Centrum voor Wiskunde en Informatica
Centre for Mathematics and Computer Science

A.P. van der Plas

A semiparametric model for citation counts

Department of Mathematical Statistics Report MS-R8814 September

The Centre for Mathematics and Computer Science is a research institute of the Stichting Mathematisch Centrum, which was founded on February 11, 1946, as a nonprofit institution aiming at the promotion of mathematics, computer science, and their applications. It is sponsored by the Dutch Government through the Netherlands Organization for the Advancement of Pure Research (Z.W.O.).

A Semiparametric Model for Citation Counts

Adriaan P. van der Plas
Centre for Mathematics and Computer Science
P.O. Box 4079, 1009 AB Amsterdam, The Netherlands

The pattern of yearly citation counts that a paper in a certain scientific field receives expresses in some way the importance of that paper. In this paper this pattern is modelled as a discrete time stochastic process with a conditional distribution dependent on a few parameters characterzing that scientific field.
For a particular dataset we estimate the parameters and we perform several tests of the model, which seems to fit the data quite well.
Finally several possible improvements of the model and applications to scientometry are discussed.

1980 Mathematics Subject Classification: 62P25, 62M99.
Key words and phrases: citation counts, expected citation patterns, mixture of distributions, semi-parametric model, identification, bootstrap.

0. Preface

This research on citation counts of scientific publications resulted from a co-operation with H.F. Moed and A.F.J. van Raan of the Research Policy Unit of the University of Leiden.

1. Introduction

1.1.

The pattern of the series of yearly citation counts that a paper in a certain scientific field receives, expresses in some way the importance of that paper. Therefore it is important to analyse and describe this pattern. Since it is reasonable to assume that there is arbitrariness in the series of citation counts one should give such a description by a (discrete time) stochastic process.

The aim of this paper is to model the series of (yearly) citation counts of a paper from a certain scientific field as a discrete time process with a conditional distribution dependent on a few parameters characterizing that scientific field.

1.2.

Although several authors tried to model such series of citation counts by a discrete time stochastic process, it seems to us that either their assumptions made are too simple, or no explicit reference to time is made. We recall the papers of Chang (1975) and De Solla Price (1976).

Chang modelled the series of citation counts by an inhomogeneous Poisson process in discrete time: the yearly number of citations has a Poisson distribution with mean decreasing exponentially fast as time proceeds. Thus between different years the citation counts are independent, which seems to be unrealistic. (In the case of so called key papers the process should be homogeneous as Chang claims, but then the field of application is very limited.)

De Solla Price tries to model the citation counts as a (stochastic) pure birth process, but 'without explicit reference to time as a variable' as he remarks (see p. 303 of De Solla Price (1976)). So the dependence structure of the process over time is in fact not analysed. It should be noted however that his aim was to explain the marginal stationary distribution of citation counts.

Report MS-R8814
Centre for Mathematics and Computer Science
P.O. Box 4079, 1009 AB Amsterdam, The Netherlands

1.3.

We propose a model that satisfies some obvious heuristic requirements.

Set the publication year of a paper equal to zero. Denote by N_t the number of citations, also called the citation count, received by a scientific paper in year $t+1$ after publication and let $S_t = N_0 + N_1 + \cdots + N_t$ for $t=0,1,2,....$

Heuristic reasoning leads to the following requirements that a model for citation counts of a paper should minimally satisfy.

1.3.1.

The yearly number of citations tends to increase in the first years after publication and to decrease afterwards. More precisely, there exists a positive integer τ such that the expected number of citations $\mathbb{E}N_t$ increases for t increasing and $t \leqslant \tau$ and decreases for t increasing and $t > \tau$.

The series $(\mathbb{E}N_0, \mathbb{E}N_1, ...)$ is often called the expected citation pattern. We will call the series $(N_0, N_1, ...)$ the citation pattern of a paper.

1.3.2.

Only a small number of publications within a particular scientific field will during a long time period receive each year a positive number of citations, so one should discriminate between two possible citation patterns of a paper. In the following the type of the pattern will be indicated by a dummy stochastic variable Z. Note that Z is not known beforehand, but that the history of citations will reveal the citation pattern of a paper and thus Z as time t tends to infinity.

It should be clear that a publication receiving each year a small, but positive number of citations could belong to the small group of papers positively cited during a long time period, while a publication that receives a big total amount of citations but with a citation pattern that decreases rapidly to zero in the tail, will be expected to belong to the other group. We mention here that the small group of papers exhibiting a pattern with a long positive tail were called 'key papers' by CHANG (1975).

1.3.3.

The citation count N_t received by a paper at time t depends on the series of citation counts of that paper received before time t in such a way that 'succes breeds success'. More specifically we assume that the conditional expectation of N_t given the history of citation counts $(N_0, N_1, \ldots, N_{t-1})$ and given Z is a positive real valued function of S_{t-1} and Z for $t=1,2,...$, increasing in S_{t-1} and Z for fixed t.

Note that given Z this conditional expectation of the citation count at t depends on the sum of all previous citation counts, i.e. all previous citations have the same impact on the conditional expected citation count at t. One may object against such a simplifying assumption. However, since the total length of each available series of citation counts is in most cases relatively short (about 10 years), this assumption seems to be reasonable.

1.4.

The above mentioned requirements (1.3.1) and (1.3.3) allow for a first rudimentary statistical analysis. For all the statistical analysis in this paper we will use a dataset of the Subfaculty of Chemistry of the University of Leiden, consisting of citation counts of 320 publications over a time period of eleven years. (For a short description of the dataset see MOED et al. (1985) and see MOED et al. (1983) for an extensive description.)

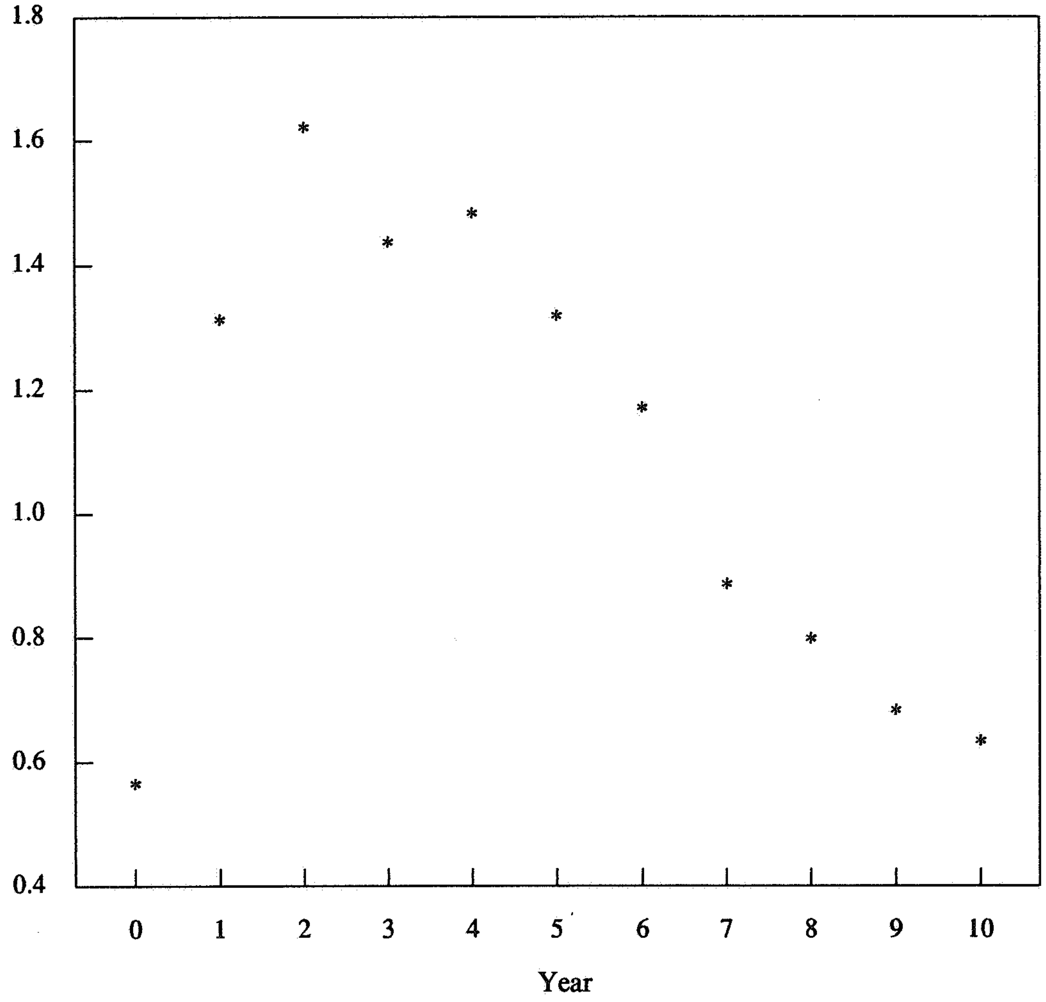
In Table (1.4.1) a random sample of 40 publications from this dataset is shown. We observe that most time series of citation counts have a very irregular pattern and that many zero counts occur.

