Cointegration in a historical perspective

H. Peter Boswijk
University of Amsterdam

Philip Hans Franses, and Dick van Dijk
Erasmus School of Economics

Econometric Institute Report 2009-08

Prepared for the special issue of the *Journal of Econometrics*
Commemorating 20 Years of Cointegration

This version: April 7, 2009

Authors' notes:

We thank Malcolm Gladwell for encouragement.

Address for correspondence is Econometric Institute, Erasmus School of Economics, PO Box 1738, NL-3000 DR Rotterdam, franses@few.eur.nl

Abstract

We analyse the impact of the Engle and Granger (1987) article by its citations over time, and find evidence of a second life starting in the new millennium. Next, we propose a possible explanation of the success of this citation classic. We argue that the conditions for its success were just right at the time of its appearance, because of the growing emphasis on time-series properties in econometric modelling, the empirical importance of stochastic trends, the availability of sufficiently long macro-economic time series, and the availability of personal computers and econometric software to carry out the new techniques.

Key words: Cointegration, citations

1. Introduction

The *Journal of Econometrics* and *Econometrica* are the two journals that contain the most cited papers in the econometrics discipline. These citation classics have in common that they mainly concern econometric techniques for the analysis of time series variables. By far the best cited time-series econometrics paper is Engle and Granger (1987). The Nobel-worthy concept of cointegration had already been introduced in Granger (1981), but the *Econometrica* paper in 1987 meant an explosive take-off for this novel idea. Many academics and practitioners resorted to the use of the cointegration technique, and theoretical developments in the area covered quite some space in econometrics conferences all over the world. A glance at the programs of the Econometric Society meetings in the 80s and 90s of the previous century, which can be found in back issues of *Econometrica*, shows that a large number of sessions were dedicated to just "Cointegration". Even today there still are workshops and sessions during conferences where new developments in cointegration are being discussed.

It is of course intriguing to ask the question why the concept of cointegration became that important and even deserved Nobel Prize recognition. A substantial part of its success undoubtedly is attributable to the elegance of the concept and the fact that it combines various streams of literature into one single framework. Another part of the success could be due to favourable circumstances at the time cointegration was discovered and put forward. In the present paper we indeed argue that cointegration could become such an important research and application area partly also because it appeared just at the right time. Our argument draws upon the discussion in Gladwell (2008), where the success factors of Microsoft and the Beatles are studied. Here we will argue that part of the success of cointegration can be found in the combination of four external factors that were prominent *when the concept first appeared*. First, in the early 1980s large macro-economic models were losing from simple time series models in terms of forecasting, although people felt that such ARIMA models were lacking economic substance. Second, due to the "discovery" of stochastic trends in macroeconomic time series a few years earlier, there was a sense of urgency for new statistical tools to analyze such data in a correct way. In a sense, when cointegration entered the stage, theoretical and applied econometricians were ready for it. Third, large enough samples of macroeconomic data were becoming available so that it started to become a meaningful exercise to explore the presence of long-run equilibrium relationships. Fourth, but certainly not least important, the computing facilities and software needed to do the calculations involved in cointegration analysis became available to a wider audience at that time, so that the methods could be widely applied.

The outline of our paper is as follows. In the next section we give a few facts and figures on the publication itself. In Section 3 we will compose our argument. In Section 4, we make an attempt to forecast when the next breakthrough, like cointegration, will happen, and what it will look like.

2. Some facts and figures

This section presents a few facts and figures to indicate how important and influential the paper of Engle and Granger (1987) has been and still is.

Insert Table 1 about here

Table 1 presents 10 most cited papers (as documented in December 2008) that have appeared in *Econometrica*. The paper of White (1980) is a clear winner, but the second best-cited paper is Engle and Granger (1987).

Insert Figure 1 about here

Modelling citations

Figure 1 depicts the annual citations to the Engle and Granger (1987) study for the years 1988 to 2006. Interestingly, the pattern mimics that of sales of new products, where usually hump-shaped patterns are found for sales and S-shaped patterns for cumulative sales. Therefore, Franses (2003) and Fok and Franses (2007) propose to model citations data with the so-called Bass (1969) model, which is frequently used for sales data on new (durable) products. This model reads as

$$\frac{C_t}{m - CC_{t-1}} = p + \frac{q}{m} CC_{t-1} \qquad (1)$$

where C_t is citations in year t and CC_{t-1} is cumulative citations up to and including year t-1. The parameter p measures the degree of innovation, the parameter q measures imitation, and *m* measures the maturity (or "saturation") level. There are various ways to create estimable expressions of (1), which mainly depend on the location and source of randomness of the process, see Boswijk and Franses (2005) for various suggestions. Simply rewriting (1) and adding an error term ε_t gives

$$C_t = pm + (q - p)CC_t - \frac{q}{m}CC_{t-1}^2 + \varepsilon_t \tag{2}$$

Estimating the parameters *m*, *p* and *q* with non-linear least squares for sample periods 1988-1995, 1988-1996, and so on, we obtain the estimates as given in Table 2.

