Modeling the diffusion of scientific publications

Dennis Fok[*]
Econometric Institute
Erasmus University Rotterdam

Philip Hans Franses
Econometric Institute and
Department of Marketing and Organization
Erasmus University Rotterdam

ECONOMETRIC INSTITUTE REPORT EI 2005-48

Abstract

This paper illustrates that salient features of a panel of time series of annual citations can be captured by a Bass type diffusion model. We put forward an extended version of this diffusion model, where we consider the relation between key characteristics of the diffusion process and features of the articles. More specifically, parameters measuring citations' ceiling and the timing of peak citations are correlated with specific features of the articles like the number of pages and the number of authors. Our approach amounts to a multi-level non-linear regression for a panel of time series. We illustrate our model for citations to articles that were published in *Econometrica* and the *Journal of Econometrics*. Amongst other things, we find that more references lead to more citations and that for the *Journal of Econometrics* peak citations of more recent articles tend to occur later.

Key words: Diffusion of innovations; Multi-level regression

[*] We thank two anonymous referees for their helpful comments and Merel van Diepen for excellent research assistance. The data and the program used to estimate the parameters can be obtained from the authors. Address for correspondence: P.H. Franses, Econometric Institute, Erasmus University Rotterdam, P.O. Box 1738, NL-3000 DR Rotterdam, The Netherlands, `e-mail: franses@few.eur.nl`.

1 Introduction

Citations to scientific publications like journal articles often show characteristics that bear similarities with the diffusion of a new product. Shortly after publication, there are not many citations. Then, the number of citations starts to grow, and after a few years, citations may peak. Finally, after this peak, citations eventually level off towards zero. The reason for this may vary across articles. The article may become outdated or it may be replaced by better research. On the other hand, it may be the case that the article becomes so well known that citations are not needed anymore. Strictly speaking, one then has an implicit citation process with a total number of citations that approaches infinity. In the present paper, the primary variable of our interest is the number of observed citations, which likely has an upper limit.

A visual characteristic of a typical observed cumulative citation series is that it follows an S-shaped pattern, which starts at zero and levels off to some upper bound. This upper bound can be called the level of maturity or the ceiling. Various models can describe an S-shaped diffusion pattern. Examples of these models are the logistic model, the Gompertz model, the Bass model and various of its generalizations, see Meade and Islam (1998) for a survey, among others.

The model that is most often used in new product diffusion modeling is the Bass (1969) model. The main reason for this is that it finds its origin in a formal theory of product diffusion, and that the model parameters have an easy to understand interpretation in terms of innovation and imitation effects. There are various empirical versions of this model, and these are all rather easy to use, see Mahajan, Muller and Bass (1993) for a survey. The basic Bass model contains only three parameters. Non-linear 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. Hence, diffusion data, when summarized by a Bass model, can be characterized by a small number of parameters. Using these parameters, one can easily compare various diffusion series.

In this paper we examine the characteristics of the diffusion process of scientific publications, where we choose to consider two econometrics journals. More precisely, we consider 411 articles that have been published in *Econometrica* in the years 1987 to 1995 and 116 articles in the *Journal of Econometrics* for 1988 to 1995. We choose *Econometrica* and the *Journal of Econometrics* as they are widely regarded as the leading journals in econometrics. It should be mentioned though that the *Journal of Econometrics* includes many articles which obtain zero or only a few citations, which prohibits the use of a Bass model, and hence the smaller number of included cases. Hence, we expect our empirical results for *Econometrica* to be more reliable.

We aim to describe the citation process over time of these 411 and 116 articles, where we have collected the citations up to and including 2001 for *Econometrica* and up to and including 2002 for the *Journal of Econometrics*. These citations do include self-citations, although the amount of self-citations is not large. For a few cases, we checked the robustness to the inclusion of self-citations, and we did not find strong signs that conclusions change substantially. We consider articles published up to and including 1995, as we find that peak citations typically tend to occur no earlier than after 5 to 6 years. Our decision is guided by the well-known fact that estimation routines for the Bass model deliver very inaccurate estimates if one only has data before the peak citations, see van den Bulte and Lilien (1997).

The data we analyze constitute an unbalanced panel of time series. A direct comparison of the cumulative number of citations over the years would therefore not be fair. It is our aim to provide generalizing statements about the diffusion process of the citations, while correcting for the time the article has been available. Our statements concern the link between the characteristics of the articles and the observable key features of the individual diffusion series. For example, we address the question whether more authors or more references lead to more citations. Also, did the diffusion process change over time? Do more recent articles get cited less often these days, and do peak citations occur more early in the process? When answering these questions we have to keep in mind that recent articles of course have had less opportunities to be cited than articles published earlier. Hence, we aim to summarize the data in a concise way, while preserving the opportunity to say something about all articles jointly.

One approach could be to consider a separate model for each of the articles. The resulting estimates can, in a second round, be regressed on another set of explanatory variables[1]. Strictly speaking, this is not a sound strategy as it is assumed for the second-round model that the estimated parameters are observed regressors, and thereby one assumes their uncertainty to be absent. In other words, this approach leads to too narrow confidence bounds in the second stage-regression model. One way to solve this problem is to resort to correcting the standard errors using for example instrumental variable methods. However, we believe that our approach to be discussed next is simpler as it jointly deals with the two parts of the model. In fact, subsequent statistical inference turns out to be not too complicated.