TABLE 1.4.1. Forty series of citation counts

					Year					
0	1	2	3	4	5	6	7	8	9	10
0	5	0	1	1	2	0	1	2	1	1
0	2	5	4	3	2	1	1	2	0	6
0	2	2	0	0	0	0	0	1	0	0
0	0	0	1	0	0	0	0	0	0	0
0	0	0	0	2	1	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0
0	0	1	1	0	0	0	1	0	0	1
0	0	0	1	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	1	0	0
0	0	0	2	2	0	1	2	0	0	1
0	0	0	0	0	0	0	0	0	0	0
1	0	1	2	2	0	2	0	0	2	2
1	2	3	1	4	1	1	0	1	0	0
0	0	0	3	2	1	0	2	0	0	0
0	1	1	1	2	1	1	0	2	1	2
0	3	6	1	2	2	0	0	0	0	0
2	1	4	5	3	2	2	0	0	2	1
1	0	1	0	1	0	0	0	1	0	0
0	3	6	4	8	3	0	1	2	0	1
1	1	4	7	6	8	7	3	6	8	5
0	1	4	0	0	3	1	1	0	2	0
0	1	2	4	2	1	3	1	1	2	0
0	1	1	1	1	0	1	0	1	0	0
0	0	0	0	0	0	0	0	0	1	0
0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	1	0	0	0	0	0
0	1	0	0	0	0	0	1	2	0	0
0	0	0	0	0	0	0	1	0	1	0
0	0	0	0	0	0	0	0	0	0	0
1	4	1	0	2	2	2	0	0	2	0
3	12	15	13	18	18	20	14	10	13	8
0	0	1	1	1	1	0	2	0	0	0
0	2	0	3	1	1	1	2	2	0	3
1	0	0	0	1	0	0	0	1	3	1
0	0	2	0	0	0	1	1	1	0	0
0	2	1	1	0	1	0	1	1	0	1
3	0	3	6	4	2	2	1	1	1	0
1	0	0	0	1	1	1	0	0	0	0
2	0	0	1	1	1	1	0	1	0	1
0	1	0	0	0	0	0	0	0	1	0

If we assume that the time series of citation counts are stochastically independent, then the average citation count at time t is a good estimate for the expectation of the number of citations at t for $t=0,1,2,...$. Calculating these averages with our sample of 320 publications the expected pattern of requirement (1.3.1) is seen to be confirmed for $t \geqslant 3$ (see Figure (1.4.2)).

FIGURE 1.4.2. The average citation pattern



Requirement (1.3.3) implies a positive correlation between N_t and S_{t-1} for $t=1,2,...$. Calculation of the sample correlation ρ_t between N_t and S_{t-1} shows a positive correlation for all $t;t=1,2,\dots,10$ (see Table (1.4.3)).

TABLE 1.4.3. Correlation ρ_t between N_t and S_{t-1}

t	1	2	3	4	5	6	7	8	9	10
ρ_t	.3863	.6432	.6991	.7413	.7361	.7468	.7436	.7459	.6919	.6739

A rudimentary statistical analysis of requirement (1.3.2) is possible, but quite difficult. Of course $(t+1)^{-1} \cdot S_t$ should be a good indicator for Z, but unfortunately the length of the series of citation counts for our dataset are too short.

1.5.

We will now come to some comments of a theoretical and/or practical nature in modelling series of yearly citation counts.

1.5.1.
For several reasons it seems impossible to construct a parametric model for the citation counts received by a paper in the first years after publication, i.e. the so called 'head' of the series of citation counts. We mention that the real date of publication of a paper is not known (see also APPENDIX (7.2)) and that some authors of papers are more well-known then others within a scientific community, which may especially influence the citation counts of a paper in the first years after publication. Note also that requirement (1.3.1) is only confirmed for $t \geqslant 3$ for our dataset. So, the statistical model for the series of citation counts received by a paper should be of the semi-parametric type, i.e. a statistical model with a non-parametric part for the head and a parametric part for the rest of the pattern. (For a general overview of semi-parametric models see WELLNER (1985)).

1.5.2.
Observing that the probability of one citation of a paper is small for each year and the number of possible citers is large we are now ready to formulate the parametric part of a statistical model that satisfies the requirements (1.3.1) to (1.3.3) and the above remarks.

Let $\mu,\alpha,\beta>0$ and suppose for given t_0 (e.g. $t_0=3$) that $(N_0,N_1,\ldots,N_{t_0})$ with an arbitrary probability distribution forms the head of the citation pattern, then recursively for $t=t_0+1,t_0+2,\ldots$ the number of citations N_t given the history of citation counts $(N_0,N_1,\ldots,N_{t-1})$ and Z has a (conditional) Poisson distribution with expectation

$$\lambda_t(Z)=\mu Z+\alpha \exp(-\beta t)\cdot\sum_{s=0}^{t-1} N_s .$$

Note that for $t>t_0$ the conditional probability distribution of N_t given $(N_0,N_1,\ldots,N_{t-1})$ and Z is equal to the conditional probability distribution of N_t given $S_{t-1}=\sum_{s=0}^{t-1} N_s$ and Z, i.e. given Z all previous citations have the same impact on the probability distribution of N_t. (See also (1.3.3).)

1.6.
This paper is further organized as follows. In Chapter 2 we formulate the model, discuss several probabilistic properties and make several remarks on the identification of the parameters. In Chapter 3 the estimation procedure is described and the results are given. Chapter 4 deals with testing of the model. In Chapter 5 several methods to improve the model are discussed. Finally we briefly discuss some applications to scientometry in Chapter 6.

2. THE MODEL

2.1.
In this chapter we give a mathematical formulation of the model and give a precise reformulation of the contents in the requirements (1.3.1) and (1.3.2) of the INTRODUCTION. These requirements are consequences of the constructed model, the interested reader is refered to VAN DER PLAS (1988) for proofs. In section (2.5) we mention that the parameters are identified and in section (2.6) we give some formulas that we need for later computations.

2.2. Construction of the stochastic process $\mathbf{N}$
Set the publication year of a scientific paper equal to zero. Denote by N_t the number of citations received by a scientific paper in year $t+1$ after publication for $t=0,1,2,\ldots$. We will interpret the publication of a paper as one citation in year one, i.e. $N_0 \geqslant 1$. We consider a stochastic process of citation counts $\mathbf{N}=\{N_t;t=0,1,2,\ldots\}$ with values in the non-negative integers for each t.

Let $\mathbf{N}_t=(N_0,N_1,\ldots,N_t)$ for $t=0,1,2,\ldots$, i.e. the history of the stochastic process of citation counts up to and including time t.

Suppose that for each particular scientific field there exists a fixed $t_0(t_0>0)$ such that the configuration $\mathbf{N}_{t_0}$ constitutes the head of the citation pattern of a paper from that scientific field. Let $\mathbf{N}_{t_0}$ have probability distribution F such that the mathematical expectation $\mathbb{E}(N_0+N_1+...+N_{t_0})<\infty$ and such that $N_0 \geqslant 1$ with probability 1 (w.p. 1).

Let Z be an unobservable zero-one stochastic variable with $\mathbb{P}[Z=1]=\epsilon$ for fixed $\epsilon \in (0,1)$.

We assume that Z and $\mathbf{N}_{t_0}$ are independent, i.e. the conditional probability

$$\mathbb{P}[Z=1 \mid \mathbf{N}_{t_0}]=\mathbb{P}[Z=1]=\epsilon . \tag{2.2.1}$$

Finally let for $t=t_0+1,t_0+2,...$ the conditional distribution of N_t given $\mathbf{N}_{t-1}$ and Z be Poisson-$\lambda_t(Z)$, where

$$\lambda_t(Z)=\mu Z+\alpha \exp(-\beta t) \cdot \sum_{s=0}^{t-1} N_s \quad \text{for } \mu,\alpha,\beta>0 . \tag{2.2.2}$$

We have constructed the stochastic process $\mathbf{N}$, since we are now able to calculate the probability distribution of $\mathbf{N}_t$ for each t for given probability distribution F of $\mathbf{N}_{t_0}$ and for given parameter $\theta=(\epsilon,\mu,\alpha,\beta)$.

For $n_s=0,1,2,...$ if $s \geqslant 1$, $n_0=1,2,...$ and setting $\mathbf{n}_t=(n_0,n_1,\ldots,n_t)$ for $t=0,1,2,...$ we have

$$\begin{aligned}\mathbb{P}_{\theta,F}[\mathbf{N}_t=\mathbf{n}_t]&=\mathbb{P}_F[\mathbf{N}_{t_0}=\mathbf{n}_{t_0}] \cdot \mathbb{P}_{\theta,F}[\mathbf{N}_t=\mathbf{n}_t \mid \mathbf{N}_{t_0}=\mathbf{n}_{t_0}]\\ &=\mathbb{P}_F[\mathbf{N}_{t_0}=\mathbf{n}_{t_0}]\sum_{z=0}^{1}\mathbb{P}_\epsilon[Z=z \mid \mathbf{N}_{t_0}=\mathbf{n}_{t_0}]\mathbb{P}_{(\mu,\alpha,\beta)}[\mathbf{N}_t=\mathbf{n}_t \mid \mathbf{N}_{t_0}=\mathbf{n}_{t_0} \wedge Z=z] \qquad (2.2.3)\\ &=\mathbb{P}_F[\mathbf{N}_{t_0}=\mathbf{n}_{t_0}]\sum_{z=0}^{1}\mathbb{P}_\epsilon[Z=z]\prod_{s=t_0+1}^{t}\mathbb{P}_{(\mu,\alpha,\beta)}[N_s=n_s|\mathbf{N}_{s-1}=\mathbf{n}_{s-1}\wedge Z=z]\end{aligned}$$

where

$$\mathbb{P}_{(\mu,\alpha,\beta)}[N_s=n_s \mid \mathbf{N}_{s-1}=\mathbf{n}_{s-1} \wedge Z=z] = \exp\{-\lambda_s(z)\} \cdot \frac{\{\lambda_s(z)\}^{n_s}}{n_s!} \quad \text{for } z=0,1 . \tag{2.2.4}$$

2.3. Some remarks on the model

2.3.1.

The assumption of the independence of the zero-one random variable Z and the stochastic vector $\mathbf{N}_{t_0}$, and therefore of Z and $N_0+N_1+...+N_{t_0}$ is really a simplification and heuristically seen to be unrealistic, since one expects that the sum of the number of citations in the head of the pattern will be positively correlated with Z. Nevertheless to avoid serious complications in the statistical analysis of the model we will stick to this assumption (see also chapter 5).

2.3.2.

In our model five parameters occur: the distribution function F of $\mathbf{N}_{t_0}$,$\epsilon \in (0,1)$ and the positive real valued parameters μ,α and β. Such a statistical model is called 'semi-parametric' since the parameter F is 'non-parametric'. (See WELLNER (1985) for an overview of semi-parametric models).

From a statistical point of view the parameters of interest are ϵ,μ,α and β. The estimation of μ,α, and β is made easy by the assumption that the random variable Z and the random vector $\mathbf{N}_{t_0}$ are independent, implying that we may work conditionally on $\mathbf{N}_{t_0}$. The infinite dimensional parameter F then disappears from the probability density of the data.