Insert Table 2 about here

A closer look at the parameter estimates suggests that until the year 2000, they seem to converge to about 0.032 (*p*), 0.287 (*q*) and 2800 (*m*), but after the millennium the estimates for *q* and *m* start to change again. Perhaps there is a structural break in the parameters around 2000. One way to meet this is to allow a one-time change in the maturity level by replacing *m* with m_t given by

$$m_t = m_1 + m_2 I_t[t \geq 2000] \tag{3}$$

where $I_t[.]$ is an indicator function which is equal to one when its argument is true and zero otherwise. Incorporating (3) into (2) gives estimates (based on the sample 1988-2008) of p= 0.032, q=0.251 and, most interestingly, m_1=3006 and m_2=1488, giving a total estimated maturity level of citations to Engle and Granger (1987) of 4494. These estimation results thus show that around 2000 the citations process was on its way to level off, but then a new boost in citations came about, which led the classical paper even to have a second life. A closer look at the econometrics literature shows that around that time many studies emerged on concepts as fractional cointegration, non-linear cointegration, and panel cointegration, and these may have caused this second life.

Citing papers that are classics too

Not only does the Engle and Granger (1987) paper attract an impressive number of citations, also various papers in the area of unit roots and cointegration have become citation classics themselves.

Insert Table 3 about here

Table 3 documents a selection from the 20 best-cited papers (measurement done in December 2008) in the *Journal of Econometrics*. Clearly, several of these build on the work of Engle and Granger (1987). Clearly, some citation classics on cointegration also appeared in other journals, notably Stock and Watson (1988).

3. Why did it fly?

The text that accompanies the announcement of the Nobel Prize in 2003[1] clearly outlines what the concept of cointegration is, how one can estimate and interpret the parameters in the error correction model, how forecasts can be improved when cointegration is imposed, how important it is for empirical data which oftentimes have unit roots, and how cointegration unifies literatures on economic theory (equilibrium across variables), on time series (data have stochastic trends, yet they share common properties) and on econometrics (deleting the error correction term means mis-specification). But, what the text does not say is why the concept of cointegration was *that* successful. Not only has the paper been cited many times, it also paved the way for other papers that became citation classics, and in fact, cointegration dominated the econometrics research agenda for at least two decades.

We argue that part of the success of cointegration can be attributed to the simple fact that it appeared at the very right moment. All circumstances were perfect. Let us discuss a few of these.

[1] See http://nobelprize.org/nobel_prizes/economics/laureates/2003/ecoadv.pdf

Circumstances

First, with the introduction of the influential book of Box and Jenkins (1970) there emerged an increased interest in analyzing time series data and using rather simple models for out-of-sample forecasting. Indeed, the proposed ARIMA models turned out to deliver high quality forecasts, and in fact, in those days these time series forecasts were observed to be much better than those from large-scale macro-econometric models containing hundreds of equations and equally sized numbers of unknown parameters. Even though it can be argued that large-scale simultaneous equations models can be written as VAR models, which in turn can be written as ARIMA type models (Sims, 1980 and Zellner and Palm, 1974), the sheer infinite distance between large models and ARIMA forecast schemes created a need for "models in between". Possible candidates for this were the single-equation error correction models such as the well-known Davidson *et al.* (1978) consumption function. In fact, as discussed by Granger (2009), it was exactly the confrontation of such models with unit-root processes (as discussed next) that led to the notion of cointegration.

Second, initiated by Dickey and Fuller (1979, 1981)'s innovative work on testing for unit roots in time series data and the application of their tools to US macroeconomic data in Nelson and Plosser (1982), there seemed to be an acute problem with analyzing such data. Before then, all economic time series data were supposed to be governed by deterministic trends, while suddenly the word was that they all had stochastic trends, aka unit roots. If that were true, then all previously constructed models were created using the wrong statistical tools, as Phillips (1986) showed that statistical theory for regressions with unit-root time series is markedly different than standard theory. In short, the feeling was "we did it all wrong"! Yet, at the same time, the urgent question was: "How should it be done then?"

