Multivariate Statistics

R. H. Baayen
Karl Eberhards University, Tübingen and University of Alberta, Edmonton

Introduction

Multivariate analysis deals with observations made on many variables simultaneously. Data sets with such observations arise across many areas of linguistic inquiry. For instance, (Jurafsky, Bell, Gregory, & Raymond, 2001) provide an overview of the many factors that co-determine a word's acoustic duration (including its neighboring words, syntactic and lexical structure, and frequency). The importance of these factors is determined with the help of multiple regression modeling of data extracted from speech corpora. Koesling, Kunter, Baayen, and Plag (2012) used multivariate analysis to study the pitch contours of English tri-constituent compounds, with as predictors not only time and compound structure, but also speaker, word, a word's frequency of occurrence, and the speaker's sex. In morphology, the choice between two rival affixes can depend on a wide range of factors, as shown for various Russian affix pairs by Janda et al. (2012). F. Jaeger (2010) showed that whether the complementizer *that* is present in an English sentence depends on more than 15 different factors. Gries (2003) and Bresnan, Cueni, Nikitina, and Baayen (2007) clarified the many factors that join in determining the choice of particle placement and the dative constructions respectively. In psycholinguistics, multivariate methods are becoming increasingly important (see, e.g. Kuperman, Schreuder, Bertram, & Baayen, 2009, for eye-tracking research), especially with the advent of so-called megastudies (Balota, Cortese, Sergent-Marshall, Spieler, & Yap, 2004). Multivariate methods have a long history of use in sociolinguistics (Sankoff, 1987), and play an important role in present-day dialectometry (Wieling, n.d.). What is common across all these studies is that they address linguistic phenomena for which monocausal explanations fail. Many phenomena can only be understood properly when a great many explananda are considered jointly. This is where multivariate statistics come into play.[1]

Table 1 presents a general description of a multivariate data set with n cases or observational units, presented on the rows. Observations on k different random variables $X_1, X_2, \ldots, X_k$ (presented in the columns) describe the properties of a given case. These properties can be numerical (e.g., acoustic duration in ms., frequency of occurrence in a 100 million word corpus, a response latency in a word naming experiment) or categorical (e.g., word category, discourse type, the sex of a speaker, dialect). Categorical predictors are referred to as *factors*. The values that a factor can assume are known as its *levels*. For

[1] This chapter assumes familiarity with all concepts discussed in chapters 1 (Descriptive Statistics) and 15 (Basic Significance Testing).

instance, in a given data set, a factor such as (major) Word Category may have as its levels *noun, verb*, *adjective*, and *adverb.*

Table 1: A multivariate data set with n cases (rows) and k variables (columns).

Cases	Variables			
	X_1	X_2	$\dots$	X_k
1	x_{11}	x_{12}	$\dots$	x_{1k}
2	x_{21}	x_{22}	$\dots$	x_{2k}
3	x_{31}	x_{32}	$\dots$	x_{3k}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	x_{n1}	x_{n2}	$\dots$	x_{nk}

The objective of multivariate analysis is to clarify how the variables pattern together and how they might distinguish the different cases on which the variables are observed. Most data sets are multivariate, and for a proper understanding of the structure of the data, it is often most informative to consider the different variables simultaneously.

Multivariate data sets fall into two main classes. On the one hand, we have data sets for which all variables are equally important. For such data sets, the primary interest will be in how the variables pattern together, and how they group or cluster the different cases, or on the causal relations between the variables.

On the other hand, interest may focus on how a specific variable, henceforth the *response*, is predicted from the other variables, henceforth the *predictors.* The response can be numeric or categorical. In the latter case, the goal of the analysis can be described as classification, i.e., the assignment of the different cases to the different classes defined by the levels of the response. However, not only the accuracy of the predictions, whether continuous or categorical, is of interest, but also how the variables pattern together to yield the predictions. This chapter provides an overview of some important methods for analyzing data with a specific response variable.

Within the limits of a single chapter, it is impossible to do justice to the full richness of the individual methods. The goal of this chapter is to provide the reader with the gist of the different approaches, and to provide examples that illustrate what can be accomplished with these methods. For further details, references are provided to both book-length introductions and to published studies using these methods.

Data sets from experiments often have a repetitive structure that requires special attention. Consider an experiment in which 20 different speakers read aloud 15 different words. Such a data set will have the structure shown in Table 2, where for each Subject (speaker) there are $n = 15$ cases, one for each Item (word), and where for each Item (word) there are $g = 20$ cases, one for each Subject. Experimental designs with this kind of repetitive structure are known as *repeated measures* designs.

Factors such as Subject and Item typically have many levels, which distinguishes them from factors such as Word Category (noun, verb, adjective, adverb) or the speaker's sex (female, male). Furthermore, subjects and items are — ideally — sampled randomly from populations that have many more members than the subjects and items that happen to have

Table 2: A repeated measures data set with gn cases with observations on k variables collected for n items and g subjects.

Cases	Response	Predictors					
	Y	X_1	X_2	$\dots$	X_k	Subject	Item
1	Y_1	x_{111}	x_{121}	$\dots$	x_{1k1}	1	1
2	Y_2	x_{211}	x_{221}	$\dots$	x_{2k1}	1	2
3	Y_3	x_{311}	x_{321}	$\dots$	x_{3k1}	1	3
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	1	$\vdots$
n	Y_n	x_{n11}	x_{n21}	$\dots$	x_{nk1}	1	n
n+1	Y_{n+1}	x_{112}	x_{122}	$\dots$	x_{1k2}	2	1
n+2	Y_{n+2}	x_{212}	x_{222}	$\dots$	x_{2k2}	2	2
n+3	Y_{n+3}	x_{312}	x_{322}	$\dots$	x_{3k2}	2	3
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	2	$\vdots$
2n	Y_{2n}	x_{n12}	x_{n22}	$\dots$	x_{nk2}	2	n
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	
(g-1)n+1	$Y_{(g-1)n+1}$	x_{11g}	x_{12g}	$\dots$	x_{1kg}	g	1
(g-1)n+2	$Y_{(g-1)n+2}$	x_{21g}	x_{22g}	$\dots$	x_{2kg}	g	2
(g-1)n+3	$Y_{(g-1)n+3}$	x_{31g}	x_{32g}	$\dots$	x_{3kg}	g	3
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	g	$\vdots$
gn	Y_{gn}	x_{n1g}	x_{n2g}	$\dots$	x_{nkg}	g	n

been used in the experiment. By contrast, the levels 'female' and 'male' exhaust the levels of the speaker's sex, there are no other levels in the population. Factors such as Subject and Item are referred to as *random effect factors*, and factors such as Sex as *fixed-effect factors*. In data sets with subjects and items, the other predictors can quantify properties of the subjects (e.g., age in years, sex, native speaker of English), properties of the items (e.g., a word's frequency, its word class, whether it is morphologically complex), or properties of the experiment (e.g., the number of trials a subject is in the experiment when a given sentence is presented). For the example presented in Table 2, all the predictors X_i are bound to the items and represent properties of these items, as indicated by the indexation of the first subscript of the predictor values $x_{...}$.

A great variety of multivariate techniques is available to the researcher, see, e.g., Venables and Ripley (2002) and Everitt (2005) for overviews. It will often be useful to study a given data set with more than one technique, as the strengths of one technique may counterbalance the weaknessses of the other. Only a subset of the available multivariate statistical methods is described in this chapter, which focuses on multiple regression and classification models.

Multiple Regression

Basic concepts

When the response variable is a measurement (e.g., acoustic duration in ms., response latency, pitch), and when there are no repeated measures, a multiple regression analysis models the response Y as a function of a weighted sum of the predictors and Gaussian (normally distributed) by-observation noise (ϵ).

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \epsilon, \qquad \epsilon \sim \mathcal{N}(0, \sigma^2) \tag{1}$$

When all predictors X_i are numerical, the analysis is described as a regression analysis. When all predictors X_i are factors, the analysis is referred to as analysis of variance. When both factorial predictors and numerical predictors are combined, the analysis is an analysis of covariance.

The goal of regression modeling is to approximate the observed values of the response as precisely as possible by decomposing the response into a weighted sum of the predictors. Models as defined by (1) make several important simplifying assumptions that facilitate the estimation of the model's parameters (the coefficients $\beta_0, \beta_1, \ldots, \beta_k$). First, the contribution of each predictor is assumed to be linear. When there is only one predictor, the fitted values are on part of a straight line. When there are two predictors, the fitted values are located on part of a flat surface. For more predictors, the fitted values are part of a flat hypersurface. Second, the errors (the difference between the observed and fitted values) are supposed to follow a normal distribution with mean zero and some unknown standard deviation (to be estimated from the data). Third, the errors are assumed to be independent, and all are supposed to follow the same normal distribution. This means that wherever one inspects the positioning of the observed data points with respect to the fitted line, plane, or hyperplane, one finds a cloud of points around the predicted values that is equally thick everywhere.

The regression model (1) specifies how the observed responses Y can be approximated given the values of the predictors X_i, $i = 1, \ldots, k$. Analysis of variance is a special case of regression in which the fitted values are the group means defined by the levels of the factorial predictors. For instance, given two factors with two and three levels respectively, there are six group means. The regression equation for analysis of variance specifies how these group means can be constructed. There are many ways in which this can be achieved, all of which recode factor levels numerically using *dummy coding.* Here, we focus on *treatment coding*, which offers the advantage of clarity of interpretation for analysis of covariance. Analyses using treatment coding select one group mean as point of departure, and specify coefficients that quantify the differences between this group mean and the other group means. Figure 1 and Table 3 illustrate the basic principles.

First consider Figure 1. The left panel shows a standard regression line, for 10 equally spaced values on the horizontal axis. The intercept β_0 of this line is at 1, and its slope β_1 equals 2. The right panel shows a similar line connecting two group means with values 1 for level a and 3 for level b. Since the group means of the two levels are exactly 1 unit apart on the horizontal axis, the difference between the two group means, 2, is equal to the slope of the line connecting the two group means. For both regression and analysis of variance, the same regression equation holds: $Y = 1 + 2X + \epsilon$. When a factor has more than two levels, say m, then there are $m-1$ contrasts with the reference level, which are represented on $m-1$

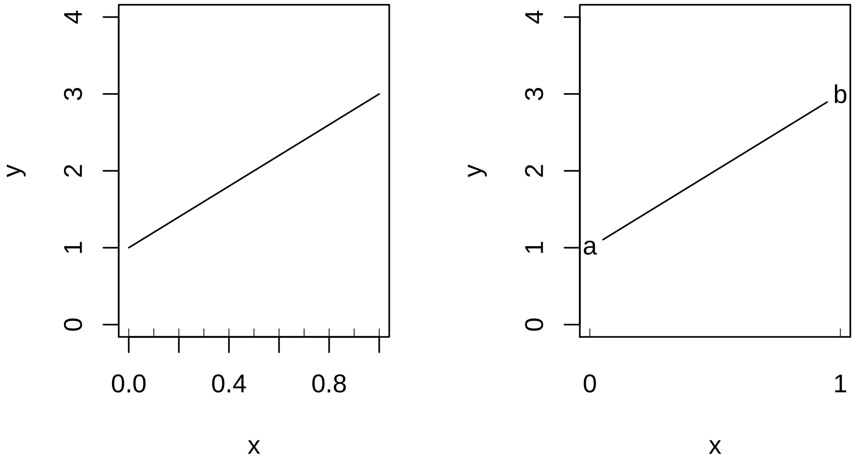

Figure 1. A regression line (left) and a factorial contrast between a reference group mean a on the intercept and a group mean b. The difference between the two group means, the contrast, is equal to the slope of the line connecting a and b: 2. Both the regression line and the line connecting the two group means are described by the line $y = 1 + 2x$.

orthogonal dimensions. Thus, a 'univariate' one-way analysis of variance with a single factor with more than two levels is recoded under the hood as a multivariate regression model.

