

How to analyse electrophysiological responses to naturalistic language with time-resolved multiple regression

Jona Sassenhagen

To cite this article: Jona Sassenhagen (2018): How to analyse electrophysiological responses to naturalistic language with time-resolved multiple regression, Language, Cognition and Neuroscience, DOI: [10.1080/23273798.2018.1502458](https://doi.org/10.1080/23273798.2018.1502458)

To link to this article: <https://doi.org/10.1080/23273798.2018.1502458>



View supplementary material [↗](#)



Published online: 01 Aug 2018.



Submit your article to this journal [↗](#)



Article views: 86



View Crossmark data [↗](#)

REGULAR ARTICLE



How to analyse electrophysiological responses to naturalistic language with time-resolved multiple regression

Jona Sassenhagen 

Department of Psychology, Goethe University Frankfurt, Frankfurt, Germany

ABSTRACT

Naturalistic language processing cannot be approached with the analysis methods constructed to handle well-controlled experiments. Language is a multi- and cross-level phenomenon, with sequential interdependencies and correlations between various lexical dimensions. A recently-developed method allows the analysis of neural time series during natural story comprehension: time-resolved multiple regression. It consists in modelling continuous brain recordings with multiple regression after embedding linguistic features in a temporal-extension matrix (a distributed-lags model). It identifies neural correlates of linguistic processes, accounting for temporal interdependencies – simultaneously for, e.g. acoustics, phonology and semantics. This has resulted in impactful discoveries about how brains process coherent speech, potentially broadening the class of phenomena that can be studied. I discuss the method conceptually, highlight caveats, and relate it to similar as well as to traditional methods, all with a particular consideration for analysing the processing of coherent narratives. In a practical example, the word frequency-dependent N400 effect is estimated from a half-hour continuous narrative.

ARTICLE HISTORY

Received 28 February 2018
Accepted 13 June 2018

KEYWORDS

ERP; story comprehension;
EEG/MEG; rERP

1. Introduction

1.1. Factorial and model-free analyses in neurolinguistic research


While scientific interest in naturalistic language is growing, the complexity of the research object is high (Brennan, 2016). Why – and how – then investigate naturalistic language? Often, for studying neural correlates of sentence or text processing, sentences are employed as carriers of *critical positions*, and neural activity is investigated only in a short time window around these. For example, in Kutas and Hillyard (1980), a seminal study for Magneto-/Electroencephalography (MEG/EEG), specifically Event-related Potential/Field research (ERP/ERF), sentences were presented word-by-word, and from each sentence, one specific word that is semantically incongruent with the preceding part of the sentence is selected. Neural activity following many such incongruent words is averaged, and compared to similarly processed activity – surrounding individual words – from a control condition. By employing a *minimal contrast* between the manipulated and the control condition, it is – at least in principle – possible to isolate just the factor of interest, allowing causal inference.

Alternatively, researchers can manipulate either the context for a text or the text itself (Hasson, Yang, Vallines, Heeger, & Rubin, 2008; Lerner, Honey, Silbert, & Hasson,

2011; St George, Kutas, Martinez, & Sereno, 1999; e.g. St George, Mannes, & Hoffnann, 1994), usually comparing unmanipulated language to language that is manipulated, or put into an improper context. In such model-free analyses, neural activity over the whole text can be investigated, but must be aggregated in some form. But aggregation over the entire text rules out temporally specific claims. Minimal-contrast experiments allow high temporal precision, and in the optimal case, a minimal-contrast design picks out just one clearly delimited aspect of language processing – allowing, unlike observational analyses (such as the method described here), causal inference.

However, the minimal-contrast approach inherently carries a number of downsides – especially if a researcher is interested in naturalistic stimuli. First, much of the above work can be considered quasi-experimental in the sense that it is usually not possible to manipulate only the aspect of language that is the object of research, as would be required for fully experimental work (with random assignment to treatments). This is because in the case of language, stimulus materials cannot be constructed wholly anew, but must be selected from the pool of existing, e.g. words. Even if new stimuli are created, as in the case of pseudowords, they exist in a network of rich prior associations – i.e. their phonological neighbourhoods are inhabited by lexical items. An appropriate term for this

CONTACT Jona Sassenhagen  jona.sassenhagen@gmail.com

 Supplemental data for this article can be accessed at <https://doi.org/10.1080/23273798.2018.1502458>

© 2018 Informa UK Limited, trading as Taylor & Francis Group

class of experiments is thus quasi-experimental minimal-contrast designs/QEMCs. While this section lays out some drawbacks of the QEMC approach in the context of analysing neural correlates of continuous (spoken) language, the QEMC approach, pioneered by, e.g. Kutas and Hillyard (1980), has undeniably laid the foundation for the study of language in the brain, and will continue to play an important role. However, for the present scenario, they have further downsides.

To ensure that contexts are randomised besides for the factors of interest, QEMCs require specifically constructed stimuli. Therefore, they by necessity do not allow for naturalistic stimuli, or for investigating neural responses to arbitrary stimuli that are of inherent interest (for example, a specific, unique text). Second, they are highly inefficient: the analysed neural responses comprise only a small fraction of the brain activity recorded while subjects attended to stimulus presentation. Next, they are inefficient in that each experiment is specifically designed, ideally to test one specific theoretical hypothesis. Afterwards, the dataset is usually of little to no further use. Moreover, datasets tend to be relevant only in the context of a specific research framework/theory: what is a meaningful and informative manipulation in the context of one theory can be uninterpretable for another, or not conclusive on which framework best explains the phenomenon.

1.2. Specific problems in analysing continuous, naturalistic language

For any analysis method, investigating continuous narratives brings with it a specific set of challenges. Two

concerns are inherent to the nature – the complexity – of language itself: the syntagmatic problem; that words are immediately followed by other words, with statistical regularities regarding which words follow which; and the paradigmatic problem; that all linguistic entities – e.g. words or phonemes – differ on multiple dimensions (e.g. concreteness, corpus frequency), which in turn are correlated within words.

On the brain side, linguistic input – such as hearing a word – induces a cascade of brain activity, partially reflected in EEG, MEG and other measurements. Some neural responses induced by specific words last longer than the words itself; e.g. correlates of syntactic load can linger for multiple seconds (e.g. Fiebach, Schlesewsky, & Friederici, 2002). Consequently, neural activity observed during word W_n will consist of a mixing of late neural responses to words W_{n-1} , W_{n-2} , ..., neural responses to word W_n itself, and neural responses to following words W_{n+1} , ... The same goes for smaller (faster) and larger (slower) units of language, i.e. syllables and phonemes, syntactic phrases and prosodic units, and for reading and signing as well as for speech. Also, processing of each linguistic unit requires processes on multiple levels: e.g. phonological processing, lexical access, semantic and syntactic integration ... Thus, neural activity at each time point is the sum of neural correlates on multiple levels to multiple words, syllables, and phonemes (see Figure 1).

If there were no statistical patterns in sequences of words (or phonemes, ...), this would not pose a major problem: such unsystematic confounding activity would appear as unrelated noise, and would disappear under

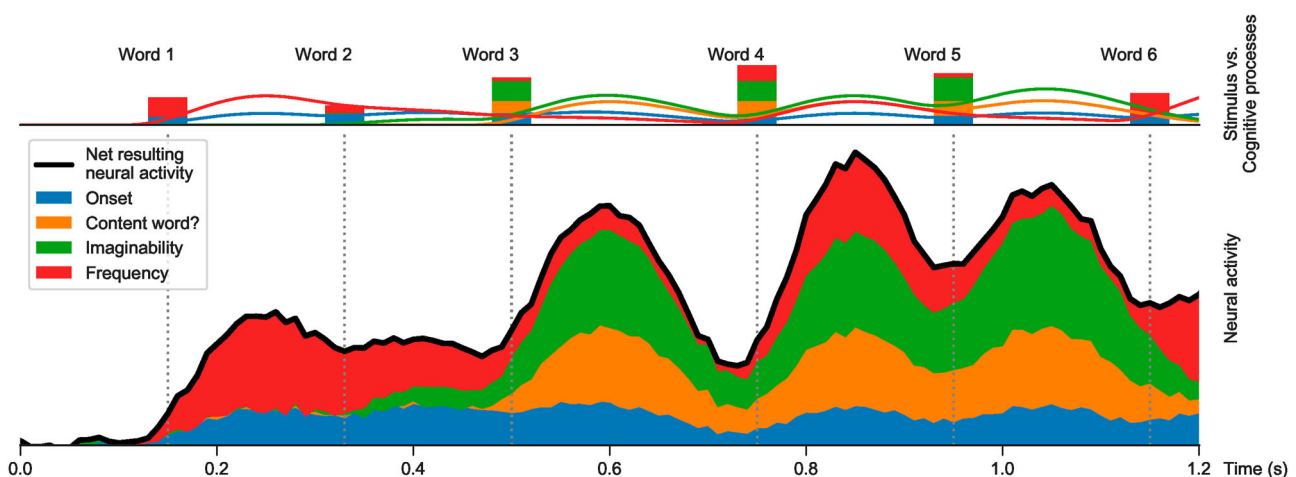


Figure 1. Visualisation of how, in a simulated example, distinct neural responses to specific linguistic features could sum up to the scalp-recorded signal. In the top row, stacked barplots show to what extent each word of an example sentence exhibits each included linguistic feature (colour-coded). Coloured line plots (see online version for colour reproduction), show individual, idealised neural response to each feature – stronger if the feature is expressed more strongly. The top row shows how the summation of these signals (coloured shapes), plus noise, sums up to the net observed signal (thick black line). Neural correlates of earlier events still influence the net signal at later positions.

averaging. Such orthogonality can be (partially) achieved with QEMC designs. But in linguistic material not specifically designed for QEMC purposes, it is not given: language is strongly characterised by statistical patterns in sequences of elements (see e.g. the Appendix for a demonstration of statistical patterns in ordinary narratives). An obvious phenomenon is the alternation of content and function words, but there are also more interesting, subtle, and more predictive patterns, e.g. basic word order regularities (with, e.g. – in Indo-European languages – nouns often following adjectives, but rarely the reverse; or, e.g. more “actor-like” words tending to precede, more “patient-like” words tending to follow verbs in nominative-accusative Subject-Verb-Object languages; Bornkessel-Schlesewsky & Schlewsky, 2009). This interacts with the paradigmatic problem: many syntagmatic regularities manifest across multiple correlated features. Thus, studying naturalistic language requires unmixing of multiple temporally overlapping responses.