2.3.3.
From the above construction of the stochastic process N it should be clear that the conditional probability distribution of N_t given $\mathbf{N}_{t-1}$ is a mixture of (conditional) Poisson distributions for $t>t_0$. This observation greatly simplifies the solution of the identification problem of the parameters ϵ,μ,α and β, since the mixing distribution of a mixture of Poisson distributions is identified. (See FELLER (1943) or for a more general treatment of identification problems of mixtures TEICHNER (1961).)

Note that for $t=t_0+1$ and given $\epsilon\in(0,1)$ the mixing distribution is given by $\mathbb{P}[Z=1]=\epsilon$ and that for $t=t_0+1,t_0+2,...$ and given ϵ,μ,α and β the mixing distribution may be computed by the formulas (2.6.5), (2.6.6) and (2.6.7) given below.

2.3.4.
Since $N_0\geqslant 1$ we have $N_0+N_1+...+N_t\geqslant 1$ for all t and therefore $\lambda_t(Z)>0$ for $t>t_0$, implying that the conditional probability that $N_t=0$ given $(\mathbf{N}_{t-1},Z)$ is strictly smaller then 1, i.e. $\mathbb{P}[N_t=0\,|\,\mathbf{N}_{t-1},Z]<1$ for $t>t_0$. Apart from theoretical reasons to take N_0 at least positive, i.e. publication of a scientific paper should be rewarded in some sense, the mathematical reason should now be obvious. For suppose $N_0=0$ with positive probability, then $N_0+N_1+...+N_{t_0}=0$ with positive probability and in that case λ_{t_0+1} takes the value 0 for $Z=0$, implying that $N_{t_0+1}=0$ w.p.1. on $\{Z=0,\ N_0+N_1+...+N_{t_0}=0\}$. Iterating we find that $N_t=0$ w.p.1. for $t>t_0$ on $\{Z=0,\ N_0+N_1+...+N_{t_0}=0\}$.

2.4. Mathematical formulation of the requirements (1.3.1) and (1.3.2)
Let the conditional expectation of $N_t-\mu Z$ given Z be denoted by $g_t(Z)$, i.e. $g_t(Z)=\mathbb{E}[N_t-\mu Z\,|\,Z]$ for $t=0,1,2,....$ We have the following results:

2.4.1.
For each Z there exists a positive finite $C=C(Z)$ such that

$$g_t(Z)<C\ t\ \exp\,(-\beta t)$$

This result shows that $\lim_{t\to\infty}\mathbb{E}[N_t\,|\,Z]=\mu Z$ w.p.1. exponentially fast.

2.4.2.
For each Z there is a time $\tau(Z)$ such that $g_t(Z)$ is increasing in t for $t_0<t\leqslant\tau(Z)$ and decreasing in t for $t>\tau(Z)$. ($\tau(Z)$ may equal t_0+1, but is finite.)

If we define $g_t=\mathbb{E}g_t(Z)=\mathbb{E}[N_t-\mu Z]$ then we have an unconditional, analogous result for g_t as described in (2.4.1) with Z replaced by $\epsilon=\mathbb{P}[Z=1]$. Furthermore there is a finite time τ such that g_t is increasing in t for $t_0<t\leqslant\tau$ and decreasing in t for $t>\tau$. These unconditional results concerning g_t were roughly spoken confirmed for $t>2$ by the statistical analysis in (1.4) of the Introduction.

2.4.3.
For given $\sum_{s\leqslant t_0}g_s(Z)$ and given (μ,α,β) one can calculate $\tau(Z)$; similarly given $\sum_{s\leqslant t_0}g_s$ and given $(\epsilon,\mu,\alpha,\beta)$ one can calculate τ. For $t>t_0$ we have

$$\mathbb{E}[N_t\,|\,\mathbf{N}_{t-1},Z]=\mu Z+\ \exp\,(-\beta t)\sum_{s=0}^{t-1}N_s$$

and so

$$\mathbb{E}[N_t\,|\,Z]=\mu Z+\ \exp\,(-\beta t)\sum_{s=0}^{t-1}\mathbb{E}[N_s\,|\,Z]\,,$$

or

$$g_t(Z)= \exp(-\beta t)\sum_{s=0}^{t-1}\{g_s(Z)+\mu Z\}\ ,$$

yielding for $t>t_0+1$ the recursion formula

$$\begin{aligned} g_t(Z) &= \exp(-\beta)\cdot g_{t-1}(Z)+\alpha\exp(-\beta t)\{g_{t-1}(Z)+\mu Z\} \\ &=\{\exp(-\beta)+\alpha\exp(-\beta t)\}g_{t-1}(Z)+\alpha\exp(-\beta t)\cdot\mu Z\ , \end{aligned}$$

or

$$g_t(Z)-g_{t-1}(Z)=\{-1+\exp(-\beta)+\alpha\exp(-\beta t)\}g_{t-1}(Z)+\exp(-\beta t)\cdot\mu Z\ . \tag{2.4.4}$$

Remark that $g_t(Z)>0$ for $t>t_0$ and that the map $f(t)=-1+\exp(-\beta)+\alpha\exp(-\beta t)$ is decreasing as t increases to $-1+\exp(-\beta)<0$.
On $\{Z=0\}$ one easily finds

$$\tau(0)=\text{ entier }[\beta^{-1}\ln\alpha(1-\exp(-\beta))^{-1}]\ ,$$

and on $\{Z=1\}$ a numerical approximation algorithm will give $\tau(1)$. Note that from (2.4.4) follows $\tau(1)\geqslant\tau(0)$ and note also that $\tau(Z)$ is decreasing in β and increasing in α, as should be expected.
Taking the mathematical expectation with respect to Z in (2.4.3) will give the recursion formula for computation of τ. We note here that τ is approximately equal to $\tau(0)$ since $\mu\epsilon=\mathbb{E}\mu Z\ll\max_s g_s$ in practice.

2.4.4.
In requirement (1.3.2) it was stated that Z will be revealed as $t\to\infty$ and in fact our model will predict Z exponentially fast as $t\to\infty$. Obviously $t^{-1}\cdot(N_1+N_2+...+N_t)$ for a large observation time t of a series of citations will give a prediction for μZ and indeed it can be shown that

$$\lim_{t\to\infty} t^{-1}(N_1+N_2+...+N_t)=\mu\, Z \qquad w.p.\ 1. \tag{2.4.5}$$

However the 'natural' prediction of Z is $\mathbb{E}[Z\,|\,\mathbf{N}_t]=\mathbb{P}[Z=1\,|\,\mathbf{N}_t]$ and this prediction can be shown to converge exponentially fast to Z. We have

$$\limsup_{t\to\infty} t^{-1}\log|Z-\mathbb{E}[Z\,|\,\mathbf{N}_t]|=-\mu \qquad w.p.\ 1. \tag{2.4.6}$$

Roughly speaking $\mathbb{E}[Z\,|\,\mathbf{N}_t]$ converges faster to Z as μ increases, where $\mu=\lambda_t(1)-\lambda_t(0)$ for $t>t_0$, i.e. the difference between the conditional expectations of N_t given $(\mathbf{N}_{t-1},Z=1)$ and $(\mathbf{N}_{t-1},Z=0)$ for $t>t_0$.

2.5.
We are interested in a consistent estimator for the parameter $\theta=(\epsilon,\mu,\alpha,\beta)$ in our model and thus the parameter θ should be at least identifiable, i.e. different values of the parameter lead to different conditional probability distributions of $\mathbf{N}_t$ given $\mathbf{N}_{t_0}$. We observe a sequence $\mathbf{N}_t=(N_0,N_1,\ldots,N_t)$ for some given observation time $t>t_0$ since Z is unobservable. Therefore for given θ and for $t>t_0$ the conditional probability distribution of $\mathbf{N}_t$ given $\mathbf{N}_{t_0}$ is a mixture of the conditional probability distributions of N_t given $\mathbf{N}_{t_0}$ and $Z=1$, respectively N_{t_0} and $Z=0$ with, mixing probability distribution given by $\epsilon=\mathbb{P}[Z=1]$ (see section (2.2)), and one should really be concerned about the identification of θ. Here, we only note that it can be shown that θ is identified if and only if the observation time $t\geqslant t_0+2$ (see also the remark (2.3.3)).

2.6.
In this section we give some formulas, which we will freely use in the following chapters. Since N_t given (N_{t-1},Z) is Poisson-$\lambda_t(Z)$ distributed for $t>t_0$ with

$$\lambda_t(Z)=\mu Z+\alpha\exp(-\beta t)\cdot\sum_{s=0}^{t-1}N_s,\qquad\text{for given }\ \mu,\alpha,\beta>0$$

we have

$$\mathbb{E}[N_t\,|\,\mathrm{N}_{t-1},Z]=\operatorname{var}[N_t\,|\,\mathrm{N}_{t-1},Z]=\lambda_t(Z)\qquad\text{for }\ t>t_0\,. \tag{2.6.1}$$

Therefore

$$\mathbb{E}[N_t\,|\,\mathrm{N}_{t-1}]=\mu\mathbb{E}[Z\,|\,\mathrm{N}_{t-1}]+\alpha\exp(-\beta t)\sum_{s=0}^{t-1}N_s \tag{2.6.2}$$

and

$$\begin{aligned}\operatorname{var}[N_t\,|\,\mathrm{N}_{t-1}]&=\mathbb{E}[\operatorname{var}(N_t\,|\,\mathrm{N}_{t-1},Z)\,|\,\mathrm{N}_{t-1}]+\operatorname{var}[\mathbb{E}(N_t\,|\,\mathrm{N}_{t-1},Z)\,|\,\mathrm{N}_{t-1}]=\\&=\mathbb{E}[\mathbb{E}(N_t\,|\,\mathrm{N}_{t-1},Z)\,|\,\mathrm{N}_{t-1}]+\operatorname{var}[\mu Z+\alpha\exp(-\beta t)\cdot\sum_{s=0}^{t-1}N_s\,|\,\mathrm{N}_{t-1}]\\&=\mathbb{E}[N_t\,|\,\mathrm{N}_{t-1}]+\operatorname{var}[\mu Z\,|\,\mathrm{N}_{t-1}]\\&=\mathbb{E}[N_t\,|\,\mathrm{N}_{t-1}]+\mu^2\operatorname{var}[Z\,|\,\mathrm{N}_{t-1}]\qquad\text{for }\ t>t_0\,.\end{aligned}$$

