On the diffusion of scientific publications; The case of Econometrica 1987*

Philip Hans Franses†

Econometric Institute, Erasmus University Rotterdam

Econometric Institute Report 2002-16

Abstract

This paper documents that salient features of (time series of annual) citations to scientific publications might be captured by a Bass type diffusion model. This is particularly useful as it allows for a comparison of these features across journals, across disciplines and over time. For the illustrative case of Econometrica 1987, it is found that the peak in citations occurs at 6.5 years, on average. Also, it is found that after 14 years there is only a little gap between cumulative citations and the estimated total cumulative amount, suggesting that on average the impact of these articles lasts for about 15 years or so. Finally, it appears that these features can partly be explained by the size of the articles, as it is found that longer papers get more citations and peak later.

---

*I thank Jesse de Klerk for his assistance with constructing the database. All computations are performed using Eviews (4.1). I thank Jan Brinkhuis, Fransje Akveld and Peter Boswijk for helpful suggestions.

†Econometric Institute, Erasmus University Rotterdam, P.O. Box 1738, NL-3000 DR, Rotterdam, The Netherlands, email: franses@few.eur.nl

1 Introduction

Citations of scientific publications can show characteristics that bear similarities with the diffusion of a new product. In the first period after publication, there are not many citations. After a few years, citations may peak, and after this peak citations eventually fall back to zero. Consequently, a visual characteristic of the associated cumulative citation series is that it follows an S-shaped pattern, which starts at zero and levels off to some upper bound, which can be called a saturation level. This upper bound may depend on various aspects, such as the publication intensity, that is, the amount of articles which are published each year on the same topic, perhaps the features of the article itself, and the speed at which new developments take place in the discipline of interest.

The empirical analysis in this paper concerns the 1987 volume of Econometrica for illustrative purposes. To highlight the data features, consider the graphs in Figures 1 and 2. These first two graphs concern the citations and cumulative citations, respectively, to an article by Aumann (R.J. Aumann, Correlated equilibrium as an expression of Bayesian rationality, Econometrica, 55, 1-18). Similar graphs for the article by Engle and Granger (R.F. Engle and C.W.J. Granger, Co-integration and error correction: Representation, estimation and testing, Econometrica, 55, 251-276) appear in Figures 3 and 4. Clearly, the graphs in these figures display the pattern as suggested. The citations series displays an inverted u-shape pattern, while the cumulative series has a tendency to show an S-shaped pattern. Of course, the inflection point of cumulative citations should correspond with peak citations.

It might be of interest to examine the characteristics of the diffusion process of scientific publications. Preferably, these characteristics should allow for a comparison across journals, across disciplines, and over time. Important features of citations data are the time between publication and peak citations and the total cumulative citations at some moment as a fraction of total eventual cumulative citations. The model to be used in this paper allows for the estimation of these two features, and others. Additionally, it would be interesting to see if these features can be explained, for example, by characteristics of the article or the editorial process, assuming one