Recently, time-resolved multiple regression has gained prominence (Dufau, Grainger, Midgley, & Holcomb, 2015; Frank, Otten, Galli, & Vigliocco, 2015; Gonçalves, Whelan, Foxe, & Lalor, 2014; see also Hauk, Davis, Ford, Pulvermüller, & Marslen-Wilson, 2006; Lalor & Foxe, 2010; Smith, 2012; Smith & Kutas, 2015b). This method addresses the above-mentioned concerns. It allows model-based investigation of arbitrary stimuli, including naturalistic texts. It considers the entire recorded segment, instead of only carefully matched, isolated positions. It facilitates sharing and re-using datasets, aggregating analyses over multiple texts. And it is inherently “Big Data”-capable: it can question a dataset for multiple research topics, multiple levels of linguistic representation, bringing with it the possibility of combining multiple researchers’ data to validate models.

2. Time-resolved regression

In the following, I briefly sketch the historical and mathematical background of the method, how it relates to other approaches, how it is conducted in practice, and important caveats to be considered when conducting time-resolved regression analyses.

2.1. The approach

2.1.1. Historical note

Time-resolved regression has been proposed independently multiple times. This brings with it a degree of terminological confusion. The first development of this statistical approach per se seems to have been the distributed-lags model in economics (Almon, 1965; Eisner, 1960) about 80 years after the initial

development of regression analysis (Galton, 1886). In language, *spectro-temporal response functions* (sTRFs) or *receptive fields* (Aertsen, Johannesma, & Hermes, 1980) were developed – a special case of time-resolved regression; a single feature – e.g. the speech envelope – is correlated with the neural signal at multiple lags. These models have laid the basis for encoder analysis of neural responses, influential in auditory processing (e.g. Holdgraf et al., 2017). Multivariate regression approaches to analysing lists of individual words – e.g. containing multiple predictors/linguistic features in one regression analysis, but not taking into account temporal overlap – were proposed by Hauk et al. (2006). Around 2010, Smith (2012; see also Smith & Kutas, 2015b) and Lalor and Foxe (2010) independently worked on the multivariate case, developing the theoretical background of the approach. Smith & Kutas proposed the name “regression-based Event-Related Potential” (rERP), Lalor & Foxe “multivariate Temporal Response Function” (mTRF). In addition to Lalor & Foxe’s work on audiovisual speech perception, continuous-time analyses of EEG data were extended to the visual domain (e.g. VanRullen & Macdonald, 2012). It is therefore not obvious which name should be given preference – Distributed Lags Models? rERP? mTRF? Encoder Models (see below)? Here, the term Time-resolved regression is preferred, as a neutral, descriptive and general term.

2.1.2. Procedure

The procedure consists in embedding a multivariate design matrix in a temporal delay matrix and aligning it with the observed neural time series. The original design matrix P contains word-level information (and/or lower- or higher-level information, e.g. syllabic or phonemic features). Each unit of interest, e.g. each word, corresponds to one row, each column to a feature of interest – including categorical predictors such as content vs. function word, and continuous predictors such as corpus frequency. For word list data, it suffices to estimate one regression model at each time point (Hauk et al., 2006; Smith & Kutas, 2015a). For the extension to continuous, overlapping speech, the delay of features is required.

For fMRI-experienced researchers, it can be helpful to understand the need for a temporal embedding as resulting from the lack of a canonical response function for MEG/EEG data. Instead, the temporal shape of the response has to be estimated from the data. The underlying mathematics are discussed in detail in the appendix, including considerations of specific aspects of the implementation. However, the basic formulation of the distributed-lags temporally resolved analysis

corresponds to equation 1 (note that in the case of language research, both positive and negative lags will be used to model a baseline period):

$$y = \beta_0 + \alpha_0 x_t + \alpha_1 x_{t-1} + \dots + \alpha_n x_{t-n} + \varepsilon_t \quad (1)$$

for positive and/or negative lags n . That is, to account for neural responses developing over time and for the problem of overlap, one regression coefficient is calculated for each lag by adding a time-shifted version of each predictor column. Expressed in matrix form, the full predictor matrix X is constructed by taking the columns of P and adding, for each column c in P , n time-lagged versions of c . That is, if the first column of P is $[0, 0, 1, 0, 0, 0]$, indicating that the event expressed in this column occurred first at the third sample. To investigate lags $-1, 1$ and 2 , one adds $c_{-1} = [0, 1, 0, 0, 0, 0]$, $c_1 = [0, 0, 0, 1, 0, 0]$ and $c_2 = [0, 0, 0, 0, 1, 0]$, leading to:

$$X = \begin{matrix} & c_{-1} & c & c_1 & c_2 \\ \begin{matrix} t1 \\ t2 \\ t3 \\ t4 \\ t5 \\ t6 \end{matrix} & \begin{pmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix} \end{matrix} \quad (2)$$

Each column now encodes if a specific event – such as a word onset – happened at a relative lag to the time point associated with the row. I.e. a column for a delay corresponding to 100 msec lag is 1 if and only if a word onset happened 100msec before, and otherwise zero. This means that the coefficient corresponding to this column will capture variance associated with the neural consequences of word onsets 100 msec after the fact. One can also encode continuous features in this manner, such as the predictability of each word; e.g. for a word with cloze probability 0.85, followed after 3 samples by a word with cloze probability 0.3:

$$X = \begin{matrix} & p_{-1} & p & p_1 & p_2 \\ \begin{matrix} t1 \\ t2 \\ t3 \\ t4 \\ t5 \\ t6 \\ t7 \\ \vdots \end{matrix} & \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0.85 & 0 & 0 & 0 \\ 0 & 0.85 & 0 & 0 \\ 0 & 0 & 0.85 & 0 \\ 0.3 & 0 & 0 & 0.85 \\ 0 & 0.3 & 0 & 0 \\ 0 & 0 & 0.3 & 0 \\ \vdots & \vdots & \vdots & \vdots \end{pmatrix} \end{matrix} \quad (3)$$

These sub-matrices – each corresponding to the temporal embedding of one feature – are concatenated, resulting in the multivariate design matrix X (see Figure 2, centre). X is aligned with the array containing the neural data Y (see Figure 2, right), and both are treated as a multiple

regression problem to calculate coefficient series B (see Figure 2, bottom), i.e. solving $Y \sim XB$ for B . For a detailed treatment, see the appendix.

Linear regression then partitions the variance in the neural signal Y amongst the predictors, in effect assuring that the intermixed signal shown in Figure 1 (centre) can be separated into the underlying components, without misattributing variance due to within- and between-word correlations; e.g. taking into account that function word-associated neural signals tend to reach into the time window where content word-associated neural signals unfold, and that words that occur rarely in corpora are also acquired later.

The resulting regression coefficients B are grouped by the basic, non-delayed features they correspond to, i.e. multiple lags of one basic feature are joined, corresponding to one single time series per basic feature each. This time series corresponds to the temporal impulse response function for the event described in the original design matrix column; i.e. the series of coefficients corresponding to word frequency indicates how the neural signal changes depending on the frequency of a word. If the coefficient at a lag of 400 msec is 3, this corresponds to brain activity following each word being larger than baseline by 3 units for each unit of word frequency. In the case of a categorical predictor, i.e. content words, it corresponds, roughly, to the neural consequences to content words, after having taken into account all other aspects covered by the model – e.g. word onset effects and word frequency. Thus, they are conceptually very close to difference wave ERP/ERFs. However, they do not require perfect matching of the two conditions (see Sassenhagen & Alday, 2016), only that a well-specified statistical model is employed (see below); and they allow for continuous predictors, instead of categorical distinctions.

The analysis is repeated for each sensor, so that the coefficient series of each analysis correspond to the specific response properties at each sensor. The results are a linear function of the data, which enables a range of conventional analyses just as one would with ordinary ERP/ERFs; in particular, baselining, source localisation and across-subjects parametric statistics.

2.2. Relation to other approaches

2.2.1. Regression vs. Averaging

As noted by Smith (2012), in the non-overlapping case and for categorical data, regression and averaging (i.e. ERP/ERF) result in precisely the same numerical results (roughly because, in these specific contexts, the mean value minimises the sum of squares). In many other contexts, the two differ. For example, averaging inherently cannot model parametric modulations; i.e. it cannot