Note that the second term after the second equality sign equals $\operatorname{var}[\mu Z\,|\,\mathrm{N}_{t-1}]$ since $\sum_{s=1}^{t-1}N_s$ given N_{t-1} is fixed.
Note also that

$$\mathbb{E}[Z\,|\,\mathrm{N}_{t-1}]=\mathbb{P}[Z=1\,|\,\mathrm{N}_{t-1}] \tag{2.6.3}$$

since $Z\in\{0,1\}$ and so

$$\operatorname{var}[Z\,|\,\mathrm{N}_{t-1}]=\mathbb{E}[Z\,|\,\mathrm{N}_{t-1}]\cdot\{1-\mathbb{E}[Z\,|\,\mathrm{N}_{t-1}]\} \tag{2.6.4}$$

where $\mathbb{E}[Z\,|\,\mathrm{N}_{t-1}]$ may be computed by

$$\mathbb{P}[Z=1\,|\,\mathrm{N}_{t-1}]=\epsilon\cdot P_{(\mu,\alpha,\beta)}(\mathrm{N}_{t-1})\cdot\{P_\theta(\mathrm{N}_{t-1})\}^{-1} \tag{2.6.5}$$

with

$$P_{(\mu,\alpha,\beta)}(\mathbf{n}_{t-1})=\mathbb{P}[\mathrm{N}_{t-1}=\mathbf{n}_{t-1}\,|\,\mathrm{N}_{t_0}=\mathbf{n}_{t_0},\ \ Z=1] \tag{2.6.6}$$

and

$$P_\theta(\mathbf{n}_{t-1})=\mathbb{P}[\mathrm{N}_{t-1}=\mathbf{n}_{t-1}\,|\,\mathrm{N}_{t_0}=\mathbf{n}_{t_0}] \tag{2.6.7}$$

for $t>t_0$.

3. Estimation of θ

3.1.
In this chapter we describe the estimation procedure and the asymptotic behaviour of the estimator for the parameter θ of the model based on n i.i.d. (independent and identically distributed) copies of an observation sequence N_t of citation counts, t is fixed (see section (3.2)). Since θ is identified if and

only if $t \geqslant t_0+2$ (see (2.5)) we will always assume that the fixed observation time $t \geqslant t_0+2$.
In section (3.3) we give the results.

3.2.
Consider the stochastic process N constructed in chapter 2 with parameter (θ,F); $\theta=(\epsilon,\mu,\alpha,\beta)$ and let N_t be a finite observation sequence of the process for fixed t.
Now from (2.2.3) it follows that

$$\mathbb{P}_{\theta,F}[\mathrm{N}_t=\mathbf{n}_t]=\mathbb{P}_F[\mathrm{N}_{t_0}=\mathbf{n}_{t_0}]\cdot\mathbb{P}_\theta[\mathrm{N}_t=\mathbf{n}_t\,|\,\mathrm{N}_{t_0}=\mathbf{n}_{t_0}]$$

implying that for one observation sequence N_t of citation counts the log likelihood for θ given the head of the citation pattern N_{t_0} contains all information about θ. Using (2.2.3) again this (conditional) log likelihood is easily seen to be proportional to

$$\log\,[\epsilon\cdot\prod_{s=t_0+1}^{t}\{\lambda_s(1)\}^{N_s}\cdot\exp\{-\lambda_s(1)\}+(1-\epsilon)\prod_{s=t_0+1}^{t}\{\lambda_s(0)\}^{N_s}\cdot\exp\{-\lambda_s(0)\}]\,.$$

For fixed t let $\{\mathrm{N}_{t,k}\}_{k=1}^{n}$ be an i.i.d. sample of order n of the model with parameters θ and F. Denote for the k-th observation sequence $\mathrm{N}_{t,k}$ the log likelihood for θ given $\mathrm{N}_{t_0,k}$ by $l_k(\theta;\mathrm{N}_{t,k})$, then the log likelihood $l_n(\theta;t)$ of θ for the total sample $\{\mathrm{N}_{t,k}\}_{k=1}^{n}$ given $\{\mathrm{N}_{t_0,k}\}_{k=1}^{n}$ is given by

$$l_n\,(\theta;t)=\sum_{k=1}^{n}l_k(\theta;\mathrm{N}_{t,k})$$

where $l_k(\theta;\mathrm{N}_{t,k})$ is proportional to

$$\log\,[\epsilon\cdot\prod_{s=t_0+1}^{t}\{\lambda_{s,k}(1)\}^{N_{s,k}}\cdot\exp\{-\lambda_{s,k}(1)\}+(1-\epsilon)\cdot\prod_{s=t_0+1}^{t}\{\lambda_{s,k}(0)\}^{N_{s,k}}\exp\{-\lambda_{s,k}(0)\}]$$

with

$$\lambda_{s,k}(z)=\mu z+\alpha e^{-\beta s}\sum_{u=0}^{s-1}N_{u,k}\quad\text{for } s>t_0\quad\text{and } z=0,1\,.$$

Let θ_0 be the true parameter value and denote by $\hat{\theta}_n(t)$ the maximum likelihood estimator (MLE) for θ based on $l_n(\theta;t)$ for arbitrarily fixed $t \geqslant t_0+2$. According to standard statistical theory we have that for fixed t the MLE $\hat{\theta}_n(t)$ for θ is a strongly consistent, efficient and asymptotically normal estimator for θ_0 as $n\to\infty$, i.e. for fixed t we have

$$\lim_{n\to\infty}\hat{\theta}_n(t)=\theta_0\qquad w.p.\ 1.\quad\text{and}\tag{3.2.1}$$

$$\lim_{n\to\infty}\mathcal{L}\,[\sqrt{n}(\hat{\theta}_n(t)-\theta_0)]=\mathfrak{N}\,(0,I^{-1}(\theta_0))\,,$$

where

$$I(\theta_0)=\mathbb{E}\frac{\partial l(\theta_0;\mathrm{N}_t)}{\partial\theta}\cdot\left[\frac{\partial l(\theta_0;\mathrm{N}_t)}{\partial\theta}\right]^T$$

is the Fisher information matrix.
Note that the expansion for $I(\theta_0)$ is complicated for our model. However by (3.2.1) and since $I(\theta_0)=-\mathbb{E}\frac{\partial^2}{\partial\theta\partial\theta^T}\,l\,(\theta_0,\mathrm{N}_t)$ also, we may approximate the information matrix in practical situations by the observed Fisher information matrix, i.e.

n^{-1} times the matrix of second derivatives of $l_n(\theta;t)$ computed for $\theta=\hat{\theta}_n(t)$.

In the results in section (3.3) the variance-covariance matrix of $\sqrt{n}(\hat{\theta}_n(t)-\theta_0)$ is estimated by the inverse of the above mentioned observed information matrix.

3.3.

We computed the MLE for θ with an optimization procedure for the log likelihood that used the analytically given first derivatives and numerically computed the second derivatives of the log likelihood. Note that these second derivatives immediately gives the observed information matrix.

For the dataset we used, consisting of 320 series of citation counts of length 11, we took $t_0=2$ (see also section (1.4)). The results for the maximum likelihood estimator (MLE) $\hat{\theta}=\hat{\theta}_{320}(10)$ with estimated standard deviations (DEV) are shown in Table (3.3.1). The estimated correlation matrix CORR for θ is shown in Table (3.3.2).

	ϵ	μ	α	β
MLE	.0406	2.4040	.7217	.2961
DEV	.0152	.4709	.0439	.0102

TABLE 3.3.1. MLE and DEV for $\theta=(\epsilon,\mu,\alpha,\beta)$

	ϵ	μ	α	β
ϵ	1			
μ	—0.5827	1		
α	—0.0255	—0.0613	1	
β	.1375	.1390	.9197	1

TABLE 3.3.2. CORR for θ

From Table (3.3.1) we observe that the parameters α and β are well estimated and nonsurprisingly that the parameters ϵ and μ of the mixing distribution are less well estimated. Note that the parameter μ is estimated by the (estimated) fraction 0.0406 of a total number of 320 series of citation counts, i.e. μ is estimated by 13 series of citation counts.

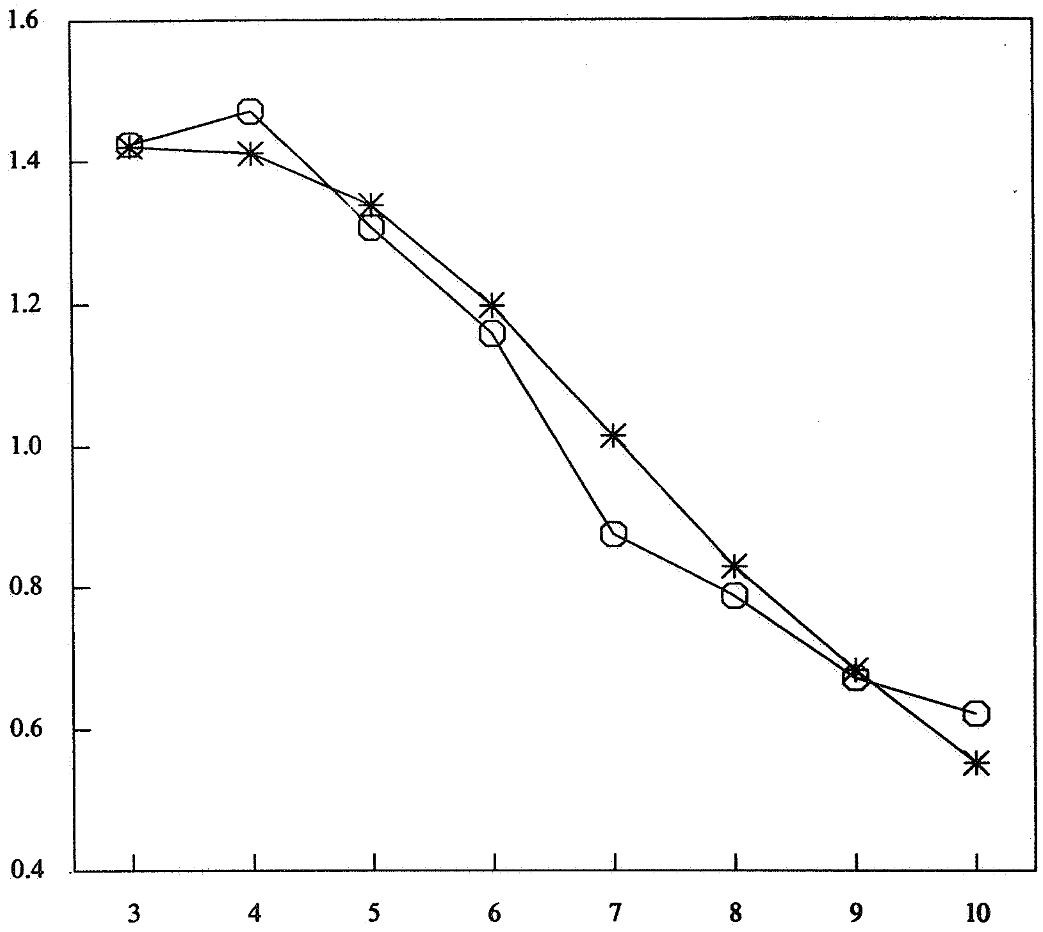
From the construction of the model, see also the likelihood in (3.2), it should be clear that the parameters ϵ and μ will be negatively correlated, resp. that the parameters α and μ will be positively correlated. Indeed Table (3.3.2) shows a negative correlation -0.5827 between ϵ and μ and unfortunately a very high positive correlation .9197 between α and β.