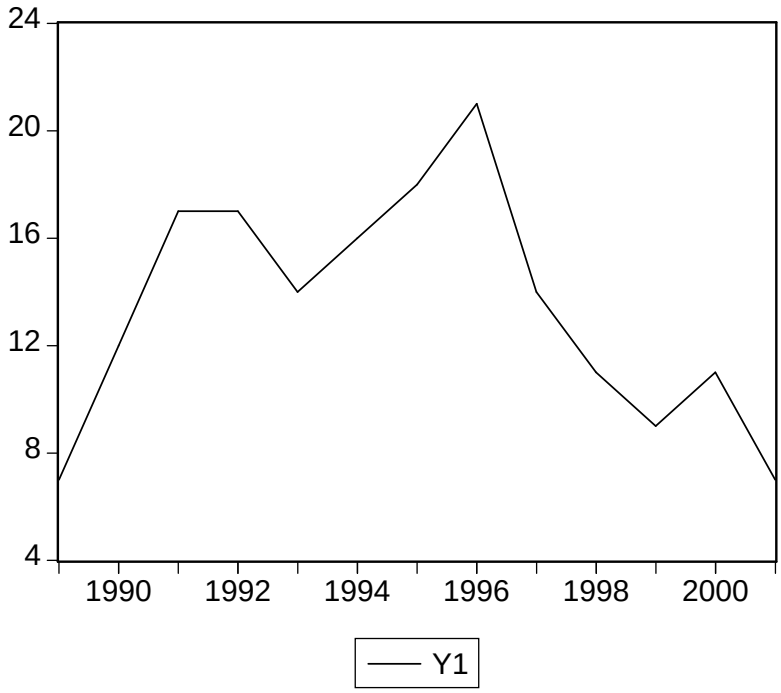

Figure 1: Number of citations of Aumann's article

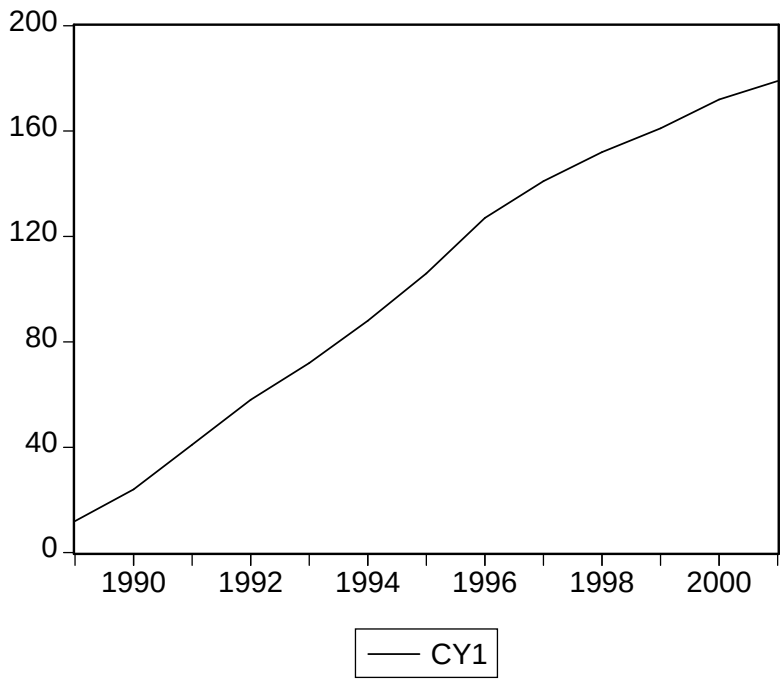

Figure 2: Number of citations of Aumann's article, cumulative up to December 2001

