Embracing chaos: using elastic net regression to analyse unbalanced datasets with thousands of predictors

Researchers tend to want to create order with their studies: unequivocal scientific conclusions, hinging on a clear-cut structure. This desire for order stands in stark contrast with reality, which is rather chaotic. The world's disorderliness is reflected in the datasets we find in the humanities, which are often unbalanced and/or messy. While it is possible to ignore (part) of this chaotic reality in one's analysis, ideally, the analysis is compatible with the variability inherent in humanities datasets.

In this paper, we present one way to cope with the richness of humanities' datasets: elastic net regression. Elastic net regression is a regression analysis technique which incorporates so-called 'regularisation' into its model fitting procedure. In contrast to more traditional regression techniques like linear or logistic regression, elastic net regression adds a penalisation factor to the maximum likelihood procedure used to compute the model parameters. The practical benefit of this procedure is that in low-data situations, which are common in humanities datasets, fair and generalisable predictions are guaranteed through 'shrinkage', an attenuation of model coefficients. In this way, we avoid extreme and unreliable model predictions stemming from a lack of data points. At the same time, not all predictors specifically need to play a role in an elastic net regression model. The model fitting procedure can attenuate predictors to the extent that they are reduced to zero, which in practice implies that that predictor is disabled. This means the model can itself decide what predictors are important in an analysis, alleviating the need to devise a specific structure or ranking of importance beforehand.

Elastic net regression is typically combined with k-fold cross-validation, a technique where different parts of the data are rotated in series to train the model. The benefit of this procedure is that it allows one to have a dataset with n observations and p variables, even when $n < p$. This feature is very useful for humanities data, as it allows one to enter a multitude of predictors in a regression model at once, again without having to decide on a particular structure beforehand. The possible applications for the humanities (entering thousands of words, people, places, texts… into a single analysis) are manifold. Additional variables can still be added to the model, which means the analysis can remain under multifactorial control.

To gain insights on the predictors we include in an elastic net regression model, we extract a model's coefficients. These coefficients serve as corrections to the intercept, the baseline value of a model. As such, these corrections can be used to infer the preferences of specific predictors. To illustrate this principle, we look at two previous studies by the authors: one in the field of culturomics (XXXXXXX XXXXa) and one in the field of linguistics (XXXXXXX n.d.).

In the first case study (published as XXXXXXX XXXXa), lasso regression, a subset of elastic net regression, is used to gauge the Dutch collective 'cultural memory'. The goal of the study is to determine which historical figures were characteristic of a specific time period through a query of the Dutch cultural journal *De Gids*. More specifically, it is measured quantitatively how characteristic each person from a predetermined list of historic figures is for the long 19th century (1837–1914) and the short 20th century (1915–1999) respectively. The operationalisation is simple: the 7000+ articles from *De Gids* count as attestations (rows), while the 264 historical figures from the predetermined list serve as predictors (columns). If a historical figure is present in an article, the value for their column is encoded 1 (as opposed to 0). The response variable for each attestation is the macroperiod the article belongs to. To get a better idea of the input data, please refer to Table 1.

Century	**Article**	**Alexander The Great**	**Ambiorix**	**…**	**William Shakespeare**
19th	#1	1	0	…	1
20th	#2	0	0	…	1
19th	#3	0	0	…	1

Table 1: An (abridged) example of the input data for the XXXXXXX (XXXXa) culturomics study.

With an intercept of 0.213 on the logit scale, and positive values showing a skew towards the 20th century, on average, the historical figures in the study pull towards the 20th century slightly.

Not all figures have correction terms in the model: 77 (29%) were eliminated, which means having corrections for those figures did not improve the model's quality. The coefficients of the remaining historical figures can be interpreted as follows: figures with negative model coefficients pull towards the long 19th century, while figures with positive model coefficients pull towards the short 20th century. The larger the absolute value of the coefficient, the stronger the association is. Refer to Figure 1 for an (abridged) visualisation of the coefficients of the lasso model, converted to probabilities. These probabilities take the intercept baseline into account.

