., papers of an author). The sources are independent of one another and their random outputs follow a common frequency distribution (f_j), that is, the random variables (X_i) are mutually independent and, for each $i = 1, \dots, M$,

$$\mathsf{P}(X_i = j) = f_j \qquad (j \in \mathbb{N}_0).$$

Remark 3.1. To streamline the notation, we keep writing $j \in \mathbb{N}_0$, wherein the zero-truncated case is included with $f_0 = 0$.

Remark 3.2. The *item production model* introduced above can be rephrased as the classic *occupancy problem*, dealing with independent allocation of M particles over infinitely many boxes with probability distribution (f_j) (Gnedin *et al.*, 2007).

We assume that the distribution (f_j) has finite mean,

$$\mu := \mathsf{E}(X_i) = \sum_{j=0}^{\infty} j f_j < \infty. \tag{3.1}$$

The total (random) number of produced items is given by the sum of the outputs,

$$N = \sum_{i=1}^{M} X_i, \tag{3.2}$$

with the expected value

$$\mathsf{E}(N) = \sum_{i=1}^{M} \mathsf{E}(X_i) = M\mu. \tag{3.3}$$

It is useful to represent each X_i via "scanning" across the range of possible values j,

$$X_i = \sum_{j=0}^{\infty} j I_{\{X_i=j\}}, \tag{3.4}$$

where I_A denotes the indicator of event A (i.e., with values 1 if A occurs and 0 otherwise). Of course,

$$\mathsf{E}\big(I_{\{X_i=j\}}\big) = \mathsf{P}(X_i = j) = f_j \qquad (j \in \mathbb{N}_0). \tag{3.5}$$

Consider the multiplicity M_j of output size $j \in \mathbb{N}_0$ in the pooled production of items (X_i),

$$M_j := \#\big\{i \in \{1, \dots, M\} \colon X_i = j\big\} = \sum_{i=1}^{M} I_{\{X_i=j\}} \qquad (j \in \mathbb{N}_0). \tag{3.6}$$

Using (3.5), we find the expectation

$$\mathsf{E}(M_j) = \sum_{i=1}^{M} \mathsf{E}\big(I_{\{X_i=j\}}\big) = M f_j \qquad (j \in \mathbb{N}_0). \tag{3.7}$$

Note that the random variables (M_j) are not independent; indeed, they sum up to the number of sources,

$$\sum_{j=0}^{\infty} M_j = \sum_{j=0}^{\infty} \sum_{i=1}^{M} I_{\{X_i=j\}} = \sum_{i=1}^{M} \sum_{j=0}^{\infty} I_{\{X_i=j\}} = \sum_{i=1}^{M} 1 = M.$$

From the interpretation of the multiplicities M_j, it is evident that the total (random) number of produced items is given by

$$N = \sum_{j=0}^{\infty} j\, M_j. \tag{3.8}$$

The same can be easily obtained using definition (3.2) and decompositions (3.4) and (3.6),

$$N = \sum_{i=1}^{M} X_i = \sum_{i=1}^{M} \sum_{j=0}^{\infty} j\, I_{\{X_i=j\}} = \sum_{j=0}^{\infty} j \sum_{i=1}^{M} I_{\{X_i=j\}} = \sum_{j=0}^{\infty} j\, M_j.$$

The expected value of N can then be expressed using (3.7) and (3.1),

$$\mathsf{E}(N) = \sum_{j=0}^{\infty} j\, \mathsf{E}(M_j) = M \sum_{j=0}^{\infty} j f_j = M\mu, \tag{3.9}$$

which is, of course, the same as (3.3).

Remark 3.3. In view of formulas (3.3) and (3.9), the sample mean $\hat{\mu} = N/M$ is an unbiased estimator of the expected value μ, possessing all standard properties such as consistency and asymptotic normality. The advantage of this estimator is that it is *non-parametric*, in the sense that it does not require knowledge of any distributional model (f_j) behind the production output data.

Example 3.1. To illustrate this notation, consider a mock example of citation data. We interpret papers as sources and citations as items produced by sources. Suppose there is a single author with $M = 8$ papers and the following citation counts for each of these papers:

i	1	2	3	4	5	6	7	8
X_i	2	1	4	2	0	1	0	0

The total number of citations is

$$N = \sum_i X_i = 2 + 1 + 4 + 2 + 0 + 1 + 0 + 0 = 10.$$

The same data can be represented using the multiplicities of specific citations:

j	0	1	2	3	4	≥ 5
M_j	3	2	2	0	1	0
M_j/M	0.375	0.250	0.250	0.000	0.125	0.000

Here, the multiplicity M_j is the number of papers with exactly j citations ($j = 0, 1, 2, \dots$). In particular, $M_0 = 3$ because there are three papers that have not been cited, and $M_3 = 0$ because none of the papers have been cited three times. The ratios M_j/M are the relative frequencies of having j citations, respectively.

3.1.3 The likelihood

Suppose that the theoretical frequencies (f_j) depend on one or more model parameters, $f_j = f_j(\boldsymbol{\theta})$, where $\boldsymbol{\theta} = (\theta_1, \dots, \theta_r) \in \mathbb{R}^r$. With the observed data $\boldsymbol{X} = (X_1, \dots, X_M)$ consisting of independent values with common distribution $(f_j(\boldsymbol{\theta}))$, the likelihood is given by the product rule,

$$\mathcal{L}(\boldsymbol{\theta}; \boldsymbol{X}) = \prod_{i=1}^{M} f_{X_i}(\boldsymbol{\theta}). \tag{3.10}$$

The maximum likelihood estimate of the vector parameter $\boldsymbol{\theta}$ can then be obtained either by maximising the function (3.10) directly (e.g., using the R command `optim`) or by solving numerically the likelihood equation $\partial\mathcal{L}/\partial\boldsymbol{\theta} = \boldsymbol{0}$, or more explicitly, the set of equations

$$\frac{\partial\mathcal{L}}{\partial\theta_k} = 0, \qquad k = 1, \dots, r.$$

As usual, it may be more convenient to work with the log-likelihood $\ell = \log\mathcal{L}$,

$$\ell(\boldsymbol{\theta}; \boldsymbol{X}) = \sum_{i=1}^{M} \log f_{X_i}(\boldsymbol{\theta}), \tag{3.11}$$

which is maximised by solving the corresponding log-likelihood equations

$$\frac{\partial \ell}{\partial \theta_k} = 0, \qquad k = 1, \dots, r.$$

Clearly, the order of terms in the product formula (3.10) (as well as in the sum formula (3.11)) is not important, so it can be rewritten using the order statistics

$$X_{1,M} \geq \cdots \geq X_{M,M},$$

obtained by arranging the sample terms X_i in non-increasing order. In particular,

$$X_{1,M} = \max\{X_i, i = 1, \dots, M\}, \qquad X_{M.M} = \min\{X_i, i = 1, \dots, M\}.$$

Thus, formula (3.10) takes the form

$$\mathcal{L}(\boldsymbol{\theta}; \boldsymbol{X}) = \prod_{i=1}^{M} f_{X_{i,M}}(\boldsymbol{\theta}). \tag{3.12}$$

Of course, this does not cause any loss of information about the unknown parameter $\boldsymbol{\theta}$, which means that the collection of order statistics $\{X_{i,M}\}$ is a *sufficient statistic* for $\boldsymbol{\theta}$, also confirmed by the factorisation rule (Garthwaite *et al.*, 2002, Theorem 2.1, p. 21). A reduction of stored information here is that we do not need to know each individual output X_i, but only an ordered collection of these outputs.

Example 3.2. The original EJP data set stores the sample $\boldsymbol{X}$, which is citations of each author's papers without any particular ordering. A reduced version of the data set is achieved by storing the multiplicities of the citations (M_j), which can also return the data to the ordered counts of citations $\{X_{i,M}\}$. The reduction in storage applies to the AMS data set as well.

In turn, collecting the counts of the order statistics $X_{i,M}$ with the same sample value $j \in \mathbb{N}_0$, expression (3.12) can be rewritten in terms of the multiplicities (M_j),

$$\mathcal{L}(\boldsymbol{\theta}; \boldsymbol{X}) = \prod_{j=0}^{\infty} f_j(\boldsymbol{\theta})^{M_j}, \tag{3.13}$$

which again appears to follow a product rule despite the fact that the multiplicities (M_j) are not independent. Formally, the expression (3.13) can be verified by expressing the frequencies in the form

$$f_x(\boldsymbol{\theta}) = \prod_{j=0}^{\infty} f_j(\boldsymbol{\theta})^{\mathbf{1}_{\{x\}}(j)},$$

where $\mathbf{1}_D(x)$ is the indicator function of set D (i.e., $\mathbf{1}_D(x) = 1$ if $x \in D$ and $\mathbf{1}_D(x) = 0$ otherwise). Substituting this into (3.10), we obtain

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta}; \boldsymbol{X}) &= \prod_{i=1}^{M} \prod_{j=0}^{\infty} f_j(\boldsymbol{\theta})^{\mathbf{1}_{\{X_i\}}(j)} \\ &= \prod_{j=0}^{\infty} f_j(\boldsymbol{\theta})^{\sum_{i=1}^{M} I_{\{X_i=j\}}} \\ &= \prod_{j=0}^{\infty} f_j(\boldsymbol{\theta})^{M_j},\end{aligned}$$

according to (3.6). Thus, formula (3.13) follows.

The representation (3.13) is useful, because the data is often aggregated by ignoring the individual outputs of the sources, instead reporting only the observed counts M_j. Again, this does not cause any loss of information about the unknown parameter $\boldsymbol{\theta}$, because, according to the factorisation rule (Garthwaite *et al.*, 2002, Theorem 2.1, p. 21), the vector of multiplicities (M_j) is a sufficient statistic for $\boldsymbol{\theta}$.

In fact, it is easy to see that the statistics $\{X_{i,M}\}$ and (M_j) are equivalent: if we know all the terms $\{X_{i,M}\}$ then we can calculate the counts (M_j), and vice versa. The choice of a particular representation of the likelihood depends on the convenience of data handling in a given format of the data set.

3.1.4 Likelihood of censored data

Sometimes the interest is in modelling a *truncated range* of observed frequencies starting from some threshold, say $j \geq j_*$. For example, this situation frequently arises when fitting a power law (see Section 4.2). In that case, the aim is to fit a conditional model $f_j^{\wedge} = \mathsf{P}(X_i = j \,|\, X_i \geq j_*)$ to the observed counts $X_i = j \geq j_*$.

Denoting the probability of threshold exceedance by $\rho = \mathsf{P}(X_i \ge j_*)$, this leads to a *likelihood of censored data*

$$\begin{aligned}\mathcal{L}(\rho, \boldsymbol{\theta}; \boldsymbol{X}) &= \prod_{i=1}^{M} (1-\rho)^{I_{\{X_i<j_*\}}} \, \big(\rho \, f^{\wedge}_{X_i}(\boldsymbol{\theta})\big)^{I_{\{X_i\ge j_*\}}} \\ &= (1-\rho)^{M-M(j_*)} \, \rho^{M(j_*)} \prod_{i=1}^{M} \big(f^{\wedge}_{X_i}(\boldsymbol{\theta})\big)^{I_{\{X_i\ge j_*\}}}, \qquad (3.14)\end{aligned}$$

where $M(j_*) := \sum_{i=1}^{M} I_{\{X_i \ge j_*\}}$ is the total number of observations (X_i) with values at least j_*. Note the binomial-type part in front of the product in (3.14). Accordingly, the log-likelihood $\ell = \log \mathcal{L}$ is given by

$$\begin{aligned}\ell(\rho, \boldsymbol{\theta}; \boldsymbol{X}) &= \big(M - M(j_*)\big) \log(1-\rho) + M(j_*) \log \rho \\ &\quad + \sum_{i=1}^{M} I_{\{X_i \ge j_*\}} \log\big(f^{\wedge}_{X_i}(\theta_j)\big). \qquad (3.15)\end{aligned}$$

Differentiating the log-likelihood $\ell = \log \mathcal{L}$ with respect to ρ, we obtain

$$\frac{\partial \ell}{\partial \rho} = -\frac{M - M(j_*)}{1-\rho} + \frac{M(j_*)}{\rho} = 0,$$

which immediately yields a familiar maximum likelihood estimator (MLE) for ρ of success-rate type,

$$\hat{\rho} = \frac{M(j_*)}{M}. \qquad (3.16)$$

The MLE of the vector parameter $\boldsymbol{\theta}$ is obtained as usual by solving the corresponding likelihood equations $\partial \ell / \partial \boldsymbol{\theta} = 0$.

3.2 Young Diagrams and Limit Shape

As already mentioned in Section 3.1.3 , it is useful to rank the sources according to their production output, that is, by considering the (descending) order statistics $X_{1,M} \ge X_{2,M} \ge \cdots \ge X_{M,M}$; for example, $X_{1,M} = \max_{1\le i\le M}\{X_i\}$ is the highest output score among M sources. The production profile is succinctly visualised by the *Young diagram* formed by the left- and bottom-aligned row blocks of unit height and lengths $(X_{i,M})$, with longer blocks positioned lower (see Figure 3.2).

In particular, blocks corresponding to the output value $j = 0$ (if it is allowed) degenerate to vertical intervals (of height 1 each) placed on top of the rest of the Young diagram along the vertical axis.

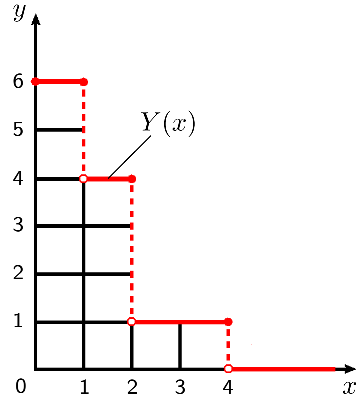


Figure 3.2: Young diagram and the boundary $Y(x)$ for $M = 6$ sources and ordered outputs $(X_{i,M}) = (4, 2, 2, 2, 1, 1)$, corresponding to counts $M_4 = 1$, $M_2 = 3$, $M_1 = 2$.

The upper boundary of the Young diagram is the graph of the (left-continuous) step function

$$Y(x) := \sum_{j \geq x} \sum_{i=1}^{M} I_{\{X_i = j\}} = \sum_{j \geq x} M_j \qquad (x \geq 0) \tag{3.17}$$

(see (3.6)).

To highlight the dependence on x, rewrite definition (3.17) in the form

$$Y(x) = \sum_{j=0}^{\infty} M_j \, \mathbf{1}_{[0,j]}(x) \qquad (x \geq 0). \tag{3.18}$$

If $M_0 > 0$ then the function $Y(x)$ has an isolated peak at $x = 0$; otherwise, $Y(x)$ is right-continuous at zero. The value at the origin is the total number of sources, i.e.,

$$Y(0) = \sum_{j=0}^{\infty} M_j = M,$$

whereas the area under the graph of $Y(x)$ equals the total number of produced items, i.e.,

$$\int_0^\infty Y(x)\,\mathrm{d}x = \sum_{j=0}^\infty M_j \int_0^\infty \mathbf{1}_{[0,j]}(x)\,\mathrm{d}x = \sum_{j=0}^\infty jM_j = N$$

(see (3.18) and (3.8)).

Setting

$$Z_i(x) := \sum_{j\ge x} I_{\{X_i=j\}} = I_{\{X_i\ge x\}} \qquad (i=1,\dots,M), \tag{3.19}$$

formula (3.17) can be expressed in the form

$$Y(x) = \sum_{i=1}^M \sum_{j\ge x} I_{\{X_i=j\}} = \sum_{i=1}^M Z_i(x). \tag{3.20}$$

The indicators $Z_1(x),\dots,Z_M(x)$ are independent and identically distributed Bernoulli random variables; specifically,

$$\mathsf{P}(Z_i(x)=1) = \mathsf{P}(X_i\ge x) = \sum_{j\ge x} f_j =: \bar F(x), \tag{3.21}$$

$$\mathsf{P}(Z_i(x)=0) = \mathsf{P}(X_i< x) = \sum_{j< x} f_j =: F(x), \tag{3.22}$$

where $\bar F(x) + F(x) = 1$ for all $x\ge 0$. Hence,

$$\mathsf{E}\bigl(Z_i(x)\bigr) = \bar F(x), \qquad \mathsf{Var}\bigl(Z_i(x)\bigr) = \bar F(x)\bigl(1-\bar F(x)\bigr) = \bar F(x)F(x).$$

Furthermore, for any $0\le x\le x'$ we have

$$Z_i(x)Z_i(x') = I_{\{X_i\ge x\}}I_{\{X_i\ge x'\}} = I_{\{X_i\ge x'\}} = Z_i(x'),$$

whence

$$\begin{aligned}\mathsf{Cov}\bigl(Z_i(x),Z_i(x')\bigr) &= \mathsf{E}\bigl(Z_i(x)Z_i(x')\bigr) - \mathsf{E}\bigl(Z_i(x)\bigr)\mathsf{E}\bigl(Z_i(x')\bigr)\\ &= \bar F(x') - \bar F(x)\bar F(x') = \bar F(x')F(x).\end{aligned}$$

It then follows easily from (3.20) that, for each $x\ge 0$,

$$\mathsf{E}\bigl(Y(x)\bigr) = M\bar F(x), \qquad \mathsf{Var}\bigl(Y(x)\bigr) = M\bar F(x)F(x). \tag{3.23}$$

and, for $0 \le x \le x'$,

$$\begin{aligned}\mathsf{Cov}\big(Y(x), Y(x')\big) &= \sum_{i,i'=1}^{M} \mathsf{Cov}\big(Z_i(x), Z_{i'}(x')\big)\\ &= \sum_{i=1}^{M} \mathsf{Cov}\big(Z_i(x), Z_i(x')\big) = M\bar{F}(x')F(x),\end{aligned} \tag{3.24}$$

A useful visual insight into the structure of the production distribution may be obtained by looking at scaled Young diagrams, with some scaling coefficients A and B,

$$\widetilde{Y}(x) = \frac{1}{B}Y(Ax) = \frac{1}{B}\sum_{j\ge Ax} M_j = \frac{1}{B}\sum_{i=1}^{M} Z_i(Ax) \qquad (x \ge 0). \tag{3.25}$$

The aim is to seek a *limit shape* $x \mapsto \varphi(x)$ such that, with suitable $A, B \to \infty$,

$$\mathsf{E}\big(\widetilde{Y}(x)\big) \to \varphi(x) \qquad (x > 0), \tag{3.26}$$

and, moreover, the random variable $\widetilde{Y}(x)$ converges to $\varphi(x)$ (in probability),

$$\widetilde{Y}(x) \xrightarrow{\mathrm{p}} \varphi(x) \qquad (x > 0). \tag{3.27}$$

Remark 3.4. The reason for restricting the range of convergence in (3.26) and (3.27) to $x > 0$ is that, in some cases, $\varphi(0) = \infty$ (e.g., see Figure 3.3).

By definition of convergence in probability, the limit (3.27) means that, for any $\varepsilon > 0$,

$$\mathsf{P}\big(\big|\widetilde{Y}(x) - \varphi(x)\big| \ge \varepsilon\big) \to 0 \qquad (x > 0). \tag{3.28}$$

Remark 3.5. It is often possible to prove a stronger limit shape result by showing uniform convergence in (3.26) and (3.28), at least away from $x = 0$ (Bogachev, 2015; Nuermaimaiti *et al.*, 2021). More precisely, this means that, for any $\delta > 0$,

$$\sup_{x\ge\delta} \big|\mathsf{E}\big(\widetilde{Y}(x)\big) - \varphi(x)\big| \to 0 \tag{3.29}$$

and, for any $\varepsilon > 0$,

$$\mathsf{P}\left(\sup_{x\ge\delta} \big|\widetilde{Y}(x) - \varphi(x)\big| \ge \varepsilon\right) \to 0. \tag{3.30}$$

A natural way to prove (3.28) is by using Chebyshev's inequality (Shiryaev, 1996, Sec. II.6, p. 192), yielding the upper bound

$$\mathsf{P}\big(|\widetilde{Y}(x) - \varphi(x)| \ge \varepsilon\big) \le \frac{\mathsf{E}\big((\widetilde{Y}(x) - \varphi(x))^2\big)}{\varepsilon^2},$$

and then decomposing the mean squared deviation $\mathsf{E}\big((\widetilde{Y}(x) - \varphi(x))^2\big)$ through the variance $\mathsf{Var}\big(\widetilde{Y}(x)\big)$ and the squared deviation $\big(\mathsf{E}\big(\widetilde{Y}(x)\big) - \varphi(x)\big)^2$. Thus, remembering the scaling (3.25) and the formulas (3.23), the proof of the limit shape result (3.28) is reduced to proving two limits,

$$\mathsf{E}\big(\widetilde{Y}(x)\big) = \frac{M\bar{F}(Ax)}{B} \to \varphi(x), \qquad \mathsf{Var}\big(\widetilde{Y}(x)\big) = \frac{M\bar{F}(Ax)F(Ax)}{B^2} \to 0. \quad (3.31)$$

Recalling that $\widetilde{Y}(x)$ is a (normalised) sum of independent indicators $Z_i(Ax) = I_{\{X_i \ge Ax\}}$, $i = 1, \dots, M$ (see (3.25)), it is natural to expect that $\widetilde{Y}(x)$ is asymptotically normal, with mean $\mathsf{E}\big(\widetilde{Y}(x)\big) = M\bar{F}(Ax)/B \sim \varphi(x)$ and variance $M\bar{F}(Ax)F(Ax)/B^2 \sim \varphi(x)/B$ (see (3.31)). However, a standard central limit theorem is not directly applicable because the "success" probability $\mathsf{P}(Z_i(Ax) = 1) = \bar{F}(Ax)$ is not constant (and, moreover, it tends to 0). We will prove such results directly for the GIGP model (Section 5.3) and the GPL model (Section 6.2.3), using the method of characteristic functions.

The notion of limit shape is motivated by similar topics in the theory of random integer partitions (Vershik, 1995, 1996). This classic example is recalled briefly in the next Section 3.3 by way of illustration, although the setting there is somewhat different from the item production model.

3.3 Example: Limit Shape of Integer Partitions

Due to the scale-free property, the power law does not change the shape after scaling, so the power law have a limit shape does not depend on scalings.

To illustrate the concept of limit shape, we start with a baseline example of the power law frequency distribution, $f_j = j^{-a}/\zeta(a)$ $(j \ge 1)$, with $a > 1$. Choose any $A \to \infty$ such that $B := M/A^{a-1} \to \infty$; that is, $1 \ll A \ll M^{1/(a-1)}$. Then the scaled expected Young diagram boundary function specialises to

$$\mathsf{E}\big(\widetilde{Y}(x)\big) = \frac{A^{a-1}}{M}\sum_{j\ge Ax}\frac{Mj^{-a}}{\zeta(a)} = \frac{1}{\zeta(a)}\sum_{j/A\ge x}\left(\frac{j}{A}\right)^{-a}\frac{1}{A} \tag{3.32}$$

$$\to \frac{1}{\zeta(a)}\int_x^\infty s^{-a}\,\mathrm{d}s = \frac{x^{-(a-1)}}{(a-1)\,\zeta(a)}, \tag{3.33}$$

using that the sum in (3.32) is the Riemann integral sum of the integral in (3.33). Thus, the limit shape exists and is given by the right-hand side of (3.33), but this is of no practical use because the scaling parameter $A\to\infty$ is arbitrary as long as $A = o(M^{1/(a-1)})$ (which confirms that the power law distribution is *scale free*).

The classic example of a frequency model possessing a meaningful limit shape comes from the theory of random integer partitions. Here, the values $j = 1, 2, \dots$ are interpreted as candidate parts into an integer partition, and the corresponding multiplicity M_j is the number of times the part j is used, respectively. In particular, if $M_j = 0$ then the value j is not involved in the partition, and it is tacitly assumed that only finitely many of M_j's are non-zero. The sum $N = \sum_{j=1}^{\infty} j\,M_j$ yields the integer being partitioned into the sum of the parts j with $M_j > 0$.

The standard model set-up there is different from the item production model described in Section 3.2. Namely, instead of the premise of M independent sources, with multiplicities (M_j) expressed by formula (3.6), the randomised partition model is defined by assuming that the multiplicities (M_j) are independent random variables with geometric distribution, $M_j \sim \mathrm{Geom}(1-z^j)$ $(j\ge 1)$, that is,

$$\mathsf{P}(M_j = m) = z^{j\,m}\,(1-z^j) \qquad (m\ge 0), \tag{3.34}$$

with the expected value given by

$$\mathsf{E}(M_j) = \frac{z^j}{1-z^j} \qquad (j\ge 1). \tag{3.35}$$

The parameter $z\in(0,1)$ is chosen specifically as

$$z = \mathrm{e}^{-\kappa/\sqrt{n}}, \qquad \kappa := \frac{\pi}{\sqrt{6}} = \sqrt{\zeta(2)}, \tag{3.36}$$

where n is an external (large) parameter.

Note that, for any $z \in (0,1)$,

$$\mathsf{P}(M_j > 0) = 1 - \mathsf{P}(M_j = 0) = 1 - (1 - z^j) = z^j,$$

and

$$\sum_{j=1}^{\infty} \mathsf{P}(M_j > 0) = \sum_{j=1}^{\infty} z^j = \frac{z}{1-z} < \infty.$$

Therefore, by the Borel–Cantelli lemma (Shiryaev, 1996, Sec. II.10, p. 255), the number of nonzero terms in the sequence of random multiplicities (M_j) is finite with probability 1.

Due to the mutual independence of M_j and the geometric marginal distributions (3.34), the probability of a given sequence of multiplicities $M_j = m_j$ $(j \ge 1)$ (with finitely many nonzero terms) is expressed as follows,

$$\mathsf{P}(M_j = m_j,\, j = 1, 2, \dots) = \prod_{j=1}^{\infty} z^{j m_j}(1 - z^j) = \frac{z^N}{G(z)}, \tag{3.37}$$

where $N = \sum_{j=1}^{\infty} j\, m_j$ and

$$G(z) = \prod_{j=1}^{\infty} \frac{1}{1 - z^j} \qquad (0 < z < 1).$$

Formula (3.37) is an instance of the so-called *Boltzmann distribution*, with roots in statistical physics (Auluck & Kothari, 1946; Vershik, 1997) and many applications in probabilistic combinatorics (Arratia *et al.*, 2003) and computing (Duchon *et al.*, 2004).

Motivation for the choice of the Boltzmann distribution (3.37) is due to the fact that its conditioning leads to the uniform distribution on the corresponding subspace. Specifically, denoting by $\mathit{\Pi}_n$ the set of all integer partitions of n, it is easy to see that the conditional probability of any partition in $\mathit{\Pi}_n$ with specific multiplicities of parts $M_j = m_j$, conditioned on $N = \sum_{j=1}^{\infty} j M_j = n$, is given by

$$\mathsf{P}\Bigl(M_j = m_j,\, j \ge 1 \,\Big|\, N = \textstyle\sum_j j\, M_j = n\Bigr) = \frac{z^n / G(z)}{(z^n/G(z)) \cdot \#\mathit{\Pi}_n} = \frac{1}{\#\mathit{\Pi}_n},$$

which is the uniform distribution on $\mathit{\Pi}_n$. Furthermore, the choice of the parameter z in the asymptotic form (3.36) is explained by the natural calibration condition

$$\mathsf{E}(N) = \mathsf{E}\Bigl(\textstyle\sum_{j=1}^{\infty} j\, M_j\Bigr) \sim n \qquad (n \to \infty). \tag{3.38}$$

Indeed, using the mean formula (3.35) and seeking the parameter z in the form $z=\mathrm{e}^{-\alpha_n}$, with $\alpha_n\to 0$, the asymptotic equation (3.38) is rewritten as

$$\mathsf{E}(N)=\sum_{j=1}^{\infty}\frac{j\,\mathrm{e}^{-\alpha_n j}}{1-\mathrm{e}^{-\alpha_n j}}=\frac{1}{\alpha_n^2}\sum_{j=1}^{\infty}\frac{\alpha_n j\,\mathrm{e}^{-\alpha_n j}}{1-\mathrm{e}^{-\alpha_n j}}\,\alpha_n\sim n. \tag{3.39}$$

Observing that the sum in (3.39) is a Riemann integral sum, it follows that

$$\begin{aligned}\sum_{j=1}^{\infty}\frac{\alpha_n j\,\mathrm{e}^{-\alpha_n j}}{1-\mathrm{e}^{-\alpha_n j}}\,\alpha_n &\to \int_0^{\infty}\frac{s\,\mathrm{e}^{-s}}{1-\mathrm{e}^{-s}}\,\mathrm{d}s\\ &=\sum_{\ell=1}^{\infty}\int_0^{\infty}s\,\mathrm{e}^{-\ell s}\,\mathrm{d}s=\sum_{\ell=1}^{\infty}\frac{1}{\ell^2}=\zeta(2)=\frac{\pi^2}{6}=\kappa^2.\end{aligned}$$

Substituting this into equation (3.39), we obtain $\alpha_n\sim\kappa/\sqrt{n}$, in line with (3.36).

The expected limit shape in the partition model can now be easily computed (Bogachev, 2015; Vershik, 1996): setting $A=B=\sqrt{n}$, we have, for any $x>0$,

$$\begin{aligned}\mathsf{E}\bigl(\widetilde{Y}(x)\bigr)=\frac{1}{B}\sum_{j\ge Ax}\mathsf{E}(M_j)&=\frac{1}{\sqrt{n}}\sum_{j\ge\sqrt{n}\,x}\frac{\mathrm{e}^{-\alpha_n j}}{1-\mathrm{e}^{-\alpha_n j}}\\ &\to\frac{1}{\kappa}\int_{\kappa x}^{\infty}\frac{\mathrm{e}^{-s}}{1-\mathrm{e}^{-s}}\,\mathrm{d}s=\frac{1}{\kappa}\sum_{\ell=1}^{\infty}\int_{\kappa x}^{\infty}\mathrm{e}^{-\ell s}\,\mathrm{d}s\\ &=\frac{1}{\kappa}\sum_{\ell=1}^{\infty}\frac{1}{\ell}\,\mathrm{e}^{-\kappa x}=-\frac{1}{\kappa}\log\left(1-\mathrm{e}^{-\kappa x}\right).\end{aligned}\tag{3.40}$$

Thus, the limit shape $y=\varphi(x)$ is given by the equation

$$y=-\kappa^{-1}\log\left(1-\mathrm{e}^{-\kappa x}\right)\qquad(x>0),\tag{3.41}$$

or, in a more symmetric form,

$$\mathrm{e}^{-\kappa x}+\mathrm{e}^{-\kappa y}=1\qquad(x,y>0),\tag{3.42}$$

where $\kappa=\pi/\sqrt{6}$ (see (3.36)). The plot of this function is shown in Figure 3.3 (red line).

Note that $\varphi(0)=\infty$. According to the calculation in (3.40), this implies that the expected value of M grows faster than $\sqrt{n}$. More precisely, we have

$$\mathsf{E}(M)=\sum_{j=1}^{\infty}\frac{\mathrm{e}^{-\alpha_n j}}{1-\mathrm{e}^{-\alpha_n j}}=\sum_{j=1}^{m}\frac{\mathrm{e}^{-\alpha_n j}}{1-\mathrm{e}^{-\alpha_n j}}+\frac{1}{\alpha_n}\sum_{j>m}^{\infty}\frac{\alpha_n\,\mathrm{e}^{-\alpha_n j}}{1-\mathrm{e}^{-\alpha_n j}},\tag{3.43}$$

where $m = [1/\alpha_n]$ and $\alpha_n = \kappa/\sqrt{n}$ (see (3.36)). Arguing as before, we see that the last sum in (3.43) converges to the integral $\int_1^\infty \mathrm{e}^{-\kappa s}\,(1-\mathrm{e}^{-\kappa s})^{-1}\,\mathrm{d}s < \infty$. Next, write

$$\sum_{j=1}^{m} \frac{\mathrm{e}^{-\alpha_n j}}{1-\mathrm{e}^{-\alpha_n j}} = \frac{1}{\alpha_n}\sum_{j=1}^{m}\frac{1}{j} + \sum_{j=1}^{m}\left(\frac{\mathrm{e}^{-\alpha_n j}}{1-\mathrm{e}^{-\alpha_n j}} - \frac{1}{\alpha_n j}\right),$$

where (Olver *et al.*, 2010, 2.10.8)

$$\sum_{j=1}^{m}\frac{1}{j} \sim \log m \sim -\log \alpha_n$$

and

$$\sum_{j=1}^{m}\left(\frac{\mathrm{e}^{-\alpha_n j}}{1-\mathrm{e}^{-\alpha_n j}} - \frac{1}{\alpha_n j}\right) \sim \frac{1}{\alpha_n}\int_0^1\left(\frac{\mathrm{e}^{-s}}{1-\mathrm{e}^{-s}} - \frac{1}{s}\right)\mathrm{d}s = O(\alpha_n^{-1}),$$

noting that the integrand function has a finite limit at zero,

$$\frac{\mathrm{e}^{-s}}{1-\mathrm{e}^{-s}} - \frac{1}{s} = \frac{s\,\mathrm{e}^{-s} - 1 + \mathrm{e}^{-s}}{s\,(1-\mathrm{e}^{-s})} = \frac{-\frac{1}{2}s^2 + O(s^3)}{s^2 + O(s^3)} \to -\frac{1}{2} \qquad (s \to 0).$$

As a result,

$$\mathsf{E}(M) \sim \alpha_n^{-1}(-\log\alpha_n) \sim \frac{\sqrt{n}}{2\kappa}\log n = \frac{\sqrt{6n}}{2\pi}\log n \qquad (n\to\infty). \tag{3.44}$$

Remark 3.6. Two different model settings discussed above — with independent outputs X_i ($i = 1, \dots, M$), like in the item production model (Section 3.1), or with independent multiplicities M_j ($j \in \mathbb{N}_0$), like in a randomised model of integer partitions (Section 3.3), are in fact closely connected and, in a sense, equivalent to one another. Indeed, randomisation of certain parameters in combinatorial structures is a frequently used technical tool (Arratia *et al.*, 2003) aiming to overcome structural constraints, such as a prescribed sum of parts in integer partitions (Bogachev, 2015; Fristedt, 1993). As another example directly related to the item production model, in the occupancy problem (see Remark 3.2) it is conventional to use the so-called *poissonisation* (Arratia *et al.*, 2003; Borisov & Jetpisbaev, 2022) by replacing the original (co-dependent) multiplicities M_j by independent Poisson random variables with mean Mf_j, respectively ($j \in \mathbb{N}_0$) Bogachev *et al.* (2008); Gnedin *et al.* (2007). In each of these settings, the anticipated equivalence is guaranteed via a suitable “bridge” between the original and

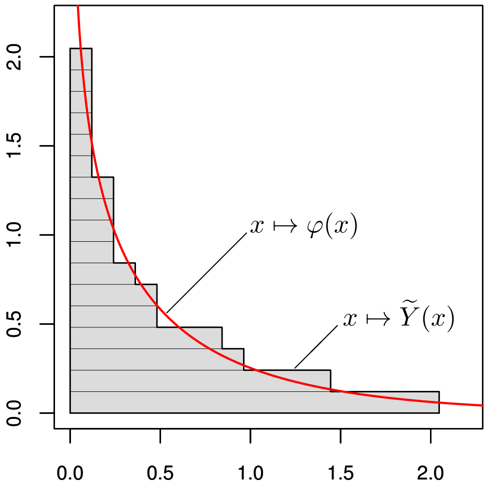


Figure 3.3: A random (simulated) Young diagram (shown as a shaded area) under the scaling $\sqrt{n}$ along both axes, with $n = 100$. Horizontal blocks correspond to the ordered outputs $X_{i,M}$ (with the sample value $M = 17$). The limit shape $\varphi(x)$ defined in (3.42) is shown as a red solid line.

randomised versions of the problem, such as a local limit theorem for the asymptotics of probabilities $\mathsf{P}\big(\sum_j jM_j = n\big)$ in the case of integer partitions (Bogachev, 2015; Fristedt, 1993), or a "depoissonisation lemma" in the occupancy problem (Bogachev *et al.*, 2008; Gnedin *et al.*, 2007).

3.4 Indexes Characterising the Item Production

In this section, we review the citation indexes, discussed in Section 2.4, in the framework of the general item production model of Section 3.1. Although these indexes (such as h-index and g-index) were historically proposed in the context of citations of scientific outputs, we argue that they make sense in the general framework and provide a useful characterisation of the output performance. In addition, a new production metric called the h_1-*index* is proposed by combining the ideas used in defining the h-index and g-index.

3.4.1 The h-index

Suppose that an author has M papers with citations $X_1, \dots, X_M$, respectively. Then the definition of the h-index given in Section 2.4.1 is expressed mathematically as follows,

$$h := \max\left\{j \ge 0\colon \sum_{i=1}^{M} I_{\{X_i \ge j\}} \ge j\right\}. \tag{3.45}$$

Using formula (3.17), the latter definition can be rewritten in either of the forms

$$h = \max\left\{j \ge 0\colon \sum_{\ell \ge j} M_\ell \ge j\right\} = \max\bigl\{j \ge 0\colon Y(j) \ge j\bigr\}. \tag{3.46}$$

Thus, the h-index is interpreted geometrically as (the size of) the largest square under the graph of the Young boundary $Y(x)$ (see Figure 3.4, top row). Such a square is often called the *Durfee square* (Yong, 2014).

It follows from (3.46) that

$$Y(h+1) - 1 < h \le Y(h). \tag{3.47}$$

Furthermore, since h and $Y(h+1)$ are integers, the left inequality in (3.47) means in fact that $Y(h+1) \le h$, so that the double inequality (3.47) can be written in a tighter form,

$$Y(h+1) \le h \le Y(h). \tag{3.48}$$

This two-sided inequality can be verified geometrically by inspection of (four) possible geometric configurations that may arise when considering the crossing of the Young diagram by the diagonal $y = x$.

3.4.2 The g-index

Suppose that an author has M papers with $X_1, \dots, X_M$ citations. Consider the corresponding order statistics, that is, order citation counts $X_{1,M} \ge \dots \ge X_{M,M}$. Then the definition of the g-index given in Section 2.4.2 can be written mathematically as follows,

$$g := \max\left\{k \ge 1\colon \sum_{i=1}^{k} X_{i,M} \ge k^2\right\}. \tag{3.49}$$

Remark 3.7. Note that, by definition (3.49), $g \le M$.

Clearly, the g-index is not smaller than the h-index of the same author (Egghe, 2006, Proposition I.2. p. 133),

$$g \ge h. \tag{3.50}$$

Indeed, if the h-index has value h then there are h papers with at least h citations each, and therefore with at least $h \times h = h^2$ citations in total. Hence, the trial value $k = h$ satisfies the inequality condition in (3.49), which implies that $g \ge k = h.$ as claimed.