Our approach is based on the general notion of a multi-level regression model. In this framework, the first-level parameters, which in our case are for example the maturity level and the location of the inflection point, are explicitly seen as functions of a set of regressors and an error term. Next, the parameters in this second level of the model are estimated directly. These parameters concern the relation between the diffusion characteristics and the article features. Estimates of, for example, the maturity level for a specific article can then be obtained using the second-stage parameters. In this paper, we put forward such a multi-level regression model for a panel of diffusion series. We should mention that an additional advantage of such an approach is that it entails possibilities for shrinkage, see also Blattberg and George (1991), among others.

In a sense, our approach bears similarities with that in Talukdar, Sudhier and Ainslie (2002). There are however three important differences. Talukdar et al. (2002) take the parameters in the original Bass model, and link these with a second set of variables. These parameters, however, have a strong non-linear effect on the diffusion process, which renders the parameters in the second-stage regression difficult to interpret. Instead, we focus on (i) the level of maturity, (ii) the fraction of cumulative citations at the peak, and (iii) the timing of the peak. These characteristics are continuous variables with a straightforward interpretation, and this facilitates an easy interpretation of the second-stage parameters. The second

[1] This rough-and-ready approach has been followed in Franses (2003), where only the 1987 volume of *Econometrica* has been analyzed.

important difference is that we rely on a recently developed alternative version of the Bass model, see Boswijk and Franses (2002). This new model deals explicitly with the nature of the error term, which should be heteroskedastic due to the very nature of the type of process. Next, this model includes an additional regressor. The third difference, and also in contrast to Lenk and Rao (1990), is that we rely on Simulated Maximum Likelihood to estimate the parameters.

The outline of this paper is as follows. In Section 2, we start off with a discussion of the single-variable Bass model, and, next, we discuss our multi-level panel model. In Section 3, we apply this model to the data at hand. We discuss some features of the data first, then present the estimation results, which we summarize in a table with, say, prototypical articles. In Section 4, we conclude with remarks.

2 The model

In this section we start off with the representation of the Bass model for a single series. Next, we put forward the representation as advocated in Boswijk and Franses (2002). We then discuss the multi-level model for a panel of diffusion series. Finally, we discuss parameter estimation of this last model.

2.1 Representation

The Bass model assumes a population of m potential adopters, where, in the context of citations, we will associate m with the maturity level. In our context, adopters should be viewed as articles which cite the articles under scrutiny. The maturity level can be viewed as the total number of citations in the long run. For each adopter, the time to adoption is a random variable with a distribution function $F(t)$ and density $f(t)$, such that the hazard rate equals

$$\frac{f(t)}{1-F(t)} = p + qF(t), \tag{1}$$

where p and q are the parameters that determine the shape of the diffusion process. The cumulative number of adopters at time t, denoted by $N(t)$, is a random variable with mean $\bar{N}(t) = E[N(t)] = mF(t)$, where t is measured in continuous time and

E denotes the expectation operator. It can be shown that the function $\bar{N}(t)$ obeys the following differential equation, that is,

$$\bar{n}(t) \equiv \frac{d\bar{N}(t)}{dt} = p[m - \bar{N}(t)] + \frac{q}{m}\bar{N}(t)[m - \bar{N}(t)], \tag{2}$$

see Bass (1969).

In the new product diffusion literature, it is common to interpret the parameter p as the innovation parameter, q as the imitation parameter, and m as the maturity level. Note that these parameters exercise a non-linear impact on the pattern of $\bar{N}(t)$ and $\bar{n}(t)$. Basic characteristics of the diffusion also non-linearly depend on p and q. For example, the inflection point T^* of $F(t)$, which corresponds with the time of peak adoptions, equals

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

A natural question is now how one can translate the theoretical model in (1) into an empirical model with parameters that can be estimated using actual discrete-time data. Bass (1969) proposes to use the cumulative number of adoptions in discrete time (N_t, for $t = 0, 1, 2, ..., T$) and the corresponding increments ($X_t = N_t - N_{t-1}$), and to consider the regression model

$$X_t = pm + (q - p)N_{t-1} - \frac{q}{m}N_{t-1}^2 + \varepsilon_t, \tag{4}$$

where $t = 1, \ldots, T$ refers to a time series measured at discrete intervals. Bass (1969) further assumes that ε_t is a standard white noise error term.