3.4.

To get a first impression if the model will possibly fit in some sense we compared the sample mean of N_t with the sample mean of $\mathbb{E}[N_t\,|\,\mathrm{N}_{t-1}]$, i.e. the conditional expectation of N_t given N_{t-1} computed for $\theta=\hat{\theta}$ for $t=3,4,\ldots,10$.

Note that $\lim\limits_{n\to\infty}\frac{1}{n}\sum\limits_{k=1}^{n}\hat{\mathbb{E}}[N_{t,k}\,|\,\mathrm{N}_{t-1,k}]=\mathbb{E}N_t$ *w.p.* 1., since $\lim\limits_{n\to\infty}\hat{\mathbb{E}}[N_t\,|\,\mathrm{N}_{t-1}]=\mathbb{E}_0[N_t\,|\,\mathrm{N}_{t-1}]$ *w.p.* 1. because $\mathbb{E}[N_t\,|\,\mathrm{N}_{t-1}]$ is continuous in θ and $\lim\limits_{n\to\infty}\hat{\theta}_n=\theta$ *w.p.* 1. and $\lim\limits_{n\to\infty}\frac{1}{n}\sum\limits_{k=1}^{n}\mathbb{E}[N_{t,k}\,|\,\mathrm{N}_{t-1,k}]=\mathbb{E}N_t$ w.p.1. ($\mathbb{E}_0[N_t\,|\,\mathrm{N}_{t-1}]$ means here computation of this conditional expectation under the true parameter value θ_0).

The results are encouraging and shown in figure (3.4.1).



$\epsilon = 0.0406$, $\mu = 2.404$, $\alpha = 0.7217$, $\beta = 0.2961$

FIGURE 3.4.1. Sample mean of N_t indicated by o and estimated sample mean of $\mathbb{E}[N_t \,|\, \mathrm{N}_{t-1}]$ indicated by $\star$

3.5.

Approximating $g_t(Z)=\mathbb{E}[N_t-\mu Z\,|\,Z]$ by $\hat{g}_t(Z)=[\hat{\mathbb{E}}[N_t-\hat{\mu}Z\,|\,Z]$ one can estimate $\tau(Z)$ by $\hat{\tau}(Z)$ for $Z=0,1$ and $t>2$ (i.e. the finite time $\hat{\tau}(Z)$ such that $\hat{g}_t(Z)$ is increasing for $2<t\leqslant\hat{\tau}(Z)$ and decreasing for $t>\hat{\tau}(Z)$).

Similarly approximating $g_t=\mathbb{E}[N_t-\mu Z]$ by $\hat{g}_t=\hat{\mathbb{E}}[N_t-\mu Z]$ for $t>2$ one can estimate τ by $\hat{\tau}$. Observing that $\hat{\mu}\,\hat{\epsilon}=0.0976\ll 1.4<\max_t \overline{g}_t$ we expect that $\hat{\tau}=\hat{\tau}(0)$. Indeed computation gives $\hat{\tau}=\hat{\tau}(0)=3$, this confirming the note at the end of section (2.4.3).

4. Testing the model

4.1.

To establish 'reasonable' tests for goodness of fit for complicated statistical models like this for sequences of citation counts is of course difficult. Therefore we will test the model with different test statistics based on the so called conditional residuals of the citation counts.

Firstly we develop tests based on summarizing the conditional residuals over all publications of the dataset for each time t. Secondly we construct tests more 'natural' for the model since these tests are based on summarizing the whole sequence of conditional residuals for one publication. Moreover we choose a basic statistic sensitive for outliers.

Since asymptotic distributions for the latter test statistics are hard to obtain we relied for these tests on the so called bootstrap method, for which numerous simulations of sequences of citation counts were needed. In fact we made 1000 simulations of the dataset consisting of 320 series of citation counts. (For a description of the procedure for the simulations the reader is refered to Appendix (6.1) and for an overview of the bootstrap method to Efron (1979) or to Bickel and Freedman (1981)). Without loss of generality and in accordance with the choice for the head of the citation counts for the dataset used we will take $t_0=2$.

4.2.

Define for $t=3,4,...$ the conditional residual U_t of the citation count N_t by

$$U_t=\sigma_t^{-1}\cdot\{N_t-\lambda_t\}\,,\quad\text{where}$$

$$\lambda_t=\mathbb{E}[\lambda_t(Z)\,|\,\mathbf{N}_{t-1}]=\mathbb{E}[N_t\,|\,\mathbf{N}_{t-1}]\quad\text{and}$$

$$\sigma_t=\sqrt{\sigma_t^2}\ \text{ with } \sigma_t^2=\operatorname{var}[N_t\,|\,\mathbf{N}_{t-1}]\,.$$

Note that λ_t and σ_t can be computed by the formulas given in (2.6).

The meaning of the conditional residuals as a basis for constructing tests for the model is given by the absolute value of these residuals U_t, which may be interpreted as a distance between N_t and the conditional expectation of N_t given the history of citation counts $\mathbf{N}_{t-1}$, relative to the conditional variance of N_t given $\mathbf{N}_{t-1}$.

Note also that the sequence $\{U_t;t=3,4,...\}$ is a martingale difference sequence since $\mathbb{E}[U_t\,|\,\mathbf{N}_{t-1}]=0$ w.p.1. and that moreover $\operatorname{var}[U_t\,|\,\mathbf{U}_{t-1}]=1$ w.p.1. because $\operatorname{var}[U_t\,|\,\mathbf{N}_{t-1}]=\mathbb{E}[U_t^2\,|\,\mathbf{N}_{t-1}]=1$ w.p.1. with $\mathbf{U}_t=(U_3,U_4,\ldots,U_t)$ for $t=3,4,....$

4.3.

For fixed t let $\{\mathbf{N}_{t,k}\}_{k=1}^{n}$ be an i.i.d. sample from the model with matched series of conditional residuals $\{\mathbf{U}_{t,k}\}_{k=1}^{n}$, where $\mathbf{U}_{t,k}=(U_{3,k},U_{4,k},\ldots,U_{t,k})$ for $k=1,2,\ldots,n$.

It is standard that $\lim\limits_{n\to\infty} n^{-1}\sum\limits_{k=1}^{n}\mathbf{U}_{t,k}=\mathbf{O}$ *w.p.* 1. and taking into account the latter note that $\lim\limits_{n\to\infty} n^{-1/2}\sum\limits_{k=1}^{n}\mathbf{U}_{t,k}$ is asymptotically distributed as a standaard multivariate normal distribution with covariance matrix I_{t-2}.

For large n therefore $V_s:=\{n^{-1/2}\sum\limits_{k=1}^{n}U_{s,k}\}^{z}$ for $s\geqslant3$ is approximatily distributed as a chi-square distribution with one degree of freedom and $W=\sum\limits_{s=3}^{t}V_s$ as a chi-square distribution with $(t-2)$-degrees of freedom. The first tests we perform are based on V_s, $s=3,4,\ldots,t$ and W.

4.4.

A more 'natural' test for the model should be based on a statistic that summarizes over the sequence of conditional residuals for one publication. We choose for that basic statistic $M=\max\limits_t |U_t|$ and for the test statistics the sample mean of M and the sample median of M. Note that M should detect outliers and that the sample mean of M is much more sensitive for outliers then the sample median. Since asymptotic distributions for these statistics are hard to obtain we relied for these tests on the above mentioned bootstrap method.

4.5.

As mentioned in the introduction (4.1) we have to rely on the bootstrap method for some tests, for which numerous simulations of sequences of citation counts are needed.

Now by the formulation of the model we may work with sequences of observations $(S_2,N_3,\ldots,N_t)$ with $S_2=N_0+N_1+N_2$. Therefore it suffices to simulate sequences $(S_2,N_3,\ldots,N_t)$ for some fixed observation time t and it is convenient to set $\mathrm{N}_t=(S_2,N_3,\ldots,N_t)$ and to let $\mathcal{G}$ indicate the probability distribution of S_2 in this and further sections.

For a dataset $\{\mathrm{N}_{t,k}\}_{k=1}^{n}$ of n i.i.d. copies of N_t let $\hat{\mathcal{G}}_n$ be the empirical distribution function of $\{S_{2,k}\}_{k=1}^{n}$ and let $\hat{\theta}_n=\theta_n(t)$ be the MLE for θ obtained from maximizing the log likelihood $l_n(\theta,t)$ of section (3.2).