3.4.3 First-order h-index

Motivated by the g-index aiming to take into account the actual citations of top-cited papers, we propose a new citation index referred to as the *first-order h-index (h_1-index)*. Suppose that an author has M publications with $X_1, \dots, X_M$ citation, then the h_1-index is defined as the maximum integer h_1 such that the total sum of citation counts of papers with at least h_1 citations each is not less than h_1^2. Mathematically, this is expressed as

$$h_1 := \max \left\{ j \ge 0 \colon \sum_{i=1}^{M} X_i \, I_{\{X_i \ge j\}} \ge j^2 \right\}. \tag{3.51}$$

Using (3.6), the sum in (3.51) is rewritten as follows,

$$\sum_{i=1}^{M} X_i \, I_{\{X_i \ge j\}} = \sum_{i=1}^{M} \sum_{\ell \ge j}^{\infty} \ell \, I_{\{X_i = \ell\}} = \sum_{\ell \ge j} \ell \sum_{i=1}^{M} I_{\{X_i = \ell\}} = \sum_{\ell \ge j} \ell \, M_\ell.$$

Hence, definition (3.51) can be expressed in terms of the multiplicities,

$$h_1 := \max \left\{ j \in [0, M] \colon \sum_{\ell \ge j} \ell \, M_\ell \ge j^2 \right\}. \tag{3.52}$$

The h_1-index combines the features of both the h-index and the g-index. It takes into account citations that are equal to or greater than a minimum threshold value of h_1 as in the h-index, while also including higher citations as in the g-index. This ensures that the h_1-index captures the impact of highly cited papers and provides a more balanced picture of their overall scholarly impact.

The general relationship between the h-index, h_1-index, and g-index is as follows,

$$h \leq h_1 \leq g. \tag{3.53}$$

Indeed, from the defining condition (3.46) we have

$$h \leq \sum_{\ell \geq h} M_\ell \leq \frac{1}{h} \sum_{\ell \geq h} \ell\, M_\ell \leq \frac{h_1^2}{h},$$

according to (3.51), and it follows that $h \leq h_1$.

Turning to the relation between h_1 and g, let the maximising sum in (3.52) be expressed as

$$h_1^2 \leq \sum_{i=1}^{M} X_i\, I_{\{X_i \geq h_1\}} = \sum_{i=1}^{J} X_{i,M}, \tag{3.54}$$

involving J higher order statistics $X_{i,M}$, each one satisfying the inequality

$$X_{i,M} \geq h_1 \qquad (i = 1, \ldots, J). \tag{3.55}$$

Now, if $J \leq h_1$ then from (3.54) we obtain

$$h_1^2 \leq \sum_{i=1}^{J} X_{i,M} \leq \sum_{i=1}^{h_1} X_{i,M},$$

and it follows from the definition (3.49) that $g \geq h_1$. Alternatively, if $J \geq h_1$ then, using (3.55), we can write

$$\sum_{i=1}^{J} X_{i,M} \geq \sum_{i=1}^{h_1} X_{i,M} \geq \sum_{i=1}^{h_1} h_1 = h_1^2,$$

and as before it follows again that $g \geq h_1$.

Remark 3.8. The confinement of the h_1-index to the range $[0, M]$ (see (3.52)) is important, for otherwise the value of h_1 could go above M, thus violating the inequality $g \geq h_1$. For instance, consider the citation output $(18, 18, 1, 1)$, then

$$\sum_{i=1}^{M=4} X_i\, I_{\{X_i \geq 6\}} = 18 + 18 = 36 = 6^2,$$

so that $h_1 = 6$. On the other hand,

$$\sum_{i=1}^{k=4} X_{i,M} = 18 + 18 + 1 + 1 = 38 \geq 4^2,$$

hence $g = 4 < h_1 = 6$.

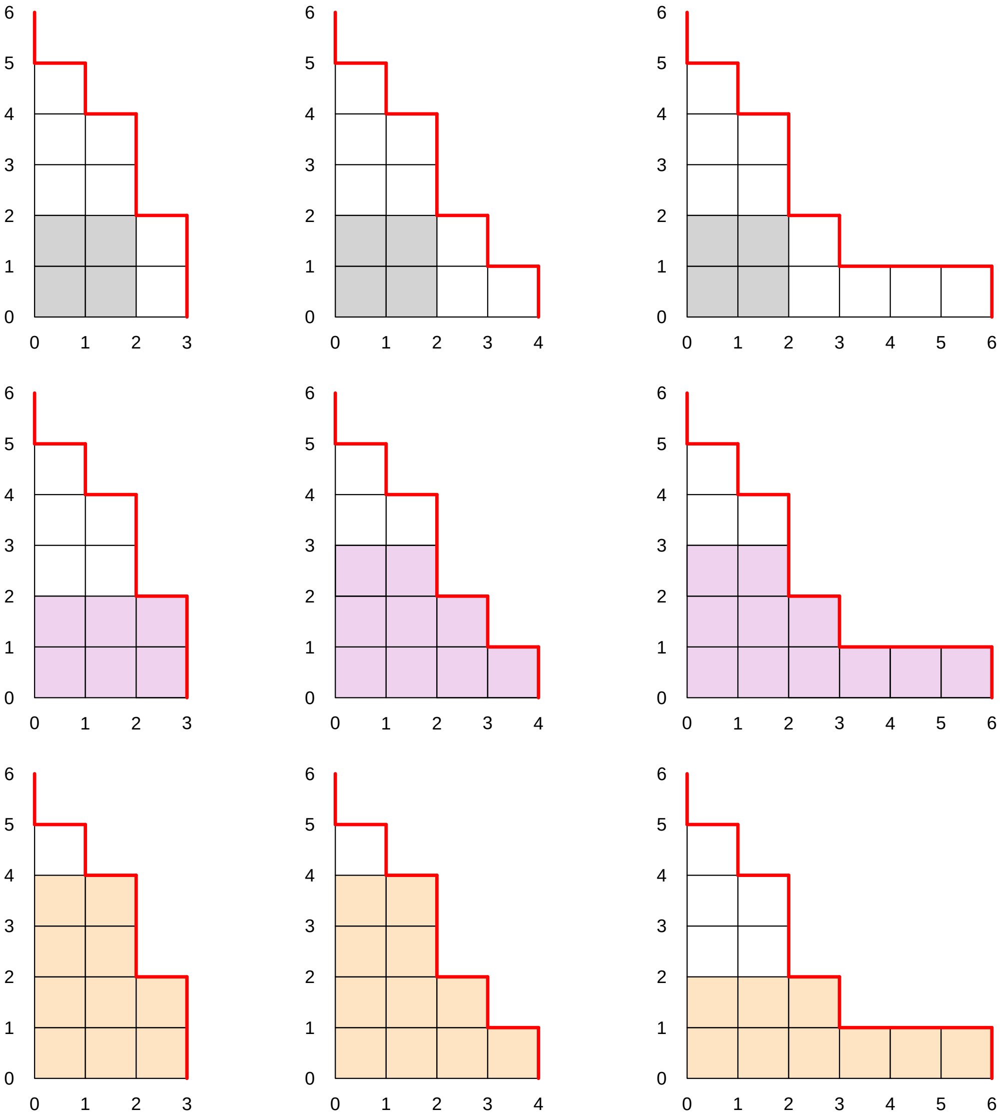


Figure 3.4: Visualisation of the citation indexes h, g, and h_1. The Young diagrams depict ordered citations of three mock authors, $(3, 3, 2, 2, 1, 0)$, $(4, 3, 2, 2, 1, 0)$ and $(6, 3, 2, 2, 1, 0)$, with $M = 6$ papers each. Shaded areas in each row show the citation counts for calculating the indexes: h-index (top row), with values $h = 2$ for each author; g-index (middle row), yielding $g = 2$, $g = 3$, and $g = 3$, respectively; and h_1-index, yielding $h_1 = 2$, $h_1 = 2$, and $h_1 = 3$, respectively.

In Figure 3.4, examples are shown of different citation outputs difference only in the citation counts of their top-cited paper (each with the same number of papers, $M = 6$), which yield the same h-index, but different g-index and h_1-index values. These examples also illustrate the inequalities (3.53).

3.5 Composite Model of Item Production

In some situations, a standard item production model with a single "battery" of sources $i = 1, \dots, M$ (as described in Section 3.1) may not be sufficient, because several such batteries are needed.

As one important example, in the APC relationship (see Section 3.1.1) papers (sources) produce citations (items) for each individual author, so in fact we are dealing with several item production models, indexed by authors (who may be assumed to be independent if we ignore multiple authorship). Another example is the statistical analysis of journal uses (see Chen's data in Section 2.2), where the journal issues are interpreted as sources and their uses as items; however, if we wanted to repeat the same analysis for a different time window, this would require considering another (independent) item production model. As our third example arising from the count statistics in literary texts (e.g., the Moby Dick data set in Section 2.2), the need for multiple item production models would arise if one wanted to consider several books by the same author or different authors.

These considerations lead to the following generalisation of the item production model with a single battery of sources (Section 3.1.2) to a *composite item production model* with multiple batteries of sources. Suppose there are K independent batteries of sources, each one following the common frequency distribution (f_j). The number of sources in each battery is denoted $M^{(k)}$, and their respective independent outputs (also independent of $M^{(k)}$) are $\bigl(X_1^{(k)}, \dots, X_{M^{(k)}}^{(k)}\bigr)$ $(k = 1, \dots, K)$. The battery sizes $M^{(k)}$ are assumed random, independent and identically distributed; we denote by $\eta = \mathsf{E}\bigl(M^{(k)}\bigr)$ their expected value. The total number of sources equals

$$M = \sum_{k=1}^{K} M^{(k)}.$$

Multiplicities of output counts $j \in \mathbb{N}_0$ in each battery are denoted $M_j^{(k)}$ $(k =$

$1,\dots,K$), and the pooled multiplicities are given by

$$M_j = \sum_{k=1}^{K} M_j^{(k)} \qquad (j \in \mathbb{N}_0).$$

Likewise, we can define the individual Young diagrams per battery, with the upper boundaries

$$Y^{(k)}(x) = \sum_{j \ge x} M_j^{(k)} \qquad (x \ge 0,\ \ k = 1,\dots,K), \tag{3.56}$$

and the pooled Young diagram with the boundary

$$Y(x) = \sum_{j \ge x} M_j \qquad (x \ge 0). \tag{3.57}$$

We can also consider the mean Young boundary of the pooled sample,

$$\overline{Y}(x) = \frac{1}{K} \sum_{k=1}^{K} Y^{(k)}(x) = \frac{1}{K} \sum_{j \ge 0} \sum_{k=1}^{K} M_j^{(k)} = \frac{1}{K} \sum_{j \ge 0} M_j \qquad (x \ge 0). \tag{3.58}$$

As an example of this notation, if there are K authors, with $M^{(1)},\dots,M^{(K)}$ papers and citation profiles $Y^{(1)}(x),\dots,Y^{(K)}(x)$, respectively, then $Y(x)$ in (3.57) corresponds to the pooled "mega-author", while $\overline{Y}(x)$ in (3.58) is related to an average author. Clearly, the latter object makes more sense if we wish to use Young diagrams to model the h-index (see Section 3.4.1).

With the aid of the indicators (cf. (3.19))

$$Z_i^{(k)}(x) := I_{\{X_i^{(k)} \ge x\}} \qquad (x \ge 0,\ \ i = 1,\dots,M^{(k)},\ \ \ k = 1,\dots,K),$$

the individual Young boundaries (3.56) can be decomposed as the sums of such indicators (cf. (3.20)) but now with a random number of terms,

$$Y^{(k)}(x) = \sum_{i=1}^{M^{(k)}} Z_i^{(k)}(x) \qquad (k = 1,\dots,K). \tag{3.59}$$

Recalling the well-known *Wald identities* (Shiryaev, 1996, Sec. VII.3, Theorem 3, p. 488), for each $k = 1,\dots,K$ we have

$$\mathsf{E}\big(Y^{(k)}(x)\big) = \mathsf{E}\big(M^{(k)}\big) \cdot \mathsf{E}\big(Z_1^{(k)}(x)\big) = \eta\,\bar{F}(x), \tag{3.60}$$

and furthermore,

$$\begin{aligned}\mathsf{E}\Big\{\big(Y^{(k)}(x)-M^{(k)}\,\mathsf{E}\big(Z_1^{(k)}(x)\big)^2\Big\} &= \mathsf{E}\big(M^{(k)}\big)\cdot\mathsf{Var}\big(Z_1^{(k)}(x)\big)\\ &= \eta\,\bar F(x)\,F(x),\end{aligned}\tag{3.61}$$

where $\bar F(x)=\sum_{j\ge x} f_j$ and $F(x)=1-\bar F(x)$ (see (3.21), (3.22)). Observe, using the first Wald identity (3.60), that

$$\mathsf{E}\big(Y^{(k)}(x)-M^{(k)}\,\mathsf{E}\big(Z_1^{(k)}(x)\big)\big)=\eta\bar F(x)-\eta\bar F(x)=0.$$

Hence, the second Wald identity (3.61) is rewritten as

$$\mathsf{Var}\big(Y^{(k)}(x)-M^{(k)}\,\mathsf{E}\big(Z_1^{(k)}(x)\big)\big)=\eta\,\bar F(x)\,F(x).\tag{3.62}$$

The new setting, with a composite item production model, leads to the question about existence of the limit shape for the sample mean diagram $\overline{Y}(x)$. The following natural result is valid.

Theorem 3.1. *Assume that in a standard (single battery) item production model, the limits (3.31) hold with some function $\varphi(x)$ and scaling coefficients $A=A_M\to\infty$ and $B=B_M\to\infty$ (i.e., chosen according to the battery size $M\to\infty$). Consider a composite item production model with $K\gg 1$ batteries of independent random sizes $M^{(k)}$, respectively, assuming that $\eta=\mathsf{E}\big(M^{(k)}\big)<\infty$. With the scaling coefficients A_η and B_η, set*

$$\widetilde{\overline{Y}}(x):=\frac{1}{B_\eta}\,\overline{Y}(A_\eta x)=\frac{1}{KB_\eta}\sum_{k=1}^{K}Y^{(k)}(A_\eta x)\qquad(x\ge 0).\tag{3.63}$$

Then $x\mapsto\varphi(x)$ is the limit shape of $\widetilde{\overline{Y}}(x)$, that is,

$$\mathsf{E}\Big(\widetilde{\overline{Y}}(x)\Big)\to\varphi(x)\qquad(x>0),\tag{3.64}$$

and moreover, for any $\varepsilon>0$,

$$\mathsf{P}\Big(|\widetilde{\overline{Y}}(x)-\varphi(x)|>\varepsilon\Big)\to 0\qquad(x>0).\tag{3.65}$$

Proof. Using formulas (3.63) and (3.60), we have

$$\mathsf{E}\Big(\widetilde{\overline{Y}}(x)\Big)=\frac{1}{KB_\eta}\sum_{k=1}^{K}\mathsf{E}\big(Y^{(k)}(A_\eta x)\big)=\frac{\eta\bar F(A_\eta x)}{B_\eta}\to\varphi(x),\tag{3.66}$$

according to the first limit in (3.31) (with η in place of M and, accordingly, A_η and B_η instead of generic A and B). This proves the first claim (3.64).

Next, we represent (3.63) as follows,

$$\widetilde{\widetilde{Y}}(x) = \frac{1}{KB_\eta}\sum_{k=1}^{K}\bigl(Y^{(k)}(A_\eta x) - M^{(k)}\bar{F}(A_\eta x)\bigr) + \frac{\eta\,\bar{F}(A_\eta x)}{B_\eta}\cdot\frac{1}{K}\sum_{k=1}^{K}\frac{M^{(k)}}{\eta}. \qquad (3.67)$$

Using mutual independence of the batteries ($k = 1, \dots, K$) and the second Wald identity in the form (3.62), we have

$$\begin{aligned}\mathsf{Var}&\left(\frac{1}{KB_\eta}\sum_{k=1}^{K}\bigl(Y^{(k)}(A_\eta x) - M^{(k)}\bar{F}(A_\eta x)\bigr)\right)\\ &= \frac{1}{K^2B_\eta^2}\sum_{k=1}^{K}\mathsf{Var}\bigl(Y^{(k)}(A_\eta x) - M^{(k)}\bar{F}(A_\eta x)\bigr)\\ &= \frac{\eta\,\bar{F}(A_\eta x)F(A_\eta x)}{KB_\eta^2} \sim \frac{\varphi(x)}{KB_\eta} \to 0,\end{aligned}$$

where we used the limit (3.66). Therefore, the first (normalised) sum on the right-hand side of (3.67) converges to zero in probability.

On the other hand, again using the limit (3.66) and the law of large numbers for the random sequence $(M^{(k)}/\eta)$ (with mean 1), we obtain

$$\frac{\eta\,\bar{F}(A_\eta x)}{B_\eta}\cdot\frac{1}{K}\sum_{k=1}^{K}\frac{M^{(k)}}{\eta} \xrightarrow{\mathrm{p}} \varphi(x).$$

Thus, the second claim (3.65) is also proved. □

3.6 Estimation of the Citation Indexes

3.6.1 Estimation of the h-index

According to (3.28), for each $x > 0$ we have an approximation

$$\widetilde{Y}(x) = B^{-1}Y(Ax) \approx \varphi(x). \qquad (3.68)$$

Applying this to the value h of the h-index, that is, such that $Y(h) \approx h$ (see (3.48)), it follows that

$$Y(h) \approx h \approx B\,\varphi(h/A). \qquad (3.69)$$

This suggests that the h-index may be approximately estimated via solving the equation

$$\varphi(s)=\frac{As}{B}. \tag{3.70}$$

Note that the solution to this equation exists and is unique, because $s\mapsto\varphi(s)$ is a non-increasing continuous function with $\varphi(0)>0$, while the right-hand side of (3.70) vanishes at zero and is continuous and increasing.

More precisely, if $s=s^*$ is the root of equation (3.70), then the h-index is estimated by setting

$$\hat{h}=As^*. \tag{3.71}$$

First, we show that the estimator (3.71) is consistent in the sense that it is asymptotically close (in probability) to the h-index on the scale B, that is,

$$\frac{\hat{h}-h}{B}=\frac{As^*-h}{B}\xrightarrow{\text{p}}0. \tag{3.72}$$

Indeed, setting $s=h/A$, from (3.48) we get

$$\frac{Y\bigl(A(s+1/A)\bigr)}{B}\le\frac{As}{B}\le\frac{Y(As)}{B},$$

that is,

$$\widetilde{Y}(s+1/A)\le\frac{As}{B}\le\widetilde{Y}(s). \tag{3.73}$$

By virtue of the limit shape result (3.27) (more precisely, its uniform version (3.30)), both the left- and right-hand sides of (3.73) are close (in probability) to $\varphi(s)$. Therefore,

$$\varphi(s)-\frac{As}{B}\xrightarrow{\text{p}}0.$$

Since the function $s\mapsto\varphi(s)$ is continuous, in view of equation (3.70) it follows

$$\frac{As^*}{B}-\frac{As}{B}\xrightarrow{\text{p}}0,$$

or, returning to h and $\hat{h}$,

$$\frac{\hat{h}-h}{B}\xrightarrow{\text{p}}0,$$

as claimed in (3.72).

Asymptotic confidence bounds for the h-index can also be constructed. To this end, recall that $\widetilde{Y}(s) = Y(As)/B$ is asymptotically normal with mean $M\bar{F}(As)/B \sim \varphi(s)$ and variance

$$\frac{M\bar{F}(As)F(As)}{B^2} \sim \frac{\varphi(s)}{B}\left(1 - \frac{B\,\varphi(s)}{M}\right)$$

(cf. (3.31)). Hence, with $s = h/A$, we have a distributional approximation

$$\frac{Y(As) - B\,\varphi(s)}{\sqrt{B\,\varphi(s)\,\bigl(1 - B\,\varphi(s)/M\bigr)}} \stackrel{d}{\approx} \mathcal{N}(0,1), \tag{3.74}$$

where $\mathcal{N}(0,1)$ is the standard normal distribution. But $Y(As) = Y(h) \approx h$ and, on the other hand, $B\,\varphi(s) \approx B\,\varphi(s^*) = As^* = \hat{h}$ (see (3.69) and (3.71)). Thus, (3.74) can be rewritten as

$$\frac{h - \hat{h}}{\sqrt{\hat{h}\,\bigl(1 - \hat{h}/M\bigr)}} \stackrel{d}{\approx} \mathcal{N}(0,1). \tag{3.75}$$

Hence, for the confidence level $(1-\alpha)\,100\%$, the confidence bounds for h are given by

$$h^{\pm} = \hat{h} \pm z_{\alpha/2}\sqrt{\hat{h}\,\bigl(1 - \hat{h}/M\bigr)}, \tag{3.76}$$

where $\hat{h} = As^*$ (see (3.71)) and $z_{\alpha/2}$ is the upper $(\alpha/2)$-quantile of the standard normal distribution, that is, the root of the equation

$$\Phi(z) = 1 - \tfrac{1}{2}\alpha.$$

For instance, if $\alpha = 0.05$ then $z_{\alpha/2} \doteq 1.9600$.

Example 3.3. In the integer partition model (see Section 3.3), the scaling coefficients are given by $A = B = \sqrt{N}$, where N is the total number of items (citations). Using the limit shape (3.41), the equation (3.70) is reduced to $\varphi(s) = s$, that is,

$$-\log\bigl(1 - \mathrm{e}^{-\kappa s}\bigr) = -\kappa s \qquad (\kappa = \pi/\sqrt{6}),$$

which solves to

$$s^* = \frac{\log 2}{\kappa} \doteq 0.5404446. \tag{3.77}$$

Thus, the general formula (3.71) specialises as

$$\hat{h} = 0.5404446\,\sqrt{N}. \tag{3.78}$$

As a real data example, we looked at one entry in the list of Fields medalists 1998–2010 in Yong (2014, Table 2, p. 1041). Specifically, we considered Richard Borcherds (1998 award), with $N = 1{,}062$ citations and $M = 32$ papers. Note that the number of papers is not provided in Yong (2014), so this had to be identified separately by using the Google Scholar. This person's actual h-index (by 2014) was $h = 14$. The estimated h-index using formula (3.78) is given by

$$\hat{h} = s^{*}\sqrt{N} \doteq 17.61219.$$

Furthermore, using formula (3.76), an asymptotic 98% confidence interval for the h-index is calculated as $[12.08, 23.11]$, which covers the true value $h = 14$ quite well. For a comparison, the 95% confidence interval calculated by Yong (2014) using combinatorial methods was $[14, 21]$, with the true value $h = 14$ sitting just on its left boundary.

Remark 3.9. In the composite item production model, with multiple batteries of sources (e.g., a group of K authors, with their papers and citations), one should work with the averaged Young diagrams $\overline{Y}(x)$. Hence, the scaling B should be replaced by B/K, with the corresponding modifications for the estimates of citation indexes. This modification is meaningful, noting that the citation indexes are individual-based metrics.

3.6.2 Estimation of the g-index

By definition of the g-index (3.49) and its graphical representation using the Young diagram (see Figure 3.5. left panel), the area of the g top blocks is at least g^2. Recall, using the limit shape result (3.27), that $\widetilde{Y}(x) = B^{-1}Y(Ax) \approx \varphi(x)$, that is,

$$Y(x) \approx B\,\varphi(x/A) =: \widetilde{\varphi}(x) \qquad (x > 0). \tag{3.79}$$

Hence, the aforementioned part of the Young diagram (i.e., below level $y = g$) can be approximated by the corresponding area below the graph of the function $\widetilde{\varphi}(x)$.

In turn, passing over to the inverse function $\widetilde{\varphi}^{-1}(y)$, the last area is expressed by integration, leading to the equation

$$\int_0^g \widetilde{\varphi}^{-1}(y)\,\mathrm{d}y = g^2. \tag{3.80}$$

Solving this equation, we obtain an estimator $\hat{g}$.

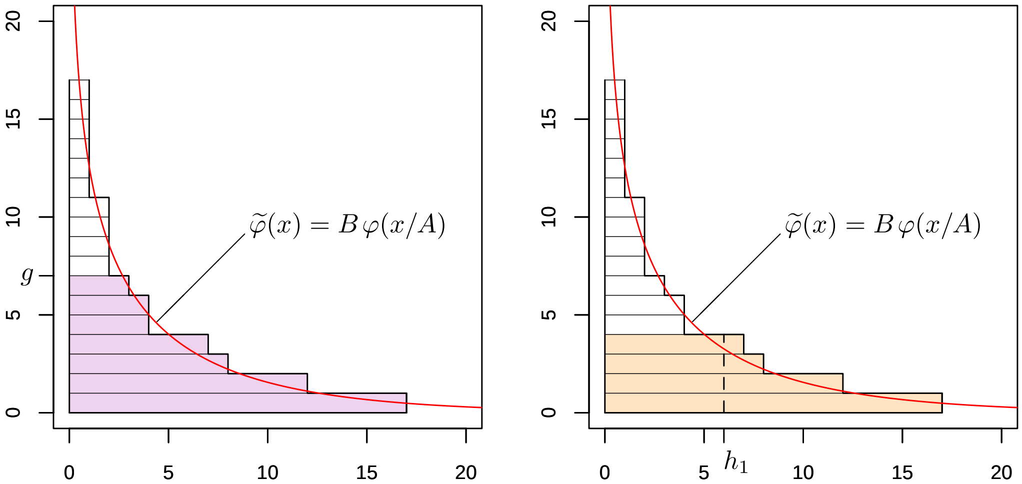


Figure 3.5: Graphical illustration of estimation of the g-index (left panel) and h_1-index (right panel) via the limit shape. The shaded parts of the Young diagram (for a mock citation output $(17, 12, 8, 7, 4, 4, 3, 2, 2, 2, 2, 1, 1, 1, 1, 1, 1)$) represent top citation counts such that the shaded area is at least g^2 or h_1^2, respectively, which are approximated by the areas below the graph of the rescaled limit shape $\widetilde{\varphi}(x)$ defined in (3.79) and the horizontal level $y = g$ (left) or $y = \widetilde{\varphi}(h_1)$ (right). These areas are expressed by formulas (3.80) and (3.86).

Example 3.4. Continuing Example 3.3 based on the integer partition model, we can estimate the g-index using formula (3.80). First, from the limit shape (3.41) we find the inverse of the function $\widetilde{\varphi}(x) = \sqrt{N}\,\varphi\bigl(x/\sqrt{N}\,\bigr)$,

$$\widetilde{\varphi}^{-1}(y) = -\frac{\sqrt{N}}{\kappa}\log\left(1-\exp\left(-\frac{\kappa y}{\sqrt{N}}\right)\right), \tag{3.81}$$

where $\kappa = \pi/\sqrt{6}$ (see (3.36)). Next, the integral in (3.80) can be calculated by means of a series expansion of the logarithm,

$$\log(1-x) = -\sum_{n=1}^{\infty}\frac{x^n}{n} \qquad (0 \le x < 1).$$

Hence, with the substitution $t = \kappa y/\sqrt{N}$,

$$\begin{aligned}\int_0^g \widetilde{\varphi}^{-1}(y)\,\mathrm{d}y &= -\frac{\sqrt{N}}{\kappa}\int_0^g \log\left(1-\exp\left(-\frac{\kappa y}{\sqrt{N}}\right)\right)\mathrm{d}y\\ &= -\frac{N}{\kappa^2}\int_0^{\kappa g/\sqrt{N}} \log\bigl(1-\mathrm{e}^{-t}\bigr)\,\mathrm{d}t\\ &= \frac{N}{\kappa^2}\sum_{n=1}^{\infty}\frac{1}{n}\int_0^{\kappa g/\sqrt{N}} \mathrm{e}^{-nt}\,\mathrm{d}t\\ &= \frac{N}{\kappa^2}\sum_{n=1}^{\infty}\frac{1}{n^2}\bigl(1-\mathrm{e}^{-nt}\bigr)\bigr|_{t=\kappa g/\sqrt{N}}\\ &= \frac{N}{\kappa^2}\bigl(\zeta(2)-\mathrm{Li}_2(\mathrm{e}^{-t})\bigr)\bigr|_{t=\kappa g/\sqrt{N}}.\end{aligned}$$

where $\mathrm{Li}_2(x)$ is the *dilogarithm function* (see Olver *et al.*, 2010, 25.12.1),

$$\mathrm{Li}_2(x) := \sum_{n=1}^{\infty}\frac{x^n}{n^2}.$$

Thus, using that $\zeta(2) = \kappa^2$, equation (3.80) takes the form

$$N - \frac{N}{\kappa^2}\,\mathrm{Li}_2\left(\exp\left(-\frac{\kappa g}{\sqrt{N}}\right)\right) = g^2. \tag{3.82}$$

From this equation, it is clear that its solution g is proportional to $\sqrt{N}$. Setting $g = z\sqrt{N}$, equation (3.82) is rewritten as

$$1 - \kappa^{-2}\,\mathrm{Li}_2\bigl(\mathrm{e}^{-\kappa z}\bigr) = z^2. \tag{3.83}$$

Recalling that $\kappa = \pi/\sqrt{6}$ and solving equation (3.83) numerically gives

$$z \doteq 0.8869923, \tag{3.84}$$

Hence we obtain the estimate

$$\hat{g} = z\sqrt{N} = 0.8869923\sqrt{N}. \tag{3.85}$$

Substituting $N = 1{,}062$, the estimate of the g-index is given by $\hat{g} = z\sqrt{N} \doteq 28.90561$.

Note that there is a striking increase from the h-index, which confirms the emphasis of the g-index on the impact of higher citations. It is also interesting that the estimated value of g is so close to the total number of papers, indicating a good balance of citations for most of the papers.

3.6.3 Estimation of the h_1-index

By definition of the h_1-index (3.49) and its graphical representation using the Young diagram (see Figure 3.5, right panel), the area of the h_1 top blocks, each at least of size h_1, is not smaller than h_1^2. Again using the rescaled limit shape (3.79), the aforementioned part of the Young diagram (i.e., below level $y = Y(h_1)$) can be approximated by the corresponding area below the graph of the function $\widetilde{\varphi}(x)$ and the horizontal level $y = \widetilde{\varphi}(h_1)$, leading to the condition

$$h_1\,\widetilde{\varphi}(h_1) + \int_{h_1}^{\infty} \widetilde{\varphi}(x)\,\mathrm{d}x = h_1^2. \tag{3.86}$$

Solving this equation, we can obtain an estimator $\hat{h}_1$.

By comparing the graphical representations of the areas involved in the estimation of g and h_1 (see Figure 3.5), it becomes evident that the area given on the left-hand side of (3.86) is the same as the area on the left-hand side of (3.80), but with $g = \widetilde{\varphi}(h_1)$. Hence, equation (3.86) can be rewritten in an equivalent form,

$$\int_0^{\widetilde{\varphi}(h_1)} \widetilde{\varphi}^{-1}(y)\,\mathrm{d}y = h_1^2. \tag{3.87}$$

Example 3.5. Adapting the calculations in Example 3.4 and writing $h_1 = w\sqrt{N}$, similarly to equation (3.82) we obtain from (3.87)

$$1 - \kappa^{-2}\,\mathrm{Li}_2\bigl(\mathrm{e}^{-\kappa\varphi(w)}\bigr) = w^2. \tag{3.88}$$

Solving this equation numerically yields

$$w \doteq 0.738981,$$

hence, with $N = 1{,}062$, we get

$$\hat{h}_1 = 0.738981\sqrt{1062} \doteq 24.08217.$$

Again, we see a substantial increase as compared to the h-index of 14, confirming the high impact of this author, although more conservative than the g-index.

Chapter 4

Power Law Model

This chapter explores the power law (PL) model, one of the popular models for scientometric data. It begins by introducing the classic power law model and the truncated power law, followed by the fitting of these models to real data. Towards the end of the chapter, the fitting of the integer partition model is compared to that of the power law model.

4.1 Power Law Distribution

Definition 4.1. A discrete random variable X with values in $\mathbb{N}$ follows a (discrete) power law with parameter $a > 1$ if the probability frequencies $f_j = \mathsf{P}(X = j)$ are given by

$$f_j = \frac{c_a}{j^a} \qquad (j \in \mathbb{N}), \tag{4.1}$$

where c_a is the normalisation constant such that

$$\sum_{j=1}^{\infty} \frac{c_a}{j^a} = 1 \quad \Rightarrow \quad c_a^{-1} = \sum_{j=1}^{\infty} \frac{1}{j^a} = \zeta(a), \tag{4.2}$$

where $\zeta(a)$ is the *Riemann zeta function* (Olver *et al.*, 2010, 25.2.1).

The power law (4.1) is also known as the *Riemann zeta distribution* or the *Zipf distribution* (Johnson *et al.*, 1994, p. 527).

Of course, the condition $a > 1$ is needed in order that the series (4.2) be convergent, so that formula (4.2) defines a proper probability distribution. In many practical examples, this parameter is in the range $2 < a < 3$ (Clauset *et al.*, 2009).

The expected value of the power law (4.1) is finite if $a > 2$, and is given by

$$\mu = \sum_{j=1}^{\infty} j f_j = \frac{1}{\zeta(a)} \sum_{j=1}^{\infty} \frac{1}{j^{a-1}} = \frac{\zeta(a-1)}{\zeta(a)}. \tag{4.3}$$

The complementary cumulative distribution function (CCDF) of the power law (4.1) with normalisation (4.2) is given by

$$\bar{F}(x) = \frac{1}{\zeta(a)} \sum_{j \geq x} \frac{1}{j^a} \qquad (x \geq 0). \tag{4.4}$$

Clearly, $\bar{F}(x) = 1$ for $0 \leq x \leq 1$. For computational convenience, the right-hand side of (4.4) may be expressed using the *Hurwitz zeta function* (Olver *et al.*, 2010, 25.11.1),

$$\zeta(a, x) = \sum_{\ell=0}^{\infty} \frac{1}{(\ell + x)^a} \qquad (x > 0). \tag{4.5}$$

Indeed, noting that, for $j \in \mathbb{N}$ and $x > 0$, the inequality $j \geq x$ is equivalent to $j \geq \lceil x \rceil$, we have

$$\sum_{j \geq x} \frac{1}{j^a} = \sum_{j \geq \lceil x \rceil} \frac{1}{j^a} = \sum_{\ell=0}^{\infty} \frac{1}{(\ell + \lceil x \rceil)^a} = \zeta(a, \lceil x \rceil). \tag{4.6}$$

Hence, for all $x > 0$ formula (4.4) takes the form

$$\bar{F}(x) = \frac{\zeta(a, \lceil x \rceil)}{\zeta(a)} \qquad (x > 0). \tag{4.7}$$

The Hurwitz zeta function $\zeta(a, x)$ can be numerically calculated in R using command `hzeta` in the `gsl` package (see more detail in Galassi *et al.* (2002)).

By taking the logarithm of (4.1) with normalisation (4.2), we get

$$\log f_j = -\log \zeta(a) - a \log j \qquad (j \in \mathbb{N}). \tag{4.8}$$

Thus, the frequency plot of the power law (4.1) in logarithmically transformed coordinates $u = \log x$, $v = \log y$ (referred to as *log-log coordinates*) lies on a straight line with equation $v = b - au$, with slope $(-a)$ and intercept $b = -\log \zeta(a)$ (for illustration, see the left panel of Figure 4.1).

Likewise, for the CCDF $\bar{F}(x)=\sum_{j\ge x} f_\ell$ we have asymptotically as $x\to\infty$,

$$\begin{aligned}\bar{F}(x)=\frac{1}{\zeta(a)}\sum_{j\ge x}^{\infty} j^{-a}&=\frac{x}{\zeta(a)}\sum_{j/x\ge 1}^{\infty}\left(\frac{j}{x}\right)^{-a}\frac{1}{x}\\ &\sim\frac{x}{\zeta(a)}\int_1^{\infty}s^{-a}\,\mathrm{d}s=\frac{x^{1-a}}{(a-1)\,\zeta(a)},\end{aligned}\tag{4.9}$$

by replacing the integral Riemann sum with the corresponding integral. Rewriting the asymptotic formula (4.9) in the log-log coordinates $u=\log x$, $v=\log y$ as before, we obtain the tail approximation

$$\log\bar{F}(x)\approx-\log\zeta(a)-\log{(a-1)}-(a-1)\log x,\tag{4.10}$$

which means that the tail of the power law plotted on the log-log coordinates is close to a straight line, $v=b-(a-1)\,u$, with slope $-(a-1)$ and intercept $b=-\log\zeta(a)-\log{(a-1)}$ (see the right panel of Figure 4.1).

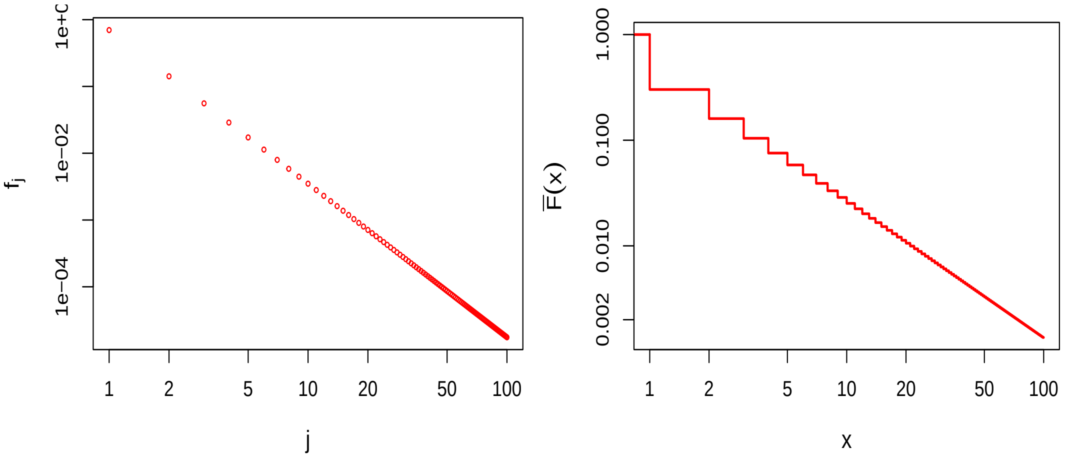


Figure 4.1: Graphical illustration of the power law distribution (with $a=2.3$). The left panel shows the frequencies (4.1) in log-log coordinates, which follow a straight line with slope $-a=-2.3$ according to (4.8). The right panel shows the CCDF plot (4.4) (also in log-log coordinates); its tail approximately follows a straight line with slope $-(a-1)=-1.3$ (see (4.10)).