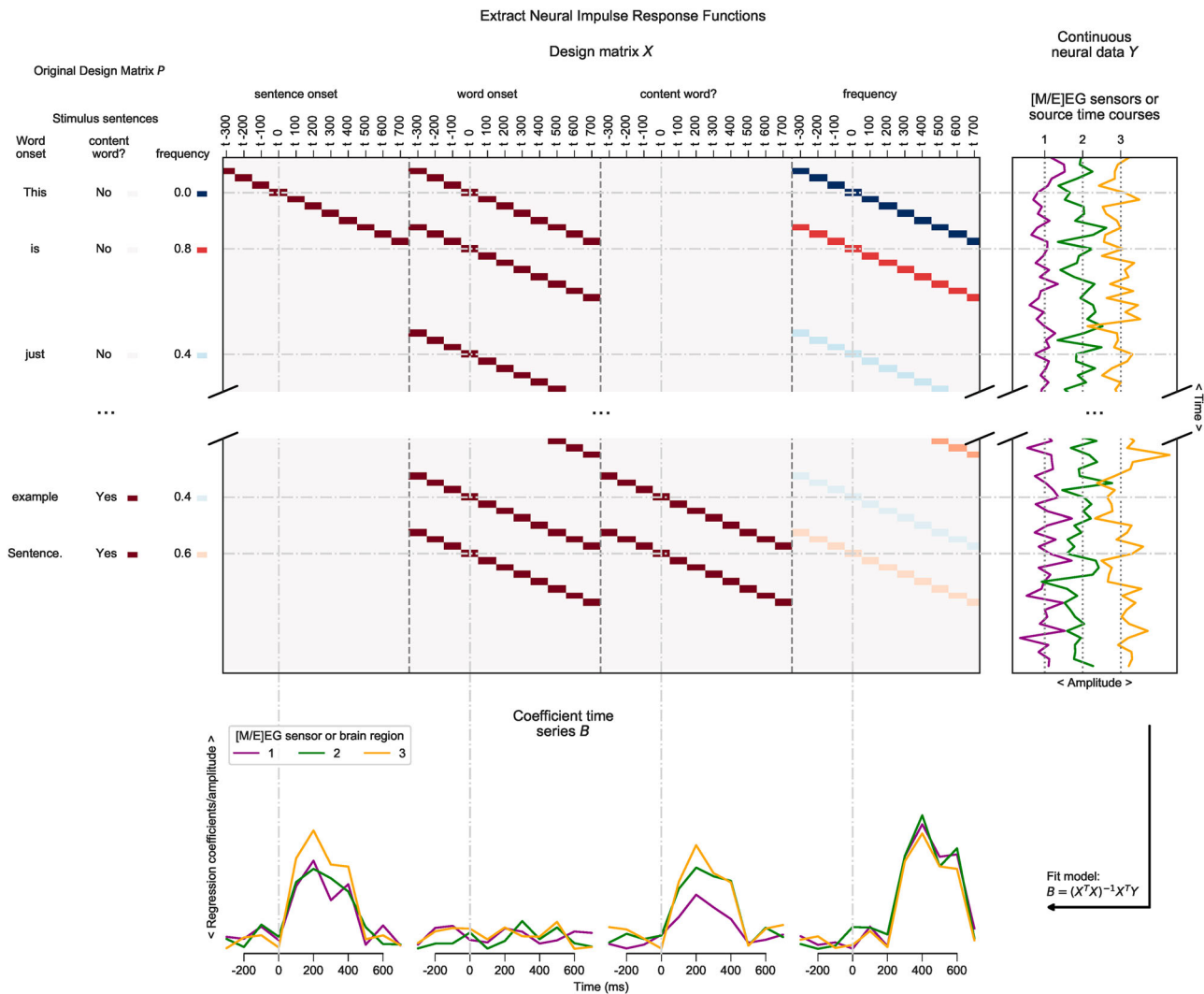


Figure 2. Extraction of Neural Impulse Response Functions. Simulated example data. Left: the original stimulus annotation. Here, an example sentence is annotated with regards to word onsets, content wordness and corpus frequency. The discontinuity in the centre indicates that (much) more material is included. Right: the observed neural data itself (Y), temporally aligned with the stimulus description. Each line reflects activity at one sensor, temporally aligned with the stimulus annotation. Centre: the stimulus annotation after alignment with the neural data, temporal embedding on a range of lags (top) and concatenation; resulting in the full design matrix X . That is, the submatrix corresponding to content word status indicates at each sample if a content word began within a certain time window. Bottom: resulting from the regression estimation procedure (bottom right), the coefficients B – one for each lag for each feature for each sensor – plotted as time series.

estimate a continuous function. Averaging also requires randomisation and perfect independence from all confounding factors to perform well; as shown here, this is not the case in continuous speech and texts. For example, averaging neural responses to all determiners in a dataset does not result in an unbiased estimate of neural responses to determiners. As determiners tend to be followed by nouns, neural responses to nouns will become hard to disentangle from determiner-induced activations. In regression-based analysis, such an unbiased estimate can be obtained: a predictor for nouns is added to a predictor for determiners; also, nouns should follow determiners either only some of

the time, and/or at sufficiently temporally variable intervals – which is the case in continuous speech, but requires adding randomly selected jitter between stimulus presentations in the case of visually presented/RSVP language stimuli. The efficacy of regression-based analysis for the purpose of identifying multiple neural processes under conditions of noise, overlap and collinearity can be demonstrated with a simulation.

Methods: For this, three datasets were constructed. In the first, white noise was generated and a simulated “ERP” impulse to word onsets was added; intervals between word onsets exceeded the length of the response, so that there was no overlap. Then, some of

the words were chosen to be mock “content” events, and a second response impulse was added on top of the onset response. For the second dataset, word onset and “content” effects were again added, but the time windows between words were shortened and randomised (inter-event durations were drawn from a normal distribution whose scale and spread corresponded to the inter-event interval length) so that there was a high degree of overlap between the simulated responses to multiple words. In the third set, a varying confounding impulse was added to each event with a probability that was dependent on both the time interval between the event and the preceding event, and if it was a “content” event. To be precise, on each event, the probability of the third impulse being added was: the duration since the last event, divided by twice the average duration; plus 0.5 if the event was a “content” event. Then, averaging and continuous-time regression estimation were applied to recover the original responses.

Results are shown in Figure 3, and indicate that in all these situations, time-resolved regression is capable of recovering the original impulses. In the non-overlapping case, it recovers the same signal as averaging and calculating a difference does. In the other cases, averaging and calculating difference waves fails to recover the original impulses, but regression *did* – with high accuracy – reconstruct the original impulses.

2.2.2. Encoding models

As noted above, the coefficients recovered by time-resolved regression can be understood as neural response functions – how does a stimulus influence brain activity? – and analysed much like, e.g. ERP/ERFs. However, the fitted model can also be used to predict neural responses given a stimulus (by multiplying the design matrix with the coefficients) – i.e. to *encode* the stimulus in brain activity. Encoding models (Holdgraf et al., 2017) are validated by how well their predictions match real brain activity. To ensure unbiasedness, this must be conducted in a cross-validation loop (Varoquaux et al., 2017) – e.g. coefficients are estimated on 2/3rds of the data, and a prediction is derived for the remaining 1/3rd (here, the crossvalidation should proceed at least block-wise, because adjacent samples in neural time series are highly correlated and thus, a shuffled crossvalidation would result in a high degree of overfitting). Then, the quality of the prediction can be evaluated, e.g. by calculating the correlation between predicted and observed data. This reduces the results of an analysis to few, or even a single number (e.g. R^2 or r per sensor). The resulting model fits are typically small – often around $r=0.1$ – because the signal to noise ratio of non-invasively measured neural time series is low and the model is set

out to explain long stretches of data, during which subjects engage in many cognitive and perceptual activities not included in the model.

This in turn allows contrasting multiple rich models. The specific problem this solves is that the temporal embedding procedure “inflates” the number of parameters. For example, if a researcher wishes to uncover if word frequency is an important predictor of neural responses, it would in principle be possible to conduct a model comparison where a measure of model fit, e.g. adjusted R^2 or a deviance-derived measure, is compared between a model including and one lacking the factor. However, the temporal embedding multiplies the number of entries in the design matrix by the number of lags, resulting in an effective difference of often hundreds of factors between two models, of which many (e.g. those modelling baseline and very late effects) are expected to make little difference. Encoding model predictions in brain activity, and checking the model fit, allows a direct comparison of much fewer numbers. Interestingly, the popular tool *Representational Similarity Analysis* (RSA; Kriegeskorte & Kievit, 2013) can be seen as a special case of an encoding model: one where both X and Y have been transformed into (dis)similarity matrices X'/Y' , and are related via a simple linear model. The similarity between the X' and Y' similarity matrices corresponds to the encoding model fit.

In sum, given the multiple regression problem $y \sim Xb$, having minimised the residual sums of squares/ $RSS = ||y - Xb||^2$, (1) time-resolved regression coefficients are the model coefficients b , (2) encoding typically concerns the quality of the fit (i.e. some function of the RSS), and (3) RSA can be seen as an investigation of the quality of fit for a specific, restricted class of encoding models: those where X and y are transformed into pairwise similarity matrices, and related to each other via simple (often univariate linear) models.

2.3. Practically speaking

In practice, conducting a time-resolved regression analysis requires a number of stages – the meaning of each of which will be discussed in greater detail below:

- Acquire data
- Choose and acquire predictors
- Build model and run regression
- Interpret
 - (i) resulting coefficients and/or
 - (ii) Encoding fit

Data acquisition can take two forms. On one hand, the flexibility of the method allows re-usage of preexisting