Recently, Boswijk and Franses (2002) modified this Bass regression model by allowing for heteroskedastic errors and by allowing for short-run deviations from the deterministic S-shaped growth path of the diffusion process, as implied by the differential equation in (2). These authors propose to consider

$$dn(t) = \alpha\left[p[m - N(t)] + \frac{q}{m}N(t)[m - N(t)] - n(t)\right]dt + \sigma n(t)^{\gamma}dW(t), \tag{5}$$

where $W(t)$ is a standard Wiener process. The parameter α in (5) measures the speed of adjustment to the path implied by the standard Bass model. Additionally, by introducing $\sigma n(t)^{\gamma}$, there is an allowance for heteroskedasticity. A useful choice is to set $\gamma = 1$. Heteroskedasticity is relevant as, towards to endpoints of the diffusion process, one has more certainty about the likely realizations of the citations and cumulative citations. Boswijk and Franses (2002) derive that the discretization of this continuous time model is

$$\Delta X_t = \beta_1 + \beta_2 N_{t-1} + \beta_3 N_{t-1}^2 + \beta_4 X_{t-1} + X_{t-1}\varepsilon_t, \tag{6}$$

where Δ denotes the first differencing operator, and where

$$\begin{aligned} \beta_1 &= \alpha p m, \quad \beta_2 = \alpha(q-p), \\ \beta_3 &= -\alpha \tfrac{q}{m}, \quad \beta_4 = -\alpha, \end{aligned} \tag{7}$$

which shows that all parameters in (6) depend on α.

2.2 Towards a multi-level regression

In our present application, we have an unbalanced panel of diffusion time series, and it is our aim to model these series jointly. In panel format, our model is

$$\Delta X_{i,t} = \beta_{1,i} + \beta_{2,i} N_{i,t-1} + \beta_{3,i} N_{i,t-1}^2 + \beta_{4,i} X_{i,t-1} + X_{i,t-1}\varepsilon_{i,t}, \tag{8}$$

where $i = 1, \ldots, N$ concerns a specific article, and $t = 1, \ldots, T_i$ with T_i the number of years in which article i could have been cited. As before, the β parameters are transformations of the underlying characteristics of the diffusion process, that is,

$$\begin{aligned} \beta_{1,i} &= \alpha_i p_i m_i, \quad \beta_{2,i} = \alpha_i(q_i - p_i) \\ \beta_{3,i} &= -\alpha_i \frac{q_i}{m_i}, \quad \beta_{4,i} = -\alpha_i. \end{aligned} \tag{9}$$

As the effects of p and q on the diffusion patterns are highly non-linear, we propose to focus on the inflection point, that is, the timing of the peak citations, T_i^*, and the level of the cumulative citations at the peak divided by m_i, denoted as f_i. Note that the Bass model imposes that $0 \leq f \leq \frac{1}{2}$. The link between p_i and q_i and the inflection point parameters is given by

$$p_i = (2f_i - 1)\frac{\log(1-2f_i)}{2T_i^*(1-f_i)}, \quad q_i = -\frac{\log(1-2f_i)}{2T_i^*(1-f_i)}, \tag{10}$$

see Franses (2003).

Combining (9) and (10), we can express the parameters in (8) in terms of the characteristics of the diffusion process. That is, we specify $\beta_{1,i}, \ldots, \beta_{4,i}$ as a function of the total number of citations (m_i), the fraction of cumulative citations at the inflection point (f_i), the time of the inflection point (T_i^*), and the speed of adjustment (α_i) of $X_{i,t}$ to the equilibrium path. These functions will be denoted as $\beta_{k,i} = \beta_k(m_i, f_i, T_i^*, \alpha_i)$.

In this paper we are interested in explaining the characteristics of the diffusion process by the characteristics of the publications. That is, we want to relate m_i, f_i, T_i^* and α_i to observable features of the articles. As mentioned before, a first and obvious approach is to consider a second-stage regression model in which the estimated first-round parameters are the dependent variables. There are two main problems with this approach. The first is that the estimated parameters from the first stage regression would be erroneously treated as given, while in reality they are, so-called, generated regressors. One may now consider the literature on generated regressors, but we believe that our multi-level model below is much simpler. A second drawback is that it can happen that the model in (8) does not deliver reliable estimation results for all N cases. This means that in some individual cases the uncertainty of parameter estimates is very large, that is, that implausible point estimates can be delivered, which in turn may lead to implausible results in the second-stage regression model.

Given this, we prefer to consider a multi-level non-linear regression model for the panel of diffusion series. The model consists of two levels and it is non-linear in its parameters, as we correlate the maturity level, timing of peak and cumulative citations at the peak, with explanatory variables. In our notation, the model is

$$\begin{aligned}\Delta X_{i,t} = \beta_1(m_i, f_i, T_i^*, \alpha_i) + \beta_2(m_i, f_i, T_i^*, \alpha_i)N_{i,t-1} + \\ \beta_3(m_i, f_i, T_i^*, \alpha_i)N_{i,t-1}^2 + \beta_4(m_i, f_i, T_i^*, \alpha_i)X_{i,t-1} + X_{i,t-1}\varepsilon_{i,t},\end{aligned} \quad (11)$$

where $\varepsilon_{i,t} \sim N(0, \sigma_i^2)$ with

$$\begin{aligned}\log(m_i) &= Z_i'\theta_1 + \eta_{1,i}, & \log(\frac{2f_i}{1-2f_i}) &= Z_i'\theta_2 + \eta_{2,i}, \\ \log(T_i^*) &= Z_i'\theta_3 + \eta_{3,i}, & \alpha_i &= Z_i'\theta_4 + \eta_{4,i}, \\ \log\sigma_i^2 &= Z_i'\theta_5 + \eta_{5,i}, \end{aligned} \quad (12)$$

where the Z_i vector contains an intercept and explanatory variables. We assume that $\eta_i = (\eta_{1,i}, \eta_{2,i}, \eta_{3,i}, \eta_{4,i}, \eta_{5,i})' \sim N(0, \Sigma_\eta)$. Furthermore, the disturbances $\varepsilon_{i,t}$

are serially independent and uncorrelated across articles. Note that the logit-type transformation of f_i ensures that $0 \leq f_i \leq \frac{1}{2}$.