According to the general formula (3.14), the likelihood of the power law model is given by

$$\mathcal{L}(a;\boldsymbol{X})=\prod_{j=1}^{\infty}\left(\frac{c_a}{j^a}\right)^{M_j}=\left(\frac{1}{\zeta(a)}\right)^{M}\prod_{j=1}^{\infty}j^{-aM_j},\tag{4.11}$$

with the log-likelihood

$$\ell(a;\boldsymbol{X}) = -M\log\zeta(a) - a\sum_{j=1}^{\infty} M_j \log j. \tag{4.12}$$

The MLE for the parameter a is obtained by taking the derivative and setting it to zero,

$$\ell'(a;\boldsymbol{X}) = -M\,\frac{\zeta'(a)}{\zeta(a)} - \sum_{j=1}^{\infty} M_j \log j = 0, \tag{4.13}$$

thus leading to the equation

$$-\frac{\zeta'(a)}{\zeta(a)} = \sum_{j=1}^{\infty} \frac{M_j}{M} \log j. \tag{4.14}$$

Noting that the derivative of the Riemann zeta function is given by

$$\zeta'(a) = -\sum_{j=1}^{\infty} \frac{\log j}{j^a},$$

and the empirical frequencies $\hat{f}_j = M_j/M$ converge to the expected frequencies $f_j = c_a j^{-a}$ (by virtue of the law of large numbers), it readily follows from the equation (4.14) that the MLE $\hat{a}$ is consistent, that is, $\hat{a}$ converges to a as $M \to \infty$.

Using an alternative representation of the likelihood via the sources outputs (X_i) (see (3.10)), the likelihood is expressed as

$$\mathcal{L}(a;\boldsymbol{X}) = \prod_{i=1}^{M} \frac{c_a}{X_i^a} = \left(\frac{1}{\zeta(a)}\right)^{M} \prod_{i=1}^{M} X_i^{-a}, \tag{4.15}$$

with the log-likelihood

$$\ell(a;\boldsymbol{X}) = -M\log\zeta(a) - a\sum_{i=1}^{M} \log X_i. \tag{4.16}$$

Hence, equation (4.13) takes the form

$$\ell'(a;\boldsymbol{X}) = -M\,\frac{\zeta'(a)}{\zeta(a)} - \sum_{i=1}^{M} \log X_i = 0, \tag{4.17}$$

that is (Johnson *et al.*, 2005; Nicholls, 1987),

$$-\frac{\zeta'(a)}{\zeta(a)} = \frac{1}{M}\sum_{j=1}^{\infty} \log X_i. \tag{4.18}$$

4.2 Truncated Power Law

In many practical cases, the power law appears to be unsuitable for smaller counts j, so a pragmatic solution is to try and fit the power law only for larger values, $j \ge j_*$, by truncating the observed frequencies with $j < j_*$ (Clauset *et al.*, 2009). A suitable value $j_* \ge 1$ is called a *truncation threshold*.

More precisely, the modified model assumes that the conditional frequencies

$$f_j^\wedge = \mathsf{P}(X_i = j \,|\, X_i \ge j_*) = \frac{\mathsf{P}(X_i = j)}{\mathsf{P}(X_i \ge j_*)} = \frac{f_j}{\bar{F}(j_*)}$$

are given, for all $j \ge j_*$, by the *truncated power law*,

$$f_j^\wedge = \frac{c_{a,j_*}}{j^a} \qquad (j \ge j_*), \tag{4.19}$$

where c_{a,j_*} is an adjusted normalising constant determined by the formula

$$c_{a,j_*}^{-1} = \sum_{j=j_*}^{\infty} \frac{1}{j^a} = \sum_{\ell=0}^{\infty} \frac{1}{(\ell + j_*)^a} = \zeta(a, j_*) \tag{4.20}$$

(cf. (4.5)). Of course, when $j_* = 1$ these probability frequencies are reduced to the non-truncated power law (4.1).

The (conditional) CCDF of the truncated power law (4.19) is given by

$$\bar{F}^\wedge(x) = \sum_{j \ge x} f_j^\wedge = \frac{1}{\zeta(a, j_*)} \sum_{j \ge x} \frac{1}{j^a} = \frac{\zeta(a, \lceil x \rceil)}{\zeta(a, j_*)} \qquad (x \ge j_*). \tag{4.21}$$

In particular, $\bar{F}^\wedge(j_*) = 1$.

The unconditional probabilities $f_j = \mathsf{P}(X_i = j)$ in the range $j \ge j_*$ are then expressed as follows,

$$\begin{aligned} f_j &= \mathsf{P}(X_i \ge j_*) \cdot \mathsf{P}(X_i = j \,|\, X_i \ge j_*) \\ &= \bar{F}(j_*)\, f_j^\wedge = \frac{\bar{F}(j_*)}{\zeta(a, j_*)\, j^a} \qquad (j \ge j_*). \end{aligned} \tag{4.22}$$

Note that no modelling assumptions are made about the expected frequencies below j_* (unless another model is deployed for that purpose). Therefore, as mentioned in Section 3.1.4, the probability $\rho = \mathsf{P}(X_i \ge j_*) = \bar{F}(j_*)$ must be treated as a parameter. Hence, one works with a likelihood of censored data (see (3.14))

$$\mathcal{L}(a,\rho;\boldsymbol{X}) = (1-\rho)^{M-M(j_*)}\,\rho^{M(j_*)}\prod_{j=j_*}^{\infty}\left(\frac{c_{a,j_*}}{j^a}\right)^{M_j}, \tag{4.23}$$

where M is the number of sources and $M(j^*) = \sum_{j=j_*}^{\infty} M_j$ is the number of sources with at least j_* items each. Accordingly, the log-likelihood is given by

$$\begin{aligned}\ell(a,\rho;\boldsymbol{X}) = \bigl(M - M(j_*)\bigr)\log(1-\rho) + M(j_*)\log\rho \\ + \log c_{a,j_*}\sum_{j=j*}^{\infty} M_j - a\sum_{j=j_*}^{\infty} M_j \log j.\end{aligned} \tag{4.24}$$

In the alternative representation of the likelihood of censored data using the outputs (X_i), formula (4.23) takes the form

$$\mathcal{L}(a,\rho;\boldsymbol{X}) = (1-\rho)^{M-M(j_*)}\,\rho^{M(j_*)}\,(c_{a,j_*})^{M(j_*)}\prod_{i=1}^{M} X_i^{-a} I_{\{X_i \geq j_*\}}. \tag{4.25}$$

4.3 Fitting the Power Law

4.3.1 Graphical methods

A natural heuristic tool to fit a power law model is by looking at the frequency plots with logarithmic scales on both axes, whereby one seeks a straight-line fit, with the slope corresponding to $(-a)$, according to (4.8) (Nicholls, 1987).

Another, more accurate approach (Clauset *et al.*, 2009), which provides the helpful smoothing of the discrete data, is via the CCDF given in (4.4). Using again the log-log plots, a good fit corresponds to a straight line, with slope $1-a$ (see (4.10)). Hence, by fitting a line to the log-transformed complementary cumulative frequencies of data and determining the slope, the parameter a of the fitted power law can be estimated.

4.3.2 Estimation of parameter a

Standard techniques for parameter estimation, such as *ordinary least squares (OLS)* or *maximum likelihood estimation (MLE)*, can be effectively employed to estimate the parameter(s) of the power law.

For the standard (untruncated) power law, the likelihood is given by either of the formulas (4.11) and (4.15) (see also (4.12) and (4.16)), which can be maximised either directly (e.g., using the R command `optim`) or by solving numerically the corresponding likelihood equation, (4.14) or (4.18).

For the truncated power law, assuming that the truncation threshold j_* is known, the parameter a can be estimated by maximising the likelihood of censored data (4.23) after substituting the MLE $\hat{\rho} = M(j_*)/M$ for the parameter $\rho = \bar{F}(j_*)$ (see (3.16)).

A useful approximation for the MLE $\hat{a}$ was given by Clauset *et al.* (2009, formula (3.7), p. 667),

$$\hat{a} \approx 1 + M \left(\sum_{i=1}^{M} \log \frac{X_i}{j_* - \frac{1}{2}} \right)^{-1}. \tag{4.26}$$

In particular, in the untruncated case (with $j_* = 1$) formula (4.26) is reduced to

$$\hat{a} \approx 1 + \frac{1}{\log 2 + M^{-1} \sum_{i=1}^{M} \log X_i}. \tag{4.27}$$

This approximation is based on a comparison with the continuous power law distribution, that is, with density $f(x; j_*) = (a-1)\, j_*^{a-1} x^{-a}$ $(x \geq j_*)$, for which the likelihood is given by (cf. (4.18))

$$\mathcal{L}(a; \boldsymbol{X}) = (a-1)^M j_*^{M(a-1)} \prod_{i=1}^{M} X_i^{-a},$$

leading to the likelihood equation

$$\frac{M}{a-1} - \sum_{i=1}^{M} \log \frac{X_i}{j_*} = 0,$$

which easily solves to the continuous power law MLE

$$\hat{a} = 1 + \frac{M}{\sum_{i=1}^{M} \log(X_i/j_*)}. \tag{4.28}$$

As explained by Clauset *et al.* (2009, Appendix B.4), this result can be satisfactorily adapted to the discrete power law by replacing j_* in (4.28) with $j_* - \frac{1}{2}$ by way of correction for discreteness, resulting in formula (4.26).

4.3.3 Estimation of the truncation threshold

According to Clauset *et al.* (2009, §3.3), joint estimation of parameter a and the truncation threshold j_* may be based on minimising the Kolmogorov–Smirnov (KS) distance between the empirical and theoretical distribution functions. In the context of the truncated power law model, the KS distance is adapted as follows,

$$D \equiv D(a, j_*) = \sup_{x \ge j_*} \left| \bar{F}_{\mathrm{obs}}(x) - \bar{F}(x) \right|, \tag{4.29}$$

where $\bar{F}_{\mathrm{obs}}(x) := \sum_{j \ge x}^{\infty} \hat{f}_j = \sum_{j \ge x}^{\infty} M_j/M$ is the empirical CCDF and $\bar{F}(x)$ is the CCDF of the fitted power law model (4.21). The joint estimate $(\hat{\jmath}_*, \hat{a})$ is the minimiser of D.

In the general case (including where j_* is unknown in advance), the R library `poweRlaw` is helpful for getting the MLE estimation of parameter a together with the optimal choice of j_* (Gillespie, 2015).

4.3.4 Estimation of the citation indexes

Egghe & Rousseau (2006) gave the formulas for estimating the h-index and the g-index under the power law model,

$$\hat{h} = M^{1/a}, \qquad \hat{g} = \left(\frac{a-1}{a-2} \right)^{1-1/a} M^{1/a}. \tag{4.30}$$

Of course, the value of parameter a should be replaced here with a suitable estimate $\hat{a}$.

In the case of a truncated power law, these formulas remain the same as long as the estimated values (4.30) prove to be above the truncation threshold j_* (see a data-based discussion in Sections 4.4.2 and 4.4.3).

4.4 Real Data Examples

The following examples illustrate the fitting results of the power law model to some real data sets introduced in Section 2.2. When using a truncated power law, we extrapolate the fitted plots below the truncation threshold to check an extended performance of the fitted model and to illustrate suitability of the estimated threshold $\hat{j}_*$.

4.4.1 Lotka's data

Lotka's data set (see Section 2.2, A) has the total number of authors $M = 6{,}891$ and the total number of papers $N = 22{,}939$. The estimated parameter $\hat{a} \doteq 1.97$ (with $j_* = 1$) is obtained using the `poweRlaw` package in R. Since $j_* = 1$, the usual (non-truncated) power law model was fitted. The outcomes of the power law fitting are graphically depicted in Figure 4.2. Note that Lotka (1926) presented the original frequencies (i.e., without the log-log transformation), and displayed only a partial range of values, $1 \leq j \leq 30$. In this representation, the power law appears to perform exceptionally well. However, using the log-log coordinates in the full range of the data (especially in , upon observing the frequencies and the complementary cumulative frequencies in log-log transformed coordinates, it is evident that the power law only fits to the initial range of the data.

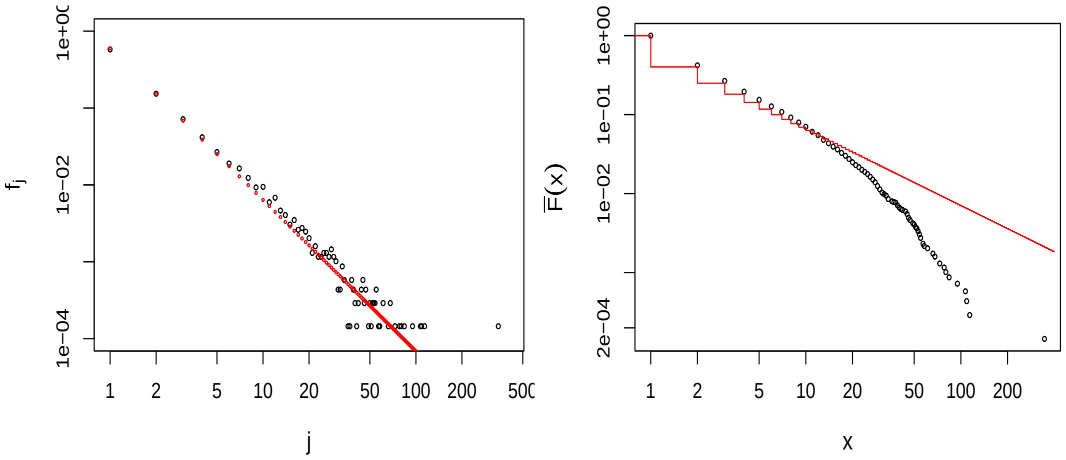


Figure 4.2: The left panel of the figure displays the empirical frequencies along with the corresponding fitted power law frequencies in log-log coordinates. In the right panel, the complementary cumulative frequencies of the number of papers of authors are presented alongside the CCDF of the power law model, both represented in log-log coordinates.

4.4.2 Moby Dick data

As mentioned in Section 2.2, the Moby Dick data is the classic count data for text analysis in informetrics. The corresponding complementary cumulative plot of this dataset (see Figure 2.4 in the lower right) exhibits a linear trend. Hence fitting the power law model to this data set is a natural process. The Moby Dick data set (see Section 2.2, C) has the total number of unique words $M = 18{,}855$ and the total number of occurrences $N = 245{,}567$. The estimated power law parameters were found to be $\hat{a} \doteq 1.95$ and $\hat{j}_* = 7$ using the `poweRlaw` package in R. The results of the power law fitting are graphically displayed in Figure 4.3. In particular, according to formula (4.22) the fitted CCDF in the range $x \geq \hat{j}_* = 7$ is given by

$$\bar{F}_{\mathrm{obs}}(j_*) \cdot \bar{F}^{\wedge}(x) = \bar{F}_{\mathrm{obs}}(j_*)\,\frac{\zeta(\hat{a}, x)}{\zeta(\hat{a}, j_*)}. \tag{4.31}$$

An interesting question is whether the fitted model (4.31) can be used to estimate the h-index. For this, we need to be able to solve the equation

$$M\,\bar{F}_{\mathrm{obs}}(j_*)\,\frac{\zeta(\hat{a}, h)}{\zeta(\hat{a}, j_*)} = h. \tag{4.32}$$

Taking $h = 7$ as a trial value and noting that $\bar{F}_{\mathrm{obs}}(7) \doteq 0.1568815$, the left-hand side of (4.32) is evaluated as

$$M\,\bar{F}_{\mathrm{obs}}(7) \doteq 2958.001 \gg 7.$$

It follows that the solution to equation (4.32) exists in the range $j \geq 7$, so the fitted model can be used for that purpose. In fact, using the estimation formula (4.30), we obtain an estimated value

$$\hat{h} = M^{1/\hat{a}} \doteq 154.6904. \tag{4.33}$$

A comparison with the true value $h = 159$ shows a remarkably accurate estimation. This is not surprising since the power law fit (even with truncation) is extremely good for the Moby Dick data set. The h-index in word frequency analysis can be a measure for evaluating the vocabulary size and the repetition of words in literature. When two books have different h-index but the same word counts in total, the following conjecture can be made: the book with a higher

h-index has a smaller vocabulary size and higher repetition of words, so this book can be recommended to the beginner of a language to read; conversely, the book with a lower h-index may have a larger vocabulary size and lower repetition in using words, so reading this book requires people with higher vocabulary levels.

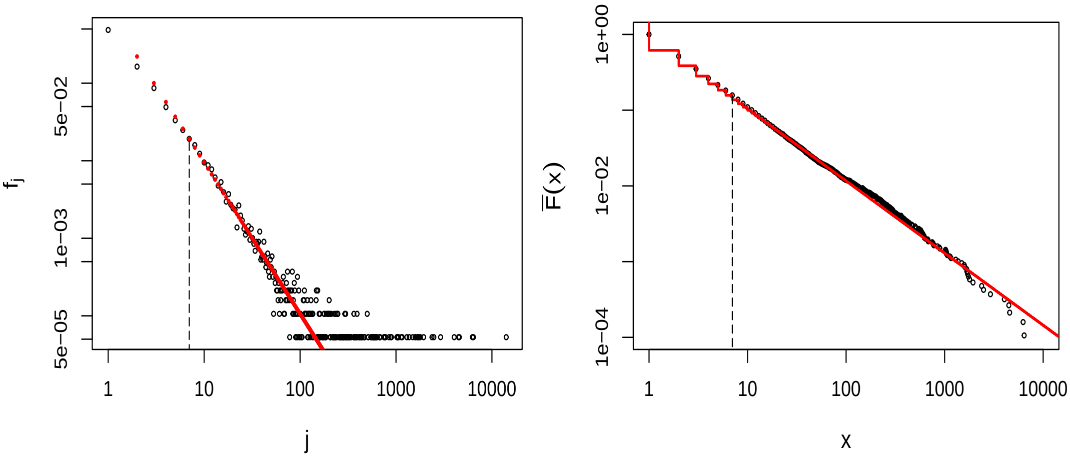


Figure 4.3: Power law (shown in red) fitted to the Moby Dick data (shown in black), with the estimated parameters $\hat{a} \doteq 1.952728$ and $\hat{\jmath}_* = 7$. The left panel displays the empirical frequencies and corresponding fitted power law frequencies, both in log-log coordinates. The right panel depicts the complementary cumulative frequencies of the words in Moby Dick along with the CCDF of the power law model, both in log-log coordinates. The dashed vertical lines correspond to the estimated truncation threshold $\hat{\jmath}_* = 7$.

4.4.3 EJP data

The EJP data set (see Section 2.2, D) comprises all papers of a "mega-author" when citing their work. Where the "mega-author" is that all papers of the group of authors are pooled together, and these papers are counted for one "mega-author". In this way the citation indexes can be estimated using the method introduced in Section 3.5. With a total of $M = 15{,}400$ papers and $N = 245{,}567$ citations, the model fitting result of the power law with a truncation to the EJP data is displayed in Figure 4.4. Using the R package `poweRlaw`, the truncated power law parameter is estimated to be $\hat{a} \doteq 2.32$, with a truncation threshold

$\hat{\jmath}_* = 48$. The dashed lines in both plots indicate that the power law is only fitted to the range $j \geq 48$, although we do extrapolate the fitted plots below the threshold to illustrate their performance there, confirming that truncation is needed for fitting the power law to the EJP data.

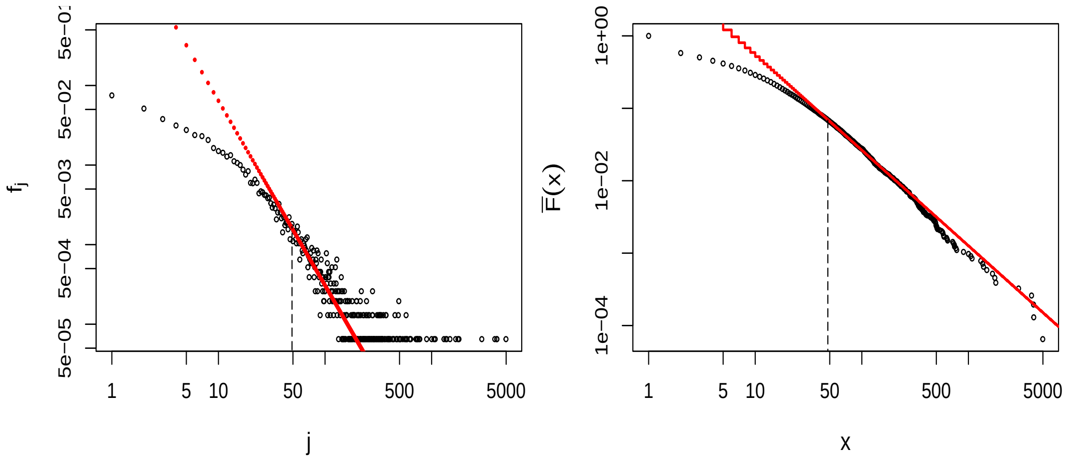


Figure 4.4: Truncated power law model (4.19) (in red) fitted to the EJP data set (in black). The left panel shows the logarithmic transformed frequencies of the EJP data and the frequencies of the truncated power law. The right panel depicts the complementary cumulative frequencies of the EJP data and the CCDF of the truncated power (4.21). The dashed vertical lines correspond to the estimated truncation threshold $j_* = 48$.

Like in Section 4.4.2, we can look at whether the fitted model admits an estimate of the h-index. Because the EJP data set requires a composite item production model, with $K = 113$ "batteries" of sources (papers), equation (3.5) should be modified to

$$\frac{M}{K}\,\bar{F}_{\text{obs}}(j_*) \cdot \frac{\zeta(\hat{a}, h)}{\zeta(\hat{a}, j_*)} = h. \tag{4.34}$$

Again taking $h = \hat{\jmath}_* = 48$ as a trial value, we compute

$$\frac{M}{K}\,\bar{F}_{\text{obs}}(48) = \frac{15400}{113} \times 0.06746753 \doteq 9.19 < 48.$$

This computation indicates that there is no solution to equation (4.34) in the range $h \geq 48$, and therefore the h-index cannot be estimated using the fitted truncated power law. For a comparison, note that the average (sample mean)

h-index (per author) of the EJP data set is $\bar{h} = 17.52$, which is much smaller than 48.

4.5 Fitting the Integer Partition Model

As demonstrated in Section 4.4, the truncated power law only fits to the tail of the EJP data but does not cover the smaller counts. This leads to the idea to check the goodness-of-fit of the integer partition model discussed in Section 3.3. As shown below, the latter model has an exponential tail decay of frequencies (i.e., faster than any power law), which renders it unsuitable for fitting the tail-end of the data; however, it may be usable for approximating the initial portion of the data.

Note that the integer partition model has no intrinsic parameters, it is calibrated through an external parameter n, which has the meaning of the expected value of the total number N of items (citations). On the other hand, the number of sources M in the partition model is not fixed, in contrast to the item production model. For the sake of comparison with other item production models such as power law, we propose to use the sample value N as a substitute for the expectation $n = \mathsf{E}(N)$ and, on the contrary, the expected value $\mathsf{E}(M)$ in place of M, or rather the asymptotic version (3.44) modified as

$$\mathsf{E}(M) \sim \frac{\sqrt{N}\log N}{2\kappa}, \tag{4.35}$$

where $\kappa = \pi/\sqrt{6}$ (see (3.36)).

4.5.1 Fitting to the EJP data

As explained above, using (4.35) the frequencies (f_j) in the integer partition model can be written in the form

$$f_j \approx \frac{\mathsf{E}(M_j)}{\mathsf{E}(M)} \sim \frac{2\kappa}{\sqrt{N}\log N} \cdot \frac{\mathrm{e}^{-\kappa j/\sqrt{N}}}{1 - \mathrm{e}^{-\kappa j/\sqrt{N}}} \qquad (j \in \mathbb{N}). \tag{4.36}$$

For $j \ll \sqrt{N}$, the model (4.36) has a power-law decay,

$$f_j \sim \frac{2}{j \log N},$$

whereas for $j \gg \sqrt{N}$ it has an exponential decay,

$$f_j \sim \frac{2\kappa\, \mathrm{e}^{-\kappa j/\sqrt{N}}}{\sqrt{N}\log N}.$$

The corresponding CCDF can be approximated using the limit shape,

$$\bar{F}(x) \approx \frac{\sqrt{N}\,\varphi(x/\sqrt{N})}{\mathsf{E}(M)} \sim -\frac{2}{\log N}\log\Big(1 - \mathrm{e}^{-\kappa x/\sqrt{N}}\Big). \tag{4.37}$$

Figure 4.5 illustrates the results of fitting the integer partition model to the EJP data. The left panel displays the frequencies of the data and the integer partition model (4.36). The right panel shows the empirical CCDF of the data and the CCDF of the fitted model (4.37). Both plots are displayed in logarithmically transformed coordinates. These plots demonstrate that the integer partition model is only suitable for fitting the initial portion of the EJP data, while it does not perform well for the tail of the data.

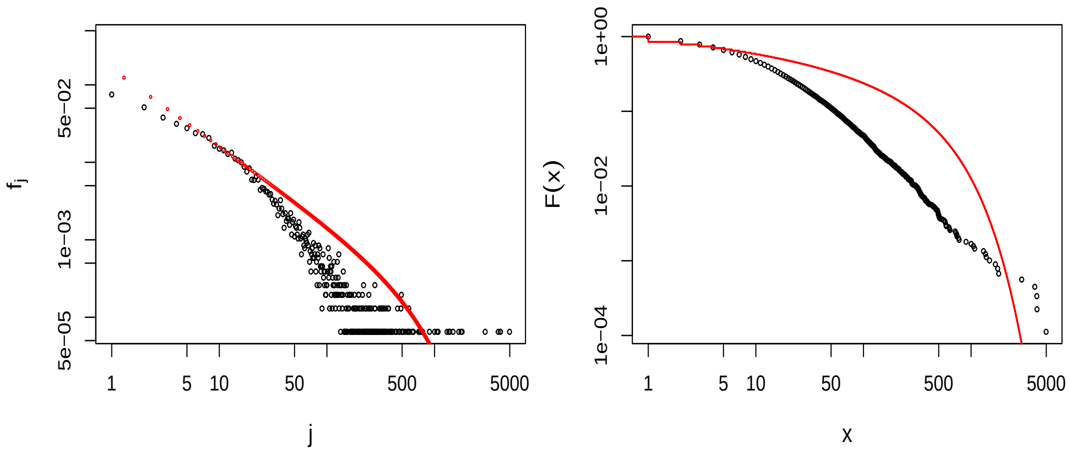


Figure 4.5: The EJP data is displayed in black, and the integer partition model is displayed in red. The left panel depicts the frequencies of the EJP data and the integer partitions, while the right panel shows the complementary cumulative frequencies of the data and the integer partition model.

4.5.2 Estimating the h and g indexes

Using the limit shape of integer partitions under the uniform measure (as $N \to \infty$), one can characterise typical asymptotic properties of citation diagrams, such as the h-index and g-index.

The average h-index and the g-index of these 113 authors from the EJP data are $\bar{h} \doteq 17.52212$ and $\bar{g} \doteq 32.28319$, respectively. The integer partition model is fitted to a "mega-author" of the EJP data, hence an averaging is needed for estimating the average h-index and average g-index. From Remark 3.9 and the estimate (3.78), the estimated average h-index is given by

$$\widehat{\bar{h}} = 0.5404446\sqrt{N/K}. \tag{4.38}$$

substituting the total number of citations $N = 245{,}567$ and the number of authors in the EJP data set $K = 113$, we obtain the estimation of the average h-index through the limit shape of the integer partition model is $\widehat{\bar{h}} \doteq 25.19399$. Similarly, again using Remark 3.9 and the estimate (3.85), the average g-index estimate is given by

$$\widehat{\bar{g}} = 0.8869923\sqrt{N/K} \doteq 41.34906.$$

The estimated average values of the h-index and g-index are relatively larger than the actual values. This discrepancy can be attributed to the integer partition model not being a perfect fit for the EJP data, as illustrated in Figure 4.5.

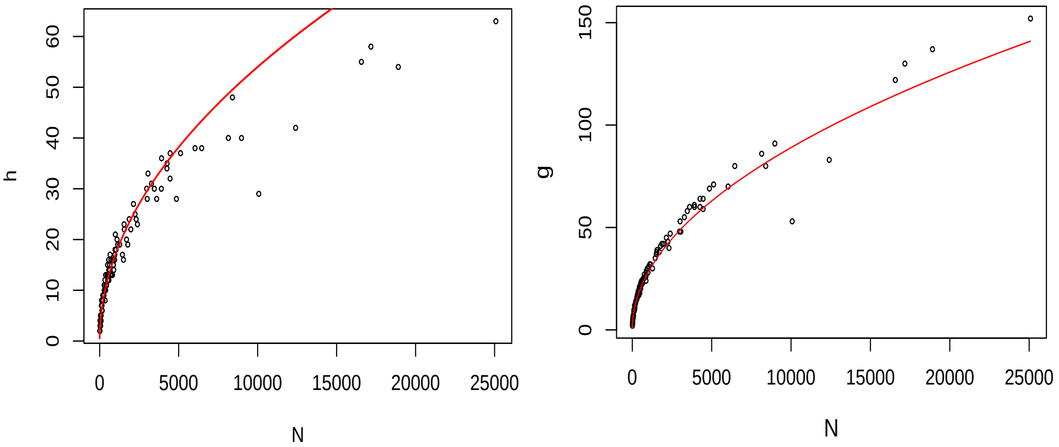


Figure 4.6: Scatter plots of the empirical and estimated h-index (left) and g-index (right) using the limit shape of integer partitions. The black circles in these plots represent the real h-index and g-index values with their corresponding number of citations N for the 113 authors included in the EJP data. The red curves represent the corresponding estimations obtained using equations (3.78) and (3.85) for the h-index and the g-index, respectively.

By focusing on individual authors rather than a "mega-author", Figure 4.6 provides a visual representation of the empirical and estimated h-index (left) and g-index (right) regarding the total number of citations N. The black circles on the plots represent the actual h-index and g-index of 113 authors from the EJP dataset. These values are calculated for each author using the definitions of the h-index and g-index as described in equations (3.45) and (3.49), respectively. The red curves depict the estimated h-index and g-index, which are generated using equations (3.78) and (3.85), respectively.

The left plot indicates that the estimation of the h-index performs well for total citation counts up until approximately 5,000 citations. However, it tends to overestimate for extremely high citation counts. This limitation of the h-index is attributed to its censoring of highly cited papers beyond the h-index. While the limit shape of integer partitions is symmetric, the citation diagram is not always symmetric for authors with exceptionally highly cited papers.

The EJP data have been fitted with two models: the truncated power law and integer partition models. The truncated power law model is effective in capturing the tail of the data for values of $j \geq 48$, as shown in Figure 4.4. On the other hand, the integer partition model is better suited for the beginning of the data with values of $j \leq 7$, as presented in Figure 4.5. Despite their respective strengths, both models are unable to capture the entire range of the data. Notably, the gap between the partition and power law models makes it impossible to accurately capture the h-index. Recall that the sample mean h-index of authors in the EJP data is given by $\bar{h} \doteq 17.52$. which is deep below the truncation threshold $j_* = 48$.

Chapter 5

Generalised Inverse Gaussian-Poisson Model

This chapter explores the *generalised inverse Gaussian-Poisson (GIGP)* model introduced by Sichel (1985). Using main strategies developed in Chapter 3, the limit shape of the GIGP model is specified. The fluctuations of this limit shape are asymptotically normal are also shown in this chapter. More precisely, for convergence to the limit shape to be valid, the number of sources should be growing fast enough. In the opposite regime referred to as "chaotic", the empirical random process is approximated by means of an inhomogeneous Poisson process. These results are illustrated using both computer simulations and some classic data sets in scientometrics dealing with citations of research papers.

This chapter has been documented as a preprint, see Bogachev *et al.* (2023).

5.1 The GIGP Distribution

5.1.1 The GIGP Frequencies

Definition 5.1. The *generalised inverse Gaussian-Poisson (GIGP)* distribution introduced by Sichel (1971, 1985) is of the form

$$f_j = \frac{(1-\theta)^{\nu/2}}{K_\nu\bigl(\alpha\,(1-\theta)^{1/2}\bigr)} \cdot \frac{\bigl(\tfrac{1}{2}\alpha\,\theta\bigr)^j}{j!}\,K_{\nu+j}(\alpha) \qquad (j \in \mathbb{N}_0), \tag{5.1}$$

where parameters have the range $\nu \in \mathbb{R}$, $\alpha > 0$ and $0 < \theta < 1$, and $K_\nu(\cdot)$ is the *modified Bessel function of the second kind* of order ν (Olver *et al.*, 2010, §10.25(i), §10.25(ii)).

The GIGP model (5.1) is a mixed Poisson distribution (Sichel, 1971),

$$f_j = \int_0^\infty \frac{\lambda^j \mathrm{e}^{-\lambda}}{j!}\, g(\lambda)\, \mathrm{d}\lambda \qquad (j \geq 0), \tag{5.2}$$

with the mixing density for the Poisson parameter λ chosen as a *generalised inverse Gaussian (GIG)* density (Johnson *et al.*, 1994, page 284))

$$g(\lambda) = \frac{\bigl(2\,(1-\theta)^{1/2}/\alpha\theta\bigr)^{\nu}}{2\,K_\nu\bigl(\alpha\,(1-\theta)^{1/2}\bigr)}\,\lambda^{\nu-1}\exp\biggl(-\frac{(1-\theta)\,\lambda}{\theta} - \frac{\alpha^2\theta}{4\lambda}\biggr) \qquad (\lambda > 0). \tag{5.3}$$

Remark 5.1. We follow the nomenclature of Sichel (1971). The connection with an alternative parameterisation (θ, ψ, χ) in Johnson *et al.* (1994) is via the maps $\theta \mapsto \nu$, $\psi \mapsto 2(1-\theta)/\theta$, $\chi \mapsto \alpha^2\theta/2$.

The normalisation in (5.3) is due to one of the integral representations for the Bessel function (Olver *et al.*, 2010, 10.32.10). Representation (5.2) explains why formula (5.1) defines a probability distribution,

$$\sum_{j=0}^\infty f_j = \int_0^\infty \sum_{j=0}^\infty \frac{\lambda^j \mathrm{e}^{-\lambda}}{j!}\, g(\lambda)\, \mathrm{d}\lambda = \int_0^\infty g(\lambda)\, \mathrm{d}\lambda = 1,$$

and it also leads to a curious identity for the Bessel functions, which does not seem to have been mentioned in the special functions literature,

$$\sum_{j=0}^\infty \frac{\bigl(\tfrac{1}{2}\alpha\,\theta\bigr)^j K_{\nu+j}(\alpha)}{j!} = \frac{K_\nu\bigl(\alpha\,(1-\theta)^{1/2}\bigr)}{(1-\theta)^{\nu/2}}. \tag{5.4}$$

From formula (5.2), the expression (5.1) is easily obtained using the normalisation of the GIG density (5.3) with parameters θ and α replaced by $\tilde{\theta} = \theta/(1+\theta)$ and $\tilde{\alpha} = \alpha\sqrt{1+\theta}$, respectively. Furthermore, formula (5.2) implies that the expected value of the GIGP distribution (5.1) coincides with that of the GIG distribution (5.3),

$$\mu = \sum_{j=0}^{\infty} j f_j = \int_0^{\infty} \sum_{j=0}^{\infty} j \, \frac{\lambda^j \mathrm{e}^{-\lambda}}{j!} \, g(\lambda) \, \mathrm{d}\lambda = \int_0^{\infty} \lambda \, g(\lambda) \, \mathrm{d}\lambda$$
$$= \frac{\alpha\,\theta}{2\,(1-\theta)^{1/2}} \cdot \frac{K_{\nu+1}\bigl(\alpha\,(1-\theta)^{1/2}\bigr)}{K_\nu\bigl(\alpha\,(1-\theta)^{1/2}\bigr)}, \tag{5.5}$$

where the last computation is based on the normalisation in (5.3) with order $\nu+1$. Expression (5.5) follows directly from the definition (5.1) by using the identity (5.4) with order $\nu+1$.

As was pointed out by Sichel (1985, p. 315), the frequencies (5.1) satisfy the recurrence relation

$$f_{j+2} = \frac{(\nu+j+1)\,\theta}{j+2}\, f_{j+1} + \frac{\alpha^2\theta^2}{4\,(j+2)\,(j+1)}\, f_j \qquad (j \in \mathbb{N}_0),$$

which can be obtained by integration by parts of the integral representation mentioned above after formula (5.3).

The tail of the GIGP distribution (5.1) has a power-geometric decay, as can be shown using Stirling's formula (Olver *et al.*, 2010, 5.11.3) and the asymptotics (B.7) of the Bessel function of large order, yielding

$$f_j \sim \frac{(1-\theta)^{\nu/2}\,\bigl(\tfrac{1}{2}\alpha\bigr)^{-\nu}}{2K_\nu\bigl(\alpha\,(1-\theta)^{1/2}\bigr)}\, j^{\nu-1}\theta^j \qquad (j \to \infty). \tag{5.6}$$

5.1.2 The boundary case $\alpha = 0$

The value $\alpha = 0$ can also be included in the GIGP class via the limit $\alpha \to 0+$. To this end, we need to consider several cases for the value of the order ν. Namely, if $\nu > 0$ then, using the small argument asymptotics of the Bessel function (see (B.2)), we obtain from (5.1)

$$f_j \sim (1-\theta)^\nu\, \frac{\Gamma(\nu+j)\,\theta^j}{\Gamma(\nu)\,j!} = \binom{\nu+j-1}{j}\,(1-\theta)^\nu\,\theta^j \qquad (j \in \mathbb{N}_0), \tag{5.7}$$

where $\Gamma(z) := \int_0^\infty s^{z-1}\,\mathrm{e}^{-s}\,\mathrm{d}s$ $(z > 0)$ is the gamma function (Olver *et al.*, 2010, 5.2.1). Formula (5.7) defines a *negative binomial distribution* with parameters ν

and θ (Johnson *et al.*, 2005, Sec. 5.1), with the expected value given by

$$\mu = \frac{\nu\,\theta}{1-\theta}. \tag{5.8}$$

The latter expression is consistent with the limit of (5.5) as $\alpha \to 0+$ (again using (B.2)). The tail behaviour of (5.7) is retrieved with the aid of Stirling's formula (Olver *et al.*, 2010, 5.11.3),

$$f_j \sim \frac{(1-\theta)^\nu\, j^{\nu-1}\theta^j}{\Gamma(\nu)} \qquad (j \to \infty), \tag{5.9}$$

which is formally in agreement with the limit of (5.6) as $\alpha \to 0+$.

However, for $\nu \le 0$ the limiting GIGP distribution degenerates to $f_0 = 1$ and $f_j = 0$ for all $j \ge 1$. Indeed, for $\nu = 0$ we get, using the asymptotic formula (B.4),

$$f_0 = \frac{K_0(\alpha)}{K_0\bigl(\alpha\,(1-\theta)^{1/2}\bigr)} \sim \frac{-\log\alpha}{-\log\bigl(\alpha\,(1-\theta)^{1/2}\bigr)} \to 1.$$

For $\nu < 0$, with the aid of the asymptotic formulas (B.1) and (B.2) we have

$$f_0 = \frac{(1-\theta)^{\nu/2}\,K_\nu(\alpha)}{K_\nu\bigl(\alpha\,(1-\theta)^{1/2}\bigr)} \sim \frac{(1-\theta)^{\nu/2}\,\frac{1}{2}\Gamma(-\nu)\bigl(\frac{1}{2}\alpha\bigr)^\nu}{\frac{1}{2}\Gamma(-\nu)\bigl(\frac{1}{2}\alpha\,(1-\theta)^{1/2}\bigr)^\nu} = 1.$$

To rectify this degeneracy, we switch to the zero-truncated GIGP distribution defined by

$$\mathsf{P}(X_i = j \,|\, X_i \ge 1) = \frac{f_j}{1-f_0} \qquad (j \in \mathbb{N})$$

and taken in the limit as $\alpha \to 0+$. We denote the resulting conditional frequencies by $(\check{f}_j)$ $(j \in \mathbb{N})$, and the corresponding expected value by $\check{\mu}$. We restrict analysis to the range $-1 < \nu \le 0$, and consider separately the cases $\nu = 0$ and $-1 < \nu < 0$ (see Remark 5.3 below for why the value $\nu = -1$ is not compatible with $\alpha = 0$).