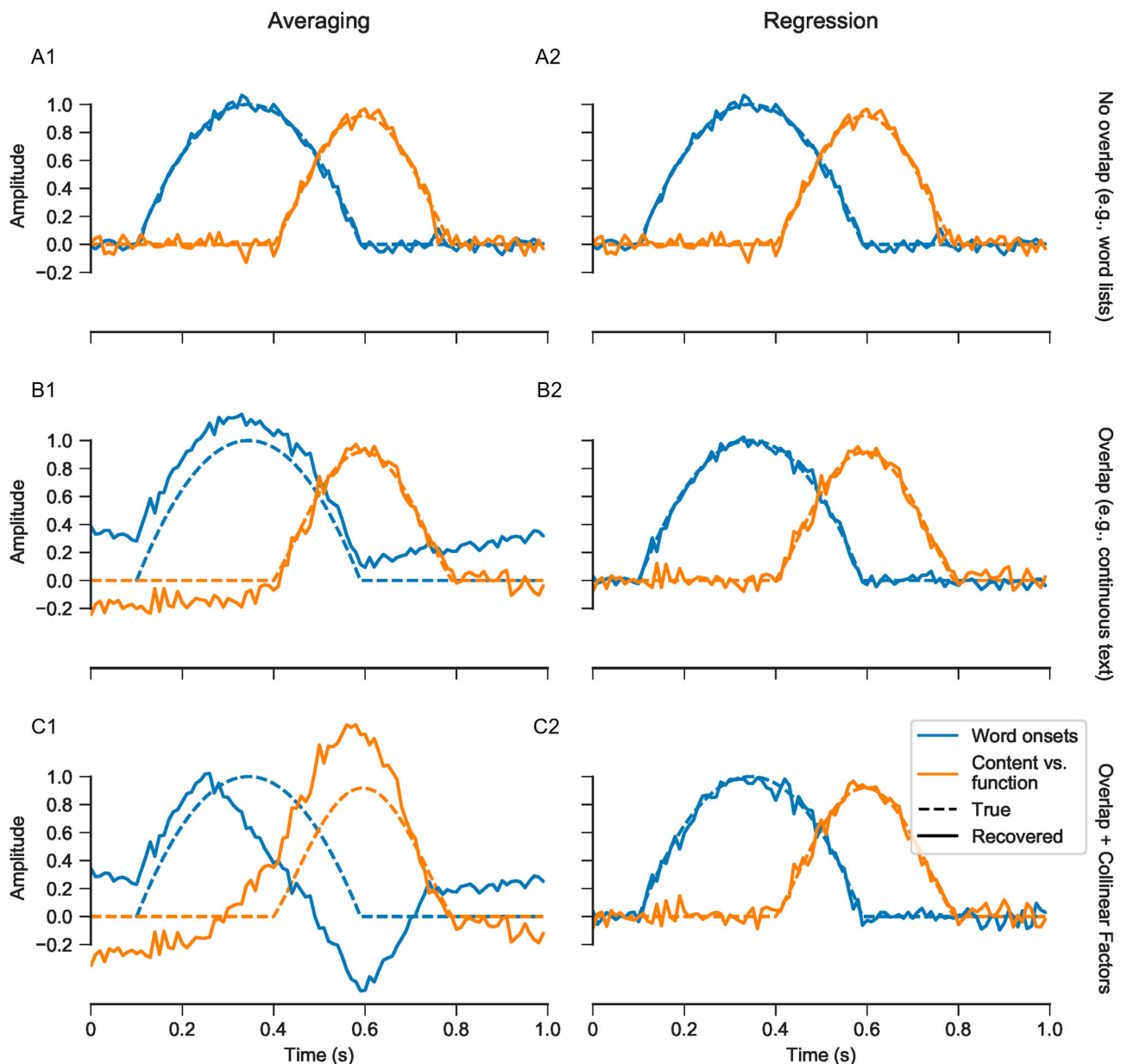


Figure 3. Simulation: recovery of impulses from continuously overlapping data. (A) two simulated “evoked responses” – “word onset” and “content word” – are added to 200 “trials” of continuous noise data, without overlap. (B): As (A), but responses overlap. (C): As (B), but a third response is added whose occurrence is modulated by the time interval between two events, and by the occurrence of the “content word” response. Left/1: Responses are recovered via averaging. That is, the “onset” response is estimated by averaging over all events; the “content” response is estimated by averaging all “content” events, and subtracting this signal from the “onset” estimate. Right/2: regression estimation with the two (A/B)/three (C) variables as predictors, and temporal embedding in a delay matrix to account for potential overlap. Resulting regression coefficients are plotted as time series. In the non-overlap case (A), averaging (A1) and regression (A2) lead to the same result: a high-quality recovery of the original responses. In the overlapping case (B), while regression again recovers the original signal (B2), the averaging result is strongly distorted (B1). In the overlapping and confounded case (C), the results of averaging are even worse (C1), but regression again recovers the true signal (C2).

datasets. On the other hand, given a sufficient language model, near-arbitrary linguistic stimuli can be presented to participants. In principle, if an unbiased estimator – i.e. linear regression, but not Ridge regression – is chosen, it is of little consequence if much data is recorded from few subjects, or few data

from many subjects, because the linear estimate will in the long run converge on the same value. However, because many predictors compete for variance, regression is a data-hungry procedure; at least half an hour of continuous narrative constitute a lower bound for many analyses.

Predictor acquisition requires deciding on what level to model the data; primarily, at continuously, a phonetic or acoustic level (i.e. considering the speech envelope); at a phonological/phonemic level or perhaps syllabic level, for example by entering phonological feature onsets (as in Di Liberto, O'Sullivan, & Lalor, 2015), and/or on a word level (see below, section *Example*). Regardless, it is strongly recommended to always include predictors for key events on all these levels: text onsets, sentence onsets, and word onsets, as well as text offsets (e.g. paragraph or sentence offsets, if followed by silence; see *Missing variables* below for what problems result otherwise). Generally, if it were advisable to match stimuli on a feature in a QEMC or other experimental paradigm, the feature should be included in the model. If word-level predictors are included, it is also recommended to include at least the predictors corresponding to the key variables known to reflect in neural activity, especially word frequency and word length. Otherwise, the corresponding variance will be misattributed; i.e. if word onsets are not accounted for, word onset-related activity will be attributed to whichever model features are best correlated with them in the final (temporally extended) model.

Building and solving the model itself can be done from scratch, but in practice consists of entering the chosen predictors, temporally aligned with the neural signal, into a software package (see below, section *Software*), which then constructs the temporally extended predictor matrix and estimates the coefficients. However, in principle, nothing particularly challenging is required: the predictor array can be constructed by, e.g. employing a Toeplitz or sparse diagonal matrix function, as can be found in most programming languages employed for data analysis; the solving itself consists of calling a linear regression routine. It is recommended to employ a closed-loop, analytic solver (see Appendix).

Interpreting the results proceeds either much like one would for an ERP/ERF analysis, including with regards to statistical analysis (e.g. see section *Example*); or model fit can be calculated (see previous section on *RSA and Encoding models*). Often, a particular coefficient set can be understood much like a difference ERP; e.g. a coefficient set for word frequency corresponds conceptually in many ways to an isolation of a word frequency effect by contrasting very frequent and very infrequent effects. However, a number of caveats must be considered.

2.4. Caveats

Generally, an understanding of the conceptual and mathematical context of multiple regression analysis is

helpful for understanding and applying time-resolved regression, especially as there are multiple cases where regression results in incorrect or easily misinterpreted results. Overviews of regression are abundant (for a recent example, see Gruber, 2013).

2.4.1. “Controlling for” confounding and non-independence with multiple regression

As noted, the time-resolved regression approach attempts to account for problems of syntagmatic and paradigmatic dependencies with multiple regression. As a reminder, multiple regression analysis can attribute the independent impact of multiple predictors X on outcome measures Y ; i.e. in this case, how neural activity Y (e.g. MEG/EEG data) is predictable by a linear sum of various linguistic (and extra-linguistic) factors quantified in X . Linear regression partitions Y into unexplained gaussian noise and the respective correlates of each predictor independently. In the optimal case, this procedure can indeed uncover the influence of each feature contained in X . However, multiple regression runs into a number of problems when (1) the components of X are not independent of each other – when the design matrix is *collinear*, (2) when predictors are not free of measurement error, (3) when there are missing predictors.

Collinearity and variance inflation The first is that the uncertainty in the estimate increases proportionally to the degree of collinearity (see Smith, 2012; Smith & Kutas, 2015b for an extensive discussion of the problem in this context). More technically, the variance of the estimated coefficients increases. If the design matrix is correctly constructed, for the purposes of estimating neural responses, this aspect is fundamentally a data problem; it can be solved – and can *only* be solved – by adding more data points.

However, some collinearity issues stem from incorrect choices when constructing the initial design matrix. When one, two, or more columns can be linearly combined so that they are redundant with another, perfect (*multi*-)collinearity results, and the model becomes impossible to estimate (because the relevant matrix becomes singular). For example, accidentally adding the same column twice; adding two highly correlated measures of the same variable, such as two measures of word frequency; adding a “content word” column, a “function word” column, and a “word onset” column (content plus function words equals essentially all words); adding a column that is *nearly* a non-zero constant, such as improperly binned corpus frequency measures (again highly correlated with word onsets); adding a column that is zero at all time points. In these cases, although often, strong regularisation can still result in estimated coefficients, the more prudent choice is to respecify the model – i.e. choosing

a different subset of columns to include (e.g. include only word onsets and a content word column, which together also exhaustively describe the words on this group of characteristics) – while avoiding the *missing variable* problem (see below).

Measurement error A second problem arises to the extent to which there is measurement error in X , or otherwise a column in X does not capture the actual characteristic of interest, but only approximates it (Westfall & Yarkoni, 2016). For example, while it is well known that neural responses to words are correlated with word frequencies (Kutas & Federmeier, 2011), no single individual will have had the exact exposure to words as are contained in the corpus; that means a corpus measure of word frequency only approximates any individual's exposure. On one hand, these measurement errors attenuate the estimates; i.e. they lead to an *underestimation* of neural responses. On the other hand, they can prevent precisely attributing neural activity to any specific of a given collection of aspects of language. For example, if in a stimulus set, age of acquisition and corpus frequency are two (highly correlated) factors that only approximate the quantity the human brain is truly sensitive to (e.g. probability, or familiarity), then multiple regression will not reliably recover the specific neural consequences of the two factors. This particularly complicates employing regression for the validation of constructs (Westfall & Yarkoni, 2016). The problem is especially important if the correlation between the factors has a *causal* origin; e.g. if it were the case that less frequent words have higher ages of acquisition *because they are less frequent*, then it can become futile to “control for frequency” when estimating the effects of age of acquisition by including both in a regression problem (Miller & Chapman, 2001).

Missing variables A related problem is that multiple regression requires that the model is well specified in the sense that *all* predictors which are reflected in the outcome are accounted for in the design matrix. This is highly unlikely to become possible in naturalistic language. As a consequence, some of the variance corresponding to factors not included in the design matrix – because they are not known, or were excluded for other reasons – will be misattributed to those that are included. This issue becomes a source of systematic errors if the excluded predictors are correlated with those that are included, in which case left-over variance will be misattributed in systematically misleading ways. For example, if a model includes a novel measure of semantic density, but not word frequency, and the novel measure is correlated with word frequency, then truly frequency-associated neural activity will show up as being dependent on semantic density. For an even more trivial example, consider a predictor for word

onsets. Especially in reading, word onsets, or fixations – as discontinuous events – trigger a cascade of neural activity. This neural activity reflects in the neural signal as variance that regression will attempt to associate to the predictors. If there are *any* word-level predictors included, they will by necessity be correlated with word onsets (e.g. because most positions – often all that are not word onsets – are zero). Thus, if no word onset predictor is included, regression will attribute the onset/fixation-locked activity to these other predictors – in practice, in the case of ordinary least-squares regression, to those that have the least variance. Thus, it is crucial to add predictors for those events which induce large neural signals and have a somewhat similar temporal structure as words, e.g. word onsets, speech onsets, and speech offsets.