The third favourable circumstance for the fly of cointegration was the availability of useful samples of data. By 1980 many countries had collected reliable quarterly macroeconomic data since the end of WWII, meaning that around thirty years of quarterly data, that is 120 observations, were available for a range of western countries (and even for Japan, see Hylleberg *et al*. 1990).

A fourth circumstance, and this is very well described in Gladwell (2008), is that the beginning of the 1980s also marked the entry of the Personal Computer (PC). Two of the three authors of the current article vividly remember seeing a PC for the first time in those days, where at the time they and all students worked at terminals that were linked to housesized mainframe computers. Suddenly, everyone could buy a PC, have it at home and at work, and use it for computations and word processing.

Insert Table 4 about here

A fifth and final circumstance, which is very much related to the previous one, is that the econometrics discipline witnessed an explosion in statistical packages and matrix programming languages that were developed for the PC and made available for free or at a reasonable price. Table 4 summarizes a few of these packages, some of which are still with us today, and clearly, they were all available at the very same time.

And then, it happened!

In the midst of the rapid developments of PCs and econometric software, the increasing availability of the relevant data, and the enormous sense of urgency felt to properly put stochastic trend data into a, not too large, multivariate model, there suddenly it was! Cointegration implied small-scale models, incorporating stochastic trend data, useful for forecast quarterly macroeconomic data, using the proper statistical tools, and…, everybody could do it! The regression-based inference was simple to carry out, simulated critical values became available, and all analysis could simply be done on a PC, at home or in the office.

Of course, matters were not immediately that simple. Data could be unreliable and still a bit too short. The discriminating power of the tests was at best rather low, and sometimes smaller than the size. Cointegration was not robust to breaks in the data or to outliers. Data could be non-linear, that is, experience asymmetry over the business cycle, and the like. All this just meant that the basic Engle and Granger (1987) proposition could be extended in an almost infinite number of ways. Workshops, conferences and special issues of the leading journals were all addressing these developments. It marked the start of careers of various academics, who are still sometimes working on these topics, even today.

Naturally, new scientific developments are also often associated with particular research environments. Important academics in the cointegration area created working conditions that attracted young academics and students, who all gathered in workshops and conferences. In those days, the places to be were San Diego, Aarhus, Oxford and Copenhagen, the respective domiciles of Robert Engle and Clive Granger, Svend Hylleberg, David Hendry and Søren Johansen. It was a spectacular period, and it really gave a boost to the econometrics discipline.

4. What will happen next?

Now we have seen that cointegration could fly not only due to its particular relevance to the econometrics discipline but also due to five favourable circumstances, we are tempted to put it into an even more historical perspective and to make a prediction of what might happen next.

Did we see it before?

To facilitate making such a prediction, it is perhaps good to look back in time and see if there have been more such revolutionary developments in the econometrics discipline. To us, it seems there have been two such developments, and interestingly enough, they share part of the favourable circumstances that were relevant for cointegration.

By the late 50s and 60s of the 20th century there were developments concerning the simultaneous equations model which mimic those of cointegration. With the advent of a few annually observed datasets (mainly covering the US economy) and with the advent of the first mainframe computers, it became possible to create multiple-equation models that could be used for forecasting and policy analysis. There also was a sense of urgency as, after the end of WWII, many countries needed tools to properly analyze economic growth and other macroeconomic figures. The key problem though was that the models had to be fitted to annually observed data, and with an annual sampling frequency many changes in macroeconomic data seemed to happen at the same time. Hence, the running model for most econometricians was the simultaneous equations model, and this involved problems for estimation. With the discovery of two-stage least squares (by Henri Theil in Rotterdam) and all its variants, suddenly these problems could be solved and the models could be used in practice. In those days, the places to be were New Haven (Yale), Chicago and the Econometric Institute in Rotterdam.

The second relevant development started more or less at the same time as cointegration by the beginning of the 1980s, and yet again in San Diego. This was the creation of the ARCH model (Engle, 1982). With the advent of detailed financial data, the urgency to measure and estimate risk and volatility of financial assets, the inclusion of ARCH estimation routines in MicroTSP and EZARCH, this also Nobel-worthy invention could fly too.