Remark 5.2. The case $\nu < -1$ with $\alpha > 0$ is excluded from consideration (see Proposition 5.2(e) and a comment before this proposition). Hence, it is of no interest for us to consider the limit $\alpha \to 0$ here.

Case $\nu = 0$

Applying the asymptotic formula (B.4), we obtain

$$1 - f_0 = \frac{K_0\bigl(\alpha\,(1-\theta)^{1/2}\bigr) - K_0(\alpha)}{K_0\bigl(\alpha\,(1-\theta)^{1/2}\bigr)} \sim \frac{\log(1-\theta)}{\log\alpha},$$

whereas (B.2) and (B.5) give for $j \geq 1$

$$f_j = \frac{\left(\frac{1}{2}\alpha\,\theta\right)^j}{j!} \cdot \frac{K_j(\alpha)}{K_0\big(\alpha\,(1-\theta)^{1/2}\big)} \sim \frac{1}{-\log\alpha} \cdot \frac{\theta^j}{j},$$

using that $\Gamma(j) = (j-1)!$. Hence,

$$\frac{f_j}{1-f_0} \sim \check{f}_j := \frac{1}{-\log(1-\theta)} \cdot \frac{\theta^j}{j} \qquad (j \in \mathbb{N}), \tag{5.10}$$

which is *Fisher's logarithmic series distribution* (Johnson *et al.*, 2005, Sec. 7.1.2). Note that the tail behaviour of (5.10) is automatically power-geometric akin to (5.9) (with $\nu = 0$). The expected value of this distribution is easily computed,

$$\check{\mu} = \frac{1}{-\log(1-\theta)} \sum_{j=1}^{\infty} \theta^j = \frac{\theta}{(1-\theta)\big(-\log(1-\theta)\big)}. \tag{5.11}$$

Case $-1 < \nu < 0$

With the aid of the asymptotic formula (B.6) we get

$$\begin{aligned} 1 - f_0 &= \frac{K_\nu\big(\alpha\,(1-\theta)^{1/2}\big) - (1-\theta)^{\nu/2}\,K_\nu(\alpha)}{K_\nu\big(\alpha\,(1-\theta)^{1/2}\big)} \\ &\sim \frac{\Gamma(\nu+1)}{(-\nu)\,\Gamma(-\nu)} \left(\tfrac{1}{2}\alpha\right)^{-2\nu} \left(1 - (1-\theta)^{-\nu}\right), \end{aligned}$$

and furthermore, for $j \geq 1$,

$$f_j \sim \frac{(1-\theta)^{\nu/2}\left(\frac{1}{2}\alpha\,\theta\right)^j}{j!} \cdot \frac{\frac{1}{2}\Gamma(\nu+j)\left(\frac{1}{2}\alpha\right)^{-\nu-j}}{\frac{1}{2}\Gamma(-\nu)\left(\frac{1}{2}\alpha\,(1-\theta)^{1/2}\right)^{\nu}} \sim \frac{\Gamma(\nu+j)\,\theta^j}{\Gamma(-\nu)\,j!}\left(\tfrac{1}{2}\alpha\right)^{-2\nu}.$$

Hence,

$$\frac{f_j}{1-f_0} \sim \check{f}_j := \frac{(-\nu)\,\Gamma(\nu+j)\,\theta^j}{\Gamma(\nu+1)\left(1-(1-\theta)^{-\nu}\right)j!} \qquad (j \in \mathbb{N}). \tag{5.12}$$

This is an *extended negative binomial* distribution (Johnson *et al.*, 2005, Sec. 5.12.2), with the expected value

$$\check{\mu} = \frac{(-\nu)\,\theta\,(1-\theta)^{-\nu-1}}{1-(1-\theta)^{-\nu}}. \tag{5.13}$$

The tail decay of the distribution (5.12) is easily obtained using Stirling's formula (Olver *et al.*, 2010, 5.11.3),

$$\check{f}_j \sim \frac{(-\nu)\,j^{\nu-1}\,\theta^j}{\Gamma(\nu+1)\left(1-(1-\theta)^{-\nu}\right)} \qquad (j \to \infty). \tag{5.14}$$

Remark 5.3. If $\nu = -1$ then, using (B.1), (B.3) and (B.4), we have

$$1 - f_0 = \frac{K_1\bigl(\alpha\,(1-\theta)^{1/2}\bigr) - (1-\theta)^{-1/2}\,K_1(\alpha)}{K_1\bigl(\alpha\,(1-\theta)^{1/2}\bigr)} \sim \tfrac{1}{2}\alpha^2\theta\,(-\log\alpha),$$

and

$$f_1 = \frac{(1-\theta)^{-1/2}\,\bigl(\tfrac{1}{2}\alpha\,\theta\bigr)\,K_0(\alpha)}{K_1\bigl(\alpha\,(1-\theta)^{1/2}\bigr)} \sim \tfrac{1}{2}\alpha^2\theta\,(-\log\alpha)\,,$$

hence

$$\frac{f_1}{1-f_0} \sim \check{f}_1 = 1.$$

Thus, the limiting conditional distribution $(\check{f}_j)$ appears to be degenerate, with all mass concentrated at $j = 1$. This is unsuitable for the modeling purposes, which explains why the “corner” case $\nu = -1$, $\alpha = 0$ is excluded from consideration.

5.1.3 Asymptotics of the GIGP mean

As indicated by the integer partition example in Section 3.3, for the existence of a meaningful limit shape, the area of the Young diagram must grow faster than the number of constituent blocks (see (3.44)). In the context of the item production model, this means that the total number of items, $N = \sum_j j\,M_j$, should be much larger than the number of sources, $M = \sum_j M_j$. Recalling from (3.3) that the expected total number of items is given by $\mathsf{E}(N) = M\mu$ (where $\mu = \mathsf{E}(X_i)$ is the expected number of items per source, see (3.1)), this implies that a suitable limiting regime is determined by $\mu \to \infty$.

In turn, from the expression (5.5) for the GIGP mean μ, one can hypothesise that the latter is achieved if $\theta \approx 1$, while the parameters α and ν are kept fixed. This can be verified (cf. Proposition 5.2 below) using the known asymptotic formulas for the Bessel function $K_\nu(z)$ with $z \to 0$, adapted to our needs in the next lemma.

Lemma 5.1. *For $\alpha > 0$ and $\nu \in \mathbb{R}$ fixed, the following asymptotics hold as $\theta \to 1-$,*

$$K_\nu\bigl(\alpha\,(1-\theta)^{1/2}\bigr) \sim \begin{cases} \tfrac{1}{2}\Gamma(\nu)\bigl(\tfrac{1}{2}\alpha\bigr)^{-\nu}(1-\theta)^{-\nu/2} & (\nu > 0), \\ \tfrac{1}{2}\bigl(-\log\,(1-\theta)\bigr) & (\nu = 0), \\ \tfrac{1}{2}\Gamma(-\nu)\bigl(\tfrac{1}{2}\alpha\bigr)^{\nu}(1-\theta)^{\nu/2} & (\nu < 0). \end{cases} \qquad (5.15)$$

Proof. The leading terms of the asymptotics (5.15) follow directly from formulas (B.2) for $\nu \neq 0$ (with the aid of (B.1) for $\nu < 0$) and (B.5) for $\nu = 0$, □

Using this lemma, we can characterise more precisely the asymptotic behaviour of the GIGP mean in the limit as $1 - \theta \to 0+$. In particular, this analysis reveals that the desired growth to infinity is in place for $\nu \geq -1$, but fails for $\nu < -1$.

Proposition 5.2. *The expected values μ and $\check{\mu}$ of the GIGP ($\alpha > 0$) and zero-truncated GIGP ($\alpha = 0$) distributions, respectively, have the following asymptotics as $\theta \to 1-$.*

(a) $\nu > 0,\ \alpha \geq 0$:

$$\mu \sim \frac{\nu}{1-\theta}. \tag{5.16}$$

(b) $\nu = 0,\ \alpha \geq 0$:

$$\mu \sim \frac{1}{(1-\theta)\bigl(-\log(1-\theta)\bigr)}, \qquad \check{\mu} \sim \frac{1}{(1-\theta)\bigl(-\log(1-\theta)\bigr)}. \tag{5.17}$$

(c) $-1 < \nu < 0,\ \alpha \geq 0$:

$$\mu \sim \frac{\Gamma(\nu+1)\bigl(\frac{1}{2}\alpha\bigr)^{-2\nu}}{\Gamma(-\nu)\,(1-\theta)^{\nu+1}}, \qquad \check{\mu} \sim \frac{-\nu}{(1-\theta)^{\nu+1}}. \tag{5.18}$$

(d) $\nu = -1,\ \alpha > 0$:

$$\mu \sim \bigl(\tfrac{1}{2}\alpha\bigr)^2\bigl(-\log(1-\theta)\bigr). \tag{5.19}$$

(e) $\nu < -1,\ \alpha > 0$:

$$\mu \sim \frac{\bigl(\frac{1}{2}\alpha\bigr)^2}{-\nu-1}. \tag{5.20}$$

Proof. Consider cases (a)–(e) using the asymptotic formulas of Lemma 5.1.

(a) For $\alpha > 0$, using the first line of (5.15) for orders ν and $\nu+1$, we have

$$\frac{K_{\nu+1}\bigl(\alpha\,(1-\theta)^{1/2}\bigr)}{K_\nu\bigl(\alpha\,(1-\theta)^{1/2}\bigr)} \sim \frac{\frac{1}{2}\,\Gamma(\nu+1)\bigl(\frac{1}{2}\alpha\,(1-\theta)^{1/2}\bigr)^{-\nu-1}}{\frac{1}{2}\,\Gamma(\nu)\bigl(\frac{1}{2}\alpha\,(1-\theta)^{1/2}\bigr)^{-\nu}} = \frac{\nu}{\frac{1}{2}\alpha\,(1-\theta)^{1/2}}, \tag{5.21}$$

where we also used the recurrence property of the gamma function, $\Gamma(\nu+1) = \nu\,\Gamma(\nu)$ (Olver *et al.*, 2010, 5.5.1). Substituting this into (5.5) gives (5.16). If $\alpha = 0$ then (5.16) readily follows from (5.8).

(b) For $\alpha > 0$, formulas (5.15) with $\nu = 0$ and $\nu = 1$ give

$$\frac{K_1\bigl(\alpha\,(1-\theta)^{1/2}\bigr)}{K_0\bigl(\alpha\,(1-\theta)^{1/2}\bigr)} \sim \frac{\bigl(\frac{1}{2}\alpha\,(1-\theta)^{1/2}\bigr)^{-1}}{-\log{(1-\theta)}}, \tag{5.22}$$

and the first formula in (5.17) follows from (5.5). If $\alpha = 0$ then formula (5.11) immediately gives the second formula in (5.17).

(c) For $\alpha > 0$, using the symmetry relation (B.1), similarly to (5.21) we obtain

$$\frac{K_{\nu+1}\bigl(\alpha\,(1-\theta)^{1/2}\bigr)}{K_\nu\bigl(\alpha\,(1-\theta)^{1/2}\bigr)} \sim \frac{\Gamma(\nu+1)\bigl(\frac{1}{2}\alpha\,(1-\theta)^{1/2}\bigr)^{-2\nu-1}}{\Gamma(-\nu)},$$

and the first formula in (5.18) then follows from (5.5). The second formula in (5.18) is immediate from (5.13).

(d) Follows from (5.5) using the symmetry relation (B.1) and the asymptotic ratio (5.22).

(e) Again using (B.1) and the first line of (5.15) with orders $-\nu > 0$ and $-\nu - 1 > 0$, we obtain

$$\frac{K_{\nu+1}\bigl(\alpha\,(1-\theta)^{1/2}\bigr)}{K_\nu\bigl(\alpha\,(1-\theta)^{1/2}\bigr)} \sim \frac{\Gamma(-\nu-1)\bigl(\frac{1}{2}\alpha\,(1-\theta)^{1/2}\bigr)}{\Gamma(-\nu)},$$

and (5.20) follows from (5.5), again using the recurrence $\Gamma(z+1) = z\,\Gamma(z)$ (Olver *et al.*, 2010, 5.5.1), now with $z = -\nu - 1$.

Thus, the proof of Proposition 5.2 is complete. □

Proposition 5.2 describes the growth of the expected value μ (for $\alpha > 0$) or $\breve{\mu}$ (for $\alpha = 0$) in terms of the small parameter $1 - \theta$. For the purposes of the GIGP model fitting, it is useful to express $1 - \theta$ through μ or $\breve{\mu}$, respectively, by solving the asymptotic equations (5.16), (5.17), (5.18), and (5.19).

Proposition 5.3. *Under the conditions of Proposition 5.2, the following asymptotics hold.*

(a) $\nu > 0,\ \ \alpha \ge 0$:

$$1 - \theta \sim \frac{\nu}{\mu}.$$

(b) $\nu=0,\ \alpha\ge 0$:

$$1-\theta\sim\frac{1}{\mu\log\mu},\qquad 1-\theta\sim\frac{1}{\breve{\mu}\log\breve{\mu}}.$$

(c) $-1<\nu<0,\ \alpha\ge 0$:

$$1-\theta\sim\left(\frac{\Gamma(\nu+1)\bigl(\frac{1}{2}\alpha\bigr)^{-2\nu}}{\Gamma(-\nu)\,\mu}\right)^{1/(\nu+1)},\qquad 1-\theta\sim\left(\frac{-\nu}{\breve{\mu}}\right)^{1/(\nu+1)}.$$

(d) $\nu=-1,\ \alpha>0$:

$$\log(1-\theta)\sim-\frac{4\mu}{\alpha^2}.$$

Remark 5.4. Formula (5.19) provides only the logarithmic asymptotics of $1-\theta$, but this suffices for the estimation purposes.

5.2 The Limit Shape in the GIGP Model

5.2.1 Scaling coefficients and the main theorem

Let the frequencies f_j ($j\in\mathbb{N}_0$) be given by the GIGP distribution formula (5.1) with parameters $0<\theta<1$, $\nu\ge-1$ and $\alpha\ge 0$, excluding the "corner" pair $\nu=-1$, $\alpha=0$. The case $\alpha=0$ is understood as the limit of conditional probabilities $\mathsf{P}(X_i=j\,|\,X_i>0)=f_j/(1-f_0)$ ($j\in\mathbb{N}$) as $\alpha\to 0+$ (see Section 5.1.2).

Given the random vector of observed multiplicities (M_j) produced by M sources, our aim is to study the asymptotics of scaled Young diagrams with the boundary (see (3.25))

$$\widetilde{Y}(x):=\frac{Y(Ax)}{B}=\frac{1}{B}\sum_{j\ge Ax}M_j=\frac{1}{B}\sum_{i=1}^{M}Z_i(Ax)\qquad(x\ge 0).\tag{5.23}$$

We proceed under the following assumptions on the limiting regime, including the specification of the scaling coefficients A and B.

Assumption 5.1. The number of sources is large, $M\to\infty$. In addition, the intrinsic parameter $\theta\in(0,1)$ is assumed to be close to its upper limit 1, that is, $\theta\to 1-$, which guarantees that the mean number of items per source is large (see Proposition 5.2).

Assumption 5.2. The x-scaling coefficient A is chosen to be

$$A = \frac{1}{-\log\theta} \sim \frac{1}{1-\theta} \to \infty \qquad (\theta \to 1-), \tag{5.24}$$

whereas the y-scaling coefficient B is specified according to particular domains in the space of parameters ν and α as follows:

(a) $\nu > 0$, $\alpha \geq 0$:

$$B = \frac{M}{\Gamma(\nu)}. \tag{5.25}$$

(b) $\nu = 0$, $\alpha \geq 0$:

$$B = \frac{M}{-\log(1-\theta)}. \tag{5.26}$$

(c) $-1 \leq \nu < 0$, $\alpha > 0$:

$$B = \frac{M\left(\frac{1}{2}\alpha\right)^{-2\nu}(1-\theta)^{-\nu}}{\Gamma(-\nu)}. \tag{5.27}$$

(d) $-1 < \nu < 0$, $\alpha = 0$:

$$B = \frac{M(-\nu)(1-\theta)^{-\nu}}{\Gamma(\nu+1)}. \tag{5.28}$$

Assumption 5.3. The y-scaling coefficient B defined in Assumption 5.2 is large, $B \to \infty$. For $\nu > 0$, this is automatic according to (5.25) (as long as $M \to \infty$), but for $\nu \leq 0$ we must assume in addition that $M \gg -\log(1-\theta)$ if $\nu = 0$ and $M \gg (1-\theta)^{\nu}$ if $\nu < 0$.

Remark 5.5. The need to impose an additional condition in Assumption 5.3 on the joint limiting behaviour of the external parameter $M \to \infty$ and the intrinsic GIGP parameter $\theta \to 1-$ for $\nu \leq 0$ shows that, in order to have a manifested limit shape in the data, the number of sources, M, must be sufficiently large. We will clarify the opposite situation below in Section 5.4.

For $\nu \geq -1$, consider the function

$$\varphi_\nu(x) := \int_x^\infty s^{\nu-1}\,\mathrm{e}^{-s}\,\mathrm{d}s \qquad (x > 0), \tag{5.29}$$

which is the *(upper) incomplete gamma function* (Olver *et al.*, 2010, 8.2.2). The following is our main result, establishing convergence in probability of the scaled Young diagrams (see (5.23)) to the limit shape $\varphi_\nu(x)$.

Theorem 5.4. *Under Assumptions 5.1, 5.2 and 5.3, for each $x > 0$ and any $\varepsilon > 0$, we have*

$$\mathsf{P}\left(\bigl|\widetilde{Y}(x) - \varphi_\nu(x)\bigr| \ge \varepsilon\right) \to 0. \tag{5.30}$$

The proof of Theorem 5.4 is postponed to Sections 5.2.3 and 5.2.4, after we illustrate convergence to the limit shape using computer simulations in the next section.

5.2.2 Graphical illustration using computer simulations

In this section, we illustrate the limit shape approximation using computer simulated data in two example cases, with $\nu = 0.5$ and $\nu = -0.5$ (see Fig. 5.1, left panels). The other parameter settings are as follows, $\alpha = 2$. $\theta = 0.99$, and $M = 1000$. The plots depict the data as the upper boundary of the Young diagram $Y(x)$ defined in (3.17) and the theoretical GIGP complementary distribution function $\bar{F}(x)$ (see (3.21) and (5.1)), along with the limit shape scaled back to the original frequencies of counts, that is, $x \mapsto B\,\varphi_\nu(x/A)$, where $A = -1/\log\theta \doteq 99.49916$ (see (5.24)) and $B \doteq 564.1896$ for $\nu = 0.5$ or $B \doteq 56.41896$ for $\nu = -0.5$ (see (5.25) and (5.27), respectively). In both cases, the plots show a very good fit of the limit shape in the bulk of the observed values.

The inspection of the tail behaviour is facilitated by observing from (5.29) that

$$\begin{aligned}\varphi_\nu(x) &= -\int_x^\infty s^{\nu-1}\,\mathrm{d}(\mathrm{e}^{-s})\\ &= x^{\nu-1}\,\mathrm{e}^{-x} + (\nu-1)\int_x^\infty s^{\nu-2}\,\mathrm{e}^{-s}\,\mathrm{d}s \sim x^{\nu-1}\,\mathrm{e}^{-x} \qquad (x\to\infty).\end{aligned}$$

Therefore, according to (5.23) and (5.30), it may be expected that, for large enough x,

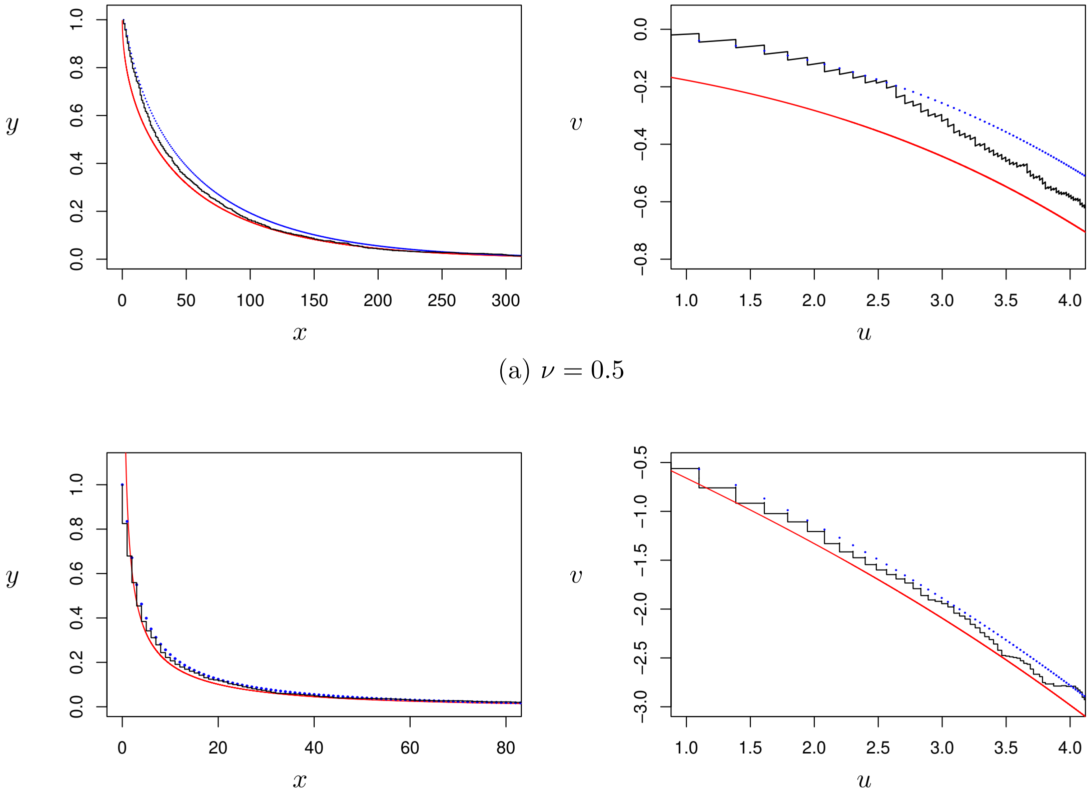
$$y = Y(Ax) \approx B\,\varphi_\nu(x) \approx B\,x^{\nu-1}\,\mathrm{e}^{-x},$$

or, taking the logarithm,

$$\log Y(Ax) + x \approx \log B + (\nu-1)\log x. \tag{5.31}$$

Hence, switching from (x, y) to the new coordinates

$$u = \log x, \qquad v = \log y + x, \tag{5.32}$$



(a) $\nu = 0.5$

(b) $\nu = -0.5$

Figure 5.1: Illustration of the limit shape approximation using $M = 1000$ random values (X_i) simulated using the GIGP model (5.1) with parameters $\theta = 0.99$, $\alpha = 2$, and (a) $\nu = 0.5$ or (b) $\nu = -0.5$. In the left panels, the black stepwise plots represent the upper boundary $Y(x)$ of the corresponding Young diagrams, together with the GIGP complementary cumulative distribution function $\bar{F}(x)$ shown as blue dotted plots, while the smooth red curves represent the scaled back limit shape, $x \mapsto B\,\varphi_\nu(x/A)$. In the right panels, the tails are shown in transformed coordinates (5.32), with the same line and colour coding.

a transformed data plot may be expected to be close to a straight line with slope $\nu - 1$, as well as the tails of the theoretical GIGP distribution function and of the limit shape alike. This is illustrated for the simulated data in Fig. 5.1 (right panels), showing a reasonable linearisation of the long tails in both cases, $\nu = 0.5$ and $\nu = -0.5$.

The graphical method described above can be used for a quick visual check of suitability of the GIGP frequency model even before estimating the model parameters, by first experimenting with the scaling coefficient $A = -1/\log\theta$ (see (5.24)) aiming to get a linearised data plot (thus producing a crude estimate for the parameter θ), followed by reading off the fitted slope (which estimates the parameter $\nu - 1$), and then exploiting the fitted intercept (close to $\log B$, see (5.31)) to get an estimate for the parameter α using one of the formulas (5.25) to (5.28). We will apply this method to some real data sets in Section 5.5.

5.2.3 Convergence of expected Young diagrams

We start our proof of Theorem 5.4 by showing that convergence to the limit shape $\varphi_\nu(x)$ holds for the expected Young diagrams. From (3.23) and (3.25), we have

$$\mathsf{E}\big(\widetilde{Y}(x)\big) = \frac{M\bar{F}(Ax)}{B}. \tag{5.33}$$

Theorem 5.5. *Under Assumptions 5.1 and 5.2, for each $x > 0$*

$$\mathsf{E}\big(\widetilde{Y}(x)\big) \to \varphi_\nu(x). \tag{5.34}$$

Remark 5.6. Note that Assumption 5.3 is not needed in Theorem 5.5.

Remark 5.7. In calculations below, we confine ourselves to the leading asymptotics (5.6) of terms in the series $\bar{F}(Ax)$ (see (5.33)). A more careful analysis involving control over the approximation errors is straightforward by using the classic Euler–Maclaurin summation formula (Olver *et al.*, 2010, §2.10(i)) and uniform asymptotic expansions of the Bessel function of large order (Olver *et al.*, 2010, §10.41(ii)).

Proof of Theorem 5.5. The proof below is broken down according to various subdomains of the parameters ν and α (see Assumption 5.2). First, we consider the cases with $\alpha > 0$, where the GIGP distribution is supported on $j \in \mathbb{N}_0$, and then switch to the boundary cases with $\alpha = 0$, where the support is reduced to $j \in \mathbb{N}$.

- $\alpha > 0$

  Using the asymptotic approximation (5.6) of the frequencies f_j (with $j \geq Ax \geq A\delta \gg 1$), from (5.33) we obtain

$$\mathsf{E}\big(\widetilde{Y}(x)\big) = \frac{M}{B}\sum_{j\geq Ax} f_j \sim \frac{M\,(1-\theta)^{\nu/2}\left(\frac{1}{2}\alpha\right)^{-\nu}}{2BK_\nu\big(\alpha\,(1-\theta)^{1/2}\big)} \sum_{j\geq Ax} j^{\nu-1}\theta^j. \tag{5.35}$$

Recalling that $A = (-\log\theta)^{-1} \sim (1-\theta)^{-1}$, for the last sum in (5.35) we have

$$A^{-\nu} \sum_{j\ge Ax} j^{\nu-1}\theta^j = \sum_{j\ge Ax} \left(\frac{j}{A}\right)^{\nu-1} \mathrm{e}^{-j/A}\,\frac{1}{A} \to \int_x^\infty s^{\nu-1}\,\mathrm{e}^{-s}\,\mathrm{d}s = \varphi_\nu(x), \tag{5.36}$$

which is evident by interpreting (5.36) as the Riemann integral sum converging to the integral on the right. Furthermore, the asymptotics of the denominator in (5.35) is obtained from formulas (5.15) (see Lemma 5.1). Hence, returning to (5.35) and recalling the definitions (5.24) of A and (5.25), (5.26), (5.27) of B, we easily obtain (5.34).

- $\alpha = 0$

  Using the tail approximations (5.9) ($\nu > 0$), (5.10) ($\nu = 0$) and (5.14) ($-1 < \nu < 0$), from (5.33) we obtain, similarly to (5.35) and (5.36),

  $$\mathsf{E}\big(\widetilde{Y}(x)\big) \sim \frac{MC_\nu(\theta)}{B} \sum_{j\ge Ax} j^{\nu-1}\theta^j \sim \frac{MC_\nu(\theta)A^\nu}{B}\,\varphi_\nu(x), \tag{5.37}$$

  where $A \sim (1-\theta)^{-1}$ (see (5.24)) and

  $$C_\nu(\theta) := \begin{cases} (1-\theta)^\nu/\Gamma(\nu) & (\nu > 0),\\ (-\log(1-\theta))^{-1} & (\nu = 0),\\ (-\nu)/\Gamma(\nu+1) & (-1<\nu<0). \end{cases}$$

  Now, using the specifications (5.25), (5.26), or (5.28), it is immediate to see that the right-hand side of (5.37) is reduced to $\varphi_\nu(x)$.

This completes the proof of Theorem 5.5. □

5.2.4 Convergence of random Young diagrams

Theorem 5.6. *Under Assumptions 5.1, 5.2 and 5.3, for each $x > 0$ the mean squared deviation of $\widetilde{Y}(x)$ from the limit shape $\varphi_\nu(x)$ is asymptotically small,*

$$\mathsf{E}\big(\big|\widetilde{Y}(x) - \varphi_\nu(x)\big|^2\big) \to 0. \tag{5.38}$$

This implies convergence in probability, $\widetilde{Y}(x)\xrightarrow{\mathrm{p}}\varphi_\nu(x)$*, that is, for each* $x>0$ *and any* $\varepsilon>0$*,*

$$\mathsf{P}\big(\big|\widetilde{Y}(x)-\varphi_\nu(x)\big|\ge\varepsilon\big)\to 0. \tag{5.39}$$

Proof. By the standard decomposition of the mean squared deviation, we have

$$\mathsf{E}\big(\big|\widetilde{Y}(x)-\varphi_\nu(x)\big|^2\big)=\mathsf{Var}\big(\widetilde{Y}(x)\big)+\big(\mathsf{E}\big(\widetilde{Y}(x)\big)-\varphi_\nu(x)\big)^2. \tag{5.40}$$

Using formulas (3.23) and (3.25), the variance term in (5.40) is estimated as follows,

$$\begin{aligned}\mathsf{Var}\big(\widetilde{Y}(x)\big)&=\frac{M\bar{F}(Ax)F(Ax)}{B^2}\\&\le\frac{M\bar{F}(Ax)}{B^2}\sim\frac{\varphi_\nu(x)}{B}\to 0,\end{aligned} \tag{5.41}$$

according to (5.33), (5.34), and also Assumption 5.3, which guarantees that $B\to\infty$. By Theorem 5.5, convergence in (5.41) is uniform on $[\delta,\infty)$, for every $\delta>0$. As for the second term on the right-hand side of (5.40), due to Theorem 5.5 it is asymptotically small, uniformly on every interval $[\delta,\infty)$. Hence, the limit (5.38) follows.

Finally, convergence in probability (5.39) is a standard consequence of (5.38) due to Chebyshev's inequality (Shiryaev, 1996, Sec. II.6, p. 192):

$$\mathsf{P}\big(\big|\widetilde{Y}(x)-\varphi_\nu(x)\big|\ge\varepsilon\big)\le\frac{\mathsf{E}\big(\big|\widetilde{Y}(x)-\varphi_\nu(x)\big|^2\big)}{\varepsilon^2}\to 0,$$

according to (5.38). □

5.3 Fluctuations of Random Young Diagrams

As mentioned in Section 3.2, $\widetilde{Y}(x)$ may be expected to be asymptotically normal, with mean $\mathsf{E}\big(\widetilde{Y}(x)\big)=M\bar{F}(Ax)/B\sim\varphi_\nu(x)$ and variance $M\bar{F}(Ax)F(Ax)/B^2\sim\varphi_\nu(x)/B$ (see (5.33), (5.34) and (5.41)). We prove this result below, using the method of characteristic functions.

Theorem 5.7. *Under Assumptions 5.1, 5.2 and 5.3, for any* $x>0$*,*

$$\varUpsilon(x):=\sqrt{\frac{B}{\varphi_\nu(x)}}\left(\widetilde{Y}(x)-\frac{M\bar{F}(Ax)}{B}\right)\xrightarrow{\mathrm{d}}\mathcal{N}(0,1), \tag{5.42}$$

where $\mathcal{N}(0,1)$ *is a standard normal law (i.e., with zero mean and unit variance), and* $\xrightarrow{\mathrm{d}}$ *denotes convergence in distribution.*

Proof. Substituting (5.23), the left-hand side of (5.42) is rewritten as

$$\Upsilon(x)=\frac{1}{\sqrt{B\,\varphi_\nu(x)}}\sum_{i=1}^{M}\bigl(Z_i(Ax)-\bar F(Ax)\bigr). \tag{5.43}$$

The characteristic function of (5.43) is given by

$$\psi(t;x):=\mathsf{E}\bigl(\mathrm{e}^{\mathrm{i}t\Upsilon(x)}\bigr)=\mathrm{e}^{-\mathrm{i}\tilde t M\bar F(Ax)}\left(1+\bar F(Ax)\bigl(\mathrm{e}^{\mathrm{i}\tilde t}-1\bigr)\right)^M, \tag{5.44}$$

where

$$\tilde t=\frac{t}{\sqrt{B\,\varphi_\nu(x)}},\qquad t\in\mathbb{R}. \tag{5.45}$$

Choosing the principal branch of the logarithm function $\mathbb{C}\setminus\{0\}\ni z\mapsto\log z\in\mathbb{C}$ (i.e., such that $\log 1=0$), we can rewrite (5.44) as

$$\log\psi(t;x)=-\mathrm{i}\tilde t M\bar F(Ax)+M\log(1+w), \tag{5.46}$$

where

$$w:=\bar F(Ax)\bigl(\mathrm{e}^{\mathrm{i}\tilde t}-1\bigr). \tag{5.47}$$

Since $A\to\infty$ and $B\to\infty$ (by Assumptions 5.2 and 5.3), we have $\tilde t\to 0$ and $w\to 0$, hence

$$\log(1+w)=w-\tfrac12 w^2+O(|w|^3).$$

Therefore, Taylor expanding $\mathrm{e}^{\mathrm{i}\tilde t}=1+\mathrm{i}\tilde t-\frac12\tilde t^2+O(\tilde t^3)$ and substituting (5.45) and (5.47), formula (5.46) is elaborated as follows,

$$\log\psi(t;x)=-\frac{M\bar F(Ax)\,F(Ax)\,t^2}{2B\,\varphi_\nu(x)}+O\left(\frac{M\bar F(Ax)}{B^{3/2}}\right)\to-\frac{t^2}{2},$$

using that $M\bar F(Ax)/B\sim\varphi_\nu(x)$ and $F(Ax)\to 1$ for any $x>0$. Thus, $\psi(t;x)\to\mathrm{e}^{-t^2/2}$, which is the characteristic function of the normal distribution $\mathcal{N}(0,1)$, as claimed. □

5.4 Poisson Approximation in the “Chaotic” Regime

In this section, we consider the case wherein Assumption 5.3 is not satisfied, so that the y-scaling coefficient B is bounded (which is only possible for $\nu \le 0$, see formulas (5.25) to (5.28)). We call this case *chaotic* because convergence of the random variable $\widetilde{Y}(x)$ to the limit shape $\varphi_\nu(x)$ does not hold here (cf. Theorem 5.6), despite convergence of the expected value $\mathsf{E}\bigl(\widetilde{Y}(x)\bigr) = M\bar{F}(Ax)/B \to \varphi_\nu(x)$ (Theorem 5.5). The root cause of this failure is that, although $\widetilde{Y}(x)$ is a normalised sum of independent Bernoulli variables $Z_i(Ax) = I_{\{X_i \ge Ax\}}$ (see (5.23)), the success probability $\mathsf{P}(X_i \ge Ax) = \bar{F}(Ax)$ tends to zero, which is not offset by a fast enough growth of the number of terms M (see Remark 5.5).

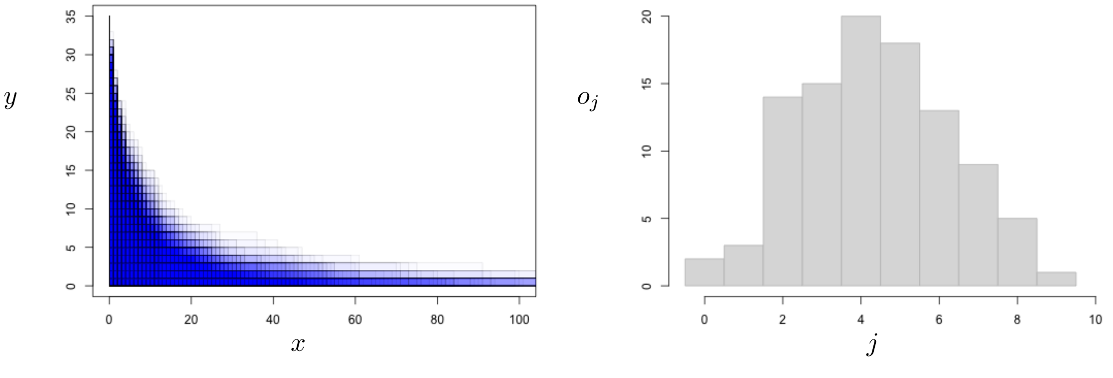
For orientation, consider a stylised case where $B = 1$, then $M\bar{F}(Ax) \to \varphi_\nu(x)$ and, according to the classic Poisson “law of small numbers” Whitaker (1914), the binomial distribution of the sum $\widetilde{Y}(x) = Z_1(Ax) + \dots + Z_M(Ax)$ is asymptotically close to a Poisson distribution with parameter $\lambda = \varphi_\nu(x)$. That is to say, the sums $\widetilde{Y}(x)$ do not settle down to a deterministic constant (like in a law of large numbers (3.27)) but, due to a persistent “small” randomness, admit a non-degenerate (Poisson) approximation without any normalisation. This observation is generalised as follows.

Theorem 5.8. *Suppose that Assumptions 5.1 and 5.2 are satisfied but Assumption 5.3 is not, so that $B = O(1)$. Then the distribution of the random variable $Y(Ax)$ for $x > 0$ is approximated by a Poisson distribution with parameter $M\bar{F}(Ax) \sim B\,\varphi_\nu(x)$ and with the corresponding error in total variation distance (or in Kolmogorov's uniform distance) bounded by $O(M^{-1}) = o(1)$.*

Proof. This is an immediate consequence of a well-known approximation for the binomial distribution of the total number of successes in n independent Bernoulli trials, with success probability p, by a Poisson distribution with parameter $\lambda = np$, with the error bounded by $\sigma^2 = np^2$ (see, e.g., Barbour *et al.* (1992); Novak (2019)). In our case, $\lambda = M\bar{F}(Ax)$ and $\sigma^2 = M\bar{F}(Ax)^2 \sim \bigl(B\,\varphi_\nu(x)\bigr)^2/M = O(1/M) = o(1)$. □

The Poisson approximation stated in Theorem 5.8 is illustrated in Fig. 5.2 using 100 simulated samples (of size $M = 35$ each) from the GIGP distribution (5.1) with parameters $\nu = -0.5$, $\alpha = 2$, and $\theta = 0.99$. The y-scaling coefficient computed from (5.27) is given by $B \doteq 1.974664$, confirming that this is a chaotic regime (i.e., where Assumption 5.3 is not satisfied). The x-scaling coefficient (5.24) specialises to $A \doteq 99.49916$. The left panel in Fig. 5.2 shows the sample Young diagrams superimposed on one another using transparent shading (in blue), so that darker places correspond to a more frequent occurrence. As anticipated, there is no convergence to a deterministic limit shape, but an emerging "typical" boundary of blue diagrams clearly indicates an expected curve establishing in the limit.

