Predicting scientific trajectories: supplemental material

Daniel E. Acuña, Stefano Allesina, Konrad P. Körding

Introduction

In this work, we rely on automatization for efficiently collecting information from the Internet. This is a fast and cheap method for constructing large bibliometrics datasets that are much less error prone when compared to manual entering.

From Academic Tree (http://www.academictree.org), we obtained names, institutions, studies, advisors, and collaborators of 38,293 profiles spanning Neurotree, (see Table 1). For each Academic Tree researcher, we downloaded the entire Scopus profile (http://www.scopus.com). For each article in Scopus, we obtained title, journal, co-authors, year of publication, and citing articles. For each citing article, we obtained title, journal, co-authors, and year of publication. A summary of the data obtained from Scopus is shown in Table 2.

We downloaded funding information from the NIH ExPORTER Data Catalog (http://exporter.nih.gov/ExPORTER_Catalog.aspx) for the years 1995 to 2010, and we matched PI names to Academic Tree profiles. We imputed missing yearly costs (funding) in years 1995 to 1999 from the average yearly funding of R01 grants in other years (US$ .24 million). A summary of the grant information is shown in Table 3.

The complete set of features is described in Table 4.

Table 1 Summary of data available for each tree

Subset	Number of profiles
Neurotree	34,873
Fly tree	2,032
Evolution	1,388

Table 2 Summary of data from Scopus

Scopus	Quantity
Articles	2,050,686
Citations	5,903,507
Authors	2,729,915

Table 3 Some summary of Neurotree, Scopus, and NIH grant funding

Data property	Number
NIH/neurotree matches	3,866
Grant cycles matched	6,876

Table 4 Features

Feature number	Definition
1	Number of articles in top journals, where top journals are defined as Nature, Science, Nature Neuroscience, PNAS, and Neuron.
2	Number of articles in 'theoretical' journals (Neural computation, network: computation in neural systems, Frontiers in computational neuroscience, journal of computational neuroscience, journal of neural engineering, neuroinformatics, PLOS computational biology)
3	Number of years in a postdoctoral position.

Feature number	Definition
4	Number of years in graduate school
5	Number of articles with a adviser is a co-author
6	Number of R-type grants
7	Total current yearly cost
8	Current *h*-index divided by career length (also known as *m*-index)
9	Mean *h*-index divided by career length of PhD supervisor
10	Proportion of articles as last author
11	Proportion of articles as first author
12	Total number of citations
13	Total number of articles
14	Average number of coauthors per article
15	Total number of journals
16	Average number of citations per article
17	Career length defined as number of year since publishing first article
18	Current *h*-index

Elastic net regularization for linear regression

We use linear regression with elastic net regularization[1] to study effect sizes and improve generalization. In particular, we use the extremely-efficient `glmnet` package [2] within the

statistical software system R [3].

Ordinary least square linear regression seeks to construct a relationship between a set of dependent variables $x_1,\ldots,x_m$ and an independent variable y through the estimation of coefficients $\beta_0,\beta_1,\ldots,\beta_m$. Additionally, since we are analyzing a time series, we can think of a linear regression as

$$y_t = \beta_0 + \sum_{i=1}^{m} \beta_i x_{it} + e_t \ ,$$

where e_t is a noise term normally distributed with zero mean and a time-independent standard deviation σ. In fact, this regression is time invariant and therefore the index t can be informally dropped. By creating a vector with an additional "bias" term $\mathbf{x} = \begin{bmatrix} 1 & x_1 & \cdots & x_m \end{bmatrix}^T$ and $\beta = \begin{bmatrix} \beta_0 & \cdots & \beta_m \end{bmatrix}^T$, we can represent the regression as

$$y = \beta^T \mathbf{x} + e \ .$$

However, when the number of features is large, features are possibly collinear, and generalization of results needs to be improved, it is possible to adaptively reduce the complexity of the model by forcing some coefficients to be small or zero. This is what elastic net regularization tries to achieve by minimizing the following objective function

$$\min_{\beta} \left(\frac{1}{2n} \sum_{i=1}^{n} \left(y_i - \beta^T \mathbf{x}_i \right)^2 + \lambda \left(\frac{(1-\alpha)}{2} ||\beta||_2^2 + \alpha ||\beta||_1 \right) \right), \qquad \lambda \geq 0, 0 < \alpha \leq 1$$

where m is the number of data points, λ controls the complexity of the model — higher values imply a simpler model — and α controls how much we expect features to be collinear (α close to 0) or irrelevant (α close to 1). In all our analyses, we set α to .2 and find the best λ by cross-validation.

Because elastic net regularized regression does not conform easily to the normality assumptions of ordinary least square regression, we cannot rely on the central limit theorem to obtain 95% confidence intervals. We instead obtain confidence intervals using 10,000 bootstrap samples. Finally, to produce the simplified regression of Box 1 in the main text, we computed the best model by backward stepwise selection from the full model [4]. We stop the back search

when five variables were left. The five-variable model had a significantly higher R^2 than the four, three, and two variable models.

References

1. Zou, H. & Hastie, T. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **67,** 301-320 (2005).

2. Friedman, J. H., Hastie, T. & Tibshirani, R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software* **33,** 1-22 (2010).

3. R Development Core Team *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna, Austria, 2011).

4. Hastie, T. *The elements of statistical learning* (Springer, New York, NY, 2009).