What will the future bring us?

In sum, new developments in econometrics seem to take off in times when there is a sense of urgency, when circumstances are perfect, and when everybody suddenly can use the new models or tools themselves.

So, the next revolution in econometrics could again be based on serious improvements in three dimensions, that is, better data, a sense of urgency and more computing power (so that everybody can do it). Better data could mean that we all have immediate access to the relevant data at a high frequency. More computing power could mean that it becomes available in personal calculators with the size of a mobile phone, which would make models to run automatically, parameters to be estimated in a split second, and model choice to be automated. People can then interact individually with model outcomes, adjust forecasts, and in a next round this expertise is incorporated in new model forecasts. The urgency could be that forecasts need to be made very often, for example in financial risk management based on high-frequency data, and by then, in the future, it has been widely recognized that models cannot do it all, and that model outcomes need an expert touch.

Table 1: Citations to *Econometrica* papers (December 2008 score)

Paper	Citations
White (1980)	4829
Engle and Granger (1987)	3816
Heckman (1979)	3498
Engle (1982)	2583
Hausman (1978)	2236
Newey and West (1987)	1976
Hansen (1982)	1886
Dickey and Fuller (1981)[2]	1731
Sims (1980)	1395
Johansen (1991)	1311

[2] The same authors (1979) published on the same topic a paper two years earlier in JASA, and this paper has now achieved 2518 citations.

Table 2: Parameter estimates of Bass (1969) model when applied to annual citations data for Engle and Granger (1987)

Sample	*p*	*q*	*m*
1988-1995	0.039	0.427	1805
1988-1996	0.035	0.357	2207
1988-1997	0.033	0.319	2495
1988-1998	0.032	0.294	2708
1988-1999	0.032	0.288	2755
1988-2000	0.032	0.287	2763
1988-2001	0.032	0.262	2938
1988-2002	0.032	0.230	3190
1988-2003	0.032	0.212	3348
1988-2004	0.031	0.189	3585
1988-2005	0.031	0.180	3683
1988-2006	0.031	0.158	3957

Table 3: Citations to *Journal of Econometrics* papers (published after 1987)

(December 2008 score)

Paper	Citations
Kwiatkoski, Phillips, Schmidt and Shin (1992)	974
Hylleberg, Engle, Granger and Yoo (1990)	394
Johansen and Juselius (1992)	331
Gonzalo (1994)	236
Levin, Lin and Chu (2002)	234
Pesaran and Smith (1995)	218

Table 4: Software development (still available)

Period	Econometric and Statistical software
1975-1979	Shazam
	Limdep
	RATS
	Autoreg (became Give and later PcGive)
1980-1985	MicroTSP (became Eviews around 1990)
	Gauss
	Stata
	Matlab

Source: Figure 1 in Ooms and Doornik (2006, page 213)

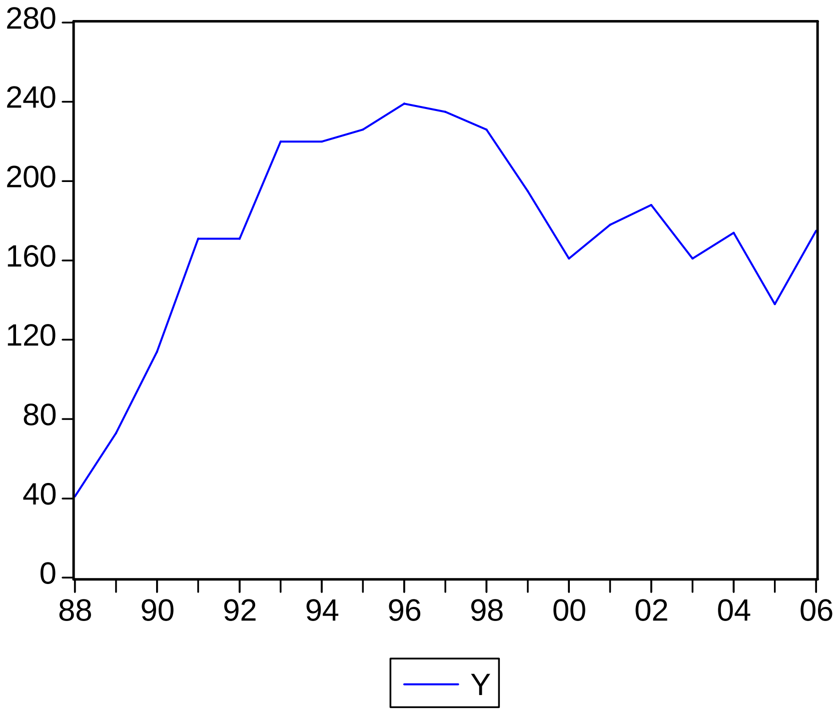

Figure 1: Annual citations to Engle and Granger (1987): 1988-2006

References

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

Boswijk, H.P. and P.H. Franses (2005), On the econometrics of the Bass diffusion model, *Journal of Business & Economic Statistics* 23, 255-268.