In the right panel of Fig. 5.2, we choose a trial value $x_0 = 0.2$ and plot a histogram for the observed frequencies of the random values $Y(Ax_0)$, where $Ax_0 \doteq 19.89983$. A visual inspection supports a reasonable match with Poisson distribution with the mean $M\bar{F}(Ax_0) \doteq 4.342498$. This is confirmed by Pearson's χ^2-test, with the bins labelled by the values of j from 0 to 9 and the respective observed frequencies o_j. Since the expected frequencies $e_0 \doteq 1.3004$ and $e_9 \doteq 3.340067$ are less than 5, we follow a common recommendation and combine the bins $j = 0$ and $j = 9$ with $j = 1$ and $j = 8$, respectively. The grouped χ^2-statistic is calculated to yield 1.972246 on $10 - 2 - 1 = 7$ degrees of freedom, with the p-value of 96.14%, so the goodness-of-fit test is comfortably passed.



(a)

(b)

Figure 5.2: Illustration of Poisson statistics in the chaotic regime. The left panel shows superimposed Young diagrams of 100 random samples of size $M = 35$ each, generated from the GIGP model (5.1) with parameters $\nu = -0.5$, $\alpha = 2$, and $\theta = 0.99$. The right panel shows the histogram of observed frequencies o_j of the random values $Y(Ax_0)$, where $Ax_0 \doteq 19.89983$. For orientation, the mean of the approximating Poisson distribution is given by $M\bar{F}(Ax_0) \doteq 4.342498$.

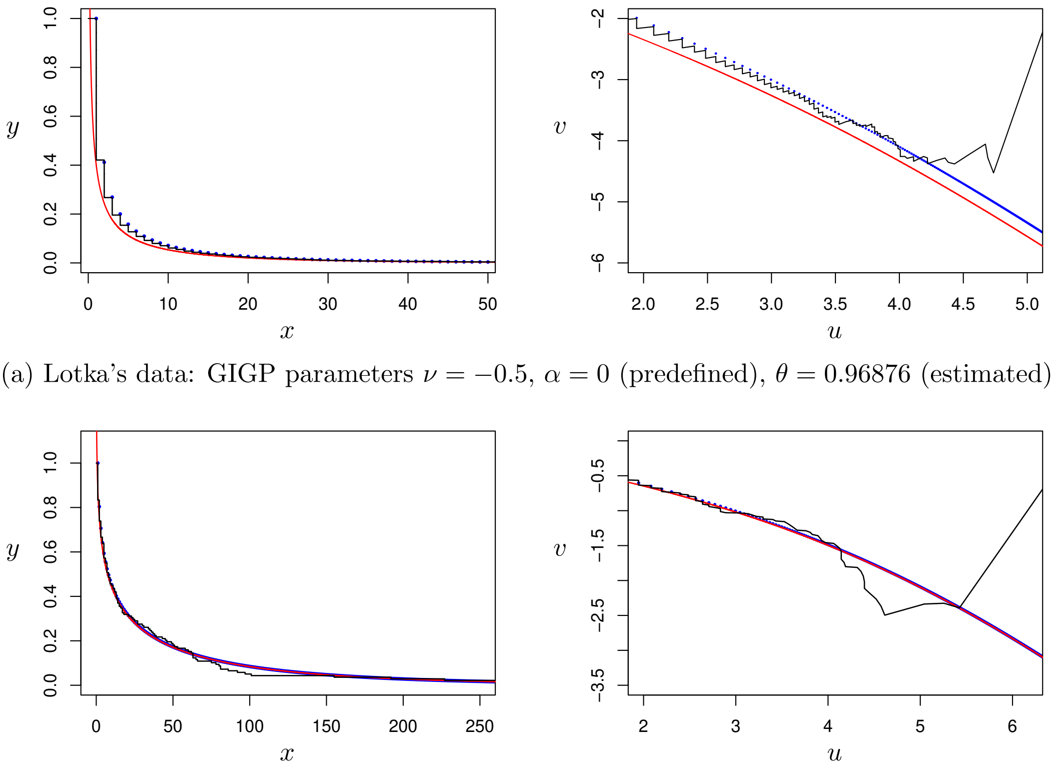
5.5 Real Data Examples

In this section, we look at how well the theoretical limit shape $\varphi_\nu(x)$ conforms to some real data sets studied earlier by Sichel (1985).

5.5.1 Lotka's data set: author productivity

We start with a classic data set considered by Lotka in his seminal paper Lotka (1926), comprising the counts of the number of papers (items) published by authors (sources) in *Chemical Abstracts* during 1907–1916. This data set is usually considered as a baseline example of the power law statistics of counts (Clauset *et al.*, 2009; Lotka, 1926), but Sichel (1985, pp. 316–317 and Table 2) argued that a GIGP model with predefined parameters $\nu = -0.5$, $\alpha = 0$ and an estimated $\theta = 0.96876$ is a better fit to the data (with the p-value of 91.8% in Pearson's χ^2-test). The distinctive difference between the two models is of course the long-tail behaviour, either power or power-geometric, respectively.

To examine the goodness-of-fit graphically, similarly to Section 4.3.1 we first plot the empirical Young diagram $Y(x)$, compared with the fitted GIGP complementary distribution function $\bar{F}(x)$ and contrasted with the theoretical limit shape scaled back to the original coordinates, that is, $x \mapsto B\,\varphi(x/A)$. Here, $M = 6\,891$, and the scaling coefficients calculated from (5.24) and (5.28) are given by $A \doteq 31.5076$ and $B \doteq 343.5839$. Although the value of A is not particularly large (because θ is not extremely close to 1), a big value of B confirms a reasonable predisposition of the data for a good limit shape approximation.



(a) Lotka's data: GIGP parameters $\nu = -0.5$, $\alpha = 0$ (predefined), $\theta = 0.96876$ (estimated)

(b) Chen's data: GIGP parameters $\nu = 0$, $\alpha = 0$ (predefined), $\theta = 0.99369$ (estimated).

Figure 5.3: GIGP model fit to real data sets. Black plots represent the data, blue dotted plots show the fitted GIGP complementary distribution functions, and smooth red lines depict the graphs of the scaled-back limit shape. The right panel shows the tail versions of these plots in transformed coordinates (5.32).

The left panel in Fig. 5.3(a) demonstrates an excellent fit to the bulk of the data for both the GIGP model as well as the limit shape (scaled back to the original coordinates, $x \mapsto B\,\varphi(x/A)$). The visual inspection of the tails in the right panel of Fig. 5.3(a) (in transformed coordinates (5.32)) confirms a good fit but only for moderately large observed values j, whereas the region $u \geq 4.5$, corresponding to values $j \geq \mathrm{e}^{4.5} \approx 90$, reveals increasing deviations from the GIGP prediction. This suggests that very large values in the tail of Lotka's data require a different fitting model, such as a stretched-exponential approximation Laherrere & Sornette (1998). Incidentally, upon a closer look at the upper extremes in the Lotka data set, there is just a handful of counts larger than 90, namely, 95, 107, 109, 114, and 346.

A surprising maximum 346 looks like a genuine outlier, three times bigger than the runner-up! Interestingly, this record is attributed to Professor Emil Abderhalden, a prolific and controversial Swiss biochemist and physiologist who worked in the first half of the 20th century (Wikipedia, 2023). Being rather extraordinary, perhaps this individual record needs to be removed from statistical analysis.

5.5.2 Chen's data set: journal use

For our second real data example, we revisit the data set from Chen (1972), considered by Sichel (1985, pp. 318–319 and Table 4) for the sake of testing a GIGP model. The data comprised counts of use (items) of physics journals in the M.I.T. Science Library in 1971, recorded per each volume (sources) taken from the shelves for reading or photocopying. The total number of sources involved (i.e., the number of volumes ever requested) was $M = 138$. Sichel fitted a GIGP model with predefined values $\nu = 0$, $\alpha = 0$ and an estimated $\theta = 0.99369$. He tested goodness-of-fit via Pearson's χ^2-test, observing a reasonably high p-value of 31.2%, thus not signalling any significant mismatch.

To cross-examine the fit using our methods, similarly as in Section 5.5.1 we plot the empirical Young diagram's boundary $Y(x)$ along with the fitted GIGP function $\bar{F}(x)$ and scaled-back limit shape $B\,\varphi(x/A)$ (see Fig. 5.3(b), left panel), where the scaling coefficients calculated from (5.24) and (5.26) are given by $A \doteq 157.9781$ and $B \doteq 27.24247$. It is worth pointing out that, in contrast to Lotka's data set considered in Section 5.5.1, here we have quite a large value for the x-scaling coefficient A but a relatively small value of the y-scaling parameter B.

At first glance, the plots in Fig. 5.3(b) (left panel) seem to conform to the GIGP model; however, one cannot help noticing a visible deviation from the theoretical prediction around the value $x = 100$. This is confirmed by looking at the tail plots in transformed coordinates (5.32) (see Fig. 5.3(b), right panel), where $x = 100$ corresponds to $u = \log 100 \doteq 4.60517$. To test statistically whether the deviations are significant, we can use the asymptotic normality of $Y(x)$ due to Theorem 5.7. Specifically, setting $x = 100$ and standardising according to formula (5.42), we calculate $\Upsilon(100/A) \doteq -3.413073$, with an extremely small p-value of 0.032%. Thus, the deviation is highly significant, which implies that the GIGP model is not an accurate fit, at least for moderately large values starting from about $x = 70$.

Chapter 6

Generalised Power Law

This section is concerned with the *generalised power law (GPL)* model for the frequency distribution of count data. This model aims to bridge small values of counts and a power type upper tail in order to overcome the limitations of integer partitions and power law models. As mentioned in the Introduction, we introduced the GPL model independently in Nuermaimaiti *et al.* (2021) but discovered later on that a similar model is known in the literature as a *hooked power law* (see, e.g., Thelwall & Wilson (2014)). The new features of our usage of the GPL model contrasted with previous work are summarised at the end of Section 6.4.5.

Unlike the integer partitions in Chapter 3 and the power law in Chapter 4, the GPL model achieved good results to model the whole range of the data, while the other models fit well only part of the data. Compared to the generalised inverse Gaussian-Poisson (GIGP) model introduced in Chapter 5, the GPL is computationally easier to fit to the real data. The performance of our model was verified by fitting GPL to the EJP data and the AMS data, respectively, which demonstrated that the GPL model works well in both cases.

6.1 Generalised Power Law Model

6.1.1 The model setup

The generalised power law (GPL) model introduced in this section is designed as a suitable extension of the classical power law model (4.1). It involves a shape parameter $a > 1$, akin to the power-law exponent in (4.1), and a scale parameter $L > 0$, tacitly assumed to be large.

Definition 6.1. We say that a discrete random variable X with values in $\mathbb{N}_0$ follows a *generalised power law* with parameters a and L if the frequencies $f_j = \mathsf{P}(X = j)$ are given by

$$f_j = C_{a,L} \left(1 + \frac{j}{L}\right)^{-a} \qquad (j \in \mathbb{N}_0), \tag{6.1}$$

where $C_{a,L}$ is a normalisation constant ensuring that (6.1) defines a proper probability distribution,

$$C_{a,L}^{-1} = \sum_{j=0}^{\infty} \left(1 + \frac{j}{L}\right)^{-a}. \tag{6.2}$$

Observe that, for smaller values of j, the GPL formula (6.1) simplifies to

$$f_j \sim C_{a,L} \left(1 - \frac{aj}{L}\right) \qquad (j \ll L), \tag{6.3}$$

using the asymptotic approximation $(1 + x)^{-a} - 1 \sim -ax$ for $x \to 0$. On the other hand, for large j formula (6.1) is reduced to a power-law dependence,

$$f_j = C_{a,L} \left(\frac{j}{L}\right)^{-a} \left(1 + \frac{L}{j}\right)^{-a} \sim \frac{C_{a,L} L^a}{j^a} \qquad (j \gg L),. \tag{6.4}$$

Thus, the GPL model (6.1) may be viewed as providing an effective sewing of the formerly truncated lower values with the power-law tail.

Assuming that $L \gg 1$ and replacing the sum in (6.2) (rearranged as a Riemann integral sum) with the corresponding integral, we can write

$$\begin{aligned} 1 = \sum_{j\ge 0} f_j &= C_{a,L}\, L \sum_{j\ge 0} \frac{1}{(1 + j/L)^a} \cdot \frac{1}{L} \\ &\sim C_{a,L}\, L \int_0^\infty \frac{\mathrm{d}x}{(1 + x)^a} = \frac{C_{a,L}\, L}{a - 1}. \end{aligned} \tag{6.5}$$

Hence, the normalisation constant $C_{a,L}$ is asymptotically evaluated as

$$C_{a,L} \sim \frac{a-1}{L}. \tag{6.6}$$

The expected value of the GPL model is given by

$$\mu = \sum_{j\geq 0} j\, f_j = C_{a,L} \sum_{j\geq 1} \frac{j}{(1+j/L)^a}. \tag{6.7}$$

Substituting equation (6.2) for $C_{a,L}$, we have

$$\mu = \frac{\sum_{j\geq 1} j\,(1+j/L)^{-a}}{\sum_{j\geq 0} (1+j/L)^{-a}}. \tag{6.8}$$

Remembering that $L \gg 1$, similarly to (6.5) we can obtain an asymptotic approximation by replacing a Riemann sum with the integral,

$$\begin{aligned} \mu &= C_{a,L}\, L^2 \sum_{j\geq 1} \frac{j/L}{(1+j/L)^a} \cdot \frac{1}{L} \\ &\sim C_{a,L}\, L^2 \int_0^\infty \frac{x\,\mathrm{d}x}{(1+x)^a} = \frac{C_{a,L}\, L^2}{(a-1)(a-2)}. \end{aligned} \tag{6.9}$$

Substituting (6.6), this simplifies to

$$\mu \sim \frac{L}{a-2}, \tag{6.10}$$

that is,

$$L \sim (a-2)\,\mu. \tag{6.11}$$

In particular, formulas (6.10) and (6.11) show that the expected value μ is large (for large scale parameter L).

6.1.2 Conceptual justification of the GPL model

To provide a meaningful motivation for the GPL formula (6.1), consider first the low-production regime, $j \ll L$. Substituting (6.6), we can rewrite the approximate expression (6.3) in an asymptotically equivalent form as

$$f_j \sim \frac{a-1}{a} \cdot \frac{a}{L} \left(1 - \frac{a}{L}\right)^j \qquad (0 \leq j \ll L). \tag{6.12}$$

Formula (6.12) has a clear interpretation: in the course of time a source (such as a newly published paper) produces items (citations) sequentially, independently of one another, with probability $p = 1 - a/L$ each (close to 1), until the string of items is terminated with probability $q = 1 - p = a/L$. Thus, the probability of producing j items is given by the geometric formula, $p^j q$. The pre-factor $(a-1)/a$ in (6.12) may be interpreted as the probability of a low-production outcome (i.e., with $j \ll L$).

Note that the geometric law in (6.12), even extended over the entire range $j \geq 0$, has the mean $p/q \sim L/a$. Compensated by the pre-factor $(a-1)/a$ it becomes

$$\frac{(a-1)\,L}{a^2} = \frac{(a-1)(a-2)}{a^2} \cdot \frac{L}{a-2} < \frac{L}{a-2} \sim \mu,$$

according to formula (6.10). That is to say, the mean number of items produced under the first regime is asymptotically smaller than the average number of items per source. This indicates that the low-production regime is not exhaustive; moreover, the long-tail regime of the power-law type provides a dominating contribution to the mean of items.

The mechanism of migration to the long-tail regime represented by the formula (6.4) is also quite clear: the independent generation of successive items with approximately constant probability of adding a new item, becomes more and more state-dependent, thus paving the way to the principle *"success breeds success"* (also known as *cumulative advantage*), which is commonly accepted to underpin the power-law behaviour (see Price (1976), Egghe (2005), Egghe & Rousseau (1995), Huber (2002)).

6.1.3 Mixed geometric distribution as an approximation of the GPL

Motivated by the approach of Sichel (1971, 1985), we propose a mixed distribution model for the item production, designed to reproduce the GPL distribution (6.1). Building on the observation in Section 6.1.2 about a geometric approximation (6.12), we assume a background geometric law for individual sources instead of a Poisson law used in Sichel (1971). Specifically, suppose that each of the M sources produces items according to a geometric law with individual parameters

p_i $(i = 1, \ldots, M)$, which are deemed to be independent sample values of a random variable p with a beta distribution, $p \sim \mathrm{Beta}(\alpha, \beta)$, that is, with density

$$g(p) = \frac{p^{\alpha-1}\,(1-p)^{\beta-1}}{\mathrm{B}(\alpha,\beta)} \qquad (0 < p < 1),$$

where $\mathrm{B}(\alpha,\beta)$ is the beta function (Olver *et al.*, 2010, §5.12),

$$\mathrm{B}(\alpha,\beta) = \int_0^1 p^{\alpha-1}\,(1-p)^{\beta-1}\,\mathrm{d}p.$$

The resulting mixed distribution is given by

$$\begin{aligned} f_j &= \int_0^1 p\,(1-p)^j\; g(p)\,\mathrm{d}p \\ &= \frac{1}{\mathrm{B}(\alpha,\beta)} \int_0^1 p\,(1-p)^j\, p^{\alpha-1}(1-p)^{\beta-1}\,\mathrm{d}p \\ &= \frac{\mathrm{B}(\alpha+1,\beta+j)}{\mathrm{B}(\alpha,\beta)} \qquad (j \in \mathbb{N}_0). \end{aligned} \tag{6.13}$$

The beta function can be conveniently expressed as (Olver *et al.*, 2010, 5.12.1)

$$\mathrm{B}(\alpha,\beta) = \frac{\Gamma(\alpha)\,\Gamma(\beta)}{\Gamma(\alpha+\beta)}, \tag{6.14}$$

where $\Gamma(z) = \int_0^\infty s^{z-1}\,\mathrm{e}^{-s}\,\mathrm{d}s$ is the gamma function. Using the recurrence relation $\Gamma(z+1) = \alpha\,\Gamma(z)$ (Olver *et al.*, 2010, 5.5.1), it is easy to check that the probabilities (f_j) satisfy the chain of recurrence relations

$$f_0 = \frac{\alpha}{\alpha+\beta}, \qquad f_{j+1} = \frac{\beta+j}{\alpha+\beta+j+1}\,f_j \quad (j \ge 0).$$

For example,

$$f_1 = \frac{\alpha\beta}{(\alpha+\beta)(\alpha+\beta+1)}.$$

To identify the asymptotic behaviour of f_j, we substitute (6.14) and use Stirling's asymptotic formula (Olver *et al.*, 2010, 5.11.3)

$$\Gamma(z) \sim \sqrt{2\pi}\, z^{z-1/2}\,\mathrm{e}^{-z} \qquad (z \to \infty),$$

which yields for large j

$$\begin{aligned} f_j &= \frac{\Gamma(\alpha+1)\,\Gamma(\beta+j)/\Gamma(\alpha+\beta+j+1)}{\Gamma(\alpha)\,\Gamma(\beta)/\Gamma(\alpha+\beta)} \\ &= \frac{\alpha\,\Gamma(\alpha+\beta)}{\Gamma(\beta)}\cdot\frac{\Gamma(\beta+j)}{\Gamma(\alpha+\beta+j+1)} \\ &\sim \frac{\alpha\,\Gamma(\alpha+\beta)}{\Gamma(\beta)}\cdot\frac{\sqrt{2\pi}\,(\beta+j)^{\beta+j-1/2}\,\mathrm{e}^{-\beta-j}}{\sqrt{2\pi}\,(\alpha+\beta+j+1)^{\alpha+\beta+j+1/2}\,\mathrm{e}^{-\alpha-\beta-j-1}} \\ &\sim \frac{\alpha\,\Gamma(\alpha+\beta)}{\Gamma(\beta)}\cdot\frac{\mathrm{e}^{\alpha+1}}{j^{\alpha+1}}\left(1-\frac{\alpha+1}{\alpha+\beta+j+1}\right)^{\alpha+\beta+j+1} \\ &\sim \frac{\alpha\,\Gamma(\alpha+\beta)}{\Gamma(\beta)\,j^{\alpha+1}}. \end{aligned} \tag{6.15}$$

Comparing with the long-tail behaviour of the GPL model (see (6.4)) we see that

$$\frac{C_{a,L}L^a}{j^a}\sim\frac{\alpha\,\Gamma(\alpha+\beta)}{\Gamma(\beta)\,j^{\alpha+1}}\qquad(j\to\infty),$$

and it follows that

$$a=\alpha+1. \tag{6.16}$$

Furthermore, recalling that $L\gg 1$ and $C_{a,L}\sim(a-1)/L$ (see (6.6)), we get

$$(a-1)\,L^{a-1}=\alpha L^\alpha\sim\frac{\alpha\,\Gamma(\alpha+\beta)}{\Gamma(\beta)}, \tag{6.17}$$

and it is clear that the parameter β must be large. Again using Stirling's formula, we obtain

$$\frac{\Gamma(\alpha+\beta)}{\Gamma(\beta)}\sim\frac{\sqrt{2\pi}\ (\alpha+\beta)^{\alpha+\beta-1/2}\,\mathrm{e}^{-\alpha-\beta}}{\sqrt{2\pi}\,\beta^{\beta-1/2}\,\mathrm{e}^{-\beta}}\sim\beta^\alpha. \tag{6.18}$$

Combining (6.17) and (6.18), we obtain $\alpha L^\alpha\sim\alpha\beta^\alpha$, so

$$L=\beta. \tag{6.19}$$

In the asymptotic calculations above, it was tacitly assumed that β is fixed. Repeating this analysis with $\beta\to\infty$ and $j\gg\beta$ (i.e., by formally combining (6.15) and (6.18)), it is straightforward to obtain the asymptotics

$$f_j\sim\frac{\alpha\,\beta^\alpha}{j^{\alpha+1}}\qquad(j\gg\beta), \tag{6.20}$$

which is consistent with the GPL asymptotics (6.4) due to the connection formulas (6.16) and (6.19).

The mixed geometric model is quite convenient analytically, because many formulas are exact rather than asymptotic. For instance, the normalising constant is expressed exactly in terms of the beta function. Similarly, the expected value of this distribution is easy to compute,

$$\mu = \sum_j j f_j = \frac{1}{\mathrm{B}(\alpha,\beta)} \sum_j j\,\mathrm{B}(\alpha+1,\beta+j).$$

Recalling the integral representation (6.13), we have

$$\begin{aligned}\sum_j j\,\mathrm{B}(\alpha+1,\beta+j) &= \int_0^1 \sum_j p\,(1-p)^j\,p^{\alpha-1}\,(1-p)^{\beta-1}\,\mathrm{d}p\\ &= \int_0^1 \frac{1-p}{p}\,p^{\alpha-1}\,(1-p)^{\beta-1}\,\mathrm{d}p\\ &= \mathrm{B}(\alpha-1,\beta+1).\end{aligned}$$

Using the formula (6.14), it is now easy to compute

$$\mu = \frac{\mathrm{B}(\alpha-1,\beta+1)}{\mathrm{B}(\alpha,\beta)} = \frac{\beta}{\alpha-1}.$$

In the original GPL model, this result is obtained only asymptotically (see (6.10)).

6.2 Limit Shape in the GPL Model

6.2.1 Convergence of random Young diagrams

In line with Chapter 3, let $\boldsymbol{X} = (X_i, i = 1,\dots,M)$ be an independent random sample of size M from the GPL distribution (see Definition 6.1), interpreted as the random item outputs produced by i-th source, respectively. Let (M_j) be the corresponding multiplicities of counts $j \in \mathbb{N}_0$ (see (3.6)). Recalling that $Y(x) = \sum_{j\ge x} M_j$ defines the upper boundary of the corresponding Young diagram (see (3.17)), consider a rescaled diagram with scaling coefficients

$$A = L, \qquad B = M, \tag{6.21}$$

that is,

$$\widetilde{Y}(x) = \frac{1}{B}\sum_{j\ge Ax} M_j = \frac{1}{M}\sum_{j\ge Lx} M_j \qquad (x \ge 0).$$

For $a > 1$, define the function

$$\varphi_a(x) := \frac{1}{(1+x)^{a-1}} \qquad (x \ge 0). \tag{6.22}$$

Our first result is the convergence to $\varphi_a(x)$ of the expected (rescaled) Young diagrams, $\mathsf{E}\big(\widetilde{Y}(x)\big)$.

Theorem 6.1. *Assuming that $L \to \infty$, we have*

$$\mathsf{E}\big(\widetilde{Y}(x)\big) \to \varphi_a(x) \quad (x \ge 0). \tag{6.23}$$

Proof. The general formula (3.31) specialises to

$$\mathsf{E}\big(\tilde{Y}(x)\big) = \frac{M\bar{F}(Ax)}{B} = \bar{F}(Lx) = C_{a,L}\sum_{j\ge Lx}\frac{1}{(1+j/L)^a}. \tag{6.24}$$

Again approximating the sum in (6.24) by an integral, we obtain, for any $x \ge 0$,

$$\begin{aligned}\sum_{j\ge Lx}\frac{1}{(1+j/L)^a} &= L\sum_{j\ge Lx}\frac{1}{(1+j/L)^a}\cdot\frac{1}{L}\\ &\sim L\int_x^\infty \frac{\mathrm{d}s}{(1+s)^a} = \frac{L}{(a-1)\,(1+x)^{a-1}} = \frac{L\,\varphi_a(x)}{a-1}.\end{aligned}$$

Returning to (6.24) and substituting (6.6), we finally get

$$\mathsf{E}\big(\tilde{Y}(x)\big) \sim \frac{a-1}{L}\cdot\frac{L\,\varphi_a(x)}{a-1} = \varphi_a(x), \tag{6.25}$$

as claimed. □

Remark 6.1. For convergence in (6.25) we only need that $L \to \infty$ — no condition on growth of $B = M$ is required. But for convergence of random functions $\widetilde{Y}(x)$ the condition $M \to \infty$ is essential (see Theorem 6.2).

We can now obtain our main result about convergence of random diagrams $\widetilde{Y}(x)$ to the limit shape $\varphi_a(x)$.

Theorem 6.2. *Assuming that $L \to \infty$ and $M \to \infty$, for each $x \geq 0$ the mean squared deviation of $\widetilde{Y}(x)$ from the limit shape $\varphi_a(x)$ is asymptotically small,*

$$\mathsf{E}\big(|\widetilde{Y}(x) - \varphi_a(x)|^2\big) \to 0. \tag{6.26}$$

This implies convergence in probability, $\widetilde{Y}(x) \xrightarrow{\text{p}} \varphi_\nu(x)$, that is, for each $x > 0$ and any $\varepsilon > 0$,

$$\mathsf{P}\big(\big|\widetilde{Y}(x) - \varphi_a(x)\big| \geq \varepsilon\big) \to 0. \tag{6.27}$$

Proof. By a standard decomposition of the mean squared deviation, we have

$$\mathsf{E}\big(|\widetilde{Y}(x) - \varphi_a(x)|^2\big) = \mathsf{Var}\big(\widetilde{Y}(x)\big) + \big(\mathsf{E}\big(\widetilde{Y}(x)\big) - \varphi_a(x)\big)^2. \tag{6.28}$$

Using formulas (3.23) and (3.25) (with $A = L$ and $B = M$), the variance term in (6.28) is estimated as follows,

$$\begin{aligned}\mathsf{Var}\big(\widetilde{Y}(x)\big) = \frac{M\bar{F}(Ax)F(Ax)}{B^2} &= \frac{\bar{F}(Lx)F(Lx)}{M}\\ &\leq \frac{\bar{F}(Lx)}{M} \sim \frac{\varphi_a(x)}{M} \to 0,\end{aligned} \tag{6.29}$$

according to (6.23) and (6.24). As for the second term on the right-hand side of (6.28), it is asymptotically small due to (6.23). Hence, the limit (6.26) follows.

Finally, convergence in probability (6.27) is a standard consequence of (6.26) due to Chebyshev's inequality (Shiryaev, 1996, Sec. II.6, p. 192),

$$\mathsf{P}\big(\big|\widetilde{Y}(x) - \varphi_a(x)\big| \geq \varepsilon\big) \leq \frac{\mathsf{E}\big(|\widetilde{Y}(x) - \varphi_a(x)|^2\big)}{\varepsilon^2} \to 0,$$

according to (6.26). □

6.2.2 Graphical illustration using computer simulation

As an example, the limit shape of the GPL model using simulation is shown in this section, with parameters $a = 2.5, L = 20$ and $M = 1{,}000$. Figure 6.1 shows the simulated data as the upper boundary of the Young diagram $Y(x)$ which is defined in (3.17) in black; and the theoretical complementary cumulative distribution function

$$\bar{F}(x) = \frac{B\varphi(\frac{x}{A})}{M} = (1 + \frac{x}{L}^{1-a}) \tag{6.30}$$

$\bar{F}(x)$ (see (3.21) and (6.1)) and the limit shape scaled back to the original frequencies of counts, that is $x \mapsto M\,\varphi_a(x/L)$ are shown in blue and red, respectively. The plots show a good fit of the limit shape in the bulk of the simulated data.

The tail behaviour of the limit shape of the GPL is inspected, using (6.25), for $x \to \infty$,

$$y = Y(Lx) \approx M\,\varphi_a(x) \approx M\,(1+x)^{1-a}\,, \tag{6.31}$$

and furthermore, by taking the logarithm,

$$\log Y(Lx) = \log M - (a-1)\log(x+1) \approx \log M - (a-1)\log x. \tag{6.32}$$

That is to say, the graph of the limit shape of the GPL model is approximately a straight line for large x in logarithmically transformed coordinates $u = \log z$, $v = \log y$. The tail of the CCDF of the GPL model is also a straight line since (3.31).

Alternatively, this can be explained by recalling that the GPL is approximately a power law for large j (see (6.4)), and the tail of the CCDF of the power law in logarithmically transformed coordinates is approximately a straight line.

6.2.3 Fluctuations of random Young diagrams

Recalling that $\widetilde{Y}(x)$ is a (normalised) sum of independent indicators $Z_i(Ax) = I_{\{X_i \ge Ax\}}$, $i = 1, \dots, M$ (see (5.23)), it is natural to expect that $\widetilde{Y}(x)$ is asymptotically normal, with mean $\mathsf{E}\big(\widetilde{Y}(x)\big) = M\bar{F}(Ax)/B \sim \varphi_\nu(x)$ and variance $M\bar{F}(Ax)F(Ax)/B^2 \sim \varphi_\nu(x)/B$ (see (6.25) and (6.29)). However, a standard central limit theorem is not directly applicable because the "success" probability $\mathsf{P}(Z_i(Ax) = 1) = \bar{F}(Ax)$ is not constant (and, moreover, it tends to 0), so we have to re-prove this statement using the method of characteristic functions.

Theorem 6.3. *Assuming that $L \to \infty$ and $M \to \infty$, for any $x > 0$ we have*

$$\varUpsilon(x) := \sqrt{\frac{M}{\varphi_a(x)}}\,\Big(\widetilde{Y}(x) - \bar{F}(Lx)\Big) \xrightarrow{\mathrm{d}} \mathcal{N}(0,1), \tag{6.33}$$

where $\mathcal{N}(0,1)$ is a standard normal law (i.e., with zero mean and unit variance), and $\xrightarrow{\mathrm{d}}$ denotes convergence in distribution.

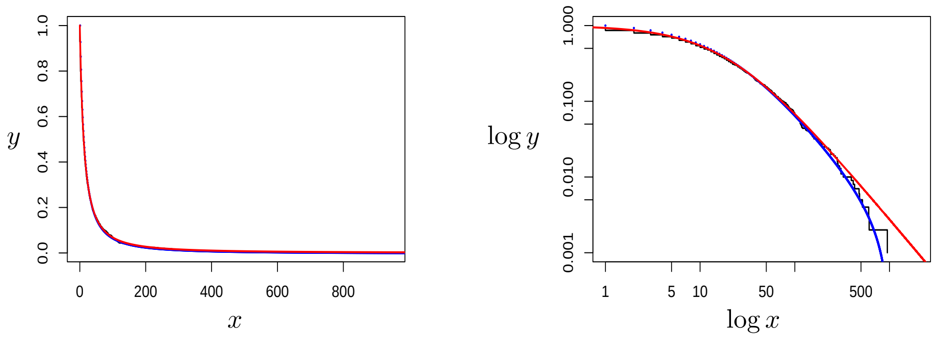


Figure 6.1: Illustration of the limit shape approximation using $M = 1{,}000$ random values (X_i) simulated using the GPL model (6.1) with parameters $a = 2.5$ and $L = 20$. In the left panel, the black stepwise plot represents the upper boundary $Y(x)$ of the corresponding Young diagrams, together with the GPL complementary distribution function $\bar{F}(x)$ shown in blue dots, while the smooth red curves represent the scaled back limit shape, $x \mapsto B\,\varphi_a(x/A)$. In the right panel, the tail is shown in logarithmic transformed coordinates, with the same line and colour coding.

Proof. Substituting (3.25), the left-hand side of (6.33) is rewritten as

$$\varUpsilon(x) = \frac{1}{\sqrt{M\,\varphi_a(x)}} \sum_{i=1}^{M} \bigl(Z_i(Lx) - \bar{F}(Lx)\bigr). \tag{6.34}$$

The characteristic function of (6.34) is given by

$$\psi(t;x) := \mathsf{E}\bigl(\mathrm{e}^{\mathrm{i}t\varUpsilon(x)}\bigr) = \mathrm{e}^{-\mathrm{i}\tilde{t}M\bar{F}(Lx)} \Bigl(1 + \bar{F}(Lx)\bigl(\mathrm{e}^{\mathrm{i}\tilde{t}} - 1\bigr)\Bigr)^{M}, \tag{6.35}$$

where

$$\tilde{t} = \frac{t}{\sqrt{M\,\varphi_a(x)}}, \qquad t \in \mathbb{R}. \tag{6.36}$$

Choosing the principal branch of the logarithm function $\mathbb{C}\setminus\{0\} \ni z \mapsto \log z \in \mathbb{C}$ (i.e., such that $\log 1 = 0$), we can rewrite (6.35) as

$$\log\psi(t;x) = -\mathrm{i}\tilde{t}M\bar{F}(Lx) + M\log\,(1 + w), \tag{6.37}$$

where

$$w := \bar{F}(Lx)\bigl(\mathrm{e}^{\mathrm{i}\tilde{t}} - 1\bigr). \tag{6.38}$$

Since $L \to \infty$, we have $\tilde{t} \to 0$ and $w \to 0$, hence

$$\log(1+w) = w - \tfrac{1}{2}w^2 + O(|w|^3).$$

Therefore, Taylor expanding $\mathrm{e}^{\mathrm{i}\tilde{t}} = 1 + \mathrm{i}\tilde{t} - \frac{1}{2}\tilde{t}^2 + O(\tilde{t}^3)$ and substituting (6.36) and (6.38), formula (6.37) is elaborated as follows,

$$\log\psi(t;x) = -\frac{\bar{F}(Lx)\,F(Lx)\,t^2}{2\,\varphi_a(x)} + O\biggl(\frac{\bar{F}(Lx)}{M^{1/2}}\biggr) \to -\frac{t^2}{2},$$

using that $\bar{F}(Lx) \sim \varphi_a(x)$ (see (6.29)) and $F(Lx) \to 1$ for any $x > 0$. Thus, $\psi(t;x) \to \mathrm{e}^{-t^2/2}$, which is the characteristic function of the normal distribution $\mathcal{N}(0,1)$, as claimed. □

6.3 Fitting the GPL Model

6.3.1 Graphical methods

A simple graphical approach to estimation of the parameter a is via a log-log transform of the data in the upper tail of the frequency range, taking advantage of a power-law approximation (6.4), and using equation (4.8). After that, the parameter L can be approximately recovered using the relation (6.31). Despite a considerable waste of data, these crude estimates may be helpful as meaningful seeds in the iterative numerical procedures, such as ordinary least squares or the maximum likelihood estimation.

6.3.2 Ordinary least squares

Another method is the ordinary least squares (OLS). Here, the square-distance between the empirical distribution $(\hat{f}_j)$ (with $\hat{f}_j = M_j/M$) from the GPL distribution (f_j) (with trial parameter values a and L) is given by

$$g(a,L;\boldsymbol{X}) := \sum_{j\ge 0}\bigl(\hat{f}_j - f_j\bigr)^2 = \sum_{j\ge 0}\left(\frac{M_j}{M} - \frac{C_{a,L}}{(1+j/L)^a}\right)^2. \tag{6.39}$$

To minimise $g(a, L)$, we let the partial derivatives equal to zero,

$$\frac{\partial g}{\partial a} = 2\sum_{j\geq 0}\left(\frac{M_j}{M} - \frac{C_{a,L}}{(1+j/L)^a}\right)\frac{C_{a,L}\log\left(1+j/L\right) - \frac{\partial C_{a,L}}{\partial a}}{(1+j/L)^a} = 0, \tag{6.40}$$

$$\frac{\partial g}{\partial L} = -2\sum_{j\geq 0}\left(\frac{M_j}{M} - \frac{C_{a,L}}{(1+j/L)^a}\right)\frac{C_{a,L}\left(aj/L^2\right) + \frac{\partial C_{a,L}}{\partial L}\left(1+j/L\right)}{(1+j/L)^{a+1}} = 0. \tag{6.41}$$

The partial derivatives of $C_{a,L}$ can be obtained from (6.2),

$$\frac{\partial C_{a,L}}{\partial a} = C_{a,L}^2\sum_{j\geq 0}\frac{\log(1+j/L)}{(1+j/L)^a}, \tag{6.42}$$

$$\frac{\partial C_{a,L}}{\partial L} = -\frac{a\,C_{a,L}^2}{L}\sum_{j\geq 0}\frac{j/L}{(1+j/L)^{a+1}}. \tag{6.43}$$

Substituting expressions (6.42) and (6.43) (together with (6.2)) into (6.40) and (6.41), the latter equations can be solved (e.g., numerically) to yield the OLS estimates for parameters a and L. Alternatively, the function $g(a, L)$ can be minimised directly, for instance, using the R function `optim`.