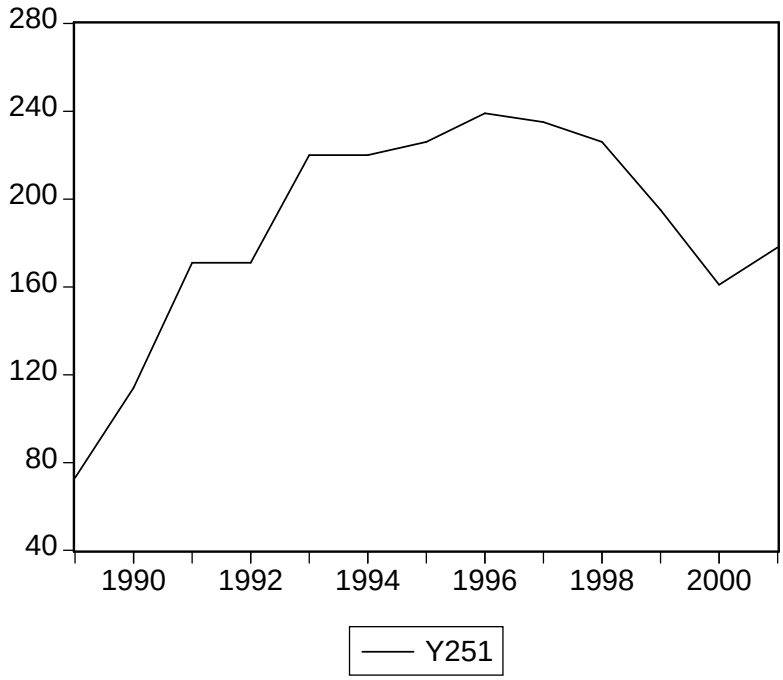

Figure 3: Number of citations of Engle and Granger's article

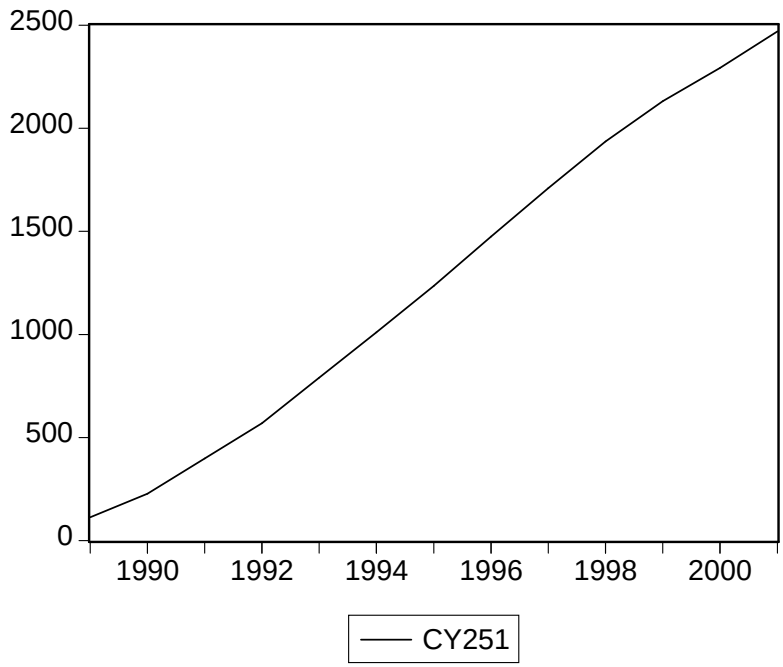

Figure 4: Number of citations of Engle and Granger's article, cumulative up to December 2001

has reliable data. This can provide useful information for a journal, for example, as it can seek to modify its editorial policy. It can also be relevant for comparing journals, where these journals may differ in the total cumulative citations per article. It may also be used to examine how and why certain articles are more successful than others, thereby perhaps providing empirical support for the insights in Van Dalen and Henkens (2001) and Klamer and van Dalen (2002), among others.

An S-shaped diffusion pattern can be described by various models, but the most popular model in the area of technological forecasting is the so-called Bass (1969) model. This model is grounded in diffusion theory, and its empirical representations are rather easy to use, see Mahajan, Muller and Bass (1993), among others. Additionally, the model contains only three parameters, which each have a sensible interpretation. Functions of these parameters can be used to estimate the timing of peak citations and the amount of cumulative citations at the time of this peak. The latter is important, as one can examine whether this timing and amount are equal across various citations series.

The outline of this paper is as follows. In Section 2, I discuss the Bass model, and, in particular, a convenient empirical representation for testing, for example, common inflection points across two or more Bass models. In Section 3, I apply this model to the articles which appeared in the 1987 volume Econometrica (55). As the present paper should foremost be viewed as a pilot study, the analysis is constrained to only these articles. It seems that already a few interesting phenomena can be observed. In Section 4, I outline various topics for further research.

2 A diffusion model

This section concerns a discussion of the Bass diffusion model. Next, some practical matters are discussed, with a focus on a useful empirical representation. This is a rewritten version of the commonly considered model, and, although nonlinear in its parameters, it can easily be programmed in currently available statistical packages.

2.1 Representation

Denote C_t, $t = 1, 2, \ldots, T$, as the citations in the period running from $t-1$ to t, and denote CC_t as the cumulative citations up to and including time t. In the empirical application below, t amounts to years and T is equal to 14, which are commonly found settings for the use of the Bass model in practice. The Bass diffusion theory states that new adoptions (or citations, as in the present paper), given that they have not been adopted before, are a function of a constant and the number of cumulative adoptions. In the notation to be relevant below, the basic Bass model reads as

$$\frac{C_t}{m - CC_{t-1}} = p + \frac{q}{m} CC_{t-1}, \tag{1}$$

where m is the saturation level, p is the coefficient of so-called external influence, measuring a constant likelihood of adoption, and q is called the coefficient of so-called internal influence, measuring the effect of other and previous adoptions on the likelihood of current adoption. All parameters take positive values.

To estimate the model parameters one can write (1) as

$$C_t = \alpha_1 + \alpha_2 CC_{t-1} + \alpha_3 CC_{t-1}^2, \tag{2}$$

where $\alpha_1 = pm$, $\alpha_2 = q - p$ and $\alpha_3 = -\frac{q}{m}$. A commonly applied empirical representation is (2) with an added error term ε_t, which is usually assumed to have mean zero and common variance σ^2. The application of nonlinear least squares to (2) gives the estimators for the parameters p, q and m.

