Splines can recover dynamic information contained in discrete data

Yuxian Liu[1] and Ronald Rousseau[2]

[1] *yxliu@tongji.edu.cn*
Tongji University, Library of Tongji University, Siping Street 1239, 200092 Shanghai (PR China)
University of Antwerp, IBW, Venusstraat 35, 2000 Antwerpen (Belgium)

[2] *ronald.rousseau@khbo.be*
University of Antwerp, IBW, Venusstraat 35, 2000 Antwerpen (Belgium)
KHBO (Association K.U.Leuven), Industrial Sciences and Technology, 8400 Oostende (Belgium)

Introduction

Cumulative citation data, like all citation data, are discrete. In order to be able to apply the methods of differential calculus we propose using cubic spline interpolation (Hildebrand, 1956), leading to a continuous approximating curve reflecting the number of cumulative citations with respect to time. We like to consider the cumulative citation curve as a curve representing the corresponding article's position in academic space. Since we use cubic splines we can calculate first and second order derivatives. The first order derivative is referred to as instantaneous velocity and the second order derivative as instantaneous acceleration. We now provide a proof that by using splines it is possible to recover information that is hidden in the original data.

Theoretical relation between academic position, instantaneous velocity and instantaneous acceleration

We denote by $cit(t)$ the total number of citations a paper (or group of papers) receives t years after publication. Spline functions are used to describe the relation between the cumulated citation frequency and time. As an example, we show the first three years' accumulated citation data of Kao's Nobel Prize winning article (Kao & Hockham, 1966), but in our investigations we actually used the complete citation history. For the first three years we have formula (1)

$$cit(t) = \begin{cases} 0.78t^3 - 0.78t + 1 & t \in [0,\ 1] \\ -1.89t^3 + 8.02t^2 - 8.80t + 3.67 & t \in [1,\ 2] \\ 3.80t^3 - 26.12t^2 + 59.47t - 41.84 & t \in [2,\ 3] \end{cases} \quad (1)$$

We denote by $v_c(t)$ the velocity at a specific time t, where the suffix c refers to citation diffusion. According to the definition of velocity we have:

$$v_c(t) = \frac{dcit(t)}{d(t)} \quad (2)$$

Similarly acceleration, denoted as $a_c(t)$, can be obtained as the second derivative of $cit(t)$.

Again, using Kao's data as an example, its velocity and acceleration are shown by formulae (3) and (4):

$$v_c(t) = \begin{cases} 2.34t^2 - 0.78 & t \in [0,\ 1] \\ -5.68t^2 + 16.03t + 8.80 & t \in [1,\ 2] \\ 11.39t^2 - 52.23t + 59.47 & t \in [2,\ 3] \end{cases} \quad (3)$$

$$a_c(t) = \begin{cases} 4.67226t & t \in [0,\ 1] \\ -11.3613t + 16.0335 & t \in [1,\ 2] \\ 22.7729t - 52.2348 & t \in [2,\ 3] \end{cases} \quad (4)$$

Instantaneous velocity is the derivative of cumulative citation with respect to time. Hence, knowing the instantaneous velocity with respect to time, we can obtain the cumulative citation by integration. This is done as follows (see equations (5) and (6)):

$$Cit_t = \int_0^t v_c^{\,t}\,dt = v_c^{\,0} + \int_0^1 v_c^{\,1}\,dt + \ldots + \int_{t-1}^t v_c^{\,t}\,dt \quad (5)$$

$$v_c^{\,t} = \int_0^t a_c^{\,t}\,dt = a_c^{\,0} + \int_0^1 a_c^{\,1}\,dt + \ldots + \int_{t-1}^t a_c^{\,t}\,dt \quad (6)$$

We checked and indeed instantaneous velocity and cumulative citation of Kao's

data satisfy equation (5), and instantaneous acceleration and the velocity satisfy equation (6). On the one hand, this proves that our methods are mathematically rigorous, and on the other hand, that the goodness-of-fit of our spline approximation is good.

Graphical representations of the spline curves

We use cumulative citation data to the most important articles written by five Chinese-American Nobel prize winners in physics to draw the spline curves (Figure 1). Obvious the splines are continuous in the data points and are even differentiable in these points. These curves represent the number of cumulative citations with respect to time.

