A primer on EEG data analysis

Fabian Dablander

University of Tübingen

Abstract

Analyzing large amounts of noisy, non-linear, high-dimensional time-series data is no easy task. The analysis of EEG data presents such a challenge. Between initial data pre-processing, statistical analysis, and the final figure presented in publications lie many subtle and not so subtle steps. Even after familiarizing oneself with these steps, methods papers proposing fancier analyses and more clever pre-processing techniques are published on a regular basis, making it difficult to keep up. Beginning researchers might find themselves confused in the jungle of EEG data analysis, paralyzed by the many routes they could take. In this paper, I try to give an overview of the EEG data analysis pipeline, from pre-processing to statistical analysis. In addition to discussing standard analysis practices and their extensions, I discuss novel approaches such as multilevel modeling, general additive mixed models, cognitive model-based approaches, as well as their main challenge: increasing the signal to noise ratio in single trials. I have included many references the interested reader may find helpful, a technical appendix that derives the basic mathematical ideas presented in the main text, and a corresponding online appendix providing intuition in Python.

Keywords: EEG, overview, machine learning, single-trial analysis, wavelets

A primer on EEG data analysis

The electroencephalogram (EEG) has long been a popular tool to measure cognitive processes on a millisecond time scale. Initially developed by Hans Berger in the 1920s, it quickly established itself as an indispensible tool for cognitive neuroscience and clinical practice (for a detailed overview of its history, see Collura, 1993).

As with any tool, EEG has both advantages and disadvantages. Understanding how the signal EEG measures is generated sheds light on the features of the tool (for a detailed overview, see Buzsáki, Anastassiou, & Koch, 2012). Briefly, postsynaptic potentials can last up to several hundreds of milliseconds, and create tiny dipoles at the neurons. When large populations of spatially aligned neurons receive similar inputs (excitatory or inhibitory), their dipoles are summed and produce an event-related component. Unfortunately, an EEG electrode does not only measure the dipole that is created at its scalp location. Instead, because the brain is a conductive medium, the electricty due to one dipole spreads across the entire scalp, a phenomenon termed *volume conduction*.

Resultingly, because EEG measures neural activity directly, it yields very high temporal resolution, allowing us to measure brain dynamics as they evolve (da Silva, 2013). This is in contrast to fMRI, which relies on blood flow, or PET and NIRS, which rely on metabolic activity. Additionally, EEG is non-invasive, and fairly inexpensive. Due to volume conduction, however, its spatial resolution is very poor[1]. While we can determine wich recorded activity belongs to which electrode – *topographical localization* – a much harder challenge is to determine to which physical neural source it can be attributed to – *brain localization*.

While EEG has been in use by researchers and clinicians for many decades, recent years have seen a rise of applications aimed at the mainstream market (from games to meditation), and companies aimed at the open-source, "do-it-yourself" community have cropped up (see OpenBCI). A main challenge for increased usability and popular uptake

[1] It is so poor that, "if you had data only from one electrode, you would not be able to determine with any reasonable accuracy where in the brain that signal was generated." (Cohen, 2014a, p. 26-27)

is to achieve relatively noise-free recordings from mobile EEG systems (which does seem in reach, see Ries, Touryan, Vettel, McDowell, & Hairston, 2014; Mullen et al., 2015).

As with any measure that allows going beyond behavior, measuring cognitive processes on a more fine-grained scale, EEG records massive amount of data; per electrode (commonly 32-256), per subject, and per time-point (sampling rates of up to 500Hz are common). Figure 1 below shows raw EEG data.

How do we pre-process, clean, and analyze such large amounts of data? How do we navigate the jungle of EEG data analysis without being paralysed by its many routes? In this paper, I provide an overview of those challenges, linking to additional ressources for more in-depth treatment. While useful for the beginning researcher, I also hope that some ideas, perspectives, or references are new even to the experienced researcher.