The original Bass model concerns an expression like (2), but then in continuous time. This is a differential equation, and when this is solved, one obtains an expression which suggests that the cumulative process has an S-shaped pattern indeed, see Appendix 1 for more details. Given the solution to the differential equation, one can derive that the point of inflection of CC_t and hence the peak of citations C_t occurs at time T^*, which is equal to

$$T^* = \frac{1}{p+q} \log(\frac{q}{p}), \tag{3}$$

which assumes that the parameter q is larger than p for the inflection point to occur within sample. Notice that this inflection point is a function of p and q only. At the

time of peak citations, the cumulative citations m^* equal

$$m^* = m(\frac{1}{2} - \frac{p}{2q}). \tag{4}$$

Note that m^* is always smaller than $\frac{m}{2}$, that is, the S-shaped pattern implied by the Bass model is not symmetric. In practice, one would be interested in the fraction $\frac{m^*}{m}$, which is the fraction cumulative citations at peak time of the total amount of eventual citations. Finally, the peak citations are equal to

$$\frac{m}{4q}(p+q)^2, \tag{5}$$

see Bass (1969), and see Appendix 1 for a few details.

2.2 Practical matters

One can be interested in the parameters p, q, and m. There are various studies which aim to link the estimated values for various diffusion processes with explanatory variables, see for example Parker (1994) for a survey of the practical use of the Bass model. Note, however, that the parameters p and q have a nonlinear effect on how C_t and CC_t evolve over time, see for example the expression for peak citations above. Hence, it is unclear what a shift of p to, say, $2p$ means for the location of peak citations, and consequently whether linear regression models explaining estimated parameters are easy to interpret along these lines.

It seems perhaps more useful not to consider p and q, but the parameters T^* and m^* directly, as linear functions of these parameters seem more easy to interpret. The basic parameters can be obtained from applying nonlinear least squares to (2), and estimators for $\hat{p}$, $\hat{q}$, and $\hat{m}$ can be derived from $\hat{\alpha}_1$, $\hat{\alpha}_2$ and $\hat{\alpha}_3$. The delta method gives the estimated standard errors. If one is interested in the estimates of the inflection point and the number of cumulative citations at the peak, one can again use $\hat{p}$, $\hat{q}$, and $\hat{m}$ to compute $\hat{T}^*$ and $\hat{m}^*$, again using the delta method.

When examining if estimated Bass models have features in common, one can see if there are common values of saturation levels or common values of internal and external influence parameters. However, when one aims to test if two or more inflection points occur at the same time, the p and q parameters do not have to be

equal across equations. For example, it suffices for two citations series that

$$\frac{1}{p_1+q_1}\log(\frac{q_1}{p_1}) = \frac{1}{p_2+q_2}\log(\frac{q_2}{p_2}). \tag{6}$$

This restriction is rather inconvenient, as it cannot be written as a closed-form expression for, say, p_1 being a function of the other three parameters. To circumvent this problem, one might better consider the diffusion model when it is already written in terms of the parameters m, m^* and T^*. In Appendix 2 it is shown that the parameters α_1, α_2 and α_3 in (2) can be written as the following functions of m, m^* and T^*, that is,

$$\alpha_1 = \frac{m(2m^*-m)}{T^*(m-m^*)}\log(1-\frac{2m^*}{m}), \tag{7}$$

$$\alpha_2 = \frac{-m^*}{T^*(m-m^*)}\log(1-\frac{2m^*}{m}), \tag{8}$$

and

$$\alpha_3 = \frac{1}{2T^*(m-m^*)}\log(1-\frac{2m^*}{m}). \tag{9}$$

These parameter restrictions can easily be programmed using a standard statistical package like Eviews 4.1. Also, the subsequent expressions can be used to compare inflection points across equations, and to test whether T^* or $\frac{m^*}{m}$ are equal across citations series.

3 Econometrica 1987

This section considers an application of the ideas discussed above to observations on citations to articles in volume 55 of Econometrica. First, I discuss some general features of the data. Next, I will discuss the estimation results, and finally, I will correlate estimated parameters with various possibly explanatory variables.