Relation between real changes and the changes deduced from spline-fitted curves

Does instantaneous velocity really reflect realistic rates of change or are they just artefacts of the used method? We know that using one year as a time unit, the velocity in a specific year is essentially the

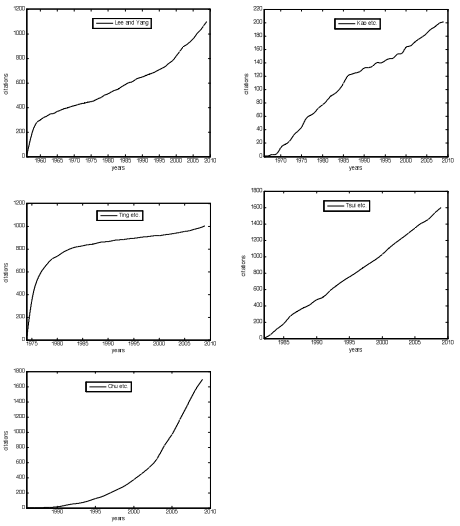

Figure 1. Spline-fitted curves of cumulative citations of the main articles of five Chinese-American Nobel Prize winners in physics

number of citations obtained in that year. So we can test if the velocity we calculate by our method is really consistent with the number of citations that the article receives in that specific year.

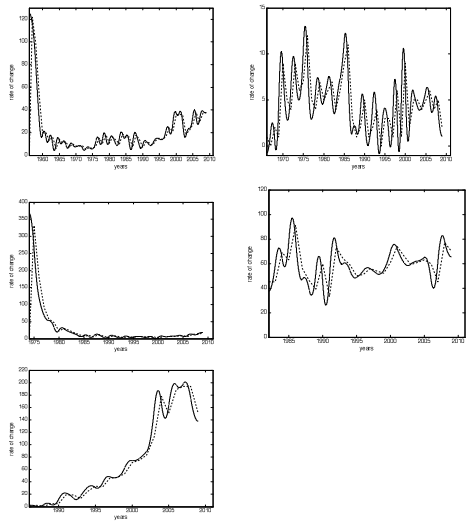

Figure 2. Yearly number of citations and velocity to main articles of five Chinese-American Nobel Prize winners in physics

Figure 2 shows the original yearly number of citations to the most important article written by five Chinese-American Nobel winners in Physics (connected by dashed line segments) and instantaneous velocity curves (full lines) derived from our spline-fitted curves. Clearly differences are very small. For more details, including the original citation data, we refer to (Liu, 2011).

The meaning of the rate of change

Describing the development of science is one aspect of informetrics. How citations change over time is a key factor to describe this development. (Buter et al., 2011) defined growth rates and try to use this concept to describe how different disciplines converge over time. However, they use discrete data in such a way that the growth rate can only be obtained in specific points. In the context of epidemic models colleagues do use continuous rates of change. Such epidemic models are based on several assumptions, one being that all individuals in the population are considered to have an equal probability of contracting the disease. In our view, this assumed specific probability yields an unrealistic model. The high consistency

between instantaneous velocity and yearly number of citations yields the promise to be able to describe the actual growth process. We claim that this approach can be used to investigate how science develops over time.

Conclusion

We have shown why the use of splines is an interesting method to describe discrete, cumulative data, even if citation data can be connected directly by a polygonal line. Such a polygonal line contains no change in information because its first derivative is a constant and its second derivative is zero (in the points where these derivatives exist). A spline-fitted curve however can represent changing information. All information about change resides in the original data. Splines not only reflect this change, but keep this information. By integration and the use of differential calculus we can revive and reproduce the original information. In this sense, we like to say that splines behave like a hologram: both are able to recover the original information.

Acknowledgments

The work of R.R. is supported by an international grant from the National Natural Science Foundation of China (NSFC Grant No 7101017006).

References

Buter, R.K., Noyons, E.C.M. & van Raan, A.F.J. (2011). Searching for converging research using field to field citations. *Scientometrics,* 86(2), 325-338.

Hildebrand, F.B. (1956). Introduction to numerical analysis. New York: McGraw-Hill.

Kao, K.C & Hockham, G.A. (1966). Dielectric-fibre surface waveguides for optical frequencies. *Proceedings of the Institution of Electrical Engineers – London*, 113(7), 1151-1158.

Liu, YX. (2011). The diffusion of scientific ideas in time and indicators for the description of this process. Doctoral thesis. Antwerp University.

Liu YX. & Rousseau, R. (2011). Visualizing discrete data by spline functions. Submitted.