6.3.3 Maximum likelihood

According to (3.13), the likelihood of the GPL model is represented as

$$\mathcal{L}(a, L; \boldsymbol{X}) = \prod_{j\geq 0} f_j^{M_j} = C_{a,L}^M\prod_{j\geq 1}\left(1+\frac{j}{L}\right)^{-aM_j}, \tag{6.44}$$

with the log-likelihood

$$\ell(a, L; \boldsymbol{X}) = M\log C_{a,L} - a\sum_{j\geq 1} M_j\log\left(1+\frac{j}{L}\right). \tag{6.45}$$

To maximise (6.45), we set the partial derivatives equal to zero,

$$\frac{\partial \ell}{\partial a} = \frac{M}{C_{a,L}}\cdot\frac{\partial C_{a,L}}{\partial a} - \sum_{j\geq 1} M_j\log\left(1+\frac{j}{L}\right) = 0, \tag{6.46}$$

$$\frac{\partial \ell}{\partial L} = \frac{M}{C_{a,L}}\cdot\frac{\partial C_{a,L}}{\partial L} + \frac{a}{L}\sum_{j\geq 1}\frac{j/L}{1+j/L} = 0, \tag{6.47}$$

where the partial derivatives of $C_{a,L}$ are given in (6.42) and (6.43). By solving these equations numerically, we obtain the MLEs $\hat{a}$ and $\hat{L}$.

A simple asymptotic analysis of the likelihood equations (6.46) and (6.47) may be helpful. Namely, taking advantage of the approximate relation (6.6), the log-likelihood is written as

$$\ell(a, L; \boldsymbol{X}) \approx M \log \frac{a-1}{L} - a \sum_{j \geq 1} M_j \log \left(1 + \frac{j}{L}\right),$$

Then equations (6.46) and (6.47) simplify to

$$\frac{\partial \ell}{\partial a} \approx \frac{M}{a-1} - \sum_{j \geq 1} M_j \log \left(1 + \frac{j}{L}\right) = 0, \tag{6.48}$$

$$\frac{\partial \ell}{\partial L} \approx -\frac{M}{L} + \frac{a}{L} \sum_{j \geq 1} M_j \frac{j/L}{1 + j/L} = 0. \tag{6.49}$$

Hence, we obtain the approximate maximum likelihood equations,

$$\sum_{j \geq 1} \frac{M_j}{M} \log \left(1 + \frac{j}{L}\right) = \frac{1}{a-1}, \tag{6.50}$$

$$\sum_{j \geq 1} \left(\frac{M_j}{M} \cdot \frac{j/L}{1 + j/L}\right) = \frac{1}{a}. \tag{6.51}$$

Eliminating parameter a gives a closed equation on L,

$$\left(\sum_{j \geq 1} \left(\frac{M_j}{M} \cdot \frac{j/L}{1 + j/L}\right)\right)^{-1} - \left(\sum_{j \geq 1} \frac{M_j}{M} \log \left(1 + \frac{j}{L}\right)\right)^{-1} = 1. \tag{6.52}$$

This equation can be solved numerically to yield an estimate $\hat{L}$, and then from (6.51) we get

$$\hat{a} = \left(\sum_{j \geq 1} \left(\frac{M_j}{M} \cdot \frac{j/\hat{L}}{1 + j/\hat{L}}\right)\right)^{-1}. \tag{6.53}$$

6.3.4 Method of moments

A more precise way of representing the expected value $\mu = \mathsf{E}(X)$ without replacing sums with integrals, compared to equation (6.9), is given by

$$\mu = \frac{\sum_{j \geq 1} j \, (1 + j/L)^{-a}}{\sum_{j \geq 0} (1 + j/L)^{-a}} \tag{6.54}$$

This relation can be viewed as a first-order moment equation, which can be used for parameter estimation, whereby the theoretical mean μ is replaced by its sample mean value, $\hat{\mu} = N/M$,

$$N = M\,\frac{\sum_{j\geq 1} j\,(1+j/L)^{-a}}{\sum_{j\geq 0}(1+j/L)^{-a}}.$$

Using (6.10), a simplified (asymptotic) version of this equation reads

$$N = \frac{ML}{a-2}. \tag{6.55}$$

Since we have two parameters, a and L, we need another equation to close the system of equations. We cannot use the second-order moments, since they may not exist (if $a \leq 3$). Instead, we can use a different statistic related to the occupation problem we are considering. One such statistic is the number of occupied boxes — in our case, the number of distinct output values j produced by the M sources,

$$W = \#\{M_j > 0\} = \sum_{j\geq 0} I_{\{M_j>0\}}. \tag{6.56}$$

It is easy to see (Karlin, 1967, p. 381) that the expectation of W is given by

$$\mathsf{E}(W) = \sum_{j\geq 0}\bigl(1-(1-f_j)^M\bigr).$$

Note that by the binomial theorem,

$$(1-f_j)^M = \sum_{k=0}^{M}\binom{M}{k}(-f_j)^k, \tag{6.57}$$

while by the Taylor expansion of exponential,

$$\mathrm{e}^{-Mf_j} = \sum_{k=0}^{\infty}\frac{(-Mf_j)^k}{k!}. \tag{6.58}$$

From (6.57) and (6.58), also using the approximation $\binom{M}{k} \approx \frac{M^k}{k!}$, we obtain

$$\sum_{j\geq 0}\bigl\{(1-f_j)^M - \mathrm{e}^{-Mf_j}\bigr\} \to 0 \qquad (M\to\infty).$$

Hence,

$$\mathsf{E}(W) = \sum_{j\geq 0}\bigl(1-\mathrm{e}^{-Mf_j}\bigr) + o(1).$$

Thus, replacing this expectation with the sample value of W (see the definition (6.56)), the required second equation for estimation of the model parameters takes the form

$$W = \sum_{j\geq 0}\left(1-\mathrm{e}^{-Mf_j}\right).$$

We can simplify this problem even further by using an asymptotic expression for $\mathsf{E}(W)$ (Karlin, 1967, Example 4, p. 378, and Theorem 1′, p. 381) and also recalling (6.6),

$$\mathsf{E}(W) \sim \Gamma\bigl(1-\tfrac{1}{a}\bigr)\,(a-1)^{1/a}\; M^{1/a}L^{1-1/a}.$$

So for the estimation purposes we can use the equation

$$W = \Gamma\bigl(1-\tfrac{1}{a}\bigr)\,(a-1)^{1/a} M^{1/a}L^{1-1/a}. \tag{6.59}$$

For instance, expressing L from (6.55), we obtain a closed equation on a,

$$W = \Gamma\bigl(1-\tfrac{1}{a}\bigr)\,(a-1)^{1/a}\,(a-2)^{1-1/a}N^{1-1/a}M^{-1+2/a}$$

which can be solved numerically.

6.4 Real Data Examples

In the following examples, we illustrate the fitting result of the GPL model to real data sets introduced in Section 2.2.

6.4.1 Lotka's data

We fit the GPL model to Lotka's data set (see Section 2.2, A), the estimated parameters of the GPL model $\hat{a} \doteq 2.511331$ and $\hat{L} \doteq 2.280752$ are obtained using the OLS with the aid of `optim` function in R. The fitted result is depicted in Figure 6.2.

The GPL model performs reasonably well at the beginning of the data. For $j \gg L$, the GPL is approximated by a power law (see (6.4)), so for large j the

GPL exhibits a power tail. However, Lotka's data starts the power law behaviour from the beginning of the data, so the estimated parameter $\hat{L} = 2.280752$ is relatively small. The fitting result of the GPL to Lotka's data in Figure 6.2 did not show a significant improvement compared to the result obtained by fitting the power-law model to the data, as shown in Figure 4.2.

Lotka's data does not have a power law tail from observing the right panel of Figure 6.2. Compared to the GIGP model, which has a power-geometric tail that fits to Lotka's data better (see the first row of Figure 5.3).

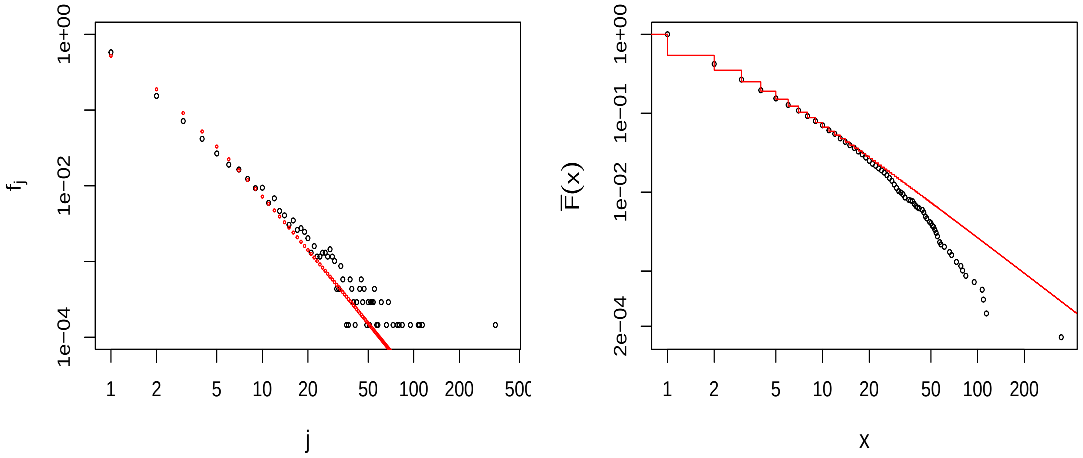


Figure 6.2: The GPL model (red) fitted to Lotka's data (black), with the estimated parameters $\hat{a} \doteq 2.511331$ and $\hat{L} \doteq 2.280752$. The plots depict frequencies (left panel) and complementary cumulative frequencies (right panel), shown in log-log coordinates.

6.4.2 Chen's data

The GPL is fitted to Chen's data (Section 2.2, C) using two different estimation methods, ordinary least squares (OLS) and maximum likelihood estimation (MLE). The top row of Figure 6.3 illustrates the fitting result of the GPL using OLS to estimate the parameters, with the resulting estimates of $\hat{a} \doteq 1.841551$ and $\hat{L} \doteq 6.845811$. The OLS method captures reasonably well the bulk of the data within the range $1 \leq j \leq 70$. The bottom row of Figure 6.3 displays the fitted GPL obtained using MLE, with estimated parameters $\hat{a} \doteq 2.232609$ $\hat{L} \doteq 12.967$.

The MLE method provides a better fit for the tail of the data than OLS. However, there are visible deviations in the middle part of the data, also observed in fitting using the GIGP model (see Figure 5.3 (b)).

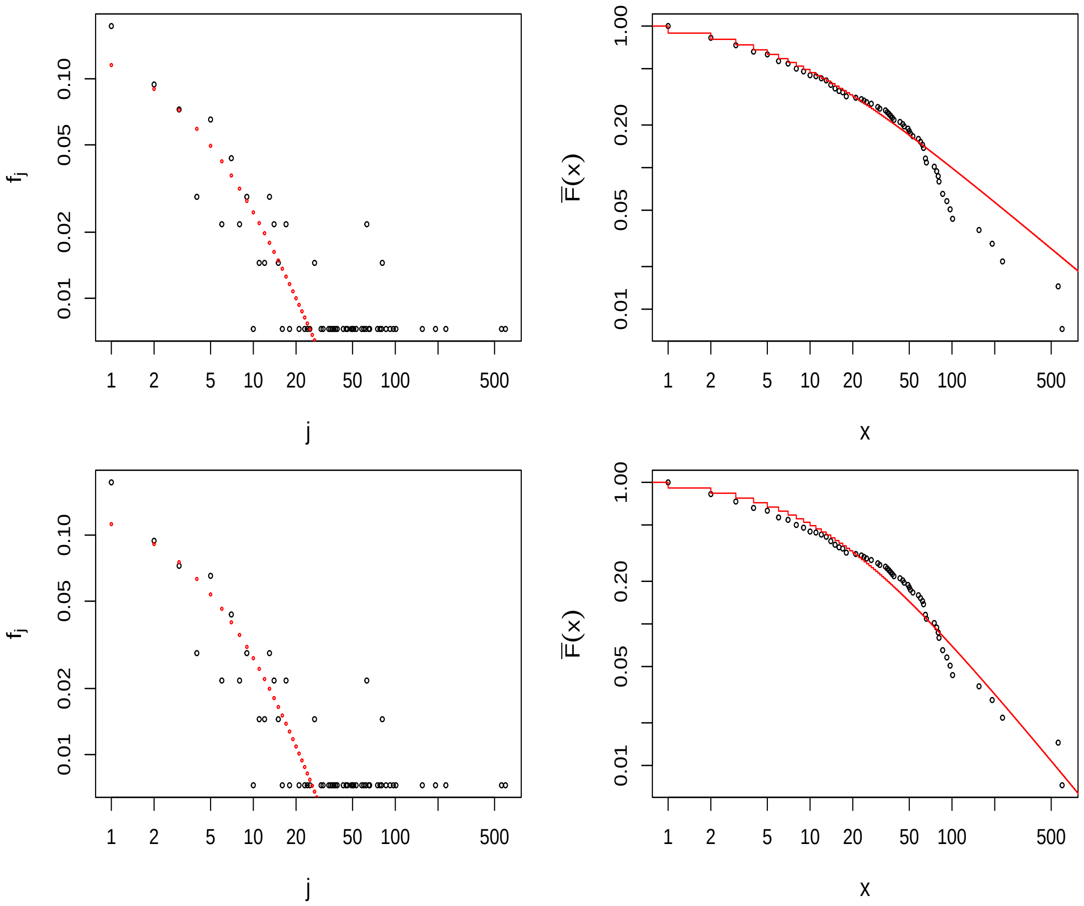


Figure 6.3: The GPL model (red) fitted to Chen's data (black). The top row depicts the GPL with estimated parameters $\hat{a} \doteq 1.841551$ and $\hat{L} \doteq 6.845811$ using OLS. The bottom row depicts the GPL with estimated parameters $\hat{a} \doteq 2.232609$; $\hat{L} \doteq 12.967$ using MLE. The plots depict frequencies (left panel) and complementary cumulative frequencies (right panel), shown in log-log coordinates.

6.4.3 EJP data

The EJP dataset (see Section 2.2, D) comprises $K = 113$ authors, $M = 15{,}400$ papers, and $N = 245{,}567$ citations, including $M_0 = 6{,}472$ papers with zero citations.

Noting that the observed frequency of zero-count papers $\hat{f}_0 = M_0/M = 0.42$ is relatively high, we compared the fitting results with and without zero counts, which suggested that the value at $j = 0$ is worth excluding. For a clearer visualisation, Figure 6.4 (left) includes zero counts, making it evident that zero-count citations appear to be an outlier.

Using the OLS method, the estimated parameters are $\hat{a} \doteq 2.175316$ and $\hat{L} \doteq 16.35075$. Figure 6.4 shows that the fit is remarkably accurate, especially over a large initial part of the citation spectrum (up to around $j = 500$).

Compared with fitting the truncated power law model (Figure 4.4) and integer partition model (Figure 4.5), the GPL model covered the whole range of the EJP data.

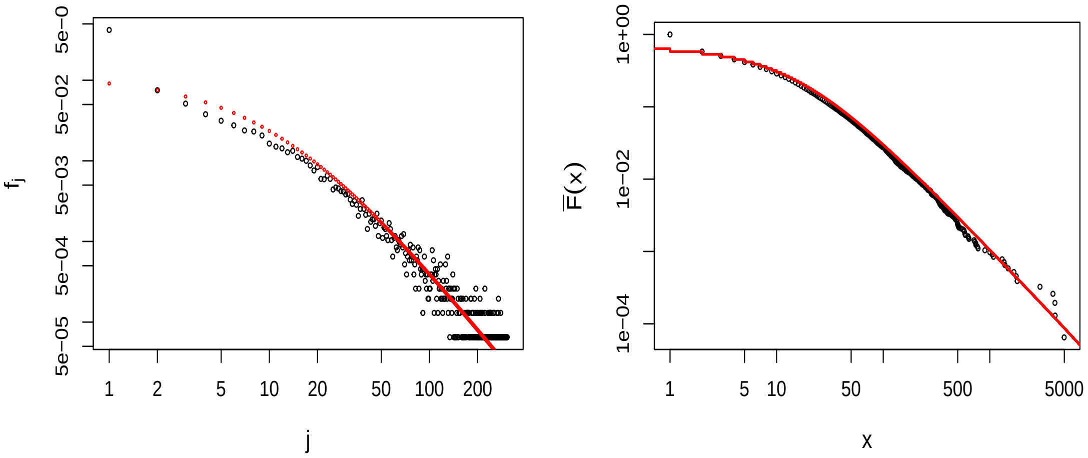


Figure 6.4: The GPL model (red) fitted to the EJP data (black), with the estimated parameters $\hat{a} \doteq 2.175316$ and $\hat{L} \doteq 16.35075$. The plots depict frequencies (left panel) and complementary cumulative frequencies (right panel), shown in log-log coordinates.

Substitute (6.22) and (6.21) into (3.70), we obtain

$$\frac{1}{(1+h)^{a-1}} = \frac{Lh}{M} \tag{6.60}$$

There are $K = 113$ authors in the EJP data, according to remark 3.9, B is replaced by B/K, so (6.60) is written as

$$\frac{1}{(1+h)^{a-1}} = \frac{Lh}{M/K}, \tag{6.61}$$

substituting $\hat{a} \doteq 2.175316$, $\hat{L} \doteq 16.35075$, $M = 15{,}400$ and $K = 113$, the estimation of the h-index in the EJP data is given by $\hat{h} \doteq 22.754$. Compared to the estimated h-index from the integer partition model (25.19399), the estimation from the GPL model is the closer to the real average h-index 17.52.

6.4.4 AMS data

The AMS data (see Section 2.2, E) comprises $K = 3{,}089$ authors, $M = 316{,}361$ papers, and $N = 12{,}351{,}608$ citations, including $M_0 = 101{,}576$ papers with zero citations (i.e., 32.11%). Figure 6.5 depicts the results of fitting the GPL to the AMS data; using the OLS the estimates are $\hat{a} \doteq 2.225705$ and $\hat{L} \doteq 18.04318$. Similarly to the EJP data set, a big percentage of zero citations suggests that the count m_0 may be an outlier worth omitting from fitting. This has been confirmed by trying both fits, with and without M_0. Indeed, as shown in Figure 6.5 (left), the value $j = 0$ appears to be outlier off the GPL fit, with the rest of the data being fitted very well. The GPL fitted the AMS data accurately around $1 \leq x \leq 10{,}000$. However, as seen in the right panel of Figure 6.5, the tail of the AMS data exhibits a faster decay than the tail of the fitted GPL.

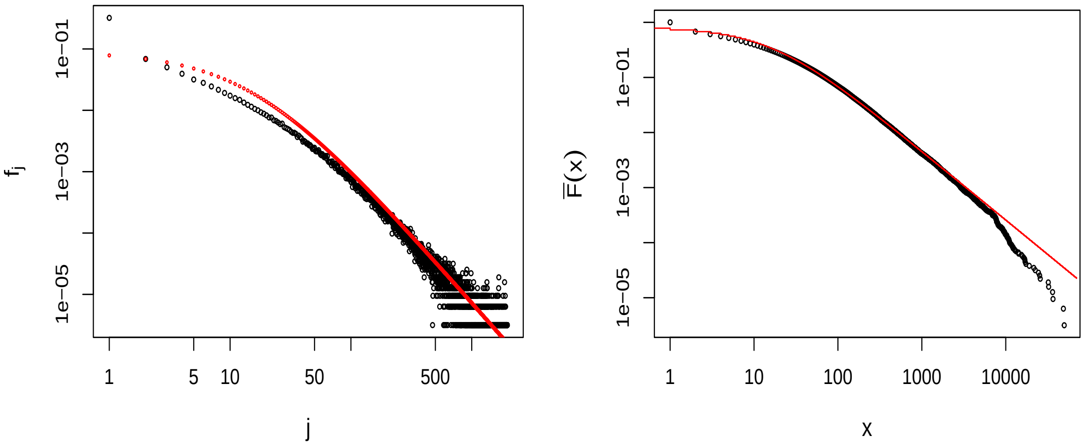


Figure 6.5: The GPL model (red) fitted to the AMS data (black), with the estimated parameters $\hat{a} \doteq 2.225705$ and $\hat{L} \doteq 18.04318$. The plots depict frequencies (left panel) and complementary cumulative frequencies (right panel), shown in log-log coordinates.

6.4.5 GPL fitting lessons

As a summary from the GPL fitting to various data sets reported above, some conclusions are as follows. For the data sets that manifest a relatively flat (linear) decay at a smaller range of count values, the GPL model works extremely well by capturing a transition to a power decay range and thus extending to a considerable bulk of the data, if not the entire data set (like with the EJP data). A remarkable additional benefit of the GPL model (e.g., evidenced in the EJP and AMS cases) is that it reveals the zero-inflated feature of the data, which can be meaningfully interpreted (e.g., as an excessive proportion of papers that acquire no citations).

In some cases, such as the Lotka data, no separate modelling if the initial range is needed, since the bulk of the data including the initial range is successfully modelled by the non-truncated power law. On the other hand, the GPL cannot model non-power tails for larger values of j (e.g., for the AMS data set), which suggests that different models should be used, such as the GIGP model (with a power-geometric decay) or a stretched exponential model advocated by Laherrere & Sornette (1998). It should be mentioned though that the GPL model is a lot more straightforward to fit by having a much simpler analytic expression with just two parameters, whereas the GIGP model is expressed via the non-elementary Bessel function, and moreover, some of its three parameters often need to be predefined *a priori* (Sichel, 1985).

However, instead of treating each range of the data separately by a specially selected model each, we argue using a successful example of the "sewing" GPL that a synthetic model combining the GPL with the stretched exponential model via a functional form resembling (6.1), would be the most flexible fitting tool, allowing to capture both transitions, from the initial "flat" domain to the power law and then to the stretched exponential one. It would be interesting to elaborate this idea in our future research.

In conclusion of this section, let us summarise the new contributions due to the GPL model developed in the present thesis as compared to the previous applications of the hooked power law. The latter was initially proposed for fitting to the web links data by Pennock *et al.* (2002); it was also useful in modelling zero-inflated citation data (Shahmandi *et al.*, 2020; Thelwall & Wilson, 2014).

In our work, we have found the limit shape of the GPL model (Section 6.2.1) and proved asymptotic normality of fluctuations around it (Section 6.2.3); these results play an important part in the model diagnostics and fitting to the data. Besides, we have provided a conceptual motivation of the GPL model as a mixed geometric distribution with a beta mixing density (Section 6.1.3). Incidentally, the GPL model was instrumental in the discovery of a zero-count "outlier" in the EJP and AMS data sets (Sections 6.4.3 and 6.4.4).

Chapter 7

Modelling Temporal Dynamics of Citations

This chapter begins with an exploratory data analysis of temporal citation data and applies survival analysis to analyse the time of the first citation of papers after their publication. The chapter then proceeds to study the time evolution of citations using the Hawkes point process.

7.1 Exploratory Data Analysis

The time evolution of citations has always been a popular topic in scientometric research (see Egghe & Rao, 2001; Egghe *et al.*, 1992; Wang *et al.*, 2013, etc.). Note that in our analysis of the GIGP model in Chapter 5, we confined ourselves to the case where the GIGP parameters are fixed constants; however, in the original paper by Sichel (1985) an attempt was made to address the temporal aspect by allowing the parameters to be time-dependent. It would be interesting to pursue further research in this direction by looking into the possible time evolution of the GIGP scaling coefficients leading to the limit shape. In a related development, there has been research on growing Young diagrams (see Eriksson & Sjöstrand, 2012; Krapivsky, 2021). In principle, these results may be useful in describing the time evolution of citations, but we did not pursue this direction in the thesis, given a poor performance of the integer partition model evidenced by our experiments with the EJP citation data.

The present chapter utilises an extended EJP data set (see Section 2.2, D′). The publication year of each paper is designated as Year $A = 0$.

The `get_article_cite_history` command in the `scholar` package of R (Yu *et al.*, 2016) allows for the retrieval of dynamic citation data from Google Scholar, given that the publication ID is known. The corresponding ID can be obtained through the `get_publications` command within the same R package. Nevertheless, despite this option, we find that accessing the dynamic data from WoS is a more convenient and efficient method. WoS provides a user-friendly interface that allows for easy access to yearly citation records of papers. The citation data is displayed publicly and presented in numerical form upon clicking the citations record of each paper. In contrast, Google Scholar only presents the historical total citation records of authors in histograms, necessitating web scraping to obtain precise citation history data for individual publications.

Figure 7.1 depicts the accumulation of citations over time. The left panel displays the citations of 100 papers that have experienced growth in citations.The selection of 100 papers was made randomly from a data set of 3,951 papers using the R command `sample.int(3591,100)`; this was done to improve clarity, as it is more comprehensible than including all 3,951 papers. The plot reveals that the citations of different papers are distributed broadly. The right panel displays citations for one of these randomly selected papers. This plot reveals that a paper may take several years to become recognised and accumulate citations. To investigate this observation further, Figure 7.2 presents a histogram showing the distribution of the number of years before a paper receives its first citation.

The distribution of citations over time after publication is a topic of interest in our investigation. Figure 7.3 displays histograms of the total citations received by papers at one year, five years, ten years, fifteen years, twenty years, and twenty-five years after publication. Generally, these plots exhibit a similar right-skewed, long-tail curve shape, with the curve shifting to the right as the years progress. Moreover, the peak of these histograms is consistently not at 0. Note that we only consider the publication year, ignoring the publication month. This data limitation implies that even papers published in January and December of the same year are grouped. However, as time passes, the month of publication becomes less critical, thus rendering this limitation insignificant.

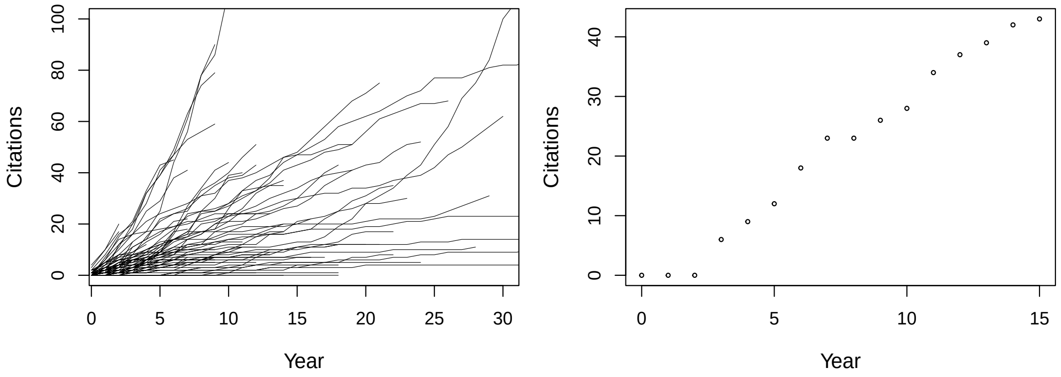


Figure 7.1: Plots illustrating accumulation of citations over time for a set of randomly chosen papers from extended EJP data, with the year axis defined relative to the year of publication (i.e., year 0 corresponds to the year of publication). The left panel displays the citation counts for 100 randomly selected papers, with each line representing an individual paper. The right panel shows the citation growth of a single randomly chosen paper over time since publication.

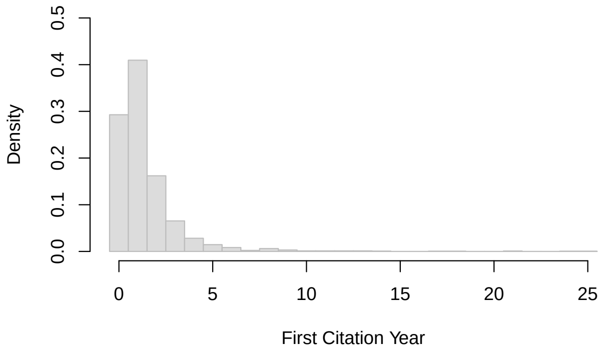


Figure 7.2: Histogram of the distribution of years before a paper receives its first citation after publication.

The number of papers for collecting citations varies due to differing publication years. Specifically, papers published one year prior to the data collection possess a citation record of only two years. However, to analyse citations a decade after publication, a minimum of eleven years of citation records are necessary. Consequently, the amount of data available for citation analysis after ten years is less than after one year. This explains the greater amount of noise present in the

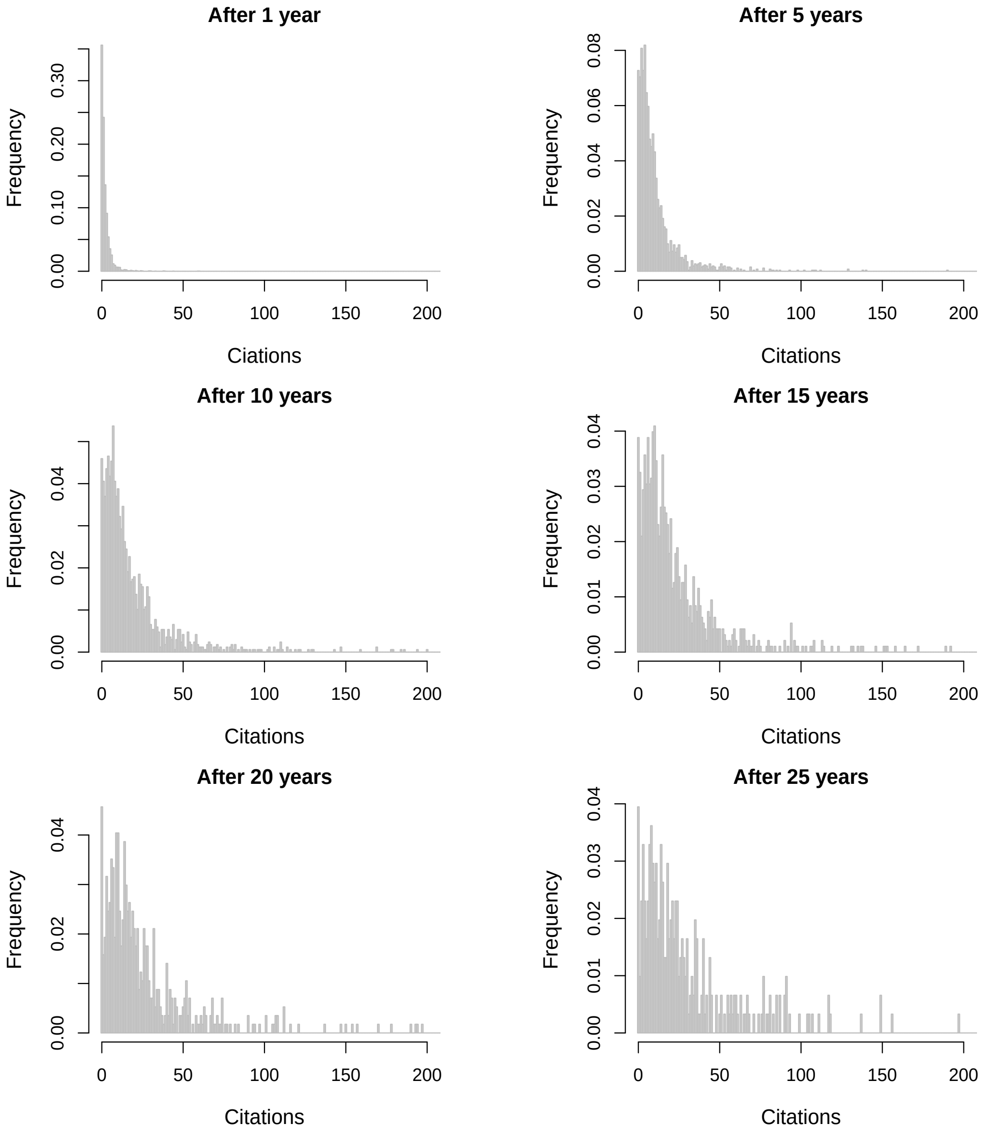


Figure 7.3: Histograms of citations after 1, 5, 10, 15, 20 and 25 years since publication. Note that the noise in the histograms increases in later panels due to the smaller sample sizes resulting from fewer papers having longer records of citations.

histograms displayed in the latter panel of Figure 7.3. Specifically, the numbers of papers available for citation analysis after one, five, ten, fifteen, twenty, and twenty-five years are 3,475, 2,613, 1,676, 953, 569, and 304, respectively.

The scatter plot presented in Figure 7.4 displays the relationship between

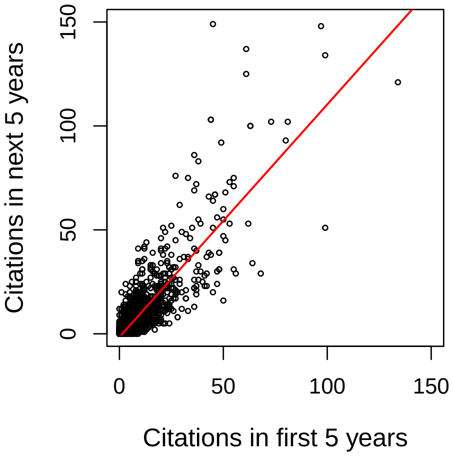


Figure 7.4: Scatter plot of citations received in the first 5 years since publication versus those in the next 5 years. The red line fitted using linear regression has a slope of 1.1189.

the citations accumulated in the first five years and in the next five years after publication. The plot consists of 1,676 papers published at least ten years ago and for which citation data is available for both the fifth and tenth year post-publication. By applying a linear regression on these data, we obtained the slope 1.1189, which indicates that in the following five years, these papers receive citations slightly more than in the first five years. The 95% confidence interval to regression coefficient of 1.1189 is $(1.0865, 1.1513)$. Although there is a general trend, in that papers receiving more citations in the first five years also receive more citations in the next five years, this relationship does not necessarily hold for all papers. It is important to note that citations received in the same year of publication have been excluded to ensure that both five-year periods have the same length of time.

Figure 7.5 provides further insights into the data set by displaying the mean and median of citations over time. As seen in these plots in Figure 7.5, there is a general upward trend in both the mean and median of citations with minor fluctuations at the beginning. However, towards the end of the time period, there are larger fluctuations in the data, which could be attributed to the reduced number of papers with available citation records for longer time periods. The

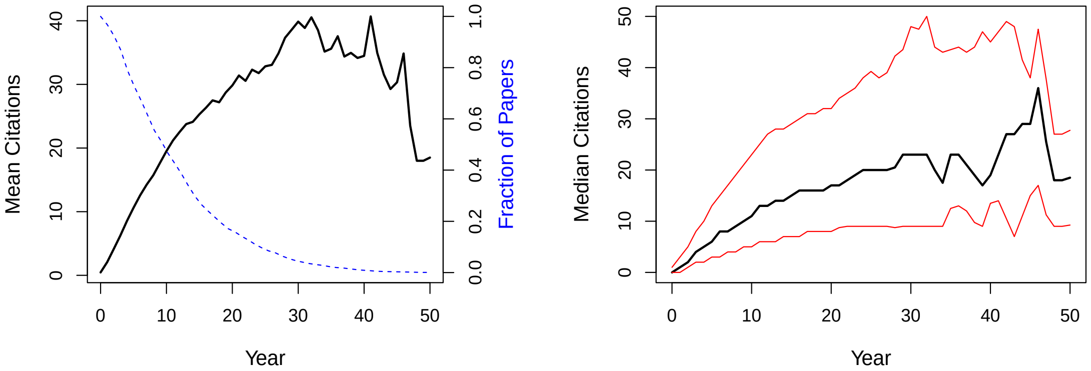


Figure 7.5: Statistics of citations received by papers published over the years. The left panel shows the mean of citations changing by years, represented by a black solid line, along with the fraction of the number of papers published corresponding years ago, denoted by a blue dashed line. The right panel illustrates the median changing by years, represented by a black solid line, along with the 25% and 75% percentiles represented by red lines.

fraction is one for year zero, as all papers have data for their year of publication.

As noted in the discussion of Figure 7.1, papers often require some time before being discovered and cited. Figure 7.2 examines the distribution of the time it takes for a paper to receive its first citation after publication. As shown in the histogram, the most common scenario is for papers to receive their first citation after one year of publication.

One interesting observation from the data set is that the paper by Fill (1988) took 25 years to receive its first (self-)citation in 2013, followed by only one further citation in 2018 (https://mathscinet-ams-org.eu1.proxy.openathens.net/mathscinet/search/publications.html?refcit=958208&loc=refcit). This observation highlights the fact that some papers may take a considerable amount of time to be discovered and cited, and that self-citations can also play a role in the citation patterns of a given paper.

A related phenomenon known in bibliometrics as the “sleeping beauty” is where a paper may not generate any citations for a very long time, but then suddenly it is noticed by the research community and then starts receiving citations (Ke *et al.*, 2015). One of the well-known examples is with a French mathematician Paul Painlevé who discovered in the 1900s (Painlevé, 1902) a classification of solutions to second-order nonlinear differential equations in the complex plane, the so-called *Painlevé transcendents*, but mathematicians and physicists only appreciated this work after more than 70 years (McCoy *et al.*, 1977).

Figure 7.6 provides a more detailed view of the citation record for a subset of papers. The left panel of the figure illustrates the citations received by 100 randomly selected papers each year, with each line representing a single paper. The red line depicts the mean of the annual citations received by these papers. The right panel of the figure displays the citations received by a single paper each year. Notably, Figure 7.6 presents the citations received by papers on an annual basis, in contrast to Figure 7.1 which shows the accumulated citations over time.

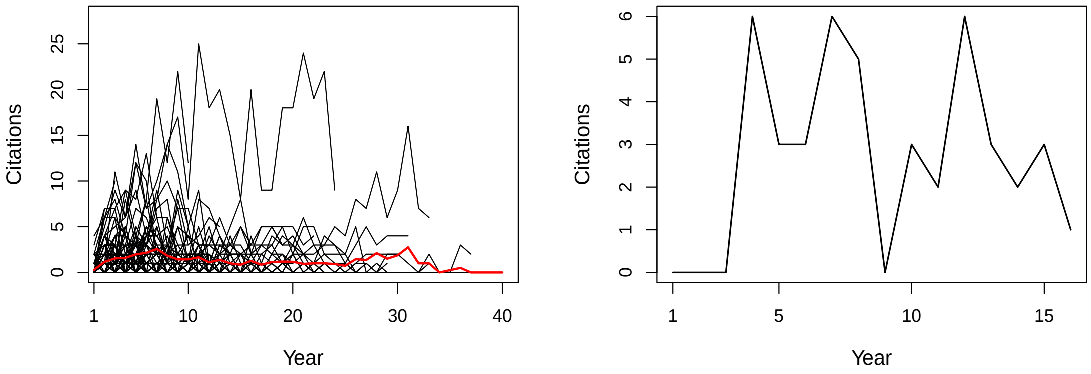


Figure 7.6: Annual citation records for a set of papers. The left panel displays the annual citation count for 100 randomly selected papers, with each line representing a single paper and the red line indicating the average number of annual citations. In the right panel, the annual citation record of a single randomly chosen paper is presented. It is worth noting that unlike other figures that show accumulated citations, this figure shows the citations received every year.