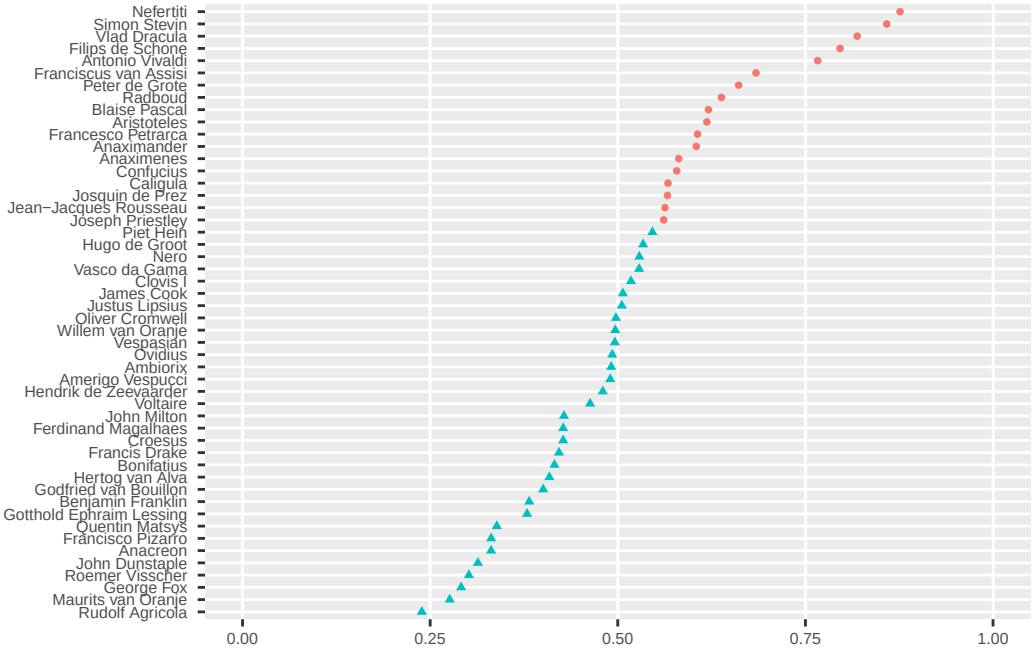

Figure 1: A sample of 50 historical figures from the culturomics study. The logit values have been converted to probabilities. Triangles pull towards the 19th century, circles pull towards the 20th century.

In XXXXXXX (XXXXa), the coefficient pulls from the lasso regression are used in tandem with weighted linear regression analysis to further investigate whether the importance of historical figures from certain domains (scientists, political leaders, religious leaders) changed from time period to time period. This indeed seems to be the case: where in the earlier period, political leaders were more important, later issues of *De Gids* gave scientists more prominence. This shows that elastic net regression can be used to bootstrap a more elaborate, structured culturomics analysis, even despite the initial plethora of historical names present in the corpus.

In the second case study (to be published as XXXXXXX n.d.), preferences of participles in verbal clusters found in subordinate sentences. Consider the following two examples:

1. Ik weet dat hij vader **is geworden**.
   I know that he father has become. (literally)

2. Ik weet dat hij vader **geworden is**.
   I know that he father become has. (literally)

In these verbal clusters, two word orders are possible: auxiliary + participle (example 1, red order), and participle + auxiliary (example 2, green order). There are many factors influencing the choice between red and green: lectal factors, prosody, frequency, priming, but also semantics (De Sutter et al. 2005, Bloem 2021, XXXXXXX XXXXb). The idea is that the choice of the verb in the cluster also influences whether the red or green word order is used. We queried the SoNaR corpus of Dutch (Oostdijk et al. 2008, 2014) for all red or green verbal clusters in subordinate sentences, yielding 236,408 attestations: 150,574 of the red word order, 85,834 of the green word order. Again, elastic net regression is used, this time to investigate the relative preferences of Dutch verbs. Each participle is encoded as a predictor, which is then either removed by the model or retained, with a red or green association respectively. 1,121 verbs (53%) were removed, 998 verbs (47%) were retained. Of the retained verbs, 452 verbs (45%) skew towards the red word order, 546 (55%) skew towards the green word order.

The pulls of the verbs were used in combination with distributional semantics, quantitative encodings of meaning (Montes 2021), to gain an insight to a possible relation between the meaning of a verb and its preference for either the red or green word order. If we visualise the semantic vectors using dimension reduction (in this case MDS: Carroll & Arabie 1998), and then generalise the locations of these verbs using a Generalised Additive Model (Hastie & Tibshirani 1987), we see that there are indeed semantic areas with relative red or green preferences (see Figure 2). While it is difficult as humans to interpret these areas (quantitative semantics is rather opaque), the 'semantic pockets' in the plot as they appear in Figure 2 prove that a relation between meaning and word order preference does exist. Again, elastic net regression was used here to bootstrap a larger analysis on a dataset with very unbalanced frequencies of, in this case, verbs.