Table 3: An example of treatment dummy coding for two-way analysis of variance.

Cases	A	B	X_1	X_2	X_3	X_4
1	a	e	1	0	0	0
2	a	f	1	0	0	1
3	b	e	1	1	0	0
4	b	f	1	1	0	1
5	c	e	1	0	1	0
6	c	f	1	0	1	1

Table 3 illustrates dummy coding for a fictive data set with 6 cases and two factorial predictors, one with three levels, and one with two levels. The reference group mean is represented by A=a and B=e. Each of the other five group means is defined by a unique combination of the dummy predictors X_2, X_3 and X_4. The multiple regression equation

$$Y = \beta_0 X_1 + \beta_1 X_2 + \beta_2 X_3 + \beta_3 X_4 + \epsilon, \qquad (2)$$

together with the dummy coding of Table 3, defines the group means listed in Table 4.

The regression equations (1) and (2) define flat planes in two or more dimensions. In the case of regression, the fitted data points are on such planes, whereas in the case of

Table 4: Predicted group means given the dummy coding in Table 3 and regression equation (2).

Cases	A	B	predicted group mean
1	a	e	β_0
2	a	f	$\beta_0 + \beta_3$
3	b	e	$\beta_0 + \beta_1$
4	b	f	$\beta_0 + \beta_1 + \beta_3$
5	c	e	$\beta_0 + \beta_2$
6	c	f	$\beta_0 + \beta_2 + \beta_3$

analysis of variance, the predicted group means are located on these planes. However, the assumption that the regression surfaces are flat (hyper)planes is often too simplistic. The standard linear model allows the user to relax this assumption by introducing *multiplicative interactions*. For a regression model with predictors X_1 and X_2, the interaction is obtained by adding a third predictor which has as its values the product of the values of X_1 and X_2:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon. \tag{3}$$

The left and center panels of Figure 2 visualize the general regression surface defined by a multiplicative interaction. The left panel plots, for different values of X_2, the regression line for the response as a (still linear) function of X_1. For large negative values of X_2, the effect of X_1 has a negative slope. As the values of X_2 increase, this effect reverses and the slope becomes positive. The center panel uses a contour plot to present the joint effect of X_1 and X_2. Contour lines connect points with the same fitted value. Lighter shades of gray indicate higher values, darker shades of gray indicate lower values. The contour plot, which visualizes a hyperbolic plane, is easier to interpret, as it allows the analyst to compare the joint effect of X_1 and X_2 on the response for any pairs of (x_1, x_2) values. It should be kept in mind that for a given data set, only part of a hyperbolic plane is used, for instance, part of the upper left corner, just as when fitting data points to a straight line, only a line segment, i.e., only part of an infinitely long line, is used.

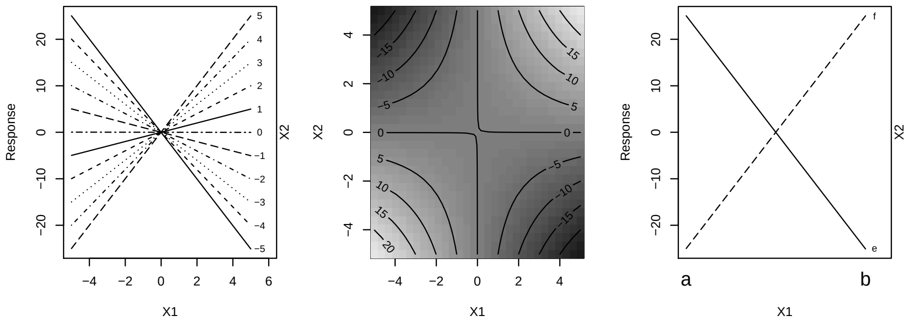

Figure 2. Multiplicative interactions in the linear model.

The third panel of Figure 2 illustrates the corresponding so-called *crossover interaction* for two factorial predictors, each with two levels. The effect of X_2 reverses for the levels of X_1. A model with an interaction for the 2 by 3 design described in Table 3 is

$$Y = \beta_0 X_1 + \beta_1 X_2 + \beta_2 X_3 + \beta_3 X_4 + \beta_4 X_2 X_4 + \beta_5 X_3 X_4 + \epsilon, \tag{4}$$

and the expected group means are listed in Table 5. The weights β_4 and β_5 break the parallelism of the effect of factor B within each level of A. In Table 4, where there is no interaction, the contrast between the means for levels e and f is always the same, irrespective of the levels of A, and equal to β_3. With the interaction, as shown in Table 5, the effect is modified by β_4 for factor level b (of A) and by β_5 for factor level c (of A).

Table 5: Predicted groupmeans for the data in Table 3 given the regression equation (4).

Cases	A	B	predicted group mean
1	a	e	β_0
2	a	f	$\beta_0 + \beta_3$
3	b	e	$\beta_0 + \beta_1$
4	b	f	$\beta_0 + \beta_1 + \beta_3 + \beta_4$
5	c	e	$\beta_0 + \beta_2$
6	c	f	$\beta_0 + \beta_2 + \beta_3 + \beta_5$

A linear model can comprise both numeric and factorial predictors. When the effect of a numeric covariate varies depending on the specific level of a given factor, we have an interaction of that covariate by the factor. An example is presented in Table 6 and Figure 3.

Figure 3 depicts a regression line with a positive slope for level a of factor A, but a negative slope for level b. Table 6 shows the dummy coding for this data set, with a column of ones for the intercept, a contrast for level b of A, the values of the covariate (X_3), which repeat within the levels of A, and the multiplicative interaction (X_2X_3). The regression equation for this example is

$$Y = \beta_0 X_1 + \beta_1 X_2 + \beta_2 X_3 + \beta_3 X_2 X_3 + \epsilon. \tag{5}$$

The regression lines in Figure 3 are described by the equations $y = 2 + 3x$ for level a and $y = 6 - 2x$ for level b. The β weights for these data given the dummy coding in Table 6 and equation (5) are as follows. Given a as reference level, the intercept of the model will be the intercept of the regression line for level a, so $\beta_0 = 2$. The intercept for the second regression line is at 6, hence β_1, which quantifies the difference between the two group means when the covariate X_3 is 0, i.e., where the regression lines cross the Y-axis, equals 4. The slope of the line for level a is $3 = \beta_2$. Finally, the slope for the regression line for level b is -2, a difference of -5 with the slope for level a. This difference is the contrast for the slopes of the two lines, hence $\beta_4 = -5$. Note that since the product X_2X_3 is zero for factor level a, the coefficient for X_2X_3 serves as a correction (i.e., a contrast) on the slope for the regression line for level a, but, as required, only within level b.

Given a regression equation for a given data set, a first question that arises is how to estimate the parameters of the model. Fortunately, excellent algorithms are available for

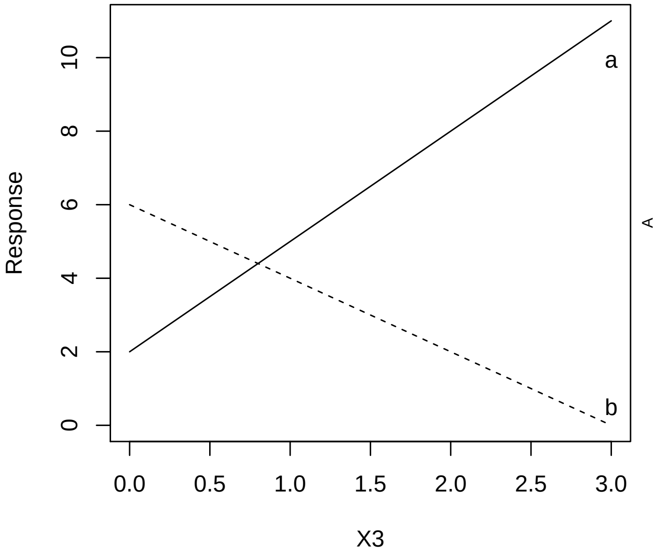

Figure 3. Example of an interaction of a factor and a covariate in an analysis of covariance.

Table 6: An example of treatment dummy coding for an analysis of covariance with an interaction.

Cases	A	X_1	X_2	X_3	X_2X_3
1	a	1	0	1	0
2	a	1	0	2	0
3	a	1	0	3	0
4	b	1	1	1	1
5	b	1	1	2	2
6	b	1	1	3	3

doing this, which have been implemented in many software packages. Although the mathematics for simple linear regression and analysis of variance are relatively straightforward, the more sophisticated algorithms underlying mixed-effects regression models and generalized additive models, which will be discussed below, require substantial training in mathematics. Fortunately, these models can be used responsibly without having to know the details of the underlying mathematical theory. In the case of analysis of (co)variance, dummy coding can be either hand-crafted by the analyst, or a specific dummy coding scheme can be specified, with the actual creation of dummy variables being left to the software.

When fitting a regression model to the data, the software will generally return several kinds of information to the user. First, information is provided about the estimated values of the coefficients for the intercept, the slopes, and the factor contrasts. For a given coefficient, a measure is provided about the uncertainty of the estimate in the form of a standard error. The ratio of the estimate and its standard error yields a statistic that follows a t-distribution. If the observed value of the t statistic is far out in one of the tails of the distribution, there is reason for surprise about the magnitude of the estimate, and a p-value based on the t-distribution will allow the researcher to evaluate whether a coefficient is surprisingly different from zero.

By way of example, consider a study of pitch (F0) in English tri-constituent compounds (Koesling et al., 2012). Three predictors are of interest here: *Time*, *Sex* (female versus male), and *Branching*. Branching is a factor that distinguishes between four kinds of compound stress patterns on the basis of branching direction (left or right) and location of the stress (first, second, or third noun):

code	branching direction	stress pattern	example
LN1	left	[ŃN]N	[háy fever] treatment
LN2	left	[NŃ]N	[science fíction] book
RN2	right	N[ŃN]	business [crédit card]
RN3	right	N[NŃ]	family [Christmas dínner]

We expect pitch (in semitones) to decline over time. Pitch is also expected to be lower for men than for women. Of main interest is whether there are significant differences in pitch contours for the different types of compounds as distinguished by the factor Branching.

Table 7 presents the coefficients of a simple main effects model fitted to the data, described (using the symbolic description language of

$$\texttt{Pitch} \sim \texttt{Time + Sex + Branching}. \tag{6}$$

In this model formula, the intercept and the error term are not mentioned explicitly. Nevertheless, any software package will provide the analyst with estimates of both. In Table 7, the intercept (93.1331) represents the pitch predicted at word onset for female speakers for branching condition LN1. The negative slope for Time (-0.0327) indicates that pitch decreases over time, as expected. For male speakers, the intercept has to be lowered by 9.9297, again as expected. The three contrasts in the last part of Table 7 specify the difference in pitch between LN2 and LN1, between RN2 and LN1, and between RN3 and LN1. The small standard errors, the large t-values, and the small p-values suggests that all coefficients are significant.