Denote for $m=1,2,\ldots$ the m-th simulation of the dataset $\{\mathrm{N}_{t,k}\}_{k=1}^{n}$ under $\hat{\theta}_n$ and $\hat{\mathcal{G}}_n$ by $\{\mathrm{N}_{t,k}^{(m)}\}_{k=1}^{n}$. So $\mathrm{N}_{t,k}^{(m)}=(S_{2,k}^{(m)},N_{3,k}^{(m)},\ldots,N_{t,k}^{(m)})$ is the k-th simulated sequence of citation counts in the m-th simulation of the dataset. (For a description of the simulation procedure see Appendix (6.1)). Furthermore indicate the conditional residuals and all derived statistics for the m-th simulated sample with an upper-index m between brackets. For instance $U_{s,k}^{(m)}$ is the conditional residual of $N_{s,k}^{(m)}$ given $\mathrm{N}_{s-1,k}^{(m)}$ computed for $\theta=\hat{\theta}_n$ at time s, $M_k^{(m)}=\max\limits_s |U_{s,k}^{(m)}|$ etc. Finally let $\hat{\theta}_n^{(m)}=\hat{\theta}_n^{(m)}(t)$ be the MLE for θ for the m-th simulated dataset based on the log likelihood $l_n^{(m)}(\theta,t)$.

4.6.

Since we do not know the real parameter value θ_0 we have to approximate the conditional residuals and therefore all test statistics derived from these residuals by computing them for $\theta=\hat{\theta}_n$. This is justified because $\lim\limits_{n\to\infty}\hat{\theta}_n=\theta_0$ w.p.1. Consequently we should compare these test statistics computed in $\hat{\theta}_n$ with the test statistics obtained from the simulated samples computed in their matched MLE-values for θ, so in the m-th simulation of the dataset computed under $\hat{\theta}_n^{(m)}$. We will indicate these 'approximate' test statistics by placing a 'hat' on the original test statistic. So $\hat{U}_{s,k}^{(m)}$ means $U_{s,k}^{(m)}$ computed for $\theta=\hat{\theta}^{(m)}$, $\hat{M}_k^{(m)}=\max\limits_s |\hat{U}_{s,k}^{(m)}|^{s,k}$ etc.

We computed the so called p-values for the test statistics. Note that if $\hat{T}$ is a test statistic based on the original dataset and $\hat{T}^{(m)}$ the matched test statistic for the m-th simulation of that dataset, then the (estimated) p-value is the proportion of $\hat{T}^{(m)}$ greater or equal to $\hat{T}$, i.e. for a 1000 simulations this p-value is $(1000)^{-1}\cdot\#\{m:\hat{T}^{(m)}\geqslant\hat{T}\}$.

4.7.

In this section we give the results for the tests of the first type of section (4.3). Note that the asymptotic distributions for these statistics are derived under the true, but unknown parameter value θ_0 and that we have not derived the asymptotic distribution for these statistics under the MLE $\hat{\theta}_n$ for θ. If possible we will therefore test nonconservatively and for comparison we also computed the simulated p-values (see the last paragraph at the end of section (4.6)).

Note again that the dataset we used, consists of 320 series of citation counts over a time period of eleven years and that $t_0=2$.

The first tests are based on V_t for $t=3,4,\ldots,10$. Comparing $\hat{V}_t$ with a chi-square distribution with one degree of freedom we find the p-value $\hat{p}_\chi$ and also comparing $\hat{V}_t$ with the simulated $\{\hat{V}_t^{(m)}\}_{m=1}^{1000}$

we find the simulated p-value $\hat{p}_{\mathrm{sim}}$. The results are shown in table (4.7.1). Note that $\hat{p}_{\chi}>\hat{p}_{\mathrm{sim}}$ as we should expect.

t	3	4	5	6	7	8	9	10
$\hat{V}_t$	.1929	.0691	.3715	.1852	2.1680	.0175	.0460	1.9211
$\hat{p}_{\chi}$	.6605	.7921	.5422	.6669	.1409	.8948	.8302	.1657
$\hat{p}_{\mathrm{sim}}$	.5400	.7500	.4960	.6590	.1160	.8790	.81110	.1170

TABLE 4.7.1. p-values for V_t

The second test statistic is based on $W=\sum_{s=3}^{10} V_s$. Nonconservatively testing $\hat{W}=4.9712$ with a chi-square distribution with $8-4=4$ degrees of freedom (4 is the number of parameters in θ) gives the p-value $\hat{P}_{\chi}=.2903$ and also comparing $\hat{W}$ with $\{\hat{W}^{(m)}\}_{m=1}^{1000}$ gives the simulated p-value $\hat{P}_{\mathrm{sim}}=.588$. Note that $\hat{P}_{\mathrm{sim}}$ is much bigger than $\hat{P}_{\chi}$.

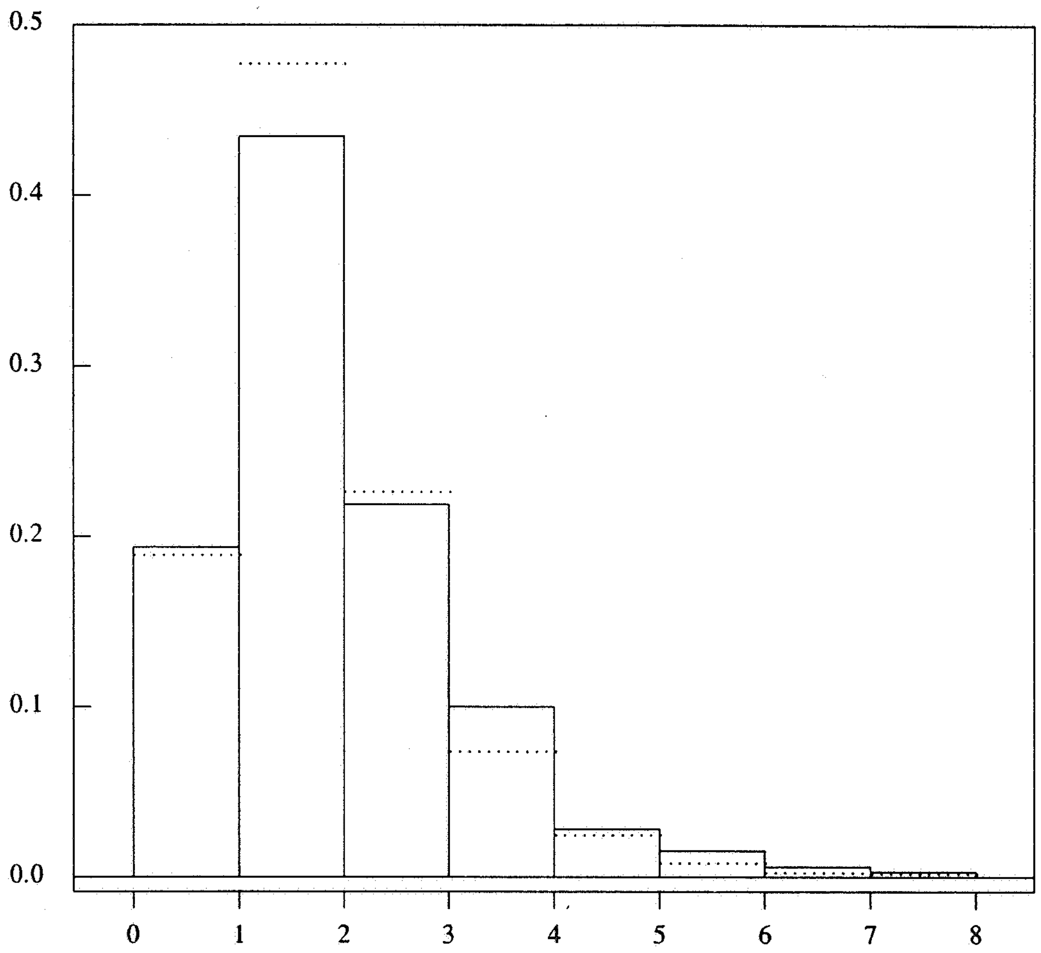
4.8.

To get a first impression about tests of the second type of section (4.4) we made a histogram of $M=\max_t|U_t|$ based on the original sample and for comparison a histogram based on all simulated samples.

Let $\hat{g}=(\hat{g}(1), \hat{g}(2), \ldots, \hat{g}(8))$, where $\hat{g}(i)$ denote the relative freqency of $\{\hat{M}_k\}_{k=1}^{n}$ with values greater or equal to $i-1$ and smaller then i for $i=1,2,\ldots,7$ and $\hat{g}(8)=1-\sum_{i=1}^{7}\hat{g}(i)$. Similarly let $\bar{g}=(\bar{g}(1), \bar{g}(2), \ldots, \bar{g}(8))$ denote the vector of relative frequencies obtained from the simulated samples, e.g. for 1000 simulations of the dataset we have $\bar{g}(1)=(1000)^{-1}\sum_{m=1}^{1000}\hat{g}^{(m)}(1)$ with $\hat{g}^{(m)}(1)=n^{-1}\cdot\#\{k\colon 0\leqslant\hat{M}_k^{(m)}<1\}$. Note that $\bar{g}(i)$ is approximately equal to $\mathbb{P}_{\hat{\theta}_0,\hat{G}_n}[M\in[i-1, i>]$ for a large number of simulations of the dataset $(i=1,2,\ldots,8)$.

For our dataset with $n=320$ the results for $\hat{g}$ and for $\bar{g}$ for the 1000 performed simulations are shown in figure (4.8.1): $\bar{g}$ dashed.