A special case of this assumption is that an appropriate number of lags must be chosen. For example, if word frequency results in neural consequences up to 2 s after word onsets, but only lags up to 1 s are included, then the remaining variance of late word frequency effects will partially be misattributed to following words.

However, in the cases where there is no causal relationship, where the predictors are measured with negligible measurement error (e.g. when the predictors are well specified), where all major predictors that are correlated with linguistic predictors have been included, and where the collinearity-induced variance inflation problem has been ameliorated by collecting sufficient data, multiple regression is indeed capable of solving the task of identifying and separating out the consequences of a large number of characteristics of language from a multiplexed signal.

To visualise these noted potential issues, another simulation was conducted. **Methods** Data was simulated as in the data underlying Figure 3. However, in three additional scenarios, measurement error, collinearity and missing variables were modelled. To simulate measurement error, random (normally distributed, $SD=1$) noise was added to the non-zero values of one predictor before temporal embedding. To simulate collinearity, the third predictor was constructed by adding random noise to the second predictor, leading to a correlation of $r>0.99$. To simulate a missing variable, taking these correlated predictors, the regression was conducted while omitting the third variable. **Results** In contrast to the estimation of the simulated neural impulses under optimal conditions (Figure 4(A)), the estimated response for the feature measured under noise was attenuated, i.e. biased towards zero (Figure 4(B2)); however, the general shape was not distorted. No bias or *systematic* distortion can be observed for highly collinear variables (Figure 4(C2)); however, the variance of the estimates is very high, i.e. the estimated impulse

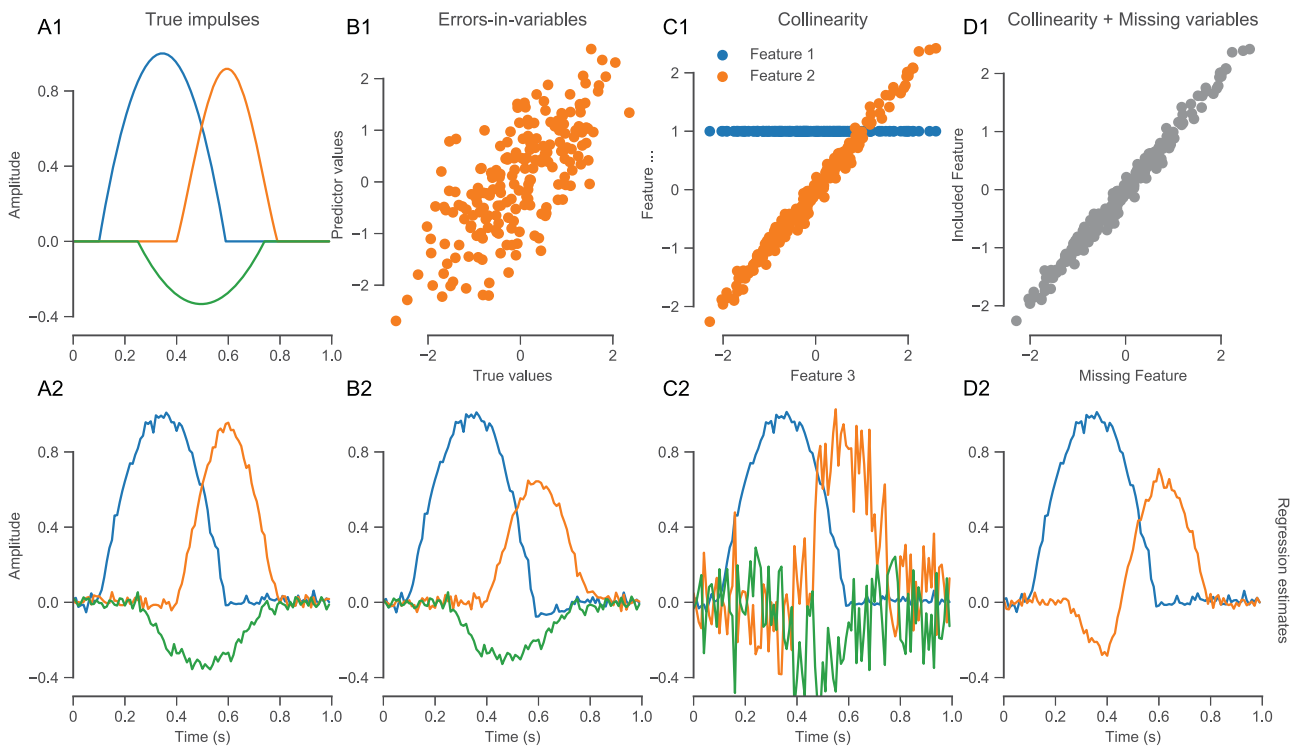


Figure 4. Simulation: recovery of impulses under adverse circumstances. Top: illustration of problem. Bottom: regression estimate of impulses. (A): as in Figure 3(C2), but the third feature is also shown. (B): The predictor contains errors. (B1) shows the (imperfect) correlation between the true value (i.e. the true familiarity or unconditional probability) on the x-axis and the value that is entered into the regression model (i.e. corpus frequencies) on the y-axis. (B2) shows that the estimate of the effect (positive peak around 600 msec) is not distorted in shape, but attenuated. (C): high collinearity, i.e. two features present in the model are highly correlated (C1). The regression estimates for the correlated effects are not systematically biased, but show high variance (noisy lines in (C2)). (D): as in (C), but one of the two correlated features is not included in the model. The estimate of the remaining feature is strongly distorted (note biphasic negative-positive component in (D2)).

shows a large degree of noise. Finally, a gross distortion of the estimate occurs if a factor is excluded that is correlated to one that is included (Figure 4(D)).

2.4.2. Assumptions of linear regression, and alternatives

In principle, linear regression makes a number of further assumptions, in particular that the residuals are linear, independent, normally-distributed and homoskedastic. In practice, this class of assumptions is of comparatively minor importance in this context – either (1) because they affect only direct inference on the regression (i.e. confidence bands on individual coefficients), whereas in time-resolved regression, inference is preferably handled across subjects, or with nonparametric statistics, (2) because of the typically massive number of data points. On the other hand, many of the tools typically employed for diagnosing problematic regressions cannot be used in this context because the number of estimated coefficients is much too large (often many hundreds or thousands, multiplied by the number of channels). For example, it is typically infeasible to inspect a plot of the residuals, because that would

entail thousands of residual plots (one per electrode, time lag and condition).

While it is not an assumption, by its form, linear regression specifies a linear relationship between predictors and outcomes; that is, the effect of a given shift in a predictor is independent of the value of the predictor. In the case of analysing neural correlates of naturalistic language, this is only guaranteed for binary predictors (i.e. word onset, content wordness, noun vs. not noun, voiced vs. unvoiced). Continuous predictors could in principle be expected to have any form of impact, and linearity is extremely unlikely (as it would imply that, e.g. in the case of a non-zero coefficient for word frequency, for an arbitrarily rare word, an arbitrarily large neural response would be elicited, which is physiologically impossible). Most likely, many relationships between psycholinguistic variables and their neural consequences are nonlinear – although there are known cases where a relationship is very close to linear in the relevant space, e.g. cloze probability and the N400 (Kutas & Federmeier, 2011).

It is not straight-forward to assess the specific nature of the dependence. One possibility is to construct dummy

predictors for quantile bins, e.g. the quartiles or deciles, and investigating the resulting coefficients for each quantile bin (see below in the Example). If an estimate for the nature of the relationship has been made (on an independent dataset), transformations can be applied to the predictors – e.g. squaring the values, or taking their logarithm. However, when values cluster in a narrow range – i.e. if they follow e.g. a Zipfian distribution, or when they are log scaled –, the corresponding column can be highly collinear with a word onset predictor.

Nonlinear models could in principle estimate the shape of a nonlinear effect. For this, a range of tools for nonlinear estimation is available, from Kernel Ridge Regression to Random Forests (Hastie, Tibshirani, & Friedman, 2009). However, in practice, these methods largely do not scale well enough to large problem sizes as would be required for time-resolved regression. The *Example* below presents one way of estimating the shape of the relationship between a continuous predictor and the strength of the pattern by relying on extensive dummy-coding and cross-validated pattern matching.

A similar problem concerns advanced feature selection techniques such as the LASSO, which cannot be fit with one-pass closed-form solutions like least-squares estimation for a multi-columnar Y . An exception is Ridge Regression (notably, a *linear* model that does not perform, but only assist in feature selection). Ridge regression (Hoerl & Kennard, 1970) can help in fitting badly-conditioned problems by adding a small regularisation term for details and alternatives, see the supplement – although a failing Ordinary Least Squares usually indicates a problem in the model specification that should be fixed, not simply resolved by adding additional constraints. An in-depth discussion is provided in the appendix. For an overview of various regularised approaches, see Wong et al. (2018).

Mixed models with item effects have recently become common in language research (Baayen, Davidson, & Bates, 2008). The welcome tendency to improve ecological validity however is not readily integrated with time-resolved regression on a word level. The temporal embedding multiplies the number of columns in the design matrix by a large factor. This means that for now the established techniques for estimating mixed models typically cannot be leveraged for time-resolved models with overlap correction, because the resulting models would require too much computer memory. The reduced (non-overlap correcting) approach can however readily integrate mixed models.

2.4.3. Artifactual data

In averaging approaches, epochs with artefacts can be discarded. But when dealing with continuous data, the

problem of artefact contamination is much less straight-forward. In epoched data, it is only required to make a categorical decision on each epoch: reject, or keep. In continuous data, each artefact requires a decision on what exact data segment to discard. In principle, robust regression methods could ameliorate this issue by reducing the dependence of regression coefficients on outliers (the *leverage* of these data points). In practice, robust regression methods do not scale to the size of the problem (see the appendix for an introduction of the technical problem). Thus, (1) researchers must select a criterion for marking contaminated stretches of data; (2) artefact *correction*, over artefact rejection, becomes imperative. For this reason, it is strongly recommended to employ e.g. independent component analysis (Jung et al., 2000) to attenuate eye artefacts. Artifactual segments may still be rejected, but need to be taken into account when constructing the design matrix: to correctly model overlap, rejection has to be performed after the temporal embedding.