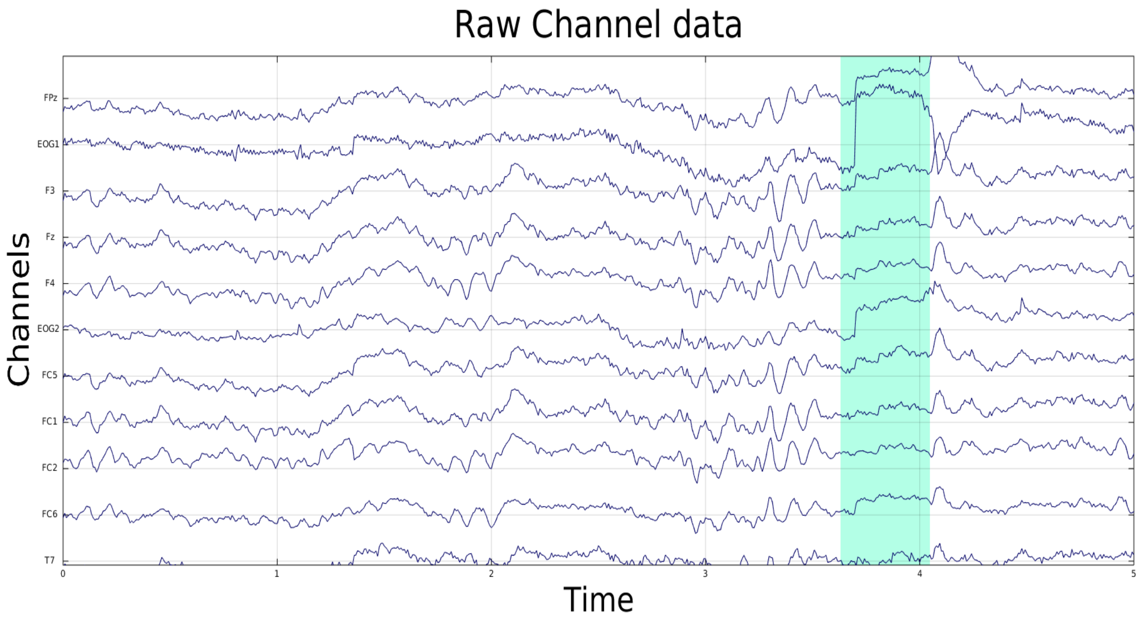

Figure 1. Shows non-epoched, raw channel data for a select group of electrodes. Shaded in blue is a blink artefact.

Overview of this paper

There are many different ways to analyze EEG data, and to do justice to the complexity would require several books (Cohen, 2014a; Luck, 2014; Freeman & Quiroga, 2012). Consequently, I will mainly focus on *event related potentials* (ERPs) and their analysis.

Event-related potentials are changes in the ongoing EEG activity due to stimulation; exogenous ERPs are elicitated by the visual, auditory, or somasensory properties of a physical stimulus. More interestingly, endogenous ERPs are elicitated by internal brain processes and can be used to study cognitive processes (thee exact origin of ERPs is debated, see Bastiaansen, Mazaheri, & Jensen, 2012).

In the first part of the paper, I will give an overview on different way of representing the multidimensional EEG signal and the pre-processing pipeline including various ways of normalization, artefact detection, and filtering. Explaining the standard statistical approach – averaging over trials and individuals – and a recently suggested "massive univariate" alternative (Groppe, Urbach, & Kutas, 2011a), I will argue that a regression based formulation allows for more powerful modeling (see Smith & Kutas, 2015a). Within this framework, I will try to make the case for non-linear, multilevel modeling of the entire ERP wave (see Tremblay & Newman, 2015). A brief section hinting at the machine learning techniques used in research on brain-computer interfaces (BCIs) is followed by a short section on cognitive-model based approaches (see Forstmann & Wagenmakers, 2015). Lastly, I will briefly present approaches aimed at increasing the signal to noise ratio in single trials.

While not discussing means to increase the spatial localization of EEG in detail (for an overview, see Jatoi, Kamel, Malik, Faye, & Begum, 2014), I will hint at how using different spatial filters and pre-processing methods in general may influence the conclusions drawn from the data (see also Cohen & Gulbinaite, 2014).

There is a crucial issue with (EEG) data analysis of which the reader should be aware, but which will not be discussed in detail in this paper. As with any method that requires many pre-processing steps, the choice of algorithms and statistical analyses are enormous (Carp, 2012; Cohen, 2014a), and hold many degrees of freedoms or "forks in the garden" (Gelman & Loken, 2014) which allow researchers to present anything as statistically significant (Simmons, Nelson, & Simonsohn, 2011; Luck & Gaspelin, in press). Open data, reproducible code, and registration of one's methodology, pre-processing pipeline, and statistical analysis prior to data collection alleviate these issues and are important steps towards a more transparent and solid science. A new publishing format called *Registered Reports* implementing these changes has already been adopted by journals from various fields (e.g., Chambers, 2013; King et al., 2016); see also https://osf.io/8mpji/wiki/home/.

While not discussing the mathematical ideas in the main text in detail, a corresponding technical appendix includes their derivations, while an online appendix provides intuition in Python. All materials can be found on https://github.com/fdabl/EEGpaper.

Preliminaries and pipeline

Signal representation

The EEG signal is multidimensional, encoding information in frequency, amplitude, phase, and time. Taking a look at the sine waves plotted in figure 2, frequency refers to the number of oscillations per second, measured in Hertz. Peak-to-peak amplitude denotes the height of the wave, and is commonly squared to yield power, which is a measure of the amount of energy in a certain frequency band at a certain point in time. Phase is a circular measure, and gives the offset of the signal; thinking in terms of complex numbers, power is the radius squared, and phase is the angle. Depending on what features we want to highlight, the EEG signal can be represented in different domains. Figure 2 visualizes those domains in the first row, while the second row shows their respective bases.