Table 7 does not list all possible contrasts between the four branching conditions. Of the $\binom{4}{2} = 6$ possible contrasts, only three are listed. For instance, no information is provided as to whether there is a real difference between the RN2 and RN3 conditions. In addition, it would be useful to know which contrasts remain significant after being corrected for multiple comparisons. Figure 4 presents each of the six contrasts together with its 95% confidence interval, using Tukey's all-pairs comparison method (Hothorn, Bretz, & Westfall, 2008). Of the four contrasts, only those between RN3 and LN2, and RN3 and RN2 do not reach significance, as their confidence intervals straddle zero.

The model considered thus far assumes that the slope of the effect is the same across all branching conditions, and the same across female and male speakers. This is a simplifying

	Estimate	Std. Error	t value	p value
Intercept	93.1331	0.0437	2131.1298	0.0000
Time	-0.0327	0.0005	-59.6105	0.0000
Sex=male	-9.9297	0.0321	-309.4371	0.0000
Branching=LN2	0.3244	0.0447	7.2507	0.0000
Branching=RN2	0.4603	0.0447	10.3051	0.0000
Branching=RN3	0.4279	0.0447	9.5816	0.0000

Table 7: Coefficients of an analysis of covariance model fitted to the pitch of English tri-constituent compounds.

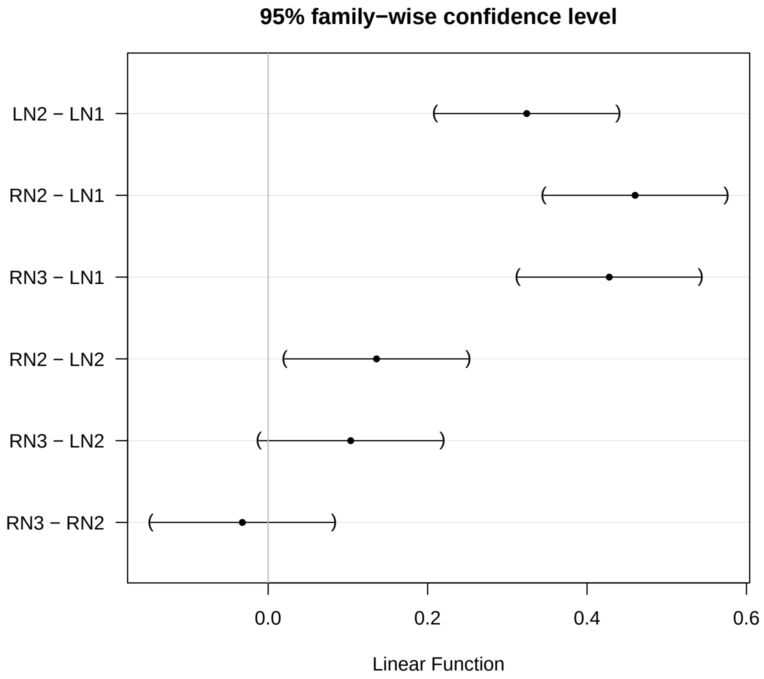

Figure 4. Tukey all-pairs confidence intervals for contrasts between mean pitch for different branching conditions across English tri-constituent compounds.

assumption, and we need to check whether it is justified by allowing interactions of *Time* by *Branching* and *Time* by *Sex* into the model. It turns out that both interactions improve the fit of the model to the data. In order to assess the importance of the various terms in the model, we compare a sequence of nested models, step by step increasing in complexity. For the present data, the sequence of models (where we make explicit the presence of an intercept term by adding 1 to the model formula)

$$\text{Pitch} \sim 1$$
$$\text{Pitch} \sim 1 + \text{Time}$$
$$\text{Pitch} \sim 1 + \text{Time} + \text{Sex}$$

$$\text{Pitch} \sim 1 + \text{Time} + \text{Sex} + \text{Branching}$$
$$\text{Pitch} \sim 1 + \text{Time} + \text{Sex} + \text{Branching} + \text{Time} : \text{Branching}$$
$$\text{Pitch} \sim 1 + \text{Time} + \text{Sex} + \text{Branching} + \text{Time} : \text{Branching} + \text{Time} : \text{Sex}$$

is evaluated statistically by the sequential F-tests listed in Table 8, to which the reduction in AIC has been added as a measure of variable importance.

	Res. Df	Df	F-value	p-value	reduction in AIC
intercept	47573				
+time	47572	1	3549.746	0.000	1159.878
+sex	47571	1	95840.330	0.000	52383.491
+branching	47568	3	44.433	0.000	127.023
+time*branching	47565	3	11.822	0.000	29.456
+sex*branching	47562	3	2.794	0.039	2.383

Table 8: Sequential model comparison for Pitch in English tri-constituent compounds (number of observations: 47574)

The column labeled 'Res. Df' lists the residual degrees of freedom, which is equal to the number of observations in the data minus the number of parameters. The first model, which has an intercept only (which in this case represents the grand average) has only one parameter (the intercept), and hence 47574-1=47573 residual degrees of freedom. The second column, `Df`, lists the number of parameters required when adding one or more predictors. For *Time*, which requires a slope coefficient, one additional parameter is required. For *Branching*, which has 4 levels, 3 contrast coefficients are required when it is added in as a simple main effect. The column with p-values is obtained from the F statistics given 'Df' and 'Res. Df'. The final column lists the change in AIC, Akaike's information criterion, which is defined as

$$AIC = 2k - 2\ln(L) \tag{7}$$

where L denotes the likelihood of the model and k denotes the number of parameters. The AIC measure describes the tradeoff between a model's accuracy and its complexity. On the one hand, a model should be as accurate as possible. At the same time, the model should be as simple as possible. Simpler models have lower k, more accurate models have higher L. In other words, the AIC measure penalizes models for their complexity. Lower values of AIC indicate a better fit of the model. The greater the reduction in AIC obtained by adding a term to the model equation, the better the relative goodness of fit of the model. Furthermore, the greater the reduction in AIC is, the more important a term is.

For a set of n models with AIC values AIC_1, AIC_2, ..., AIC_n, we can select the model with the smallest AIC (model AIC_{min}), and calculate *evidence ratios* (ER)

$$\text{ER} = \exp\left(\frac{\text{AIC}_i - \text{AIC}_{\text{min}}}{2}\right) \tag{8}$$

that express the relative probability that the model with the minimum AIC is more likely to provide a more precise model of the data.

From Table 8 it is therefore immediately clear that *Sex* is the most important predictor, followed by *Time* and, at a distance, by *Branching*. The interaction of *Sex* by *Branching* adds only a small improvement to the model's goodness of fit. Its evidence ratio, $\exp(2.383/2) = 3.32$ nevertheless indicates that this model is approximately three times as likely to provide a description of the data that loses less information about the data as the model without the interaction.

A question with no definite answer is how to find the model that best describes the data. There are automatized search procedures that start with the simplest possible model and keep adding main effects and interactions until there is no significant improvement in goodness of fit. Instead of forward stepwise model selection, one can start with the most complex model and remove superfluous predictors until the simplest yet adequate model is obtained. Backward and forward selection heuristics can be combined. Some researchers prefer to use code that works through all possible models and then select the model with the best fit (see, e.g., Kuperman & Van Dyke, 2011; Lumley & Miller, 2009). Other researchers such as Harrell (2001) argue that only one model should be fit to the data, as *p*-values become meaningless when large numbers of models are fitted and compared. The present author favors hypothesis-driven exploration of the data, with theoretically potentially relevant predictors being added successively to the model specification. Further motivation of this research strategy is deferred till after discussion of generalized additive mixed modeling. However, irrespective of how a final model for the data is obtained, replication studies will be crucial for consolidating the validity of the conclusions reached.

In the absence of new data, bootstrap validation is one way in which the stability of the model parameters can be evaluated. Bootstrap validation fits a given model to a large number of bootstrap samples. Each bootstrap sample is a sample with replacement from the original data points. Some observations will appear more than once in a given bootstrap sample, and other observations will not appear at all. These observations constitute unseen, new data points. The accuracy of the model fitted to the bootstrap sample can be gauged by comparing its predictions with the actual values of the response for the unseen data points. Averaging across all bootstrap samples yields information about the extent to which the model overfits the data as well as about which predictors are significant across the bootstrap runs (see, e.g., Harrell, 2001) for detailed examples of bootstrap validation.

Data points that are located outside the cloud of data points have the potential of seriously distorting a regression model. There are several measures that help protect against overly influential outliers. First, if a predictor has a highly skewed distribution, a square root transformation or a log transformation may result in a more symmetrical distribution. For instance, word frequency distributions have a long right tail, and without a logarithmic transform, a small minority of very high frequency words will adversely dominate the regression model. Second, if the distribution of the response is highly skewed, a transformation rendering it more normal may be necessary (Box & Cox, 1964; Venables & Ripley, 2002). Without an appropriate transformation of the response, the distribution of the residuals will be non-normal, violating the fundamental assumption of multiple regression that the errors should be identically distributed. Third, one can inspect the *leverage* of the data points to identify potentially harmful outliers. The leverage of a data point quantifies how much the parameters of the model would change if the data point were not included when fitting the model. The greater this change, the more likely the data point is an outlier (see, e.g., Chatterjee, Hadi, & Price, 2000, for detailed discussion).

Finally, it is worth noting that regression modeling cannot tease apart the effect of predictors that are very strongly correlated. Data sets with highly correlated predictors are described as *collinear*. By way of example, a data set with four frequency measures taken from different corpora of contemporary written English would be highly collinear. A consequence of collinearity is that the coefficients for collinear predictors may be significant but with counterintuitive signs. For instance, frequency as a predictor for response latencies in various psycholinguistic tasks usually has a negative coefficient, indicating that as frequency increases, processing speed decreases. When two highly correlated frequency measures are entered into the regression equation, one will have the expected negative coefficient, but the other may have a significant positive coefficient. Jointly, the two highly correlated predictors provide a better fit, but from a cognitive perspective, the coefficients are no longer interpretable. The phenomena of *suppression* and *enhancement* in regression are well described in Friedman and Wall (2005). When the goal of modeling is to obtain accurate predictions, the adverse consequences of enhancement and suppression are not a concern. However, for the model coefficients to remain interpretable, the analyst has several choices. First, centering and scaling (subtracting the mean, and dividing by the standard deviation) often substantially reduces collinearity. This may not be sufficient for very strongly correlated predictors. In this case, the simplest option is to consider only one of a set of collinear predictors, e.g., select only one of the four frequency measures for inclusion in the model specification. Alternatively, a dimension reduction technique such as principal components analysis can be used to obtain a new frequency measure that combines the strengths of the four separate frequency variables.

Mixed-Effects Modeling

For repeated measures designs (see Table 2 above), the standard linear model is inappropriate. Although one could use dummy coding for subjects or items, this comes with several important disadvantages. First, the dummy coding will tune the model to the subjects and items in the experiment, but it will not allow inferences beyond exactly these subjects and items. The model does not generate predictions about unseen subjects and unseen items. Second, the standard linear model does not allow the user to gain insight into the correlational structure in the data with respect to subjects and items.