3.1 The data

Econometrica is the leading journal in econometrics. It has the highest impact score (according to the Social Science Citation Index) across econometrics journals, and

Table 1: Descriptive statistics of Econometrica 1987 articles

Variable	Mean	Median	Max.	Min.	Sd
All articles (72)					
Pages	17.597	17	35	2	8.577
Authors	1.625	1	4	1	0.721
Note	0.194	0	1	0	0.399
Articles used for modeling (41)					
Pages	18.683	17	35	5	7.983
Authors	1.512	1	3	1	0.597
Note	0.122	0	1	0	0.331
Citations, 2001	159.976	42	2470	9	408.640
Citations, 2001, logs	3.974	3.738	7.811	2.197	1.288

it has maintained this high level for decades. Volume 55 of Econometrica (1987) contains a few classic articles, which are still cited today. Examples are the two articles of Aumann and Engle and Granger mentioned before, where the last article receives 2470 citations for the sample considered. The data are obtained from the Web of Science (of the International Scientific Institute) in January and February 2002. The data concern all annual citations in the years 1988 through 2001, hence the sample size of 14. Table 1 provides some key statistics of the articles.

Volume 55 of Econometrica contains 72 articles in 6 issues. As the citations data are observed per year, no distinction between these issues is made in subsequent analysis, in terms of timing of publication. The first panel of Table 1 gives key statistics on the number of pages, the number of authors and on whether the article appeared as a note or not. The mean length of an article is about 18 pages, and in many cases, the papers are single-authored. The correlation between the number of pages and the number of authors is 0.085, between pages and note it is -0.690 and between authors and note it is -0.086.

In the subsequent empirical analysis, not all 72 articles will be used. There are 8 articles, which have less than 7 citations over the 14 years, and it seems unwise to consider empirical Bass models for these series. Next, there are 23 cases (of the

remaining 68) for which the Bass model does not seem to fit well. Usually this means that the q parameter gets estimated below zero, or that the q parameter gets estimated below p, or that there is simply no convergence of the estimation routine. In all other 41 cases, the Bass model fits the data well, and this amounts to 64% of all articles.

The second panel of Table 1 contains the characteristics of these 41 articles. The first three rows of this second panel suggest that the characteristics of these 41 articles are not much different from those of the 72 articles, and hence I assume the 41 articles constitute a representative sample. The last two rows of Table 1 give the cumulative citations in 2001 and this variable in natural logs. The distribution of these cumulative citations is clearly skewed, and a log transformation seems to render more symmetry. The mean of the total citations in 2001 is about 160, and this is close to the exemplary case of the citations to Aumann's paper.

3.2 Estimation results

For each of the 41 series, I estimate the parameters of the basic Bass model, in the original format (2) and in the format as in (2) with (7), (8) and (9), which gives the time of the peak and the cumulative citations (as a fraction of the estimated total number of citations) at this peak. A summary of these estimation results appears in Table 2.

The average value of the estimated saturation levels is close to 200, while the median is about 52. The saturation level of the most cited Engle-Granger paper is about 2940, which suggests that this paper is currently getting close to this level. The third row of Table 2 contains an estimate of the rate of the cumulative citations in 2001 over the estimated saturation level. This number is close to 0.850, which suggests that Econometrica Volume 55 articles take 14 years to obtain 85% of their total citations. Note that it can happen that the saturation level is estimated to be lower than the actual cumulative citations in 2001 (see Max is 1.026). This might be due to the problems mentioned in van den Bulte and Lilien (1997).

The value of p is estimated as 0.055 on average, while the average estimated q equals 0.239. These values are often found in practice for the Bass model, see for ex-

Table 2: Descriptive statistics of parameter estimates in Bass diffusion models

Variable	Mean	Median	Max.	Min.	Sd
Unrestricted Bass models					
Saturation level	197.827	52.210	2938.130	8.890	494.441
Saturation level, logs	4.156	3.955	7.986	2.185	1.347
Rate in 2001	0.847	0.885	1.026	0.545	0.146
External effect (p)	0.055	0.046	0.137	0.018	0.035
Internal effect (q)	0.239	0.237	0.444	0.075	0.089
Inflection point (in years)	5.674	5.411	11.308	0.459	2.872
Fraction (as % of saturation level)	0.364	0.402	0.468	0.041	0.108
R^2	0.356	0.310	0.881	0.024	0.223
Restricted fractions at peak					
Saturation level	192.908	46.781	2948.319	8.674	498.509
Saturation level, logs	4.099	3.845	7.989	2.160	1.339
Rate in 2001	0.894	0.923	1.139	0.509	0.138
Inflection point	6.604	6.421	11.417	3.334	2.023

ample Parker (1994), thereby tentatively confirming that a citations process mimics characteristics of the diffusion of, for example, new durable consumer products. The inflection point seems to occur approximately at 5.5 years after publication. Finally, the R^2 of the estimated models is on average equal to 0.356, with cases where this value is as high as 0.881 or as low as 0.024.