Time domain. The first, vertical part of figure 2 shows the trial-averaged ERP waveform for a single channel and five individual trials below. The power of the individual trials is much higher, and they look much noisier. Averaging is the most common approach of increasing the signal to noise ratio, which grows with the square root of the number of trials, $\sqrt{N}$. However, averaging in the time domain can lead to severe distortions when single trials differ in their onset, i.e. are not phase-locked, also referred to as *latency jitter*. While there are some remedies (see Luck, 2014, pp. 271), time-frequency based averaging has become the dominant approach over the last decade.

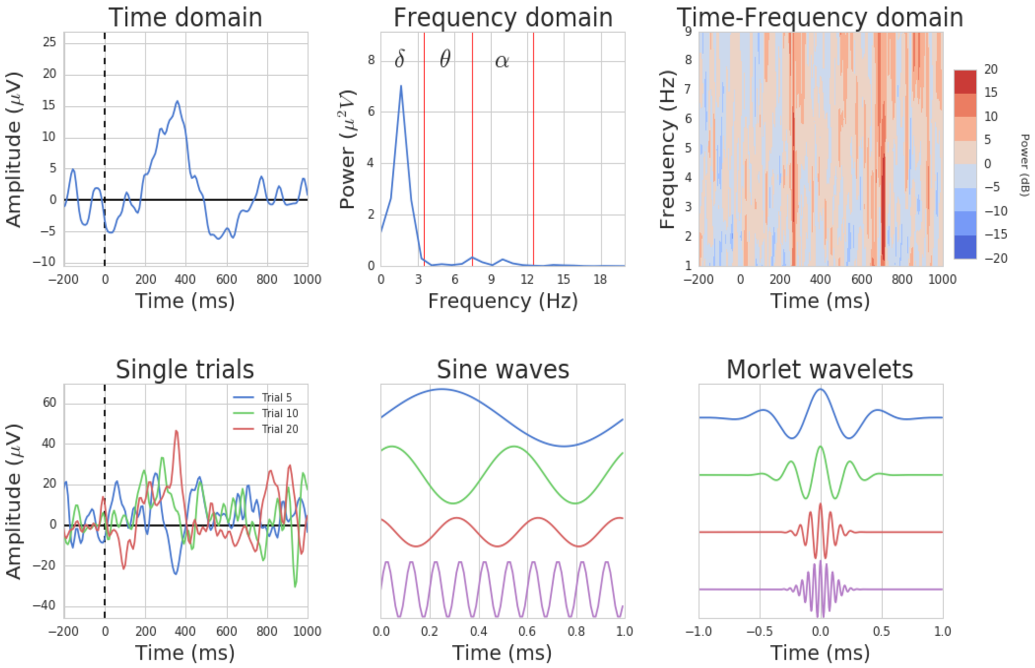

Figure 2. First vertical part shows the averaged ERP waveform of electrode FPz based on all trials, and three single trials below. The second part shows the power spectrum of the average ERP via a Fourier transform; plotted below are sine waves of varying amplitude, frequency, and phase. The last part shows the result of time-frequency decomposing the single trials and subsequently averaging them. Power values are decibel normalised against the baseline; plotted below are Morlet wavelets varying in width and frequency. EEG data are taken from the EEGLAB tutorial, and are not pre-processed except bandpass filtered ($1 - 25$Hz), but the general theme becomes clear.

In the time-domain, researchers look at peaks of the waveform such as P300 or N100, which are named after their polarity and latency; although there are different naming conventions (see Luck, 2014, pp. 72), inconsistently applied[2]. Research on various ERP components is vast; for a whole book on “ERPology”, see Luck and Kappenman (2011).

Frequency domain. Oscillatory brain activity is important for the functional communication between populations of neurons, and is thus integral to sensory and cognitive functions (Cohen, 2015; Başar, Başar-Eroglu, Karakaş, & Schürmann, 2001).

[2] The inconsistency in terminology across a broad range of concepts is one main challenge of EEG research (see Cohen & Gulbinaite, 2014).

By viewing the ERP waveform in the time-domain, it is difficult to extract the frequencies at which neurons oscillate. Using the Fourier transform (see technical appendix), the EEG signal can be decomposed into sine waves of different frequencies, amplitudes, and phases. The second part of figure 2 shows the Fourier decomposition of the averaged ERP, and sine waves varying in parameters. The most common frequency bands are denoted delta ($0 - 3.5$Hz), theta ($3.5 - 7.5$Hz), alpha ($7.5 - 12.5$Hz), beta ($12.5 - 30$Hz), gamma ($30 - 60$Hz; classification taken from Freeman & Quiroga, 2012, p. 31)[3]. Corresponding to event related potentials in the time-domain, researchers can look at event-related oscillations in the frequency domain (EROs; Herrmann, Rach, Vosskuhl, & Strüber, 2014). Note that the Fourier transform is a mathematical tool, and it describes which kind of sine waves we would need to reconstruct the signal; observing a sine wave with a certain frequency resulting from a Fourier transformation does not necessarily entail that the brain oscillated at that frequency (see Luck, 2014, p. 223-226).