We looked for a typical period without citations to define the “death” of a paper. Figure 7.7 shows the distribution of gap years between two citations. Even though the largest gap is 25 years in this data set, we can define the end point after 25 years with zero citations; it is not necessarily true in other data sets.

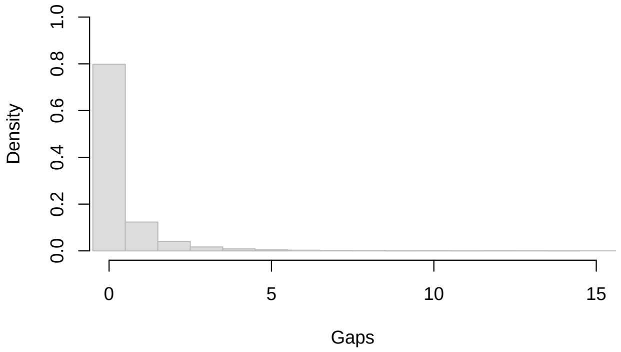


Figure 7.7: Histogram of years in between two citations. Note that this is not the same as the gaps between publications to the first citation shown in Figure 7.2.

7.2 Using Survival Analysis

7.2.1 Motivation

In this section, we use survival analysis to model the time until the first citation of a paper (interpreted as "death", or end event). Out of all the papers in our data set, 90.2% have been cited before 2022. However, for the remaining papers, the timing of their potential future citations is unknown, which means they are treated as censored. To address this issue, we conduct a study focusing on papers published between 1900 to 2022, and investigate their first citation. Papers that have yet to receive a first citation are treated as "surviving" and the first citation of a paper as interpreted as an "end event". Papers that had already received their first citation before 1900 are excluded from our study (as left-censored).

Survival analysis is a set of statistical techniques commonly used to model time-to-event data. In classic survival analysis, it is often encountered that some individuals under study do not experience the event of interest before the end of the observation, or withdraw from the study for unrelated reasons. As a result, the exact time of death is not known for such individuals, which leads to incompleteness of information. This is known as *censoring*. A common type of censoring in survival analysis is *right censoring*, which occurs when the observed time for an individual is less than their actual time to the event of interest (Collett, 2015).

There are seven different situations in our data that can result in censoring. Specifically, a paper is considered censored if the end event (i.e., the first citation) does not occur during the study period.

1. The paper was published and cited before 1900 but not since 1900 until 2022. Thus, the event of interest happened before the beginning of the study, which is an example of left censoring. This thesis has no such a case since the WoS only records data from 1900.

2. The paper was published before 1900 and cited after 1900 but before 2022. We treated these papers as if they were published in 1900, since the observation started in 1900.

3. The paper was published before 1900 and not cited until the end of the study in 2022. In this case, the paper is right-censored. Similarly as for Type 2, we assumed these papers were published in 1900.

4. The paper was published in 1900 and received a citation during the study, but the end of study coincided with the time of citation. This is a non-censored observation.

5. The paper was published in 1900 and did not receive citations until the end of the study. Such a paper is right-censored.

6. The event occurred during the study as the paper was published after 1900 and cited before 2022.

7. The paper was published after 1900 and before 2022, but not cited until the end of the study. In this case, the paper is right-censored.

Figure 7.8 illustrates various survival histories of first citations, including all seven aforementioned scenarios.

In what follows, we treat the duration of time elapsed before a paper receives its first citation as a survival time. Hence, the data comprising such times for a set of papers may be studied using statistical survival analysis.

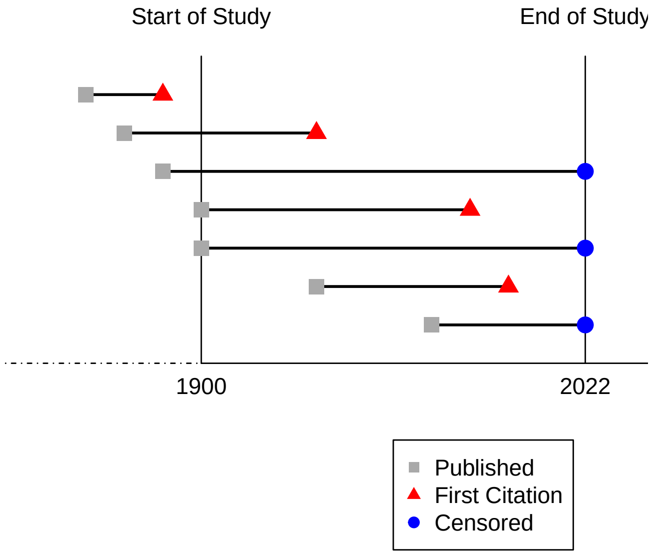


Figure 7.8: Different survival histories with censoring. Here, “death” is interpreted as a first citation after publication of the paper.

7.2.2 Basic concepts: survival function and hazard rate

Assume that papers receive their citations independently of one another. Let T denote the time when a paper receives its first citation after publication. This is interpreted as “death” and the duration from 0 to T as the “survival time”. Assume that the random survival time $T_1, T_2, \dots$ have the same distribution as T. Suppose that the random variable T has continuous distribution function with the probability density $f(t)$,

$$F(t) := \mathsf{P}(T < t) = \int_0^t f(u)\,\mathrm{d}u \qquad (t \geq 0). \tag{7.1}$$

The survival function is then defined as

$$S(t) := \mathsf{P}(T \geq t) = 1 - F(t) \qquad (t \geq 0). \tag{7.2}$$

Note that $S(0) = \mathsf{P}(T \geq 0) = 1$, since $T \geq 0$. If “death” is a certain event (i.e., it occurs with probability 1, sooner or later), then $S(\infty) = \lim_{t\to\infty} S(t) = 0$. However, if this is not the case (e.g., with a first citation which may never occur) then $S(\infty) > 0$.

The *mean survival time* is given by

$$\mathsf{E}(T) = \int_0^\infty t\, f(t)\, \mathrm{d}t = \int_0^\infty S(t)\, \mathrm{d}t, \tag{7.3}$$

where the second formula is obtained by integration by parts. The *median survival time* μ is defined by the property $\mathsf{P}(T < \mu) = \mathsf{P}(T \geq \mu)$, that is, $1 - S(\mu) = S(\mu)$. Hence, μ is the root of the equation

$$S(\mu) = 0.5. \tag{7.4}$$

The *hazard function* $h(t)$ (also called *hazard rate*) is defined as the conditional probability of death at time t,

$$h(t) := \frac{\mathsf{P}(T \in [t, t + \mathrm{d}t)\,|\, T \geq t)}{\mathrm{d}t}. \tag{7.5}$$

Here, $\mathrm{d}t$ is an infinitely small time increment, so the expression on the right-hand side should be understood as the limit as $\mathrm{d}t \to 0$.

Expressing the conditional probability in definition (7.5), we have

$$\frac{\mathsf{P}(t \leq T < t + \mathrm{d}t)}{\mathsf{P}(T \geq t)\, \mathrm{d}t} = \frac{S(t) - S(t + \mathrm{d}t)}{S(t)\, \mathrm{d}t} = \frac{-S'(t)\, \mathrm{d}t}{S(t)\, \mathrm{d}t}.$$

Recalling that $S'(t) = -f(t)$, this gives an explicit formula for the hazard rate,

$$h(t) = -\frac{S'(t)}{S(t)} = \frac{f(t)}{S(t)} \qquad (t \geq 0). \tag{7.6}$$

This formula can be rewritten in a more intuitive way as

$$h(t)\, \mathrm{d}t = S(t) \cdot f(t)\, \mathrm{d}t,$$

where the term $S(t)$ on the right represents survival up to t and the second term $f(t)\, \mathrm{d}t$ stands for instantaneous death thereafter.

Formula (7.6) can be viewed as a differential equation for $y = S(t)$. Note that

$$h(t) = -\frac{S'(t)}{S(t)} = -\frac{\mathrm{d}}{\mathrm{d}t}\big(\log S(t)\big). \tag{7.7}$$

Integrating (7.7) and using the initial condition $S(0) = 1$, we obtain

$$\log S(t) = -\int_0^t h(u)\, \mathrm{d}u = -H(t). \tag{7.8}$$

where $H(t)$ is the cumulative (integrated) hazard function,

$$H(t) = \int_0^t h(u)\,\mathrm{d}u. \tag{7.9}$$

Thus, the survival function is uniquely determined by the hazard function according to the formula

$$S(t) = \exp\left(-\int_0^t h(u)\,\mathrm{d}u\right) = \mathrm{e}^{-H(t)} \qquad (t \geq 0). \tag{7.10}$$

7.2.3 Kaplan–Meier survival estimator

Consider a group of n subjects with a recorded survival time $t_1, t_2, \ldots, t_n$. Among these time points, there are $m \leq n$ cases of death, with the associated times arranged in ascending order denoted by $t_{(1)}, t_{(2)}, \ldots, t_{(m)}$, where $t_{(0)} := 0$ and $t_{m+1} := \infty$. For $i = 0, 1, \ldots, m$, let n_i denote the number of subjects alive just before $t_{(i)}$, d_i denote the number of death at time $t_{(i)}$, and c_i denote the number of censored subjects at or after $t_{(i)}$ but before $t_{(i+1)}$. It follows that $n_{i+1} = n_i - c_i - d_i$. The probability of death during each interval $[t_{(i)}, t_{(i+1)})$ is estimated by

$$\hat{q}_i = \frac{d_i}{n_i},$$

and the probability of survival is estimated by

$$\hat{p}_i = 1 - \hat{q}_i = \frac{n_i - d_i}{n_i}.$$

Note that the censored subjects, denoted by c_i, are accounted for in the total number of subjects at risk n_i (just before $t_{(i)}$), so they are exposed to hazard at the time of death $t_{(i)}$. To prevent ambiguity, if a censored survival time coincides with the death time $t_{(i)}$, it is presumed that death happens before any censored times, with the latter occurring immediately thereafter.

Then a product-type of estimation for the survival function, called the *Kaplan–Meier (KM) estimator*, is given by

$$\hat{S}(t) = \prod_{i=0}^{k} \frac{n_i - d_i}{n_i} \qquad (t_{(k)} \leq t < t_{(k+1)}). \tag{7.11}$$

Note that $\hat{S}(t) = 1$ for $0 \le t < t_{(1)}$, since $d_0 = 0$. In the particular case with no censoring (i.e., all $c_i = 0$), we have $n_{i+1} = n_i - d_i$ and the Kaplan–Meier formula (7.11) is reduced to the usual empirical distribution function estimator,

$$\hat{S}(t) = \frac{n_{k+1}}{n} \qquad (t_{(k)} \le t < t_{(k+1)}).$$

The fundamental formula (7.11) can be intuitively justified as follows. Denote $A_i := \{T > t_{(i)}\}(i = 0, 1, \ldots, m)$, then

$$\mathsf{P}(A_i|A_{i-1}) \approx 1 - \frac{d_i}{n_i} = \frac{n_i - d_i}{n_i}. \tag{7.12}$$

Then, viewing the survival experience sequentially, The probability of event A_k can be represented as a product of conditional probabilities accounting for increasing history,

$$\begin{aligned}\mathsf{P}(A_k) &= \mathsf{P}(A_k|A_{k-1}) \times \mathsf{P}(A_{k-1}) \\ &= \mathsf{P}(A_k|A_{k-1}) \times \mathsf{P}(A_{k-1}|A_{k-2}) \times \mathsf{P}(A_{k-2}) \\ &= \cdots = \mathsf{P}(A_k|A_{k-1}) \times \mathsf{P}(A_{k-1}|A_{k-2}) \times \cdots \times \mathsf{P}(A_1|A_0) \times \mathsf{P}(A_0). \end{aligned} \tag{7.13}$$

Noting that $\mathsf{P}(A_0) = 1$ and using the estimate (7.12), this product is approximated as

$$\mathsf{P}(A_k) = S(t_{(k)}) \approx \prod_{i=1}^{k} \frac{n_i - d_i}{n_i}, \tag{7.14}$$

which leads to the Kaplan–Meier product estimator (7.11).

The KM survival estimator can be calculated using command `survfit` under `survival` library in R (Therneau, 2022).

The variance of the Kaplan-Meier estimator is given by Greenwood's formula:

$$\mathsf{Var}\{\hat{S}(t)\} \approx \{\hat{S}(t)\}^2 \sum_{i=0}^{k} \frac{d_i}{n_i(n_i - d_i)}, \qquad t_{(k)} \le t < t_{(k+1)} \tag{7.15}$$

7.2.4 Comparison of groups and the Cox proportional hazards model

There is often a need to compare various survival data sets, e.g. to determine if different groups of data are statistically similar (i.e. belong to the same general population) or there are significant differences. One commonly used method is log-rank test. Suppose we want to test the hypothesis of homogeneity of two groups, that is, that survival times observed in groups 1 and 2 have the same theoretical survival functions,

$$H_0 \colon S_1(t) \equiv S_2(t) \qquad (t \geq 0).$$

Let $(t_{(i)})$ be a pooled set of death times in the union of the two groups, and denote by n_i and d_i the corresponding pooled numbers at risk and numbers of death at time $t_{(i)}$. Also, denote by n_{1i} and n_{2i} the individual numbers at risks, and by d_{1i} and d_{2i} the individual numbers of deaths in the groups. Note that under the null hypothesis H_0 the probability of death at time $t_{(i)}$ can be estimated from the pooled data as $\hat{p}_i = d_i/n_i$, while the expected shares of deaths in the groups will be proportional to $\hat{p}_i$ and their numbers at risk, n_{1i} and n_{2i}, respectively.

The log-rank test is based on the chi-squared type statistic

$$W := \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2}, \tag{7.16}$$

where O_1 and O_2 are the observed numbers of deaths in groups 1 and 2,

$$O_1 = \sum_{i=1}^{m} d_{1i}, \qquad O_2 = \sum_{i=1}^{m} d_{2i},$$

and E_1 and E_2 are the expected numbers of deaths in the groups under the null hypothesis H_0,

$$E_1 = \sum_{i=1}^{m} \frac{n_{1i} d_i}{n_i}, \qquad E_2 = \sum_{i=1}^{m} \frac{n_{2i} d_i}{n_i}.$$

It can be shown that W has approximately a chi-squared distribution χ_1^2 with one degree of freedom. Hence, at a significance level α, we reject the null hypothesis H_0 if $W_L > k_{1-\alpha}$, where $k_{1-\alpha}$ is the $(1-\alpha)$-quantile of χ_1^2.

A more flexible approach, which allows to assess the impact of various covariates on survival, is based on the *Cox proportional hazards model.* It is based on the idea to model the hazard by separating the impact of explanatory variables from the time dependence, represented by a *baseline hazard rate.* More precisely, the corresponding modelling assumption is that the hazard rate of the i-th observed individual ($i = 1, \ldots, n$) may be represented in the form

$$h_i(t) = \mathrm{e}^{\beta_1 x_{i1} + \cdots + \beta_p x_{ip}} h_0(t) \qquad (t \geq 0), \tag{7.17}$$

where the quantity $\beta_1 x_{i1} + \cdots + \beta_p x_{ip}$ is called the *risk score* (Cox, 1972). Note that under this model, the hazard ratio between any two individuals is a constant not depending on time,

$$\frac{h_i(t)}{h_j(t)} = \exp\Bigl(\beta_1 (x_{i1} - x_{j1}) + \cdots + \beta_p (x_{ip} - x_{jp})\Bigr) \qquad (t \geq 0).$$

This explains the term "proportional hazards". Note that no modelling assumptions are made about the baseline hazard rate $h_0(t)$ in (7.17), which is why the Cox model is often referred to as *semi-parametric*, that is, half-parametric due to the regression part in (7.17) and half-non-parametric due to the unspecified function $h_0(t)$.

The regression parameters β_k can be estimated by maximising a properly constructed partial likelihood (see more details in Collett (2015)). The R command `coxph` can be used to fit the Cox proportional hazards model. It also returns the estimated standard errors of the estimates $\hat{\beta}_k$ and the corresponding p-values for testing the hypotheses $H_{0k}\colon \beta_k = 0$, that is, that the corresponding covariate x_k is not included in the model, which is just another way to say that this covariate is not influential and, therefore, may be omitted from the model.

The Cox proportional hazards model can be used to compare two groups; to this end, let the variable x be defined so that the values $x = 0$ and $x = 1$ correspond to being in group 1 or group 2, respectively. Here, formula (7.17) specialises as follows,

$$h_i(t) = \mathrm{e}^{\beta x_i} h_0(t) \qquad (t \geq 0), \tag{7.18}$$

where $x_i = 0$ or $x_i = 1$ depending on whether the i-th individual is in group 1 or group 2, respectively. Thus, the Cox framework provides a useful regression-based alternative to a non-parametric log-rank test.

7.2.5 Application to first citation data

The survival analysis is applied to the extended EJP data (see Section 2.2 D') in this section. Set the publication year of papers as year 1. Figure 7.9 shows the estimated survival function plot in terms of the time to first citations of papers from the extended EJP data, i.e., the probability of not cited after publications the corresponding year. The median survival time of receiving the first citation is two years.

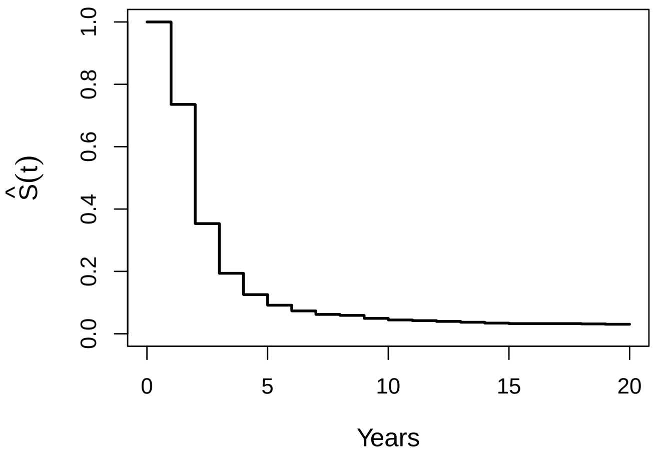


Figure 7.9: Estimated survival plot for the extended EJP data using the Kaplan–Meier estimator. The solid black line is the probability of not having any citations yet in the corresponding year.

Suitable covariates may be included in the survival model and assessed in terms of their importance. In the context of scientific production, Nane (2015) considered some features as covariates: collaboration type (international or national), document type, the number of authors, and the field of research. Unfortunately, due to limitations of the Web of Science, only the number of authors is available and, therefore, can be used as a covariate. In addition, the length of the paper is included as a covariate.

Using our data set, we experimented with considering the number of collaborators as one of the covariates. With the aid of Excel, the number of authors of a paper is counted by counting the semicolons and plus one in the cell of the authors, i.e., using command `= Len(A1) - Len(Substitute(A1, ";", "")) + 1`.

The left panel of Figure 7.10 shows the distribution of collaborators of a paper in a histogram. Papers with two authors are the majority. The mean number of collaborators in this data is 2.93, and the median is 2.

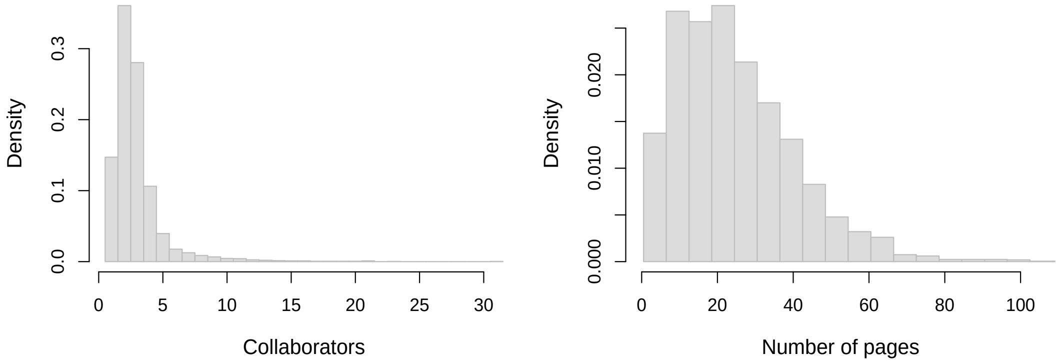


Figure 7.10: Histograms of the number of collaborators (left panel) and the number of pages (right panel) of a paper.

One of these papers has 2,333 collaborators, which is rather exceptional for this data set, so we excluded it from our analysis. This paper is based on an ATLAS collaboration (see `https://en.wikipedia.org/wiki/ATLAS_experiment#ATLAS_detector`), where ATLAS is an abbreviation of *A Toroidal LHC Apparatus.*

It is worth noting that publications in high-energy physics, particle physics, and cosmology often have huge author lists. This is because such work may have been done on big installations such as the Large Hadron Collider (LHC), in which case the publication would include the entire personnel involved in such work.

First, suppose that papers are divided into two categories according to the number of authors. Specifically, single-authored papers are in one category, and papers with collaboration (i.e., with more than one author) are in another category. The left panel of Figure 7.11 shows the plot of the survival functions estimated by the KM estimator of the extended EJP data in two different groups. The single-authored papers have a higher survival probability, which means that the single-authored papers have a lower probability of being cited earlier than collaborative papers. In contrast, survival probability for collaborative papers is lower than that for single-authored papers; this means that collaborative papers tend to be cited earlier in this data set.

To test the significance of differences statistically, we applied the log-rank test using R; the observed statistic value is $W_{\text{obs}} = 47.9$, with the corresponding p-value of 4×10^{-12} (based on a chi-squared distribution with one degree of freedom). Hence, we strongly reject the null hypothesis, which means that these two groups have a significant difference with regard to survival.

We also fitted the Cox proportional hazards model (7.18), with the estimated regression coefficient $\hat{\beta} = 0.34232$. The corresponding null hypothesis $H_0\colon \beta = 0$ is strongly rejected with the p-value 9.869×10^{-13}, which confirms the log-rank test conclusion.

Similarly, we looked into survival differences with regard to the length of the papers. The right panel of Figure 7.10 shows the histogram of the number of pages of these publications. The mean paper length is 25.07, and the median is 22. In some journals, submitted manuscripts are required to be no more than 6 pages, hence in this section papers are classified into two groups in terms of their length, long papers (longer than 6 pages) and short papers (6 pages or less). This classification can be another covariate in the survival analysis. The middle panel of Figure 7.11 shows the estimated survival plots of the extended EJP data for these two categories. Perhaps unexpectedly, it appears that shorter papers have higher chances of survival, which means that they have a lower chance of being cited earlier as compared to longer papers in this data set.

Applying the log-rank test in R, we obtained the observed statistic value $W_{\text{obs}} = 78.4$, with the p-value less than 2.2×10^{-16} (based on the approximately chi-squared distribution with one degree of freedom). Hence, the null hypothesis is very strongly rejected, which means that long and short papers are significantly different in terms of timing to first citation.

As a cross check, by fitting the Cox proportional hazards model (7.18) and testing the null hypothesis $H_0\colon \beta = 0$, we obtained the p-value less than 2.2×10^{-16}, which confirms that the null H_0 is strongly rejected and, therefore, these two groups are significantly different.

Lastly, we look at survival probability differences when both covariates are included, accounting for the paper length and collaboration, respectively. That is, papers are now divided into four categories: (i) short single-authored, (ii) short collaborative, (iii) long single-authored, and (iv) long collaborative. The result is shown in the right panel of Figure 7.11. Short single-authored papers have the highest survival probability, followed by short collaborative papers, long single-authored papers, and finally long collaborative papers. This means that long collaborative papers have a higher chance to receive their first citation sooner, as compared to the other three categories in this data set. In addition, short papers and long papers have clear separation at the end in this plot. By applying the log-rank test, we obtain $W_{\text{obs}} = 121$, with the corresponding p-value less than 2×10^{-16} (based on a chi-squared distribution with 3 degree of freedom). Therefore, we strongly reject the null hypothesis and these four groups are significantly different.

Cross checking by fitting the Cox proportional hazards model (7.18), we strongly reject the null hypothesis $\beta = 0$ with obtained p-value less than 2.2×10^{-16}, which confirms that these four groups are different.

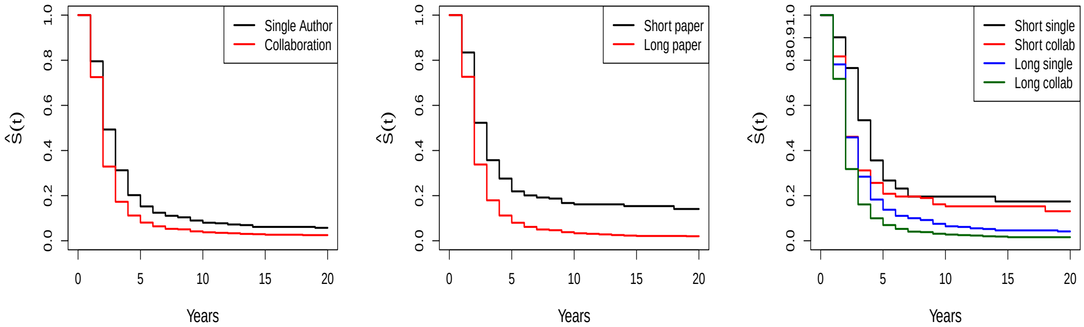


Figure 7.11: Survival plots until first citation for papers categorised according to different covariates. Left panel: single-authored (black) or collaborative (red). Middle panel: short (at least six pages, black) or long (more than six pages, red). Right panel: short single-authored (black), short collaborative (red), long single-authored (blue), and long collaborative (green).

We conjecture that the above results may be explained by noting that the more authors a paper has, the higher chance is there to disseminate the paper to the public (e.g., in conferences and seminars). So it will be known by more people, which may push the paper to be cited by others. On the other hand, our findings about the impact of length of the paper could be explained by noting that a longer paper has more content, so the perspectives for being cited are more than with shorter papers. In the present thesis, these conjectures have not been tested from the data, it is worth checking these points in the future.

As a concluding comment, survival analysis conducted above could be improved by inclusion of left-truncated papers instead of omitting them as in the present study, but the analysis will become more complicated.

7.3 Citations as a Point Process

In this section, we consider dynamic citations as a point process. Point processes are stochastic processes employed in modelling events that occur at irregular intervals concerning either the temporal or spatial axis (Daley & Vere-Jones, 2003). Before generalising to the Hawkes process, simpler models of Poisson and inhomogeneous Poisson processes are introduced.

Following Daley & Vere-Jones (2003), let $(T_i) = (0 < T_1 < T_2 < \dots)$ denote the consecutive random times of occurrence of some events (such as "arrivals"). It is assumed that there are no ties, so the next arrival time is strictly larger than the previous one, $T_i < T_{i+1}$. Such point processes are called "simple". Let N_t be the accumulated number of arrivals over time interval $[0, t]$,

$$N_t := \sum_{(T_i)} I_{\{T_i \le t\}} \qquad (t \ge 0). \tag{7.19}$$

The counting random process (N_t) is piecewise constant, with unit jumps at the arrival times T_i.

In practice, the data may include multiple items at each arrival (see Figure 7.12), which may be due to the problem setting (e.g., number of cars involved in a traffic accident) or because of data aggregation (e.g., monthly or yearly data). In such cases, one may use compound processes, where the number of items ν_i in arrival T_i is assumed independent of T_i and modelled via a certain distribution on top of the background point process (T_i), for example, geometric, $\mathsf{P}(\nu_i = k) = \alpha\,(1-\alpha)^{k-1}$, or conditional Poisson, $\mathsf{P}(\nu_i = k) = \left(1 - \mathrm{e}^{-\mu}\right)\mu^k\,\mathrm{e}^{-\mu}/k!$ $(k = 1, 2, \dots)$.

7.3.1 Poisson process

Let τ_i be a random "waiting" time between two consecutive arrivals at times T_{i-1} and T_i $(i = 1, 2, \dots)$, where we set $T_0 = 0$. Then the arrival times are expressed

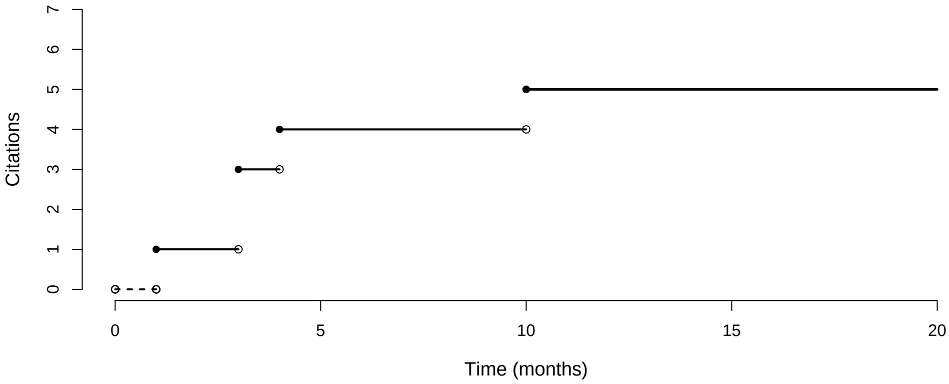


Figure 7.12: Cumulative number of citations N_t over time t with mock arriving time $T = (1, 3, 3, 4, 10)$. The jump points represent the instants of time when new citations arrive.

as $T_n = \sum_{i=1}^{n} \tau_i$. The random variables (τ_i) are also called *inter-arrival times*.

Definition 7.1. The point process (T_i) defined in (7.19) is called a *(homogeneous) Poisson process* with intensity $\lambda > 0$ if inter-arrival times (τ_i) are mutually independent and follow exponential distribution with parameter $\lambda > 0$, that is,

$$\mathsf{P}(\tau_i > t) = \mathrm{e}^{-\lambda t} \qquad (t \geq 0).$$

The probability density function of τ is given by

$$f_\tau(t) = \lambda \mathrm{e}^{-\lambda t} \qquad (t \geq 0), \tag{7.20}$$

with the expected value

$$\mathsf{E}(\tau) = \int_0^\infty t f_\tau(t)\,\mathrm{d}t = \lambda \int_0^\infty t\,\mathrm{e}^{-\lambda t}\,\mathrm{d}t = \frac{1}{\lambda}. \tag{7.21}$$

The Poisson process is *memoryless*, that is, the distribution of future inter-arrival times does not depend on the earlier arrivals (i.e., past history of the process). This is illustrated by calculating a conditional distribution of the remaining waiting time $\tau - s$ given that $\tau > s$,

$$\begin{aligned}\mathsf{P}(\tau > t + s \,|\, \tau > s) &= \frac{\mathsf{P}(\tau > t + s, \tau > s)}{\mathsf{P}(\tau > s)} = \frac{\mathsf{P}(\tau > t + s)}{\mathsf{P}(\tau > s)} \\ &= \frac{\mathrm{e}^{-\lambda(t+s)}}{\mathrm{e}^{-\lambda s}} = \mathrm{e}^{-\lambda t} = \mathsf{P}(\tau > t). \end{aligned} \tag{7.22}$$

Thus, $\mathsf{P}(\tau > t + s | \tau > s) = \mathsf{P}(\tau > t)$, which means that the previous waiting of at least time s does not change the exponential distribution of the remaining time, as if the waiting is renewed at time s.

A useful alternative description of the Poisson process is through the consideration of possible jumps within an infinitely small amount of time $\mathrm{d}t$, leading to the following asymptotic formulas,

$$\begin{cases} \mathsf{P}(N_{t+\mathrm{d}t} = n+1 \,|\, N_t = n) = \lambda\,\mathrm{d}t + o(\mathrm{d}t), \\ \quad \mathsf{P}(N_{t+\mathrm{d}t} = n \;|\, N_t = n) = 1 - \lambda\,\mathrm{d}t + o(\mathrm{d}t), \\ \mathsf{P}(N_{t+\mathrm{d}t} \geq n+2 \,|\, N_t = n) = o(\mathrm{d}t). \end{cases} \tag{7.23}$$

Thus, a jump by one unit occurs with probability proportional to time $\mathrm{d}t$, with intensity λ as the proportionality coefficient, with probability of making more than one jump being of higher order of smallness, $o(\mathrm{d}t)$.

To obtain the likelihood, assume that over the time interval $[0, t]$, we observed $N_t = N$ arrivals at times $0 = T_0 \leq T_1 < T_2 < \cdots < T_{N_t} \leq t$. Then, using the independence of the waiting times $\tau_i = T_i - T_{i-1}$, the likelihood is given by

$$\mathcal{L}(\lambda; N, \boldsymbol{T}) = \lambda\,\mathrm{e}^{-\lambda T_1} \times \lambda\,\mathrm{e}^{-\lambda(T_2 - T_1)} \times \cdots \times \lambda\,\mathrm{e}^{-\lambda(T_N - T_{N-1})} \times \mathrm{e}^{-\lambda(t - T_N)},$$

where the last factor accounts for no arrivals from T_N till t. Simplifying, this yields

$$\mathcal{L}(\lambda; N, \boldsymbol{T}) = \lambda^N\,\mathrm{e}^{-\lambda t}.$$

Hence, the log-likelihood $\ell = \log \mathcal{L}$ is

$$\ell(\lambda; N, \boldsymbol{T}) = N \log \lambda - \lambda t.$$

It is easy to find the maximum likelihood of λ,

$$\ell'(\lambda; N, \boldsymbol{T}) = \frac{N}{\lambda} - t = 0 \quad \Rightarrow \quad \hat{\lambda} = \frac{N}{t}.$$

Thus, the MLE $\hat{\lambda}$ is given by the mean number of arrivals per unit time, which is consistent with the meaning of λ as the arrival intensity.

The *inhomogeneous Poisson process* is defined similarly, but it allows the intensity to be a (deterministic) function of time, $\lambda = \lambda(t)$. The simple choice of the time-dependent intensity is to assume that it is piecewise constant between some change-points (Gyarmati-Szabó *et al.*, 2011).

The equations (7.23) are transformed accordingly,

$$\begin{cases} \mathsf{P}(N_{t+\mathrm{d}t} = n+1 \,|\, N_t = n) = \lambda(t)\,\mathrm{d}t + o(\mathrm{d}t), \\ \mathsf{P}(N_{t+\mathrm{d}t} = n \,\,|\, N_t = n) = 1 - \lambda(t)\,\mathrm{d}t + o(\mathrm{d}t), \\ \mathsf{P}(N_{t+\mathrm{d}t} \geq n+2 \,|\, N_t = n) = o(\mathrm{d}t). \end{cases} \tag{7.24}$$

Note that this process is also memoryless, since the intensity $\lambda(t)$ does not depend on the past history.

A more general class of point processes *with memory* can be defined using the notion of conditional intensity,

$$\lambda(t \,|\, \mathcal{H}_t) = \frac{\mathsf{P}\big(N_{t+\mathrm{d}t} - N_t = 1 \,\big|\, \mathcal{H}_t\big)}{\mathrm{d}t}, \tag{7.25}$$

where $\mathcal{H}_t$ is the history of the process up to time t, which comprises all past arrival times $T_1, T_2, \ldots, T_{N_t}$. Accordingly, the equations (7.24) are further modified,

$$\begin{cases} \mathsf{P}(N_{t+\mathrm{d}t} = n+1 \,|\, N_t = n, \mathcal{H}_t) = \lambda(t|\mathcal{H}_t)\,\mathrm{d}t + o(\mathrm{d}t), \\ \mathsf{P}(N_{t+\mathrm{d}t} = n \,\,|\, N_t = n, \mathcal{H}_t) = 1 - \lambda(t|\mathcal{H}_t)\,\mathrm{d}t + o(\mathrm{d}t), \\ \mathsf{P}(N_{t+\mathrm{d}t} \geq n+2 \,|\, N_t = n, \mathcal{H}_t) = o(\mathrm{d}t). \end{cases} \tag{7.26}$$

This general approach can be made more specific by defining the Hawkes process, considered in the next section.

7.3.2 Hawkes process

The class of Hawkes processes was introduced by Hawkes (1971) with the aim to capture the possible dependence on past event. The key idea is to model the conditional intensity of the process as a linear combination of inputs from past arrivals using a certain self-exciting kernel. Examples of practical applications of the Hawkes processes are ubiquitous, including modelling of earthquake aftershocks (Vere-Jones & Ozaki, 1982), social media activity (Rizoiu *et al.*, 2018), epidemic dynamics (Browning *et al.*, 2021), and many more. The similarity of the self-exciting property of the Hawkes processes with the scientometric principle “success breeds success” makes the choice of this class of point processes particularly attractive for modelling dynamic citation data.

Definition 7.2. A point process is called a *Hawkes process* if the conditional intensity function (7.25) is of the form

$$\lambda(t\,|\,\mathcal{H}_t) = \lambda_0(t) + \sum_{T_i<t} k(t-T_i), \tag{7.27}$$

where $\lambda_0(t)$ is a (deterministic) base intensity function and $k(t)$ is the memory kernel.

Remark 7.1. The kernel $k(t)$ is usually assumed to be a non-increasing function, which is motivated by the natural assumption that older events have a smaller impact on future arrivals.

The interpretation of the kernel $k(t)$ is that it accounts for earlier events at times $T_i < t$, which influence the new arrivals through the updated intensity $\lambda(t\,|\,\mathcal{H}_t)$ at $t \geq 0$ (hence, the term "self-exitation"). When the kernel vanishes, $k(t) \equiv 0$, the Hawkes process is reduced to an inhomogeneous Poisson process with intensity $\lambda_0(t)$.

A popular choice of the kernel is an exponential function, which is given by

$$k(t) = a\,\mathrm{e}^{-bt} \qquad (t \geq 0), \tag{7.28}$$

where $a \geq 0$, $b > 0$ and $a < b$. Another choice widely used in the literature is a power-law kernel,

$$k(t) = \frac{\beta}{(t+\delta)^{\alpha}+1} \qquad (t \geq 0), \tag{7.29}$$

where $\beta \geq 0$, $\alpha, \delta > 0$ and $\beta < \alpha\delta^{\alpha}$.

7.3.3 Hawkes process with multiple arrivals

As already mentioned at the beginning of Section 7.3, in practical applications it is often the case that the data records are aggregated, which leads to ties in the data. This occurs in our citation data set, as the numbers of collected citations are reported on a monthly basis. On the other hand, multiple items at arrival cannot be ignored: having more citations in the current month makes it more likely that there will be further citations, thus enhancing the self-excitation mechanism.

To address this complication, we propose to understand the summation in formula (7.27) as extending over all individual arrivals, with account of their multiplicities. That is to say, if there are ν_i items arriving at time T_i, then the conditional intensity is given by

$$\lambda(t\,|\,\mathcal{H}_t) = \lambda_0(t) + \sum_{T_i<t} \nu_i\, k(t-T_i). \tag{7.30}$$

Note that a similar approach is used for discrete-time Hawkes processes (Browning *et al.*, 2021).