Mixed-effects models (Pinheiro & Bates, 2000; West, Welch, & Galecki, 2007), i.e., regression models that combine fixed-effect factors with random-effect factors such as subjects and items, treat random-effect factors as sources of random variation in the data. This random variation can manifest itself at various "sites" in a regression model. First, it can be tied to the intercept, in which case the intercept has to be adjusted upwards or downwards depending on which unit (level of a random-effect factor) was sampled for the experiment. For instance, if the response is the duration in ms of the vowel in a speech corpus, it is important to bring the speaker into the model as a random-effect factor, as speakers have different speech rates. Given an estimate of the average speech rate in the population, represented by the intercept (β_0) in the regression equation, the speech rate of a specific individual speaker can be obtained by taking the average speech rate β_0 and adjusting it upwards (for slow speakers) or downwards (for fast speakers) by an amount b_{0i} for speaker i. The mixed-effects regression model assumes that the b_{0i} adjustments follow a normal distribution with mean zero and unknown standard deviation that will be estimated from the data. In other words, the variation in speech rates is modeled as Gaussian noise around the population speech rate.

Such Gaussian noise need not be restricted to the intercept, it can extend to slopes and contrasts. For instance, the effect of frequency of occurrence can be stronger for some subjects, and weaker for others. This subject variation can be represented as Gaussian noise around the population slope for frequency. These considerations lead to the general mixed-effects regression equation

$$Y = (\beta_0 + b_0) + (\beta_1 + b_1)X_1 + \ldots + (\beta_k + b_k)X_k + \epsilon, \tag{9}$$

where $\epsilon, b_0, b_1, \ldots, b_k$ all are normally distributed with mean zero and unknown (and generally different) standard deviations. Table 9 charts the individual adjustments b_{ij} for coefficients $j = 0, 1, \ldots, k$ (columns) and subjects $i = 1, 2, \ldots, s$ (rows). Because the adjustments $b_{i.}$ are estimated for the same subject i, it is possible that any pair of (column) vectors of adjustments $\{b_{.n}, b_{.m}\}$ are correlated. As a consequence, the specification of a mixed-effects model is complete only with the matrix of pairwise correlations of the b (column) vectors. Which standard deviation and correlation parameters are actually required for a given dataset is an empirical issue. Generally, adjustments to the intercept (*random intercepts*) for subjects and items lead to substantially better models, less often, but regularly, adjustments to slopes (*random slopes*) are also well supported. In the literature, the adjustments are referred to as best linear unbiased predictors (BLUPs) or as posterior modes.

Table 9: Notation for adjustments to intercept and predictors.

level	random intercepts	random slopes			
number	b_0	b_1	b_2	$\ldots$	b_k
1	b_{01}	b_{11}	b_{21}	$\ldots$	b_{k1}
2	b_{02}	b_{12}	b_{22}	$\ldots$	b_{k2}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
s	b_{0s}	b_{1s}	b_{2s}	$\ldots$	b_{ks}

To illustrate mixed-effects modeling, consider again the study on pitch on English triconstituent compounds. The AIC for the best model obtained above is 252840.7. When we add random intercepts for *Subject* and for *Item*, the AIC of the model is 202574, a decrease of no less than 50266.7. A further improvement of the model is obtained by adding random slopes (contrasts) for *Sex*, which reduces the AIC by 539. The significance of the additional random effects structure (here, a standard deviation for the by-word adjustments for *Sex* and a correlation parameter for the by-word adjustments to the intercept and *Sex*) is assesed with a likelihood ratio test. The likelihood ratio statistic, defined as twice the difference of the log likelihoods of the more complex model (model 2) and the simpler model (model 1),

$$2\ln(L2/L1) = 2[\log(L2) - \log(L1)] \tag{10}$$

Groups	Name	Std.Dev.	Correlation
Word	Intercept	0.58029	
	Sex=male	0.50815	-0.703
Speaker	Intercept	3.01529	
Residual		2.01422	

Table 10: Standard deviations and correlation parameter for the random-effects structure of the mixed-effects model fitted to the pitch of English tri-constituent compounds.

follows a chi-squared distribution with as degrees of freedom the difference in the number of parameters. For the present data, the LRT statistic is 542.69, the number of additional parameters is 2, and the corresponding extremely small p-value indicates that the second model improves significantly on the goodness of fit.

One consequence of including random intercepts and slopes is that the interaction of *Branching* by *Sex*, which was the weakest predictor (see Table 8), is no longer significant. With the individual speakers in the model, the factor *Sex*, which groups speakers into females and males, becomes less important. An important general methodological issue that is illustrated by this example is that analyses that fail to bring subject and item random intercepts and slopes into the model may be anti-conservative, i.e., they may produce p-values that are smaller than they should be, and therefore may mislead the analyst into believing that a non-significant effect would be significant.

The estimates of the standard deviations and the correlation parameter are listed in Table 10. When reporting a mixed-effects model, it is essential to report these parameters as they are an intrinsic part of the model and provide the reader with insight into the magnitude of the different sources of random variation in the data and their interrelations.

The large negative correlation for the by-word random intercepts and random contrasts for *Sex* invites further interpretation. Figure 5 presents a scatterplot of the words in the plane spanned by the two dimensions of word-related variability in pitch. The horizontal axis represents the by-word random intercepts, which are calibrated for the reference level of *Sex*: *female*. The vertical axis represents the additional by-word adjustment required for the male speakers. Recall that male speakers have lower pitch, represented in the model by a downward shift of the population intercept for males. For some words, the shift for males is not down far enough, for others, it is too far down. The by-word random contrasts on the vertical axis show, for each word, how the intercept for the males has to be fine-tuned. In the lower right of the scatterplot, we find words such as *cream cheese recipe* and *student season ticket* for which females have a higher than average intercept (a large value on the horizontal axis), whereas in the upper left, one finds compounds such as *money market fund* and *pilot leather jacket* for which females have a lower than average intercept. Conversely, compared to the female baseline, the males have a higher than average pitch for the latter words, and a lower than average pitch for the former. What seems to be going on here is that pitch rises for words that speakers find more exciting and interesting. However, what is exciting and interesting differs between the sexes. Males show a clear disinterest in *woman fruit cocktail*, while the female disinterest in *money market fund* is absent for the males.

An example of more complex by-subject random-effects structure can be found in a

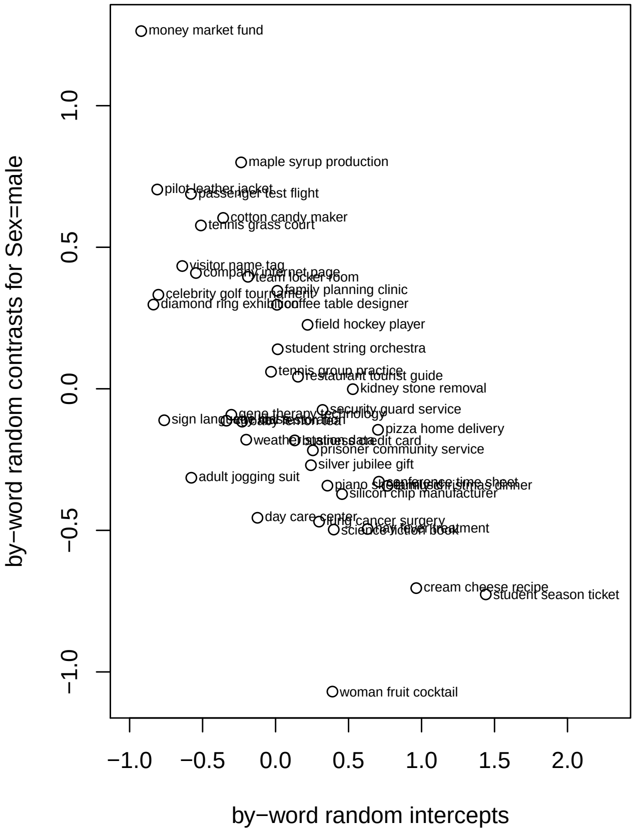

Figure 5. Correlation of the by-word random intercepts and the by-word random slopes for Sex=male in the linear mixed-effects model fitted to the pitch of English tri-constituent compounds.

large-scale self-paced reading study reported in Baayen and Milin (2010). Of interest here are two numerical predictors for the self-paced reading latencies: a word's frequency and its number of morphemes. The (log-transformed) frequency measure represents how practiced a word is, the morpheme count is a measure of its morphological complexity. By-subject variation with respect to these predictors indicates would imply that the experiment is picking up on by-subject variability in using (remembering) the words, as well as by-subject variation in the ability to deal with morphological complexity.

Figure 6 presents the by-subject random effects structure characterizing this data set. The top panels show the BLUPs, the bottom panels the subject-specific coefficients (the BLUPs incremented with the corresponding population mean values for the intercepts and slopes). In these scatterplot, dots represent subjects. All that changes between the upper and lower panels is position with respect to the vertical axis. The left panels indicate that fast responders (with a small BLUP, i.e., a small coefficient for the intercept) are least delayed by the number of morphemes in a word. Conversely, the slow responders are the

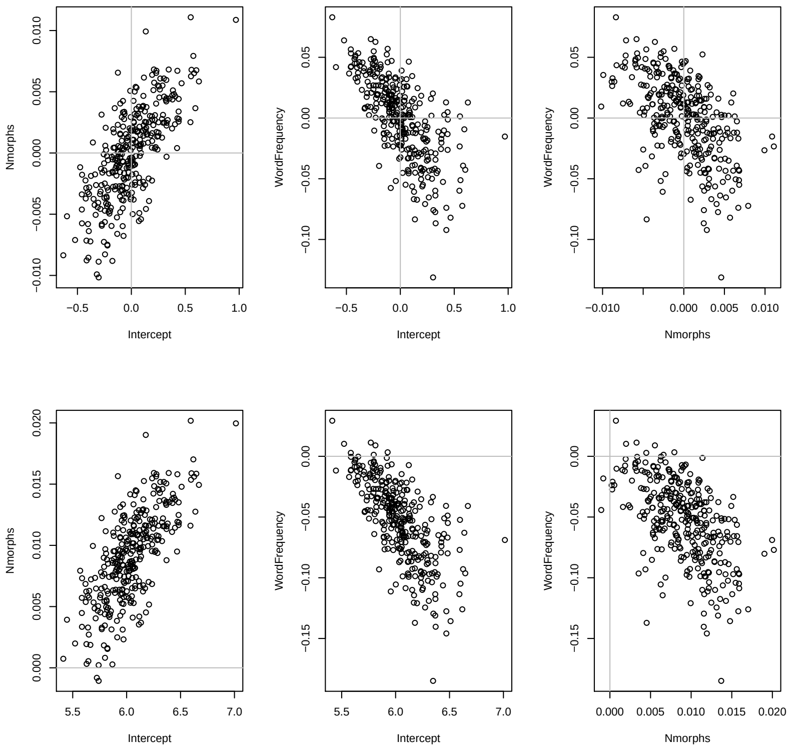

Figure 6. Random effects structure for subject. Upper panels: correlations of the BLUPs, Lower panels: correlations of the by-subject coefficients.

ones who are delayed most by morphological complexity. The center panels show that fast responders (low values for the intercept) have little or no facilitation from word frequency. Slow responders, on the other hand, show healthy facilitation from word frequency. The right panels point to a trade-off between word frequency and morphological complexity, such that subjects who are least affected by morphological complexity are also the subjects with the weakest, if any, facilitation from word frequency. The importance of mixed-effects models for language studies is that they clarify not only the main trends in the population, but also the correlational structure tied to subjects and items. For the present example, the trade-off between storage and computation across the subject population is one of the most interesting findings of the study.