Time-Frequency domain. The Fourier transform assumes a stationary signal i.e., that properties of the signal do not change over time, an assumption clearly violated in EEG. Additionally, it provides no straightforward means to extract time information; for example, it is not at all clear whether the frequency component of, say, 5Hz is constant over time or whether it occurs in short bursts of a few milliseconds. This is because the Fourier transform uses sine waves as basis which do not have a temporal localization. To quantify the time dynamics of different frequency components, one can use the short-time Fourier transform, multitaper methods, or other techniques (see Cohen, 2014a, ch. 15-17; for a technically detailed comparison, see Wacker & Witte, 2013). In recent years, however, wavelets have become the dominating approach for time-frequency decomposition (see also Freeman & Quiroga, 2012, ch. 2-4; Herrmann et al., 2014).

There are many different wavelet families, but not all are useful for EEG data. Of

[3] These are not arbitrary groupings, but result from neurobiological mechanisms of the brain which seem to be universal across the mammalian species (Buzsáki, Logothetis, & Singer, 2013).

central importance is that the wavelet looks similar to the signal we want to analyze. For EEG data, (complex) Morlet wavelets are commonly used, which are the result of convoluting a (complex) sine wave with a Gaussian window. The frequency of the sine wave determines at which frequency we decompose, while the width of the Gaussian window controls the temporal resolution; the number of frequencies and the widths of the Gaussian windows are non-trivial parameters to set, and can markedly impact the results obtained (see Cohen, 2014a, pp. 168). For more on wavelets, see Strang (1994), Graps (1995) and Freeman and Quiroga (2012, ch. 3-4). For a review of applications and measures used with time-frequency analyses, see Roach and Mathalon (2008).

To avoid pitfalls with averaging in the time-domain due to non-phase locked activity, single-trials are commonly transformed into the time-frequency domain and the resulting power values, which are non-negative, are subsequently averaged (Herrmann et al., 2014). The third part of figure 2 gives the result of this procedure; shown below are members of the *Morlet wavelet family* varying in frequency and width. There is an inherent trade-off, known as the *Gabor limit*: the smaller the width of the Gaussian window, the better the resolution in time – at the cost of the resolution in frequency.

Pre-processing

Re-referencing[4]. It is important to stress that EEG measures electric *potential* between electrodes and a reference electrode in (micro) Volt. Thus it is a relative measure, and can be subject to big and subtle changes. More broadly, an *EEG channel* is comprised of an active, ground, and reference electrode. We can use the term *absolute voltage* to refer to the potential between an active electrode and the average of the rest of the head. Let A and G denote absolute potentials of an active electrode and the ground. We could have the amplifier record $A - G$, the voltage between A and G; however, because of the way amplifiers work, G is contaminated with electrical noise, rendering this approach futile. By introducing another electrode R, our reference electrode, we can compute the difference between the voltages,

[4] This section is heavily influenced by Luck (2014, p. 150-164).

$(A - G) - (R - G) = A - R$, and thus eliminate electrical noise. Amplifiers operating this way are called *differential amplifiers*.

Some recording systems set the reference electrode themselves (e.g., Cz), but for subsequent data analysis one usually wants to change this. This process is called *re-referencing*, and can be done offline via simple operations. For example, re-referencing to the electrode of the right mastoid is done by subtracting its potential from all other electrodes (e.g., $A_m = (A - R) - (M - R) = A - M$). An appealing choice is to re-reference to the average of all channels because this does not bias the result to any hemisphere, and also minimizes noise. However, as apparently not many researchers are aware of (see Luck, 2014, pp. 162), this can have subtle and not so subtle side-effects. For the latter, note that taking the average as reference impedes quantitative comparisons across experiments, because the ERP waveform will be influenced by the place and number of channels recorded. For an overview of these and other issues, see Dien (1998).

In sum, choosing a reference is an important pre-processing step that can drastically influence the results, and should be done carefully.

(Temporal) Filtering. Filtering is usually done on the continuous, non-epoched EEG signal and thus is the first step in the preprocessing pipeline; although being a linear operation, the exact order does not matter (see Luck, 2014, pp. 246). There are two major use-cases for filtering (see also Luck, 2014, ch. 7). Note that the Nyquist-Shannon sampling theorem states that analog signals are adequately sampled only at a frequency at least twice as high as the original signal. In the recorded signal, higher frequencies will appear as artefactual low frequencies, a phenomenon called *aliasing*[5], This is seen as slow drift in the EEG signal, and is removed by *high-pass* filters of about 0.1Hz. The second main use case is to reduce noise, and here high-pass filters can help attenuate low frequencies caused by the skin. Additionally, components with frequencies higher than 100Hz are often muscle artefacts, and can be removed by *low-pass* filters. Filters in the frequency domain have an equivalent

[5] For a beautiful demonstration, see here.

representation in the time-domain. In the time-domain, filtering is done by computing a running average or convolution, which results in *temporal smearing* – high-frequency noise is attenuated, and the waveform looks nice and smooth. While decreasing temporal resolution, it increases frequency resolution. Filtering is a distortion of the signal. Especially high-pass filters of 0.1Hz can be problematic; they can increase statistical power (Kappenman & Luck, 2010), but might also significantly distort the ERP waveform (Tanner, Morgan-Short, & Luck, 2015).