2.3 Parameter estimation

The parameters in our multi-level model are now contained in θ_1 to θ_5 and Σ_η. Estimates of these parameters can be obtained through maximum likelihood estimation. The likelihood function of the model equals

$$\ell = \prod_{i=1}^{N} \int_{\eta_i} \ell_i(\eta_i)\phi(\eta_i; 0, \Sigma_\eta) d\eta_i, \tag{13}$$

with

$$\ell_i(\eta_i) = (2\pi)^{-T_i/2}\sigma_i^{-T_i} \exp(-\frac{1}{2}\sum_{t=1}^{T_i}\left(\frac{e_{i,t}(\eta_i)}{X_{i,t-1}\sigma_i}\right)^2) \tag{14}$$

where $\phi(\eta_i; 0, \Sigma_\eta)$ denotes the density function of a 5-variate normal distribution with mean 0 and covariance matrix Σ_η evaluated at η_i, $\ell_i(\eta_i)$ is the likelihood contribution of article i conditional on η_i, and $e_{i,t}(\eta_i)$ is the (unstandardized) residual of (11) given η_i. Note that σ_i^2 also depends on η_i, as from (12) it follows that $\sigma_i^2 = \exp(Z_i'\theta_5 + \eta_{5,i})$.

The integral in (13) cannot be solved analytically. To obtain parameter estimates we opt for Simulated Maximum Likelihood, see for example Gourieroux and Montfort (1996). To reduce the variance of the likelihood simulator we use Importance Sampling, see Kloek and van Dijk (1978) and Geweke (1989). To this end, we rewrite the likelihood function as

$$\ell = \prod_{i=1}^{N} \int_{\tilde{\eta}_i} \frac{\ell_i(\Sigma_\eta^{1/2}\tilde{\eta}_i)\phi(\tilde{\eta}_i; 0, \mathbf{I})}{g(\tilde{\eta}_i; m_i, S_i)} g(\tilde{\eta}_i; m_i, S_i) d\tilde{\eta}_i, \tag{15}$$

where $\Sigma_\eta^{1/2}$ is the Choleski decomposition of Σ_η and where $g(\tilde{\eta}_i; m_i, S_i)$ denotes the importance function which is set to the normal density with mean m_i and variance S_i. To approximate the likelihood we use

$$\tilde{\ell} = \prod_{i=1}^{N} \frac{1}{K} \sum_{k=1}^{K} \frac{\ell_i(\Sigma_\eta^{1/2}\tilde{\eta}_i^{(k)})\phi(\tilde{\eta}_i^{(k)}; 0, \mathbf{I})}{g(\tilde{\eta}_i^{(k)}; m_i, S_i)}, \tag{16}$$

where $\tilde{\eta}_i^{(k)}$ is a draw from $g(\tilde{\eta}_i; m_i, S_i)$. To reduce the sampling variance we set m_i and S_i such that the importance function closely resembles the likelihood contribution conditional on $\tilde{\eta}_i$ for each article. Appropriate values for m_i and S_i can be obtained using the following iterative scheme, that is, (i) set $m_i = 0$ and $S_i = \mathbf{I}$, (ii) simulate $\tilde{\eta}_i^{(k)}$, $k = 1, \ldots, K$ from $g(\tilde{\eta}_i; m_i, S_i)$, (iii) calculate

$$w_i^{(k)} = \frac{\ell_i(\Sigma_\eta^{1/2}\tilde{\eta}_i^{(k)})\phi(\tilde{\eta}_i^{(k)}; 0, \mathbf{I})}{g(\tilde{\eta}_i^{(k)}; m_i, S_i)}, \tag{17}$$

(iv) update location and scale parameters

$$m_i = \frac{\sum_{k=1}^K w_i^{(k)}\tilde{\eta}_i^{(k)}}{\sum_{k=1}^K w_i^{(k)}}, \quad S_i = \frac{\sum_{k=1}^K w_i^{(k)}(\tilde{\eta}_i^{(k)} - m_i)(\tilde{\eta}_i^{(k)} - m_i)'}{\sum_{k=1}^K w_i^{(k)}}, \tag{18}$$

and (v) go to (ii). In practice only a few iterations are necessary to obtain appropriate values for m_i and S_i. Finally, parameter estimates of the model are obtained by numerically maximizing $\log\tilde{\ell}$ over θ_1 to θ_5 and the parameters contained in Σ_η. As the optimal location and scale parameters of the importance function depend on the vector of parameters at which the likelihood is evaluated, m_i and S_i will have to be updated a few times during the maximization.