Box, G.E.P. and G.M. Jenkins (1970), *Time Series Analysis: Forecasting and Control*, San Francisco: Holden-Day.

Davidson, J.E.H., D.F. Hendry, F. Srba, and J.S. Yeo (1978), Econometric modelling of the aggregate time-series relationship between consumers' expenditure and income in the United Kingdom, *Economic Journal* 88, 661-692.

Dickey, D.A. and W.A. Fuller (1979), Distribution of the estimators for autoregressive time series with a unit root, *Journal of the American Statistical Association* 74, 427–431.

Dickey, D.A and W.A. Fuller (1981), Likelihood ratio statistics for autoregressive time series with a unit root, *Econometrica* 49, 1057-1072.

Engle, R.F. (1982), Autoregressive conditional heteroskedasticity with estimates of the variance of U.K. Inflation, *Econometrica* 50, 987-1008.

Engle, R.F. and C.W.J. Granger (1987), Co-integration and error-correction: representation, estimation, and testing, *Econometrica* 55, 251-76.

Engle, R.F. and B.S. Yoo (1987), Forecasting and testing in co-integrated systems, *Journal of Econometrics* 35, 143-159.

Fok, D. and P.H. Franses (2007), Modeling the diffusion of scientific publications, *Journal of Econometrics* 139, 376-390.

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

Gladwell, M. (2008), *Outliers: The Story of Success*, New York: Little, Brown and Company.

Gonzalo, J. (1994), Five alternative methods of estimating long-run equilibrium relationships, *Journal of Econometrics* 60, 203-233.

Granger, C.W.J. (2009), Some thoughts on the development of cointegration, *Journal of Econometrics*, this issue.

Hansen, L.P. (1982), Large sample properties of generalized method of moments estimators, *Econometrica* 50, 1029-1054.

Hausman, J.A. (1978), Specification tests in econometrics, *Econometrica* 46, 1273-1291.

Heckman, J.J. (1979), Sample selection bias as a specification error, *Econometrica* 47, 153-161.

Hylleberg, S., R.F. Engle, C.W.J. Granger and B.S. Yoo (1990), Seasonal integration and cointegration, *Journal of Econometrics* 44, 215–238.

Johansen, S. (1991), Estimation and hypothesis testing of cointegration vectors in Gaussian vector autoregressive models, *Econometrica* 59, 1551-1580.

Kwiatkowski, D., P.C.B. Phillips, P. Schmidt and Y. Shin (1992), Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root?, *Journal of Econometrics* 54, pp. 159–178.

Levin, A., C. F. Lin, and C. Chu (2002). Unit root tests in panel data: Asymptotic and finite-sample properties, *Journal of Econometrics* 108, 1–24.

Nelson, C.R. and C.I. Plosser (1982), Trends and random walks in macroeconomic time series. *Journal of Monetary Economics* 10, 139-162.

Newey, W.K. and K.D. West (1987), A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica* 55, 703–708.

Ooms, M. and J.A. Doornik (2006), Econometric software development: Past, present and future, *Statistica Neerlandica* 60, 206-224.

Pesaran, M.H. and R. Smith (1995), Estimating long-run relationships from dynamic heterogeneous panels, *Journal of Econometrics* 68, 79-113.

Phillips, P.C.B. (1986), Understanding spurious regressions in econometrics, *Journal of Econometrics* 33, 311-340.

Sims, C.A. (1980), Macroeconomics and reality, *Econometrica*, 48, 1-48.

Stock, J. and M. Watson (1988), Testing for common trends, *Journal of the American Statistical Association* 83, 1097–1107.

White, H. (1980), A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity, *Econometrica* 48, 817-838.

Zellner, A. and F. Palm (1974), Time series analysis and simultaneous equations models, *Journal of Econometrics* 2, 17-54.