2.4.4. Summary of caveats

In sum, when conducting time-resolved regression analyses, it is recommended to

- respect that even the best-controlled time-resolved regression analysis is a correlational, not a truly experimental, analysis
- minimise and understand errors in the model
 - (i) minimise collinearity
 - (ii) do not overinterpret results when collinearity is present
 - (iii) do not leave out predictors corresponding to factors that affect the signal
 - (a) ... especially not when left-out factors are correlated with factors that are present
 - (b) always add word onsets, sentence and text onsets and offsets to the model
 - (iv) understand the implications of nonlinearities and models without item effects
- perform artefact correction

3. Impact

3.1. Previous results

An impressive and representative demonstration of the capabilities of the mTRF and encoding approaches is provided by Di Liberto et al. (2015). Di Liberto and colleagues engage the question if there is a categorical representation of entities motivated by linguistic theory in brain activity. They conduct a phoneme-level analysis, coding for either individual phonemes, or phonological

features, and analyse 128-channel EEG data from a natural story listening study. Investigating the model coefficients, they illustrate the brain responses elicited by individual phonemes or features. Then, they encode these models, investigate their fit to observed brain data, and compare the encoding results to those from the encoding of the speech envelope (Ghitza, Giraud, & Poeppel, 2013) as a continuous feature. They find that models containing phonological features improve model fit substantially (obtaining correlations of $r > 0.1$), over and beyond the speech envelope. These results are taken to indicate a partially categorical representation of phonemes in brain correlates of the processing of spoken narratives. From the same group, Broderick, Anderson, Di Liberto, Crosse, and Lalor (2017) have employed mTRF estimation to track a neural signature of semantic similarity in coherent narratives – demonstrating the applicability of the method from phonemic to semantic level.

3.2. Available Software

The growing interest in time-resolved regression analysis of [M/E]EG data is reflected in a number of open-source toolkits including or focusing on it. A dedicated MATLAB toolbox has been developed for distributed-lags models of EEG data by Crosse, Di Liberto, Bednar, and Lalor (2016). Crosse et al.'s multivariate Temporal Response Function toolbox is particularly attractive for researchers already employing MATLAB-based analysis tools, such as Fieldtrip (Oostenveld, Fries, Maris, & Schoffelen, 2011) and EEGLAB (Delorme & Makeig, 2004). A direct EEGLAB-based suite has also been contributed (Burns, Bigdely-Shamlo, Smith, Kreutz-Delgado, & Makeig, 2013). A stand-alone Python module has been developed by Nathaniel J. Smith – available at github.com/njsmith/pyrerp –, but while fully functional, is no longer actively maintained. The MNE-Python suite (Gramfort et al., 2013) presents three different implementations of time-resolved regression (see Table 1) – in a highly flexible and performant implementation that integrates well with the general machine learning toolbox scikit-learn (Pedregosa et al., 2011). Moreover, the basic principles of time-resolved regression are straight-forward, and if a solver is available and some support for diagonal matrices is present, can be conducted in ~20 lines of code.

4. Example

Having discussed the potential and possible drawbacks of the method, to illustrate the approach, data from a half-hour naturalistic story listening paradigm are analysed. The (well-established) dependence of EEG amplitude on word frequency is recovered.

4.1. Methods

4.1.1. Stimulus presentation and data collection

Twenty subjects (14 female, mean age 24.1, range 19–34 years) participated in the study. All were right-handed, and native German speakers. None of the participants reported any hearing or visual deficits, relevant medical or psychiatric illnesses. Participants gave written informed consent prior to taking part in the experiment. The study was approved by the local ethics committee.

After application of the EEG set (32-channel Brainproducts BrainAmp), participants watched a video of a professional actor telling a story. The story text was adapted from the short story “Der Kuli Kimgun” by Max Dauthendey (Alday, Schlesewsky, & Bornkessel-Schlesewsky, 2017; Nagels et al., 2013; for previous analyses of experimental data employing this stimulus set, see, e.g. Whitney et al., 2009). The actor gesticulated freely to enrich the narrative. Some foreign or highly infrequent words had been replaced with more familiar words. The story was 3582 words long, corresponding to 31:12 min. It was presented in 16 segments of approximately 2 min length each, played consecutively and in order. Participants were asked to passively attend to the narrative.

The recorded datasets were processed in MNE-Python (Gramfort et al., 2013) and scikit-learn (Pedregosa et al., 2011). Analysis code is (together with code for all simulations) made available in the accompanying github repository at github.com/jona-sassenhagen/rep_review.

4.1.2. Preprocessing

For each dataset, data was downsampled to 100 Hz, re-referenced to average reference, and bandpass filtered between 0.1 and 12 Hz. ICA (Jung et al., 2000) was used to correct eye artefacts; after FastICA decomposition (Hyvarinen, 1999) of the continuous EEG activity, EOG-associated independent components were identified with a semi-automatic template-based algorithm (Viola et al., 2009) implemented in MNE-Python, and removed.

Table 1. Time-resolved regression in MNE-Python.

Function	Example	Purpose
<code>mne.stats.regression.linear_regression</code>	https://bit.ly/2MMYFhz	rERP
<code>mne.stats.regression.linear_regression_raw</code>	https://bit.ly/2Kt6MhP	rERP w/ overlap
<code>mne.decoding.ReceptiveField</code>	https://bit.ly/2lQchS0	Encoding
<code>mne.decoding.TimeDelayingRidge</code>	–	General framework

4.1.3. Analysis

Time-resolved regression analysis was conducted via the `linear_regression_raw` function in MNE-Python. To keep the analysis simple, a very abbreviated model was selected. The following values were added to the model: dummy predictors for word onsets, sentence onsets, segment onsets, segment offsets, sentence offsets, and as continuous covariates: word lengths, sentence positions (i.e. for the 4th word in a sentence, a 4), and segment positions (i.e. for the 17th word in the 3rd segment, a 17). Finally, word frequency was added. In one model, word frequency was included as the corpus rank bin count from the Projekt Deutscher Wortschatz (Goldhahn, Eckart, & Quasthoff, 2012), with higher ranks (1–24) indicating less frequent words. Words not found in the corpus were coded as 2 + the highest rank. In another model, instead of a continuous predictor, dummy predictors were added for each word frequency bin, resulting in 25 categorical predictors. All these predictors were coded for each word, resulting in design matrices with 3582 rows and 9 columns for the first model, and 34 columns for the second. All features were scaled to the 0–1 range. In a third model, referential status was modelled (see Appendix).

Word features were temporally aligned with the recorded EEG signal, and embedded in a distributed-lags temporal embedding matrix, with lags for each word ranging from 1 s before to 2 s, after each event, leading to a design matrix of $(3s * 100 \text{ Hz} * \text{columns}) * (1932s * 100 \text{ Hz}) = \sim 500,000,000$ to $2,000,000,000$ entries for each dataset after temporal embedding, with a highly sparse structure (most entries being zero). For solving the linear system, a Ridge regression from scikit-learn was selected, a Cholesky solver was chosen and the Ridge coefficient set to 1. The Cholesky algorithm was selected because it is highly performant, in particular with sparse predictor arrays. Nooptimisation of the Ridge procedure was chosen because optimal prediction performance was not the goal, and because reasonable defaults often out-perform hyperparameter tuning in neuroimaging contexts (Varoquaux et al., 2017).

Evoked responses were estimated by fitting the regression model to the neural data and grouping the resulting coefficient time series by predictor, resulting in 9/34 event-related regression coefficient sets per dataset (as in Figure 2, bottom row). These were baseline corrected to the pre-stimulus time window. Word onset and frequency estimates, averaged across subjects, were plotted as butterfly ERPs (see Figure 5(A, B)).