Under the usual regularity conditions, the SML estimator is consistent for $N \to \infty$ and $K \to \infty$. Furthermore, the estimator is asymptotically normal distributed. The standard errors can be computed using the so-called sandwich or *robust asymptotic covariance matrix* estimator recommended by McFadden and Train (2000), see Newey and McFadden (1994) for a general discussion. In our two-stage model the covariance matrix of the parameter estimates can be estimated by

$$\widehat{\text{Var}}(\omega) = \left[-\frac{\partial \log\tilde{\ell}}{\partial\omega\partial\omega'}\right]^{-1}\left[\sum_{i=1}^N\left(\frac{\partial\log\tilde{\ell}_i}{\partial\omega}\right)\left(\frac{\partial\log\tilde{\ell}_i}{\partial\omega}\right)'\right]\left[-\frac{\partial\log\tilde{\ell}}{\partial\omega\partial\omega'}\right]^{-1}, \tag{19}$$

where the vector ω contains all parameters of the model, including those in Σ_η, and where $\tilde{\ell}_i$ denotes the (simulated) likelihood contribution of article i. This estimator of the covariance matrix is to be preferred over the usual negative inverse of the Hessian of the likelihood, as the latter underestimates the covariance matrix for finite K as shown by Newey and McFadden (1994).

3 Empirical results

In this section, we apply our multi-level non-linear regression model to the panels of articles in *Econometrica* and the *Journal of Econometrics*. First, we discuss some descriptive statistics of the data. Next, we present the estimation results for our model.

3.1 The data

We collected annual citations data using the Social Science Citation Index (SSCI) for articles published in *Econometrica* and the *Journal of Econometrics.* For *Econometrica*, the first volume we analyze is 1987 and we have the citations up to and including 2001. For the *Journal of Econometrics* we start our analysis in 1988 and consider citations up to and including 2002.[2] Preliminary analysis of individual series indicated that peak citations tend to occur 5 to 7 years after publication. It is well known from the new product diffusion literature that it is very difficult, if not impossible, to estimate the location of the inflection point of the diffusion if it did not yet occur, see for example Mahajan, Muller and Bass (1993). Hence, we decide not to include articles published after 1995, so all articles could receive at least 6 years of citations. Finally, we include only those articles which received a minimum amount of 10 citations, as otherwise there would be difficulties estimating the model parameters. Hence, all forthcoming results concern the citations to an article, given that there are enough citations.

In Tables 1 and 2 we summarize some descriptive statistics of the 411 relevant articles for *Econometrica.* These statistics concern the number of pages, the number of authors (with an obvious minimum of 1), the number of references and the number of citations cumulative up to and including 2001. The first three variables will be included as the explanatory variables (Z_i) in the second level of our model.

In Tables 3 and 4 we give the same descriptive statistics for the *Journal of Econometrics.* We see that there are not many differences across the two sets of tables, in terms of the number of pages, authors and references. And, similar to *Econometrica*, we also note that the number of pages has increased over time. It

[2] We have easy access to citations data for both journals for the period from 1988 onwards. However, we are aware that in 1987 there were two publications in *Econometrica* with exceptional amounts of citations, that is, close to 60 and 25 times the median value. This is, relatively speaking, far more than any paper in the 1987 issues of the *Journal of Econometrics.* Hence, we decided to include this year of *Econometrica* articles as well, even though data collection in this case involved rather time-consuming manual labour.

should again be mentioned here that we only consider 116 articles in the *Journal of Econometrics* as only these receive a substantial amount of citations.

Clearly, the most cited article in the last 15 years in *Econometrica* is the paper on error correction and cointegration by Robert Engle and Clive Granger. The distribution of the citations in Tables 1 to 4 appears to be rather skewed, hence we also present the median values. The median number of citations seems to decrease over the years, with about 40 in the beginning, and about 20 at the end. This is of course at least partly due to the fact that more recent articles simply could not receive that many citations as older articles. To examine to what extent recent articles truly have a smaller citation potential, we consider our multi-level model, as it allows us to evaluate all diffusion series over the years.

We also observe that the median number of references has increased, and also that articles seem to have become longer. The number of authors seems to be rather constant over time.

From the literature on citations, see for example van Dalen and Henkens (2002) and the cited references therein, we can put forward the following conjectures. First, longer articles with more references and also articles with more authors tend to get more citations. The latter can be a result of self-citations, but it can also be due to network effects as more authors can give more presentations at seminars and conferences and as they each may have more students who might cite their work. This means that the corresponding variables are expected to have a positive effect on $\log m_i$. More cumulative citations, at the end of the diffusion process, can be obtained by having an early peak with low relative citations, such that it takes a longer time and many citations to eventually arrive at the maturity level. However, it can also be obtained by a late peak with a high number of relative cumulative citations. In the first case, the journal under scrutiny can be seen as a journal with an immediate impact on a small group of early adopters of an article and a larger group of the so-called late majority. In the second case, there is a larger group of early adopters.