Epoching. The recording system continuously samples the EEG signal; for subsequent analysis, however, we need to know when a trial started and ended. For this reason one sends event triggers during the experiment, for example indicating stimulus onset. *Epoching* refers to time-locking the signal to stimulus onset (for each trial), such that $t = 0$ at this point. In other words, epoching is the process of going from a two-dimensional *channel* × *time* to a three-dimensional *channel* × *time* × *trial* representation. The *baseline* refers to time before stimulus, see also figure 2.

Baseline correction. Factors not related to the experimental procedure like skin hydration and static charges in the electrodes induce an amplitude offset in the EEG signal. Baseline correction is applied to each epoch in order to mitigate these issues by subtracting the mean pre-stimulus amplitude from the post-stimulus amplitude. However, this seemingly innocuous procedure can have some subtle implications (see Luck, 2014, pp. 251). For example, if the inter-trial interval is set too short, signal from the previous trial can overlap with the baseline of the current trial, and bias the baselining procedure (Woldorff, 1993; see also Smith & Kutas, 2015b).

Note that time-frequency based analyses often do not require baseline correction because different frequency components are isolated, and those corresponding to low drifts or vertical offsets can subsequently be ignored.

Artefact rejection and detection. Artefacts are systematic or unsystematic distortions of the EEG signal which decrease the signal to noise ratio, can lead to systematic biases when comparing experimental conditions, and can suppress sensory input (e.g. through blinks; see also Luck, 2014, ch. 6). Artefact *rejection* entails discarding artefact contaminated trials (across all channels) based on certain amplitude thresholds; for a review on dealing with eye movement induced artefacts, see Plöchl, Ossandón, and König (2012). Artefact *detection* refers to identifying the influence of artefacts and subsequently subtracting them from the signal. On this front, methods based on independent component analysis (ICA; see technical appendix) are most popular. ICA is a blind source seperation method that finds maximally statistically independent sources of variance in the EEG signal[6]. It results in a set of weights that indicate the component of each electrode in the signal. Artefactual components are thus identified and can be subtracted (Delorme, Sejnowski, & Makeig, 2007).

While on the topic, let me note that another use of ICA is to submit the components to subsequent data analysis instead of the *sensor* data. This is called *spatial filtering*, and there are other methods like principal component analysis (PCA), which, like ICA, only uses statistical information in the data, or the Surface Laplacian, which uses physical properties such as the interelectrode distances to project the data onto a new space (see Cohen, 2014a, ch. 23-24; Cohen, 2014b).

Baseline normalization. While baseline correction is a linear operation, baseline normalization is a non-linear operation that is commonly carried out when doing time-frequency based analyses. EEG signal has a $1/f$ scaling problem – its power decreases non-linearly with increasing frequency. This impedes quantitative comparisons of power across frequency bands and violates common parametric assumptions because raw power values are skewed and non-negative. To alleviate this problem, several transformations have been proposed (see Cohen, 2014a, ch. 18). Here I just mention the popular decibel conversion, $dB_{tf} = 10 \cdot \log(\frac{\text{post}_{tf}}{\overline{\text{pre}}_{tf}})$, where $\overline{\text{pre}}$ denotes the average baseline amplitude, post denotes post-stimulus activity, and t and f index time and frequency points, respectively. The time-frequency decomposition in figure 2 shows decibel-converted power. The resulting scale is logarithmic and relative to the baseline. Needless to say, this means that choosing the baseline becomes a critical issue.

[6] Statistical independence can be quantified by mutual information, a concept from information theory (see Hyvärinen & Oja, 2000). However, there are many different algorithms for ICA. Delorme, Palmer, Onton, Oostenveld, and Makeig (2012) compared and scored 18 of them.

Other transformations such as percent change or the Z-transform give different results, and thus one needs to be aware of those differences. Similarly, these transformations might not be innocuous. For example, Hu, Xiao, Zhang, Mouraux, and Iannetti (2014) found that using the percent change transformation leads to a bias in estimating event-related desynchronization and synchronization.

Above I have merely scratched the surface of the preprocessing pipeline. Steps that are experiment, task, or equipment specific further increase the list of preprocessing steps. Concluding, preprocessing holds various pitfalls and many degrees of freedom (for a quantification in fMRI research, see Carp, 2012). An important step towards a more standardized approach towards early stage preprocessing is taken by Bigdely-Shamlo, Mullen, Kothe, Su, and Robbins (2015).