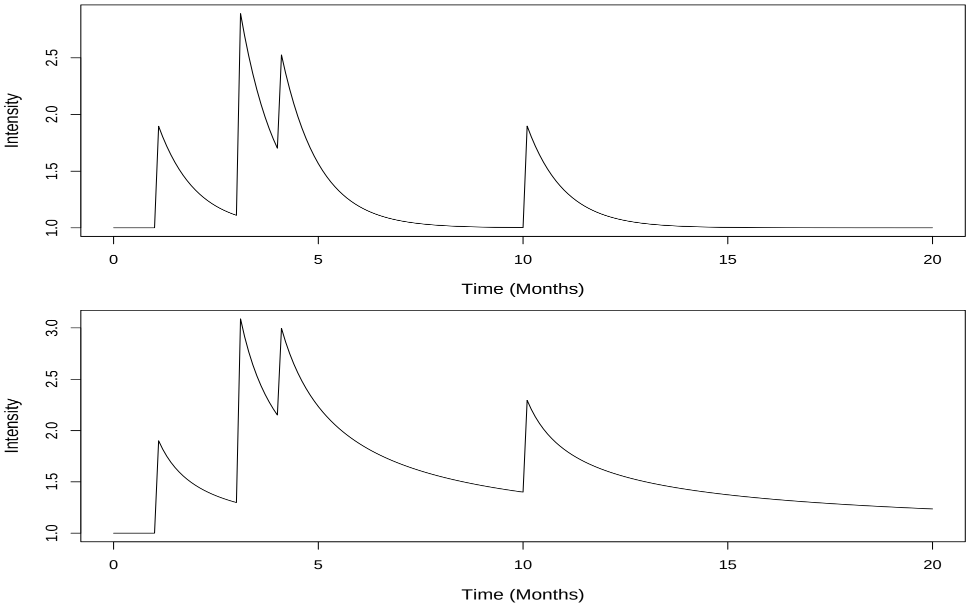


Figure 7.13: Intensity plot of the Hawkes process on a mock example, corresponding to Figure 7.12. The upper panel is with the exponential kernel ($\lambda_0 = 1$, $a = 1$, $b = 1.1$), and the lower panel is with the power-law kernel ($\lambda_0 = 1$, $\alpha = 1$, $\beta = 1$, $\delta = 1.1$).

For illustration, plots of intensity (7.30) with an exponential kernel (7.28) and a power law kernel (7.29) are depicted in Figure 7.13. Given a mock arriving time $T = (1, 3, 3, 4, 10)$, for simplicity assuming that the background intensity λ_0 is a constant. For an illustration (not fitting) Figure 7.13 depicted two different intensity as a function of time with two exponential and power law kernel functions given (7.28) and (7.29). The intensity function of these two examples are

$\lambda(t|\mathcal{H}_t) = 1 + \sum_{T_i<t} \mathrm{e}^{-1.1t}$ and $\lambda(t|\mathcal{H}_t) = 1 + \sum_{T_i<t} \frac{1}{(t+1.1)^2}$, respectively. At each time T_i, intensity increase when a new event arrives. Then while waiting for the next arrival, the intensity reduces in the exponential way (upper panel) or in a power law way (the lower panel).

7.3.4 Maximum likelihood estimation

Consider events (arrivals) happening within the time interval $[0, t]$, and let $0 \leq T_1 < T_2 < \cdots < T_{N_t} \leq t$ be the corresponding arrival times. Considering arrivals sequentially and using the decomposition into a product of conditional probabilities accounting for increasing history (cf. (7.13)), the likelihood corresponding to the observed value $N_t = N$ and the data $\boldsymbol{T} = (T_1, T_2, \ldots, T_N)$ is given by

$$\begin{aligned}
\mathcal{L}(\boldsymbol{\theta}; N, \boldsymbol{T}) = \exp&\left(-\int_0^{T_1} \lambda(u\,|\,\mathcal{H}_0)\,\mathrm{d}u\right) \lambda(T_1\,|\,\mathcal{H}_{T_1}) \\
&\times \exp\left(-\int_{T_1}^{T_2\,|\,\mathcal{H}_{T_2}} \lambda(u\,|\,\mathcal{H}_{T_1})\,\mathrm{d}u\right) \lambda(T_2\,|\,\mathcal{H}_{T_2}) \\
&\times \cdots \times \exp\left(-\int_{T_{N-1}}^{T_N} \lambda(u\,|\,\mathcal{H}_{T_{N-1}})\,\mathrm{d}u\right) \lambda(T_N\,|\,\mathcal{H}_{T_N}) \\
&\times \exp\left(-\int_{T_N}^{t} \lambda(u\,|\,\mathcal{H}_{T_N})\,\mathrm{d}u\right) \\
= \exp&\left(-\int_0^{t} \lambda(u\,|\,\mathcal{H}_0)\,\mathrm{d}u\right) \prod_{i=1}^{N} \lambda(T_i\,|\,\mathcal{H}_{T_i}).
\end{aligned} \tag{7.31}$$

Hence, the log-likelihood $\ell = \log \mathcal{L}$ is given by

$$\ell(\boldsymbol{\theta}; N, \boldsymbol{T}) = -\int_0^t \lambda(u\,|\,\mathcal{H}_0)\,\mathrm{d}u + \sum_{i=1}^{n} \log \lambda(T_i\,|\,\mathcal{H}_{T_i}). \tag{7.32}$$

Substituting a particular functional form of the background intensity $\lambda_0(t)$ (e.g., for simplicity assuming that it is constant, $\lambda_0(t) = \lambda$) and using the parametric kernel (7.28) or (7.29), the log-likelihood (7.32) can be numerically maximised using the `optim()` function in R.

If the Hawkes model with multiple arrivals is required, with the corresponding data $(\boldsymbol{T}, \boldsymbol{\nu}) = \big((T_i, \nu_i)\big)$ (see Section 7.3.3), then the log-likelihood $\ell(\boldsymbol{\theta}; N, \boldsymbol{T}, \boldsymbol{\nu})$ is again given by formula (7.32) but with the conditional intensity $\lambda(t\,|\,\mathcal{H}_t)$ modified according to formula (7.30).

7.3.5 Application to citation modelling

It is a natural idea to model citation arrivals as a point process. Set a month as a time unit, and set the month the paper is published as 0. Suppose a paper is published in January of a certain year, set this month $T_0 = 0$. Then it received its first single citation in February the same year, i.e., $N_1 = 1$; two citations in April, i.e., $N_3 = 3$; one in May, i.e., $N_4 = 4$; and one citation in November, i.e., $N_{10} = 5$. The arrivals of citations of this paper happened in months $(1, 3, 3, 4, 10)$. Figure 7.12 demonstrates citations N_t accumulated over time t in this example.

Using the same way of showing the number of citations N arrived at time t, Figure 7.14 displays examples of citations accumulated by months. These three plots used the same type of representation as in Figure 7.12; only the lines in waiting for the arrival of new events are omitted to make the plot concise. Every dot in these plots represents a citation. The left panel shows the citation growth of a paper published in January 1983, which received 7 citations until February 2023. The middle panel shows the citation growth of a paper published in January 2013, which received 50 citations until February 2023. The right panel displays a paper published in November 2005 that received 4295 citations until February 2023.

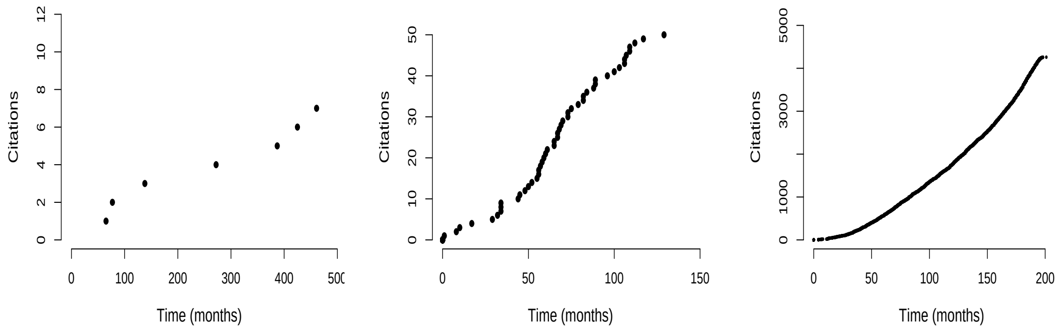


Figure 7.14: Cumulative citation plots versus time for three papers, with the total number of citations: 7 (left panel), 50 (middle panel), and 4,295 (right panel).

The Hawkes processes were fitted to the actual citation arrivals data presented in Figure 7.14. By setting the background intensity to the same function type as the kernel, albeit with a distinct parameter. Specifically, the intensity used to conform to the data for the exponential kernel is expressed as follows,

$$\lambda(t|\mathcal{H}_t) = a' \mathrm{e}^{b't} + \sum_{T_i<t} \nu_i a \, \mathrm{e}^{b(t-T_i)}. \tag{7.33}$$

The intensity with the power law kernel and the power law background is set as

$$\lambda(t|\mathcal{H}_t) = \frac{\beta'}{(t+\delta')^{\alpha'+1}} + \sum_{T_i<t} \nu_i \frac{\beta}{(t-T_i+\delta)^{\alpha+1}}. \tag{7.34}$$

Figure 7.15 depicts the result of fitting the Hawkes processes to citation arrival data with both exponential kernel and the power law kernel. Note that the dashed lines after the end of the observation in these plots do not indicate predictions.

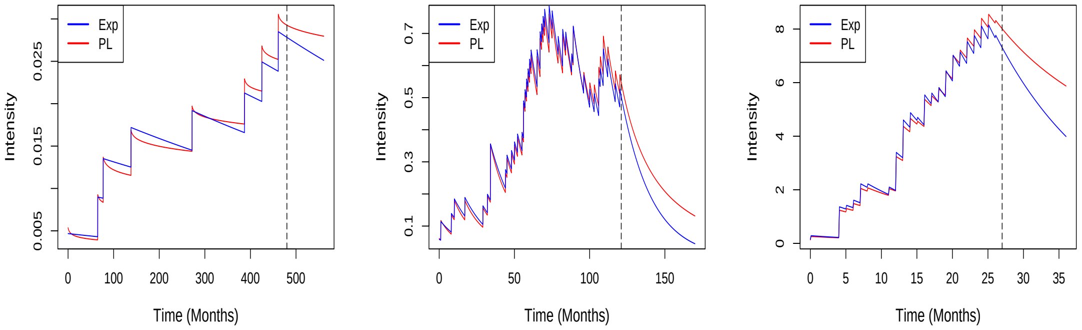


Figure 7.15: Intensity plot of the Hawkes process fitted to citations of three selected papers (cf. Figure 7.14). Blue lines are with exponential kernel function and the red lines are with the power law function. Dashed vertical lines represent termination of observations (February 2023), that is, months 480, 121, and 27, respectively. Note that the right panel only shows the first 100 citations among 4,259 citations due to computation limitations.

Chapter 8

Conclusion

8.1 What This Thesis Has Accomplished

This thesis has presented our research into statistical modelling of count data, which frequently occur in informetrics (including scientometrics and bibliometrics) as well as in many other diverse fields. We have reviewed some of the prominent existing models (such as the power law and GIGP models) through the unifying approach based on the item production model. Despite the general and transferable value of such an approach, this thesis has primarily focused on the scientific production use case, comprising authors, their papers, and the corresponding citations.

In our discussion of different models, where possible we have looked into their plausible conceptual justification, exemplified by the principle "success breeds success" for the power law and the mixing probability framework for the GIGP and GPL models. On the other hand, data analysis helps to validate the statistical model and, in particular, review its hypothetical mechanism. Specifically, in the item production model we assumed that sources are independent and produce items with the same distribution, which is clearly a simplification of real situations. Nevertheless, if the goodness-of-fit to real data is reasonable, then the researcher may use the model with confidence; however, if the model does not fit well, this signals that some of the assumptions may not be tenable or may even help detect additional data features such as interpretable outliers (see Sections 6.4.3 and 6.4.4).

Motivated by probabilistic combinatorics and theory of random integer partitions, we looked at the production profiles known as Young diagrams, and their possible scaling limits. Although the Young diagrams are essentially the complementary cumulative plots, focusing on the suitable scales provides useful insights into the data structure, which can be used for modelling and estimation of the production metrics such as h-index and g-index.

Our investigation has shown that, when fitting the power law and the integer partition model to real-life count data, neither of them provide a good match in the entire range of the data. This limitation has prompted us to propose the GPL model by modifying the power law setting, which captures a transition from a relatively flat behaviour at small counts to a power decay at the tail. Although we subsequently discovered that the GPL model had been known in the literature as a hooked power law (Thelwall & Wilson (2014); Shahmandi *et al.* (2020)), our motivation and usage of the GPL were significantly different. Fitting the GPL model to different data sets demonstrates that it performs well, often helping to highlight zero-count frequencies as outliers (like in the EJP and the AMS data). Compared to the GIGP model, the parameter estimation of the GPL model is quite straightforward and does not require complicated calculations, whereas the GIGP model, which is is expressed via the non-elementary Bessel function, leads to significant computational difficulties, so much so that in practice some of the parameters are often predefined a priori (Sichel, 1985).

Furthermore, we have identified the limit shapes in the power law, GIGP, and GPL models under appropriate scaling and proved pointwise convergence of the corresponding Young diagrams. We have also introduced a composite model of the item production model in terms of the multiple batteries of sources producing items and provided the corresponding method of finding the limit shape of the average of the composite model and the way of proving its convergence. This is useful for estimating the average production metrics in a composite model.

To compare different models through real data, we have used the web scraping technique to collect citations of authors' papers from Google Scholar and created the EJP and the AMS data sets. These data sets and other existing data sets from the literature have enabled us to evaluate the performance of different models.

In terms of the production metrics, we have proposed the h_1-index, which combines the properties of the h-index and the g-index and takes into account citations equal to or greater than h_1, while also including all citations from the top-cited papers. Previous research by Yong (2014) estimated the h-index through the limit shape of the integer partitions. In this thesis, we have presented an extension of this approach that provides estimations for the h-index, g-index, and the h_1-index using the limit shape in general. Additionally, numerical examples are provided using the limit shape of the integer partitions model.

In the study of temporal citation data, exploratory data analysis was conducted and survival analysis was applied to investigate the time to the first citation after publication. Various covariates were considered to compare the time to citations after publication in different groups. Additionally, dynamic citations were examined as a point process. Motivated by the phenomenon of "success breeds success" in citation patterns, the Hawkes process was employed to model the citation data due to its self-exciting properties and intensity dependence on the history of citations.

8.2 Specific Contributions

In this section, we summarise the main contributions of the author provided in this thesis (with according references to sections), breaking them down for convenience into the three blocks: (i) methodology, (ii) technical results, and (iii) data collection and analysis.

8.2.1 Methodology

(a) The item production model was presented systematically with a view on applications in informetrics, which included a composite item production model with multiple sources applicable to the authors-papers-citations (APC) relationship (Sections 3.1 and 3.5).

(b) A unified review of some of the existing models (such as power law and GIGP), as well as data sets, was pursued on the basis of the item production model (Sections 4.1–4.4, 5.1 and 5.5).

(c) Motivated by theory of random integer partitions, we developed and advocated the use of the approach based on finding the limit shape based on a selected model and subject to finding suitable scaling parameters (Sections 3.2, 5.2 and 6.2).

(d) Motivated by the h-index and the g-index, the h_1 index was proposed (Section 3.4.3).

(e) The GPL model was proposed by modifying the power law setting (Chapter 6).

(f) Survival analysis was employed to analyse the time-dependent first citation data, grouping papers into different categories based on covariates and comparing the discrepancy in the time taken to receive the first citation after publication (Section 7.2).

(g) Similar to the "success breeds success", the Hawkes processes exhibit a positive dependence on past events. Adapting the self-exciting property, the Hawkes processes are fitted to the time-dependent citation data. In addition, a version of kernels that allows multiple arrivals on continuous time was provided (Section 7.3).

8.2.2 Technical results

(a) The relation $h \leq h_1 \leq g$ was proved (Section 3.4.3).

(b) We provided a model-based estimator for the h-index, g-index and the h_1-index (Section 3.6).

(c) The limit shape of the power law model was found. observing that it is independent of any scaling due to the inherent scale-free nature of the model (Section 3.3).

(d) The limit shape of the GIGP and the GPL model were found with appropriate scalings and the convergence of the limit shape is proved as well as the random fluctuations of random Young diagrams (Sections 5.2, 5.3 and 6.2).

8.2.3 Data collection and analysis

(a) We fitted a regression model to verify the conjecture by Hirsch that the h-index has a square root relation with the number of citations. The regression approach was also applied to the g-index (Section 2.4).

(b) The EJP data and the AMS data were collected using the web scraping technique (data sets D & E in Section 2.2 and Appendix A).

(c) By a systematic analysis of some classical data sets, we discovered some interesting features of the data, such as a departure from the power law fit in Lotka's data for moderate tails, with superiority of the GIGP model (Sections 4.4.1 and 5.5.1).

(d) Analysis based on the GIGP limit shape revealed significant deviations within a certain range of frequencies in Chen's data (Section 7.2).

(e) Through the GPL model fit, an inflated zero count (outlier) was identified in the EJP data, as well as in the the AMS data (Sections 6.4.3 and 6.4.4).

8.3 Future Work

In Lotka's data, the smaller range of the data is the power law and the tail behave differently, which motivated "sewing" other models, such as include the stretched exponential (SE) model. The stretched exponential density is given by

$$f(x) = \frac{\gamma\, x^{\gamma-1}}{x_0^{\gamma}}\, \mathrm{e}^{-(x/x_0)^{\gamma}} \qquad (x \geq x_0), \tag{8.1}$$

where $0 < \gamma < 1$ and x_0 are two parameters of the SE distribution (Laherrere & Sornette, 1998). Combining the stretched exponential model and the power law gives the frequency

$$f_j = C j^{-a_1} \frac{j^{a_2-1}}{L^{a_2}}\, \mathrm{e}^{-(j/L)^{a_2}}, \tag{8.2}$$

which can be simplified to

$$f_j = C j^{-a}\, \mathrm{e}^{-(j/L)^{\gamma}}, \tag{8.3}$$

where C is a normalisation constant, a is the parameter for the power law, and γ and L are the parameters of the (scaled) stretched exponential part. We expect that the PL-SE model on CCDF plot is thinner than the GPL and PL but fatter than GIGP and pure exponential distribution. One may also try to combine the GPL model and SE model if the complementary cumulative frequency plot of data starts with flatter reduction and then the middle part is a straight line, then gets faster decay at the tail.

In survival analysis on citations, if data is accessible, it is worth considering covariates beyond the number of authors and the length of papers. Additional covariates, such as the gender of the (corresponding) author, the journal of publication, or the geographical region, may be considered to look at whether citations significantly differ with respect to these covariates. Some covariates, such as gender, may require manual detection, as the relevant information is not always readily apparent in the data. Adding interaction terms would also be interesting.

Aligned with the GIGP model, it would be worthwhile to explore the incorporation of time dependence within the GPL model. By retaining the limit shape and allowing for temporal variation, the model would be capable of capturing the evolution of the data over time.

Appendix A

Web Scraping: Collecting Profile IDs from Google Scholar

Provided we know the Google Scholar profile ID of a given author, we can use the help of the R package `scholar` to find the number of citations for each of their papers. However, there is no list of profile IDs of authors in R, nor any other resources. So, we need to collect a large number of Google Scholar profile IDs.

Our initial attempt to collect Google Scholar profile IDs involved doing so manually. This entailed recording names of an author from the Electronic Journal of Probability, typing the name into Google Scholar and finally cutting the ID from the web link of the page for this author. Repeating the above process one by one for each author became very time consuming, and so we decided instead to use a web scraping technique in Python and store collections of these IDs. To apply the above mentioned web scraping technique, we needed an input list of names of authors from the same academic field. One may find such a file, for example, by obtaining the records of members of the American Mathematical Society (AMS) from the AMS website. This need not be straightforward — for example, the AMS membership lists are separated by country, and only members of the AMS can see and download the file. Fortunately, due to Dr Leonid Bogachev's AMS membership, we were able to obtain such a list in this instance.

In what follows, we outline the sequence of steps for collecting Google Scholar IDs in Python.

A. Web Scraping: Collecting Profile IDs from Google Scholar

1. First, open the page of the list of authors on Google Scholar; here is the link to the page: `https://scholar.google.com/citations?view_op=search_authors&mauthors={name}&hl=en&oi=ao`. The name author should be in place of {name}. As an example, the name of the author of this thesis Ruheyan Nuermaimaiti is used, the web page we want to visit is `https://scholar.google.com/citations?view_op=search_authors&mauthors={RuheyanNuermaimaiti}&hl=en&oi=ao`. The webpage can be visited by Python using the `requests.get` command in `requests.get` package.

2. To find the resources we are interested in, right-click the page we want to scrape the data from and click "inspect" then we will see a window showing the code of this page. With moving the indicator, the corresponding section of this page is highlighted. Then one can locate the place they are interested in and observe the corresponding code. In our case, we are interested in the Google Scholar profile IDs of authors, it is 12 alphanumeric characters in the web page source listed after "user=". For example, {`<a href="/citations?hl=en&amp;user=hHkPQ4cAAAAJ"> ...</a>`, where "hHkPQ4cAAAAJ" is what we need. This step can be achieved by the Python command `split('user=')` to locate the Google Scholar ID and choose the first 12 characters after "user=" to get the 12 alphanumeric characters.

3. Save the list of Google Scholar IDs in a suitable file, such as a $*$.csv file.

Getting IDs of authors using this web scraping technique significantly reduced the time it took to collect citation data. Ideally, one would input a list of authors, and automatically yield a list of Google Scholar profile IDs which are unique and correctly correspond to the original list of authors. To achieve this objective, however, one must manually overcome various difficulties which occur both before and after web scraping.

We now document some of the aforementioned difficulties with the web scraping technique. In what follows we have listed some of the possible outcomes when we input an authors name, and in sub-lists, we have discussed some possible reasons for these outcomes. As above, ideally, the outcome is one Google Scholar profile ID, which is the correct ID, but this is not always the case. Instead, the following outcomes may be returned:

- No response
  - Some authors on the list have no Google Scholar profile.
  - There can be several choices for how one writes the name of an author. For example, one must choose the order of the first name and the last, whether the first name is abbreviated with initials or not, and whether or not to include middle names or titles. The choices made are often determined by the preference of person inputting the name into Google Scholar.
- One ID
  - This is a correct ID we are expecting.
  - This is an incorrect ID since the listed author in Google Scholar is different to the author from our original input. This can be checked, for example, if these authors work in different academic fields or work for different organisations.
- Several IDs
  - Only one of them is the correct ID. There might exist several authors who have the same name as the input.
  - None of them are the correct ID. Although there might exist several people who have the same name as the input, it is still possible none of them are the person we are looking for.

Thus, in some of the above-listed situations we may have to clean the data manually. Even though this may involve some manual processes, this web scraping technique is still more efficient than collecting the IDs purely manually.

Appendix B

Asymptotic Formulas for the Bessel Function

The following is a list of useful properties of the Bessel function $K_\nu(z)$, including some asymptotic formulas under various regimes for the argument z and the order ν. For ease of use, we collect these facts here, with reference to the NIST handbook (Olver *et al.*, 2010).

Lemma B.1 (Olver *et al.* (2010), 10.27.3)**.** *For any ν and z,*

$$K_{-\nu}(z) = K_\nu(z). \tag{B.1}$$

Lemma B.2. *Let ν be fixed and $z \to 0+$.*

(a) (Olver *et al.* (2010), 10.30.2) *If $\nu > 0$ then*

$$K_\nu(z) \sim \tfrac{1}{2}\Gamma(\nu)\left(\tfrac{1}{2}z\right)^{-\nu}. \tag{B.2}$$

(b) (Olver *et al.* (2010), 10.31.1 with the aid of 10.25.2) *If $\nu = 1$ then*

$$K_1(z) = z^{-1} + \tfrac{1}{2}z\log z + O(z). \tag{B.3}$$

(c) (Olver *et al.* (2010), 10.31.2 with the aid of 10.25.2) *If $\nu = 0$ then*

$$K_0(z) = -\log\left(\tfrac{1}{2}z\right) - \gamma + O(z^2\log z). \tag{B.4}$$

where $\gamma = 0.5772\ldots$ is Euler's constant (Olver et al. *(2010), 5.2.3). In particular,*

$$K_0(z) \sim -\log z. \tag{B.5}$$

(d) (Olver *et al.* (2010), 10.27.4 and 10.25.2 with the aid of 5.5.3) *For* $-1 < \nu < 0$,

$$K_\nu(z) = \tfrac{1}{2}\Gamma(-\nu)\left(\tfrac{1}{2}z\right)^{\nu} + \frac{\Gamma(\nu+1)}{2\nu}\left(\tfrac{1}{2}z\right)^{-\nu} + O(z^{\nu+2}). \tag{B.6}$$

Lemma B.3 (Olver *et al.* (2010), 10.41.2)**.** *If* $z \neq 0$ *is fixed and* $\nu \to +\infty$, *then*

$$K_\nu(z) \sim \sqrt{\frac{\pi}{2\nu}}\left(\frac{\mathrm{e}z}{2\nu}\right)^{-\nu}. \tag{B.7}$$

References

Arratia, R., Barbour, A.D. & Tavaré, S. (2003). *Logarithmic Combinatorial Structures: A Probabilistic Approach*. European Mathematical Society. 42, 44

Auluck, F. & Kothari, D. (1946). Statistical mechanics and the partitions of numbers. *Mathematical Proceedings of the Cambridge Philosophical Society*, **42**, 272–277. 42

Barbour, A.D., Holst, L. & Janson, S. (1992). *Poisson Approximation*. Clarendon Press, Oxford University Press. 93

Bogachev, L.V. (2015). Unified derivation of the limit shape for multiplicative ensembles of random integer partitions with equiweighted parts. *Random Structures and Algorithms*, **47**, 227–266. 39, 43, 44, 45

Bogachev, L.V., Gnedin, A.V. & Yakubovich, Y.V. (2008). On the variance of the number of occupied boxes. *Advances in Applied Mathematics*, **40**, 401–432. 44, 45

Bogachev, L.V., Nuermaimaiti, R. & Voss, J. (2023). Limit shape of the generalized inverse Gaussian-Poisson distribution. Preprint, available online: `https://arxiv.org/abs/2303.08139`. i, 77

Bonnell, J. & Ogihara, M. (2023). *Exploring Data Science with R and the Tidyverse: A Concise Introduction*. Chapman and Hall/CRC. Preprint available online: `https://ds4world.cs.miami.edu/text-analysis.html`. 13

Borisov, I. & Jetpisbaev, M. (2022). Poissonization principle for a class of additive statistics. *Mathematics*, **10**, 4084. 44

References

Browning, R., Sulem, D., Mengersen, K., Rivoirard, V. & Rousseau, J. (2021). Simple discrete-time self-exciting models can describe complex dynamic processes: A case study of COVID-19. *PLoS ONE*, **16**, 83–90. 143, 145

Chen, C.C. (1972). The use patterns of physics journals in a large academic research library. *Journal of the American Society for Information Science*, **23**, 254–270. 11, 97

Clauset, A., Shalizi, C.R. & Newman, M.E. (2009). Power-law distributions in empirical data. *SIAM Review*, **51**, 661–703. 1, 2, 3, 4, 62, 65, 66, 67, 68, 95

Coile, R.C. (1977). Lotka's frequency distribution of scientific productivity. *Journal of the American Society for Information Science*, **28**, 366–370. 2

Collett, D. (2015). *Modelling Survival Data in Medical Research*. Chapman and Hall/CRC Press, 3rd edn. 128, 135

Cox, D.R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, **34**, 187–202. 135

Daley, D.J. & Vere-Jones, D. (2003). *An Introduction to the Theory of Point Processes, Volume I: Elementary Theory and Methods*. Springer, 2nd edn. 140

Duchon, P., Flajolet, P., Louchard, G. & Schaeffer, G. (2004). Boltzmann samplers for the random generation of combinatorial structures. *Combinatorics, Probability and Computing*, **13**, 577–625. 42

Egghe, L. (2005). *Power Laws in the Information Production Process: Lotkaian Informetrics*. Elsevier. 1, 2, 29, 102

Egghe, L. (2006). Theory and practise of the g-index. *Scientometrics*, **69**, 131–152. 10, 25, 47

Egghe, L. & Rao, I.R. (2001). Theory of first-citation distributions and applications. *Mathematical and Computer Modelling*, **34**, 81–90. 121

Egghe, L. & Rousseau, R. (1995). Generalized success-breeds-success principle leading to time-dependent informetric distributions. *Journal of the American Society for Information Science*, **46**, 426–445. 3, 102

Egghe, L. & Rousseau, R. (2006). An informetric model for the Hirsch-index. *Scientometrics*, **69**, 121–129. 68

Egghe, L. *et al.* (1992). Citation age data and the obsolescence function: Fits and explanations. *Information Processing & Management*, **28**, 201–217. 121

Eriksson, K. & Sjöstrand, J. (2012). Limiting shapes of birth-and-death processes on Young diagrams. *Advances in Applied Mathematics*, **48**, 575–602. 121

Fill, J.A. (1988). Bounds on the coarseness of random sums. *Annals of Probability*, **16**, 1644–1664. 126

Fristedt, B. (1993). The structure of random partitions of large integers. *Transactions of the American Mathematical Society*, **337**, 703–735. 44, 45

Galassi, M., Davies, J., Theiler, J., Gough, B., Jungman, G., Alken, P., Booth, M., Rossi, F. & Ulerich, R. (2002). *GNU Scientific Library*. Network Theory Limited Godalming. 62

Garfield, E. (1955). Citation indexes for science: A new dimension in documentation through association of ideas. *Science*, **122**, 108–111. 10

Garthwaite, P.H., Jolliffe, I.T. & Jones, B. (2002). *Statistical Inference*. Oxford University Press, 2nd edn. 34, 35

Gillespie, C.S. (2015). Fitting heavy tailed distributions: The poweRlaw package. *Journal of Statistical Software*, **64**, 1–16. 12, 68

Gnedin, A., Hansen, B. & Pitman, J. (2007). Notes on the occupancy problem with infinitely many boxes: general asymptotics and power laws. *Probability Surveys*, **4**, 146–171. 31, 44, 45

Gupta, R.C. & Ong, S. (2005). Analysis of long-tailed count data by Poisson mixtures. *Communications in Statistics – Theory and Methods*, **34**, 557–573. 3

Gyarmati-Szabó, J., Bogachev, L. & Chen, H. (2011). Modelling threshold exceedances of air pollution concentrations via non-homogeneous Poisson processes with multiple change-points. *Atmospheric Environment*, **45**, 5493–5503. 143

Hawkes, A.G. (1971). Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, **58**, 83–90. 143

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**, 16569–16572. 10, 21, 22, 23, 24

Hirsch, J.E. (2007). Does the h index have predictive power? *Proceedings of the National Academy of Sciences of the United States of America*, **104**, 19193–19198. 22

Hirsch, J.E. (2019). h_α: An index to quantify an individual's scientific leadership. *Scientometrics*, **118**, 673–686. 23

Huber, J.C. (2002). A new model that generates Lotka's law. *Journal of the American Society for Information Science and Technology*, **53**, 209–219. 3, 102

Johnson, N.L., Kotz, S. & Balakrishnan, N. (1994). *Continuous Univariate Distributions, Volume 2*. John Wiley & Sons, 2nd edn. 3, 61, 78

Johnson, N.L., Kemp, A.W. & Kotz, S. (2005). *Univariate Discrete Distributions*. John Wiley & Sons, 3rd edn. 3, 64, 80, 81

Karlin, S. (1967). Central limit theorems for certain infinite urn schemes. *Journal of Mathematics and Mechanics*, **17**, 373–401. 113, 114

Ke, Q., Ferrara, E., Radicchi, F. & Flammini, A. (2015). Defining and identifying Sleeping Beauties in science. *Proceedings of the National Academy of Sciences of the United States of America*, **112**, 7426–7431. 126

Krapivsky, P. (2021). Stochastic dynamics of growing Young diagrams and their limit shapes. *Journal of Statistical Mechanics: Theory and Experiment*, **2021**, 013206. 121

Laherrere, J. & Sornette, D. (1998). Stretched exponential distributions in nature and economy: "fat tails" with characteristic scales. *The European Physical Journal B – Condensed Matter and Complex Systems*, **2**, 525–539. 3, 97, 119, 153

Lotka, A.J. (1926). The frequency distribution of scientific productivity. *Journal of the Washington Academy of Sciences*, **16**, 317–323. 2, 11, 27, 69, 95

Martín-Martín, A., Orduna-Malea, E., Thelwall, M. & López-Cózar, E.D. (2018). Google Scholar, Web of Science, and Scopus: A systematic comparison of citations in 252 subject categories. *Journal of Informetrics*, **12**, 1160–1177. 19

McCoy, B.M., Tracy, C.A. & Wu, T.T. (1977). Painlevé functions of the third kind. *Journal of Mathematical Physics*, **18**, 1058–1092. 127

Nane, T. (2015). Time to first citation estimation in the presence of additional information. In *The 15th International Conference on Scientometrics and Informetrics, Instanbul, Turkey*, 249–260. 136

Nicholls, P.T. (1987). Estimation of Zipf parameters. *Journal of the American Society for Information Science*, **38**, 443–445. Errata: *Ibidem*, **39** (1988), 287. 2, 3, 64, 66

Novak, S. (2019). Poisson approximation. *Probability Surveys*, **16**, 228–276. 93

Nuermaimaiti, R., Bogachev, L.V. & Voss, J. (2021). A generalized power law model of citations. In *18th International Conference on Scientometrics and Informetrics, ISSI 2021, Leuven, Belgium*, 843–848, International Society for Scientometrics and Informetrics. i, iv, 5, 39, 99

Olver, F.W., Lozier, D.W., Boisvert, R.F. & Clark, C.W. (2010). *NIST Handbook of Mathematical Functions*. Cambridge University Press. 44, 58, 61, 62, 78, 79, 80, 81, 83, 84, 86, 89, 103, 159, 160

Painlevé, P. (1902). Sur les équations différentielles du second ordre et d'ordre supérieur dont l'intégrale générale est uniforme. (French) [On differential equations of the second order and of higher order whose general integral is uniform]. *Acta Mathematica*, **25**, 1–85. 126

Pennock, D.M., Flake, G.W., Lawrence, S., Glover, E.J. & Giles, C.L. (2002). Winners don't take all: Characterizing the competition for links on the web. *Proceedings of the national academy of sciences*, **99**, 5207–5211. 5, 119

Pittel, B. (1997). On a likely shape of the random Ferrers diagram. *Advances in Applied Mathematics*, **18**, 432–488. 4

Price, D.D.S. (1976). A general theory of bibliometric and other cumulative advantage processes. *Journal of the American Society for Information Science*, **27**, 292–306. 3, 102

Price, D.J.D.S. (1965). Networks of scientific papers. *Science*, **149**, 510–515. 2, 3, 10

Pritchard, A. (1969). Statistical bibliography or bibliometrics. *Journal of Documentation*, **25**, 348–349. 9

Radicchi, F., Fortunato, S. & Castellano, C. (2008). Universality of citation distributions: Toward an objective measure of scientific impact. *Proceedings of the National Academy of Sciences of the United States of America*, **105**, 17268–17272. 23

Redner, S. (1998). How popular is your paper? An empirical study of the citation distribution. *The European Physical Journal B – Condensed Matter and Complex Systems*, **4**, 131–134. 2

Rizoiu, M.A., Lee, Y., Mishra, S. & Xie, L. (2018). Hawkes processes for events in social media. In S.F. Chang, ed., *Frontiers of Multimedia Research*, 191–218, Association for Computing Machinery (ACM) and Morgan & Claypool. 143

Rousseau, R. (2002). Lack of standardisation in informetric research. Comments on "Power laws of research output. Evidence for journals of economics" by Matthias Sutter and Martin G. Kocher. *Scientometrics*, **55**, 317–327. 2

Shahmandi, M., Wilson, P. & Thelwall, M. (2020). A new algorithm for zero-modified models applied to citation counts. *Scientometrics*, **125**, 993–1010. 5, 119, 150

Shiryaev, A.N. (1996). *Probability*. Springer, 2nd edn. 40, 42, 51, 91, 107

Sichel, H.S. (1971). On a family of discrete distributions particularly suited to represent long-tailed frequency data. In *Proceedings of the Third Symposium on Mathematical Statistics*, 51–97, Council for Scientific and Industrial Research (CSIR): Pretoria. 3, 77, 78, 102

Sichel, H.S. (1973). Statistical valuation of diamondiferous deposits. *Journal of the Southern African Institute of Mining and Metallurgy*, **73**, 235–243. 3, 4

Sichel, H.S. (1974). On a distribution representing sentence-length in written prose. *Journal of the Royal Statistical Society: Series A (General)*, **137**, 25–34. 3, 4, 5

Sichel, H.S. (1975). On a distribution law for word frequencies. *Journal of the American Statistical Association*, **70**, 542–547. 3, 4

Sichel, H.S. (1982). Repeat-buying and the generalized inverse Gaussian-Poisson distribution. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **31**, 193–204. 3

Sichel, H.S. (1985). A bibliometric distribution which really works. *Journal of the American Society for Information Science*, **36**, 314–321. iv, 3, 4, 77, 79, 95, 97, 102, 119, 121, 150

Thelwall, M. & Wilson, P. (2014). Distributions for cited articles from individual subjects and years. *Journal of Informetrics*, **8**, 824–839. 5, 99, 119, 150

Therneau, T.M. (2022). *A Package for Survival Analysis in R*. R package version 3.4-0. 133

Vere-Jones, D. & Ozaki, T. (1982). Some examples of statistical estimation applied to earthquake data: I. Cyclic Poisson and self-exciting models. *Annals of the Institute of Statistical Mathematics*, **34**, 189–207. 143

Vershik, A.M. (1995). Asymptotic combinatorics and algebraic analysis. In S. Chatterji, ed., *Proceedings of the International Congress of Mathematicians (Zürich, 1994), Vol. 2*, 1384–1394, Birkhäuser. 40

Vershik, A.M. (1996). Statistical mechanics of combinatorial partitions, and their limit shapes. *Functional Analysis and Its Applications*, **30**, 90–105. 4, 40, 43

Vershik, A.M. (1997). Limit distribution of the energy of a quantum ideal gas from the viewpoint of the theory of partitions of natural numbers. *Russian Mathematical Surveys*, **52**, 379–386. 42

Wang, D., Song, C. & Barabási, A.L. (2013). Quantifying long-term scientific impact. *Science*, **342**, 127–132. 121

Whitaker, L. (1914). On the Poisson law of small numbers. *Biometrika*, **10**, 36–71. 93

Wikipedia (2023). "Emil Abderhalden", `https://en.wikipedia.org/wiki/Emil_Abderhalden` (accessed March 10, 2023). 97

Yong, A. (2014). Critique of Hirsch's citation index: A combinatorial Fermi problem. *Notices of the American Mathematical Society*, **61**, 1040–1050. 4, 10, 23, 24, 46, 56, 151

Yu, G., Keirstead, J. & Jefferi, G. (2016). *scholar: Analyse citation data from Google Scholar*. R package, version 0.1.5. 14, 122