The dispersion around the estimated inflection points and the size of the cumulative citations at the peak does not seem very large, and hence one might now be interested in seeing whether the 41 series have features of the Bass model in common. The $\chi^2(40)$ test for the hypothesis that all fractions of total citations at the peak are the same obtains the insignificant value of 14.344. In contrast, the $\chi^2(40)$ test for equality of the inflection points obtains the value of 226.831, which is significant at the 1% level. Imposing the restriction that all fractions are equally large, the fraction gets estimated to be equal to 0.437, with a standard error of 0.002. The corresponding estimation results appear in the second panel of Table 2. With this restriction, the main empirical conclusion for the Econometrica articles is that after

Table 3: Regression results of estimated parameters on explanatory variables in restricted Bass model, with estimated heteroskedasticity-consistent standard errors in parentheses.

Dependent variable	Intercept	Authors	Pages	Note
Saturation level (logs)	2.114 (1.020)	0.390 (0.397)	0.069 (0.028)	0.846 (0.991)
	3.263 (0.549)		0.045 (0.026)	
Point of inflection	3.899 (1.423)	0.417 (0.458)	0.107 (0.053)	0.643 (1.221)
	4.974 (0.803)		0.087 (0.036)	
Rate in 2001	1.099 (0.105)	-0.024 (0.032)	-0.008 (0.004)	-0.091 (0.091)
	1.006 (0.058)		-0.006 (0.003)	

14 years, the articles are at about 89% of their total cumulative citations, and that citations peak at 6.5 years on average, and that the fraction of total citations is then about 44%.

3.3 Explanatory factors

Finally, it may be interesting to see if the estimated key parameters of the 41 Bass models can be explained by characteristics of the articles. For this purpose, I regress the estimated saturation levels, the estimated inflection points and the estimated rates in 2001, all based on the 41-equation Bass model with restricted fraction at the inflection point, on an intercept, the number of authors, the number of pages and a 1/0 dummy variable which equals 1 if the article is a note. The estimation results for the full models, and for the simplified models obtained by deleting insignificant parameters, are displayed in Table 3.

The estimation results in Table 3 all point towards the same conclusion. Whether the article is a note or not or the number of authors does not matter much. However, longer papers give more citations in the long run. Also, for these papers it takes longer for the point of inflection (or, peak citations) to occur, and the difference between cumulative citations in 2001 and the estimated saturation level is larger. Hence, more citations are to be expected.

4 Conclusion

This paper has illustrated that the salient features of citations to scientific publications can be captured by a Bass type diffusion model. Functions of the parameters have a clear-cut interpretation in terms of peak citations and relative cumulative citations. An additional property of the model is that these features can be estimated even though the diffusion process is still going on, that is, the total number of eventual citations can be estimated and does not have to be observed. When the model is written in terms of an inflection point and the fraction of total citations at this point, the model can easily be used to see if citations series have these features in common. Also, as there are not many annual observations for each case, it can enhance the statistical relevance of the model.

For the case of Econometrica 1987, it was found that the peak citations on average occur at 6.5 years, and that after 14 years there is a gap of about 11% between cumulative citations and the estimated saturation level. Even though not all citation series could be described by a Bass model, the obtained results seem representative for all articles. These features could be partly explained by the size of the articles, where a common finding was that more pages lead to more citations with a later peak.

As indicated in the introduction, this paper can be viewed as a pilot study, as there are ample opportunities to extend the ideas in this paper in various directions. It would be interesting to link the estimated features of the citation process to aspects of the editorial process, provided that one has the relevant data. Also, one might want to search for more explanatory factors than the ones considered here. In terms of econometric methods, one might also want to refine the inclusion of explanatory factors by taking into account the intrinsic randomness of the estimated parameters.

In terms of time and size, the empirical study can be extended to concern more years of the same journal, in order to see if key parameters are constant over time. It would be interesting to test if the inflection point occurs increasingly earlier, due to an increased competition across journals or an increased research intensity. Of course, one can also compare the features across journals in the same year, or

even better, across journals and across years. As such, one can identify perhaps changing competitive structures across journals. Finally, one can compare these features across disciplines. It is the intention to take up these topics in further research.