Generalized Additive Models

The preceding analyses assumed that the effects of predictors are linear, and can be described mathematically as straight lines, flat planes, or flat hyperplanes. For numeric predictors, the multiplicative interaction defines a curved surface, but when one predictor is held constant, the effect of the other predictor is still linear (cf. Figure 2). The linearity assumption may be plausible for some data, but it can be very implausible for other data sets. The pitch data discussed in the preceding sections are a case for which the linearity assumption does not make sense at all. Anyone who has ever inspected a pitch contour for English knows that pitch does not decrease linearly with time.

In order to model the functional dependency of pitch on time correctly, a flexible toolkit is required that allows the analyst to consider nonlinear functional relations in two dimensions (wiggly lines) or more than two dimensions (wiggly surfaces and hypersurfaces). Generalized additive models (GAMs, Hastie & Tibshirani, 1990; Wood, 2006) provide the user with exactly such a toolkit.

A GAM combines a standard linear model with regression coefficients $\beta_0, \beta_1, \ldots, \beta_k$ with smooth functions `s()` in one or more predictors.

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k + s(X_i) + s(X_j, X_k) + \ldots + \epsilon, \qquad \epsilon \sim \mathcal{N}(0, \sigma^2) \tag{11}$$

For smooths in one predictor, a good choice is using cubic regression splines. Cubic splines fit piecewise cubic polynomials (functions of the form $y = a + bx + cx^2 + dx^3$) to non-overlapping intervals of the predictor values, such that at the points were intervals meet, the so-called knots, the transitions are smooth (by forcing the first and second derivative to be identical). The number of knots determines the smoothness of the curve. When too many knots are used, a curve is undersmoothed, when too few knots are postulated, the curve is oversmoothed. Recent advances in the mathematics of GAM modeling (see Wood, 2006, 2011) have resulted a range of algorithms (e.g., generalized crossvalidation and relativized maximum likelihood estimation) that make the estimation of the proper number of knots part of the general parameter estimation process.

For smooths in higher dimensions with isotropic predictors (predictors expressed on the same scale, such as longitude and latitude in dialectometry), thin plate regression splines are available, which fit a wiggly regression surface as a weighted sum of geometrically regular surfaces. For both isotropic predictors as well as for predictors that are measured on different scales, tensor products provide a flexible and generally faster alternative. Tensor products define wiggly surfaces given marginal basis functions, one for each dimension of the smooth. Typically, these basis functions are themselves cubic splines, and the greater the number of knots for the different basis functions, the more wiggly the fitted regression surface will be. Recently, it has become possible to combine splines and tensor products with random-effect factors, resulting in generalized additive mixed models (GAMMs).

Returning for a final time to the pitch data, the following sequence of models relax, step by step, the linearity assumptions with which we have worked thus far. Following the notational conventions of Wood (2006), with `s()` representing a cubic regression spline (when the basis function `bs` is set to `cr`) or a random effect (when the basis is set to `re`), we have:

```
Pitch ∼ 1 + Time + Sex + Branching + Time :  Branching +
        s(Speaker, bs="re") + s(Word, bs="re") + s(Word, Sex, bs="re")
Pitch ∼ 1 + Time + Sex + Branching + Time :  Branching +
        s(Speaker, bs="re") + s(Word, bs="re") + s(Word, Sex, bs="re") +
        s(Time, bs="cr")
Pitch ∼ 1 + Sex + Branching +
        s(Speaker, bs="re") + s(Word, bs="re") + s(Word, Sex, bs="re") +
        s(Time, bs="cr", by=Branching)
```

```
Pitch ~ 1 + Sex + Branching +
        s(Speaker, bs="re") + s(Word, bs="re") + s(Word, Sex, bs="re") +
        s(Time, bs="cr", by=Branching) + s(Time, bs="cr", by=Sex)
```

The second model allows the pitch contour to be a nonlinear function of Time. The third model allows this nonlinear function to differ for the four Branching conditions. In other words, this model specification tests for an interaction of a smooth in Time by Branching condition. Separate linear terms for *Time* and its interaction with *Branching* are no longer necessary. The final model adds a further smooth to relax the assumption that the smooth in *Time* is the same for the two sexes. Table 11 indicates that these models provide increasingly good fits to the data.

	Res. Df	Df	Deviance	F	p value	change AIC
linear	192647					
+s(Time)	190218	7.64	2428.5	80.5	0.0000	588.2
+s(Time, by=Branching)	187949	22.22	2269.4	25.9	0.0000	526.6
+s(Time, by=Sex)	187308	4.79	641.1	33.9	0.0000	153.0

Table 11: Model comparision for a series of models with increasing nonlinear structure fitted to the pitch data set.

Knowing that adding a smooth results in a significantly better fit does not inform us about the shape of the shape of the nonlinearity. As cubic splines and tensor products are black boxes to the end user, there are no parameters that might inform about the functional shape of the nonlinear prediction curves or surfaces. The only way to gain insight into these shapes is through visualization. The fitted smooths for *Pitch* as a function of *Time*, for each of the four branching conditions, is shown in Figure 7. For a discussion of the interpretation of these smooths, the reader is referred to (Koesling et al., 2012).

The use of GAMs for modeling wiggly surfaces is illustrated for two data sets, one addressing auditory comprehension with EEG, the other addressing lexical diffusion in the dialectometry of Dutch.

Kryuchkova, Tucker, Wurm, and Baayen (2012) studied the comprehension of isolated words, presented over headphones, using evoked response potentials measured at the scalp. They were specifically interested in the electrophysiological response to the danger of the words' referents, as gauged by independently collected danger ratings on a 9-point Likert scale. Here, we consider a generalized additive model fitted to the microvoltages elicited at channel FC2 with a spline in `Time` for the interval $[100, 400]$ ms post stimulus onset.

```
MicroVoltage ~ s(Time) + te(Time, Danger)
```

In this model equation, `te` denotes a tensor product. Figure 8 shows how the electrophysiological response of the brain varies with time as a function of a word's danger rating score. Darker shades of green indicate lower (negative) voltages, whereas brown to white colors indicate higher (positive) voltages. Focusing on the 150–350 ms time window, the graph shows a negative inflection around 150–200 ms post stimulus onset across all danger scores,

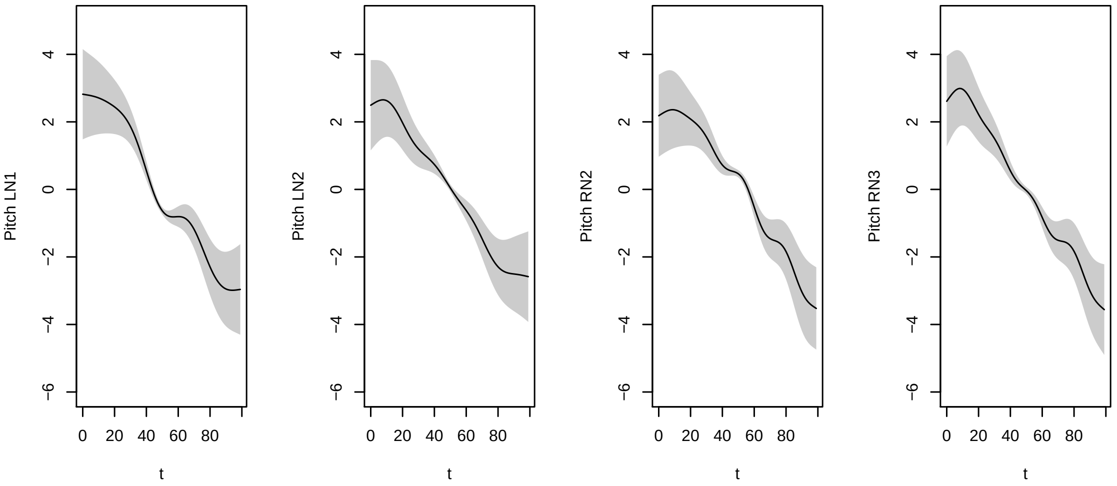

Figure 7. Fitted smooths (with 95% confidence intervals) for Pitch as a function of *Time* for the four branching conditions of the pitch data set of English tri-constituent compounds.

followed by a positive inflection. For words with higher danger ratings, this positive inflection has a reduced amplitude. This reduced posititivity in the 250–300ms time interval fits well with research on emotion processing in other modalities (see Kryuchkova et al., 2012, for further details).

It is worth noting that although one could dichotomize Danger into a factor with levels 'low' and 'high', followed by an inspection of the time intervals at which the curves for the low and high conditions diverge, the result would be a model with an inferior goodness of fit, in line with the literature on the detrimental costs of dichotomization of numerical predictors (Cohen, 1983; MacCallum, Zhang, Preacher, & Rucker, 2002; Baayen, 2010). The beauty of GAMs is that they make it possible to let the data speak for themselves without having to impose prior — often arbitrary — categorizations.

A final example of modeling nonlinear regression surfaces is based on the study of Wieling, Nerbonne, and Baayen (2011), who investigated word pronunciation distances from standard Dutch for 424 locations in the Netherlands. We focus here on an interaction of longitude, latitude, and word frequency, but note that other predictors representing socio-economic variables related to the informants can be included as well, allowing the analyst to integrate sociolinguistics with dialectometry. The model equation,

```
DialectDistance  ~  te(Longitude, Latitude, Frequency)
```

invokes a three-dimensional tensor that defines a complex hypersurface that can be represented graphically by means of a sequence of dialect maps for different frequencies, as shown in Figure 9 for four typical quantile frequencies.

The contour plots in this figure present, from left to right, the dialect distance maps for word frequency at the 0.05, 0.33, 0.66 and 0.95 quantiles. The graphs indicate that dialect leveling, which has progressed furthest for the lower frequency words, is highly regionally

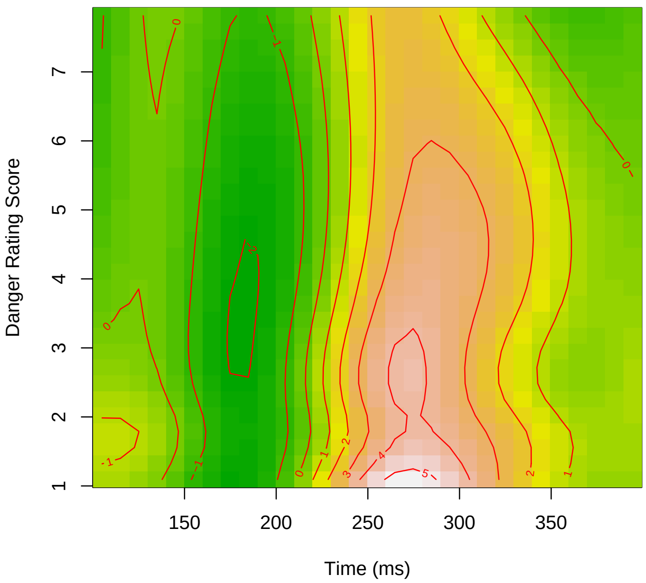

Figure 8. Tensor product for the interaction of Time by Danger Rating Score at channel FC2.

cohesive. Figure 9 fits well with the lexical diffusion model of (Wang, 1969). The greater the geographical distance from the heartland of the Dutch standard (central west), and the greater a word's frequency, the less the standard language has penetrated a speaker's lexicon.