Statistical analysis

After pre-processing, we can engage in statistical analysis, either to do confirmatory hypothesis testing, or to explore the data set and generate new hypotheses. Of course, we can and should do both, but we need to be aware that statistical procedures carry little evidential value when done in exploratory settings (De Groot, 2014). As always, replication is the best statistic.

Having said that, how can we analyze our data? Similar to the many steps involved in pre-processing, there is a huge amount of possible models we can employ – even when just looking at event-related potentials. For the remainder, it is useful to think of the EEG data as a three dimensional array indexing electrode, time, and trial (EEGLAB's convention; Delorme & Makeig, 2004).

Standard ERP analysis

Usually, in the time domain, *grand average* event-related potentials are constructed for each condition by averaging over trials and participants. Sometimes electrodes are clustered (frontal versus anterior) by additionally averaging over them. Usually, the dependent variable is peak-amplitude of the ERP component, peak-to-peak or base-to-peak amplitude differences, or the average voltage over some time window (see Luck, 2014, ch. 9). Similar to behavioral measures, the dependent variable is submitted to a standard parametric test, i.e. a t-test or ANOVA.

For example, Giraudet, Imbert, Bérenger, Tremblay, and Causse (2015) had participants do a simulated air traffic control task using two different visual notification designs – one subtle ("Color-Blink"), the other more salient ("Box-Animation"). Concurrently, participants had to perform an auditory task requiring them to react to rare pitch tones. The P300 component to these tones were taken to indicate remaining attentional ressources, and its amplitude, measured by taking the average within the 364 to 464 post-stimulus window, was greater for the salient design. This seems to establish the P300 component as a viable marker of design efficiency – good design takes up less attentional ressources.

However, as mentioned briefly above, we can also analyse the signal in the frequency domain. For an example, Goodin et al. (2012) stimulated participants with binaural beats at 7Hz (theta) and 16Hz (beta). Fourier transform was applied to the recorded signal and absolute spectral power was computed for each electrode for two frequency ranges corresponding to the binaural beats stimulation (5.5-7.5Hz for theta and 15.5-17.5 for beta). Electrodes were grouped into frontal, central, parietal, temporal, and occipital areas and subsequently averaged. Applying a within-subject ANOVA, the authors did not find differences in spectral power between stimulation (beta and theta) and control conditions, suggesting that binaural beats to not alter cognitive processes.

As an aside, note that when the parametric assumptions of statistical tests are violated, for example due to outliers, it might be invalidated and bias the conclusions drawn from the data. Recently, *robust procedures* have been suggested which are not as sensitive to parametric violations as their traditional counterparts (see Wilcox, 2012; Pernet, Chauveau, Gaspar, & Rousselet, 2011, for an implementation). For good applications, see for example Rorden, Karnath, and Bonilha (2007) and Rousselet and Pernet (2012).

Massive univariate approach

In the standard approach, researchers specify a time window in which they look at the averaged ERPs across conditions and submit this to a parametric technique such as ANOVA. Specifying a window by hand carries several problems with it; first, there is the danger of “double dipping” by selecting windows contingent on data (Kriegeskorte, Simmons, Bellgowan, & Baker, 2009; Vul, Harris, Winkielman, & Pashler, 2009) – this points to the issue of confusing exploratory and confirmatory research settings; second, one might miss interesting effects outside the specified time window, which can quickly happen even between electrodes; and third, standard analysis does not provide information about the onset and offset of the stimulus induced change in the EEG signal. Instead, the *massive univariate approach* entails conducting standard parametric tests on each time point for each electrode over the averaged ERP wave (Groppe et al., 2011a, 2011b). The massive univariate approach originated in fMRI research, where thousands of voxels are subjected to statistical tests, which requires controlling of the false positive rate (see for example Genovese, Lazar, & Nichols, 2002). Different multiple comparison methods, among them Bonferroni, permutation statistics, false discovery rate, and cluster based corrections, have been suggested to alleviate the problem (Maris & Oostenveld, 2007).

An example highlighting the advantages of the massive univariate approach is given by Scheer, Bülthoff, and Chuang (2016). Briefly, the authors had participants do a steering task while intermittently presenting distracting environmental sounds or beep tones. Using said analysis allowed them to, of course, identify the ERP components elicited by the distractors and determine their difference between the environmental and beep distractors, but also map the spatio-temporal dynamics of the ERP component. Concretely, they found that steering diminishes the amplitudes of early and late P300, as well as the re-orientation negativity to the environmental sound distractor but not to beep tones.

As an aside, if we are interested in quantifying the evidence provided by the data, *p*-values are a moot choice because they overestimate this evidence (Berger &

Delampady, 1987). Using results from Liang, Paulo, Molina, Clyde, and Berger (2008), Rouder, Speckman, Sun, Morey, and Iverson (2009) developed a Bayesian alternative to the classical t-test; however elegant the Bayesian approach may be, I am not sure it is appropriate here, especially since the focus of Bayesian techniques is not on error control.