Finally, the literature on scientific citations also suggests that more recent articles are cited less often. This is supposed to be due to the publication pressure, which has established that the editorial process slows down, see also Ellison (2002), while also the number of possible publication outlets has increased enormously. Indeed, when *Econometrica* started in 1933, there were just a few high quality journals with econometric articles, and nowadays there are many more.

A key feature of our approach is that, by focusing on the inflection point and the number of cumulative citations at this point and correlating these features with characteristics of the papers, we facilitate a comparison across papers. Hence, even though the final maturity level may differ substantially, the shape of the diffusion process may be rather similar across papers. However, the second-stage regression for the maturity level may be affected by large values of only a few papers. In fact, for this sample it may be that the Engle-Granger (1987) paper exercises an exceptional influence on the final parameter estimates. To see whether this is the case, we re-estimate the model parameters for all data except for those concerning this paper.

3.2 Estimation results

In Table 5 we report the estimation results for *Econometrica*. In this model we include in Z_i the number of pages, the number of authors, the number of references, a trend variable, and the interaction of the number of authors, the number of pages and the number of references with the trend. For all models, we use $K = 1000$ draws per article to simulate the likelihood. Furthermore, we restrict the covariance matrix Σ_η to be diagonal for computational convenience. Allowing for non-zero off-diagonal elements could well be possible but, in turn, would burden the computations substantially.

The main conclusions that can be drawn from Table 5, are that more authors, more references and more pages lead to more total citations in the end, while these effects get smaller over time. More pages also lead to a later peak of citations and also to more cumulative citations at that peak. Interestingly, the amount of references has a negative impact on these two features, although this only holds for significance levels around 20 per cent. Another result from the model in this table is the strong positive effect of the interaction between references and the trend on the fraction of cumulative citations reached at the moment of peak citations. We also estimated a model for the case without the Engle-Granger article, and we find that the parameter estimates are very similar.

The estimation results for the full model, including the interaction terms, for the *Journal of Econometrics* data are displayed in Table 6. We observe that, generally, the same type of variables has significant relevance for the variables to be explained, as we saw from Table 5. The level at the inflection point does not depend on any explanatory variables. The location of the inflection point seems to depend on the trend and on its interaction with the number of authors. Finally, the last column shows that more certainty about the diffusion process can be achieved for articles with more references, although over time this effect has become smaller.

As is common for models that are non-linear in the parameters, it is not easy to assign specific interpretation to the parameter estimates only. For that reason, we give in Table 7 important descriptive statistics of three typical articles, which are based on the estimation results in Tables 5 and 6. If we keep the number of pages fixed at 20, we see that more authors give more citations and a later peak, while, for *Econometrica*, more references also gives more citations, but now with an earlier peak. If we compare the results for the volumes of 1988 and 1995, we see interesting differences across the journals. Maturity levels for *Econometrica* have decreased over time and the timing of peak citations has not changed substantially. For the *Journal of Econometrics* we see an increase in maturity level and the timing of peak citations. In other words, if *Journal of Econometrics* articles are cited at all, they nowadays are cited more often.

4 Conclusion

In this paper we put forward a new and rather parsimonious model to summarize the salient features of an unbalanced panel with diffusion data. We illustrated this model for the diffusion patterns of *Econometrica* and *Journal of Econometrics* articles. We could see that certain aspects of the articles have an impact on the size of the citations' ceiling, the timing of peak citations, and other features. Additionally, we observed that the impact of these variables could change over the years. To better understand the model implications, we simulated the properties of three hypothetical articles. For the *Journal of Econometrics* we found that citations peak later. For *Econometrica* we found that cumulative citations have decreased over time, while for the *Journal of Econometrics* the reverse effect holds.

A first consequence of our analysis is that one might wish to reconsider the current practice in use by the SSCI. This is that journals are ranked according to citations within 2 years after publication. First of all, it might be that this number of years should not be taken as fixed over the years, but rather that it varies over time. Second, it is likely that journals vary with respect to the citation diffusion of their articles. For example, one might evaluate *Econometrica* on the basis of citations until the average timing of the peak, which is, say, 6 years. Another journal can then be evaluated during a different period. This way one accounts for the possibility that each journal might have a different type of audience with a different citation style. In fact, journals in medicine and physics have an audience that cites immediately and hence the citation scores of their journals are much higher than those in, say, economics or statistics where there is much more delay between publication and citation. One reason for this might be that researchers in medicine for example focus on similar topics due to their acute importance for human health, while researchers in statistics and economics might address a wider range of non-overlapping topics. In sum, to allow for different citation styles across journals and disciplines, one might correct for different time frames between publication and citation, and as such allow for a fairer comparison of journals across disciplines, and perhaps also within a discipline.

The present study suggests various avenues for further research, two of which will be mentioned here. The first is to see if there are generalizing statements to make about citation traditions across disciplines. For now, we only considered two econometrics journals, but one can also consider leading journals in economics, finance, marketing, regional studies and so on. Our model allows for a rather compact description of the citation process, and comparison across disciplines should be possible. The second topic concerns the role of mediating variables, like country, state, age of researcher (in terms of the maturity of career), and various aspects of the refereeing process, like time between submission and eventual publication and the number of referees. One then needs a rather detailed database, and perhaps these can made available by the editorial offices of various journals.