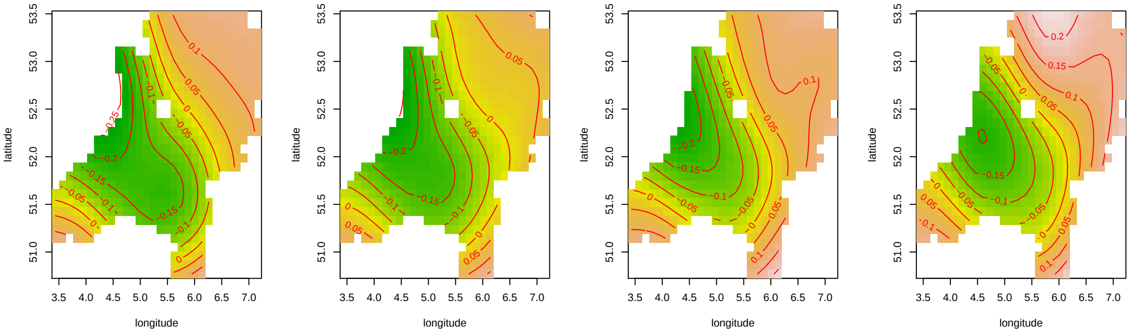

Figure 9. The pronunciation distance from standard Dutch for different quantiles of word frequency.

Generalized additive models offer the analyst a very powerful tool for understanding the structure of data sets in the language sciences. In the author's experience, models including nonlinear curves and surfaces often improve substantially over traditional models with linear effects and/or multiplicative interactions. Often, multiplicative interactions fail to detect the true but far more complex structure of the data. The results obtained with GAMs can be embarrassingly rich, in the sense that the results are far more complex than expected given current models. GAMs will often challenge the state of the art of current theories, and the author's intuition is that they may force the field to move more into the direction of dynamic systems approaches to language.

Model selection also becomes a more challenging process in the case of generalized additive modeling. Whereas for simple factorial designs it is still feasible to inspect the AIC for all possible models, this is no longer possible for GAMs. There are too many dimensions to explore, with too many options with respect to how many parameters should be invested in nonlinearities. Here, the only way to proceed is by hypothesis-driven model exploration. The three-way tensor for the Dutch dialects, for instance, was hypothesized on the basis of the theory of lexical diffusion. Higher-order interactions in theoretical hyperspace might be present (e.g., tensor products involving seven or eight predictors), but without theoretical insights to guide the analyst, the results, even if significant, would remain uninterpretable and hence not particularly helpful for the advancement of knowledge.

Classification

Thus far, we have considered numeric response variables. Response variables, however, can describe different classes of outcomes: alternative constructions, alternative affixation patterns, correct versus incorrect responses, whether an informant is a dialect speaker, near-synonyms, etc. For data sets with such response variables, the analyst may want to ascertain whether these classes are predictable from, and hence supported by, the other variables describing the properties of the individual data points. There are many different classification techniques available, here only a small subset is reviewed.

Logistic Regression

For binary response variables, i.e., variables that assume one of two values (success versus failures, correct versus incorrect responses, construction A versus construction B, etc.), an extension of the multiple regression approach known as logistic regression is often a good choice (see, e.g. T. Jaeger, 2008).

Binary response variables have the property that the variance depends on the mean. This property is easy to understand intuitively: When a success has a theoretical probability around 0.5, there will be enormous variability in the responses actually observed. But when the probability of a success is close to zero or close to one, the system will look like it is deterministic with only a little bit of leakage.

The property that the variance depends on the mean violates the fundamental assumption of the Gaussian framework of standard regression modeling, namely, that the errors are independent and identically distributed and follow a normal distribution. The solution offered by *logistic models* is to recast the dependent variable (that a novice to the field would want to cast as a proportion) in the form of a *logit*, the logarithm of the odds ratio:

$$\begin{aligned} \text{logit}(Y) &= \log\left(\frac{\text{successes}}{\text{failures}}\right) \\ &= \log\left(\frac{P}{1-P}\right), \end{aligned} \tag{12}$$

where P is the probability of success. This logit is modeled as a function of the other predictors,

$$\text{logit}(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots. \quad (13)$$

and the whole machinery of multiple regression, including mixed-effects models and generalized additive modeling, is now available to the analyst. Unlike for Gaussian models, however, there is no parameter for the error term, and errors (the difference between a predicted probability and the observed discrete outcome) are now referred to as deviances. Crucially, with logistic regression, it is the *probability* of a given class that is modelled.

Within linguistics, logistic models were pioneered by sociolinguistics under the name of variable rule analysis, see, e.g., Tagliamonte and Baayen (2012) and references cited there. As a working example, their data set on *was/were* variation is touched upon here.

The `york` data were collected to study the conditions under which *was* occurs in the spontaneous speech of inhabitants of York (UK) where the standard norm requires *were*, as in *There was still quite strong winds in these parts.* The response variable is *Form*, with levels *were* and *was.* Predictors are *Adjacency* (is the verb adjacent to its referent, with levels *adjacent* and *non-adjacent*), the informant's *Age*, and *Polarity* (*affirmative* versus *negative*). The logistic mixed-effects covariance model

```
log(was/were) ∼ Adjacency + Age*Polarity + s(Informant, bs="re")
```

is visualized in Figure 10. (In the symbolic formula of the S language, `Age*Polarity` specifies main effects for *Age* and *Polarity* as well as an interaction between these two predictors.) Figure 10 indicates that the probability of *was* is somewhat greater under non-adjacency. As indicated by the right panel, there is a substantial effect of *Age* in interaction with *Polarity.* In negative sentences, the younger informants almost categorically prefer *was* whereas the older informants prefer *were.* This effect is more muted in affirmative sentences. On the proportions scale, used in Figure 10, the effect of *Age* is non-linear. This non-linearity is due to the nonlinear nature of the transformation from logits to proportions. On the logit scale, the effect of *Age* is actually modeled (in this example) as linear.

Further examples of logistic modeling can be found in Janda et al. (2012), Bresnan et al. (2007) and F. Jaeger (2010). The latter two papers discuss more complex logistic regression models. Janda, Nesset, and Baayen (2010) discuss in detail the consequences of treatment dummy coding for the correlational random-effects structure for logistic regression models.

Polytomous Regression

Polytomous regression is a modeling option for data sets for which the response variable is discrete and has more than two levels. There are several strategies available for this kind of data. One option is to fit a series of binary logistic models contrasting one level with all the other levels, the *one versus rest* heuristic (Arppe, 2008, 2011). (For including a random-effect factor as predictor, see Faraway (2006) and Arppe (2011).) Multinomial models (Venables & Ripley, 2002; Højsgaard, Edwards, & Lauritzen, 2012) estimate the effects for all response classes simultaneously. In practice, the one-versus-rest heuristic yields results that are very similar to those of more complex methods, whereas the results tend to be more transparently interpretable. Table 12 presents a summary of the coefficients (on

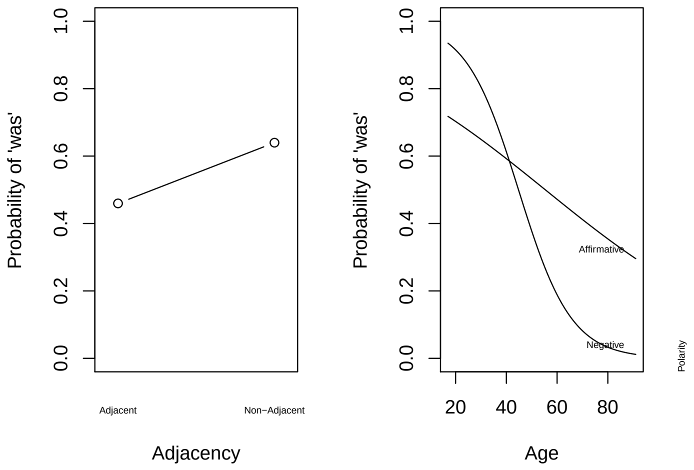

Figure 10. The probability of using *was* as a function of Age, Adjacency and Polarity.

	ajatella	harkita	miettia	pohtia
Intercept	0.76	-2.7	-2	-1.9
Agent=Group	-1.4	(0.38)	(-0.33)	0.83
Agent=Individual	(-0.066)	(-0.13)	0.69	-0.59
Patient=Abstraction	-1.6	(0.28)	0.58	1.6
Patient=Activity	-2.1	2.4	(-0.12)	0.89
Patient=Communication	-2.5	1.1	1.3	1.2
Patient=DirectQuote	-4.6	(-15)	0.8	2.9
Patient=etta.CLAUSE	0.72	-1.1	-0.61	(-0.41)
Patient=IndirectQuestion	-3	(-0.029)	1.7	1.5
Patient=IndividualGroup	0.72	(0.076)	-0.75	(-0.99)
Patient=Infinitive	1.7	(0.21)	(-15)	(-1.4)
Patient=Participle	1.5	(0.13)	(-15)	(-0.92)

Table 12: Log odds for four Finnish near-synonyms meaning *think*. Brackets mark non-significance.

the logit scale) for four Finnish near-synonyms for *think*, predicted from properties of the Agent and properties of the Patient. For completeness, we note that Arppe (2008) considers many more predictors for this lexical choice. Given the present limited set of predictors, Table 12 indicates that for patients expressing an activity, *ajatella* is dispreferred whereas *harikta* is strongly preferred.

Random forests

Regression models lose precision when, as is often the case for language data, the observations are distributed very unequally across the different predictor values. Regression models may also work less well when the data is characterized by complex interactions. In the case of the York data, for instance, there are very few instances of negative adjacent sentences, and half of the informants show no variability at all in their use of *was* versus *were*. Such very unequal and complex data may challenge the regression modeling framework.

For this kind of data, but also for data sets with relatively few observations and a great many predictors, conditional inference trees and random forests (Breiman, 2001; Strobl, Malley, & Tutz, 2009; Strobl, Boulesteix, Kneib, Augustin, & Zeileis, 2008), building on earlier work on classification and regression trees (Breiman, Friedman, Olshen, & Stone, 1984) are an excellent choice.

Conditional inference trees estimate a regression relationship by means of binary recursive partitioning. The `ctree` algorithm begins with testing the global null hypothesis of independence between any of the predictors and the response variable. The algorithm terminates if this hypothesis cannot be rejected. Otherwise, that predictor is selected that has the strongest association to the response, as measured by a p-value corresponding to a test for the partial null hypothesis of a single input variable and the response. A binary split in the selected input variable is carried out. These steps are recursively repeated until no further splits are supported.