In general, although the massive univariate approach seems somewhat inelegant, it is a good start (Guillaume Rousselet, public communication), and certainly a huge improvement over more traditional analyses.

Regression-based estimation and testing

Both approaches presented so far can be cast in a regression framework (Smith & Kutas, 2015a; Burns, Bigdely-Shamlo, Smith, Kreutz-Delgado, & Makeig, 2013), which provides a more powerful toolkit suitable to extensions (Smith & Kutas, 2015b). Before diving in, let's take a step back and ask: why is averaging a good idea? To answer this question, we have to think about our model. Implicitly in the standard approach, we assume that, on each trial for each electrode, the measured electrical activity is composed of the true activity and normally distributed noise. For this model, the mean is an unbiased and efficient estimator (but not inadmissible; for example, a hierarchical model yields better estimates, see Efron & Morris, 1977).

Taking the mean over time-locked single-trial ERP waves – that is, just constructing the average ERP – is equivalent to running a regression with only an intercept as predictor (see the technical appendix for a refresher on how regression works, and Smith and Kutas (2015a, pp. 159) for a longer explanation of why this is true).

But regression is a much more general approach. For example, the standard analysis approach does not have a natural way of representing continuous predictors. Thus it is common practice to split continous predictors into factors and continue with ANOVA (Tremblay & Newman, 2015, p. 1), which naturally decreases statistical power.

In general, when averaging over participants, we need to be conscious that we are modeling the *average participant* – on an individual level, the measure under scrutiny might look very different (for example, see Gaspar, Rousselet, & Pernet, 2011; Haegens, Cousijn, Wallis, Harrison, & Nobre, 2014)

Instead of ignoring the inter-subject variability, we can specify a model first on the trial-level, and subsequently subject these estimated beta weights to a model on the participant-level (Pernet, Chauveau, et al., 2011; Burns et al., 2013). The same methodology is dominating univariate fMRI analysis (Poline & Brett, 2012), and is referred to as *statistical parametric mapping* or GLM-approach (for General Linear Model)[7].

Non-linear multilevel models

Before moving on, we have to clarify crucial terminology. Proponents of the GLM approach often call it *hierarchical single-trial* analysis. Single-trial because for each participant seperately, we estimate the ERP waveform (1st level). Hierarchical because the result of these individual estimates are then submitted to a classical or robust parametric procedure, looking for differences on the group level (2nd level; Pernet, Sajda, & Rousselet, 2011; Pernet, Chauveau, et al., 2011). While this approach allows us to quantify the between-subject variance, I hesitate to call it hierarchical *proper*, because it does not regularize the estimates by shrinking them toward the grand mean, as is common when submitting both levels to what is frequently called a *linear mixed*, *hierarchical*, or *multilevel* model (see technical appendix; Baayen, Davidson, & Bates, 2008; Sorensen & Vasishth, 2015; Bates, Mächler, Bolker, & Walker, 2014). It further does not allow to model variation in the stimulus materials. Failing to model variability in stimuli is a long standing issue in psycholinguistics and other fields, which can lead to spurious conclusions (for example, see Judd, Westfall, & Kenny, 2012).

In addition to built-in regularization and the possibility to model all sources of variability, and in contrast to the standard approach, mixed models allow for

[7] These developments in EEG analysis seem parallel to the developments in the fMRI community, but with a time lag of about ten years. The GLM approach in fMRI has been criticized both on statistical (Monti, 2011) and conceptual (Stelzer, Lohmann, Mueller, Buschmann, & Turner, 2014) grounds.

unbalanced designs (which are common when events are time-locked based on participant's response), handle missing data, and are invariant to violations of sphericity and heteroscedasticity (Bagiella, Sloan, & Heitjan, 2000; Tibon & Levy, 2015).

Let's take another step back. When you look at an ERP waveform in the time-domain, it does not look linear at all (see figure 2). Instead of pursuing the massive univariate path – testing a time-wise normal model thousands of times – another approach is to model the entire time-course of the ERP waveform using generalized additive (mixed) models (GAMMs; Wood, 2006). Instead of forcing linearity on the data, we let the data inform the functional shape we are estimating (Tremblay & Newman, 2015). This approach inherits all benefits discussed above for the linear mixed model, but increases the modeling flexibility dramatically by getting rid of linearity. In a sense, it is the continuous version of the massive univariate approach; instead of modeling the time-course of the ERP wave by means of significant t-tests, GAMMs allow smooth estimation across participants, with regularization built-in. In this approach, we appropriately treat the data as being non-linear, high-dimensional time-series data. Similary to the massive univariate approach, which can correct for the dependency in, and clustering of the data, GAMMs can account for this auto-correlation by including an autoregressive model.

While Tremblay and Newman (2015) give an overview of applying GAMMs to ERP data, Meulman, Wieling, Sprenger, Stowe, and Schmid (2015) compare the use of traditional analysis with GAMMs by applying them to answer questions about age effects in grammar processing.

