Journal of Phonetics 71 (2018) 249–267

Contents lists available at ScienceDirect

Journal of Phonetics

journal homepage: www.elsevier.com/locate/Phonetics

Special Issue: Emerging Data Analysis in Phonetic Sciences, eds. Roettger, Winter & Baayen

Strategies for addressing collinearity in multivariate linguistic data

Fabian Tomaschek *, Peter Hendrix, R. Harald Baayen

Department of General Linguistics, University of Tübingen, Germany

ARTICLE INFO

Article history:
Received 2 October 2017
Received in revised form 3 September 2018
Accepted 10 September 2018

Keywords:
Elastic net
Supervised component generalized linear regression
Random forests
Collinearity
Concurvity
Segment duration

ABSTRACT

When multiple correlated predictors are considered jointly in regression modeling, estimated coefficients may assume counterintuitive and theoretically uninterpretable values. We survey several statistical methods that implement strategies for the analysis of collinear data: regression with regularization (the elastic net), supervised component generalized linear regression, and random forests. Methods are illustrated for a data set with a wide range of predictors for segment duration in a German speech corpus. Results broadly converge, but each method has its own strengths and weaknesses. Jointly, they provide the analyst with somewhat different but complementary perspectives on the structure of collinear data.

1. Introduction

Response measures in linguistics and phonetics are often a function not of a single predictor but of many predictors jointly, reflecting a move away from mono-causal to multifactorial explanations. For instance, reductions and deletions in speech have been shown to correlate with a range of measures which include frequencies of occurrence and conditional probabilities at word and segment level (among others Aylett & Turk, 2004; Bell, Brenier, Gregory, Girand, & Jurafsky, 2009; Gahl, 2008; Jurafsky, Bell, Gregory, & Raymond, 2000; Priva, 2015; Tremblay & Tucker, 2011). For example, Tremblay and Tucker (2011) used no less than 18 such measures to predict the durations of four-word sequences. Typically, many of the covariates included in these analyses serve as controls for potential confounds with predictors of central theoretical interest.

When predictors are completely uncorrelated and fully orthogonal, the results of a multivariable regression model and separate regressions with one predictor each will be virtually identical. Multiple regression comes into its own for data with non-orthogonal predictors. For such data, it serves as a mathematically principled arbiter for teasing apart relevant from irrelevant predictors. However, when predictors are strongly correlated, i.e., for collinear data, this arbitrage tends to result in counterintuitive and uninterpretable coefficients (Belsley, Kuh, & Welsch, 1980; Farrar & Glauber, 1967). In this study, we review statistical methods that work around this problem.

When a data set is characterized by substantial collinearity, several problems arise. First, as already mentioned, parameter estimates may assume unexpected and theoretically uninterpretable values. Second, the model fit to the data will be unstable, in the sense that removal of just a few data points may have substantial consequences for the estimates of regression parameters. This holds both for linear regression and for the linear mixed model. Third, it can happen that no predictor on its own is significant, whereas all predictors jointly are successful in explaining a significant part of the variance in the response (Chatterjee, Hadi, & Price, 2000).

In what follows, we begin with an introduction to the problem of collinearity[1] and its adverse consequences for the magnitude and sign of estimated coefficients. We then describe a data set with substantial collinearity that will serve as the test case for our analyses. Subsequently, we introduce and illustrate three

* Corresponding author.
E-mail address: fabian.tomaschek@uni-tuebingen.de (F. Tomaschek).

[1] In the context of nonlinear regression, collinearity also rears its ugly head in the form of concurvity. Concurvity can render models such as generalized additive (mixed) models unstable. We therefore briefly discuss how concurvity can be assessed, and what measures the analyst might consider when concurvity is high, in the appendix.

https://doi.org/10.1016/j.wocn.2018.09.004
0095-4470/

methods for analyzing collinear data. The first of these is a non-parametric technique from machine learning, random forests. Random forests enable the analyst to assess the relative importance of predictors. The second method is supervised component generalized linear regression (SCGLR). SCGLR performs dimensionality reduction on the predictor space, resulting in a smaller set of orthogonal predictors (the supervised components). SCGLR comes with visualization methods for inspecting how the original predictors load on the supervised components, and it provides regression coefficients for the original predictors that are properly shrunk. The third method that we discuss is the elastic net, a regularized regression technique that not only shrinks coefficients, but shrinks some of these completely to zero. This method therefore can be used to perform variable selection. For each method, we introduce the general concepts, and then illustrate its use for our example data set.

There is no fixed set of guidelines that guarantee the "correct" analysis of collinear data. George Box's famous aphorism that all models are wrong but some are useful (Box, 1976) is especially relevant with respect to models for highly collinear data. The methods we review in the present study therefore provide the analyst with a toolkit that we find useful for exploring and understanding in complementary ways to what extent, and how a response might be shaped by a set of collinear predictors.