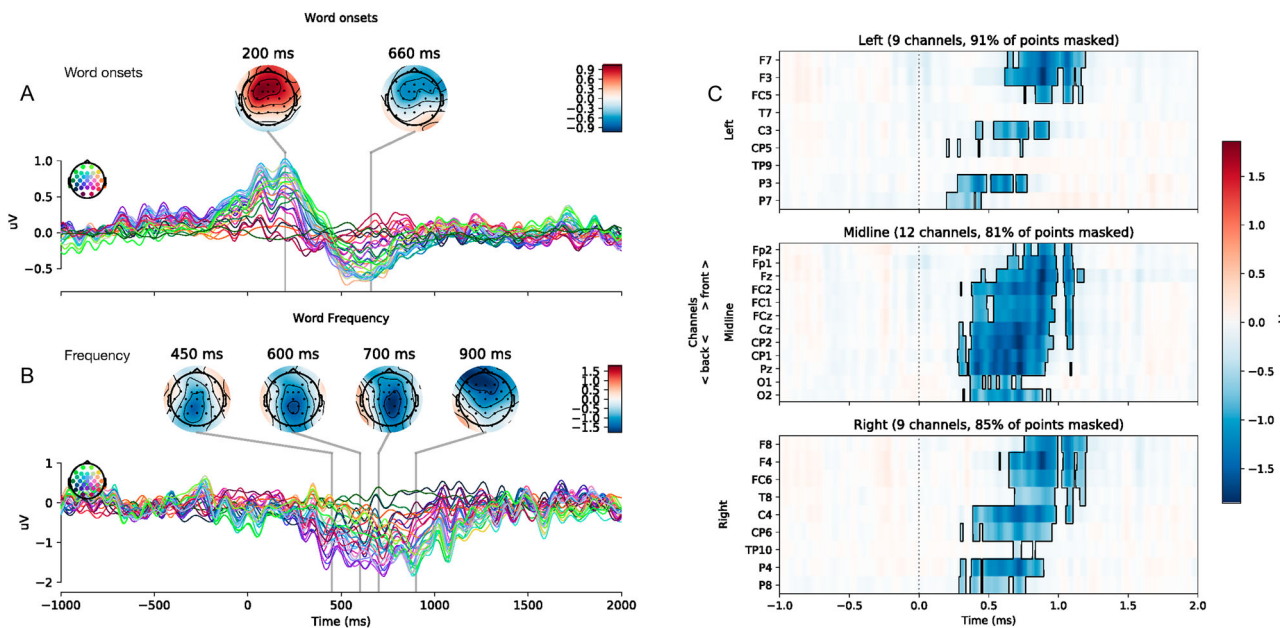


Figure 5. Coefficient time series. (A/B): Butterfly plots of coefficients, averaged across subjects, for word onsets (A) and word frequency as a continuous measure (B). Each line shows one channel; colours were chosen by converting their x, y and z coordinates to RGB values so that spatially adjacent channels share a similar colour (see inset at top left). The y axis shows the regression coefficient, which has the unit uV. The x axis shows the lags. Thus, each line shows how EEG activity at one channel depends on the word at t_0 , isolating one word characteristic by accounting for all other word characteristics and events; for word onsets, it corresponds to the activity elicited by word onsets compared to all other time points, for frequency, it corresponds to the change in EEG activity in common vs. frequent words; e.g. the difference between the most and the least frequent words is approx. 2 uV. (C): The same data as in (B), but with statistical significance masking of the frequency effect across subjects. Channels are grouped by hemisphere, stacked, and plotted as heatmaps, colour-coding for amplitude/regression coefficient size. Time/channel points not significant at $p < .05$ (TFCE; see text) where set transparent, and marked by black lines.

For statistical validation, the estimate for word frequency as a continuous covariate was subjected to a cluster-based permutation test (Maris, 2011) of the individual subject's coefficient sets against a null hypothesis of random sign. Threshold-free cluster estimation (Mensen & Khatami, 2013) was chosen to minimise parameter tuning (start: 0, step: 0.2). Clusters in the observed data exceeding the 2.5th and 97.5th percentiles in the surrogate data (1024 permutations) were collected. For each channel, time courses were plotted as heatmaps and data points exceeding the 5% threshold indicated (see Figure 5(C)). This highlights at which time points and channels the frequency estimate corresponded to spatio-temporal clusters of large amplitude unlikely under a nonparametric permutation-based null hypothesis of no relationship between EEG and frequency.

Next, for the 34-parameter model with individual frequency ranks as dummy predictors, for each bin, the results for all datasets were averaged, and the average of channels Pz and Cz (a priori assumed to reflect frequency-sensitive N400 activity; Kutas & Federmeier, 2011) was plotted for each bin (see Figure 6(A)). Then, the exact form of the relationship between EEG and frequency was quantified. For this, first, the entire coefficient set for each frequency bin was reduced to one datapoint as follows. For each dataset, the continuous frequency estimates from the first model were selected for all *other* datasets (leave-one-out cross-validation). The Frobenius inner product between this grand mean estimate and the estimates for each frequency bin for the held-out dataset was calculated, resulting in one number per dataset per frequency bin. Conceptually, this corresponds to estimating the neural correlates of word frequency, and then

estimating how strongly this response is expressed for items of a given frequency rank bins, expressed as a single number that is comparable across bins. This procedure was chosen because linear spatio-temporal filters can facilitate further analysis (Parra, Spence, Gerson, & Sajda, 2005); cross-validation was required to prevent circular analysis. The correlation between frequency ranks and the z-scored response strengths was calculated, and visualised in Figure 6(B).

4.2. Results

Word onsets reflected in an early-peaking component with a fronto-central positive maximum. EEG was sensitive to word frequency in a later, long-lasting time window from approx. 375 to 1250 msec after word onset, with a central topography shifting from an initially parietal pattern, with an N400-like spatial distribution, to a frontal pattern in the later time window. This word frequency effect was statistically significant across subjects ($p < .001$).

Investigating the form of the word frequency dependence of the EEG pattern strength (Figure 6(A)), a linear relationship was apparent (Figure 6(B)), $r = 0.873$, $p < 0.001$.

4.3. Discussion of example analysis

Here, a well-known effect was replicated to demonstrate the main points of contact, as well as the main differences, between, e.g. the QEMC difference evoked-potential approach. Time-resolved linear regression identified a dependence of EEG activity following words in the form of an N400 effect. Importantly, this N400 does not consist of a contrast between experimentally

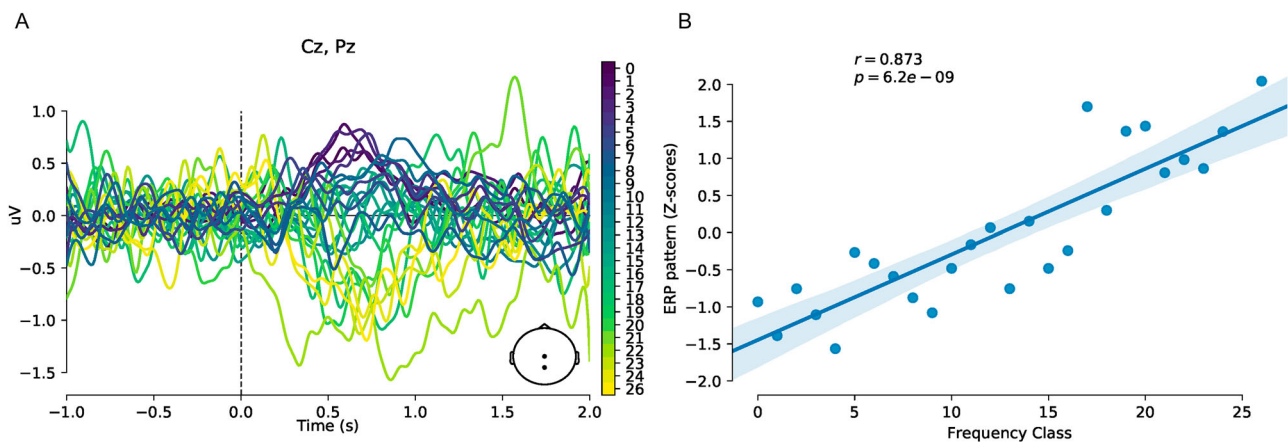


Figure 6. (A): Coefficient time series for each frequency rank bin (average, across subjects, of channels Cz and Pz). Line colour indicates frequency rank, with darker colours for more frequent words. (B): Correlation between the strength of the frequency pattern (y-axis; estimated by, within a cross-validation loop, taking the product of the coefficient maps for each bin and the continuous word frequency effect, averaging across subjects, and z-scoring), and frequency ranks bins (x-axis.)

manipulated conditions, but a regression estimation over all words occurring in a short narrative. It also accounts for effects of sentential position, avoiding the need to separately present early vs. late words in a sentence. The effect has the known characteristics of the N400 word frequency effect, and is consistent with previous regression-based estimates of the N400 from non-continuous paradigms (Kutas & Federmeier, 2011; e.g. Laszlo & Plaut, 2012).

The observed N400 pattern showed a linear dependency on frequency rank. However, this is not a perfect estimate of the word frequency effect. The aim of this analysis was to demonstrate the technique. The regression estimation takes into account only features entered into the model. Word frequency is correlated with features not included in the model, including many known to influence N400 amplitudes, such as cloze probability (Kutas & Federmeier, 2011).

5. Conclusions

Sentence processing and natural language comprehension research must not resign itself to focusing on careful experimental design of controlled experiments; another option is careful modelling of language as it is. Although this review champions the time-resolved regression side, it is not intended to claim other analysis tools are obsolete; rather, they complement each other. For example, minimal-contrast factorial designs can allow direct causal inference (unlike correlational approaches), establishing its continued role as a primary tool for researchers.

Fundamentally, a theory of language processing should be able to account for both controlled and experimentally manipulated speech, and for naturalistic speech. The main benefit of temporally resolved regression is that it can *account* for (near-)arbitrary models of language; if it can be quantified, it can be entered into the regression model, and its neural consequences (coefficient time series) and its predictive value (via encoding) can be estimated. The main drawback is the inverse: it requires good models to be applicable. Ill specified models will lead to inapplicable inference, and to uninterpretable or misleading results.

Time-resolved regression can be employed as an exploratory tool, by iteratively testing models on one or (ideally) multiple data sets, and checking which predictors are reflected in the data. It can also work as a confirmatory tool, by validating a previously constructed model on novel data. In both cases, a clear understanding of the nature of the procedure should be kept in mind.

Acknowledgments

Finally, I thank the reviewers and the action editor for their extensive, constructive commentary.

Funding

Benjamin Straube, Miriam Steines and Yifei He have made available the dataset employed here, originally obtained under a Von Behring-Roentgen-Stiftung (Project no. 59-0002; 64-0001) grant to Benjamin Straube, Helge Gebhardt and Gerhard Sammer. Alexandre Gramfort, Denis Engeman and Marijn van Vliet, Eric Larson, Jean-Rémi King and Chris Holdgraf have contributed code and conceptual support. This work was supported in part by German Research Foundation grant (BO 2471/3-2) awarded to Ina Bornkessel Schlesewsky, and by grant 617891 to Christian J. Fiebach.