Table 1: Descriptive statistics of *Econometrica* articles, 1987-1991, with cumulative citations up to and including 2001.

Year (number)	Variable	Mean	Median	Min.	Max.	St.dev.
1987 (60)	Pages	19.15	18	5	35	8.109
	Authors	1.63	1.5	1	4	0.730
	References	21.18	20.5	4	50	10.205
	Citations	128.7	39	10	2470[1]	338.9
1988 (49)	Pages	22.43	22	5	36	7.980
	Authors	1.71	2	1	3	0.606
	References	28.31	26	10	80	12.759
	Citations	55.02	41	10	272	50.05
1989 (43)	Pages	25.05	26	5	44	9.741
	Authors	1.67	2	1	3	0.672
	References	27.56	25	6	57	11.252
	Citations	78.37	42	10	604[2]	111.23
1990 (47)	Pages	22	23	3	41	8.676
	Authors	1.81	2	1	5	0.816
	References	25.75	23	2	93	16.026
	Citations	48.55	24	10	269	58.09
1991 (62)	Pages	21.58	22.5	3	42	8.051
	Authors	1.63	1	1	3	0.724
	References	28.61	27.5	4	76	15.283
	Citations	55.02	29	10	624[3]	93.94

[1] This is the famous paper on error correction and cointegration by Robert Engle and Clive Granger. An impressive runner up in that year is the paper by Whitney Newey and Ken West on HAC with 942 citations.

[2] This is the paper on unit roots and structural breaks by Pierre Perron.

[3] This paper is the cointegration paper by Soren Johansen.

Table 2: Descriptive statistics of *Econometrica* articles, 1992-1995, with cumulative citations up to and including 2001.

Year	Variable	Mean	Median	Min.	Max.	St.dev.
1992 (46)	Pages	22.17	22.5	3	42	9.986
	Authors	1.70	2	1	4	0.777
	References	25.85	23	4	103	15.374
	Citations	39.24	23	11	226	42.22
1993 (38)	Pages	26.53	27	2	38	8.598
	Authors	1.63	2	1	3	0.625
	References	32.40	27.5	2	177	27.693
	Citations	54.05	39.5	10	214	51.09
1994 (34)	Pages	29.41	29	6	54	10.890
	Authors	1.91	2	1	4	0.781
	References	35.77	33	13	82	16.423
	Citations	33.76	23.5	10	89	23.34
1995 (32)	Pages	27.69	26.5	6	61	12.337
	Authors	1.81	2	1	3	0.726
	References	32.69	30.5	7	65	14.837
	Citations	22.31	17.5	10	70	13.047

Table 3: Descriptive statistics of selected *Journal of Econometrics* articles, 1988-1991 with cumulative citations up to and including 2002.

Year (number)	Variable	Mean	Median	Min.	Max.	St.dev.
1988 (18)	Pages	22.59	20	9	43	9.219
	Authors	1.74	2	1	3	0.750
	References	32.41	24	12	108	23.858
	Citations	36.37	22	10	210	42.943
1989 (17)	Pages	18.38	18.5	5	36	7.576
	Authors	1.67	1.5	1	3	0.745
	References	21.42	17	7	46	11.906
	Citations	28.75	21.5	10	116	24.125
1990 (16)	Pages	21.28	19	11	39	6.190
	Authors	1.97	2	1	5	1.067
	References	27.08	25.5	10	62	10.623
	Citations	60.75	31	10	259	59.338
1991 (15)	Pages	23.53	22	6	54	11.312
	Authors	1.71	1.5	1	4	0.859
	References	27.47	24.5	0	77	15.734
	Citations	21.29	16.5	10	44	11.123

Table 4: Descriptive statistics of selected *Journal of Econometrics* articles, 1992-1995 with cumulative citations up to and including 2002.

Year (number)	Variable	Mean	Median	Min.	Max.	St.dev.
1992 (14)	Pages	26.89	24	13	55	9.323
	Authors	1.96	2	1	4	0.999
	References	38.70	27	9	308	53.643
	Citations	70.07	31	11	462[1]	112.878
1993 (13)	Pages	24.82	24	5	50	9.904
	Authors	1.70	2	1	4	0.717
	References	27.18	25	3	55	13.818
	Citations	30.12	25	10	72	17.961
1994 (12)	Pages	25.82	25.5	14	46	7.488
	Authors	1.75	2	1	4	0.871
	References	27.71	31	8	43	10.049
	Citations	38.14	25.5	10	159	37.843
1995 (11)	Pages	27.03	26	15	46	7.252
	Authors	1.79	2	1	4	0.760
	References	35.69	35	13	170	28.323
	Citations	25.62	19	10	85	18.479

[1] This is the review paper on ARCH models by Bollerslev, Chou and Kroner.

Table 5: Estimation results for *Econometrica*, when the trend interacts with all regressors. Standard errors are given in parentheses.