Figure 11 presents a conditional inference tree for a Russian data set (Sokolova, Janda, & Lyashevksaya, 2012; Janda et al., 2012) that addresses the question of whether verb morphology (*Verb*, with as levels the prefixes *po-, na-, za-* and zero, i.e., no prefix) co-determines the choice between theme-object versus goal-object constructions. Further predictors are *Reduced* (is the construction reduced — levels *yes* versus *no*) and *Participle* (*yes*: passive participle, *no*: active form). The ovals in the recursive partitioning graph represent the choice points, and the p-value specifies the significance of the split. The branches are labeled with the class values governing the partitioned subsets. The thermometers at the leaf nodes present the proportion of goal constructions in black and the complementary proportions of the theme construction in light grey. The tree graph presents an easy to read summary of the structure of the data. The asymmetry of the tree, with different predictors appearing in the various branches, points to a complex interaction of *Verb* by *Reduced* by *Participle*.

The accuracy of recursive partitioning trees is often close to or comparable to that of regression models. However, a conditional inference tree locally optimizes the partitioning, which may have an adverse effect on its prediction accuracy. Random forests sidestep the limitations of a single locally optimal tree by constructing a large number of conditional inference trees, resulting in a (random) forest of conditional inference trees. Each tree in the forest is grown for a subset of the data generated by randomly sampling without replacement from observations and predictors. The predictions of the random forest are based on a voting scheme for the trees in the forest: Each tree in the forest provides a prediction about the most likely class membership, and the class receiving the majority of the votes is selected as the most probably outcome. Generally, the prediction accuracy of a random forest is greater than that of the locally optimal conditional inference tree, and highly competitive with the accuracy of logistic models.

Random forests also provide insight into the relative importance of the predictors by assessing the loss of prediction accuracy when the association between a predictor and the response variable is broken by randomly permuting the values of the predictor. The greater

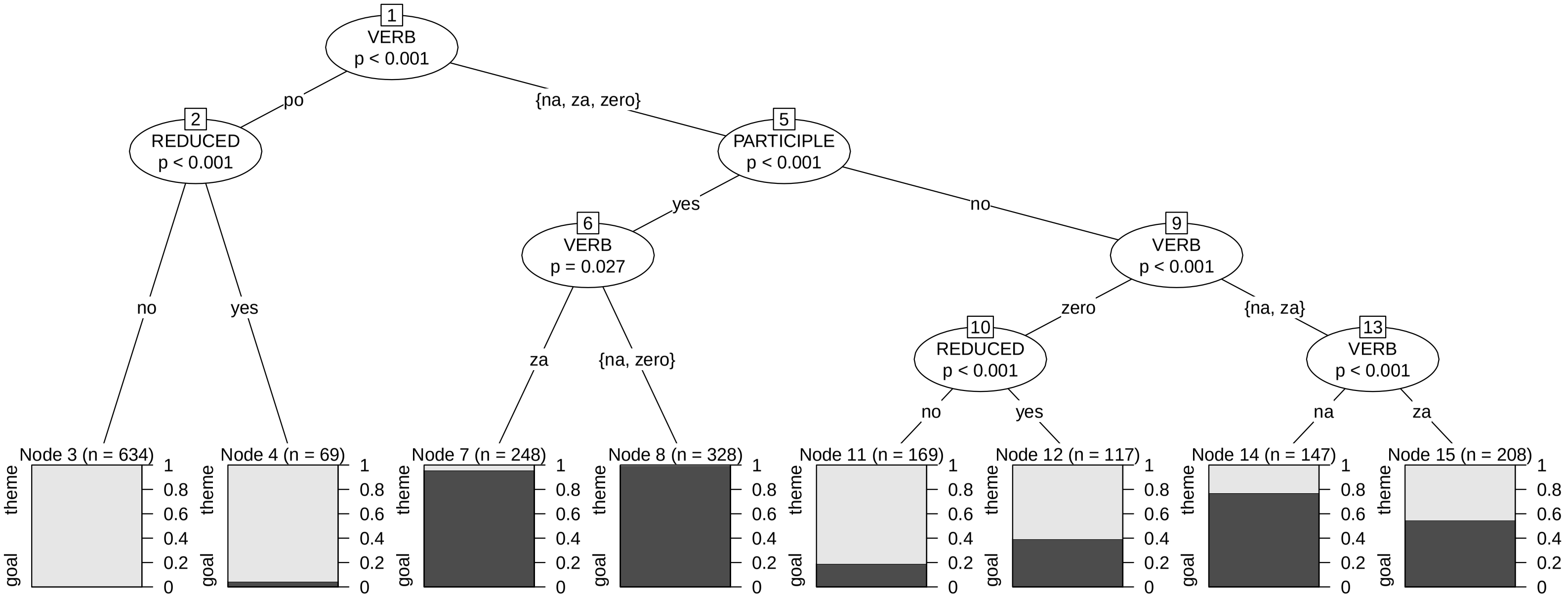

Figure 11. Recursive partitioning tree for the Russian goal/theme data.

the decrease in accuracy, the more important a predictor is. For the Russian data, the variable importance scores are 0.003 for *Reduced*, 0.076 for *Participle*, and 0.335 for *Verb*, indicating that the verb morphology is the most important predictor of the construction.

Recursive partitioning is less effective for data sets with random-effect factors. In the languages sciences, subject variability is often the strongest predictor for such data, and often one finds that the tree graphs split almost exclusively on the subjects. Furthermore, unfortunately, with large numbers of subjects and items, recursive partitioning becomes computationally prohibitive. However, when information about subjects and items is withheld, recursive partitioning trees may still provide useful information about interactions in the data that help the formulation of mixed-effects regression models.

Memory-based learning

Memory-based learning (Daelemans & Bosch, 2005), software available at `http://ilk.uvt.nl/timbl/`, is a technique that assigns a class to an observation based on the class membership of its nearest neighbors. Unsurprisingly, the accuracy of a nearest neighbor classifier depends on the definition of what constitutes a nearest neighbor. The simplest similarity metric counts the number of features that two exemplars share. (If a predictor is numeric, it has to be binned into a small number of factor levels.)

Sets of neighbors can be at various distances. Some neighbors may differ in only one predictor value, others may differ with respect to two values, etc. The set of neighbors taken into account can be restricted to the set of closest neighbors, but neighbor sets at larger distances can also be taken into account. Given a set of neighbors, an observation is assigned to that class that is best represented in this set of the nearest neighbors.

The similarity metric for neighbors can be refined in many ways. For instance, predictors (or features in the terminology of memory based learning) can be weighted for how informative they are about the response class across the data set, and further adjusted for the number of different levels of a predictor. This often result in a highly-effective classifier that is entirely competitive with the classifiers described in the preceding sections. Furthermore, memory-based learning scales up very well to large data sets and to data sets with predictors with many levels. From a theoretical perspective, memory-based learning is important because it is a computational implementation of exemplar theory, albeit only for discrete (or discretisized) data.

Examples of linguistic studies making use of memory-based learning in computational linguistics are found in (Daelemans & Bosch, 2005). Krott, Baayen, and Schreuder (2001) made use of memory-based learning to predict interfixes in Dutch compounds, Plag, Kunter, and Lappe (2007) applied it to the analysis of stress patterns in English compounds, whereas Keuleers et al. (2007) used it to study Dutch plural inflection. Keuleers (2008) provides a detailed comparison of memory-based learning with the rule-induction approach of Albright and Hayes (2003), focusing on regular and irregular verbs in English.

Naive discrimination learning

Naive discrimination learning implements a classifier based on principles of human learning as formalized in the Rescorla-Wagner equations (Wagner & Rescorla, 1972) and the equilibrium equations for the Rescorla-Wagner equations developed by Danks (2003). Currently, there is only one implementation of the naive discrimination learning, the `ndl` package (Arppe, Milin, Hendrix, & Baayen, 2011) for R (R Development Core Team, 2011). Several studies (Baayen, 2011; Janda et al., 2012) suggest that its classificatory accuracy is comparable to that of other state-of-the-art classifiers. It is mentioned here as model that offers a learning perspective on the probabilistic knowledge that speakers have of their language. For naive discrimination learning as a computational model of lexical processing, see Baayen, Milin, Filipovic Durdjevic, Hendrix, and Marelli (2011).

Table 13 lists the weights from predictor-value pairs (rows) to the four Finnish think verbs of Arppe (2008). Figure 12 shows the network layout, with darker shades of gray indicating stronger positive connections. Exactly mirroring the results with one-versus-rest polytomous regression (see Table 12), for patients expressing an activity, *ajatella* is dispreferred with a large negative weight, whereas *harkita* is favored with a strong positive weight. The total support for a given verb is obtained by adding the weights from all relevant predictor-value pairs. For instance, a patient expressing an activity and an agent expressing an individual give rise to maximal support for *harkita* (summed weights 0.42) followed at a distance by *miettia* (0.21), *ajatella* (0.20) and *pohtia* (0.17).

	ajatella	harkita	miettia	pohtia	Abbreviation
Agent=Group	0.23	0.13	0.07	0.37	AgnG
Agent=Individual	0.41	0.07	0.22	0.10	AgnI
Agent=None	0.42	0.08	0.11	0.18	AgnN
Patient=Abstraction	-0.12	0.01	0.11	0.22	PtntAb
Patient=Activity	-0.21	0.35	-0.00	0.07	PtntAc
Patient=Communication	-0.26	0.10	0.27	0.11	PtnC
Patient=DirectQuote	-0.39	-0.07	0.17	0.50	PtDQ
Patient=etta.CLAUSE	0.40	-0.05	-0.07	-0.06	P.CL
Patient=Event	0.28	-0.04	0.00	-0.03	PtnE
Patient=IndirectQuestion	-0.31	-0.01	0.37	0.17	PtIQ
Patient=IndividualGroup	0.39	-0.01	-0.08	-0.09	PtIG
Patient=Infinitive	0.51	0.00	-0.19	-0.10	PtnI
Patient=None	0.25	-0.01	0.01	-0.04	PtnN
Patient=Participle	0.49	-0.00	-0.18	-0.09	PtnP

Table 13: Naive discrimination learning weights for four Finnish near-synomyms for *think*.

Concluding remarks

This chapter has focused on multivariate regression and classification, both of which consider a response variable as functionally dependent on a set of predictors. A great many statistical methods have been developed for data sets for which there is no specific response variable, and for which the goal is to clarify how all variables pattern together. Introductions to methods for dealing with such data sets can be found in, for instance, Everitt (2005), Baayen (2008), and Højsgaard et al. (2012).