FIGURE 4.8.1. Histograms of M

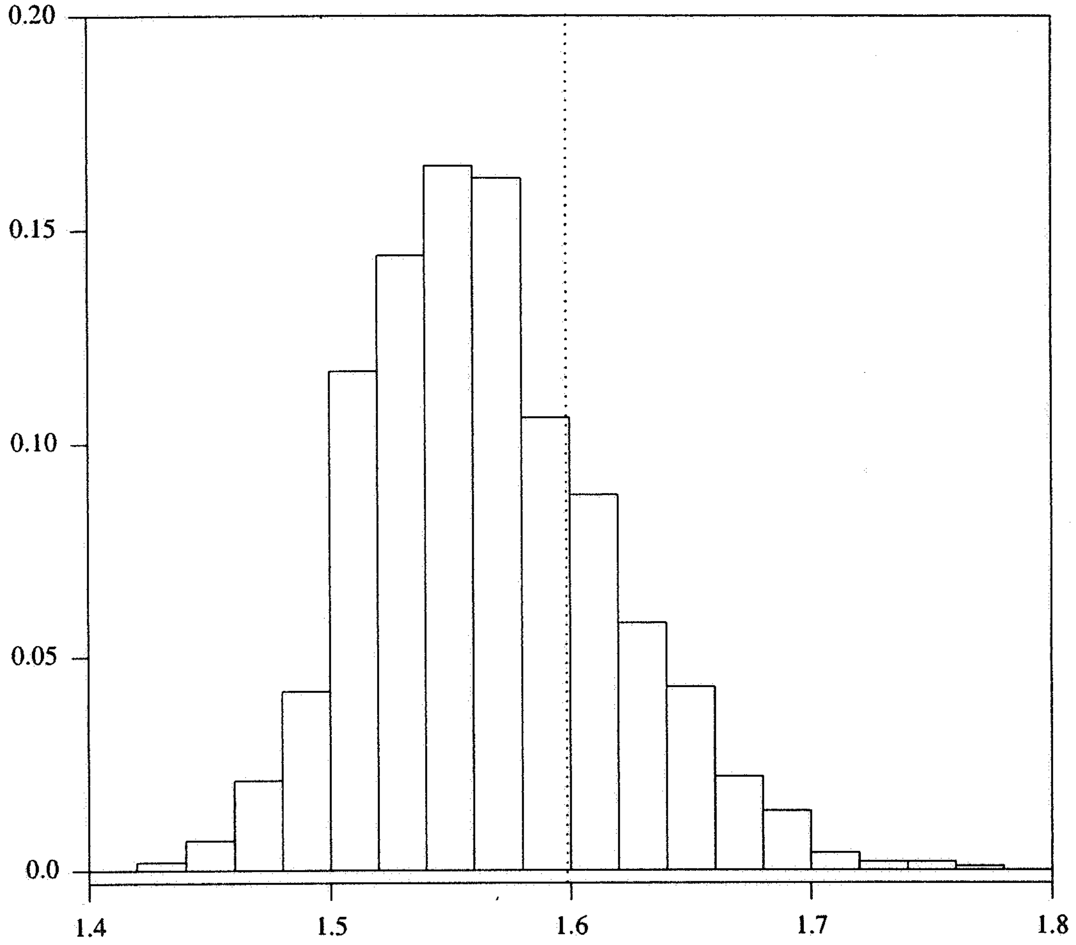


$\hat{g}$	.1938	.4344	.2188	.1000	.0281	.0156	.0963	.0031
$\bar{g}$	.1884	.4766	.2257	.0735	.0241	.0078	.0023	.0016

The figure shows a good resemblance between $\hat{g}$ and $\bar{g}$ except for the classes $i>3$, where $\hat{g}$ is obviously much bigger then $\bar{g}$ (in a relative sense).

These observations are confirmed by using the sample mean $\overline{M}$ of M and the sample median MED of M as test statistics. We find $\hat{\overline{M}}=1.9054$ with p-value $p_{\text{mean}}=0.0160$ and $\widehat{\text{MED}}=1.5974$ with p-value $p_{\text{med}}=0.2440$. A histogram of the median provided by the thousand simulations of the dataset is shown in firgure (4.7.2) with the computed value $\widehat{MED}=1.5979$ drawn dashed.

FIGURE 4.8.2. Histogram of median of M



4.8.

We conclude that the model seems to fit the data quite well. The results of the test statistics in the latter section (4.7) indicate that the model assumes too much homogeneity. We were however not able to detect outliers on qualitative or quantitative grounds. On the other hand qualitative reasoning points to some deficiencies in the model itself (see (2.3.1)) and chapter 5).

5. Some remarks

Although we believe that the model is reasonable from a theoretical and practical point of view, we will discuss in this section several ways to improve the model.

One method would be to incorporate covariates in the model and so to be able to cope with the obvious heterogeneity in datasets with a number n of citation series large enough to allow for sensible estimates of the parameters. However, as already indicated by the remark about outliers in section (4.8), we were not able to incorporate covariates. This is certainly an important theoretical question for citation analysts.

Another way should be to improve the existing model itself. The assumption of the stochastic independence between the unobserved stochastic variable Z and the head of the citation pattern N_{t_0}, and thus between Z and S_{t_0}, is certainly not realistic (see also (2.3.1) and in fact rejected by the used dataset, as we shall now show.

If we define $\hat{Z}_k = \hat{\mathbb{E}}[Z_k \mid \mathrm{N}_{t,k}]$ for $k=1,2,\ldots,n$ for a sample of n observation sequences of length $t+1$ and let $\hat{Q}_{1-\epsilon}$ indicate the $(1-\hat{\epsilon})$ sample quantile of $\hat{Z}_{(1)} \leqslant \hat{Z}_{(2)} \leqslant \cdots \leqslant \hat{Z}_{(n)}$, then $\tilde{Z}_k$ defined by

$$\tilde{Z}_k = \begin{cases} 1 & \text{if} \quad \hat{Z}_k > \hat{Q}_{1-\epsilon} \\ 0 & \text{otherwise} \end{cases}$$

is a good estimator of, or good prediction for Z_k for $k=1,2,\ldots,n$. Calculating the sample correlation r of $(\tilde{Z}_k, S_{2,k})_{k=1}^{320}$ for the used sample with $n=320$, $t=10$ and $t_0=2$ gives $r \approx 0.2$, while bootstrapping gives $r=0.0$.

So the assumption of independence between Z and N_{t_0} should be disgarded. We note that the properties of the model reviewed in section (2.4) do not depend on this assumption (see VAN DER PLAS (1988)). It should be clear however, that efficient estimation of the parameters μ,α and β is a much more complicated matter for this extension.

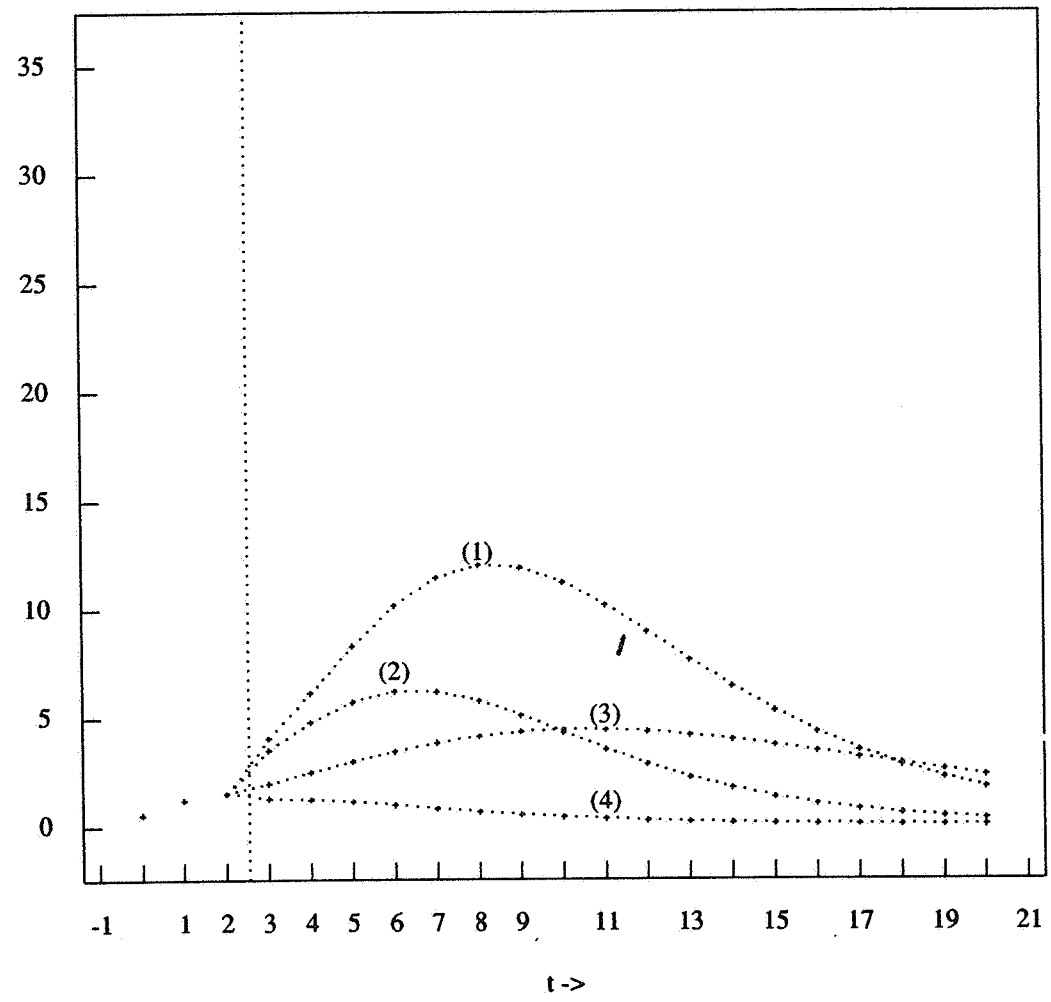
A second serious objection against the model is that Z is a dummy stochastic variable, implying that all papers with $Z=1$ have for large t about the same yearly expected number of citations. An immediate extension is to let Z take a finite or countable number of nonnegative values ($Z<\infty$). Most of the properties of the original model (see section (2.4)) are easily seen to hold for this extension (see VAN DER PLAS (1988)). It should be noted however that most series of citation counts will soon become extinct, i.e. most papers have $Z=0$ and the length of the available series is quite short in practice. Therefore from a practical point of view we do not expect that this extension of the model is useful.

6. Applications to scientometry

In this chapter we will briefly discuss the contribution of the stochastic model to questions in scientometry. Firstly scientometrists are interested in so-called different expected citation patterns. (See e.g. VLACHY (1985).) Now according to the constructed model we have a mean expected citation pattern $(\mathbb{E}N_0, \mathbb{E}N_1, \ldots)$ for a certain scientific field (see (1.3.1) and Figure (1.4.2)) and different expected citation patterns for $Z=0$, resp $Z=1$, i.e. the series $(\mathbb{E}[N_0 \mid Z=0], \mathbb{E}[N_1 \mid Z=0], \ldots)$, resp. the series $(\mathbb{E}[N_0 \mid Z=1], \mathbb{E}[N_1 \mid Z=1], \ldots)$.