	$\log m$	$\log \frac{2f}{1-2f}$	$\log T^*$	α	$\log \sigma^2$
intercept	2.485	0.102	0.352	0.858	-0.245
	(0.203)	(0.956)	(0.593)	(0.166)	(0.274)
pages[1]	4.759	11.815	8.362	1.253	-1.497
	(0.933)	(2.572)	(1.986)	(0.340)	(1.088)
authors	0.202	-0.259	0.119	0.013	0.085
	(0.059)	(0.907)	(0.513)	(0.076)	(0.124)
references[1]	3.347	-10.225	-2.748	-1.539	-1.066
	(0.475)	(3.811)	(2.066)	(0.379)	(0.797)
trend × pages[1]	-0.014	-1.383	-0.813	-0.213	0.137
	(0.278)	(0.607)	(0.448)	(0.091)	(0.248)
trend × authors	-0.024	0.038	-0.021	-0.003	0.005
	(0.023)	(0.163)	(0.077)	(0.020)	(0.031)
trend × references[1]	-0.419	1.324	0.324	0.282	0.051
	(0.191)	(0.677)	(0.413)	(0.084)	(0.167)
trend	0.069	0.086	0.139	-0.002	-0.009
	(0.061)	(0.150)	(0.090)	(0.041)	(0.068)
diag(Σ_η)	0.400	0.680	0.054	0.110	0.694

[1] Number of pages and number of references are measured in units of 100.

Table 6: Estimation results for *Journal of Econometrics*, standard errors in parentheses

	$\log m$	$\log \frac{2f}{1-2f}$	$\log T^*$	α	$\log \sigma^2$
intercept	3.316 (0.104)	1.474 (0.734)	1.982 (0.326)	1.117 (0.177)	0.022 (0.381)
pages[1]	-1.451 (1.058)	0.894 (5.250)	1.001 (2.239)	0.694 (0.789)	-0.707 (1.252)
authors	0.326 (0.098)	-0.225 (0.593)	-0.080 (0.179)	-0.286 (0.077)	0.100 (0.129)
references[1]	0.006 (0.144)	-2.279 (1.718)	-1.677 (1.018)	0.034 (0.273)	-1.500 (0.599)
trend × pages[1]	1.356 (0.399)	-0.250 (1.105)	0.538 (0.466)	-0.560 (0.260)	0.092 (0.333)
trend × authors	0.048 (0.032)	0.083 (0.126)	0.091 (0.047)	0.061 (0.034)	-0.038 (0.035)
trend × references[1]	0.104 (0.457)	0.205 (0.269)	0.387 (0.413)	-0.034 (0.061)	0.328 (0.120)
trend	-0.316 (0.029)	-0.047 (0.224)	-0.338 (0.076)	0.072 (0.062)	-0.053 (0.097)
diag(Σ_η)	0.389	0.861	0.056	0.084	0.494

[1] Number of pages and number of references are measured in units of 100.

Table 7: Descriptive statistics of typical articles, based on the estimation results in Table 5 and 6.

Pages	Authors	Refs.	Params.	Econometrica		Journal of Econometrics	
				1988	1995	1988	1995
20	2	20	m	108.08	67.64	48.52	79.61
			T^*	5.64	5.61	5.49	6.71
20	3	20	m	129.02	68.04	67.22	153.82
			T^*	6.22	5.34	5.07	11.69
20	2	30	m	144.83	67.58	48.55	85.70
			T^*	4.42	5.53	4.64	7.44

References

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

Blattberg, R.C. and E.I. George (1991), Shrinkage estimation of price and promotional elasticities - Seemingly unrelated equations, *Journal of the American Statistical Association*, 86, 304-315.

Boswijk, H.P. and P.H. Franses (2002), On the econometrics of the Bass diffusion model, Report ERS-2002-66-MKT of the Erasmus Research Institute of Management.

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

Ellison, G. (2002), The slowdown of the economics publishing process, *Journal of Political Economy*, 110, 947-993

Franses, P.H. (2003), On the diffusion of scientific publications. The case of Econometrica 1987, *Scientometrics*, 56, 29-42

Geweke, J. (1989), Bayesian inference in econometric models using Monte Carlo integration, *Econometrica*, 57, 1317-1339

Gourieroux, C. and A. Montfort (1996), *Simulation-based Econometric Methods*, Oxford University Press, Oxford

Kloek, T. and H.K. van Dijk (1978), Bayesian estimates of equation system parameters: An application of integration by Monte-Carlo, *Econometrica*, 44, 345-351

Lenk, P.J. and A.G. Rao (1990), New models from old: Forecasting product adoption by hierarchical Bayes procedures, *Marketing Science*, 9, 42-53.

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.

Meade N. and T. Islam (1998), Technological forecasting - Model selection, model stability, and combining models, *Management Science*, 44, 1115-1130.

McFadden, D. L. and K. Train (2000), Mixed MNL models for discrete response, *Journal of Applied Econometrics*, 15, 447-470

Newey, W. and D. L. McFadden (1994), Large sample estimation and hypothesis testing, in *Handbook of Econometrics*, R. F. Engle and D. L. McFadden (eds.), Amsterdam-North-Holland, 36, 2111-2245

Talukdar, D., K. Sudhir and A. Ainslie (2002), Investing new product diffusion across products and countries, *Marketing Science*, 21, 97-114.

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.