Statistics is a field in which progress is rapid. As a consequence, many new techniques

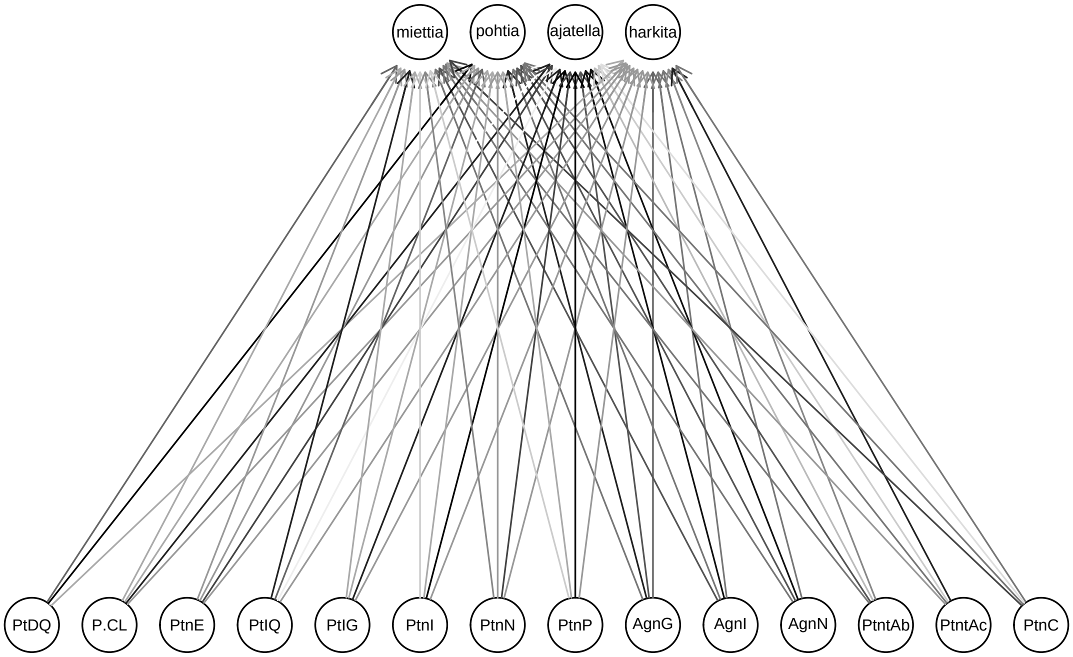

Figure 12. The NDL network for the Finnish *think* verbs. Darker shades of gray indicate stronger positive connections, lighter shades of gray larger negative connections. For the abbreviations in the nodes, see Table 13.

have become available in recent years (such as random forests and generalized additive mixed models) that considerably facilitate the analysis of language data. With the continued development of new statistical techniques that are increasingly well suited for the analysis of data from the complex dynamic systems that languages are, it will happen more often that analysts find themselves facing significant results that defy explanation within the conceptual framework within which a study was conceived. This, I believe, is good: Statistics will challenge linguistics to move beyond the boundaries of its current imagination.

References

Albright, A., & Hayes, B. (2003). Rules vs. analogy in English past tenses: A computational/experimental study. *Cognition*, *90*, 119–161.

Arppe, A. (2008). *Univariate, bivariate and multivariate methods in corpus-based lexicography. a study of synonymy.* Helsinki: University of Helsinki.

Arppe, A. (2011). polytomous: Polytomous logistic regression for fixed and mixed effects [Computer software manual]. Available from `http://CRAN.R-project.org/package=polytomous` (R package version 0.1.3)

Arppe, A., Milin, P., Hendrix, P., & Baayen, R. H. (2011). ndl: Naive discriminative learning [Computer software manual]. Available from `http://CRAN.R-project.org/package=ndl` (R package version 0.1.1)

Baayen, R. H. (2008). *Analyzing linguistic data: A practical introduction to statistics using R.* Cambridge, U.K.: Cambridge University Press.

Baayen, R. H. (2010). A real experiment is a factorial experiment? *The Mental Lexicon*, *5*(1), 149–157.

Baayen, R. H. (2011). Corpus linguistics and naive discriminative learning. *Brazilian Journal of Applied Linguistics*, *in press*.

Baayen, R. H., & Milin, P. (2010). Analyzing reaction times. *International Journal of Psychological Research*, *3*, 12–28.

Baayen, R. H., Milin, P., Filipovic Durdjevic, D., Hendrix, P., & Marelli, M. (2011). An amorphous model for morphological processing in visual comprehension based on naive discriminative learning. *Psychological Review*, *118*(3), 438-481.

Balota, D., Cortese, M., Sergent-Marshall, S., Spieler, D., & Yap, M. (2004). Visual word recognition for single-syllable words. *Journal of Experimental Psychology:General*, *133*, 283–316.

Box, G. E. P., & Cox, D. R. (1964). An analysis of transformations (with discussion). *Journal of the Royal Statistical Society B*, *26*, 211-252.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32.

Breiman, L., Friedman, J. H., Olshen, R., & Stone, C. J. (1984). *Classification and regression trees.* Belmont, California: Wadsworth International Group.

Bresnan, J., Cueni, A., Nikitina, T., & Baayen, R. H. (2007). Predicting the dative alternation. In G. Bouma, I. Kraemer, & J. Zwarts (Eds.), *Cognitive foundations of interpretation* (pp. 69–94). Amsterdam: Royal Netherlands Academy of Arts and Sciences.

Chatterjee, S., Hadi, A., & Price, B. (2000). *Regression analysis by example.* New York: John Wiley & Sons.

Cohen, J. (1983). The cost of dichotomization. *Applied Psychological Measurement*, *7*, 249–254.

Daelemans, W., & Bosch, A. Van den. (2005). *Memory-based language processing.* Cambridge: Cambridge University Press.

Danks, D. (2003). Equilibria of the Rescorla-Wagner model. *Journal of Mathematical Psychology*, *47*(2), 109–121.

Everitt, B. (2005). *An R and S-Plus companion to multivariate analysis.* London: Springer.

Faraway, J. J. (2006). *Extending linear models with R: Generalized linear, mixed effects and non-parametric regression models.* Boca Raton, FL: Chapman & Hall/CRC.

Friedman, L., & Wall, M. (2005). Graphical views of suppression and multicollinearity in multiple regression. *The American Statistician*, *59*, 127–136.

Gries, S. T. (2003). *Multifactorial analysis in corpus linguistics: a study of particle placement.* London & New York: Continuum Press.

Harrell, F. (2001). *Regression modeling strategies.* Berlin: Springer.

Hastie, T., & Tibshirani, R. (1990). *Generalized additive models.* London: Chapman & Hall.

Højsgaard, S., Edwards, D., & Lauritzen, S. (2012). *Graphicalmodels with R.* New York: Springer.

Hothorn, T., Bretz, F., & Westfall, P. (2008). Simultaneous inference in general parametric models. *Biometrical Journal*, *50*(3), 346–363.

Jaeger, F. (2010). Redundancy and reduction: Speakers manage syntactic information density. *Cognitive Psychology*, *61*(1), 23–62.

Jaeger, T. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, *59*(4), 434–446.

Janda, L. A., Nesset, T., & Baayen, R. (2010). Capturing correlational structure in Russian paradigms: a case study in logistic mixed-effects modeling. *Corpus linguistics and linguistic theory*, *6*(1), 29-48.

Janda, L. A., Nesset, T., Dickey, S., Endresen, A., Makarova, A., & Baayen, R. H. (2012). Making choices in slavic: Pros and cons of statistical methods for rival forms. *Russian Linguistics*, submitted.

Jurafsky, D., Bell, A., Gregory, M., & Raymond, W. (2001). Probabilistic relations between words: Evidence from reduction in lexical production. In J. Bybee & P. Hopper (Eds.), *Frequency and the emergence of linguistic structure* (p. 229-254). Amsterdam: Benjamins.

Keuleers, E. (2008). *Memory-based learning of inflectional morphology.* Antwerp: University of Antwerp.

Keuleers, E., Sandra, D., Daelemans, W., Gillis, S., Durieux, G., & Martens, E. (2007). Dutch plural inflection: The exception that proves the analogy. *Cognitive Psychology*, *54*, 283–318.

Koesling, K., Kunter, G., Baayen, R., & Plag, I. (2012). Prominence in triconstituent compounds: Pitch contours and linguistic theory. *Language and Speech*, in press.

Krott, A., Baayen, R. H., & Schreuder, R. (2001). Analogy in morphology: modeling the choice of linking morphemes in Dutch. *Linguistics*, *39*(1), 51–93.

Kryuchkova, T., Tucker, B. V., Wurm, L., & Baayen, R. H. (2012). Danger and usefulness in auditory lexical processing: evidence from electroencephalography. *Brain and Language*, *122*, 81–91.

Kuperman, V., Schreuder, R., Bertram, R., & Baayen, R. H. (2009). Reading of multimorphemic Dutch compounds: Towards a multiple route model of lexical processing. *Journal of Experimental Psychology: HPP*, *35*, 876–895.

Kuperman, V., & Van Dyke, J. (2011). Effects of individual differences in verbal skills on eye-movement patterns during sentence reading. *Journal of memory and language*, *65*(1), 42–73.

Lumley, T., & Miller, A. (2009). leaps: regression subset selection [Computer software manual]. Available from `http://CRAN.R-project.org/package=leaps` (R package version 2.9)

MacCallum, R., Zhang, S., Preacher, K., & Rucker, D. (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods*, *7*(1), 19–40.

Pinheiro, J. C., & Bates, D. M. (2000). *Mixed-effects models in S and S-PLUS.* New York: Springer.

Plag, I., Kunter, G., & Lappe, S. (2007). Testing hypotheses about compound stress assignment in english: a corpus-based investigation. *Corpus Linguistics and Linguistic Theory*, *3*, 199–232.

R Development Core Team. (2011). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Available from `http://www.R-project.org/` (ISBN 3-900051-07-0)

Sankoff, D. (1987). Variable rules. In U. Ammon, U. Dittmar, & K. J. Mattheier (Eds.), *Sociolinguistics: an international handbook of the science of language and society* (Vol. 1, p. 984 - 997). Berlin: De Gruyter.

Sokolova, S., Janda, L. A., & Lyashevksaya, O. (2012). The locative alternation and the russian 'empty' prefixes: A case study of the verb gruzit' 'load'. In D. Divjak & S. T. Gries (Eds.), *Frequency effects in language representation* (pp. 51–86). Berlin: Mouton de Gruyter.

Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T., & Zeileis, A. (2008). Conditional variable importance for random forests. *BMC Bioinformatics*, *9*. Available from `http://www.biomedcentral.com/1471-2105/9/307`

Strobl, C., Malley, J., & Tutz, G. (2009). An Introduction to Recursive Partitioning: Rationale, Application, and Characteristics of Classification and Regression Trees, Bagging, and Random Forests. *Psychological methods*, *14*(4), 26.

Tagliamonte, S., & Baayen, R. H. (2012). Models, forests and trees of york english: Was/were variation as a case study for statistical practice. *Language Variation and Change*, *24*, 135–178.

Venables, W. N., & Ripley, B. D. (2002). *Modern Applied Statistics with S-Plus.* New York: Springer.

Wagner, A., & Rescorla, R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning ii* (pp. 64–99). New York: Appleton-Century-Crofts.

Wang, W. S.-Y. (1969). Competing changes as a cause of residue. *Language*, *45*, 9–25.

West, B., Welch, K., & Galecki, A. (2007). *Linear Mixed Models: A Practical Guide Using Statistical Software.* London: Chapman & Hall/CRC Press.

Wieling, M. (n.d.). Voices dialectometry at the university of groningen. In C. Upton & B. Davies (Eds.), *Analysing 21st-century british english: Conceptual and methodological aspects of the bbc 'voices' project.* London: Routledge.

Wieling, M., Nerbonne, J., & Baayen, R. H. (2011). Quantitative social dialectology: Explaining linguistic variation geographically and socially. *PLoS ONE*, *6*(9), e23613.

Wood, S. (2006). *Generalized additive models.* New York: Chapman & Hall/CRC.

Wood, S. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society (B)*, *73*, 3–36.