All analyses discussed in this study are documented step by step in the Supplementary Materials, to be downloaded from https://osf.io/5merb/. For these analyses, we made use of the statistical programming environment R (R Core Team, 2018) and specialist packages available for R (introduced below).

2. Suppression and enhancement

Suppression and enhancement occur in the linear regression model when two (or more) predictors for a given response Y are strongly correlated. Take, for example, an analysis in which response times (dependent variable Y) in auditory lexical decision have to be predicted by word frequency counts in American English (predictor A) and British English (predictor B). Given that such frequency counts will tend to be strongly correlated, suppression and enhancement are likely to make the coefficients of the regression model uninterpretable. To understand why this happens, first consider the case in which we fit two one-predictor regression models to Y,

$$Y_i = \beta_0 + \beta_A A_i + \epsilon_i,\ \epsilon_i \sim \mathcal{N}(0, \sigma), \tag{1}$$

$$Y_i = \beta_0 + \beta_B B_i + \epsilon_i,\ \epsilon_i \sim \mathcal{N}(0, \sigma). \tag{2}$$

where the β_0 represent the intercepts, β_A and β_B denote the coefficients for predictors A and B, and ϵ is a Gaussian error term. When A and B are uncorrelated and completely orthogonal, the results of these two one-predictor models will almost completely identical to a multivariable regression model in Y in predicted from A and B jointly:

$$Y_i = \beta_0 + \beta_A A_i + \beta_B B_i + \epsilon_i,\ \epsilon_i \sim \mathcal{N}(0, \sigma). \tag{3}$$

In this case, the multivariable regression model has nothing to add about the effects of A and B that we did not already know from the two one-predictor analysis. However, when A and B are correlated, and not strictly orthogonal, then multiple regression comes into its own as the arbiter deciding which predictors should be given more (or less) weight. When predictors are only mildly correlated, there is little collinearity and the weights estimated by the multiple regression model (3) will make sense, but when strong collinearity is present, the resulting model will become theoretically uninterpretable.

Following Friedman and Wall (2005), we illustrate this phenomenon by varying the correlation between predictors A and B, while keeping constant the correlations between A and Y as well as the correlations between B and Y. We simulated multiple data sets with 1000 observations each, using the `mvrnorm` function from the **MASS** package (Venables & Ripley, 2002). Y, A and B are all standard normal random variables. We manipulated the correlation between A and B (r_{AB}) to range from close -1 to close to $+1$ in steps of 0.01. We fixed the correlation between B and Y at $r_{BY} = 0.3$, but considered three different correlations between A and Y : $r_{AY} = -0.3$, $r_{AY} = 0.0$ and $r_{AY} = 0.6$. When $r_{AB} = 0$, β_A is equal to r_{AY} and $\beta_B = r_{BY}$.

Fig. 1 illustrates the consequences of varying the correlation between A and B for the estimates of slopes β_A and β_B (top panels) and the corresponding t-values (bottom panels). Across all panels of Fig. 1, dashed lines represent β_A and solid lines β_B. The three values of r_{AY} are listed above their respective panels.

First consider the panels graphing coefficients against r_{AB}. When r_{AB} is zero, β_A is -0.3 when $r_{AY} = -0.3$, it is 0 when $r_{AY} = 0$, and it is 0.6 when $r_{AY} = 0.6$. As r_{BY} is fixed at 0.3, β_B is always 0.3 when $r_{AB} = 0$. When r_{AB} moves away from zero, the coefficients change, and the more extreme r_{AB} becomes, the more extreme the changes in the coefficients are. When r_{AB} approximates 1, we find large positive and negative values for both β_A and β_B. Which predictor receives a positive coefficient and which a negative depends on r_{AB}. When r_{AB} is shifted towards -1, coefficients are not enhanced, but suppressed: both β_A and β_B assume smaller values than they have when $r_{AB} = 0$. It is noteworthy that β_A is strongly enhanced even when $r_{AY} = 0$.

Estimates of the t-values associated with the coefficients also vary with r_{AB} and can be very large for extreme positive values of r_{AB}. This leads to false positives for β_A when $r_{AY} = 0$ and r_{AB} is large. In other words, the model supports a significant effect of A although there is in fact none. False negatives arise when $r_{AY} = -0.3$, $r_{BY} = 0.3$, and r_{AB} is close to -1. In other words, the model does not support a significant effect of A and B although they are in fact significantly correlated with Y. In fact, strong collinearity can give rise to a model that succeeds in explaining variance of the predictor, without a single regressor being significant (see, e.g., Chatterjee & Hadi, 2012b; Friedman & Wall, 2005; Hadi, 1988, for examples).

Large coefficients with opposite sign for strongly correlated predictors are the hallmark of collinearity. In this case, the coefficients become difficult to interpret. For the above example of American and British frequency of occurrence, one frequency measure will reveal a coefficient with the expected negative sign, but the other frequency measure will emerge with a coefficient with an uninterpretable positive sign.

When strong collinearity is present, it is important to take a step back, and to address the question of how the artifacts of strong collinearity are best avoided. Before introducing