Appendix 1

Denote $f(t)$ as the number of adoptions at time t, and $F(t)$ as the cumulative number of adoptions at time t. The solution to the continuous time Bass differential equation

$$f(t) = pm + (q - p)F(t) - \frac{q}{m}F(t)^2 \tag{10}$$

is equal to

$$F(t) = m\frac{1 - \exp(-(p+q)t)}{1 + \frac{q}{p}\exp(-(p+q)t)}. \tag{11}$$

Using that

$$\exp(-(p+q)t) = \frac{p}{q}\exp(-(p+q)(t - T^*)), \tag{12}$$

where $T^* = \frac{1}{p+q}\log\frac{q}{p}$, one can write $F(t)$ as

$$F(t) = m\frac{1 - \frac{p}{q}\exp(-(p+q)(t - T^*)}{1 + \exp(-(p+q)(t - T^*))}. \tag{13}$$

When t exceeds T^*, then $F(t)$ goes to m, while when it is much smaller, $F(t)$ goes to zero.

Using the same notation, the derivative of $F(t)$ to t is

$$f(t) = m\frac{p(p+q)^2\exp(-(p+q)t)}{(p + q\exp(-(p+q)t))^2}, \tag{14}$$

for which it is easy to verify that its derivative to t is equal to zero at the inflection point. With the expression for the inflection point it can be written as

$$f(t) = m\frac{(p+q)^2\exp(-(p+q)(t - T^*))}{q(1 + \exp(-(p+q)(t - T^*)))^2}. \tag{15}$$

It is now easy to see that

$$F(t = T^*) = m(\frac{1}{2} - \frac{p}{2q}) \tag{16}$$

and

$$f(t = T^*) = m\frac{(p+q)^2}{4q} \tag{17}$$

Appendix 2

The initial parameters are pm, $q-p$ and $-\frac{q}{m}$, where p, q and m are all assumed to be positive. Taking these combinations of parameters jointly, one can can arrive at expressions of the α parameters in terms of the saturation level m, the cumulative amount of citations at the inflection point, m^*, and the inflection point itself T^*.

It is given that

$$m^* = m(\frac{1}{2} - \frac{p}{2q}). \tag{18}$$

As $2m^* = m(1-\frac{p}{q})$, and so $\frac{2m^*}{m} = 1-\frac{p}{q}$, one has

$$\frac{p}{q} = 1 - \frac{2m^*}{m}. \tag{19}$$

It is also given that

$$T^* = -\frac{1}{p+q}\log\frac{p}{q} \tag{20}$$

Using (19), one obtains

$$p+q = -\frac{1}{T^*}\log(1-\frac{2m^*}{m}). \tag{21}$$

From (19), it also follows that $p = q(1-\frac{2m^*}{m})$. Substituting this in (21) gives

$$q - \frac{2qm^*}{m} + q = -\frac{1}{T^*}\log(1-\frac{2m^*}{m}), \tag{22}$$

or

$$q = \frac{-m\log(1-\frac{2m^*}{m})}{2T^*(m-m^*)}, \tag{23}$$

which is the first essential result.

Substituting (23) in (21) gives

$$p = \frac{2m^*-m}{2(m-m^*)}\frac{1}{T^*}\log(1-\frac{2m^*}{m}). \tag{24}$$

Using these last two expressions gives the required results.

References

Bass, F.M. (1969), A new-product growth model for consumer durables, *Management Science*, 15, 215-227.

Klamer, A. and H.P. van Dalen (2002), Attention and the art of scientific publishing, *Journal of Economic Methodology*, to appear.

Mahajan, V., E. Muller and F.M. Bass (1993), New-product diffusion models, in *Handbook of Marketing*, J. Eliashberg and G.L. Lilien (eds.), Amsterdam: North-Holland, 349-408.

Parker, P.M. Aggregate diffusion forecasting models in marketing: A critical review, *International Journal of Forecasting*, 10, 353-380.

van Dalen, H.P. and K. Henkens (2001), What makes a scientific article influential?, *Scientometrics*, 50, 455-482.

van den Bulte, C. and G.L. Lilien (1997), Bias and systematic change in the parameter estimates of macro-level diffusion models, *Marketing Science*, 16, 338-353.