ORCID

Jona Sassenhagen  <http://orcid.org/0000-0002-9935-8621>

References

- Aertsen, A., Johannesma, P., & Hermes, D. (1980). Spectro-temporal receptive fields of auditory neurons in the grassfrog. *Biological Cybernetics*, 38(4), 235–248.
- Alday, P. M., Schlesewsky, M., & Bornkessel-Schlesewsky, I. (2017). Electrophysiology reveals the neural dynamics of naturalistic auditory language processing: Event-related potentials reflect continuous model updates. *eNeuro*, 4(6), ENEURO-0311.
- Almon, S. (1965). The distributed lag between capital appropriations and expenditures. *Econometrica*, 33(1), 178–196. Retrieved from <https://doi.org/10.2307/1911894>.
- Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59(4), 390–412.
- Bornkessel-Schlesewsky, I., & Schlesewsky, M. (2009). The role of prominence information in the real-time comprehension of transitive constructions: A cross-linguistic approach. *Language and Linguistics Compass*, 3(1), 19–58.
- Brennan, J. R. (2016). Naturalistic sentence comprehension in the brain. *Language and Linguistics Compass*, 10(7), 299–313.
- Broderick, M. P., Anderson, A. J., Di Liberto, G. M., Crosse, M. J., & Lalor, E. C. (2017). Electrophysiological correlates of semantic dissimilarity reflect the comprehension of natural, narrative speech. *bioRxiv*, 193201.
- Burns, M. D., Bigdely-Shamlo, N., Smith, N. J., Kreutz-Delgado, K., & Makeig, S. (2013). Comparison of averaging and regression techniques for estimating event related potentials. *Engineering in Medicine and Biology Society (EMBC), 2013 35th Annual International Conference of the IEEE* (pp. 1680–1683). IEEE, Osaka, Japan.
- Crosse, M. J., Di Liberto, G. M., Bednar, A., & Lalor, E. C. (2016). The multivariate temporal response function (mTRF) toolbox: A matlab toolbox for relating neural signals to continuous stimuli. *Frontiers in Human Neuroscience*, 10.
- Delorme, A., & Makeig, S. (2004). EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including

- independent component analysis. *Journal of Neuroscience Methods*, 134(1), 9–21.
- Di Liberto, G. M., O'Sullivan, J. A., & Lalor, E. C. (2015). Low-frequency cortical entrainment to speech reflects phoneme-level processing. *Current Biology*, 25(19), 2457–2465. Retrieved from <https://doi.org/10.1016/j.cub.2015.08.030>
- Dufau, S., Grainger, J., Midgley, K. J., & Holcomb, P. J. (2015). A thousand words are worth a picture: Snapshots of printed-word processing in an event-related potential megastudy. *Psychological Science*, 26(12), 1887–1897.
- Eisner, R. (1960). A distributed lag investment function. *Econometrica*, 28(1), 1–29. Retrieved from <https://doi.org/10.2307/1905291>
- Fiebach, C. J., Schlesewsky, M., & Friederici, A. D. (2002). Separating syntactic memory costs and syntactic integration costs during parsing: The processing of German WH-questions. *Journal of Memory and Language*, 47(2), 250–272.
- Frank, S. L., Otten, L. J., Galli, G., & Vigliocco, G. (2015). The ERP response to the amount of information conveyed by words in sentences. *Brain and Language*, 140, 1–11. Retrieved from <https://doi.org/10.1016/j.bandl.2014.10.006>
- Galton, F. (1886). Regression towards mediocrity in hereditary stature. *The Journal of the Anthropological Institute of Great Britain and Ireland*, 15, 246–263. Retrieved from <http://www.jstor.org/stable/2841583>
- Ghitza, O., Giraud, A.-L., & Poeppel, D. (2013). Neuronal oscillations and speech perception: Critical-band temporal envelopes are the essence. *Frontiers in Human Neuroscience*, 6(January), 4–7.
- Goldhahn, D., Eckart, T., & Quasthoff, U. (2012). Building large monolingual dictionaries at the Leipzig corpora collection: From 100 to 200 languages. *LREC* (Vol. 29, pp. 31–43). Istanbul.
- Gonçalves, N. R., Whelan, R., Foxe, J. J., & Lalor, E. C. (2014). Towards obtaining spatiotemporally precise responses to continuous sensory stimuli in humans: A general linear modeling approach to EEG. *NeuroImage*, 97, 196–205. Retrieved from <https://doi.org/10.1016/j.neuroimage.2014.04.012>
- Gramfort, A., Luessi, M., Larson, E., Engemann, D. A., Strohmeier, D., Brodbeck, C., & Hämäläinen, M. (2013). MEG and EEG data analysis with MNE-Python. *Frontiers in Neuroscience*, 7.
- Gruber, M. H. (2013). *Matrix algebra for linear models*. Hoboken, NJ: John Wiley & Sons.
- Hasson, U., Yang, E., Vallines, I., Heeger, D. J., & Rubin, N. (2008). A hierarchy of temporal receptive windows in human cortex. *The Journal of Neuroscience*, 28(10), 2539–2550.
- Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). *The elements of statistical learning*. New York: Springer.
- Hauk, O., Davis, M. H., Ford, M., Pulvermüller, F., & Marslen-Wilson, W. D. (2006). The time course of visual word recognition as revealed by linear regression analysis of ERP data. *NeuroImage*, 30(4), 1383–1400.
- Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1), 55–67.
- Holdgraf, C. R., Rieger, J. W., Micheli, C., Martin, S., Knight, R. T., & Theunissen, F. E. (2017). Encoding and decoding models in cognitive electrophysiology. *Frontiers in Systems Neuroscience*, 11(61). Retrieved from <https://doi.org/10.3389/fnsys.2017.00061>
- Hyvarinen, A. (1999). Fast and robust fixed-point algorithms for independent component analysis. *IEEE Transactions on Neural Networks*, 10(3), 626–634.
- Jung, T.-P., Makeig, S., Humphries, C., Lee, T. W., McKeown, M. J., Iragui, V., & Sejnowski, T. J. (2000). Removing electroencephalographic artifacts by blind source separation. *Psychophysiology*, 37(2), 163–178.
- Kriegeskorte, N., & Kievit, R. A. (2013). Representational geometry: Integrating cognition, computation, and the brain. *Trends in Cognitive Science*, 17, 401–412.
- Kutas, M., & Federmeier, K. D. (2011). Thirty years and counting: Finding meaning in the N400 component of the event-related brain potential (ERP). *Annual Review of Psychology*, 62, 621–647.
- Kutas, M., & Hillyard, S. A. (1980). Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science*, 207(4427), 203–205.
- Lalor, E. C., & Foxe, J. J. (2010). Neural responses to uninterrupted natural speech can be extracted with precise temporal resolution. *European Journal of Neuroscience*, 31(1), 189–193.
- Laszlo, S., & Plaut, D. C. (2012). A neurally plausible parallel distributed processing model of event-related potential word reading data. *Brain and Language*, 120(3), 271–281.
- Lerner, Y., Honey, C. J., Silbert, L. J., & Hasson, U. (2011). Topographic mapping of a hierarchy of temporal receptive windows using a narrated story. *The Journal of Neuroscience*, 31(8), 2906–2915.
- Maris, E. (2011). Statistical testing in electrophysiological studies. *Psychophysiology*, 49(4), 549–565.
- Mensen, A., & Khatami, R. (2013). Advanced EEG analysis using threshold-free cluster-enhancement and non-parametric statistics. *NeuroImage*, 67(Supplement C), 111–118. Retrieved from <https://doi.org/10.1016/j.neuroimage.2012.10.027>
- Miller, G. A., & Chapman, J. P. (2001). Misunderstanding analysis of covariance. *Journal of Abnormal Psychology*, 110(1), 40–48.
- Nagels, A., Kauschke, C., Schrauf, J., Whitney, C., Straube, B., & Kircher, T. (2013). Neural substrates of figurative language during natural speech perception: An fMRI study. *Frontiers in Behavioral Neuroscience*, 7, 121. Retrieved from <https://doi.org/10.3389/fnbeh.2013.00121>
- Oostenveld, R., Fries, P., Maris, E., & Schoffelen, J.-M. (2011). FieldTrip: Open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Computational Intelligence and Neuroscience*, 2011, 1–9.
- Parra, L. C., Spence, C. D., Gerson, A. D., & Sajda, P. (2005). Recipes for the linear analysis of EEG. *NeuroImage*, 28(2), 326–341.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., & Dubourg, V. (2011). Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research*, 12, 2825–2830.
- Sassenhagen, J., & Alday, P. M. (2016). A common misapplication of statistical inference: Nuisance control with null-hypothesis significance tests. *Brain and Language*, 162, 42–45.
- Smith, N. J. (2012). *Scaling up psycholinguistics*. Proquest, UMI Dissertation Publishing.
- Smith, N. J., & Kutas, M. (2015a). Regression-based estimation of ERP waveforms: I. The rERP framework. *Psychophysiology*, 52(2), 157–168.
- Smith, N. J., & Kutas, M. (2015b). Regression-based estimation of ERP waveforms: II. Nonlinear effects, overlap correction, and practical considerations. *Psychophysiology*, 52(2), 169–181.

- St George, M., Kutas, M., Martinez, A., & Sereno, M. I. (1999). Semantic integration in reading: Engagement of the right hemisphere during discourse processing. *Brain*, 122(7 Pt), 1317–1325.
- St George, M., Mannes, S., & Hoffinan, J. E. (1994). Global semantic expectancy and language comprehension. *Journal of Cognitive Neuroscience*, 6(1), 70–83.
- VanRullen, R., & Macdonald, J. S. (2012). Perceptual echoes at 10 hz in the human brain. *Current Biology*, 22(11), 995–999.
- Varoquaux, G., Raamana, P. R., Engemann, D. A., Hoyos-Idrobo, A., Schwartz, Y., & Thirion, B. (2017). Assessing and tuning brain decoders: Cross-validation, caveats, and guidelines. *NeuroImage*, 145, 166–179.
- Viola, F. C., Thorne, J., Edmonds, B., Schneider, T., Eichele, T., & Debener, S. (2009). Semi-automatic identification of independent components representing EEG artifact. *Clinical Neurophysiology: Official Journal of the International Federation of Clinical Neurophysiology*, 120(5), 868–877.
- Westfall, J., & Yarkoni, T. (2016). Statistically controlling for confounding constructs is harder than you think. *PloS one*, 11(3), e0152719.
- Whitney, C., Huber, W., Klann, J., Weis, S., Krach, S., & Kircher, T. (2009). Neural correlates of narrative shifts during auditory story comprehension. *NeuroImage*, 47(1), 360–366.
- Wong, D. D., Fuglsang, S. A., Hjortkjær, J., Ceolini, E., Slaney, M., & de Cheveigné, A. (2018). A comparison of temporal response function estimation methods for auditory attention decoding. *bioRxiv*. Retrieved from <https://doi.org/10.1101/281345>