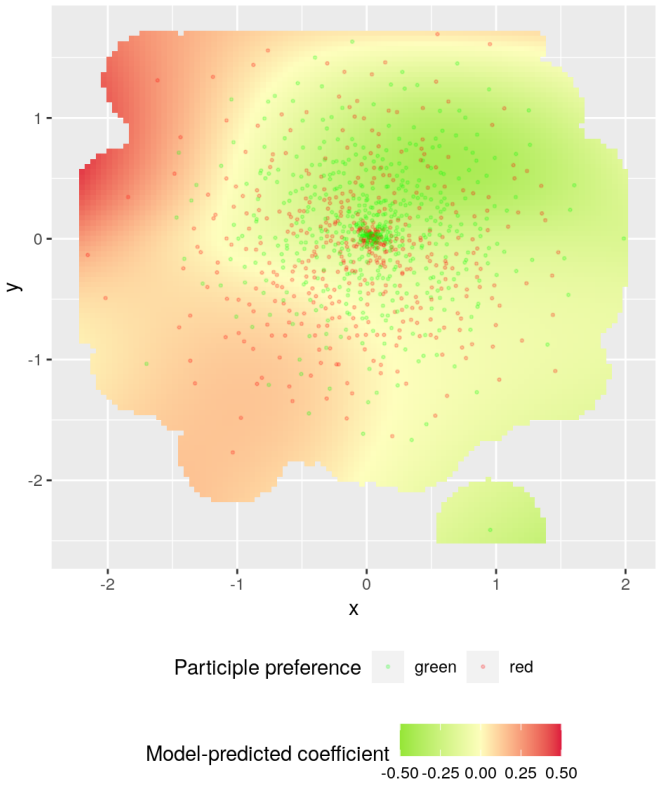

Figure 2: Visualisation of the GAM model predicting the elastic net coefficients based on the semantic space, MDS (non-eliminated types only). Each dot is a verb in the semantic space; its colour indicates its red/green preference.

The two case studies highlighted here show that elastic net regression is an interesting tool within the humanities, as it is able to cope with many of the problems typical of data in these fields. The modelling technique can serve as a diving board for further analysis with other, already established methodologies, to prove trends or relationships, as shown in the two case studies. As such, we see its potential for further use in other fields within the realm of digital humanities.

References

Bloem, Jelke. 2021. *Processing verb clusters*. LOT. https://doi.org/10.48273/LOT0586.

Carroll, J. Douglas & Phipps Arabie. 1998. Multidimensional Scaling. In Michael H. Birnbaum (ed.), *Measurement, Judgment and Decision Making* (Handbook of Perception and Cognition (Second Edition)), 179–250. San Diego: Academic Press. https://doi.org/10.1016/B978-012099975-0.50005-1.

De Sutter, Gert, Dirk Geeraerts & Dirk Speelman. 2005. *Rood, groen, corpus! Een taalgebruiksgebaseerde analyse van woordvolgordevariatie in tweeledige werkwoordelijke eindgroepen* dissertation.

Hastie, Trevor & Robert Tibshirani. 1987. Generalized Additive Models: Some Applications. *Journal of the American Statistical Association* 82(398). 371–386. https://doi.org/10.1080/01621459.1987.10478440.

Montes, Mariana. 2021. *Cloudspotting: visual analytics for distributional semantics* dissertation. https://lirias.kuleuven.be/retrieve/630179 (30 November, 2021).

Oostdijk, Nelleke, Martin Reynaert, Veronique Hoste, Henk van den Heuvel, Orphee de Clercq, Ewoud Sanders & Creative Computing. 2014. *SoNaR nieuw media corpus*. https://research.tilburguniversity.edu/en/publications/ac128452-d97c-4290-8e65-12a1462ba47d (17 May, 2023).

Oostdijk, Nelleke, Martin Reynaert, Paola Monachesi, Gertjan van Noord, Roeland Ordelman, Ineke Schuurman & Vincent Vandeghinste. 2008. From D-Coi to SoNaR: A reference corpus for Dutch. 9.

XXXXXXX, XXXXXXX. XXXXa. Xxxxxxx. *XXXXXXX*.

XXXXXXX, XXXXXXX. N.d. Xxxxxxx. Under review.

XXXXXXX, XXXXXXX. XXXXb. Xxxxxxx. *XXXXXXX* XX(X). XXX–XXX. https://XXXXXXX.com/XXXXXXX/XXXXXXX.