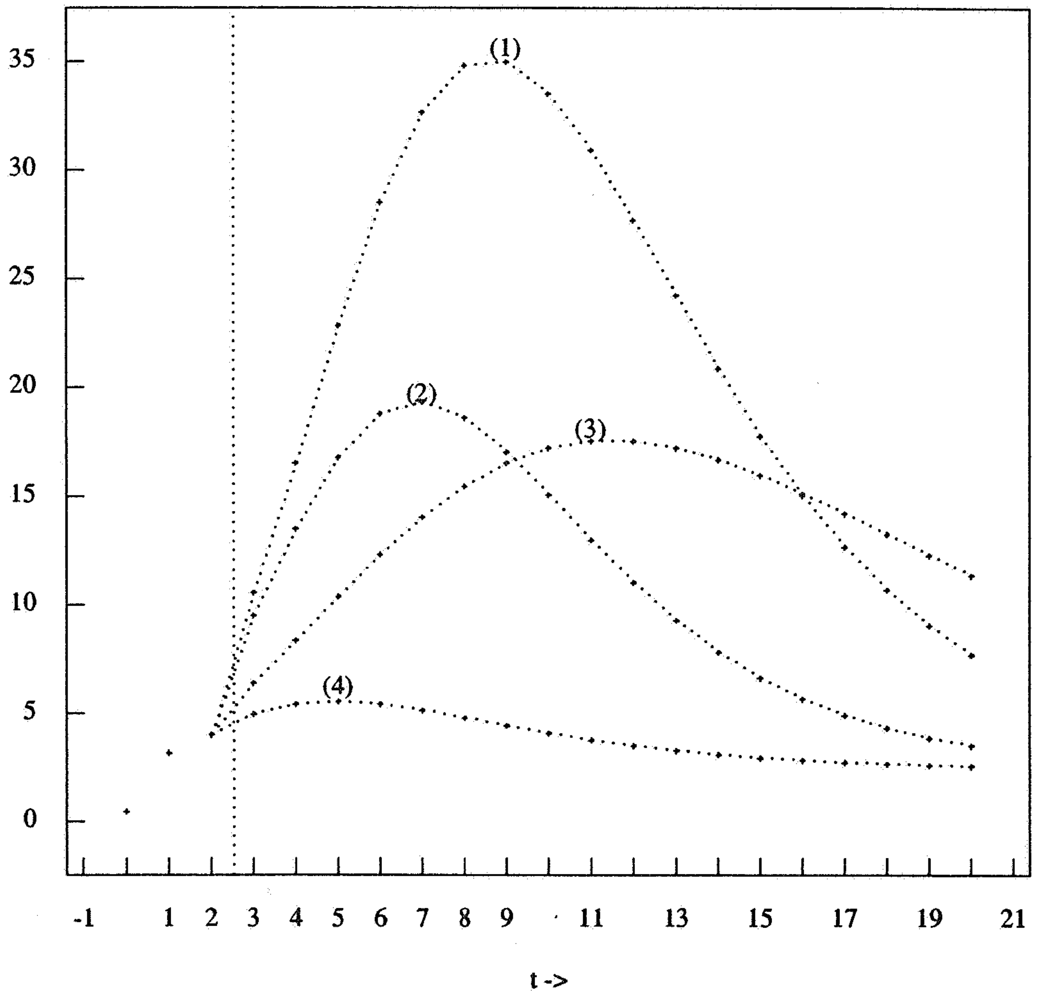
For the dataset used we computed these different expected citation patterns as follows. Estimate Z for each publication by $\hat{Z}_k$ for $k=1,2,\ldots,320$ (see Chapter 5 for the definition of $\hat{Z}_k$). Then estimate $\mathbb{E}[N_t \mid Z]$ for $Z=0,1$ and $t \leqslant 2$ by the average citation count of the group with $Z=0$, resp. $Z=1$ and for $t>2$ by the formula $\mu Z + \alpha \exp(-\beta t) \cdot \mathbb{E}[S_{t-1} \mid Z]$ with the estimated values $\mu=2.404$, $\alpha=.7217$ and $\beta=.2961$. Moreover using the above computed $\mathbb{E}[N_t \mid Z]$ for $t \leqslant 2$ and $Z=0,1$ we computed expected citation patterns for fixed $\mu=2.404$ and different values for (α,β), namely (2.0, .25); (2.0, .2961) resp. (.7217, .15). The results are presented in Figure (6.1.1) for $Z=0$ resp. Figure (6.1.2) for $Z=1$. Note that these pictures clearly show that Z and S_2 are positively correlated and that $\tau(1) > \tau(0)$ for fixed (α, β).

FIGURE 6.1.1. Expected citation patterns for $Z = 0$



$\mu = 2.404$
(1) alfa=2.0000 , beta=.2500
(2) alfa=2.0000 , beta=.2961
(3) alfa=0.7217 , beta=.1500
(4) alfa=0.7217 , beta=.2961

FIGURE 6.1.2. Expected citation patterns for $Z=1$



$\mu=2.404$
(1) alfa=2.0000 , beta=.2500
(2) alfa=2.0000 , beta=.2961
(3) alfa=0.7217 , beta=.1500
(4) alfa=0.7217 , beta=.2961

A second important question for scientometrists is the so-called lifetime distribution of citation counts for 'ordinary' publications, i.e. the publications with $Z=0$. Defining the lifetime of citation counts by the stochastic varialble $T=\min\{t : N_u=0 \text{ for } u\geqslant t\}$ we have

$$\mathbb{P}[T\leqslant t\,|\,Z=0] = \mathbb{E}[\exp\{-\Gamma_t S_{t-1}\}\,|\,Z=0] \text{ for } t\geqslant t_0\,,$$

where $\Gamma_t=\sum_{u\geqslant t}\alpha\exp\{-\beta u\}$. (See VAN DER PLAS (1988).) In other words $\mathbb{P}[T\leqslant t\,|\,Z=0]$ equals the Laplace transform of S_{t-1} in the point Γ_t for $t\geqslant t_0$. So the expected lifetime of citation counts for ordinary publications, and all other moments of T are in principle computable. Note that from the expected pattern with $Z=0$ for the estimated model we have $\mathbb{E}[T\,|\,Z=0]=20$.

Another important question called impact evaluation (see MOED et al. (1985)) can be analyzed by the probability distribution of the total number of citations for ordinary publications, i.e. the probability distribution of $\lim\limits_{t\to\infty} S_t$ on $\{Z=0\}$. We only mention here that S_t tends exponentially fast to $S=\lim\limits_{t\to\infty} S_t$ in probability. (See VAN DER PLAS (1988)). For practical purposes the probability distribution of S may be approximated via the formula

$$\mathbb{P}[S_t=s \mid S_{t-1},\ Z=0] = \begin{cases} 0 & \text{for } s<S_{t-1} \\ \mathbb{P}[N_t=s-S_{t-1} \mid S_{t-1},\ Z=0] & \text{for } s \geqslant S_{t-1} \end{cases}$$

(see also the formulas in (2.2)).

7. Appendix

7.1. *Description of simulation procedure*

Let $\mathrm{N}=\{N_t; t=0,1,2,...\}$ be the stochastic process defined in chapter 2 with parameters (θ,F), where $\theta=(\epsilon,\mu,\alpha,\beta)$ and where F is the probability distribution of the so called head N_{t_0} of N. Let $\mathcal{G}$ be the probability distribution of $S_{t_0}=N_0+N_1+...+N_{t_0}$.
The simulation procedure works as follows:

7.1.1.
Sample S_{t_0} from $\mathcal{G}$ and Z from $\mathbb{P}[Z=1]=\epsilon$.

7.1.2.
Recursively for $t=t_0+1, t_0+2,...$ sample N_t given $S_{t-1}=N_0+N_1+...+N_{t-1}$ and given Z from a Poisson-distribution with expectation $\lambda_t(Z)=\mu Z+\alpha e^{-\beta t}\sum\limits_{s<t} N_s$.

7.2.
To appoint the real date of publication of a paper in respect to citations is a serious problem since a paper may be known before the calendar date of publication in a journal and thus may be cited already before that calendar date. Note that the available dataset we used consists of citation counts over calendar years, so even the calendar data of publications are not known.

ACKNOWLEDGEMENT
The author thanks H.F. Moed, A.F.J. van Raan and also J.E.J. Oberski for stimulating conversations on citation analysis, H.C.P. Berbee, R.D. Gill and R. Helmers for stimulating discussions on several aspects of the model and R. van der Horst for performing all computations and producing the graphs.

References

BICKEL, P.J. and FREEDMAN, D.A. (1981). *Some asymptotic theory for the bootstrap,* Ann. Stat. **9**, pp. 1196-1217.

CHANG, K.H. (1975). *Evaluation and survey of a subfield of physics: magnetic resonance and relaxation studies in the Netherlands,* FOM-Report 37175, Utrecht.

DE SOLLA PRICE, D. (1976). *A general theory of bibliometric and other cumulative advantage processes,* J.A.S.I.S. **27**.

EFRON, B. (1979). *Bootstrap methods: another look at the jackknife,* Ann. Stat. **7**, pp. 1-26.

FELLER, W. (1943). *On a general class of "contagious" distributions,* Ann. Math. Stat. **14**, pp. 389-400.

MOED, H.F., BURGER, W.J.M., FRANKFORT, J.G. and RAAN, A.F.J. VAN (1983). *On the measurement of research performance: the use of bibliometric indicators,* Research Policy Unit of the University of Leiden, ISBN 90-9000552-8.

MOED, H.F., BURGER, W.J.M., FRANKFORT, J.G. and RAAN, A.F.J. VAN (1985). *The use of* *bibliometric data for the measurement of research performance,* Research Policy **14,** pp. 131-149.

TEICHER, H. (1961). *Identifiability of mixtures,* Ann. Math. Stat. **32,** pp. 244-248.

VAN DER PLAS, A.P. (1988). *Prediction in a mixture of stochastic processes, forthcoming.*

VLACHY, J. (1985). *Citation histories of scientific publications. The data sources,* Scientometrics 7, pp. 505-528.

WELLNER, J. (1985). *Semiparametric models: progress and problems,* CWI Newsletter **9.**