A note of caution. GAMMs are complex models, and an overhead in complexity must always be warranted by the problem one tries to solve. In any case, I believe mixed models and GAMMs to be a novel addition to the EEG researcher's methodological toolbox, allowing us to discover hidden patterns in the data that fuel theory building and computational modeling.

Machine learning in BCI

Not shying away from pathos, the main objective here is to *decode* brain signals in order to control machines. Frequently, the issue is one of classification: does this brain signal correspond to *this* or *that* brain state? Since this should be as efficient as possible, research in this area focuses on single-trials, instead of having the participants do many trials and average over them (Lemm, Blankertz, Dickhaus, & Müller, 2011). Because we do not increase the signal to noise ratio by averaging, the main issue is the low signal to noise ratio when focusing on single trials. Several spatial and temporal filters can be employed to the raw signal (more on that below), which leads to a subsequent high-dimensional feature space. Because of this high-dimensionality, most classifiers employed a linear, with some sort of regularization built-in (Blankertz, Lemm, Treder, Haufe, & Müller, 2011; Müller, Anderson, & Birch, 2003)[8]. Among the most popular is linear discriminant analysis (LDA), which is reasonably robust against violations of its assumption, and is optimal when they are met (Blankertz et al., 2011). For an extensive overview of different classification algorithms as applied to BCIs, see Lotte, Congedo, Lécuyer, Lamarche, and Arnaldi (2007); for an overview of BCI platforms and pipelines, see Schalk, McFarland, Hinterberger, Birbaumer, and Wolpaw (2004) and Müller et al. (2008).

Model-based cognitive neuroscience

Another, quite different path compared to traditional analysis focuses on developing mathematical models to bridge the gap between levels of analysis by analyzing behavioral and neuronal data in a joint model, using neural data to constrain behavioral models, or using the behavioral model to predict neural data (Forstmann & Wagenmakers, 2015; Turner, Forstmann, Love, Palmeri, & Van Maanen, 2016). This is an exciting development more generally, in that instead of relying on atheoretical, purely statistical models like regression, cognitive models more strongly correspond to, and instantiate a theory. Parameters of cognitive models are often directly interpretable

[8] But see the recent winning entry for the Kaggle BCI challenge here.

with respect to the theory. For a great motivating article of applying mathematical models to problems in cognitive neuroscience, see Forstmann and Wagenmakers (2015). For an introduction and review on applying cognitive models to the study of mind-wandering, see Hawkins, Mittner, Boekel, Heathcote, and Forstmann (2015).

Denoising of single-trials

Traditionally, the most powerful tool to increase the signal to noise ratio is averaging over trials; this is one of the main advantages of the standard analysis.

Averaging acts as a filter: it allows phase locked activity to pass, but blocks non-phase locked activity from contributing to the average. Because single-trial analysis cannot use such a filter, it has to rely on different techniques. One might distinguish between *sensor* versus *source-based* methods. The former measures some variable of interest, say peak measurement, in the space the EEG signal was recorded, while the latter techniques decompose the signal to estimate mathematical sources on which the peak is subsequently measured (see De Vos, Thorne, Yovel, & Debener, 2012, p. 1197). For example, De Vos et al. (2012) compared four different methods of single-trial filtering on a classification benchmark, finding that ICA based methods improve upon using the raw signal, while multiple regression and bandpass filtering did not. The authors used bandpass filtering instead of a wavelet based technique (Quiroga & Garcia, 2003) because the latter was not yet fully automatic. However, recent work extends the wavelet based approach, allowing for fully automatic denoising (Ahmadi & Quiroga, 2013). Concretely, the method works by first doing a wavelet decomposition on the average ERP, hard thresholding the wavelet coefficients incorporating their dependencies, and subsequently reconstructing the signal. While this approach is already very promising, new methods are developed at a rapid pace (for examples, see Van Vliet et al., 2016; Treder, Porbadnigk, Avarvand, Müller, & Blankertz, 2016), making an update and extension of the work by De Vos et al. (2012) desirable.

The point of this section was not to provide a detailed overview, but rather to hint at possible solutions for the problem of noise in single-trials. In any case, one may well state that single-trial denoising represents the cherry on top of the already complex EEG data analysis pipeline.

Conclusion

While EEG is a powerful tool of modern cognitive neuroscience, illuminating neural dynamics as they evolve, EEG data analysis presents are superb challenge, with many complexities and degrees of freedoms involved.

Having never taken a class on EEG data analysis nor signal processing, I have tried to give a broad overview of the basic issues, statistical approaches, new ideas, and tooling surrounding EEG data analysis. While possibly useful to the similarly uninitiated, I hope that some ideas, perspectives, or references were new even to the experienced researcher.

After presenting the basic, already quite involved pre-processing steps, I have argued that statistical approaches that collapse across participants or trials are suboptimal. An improvement upon this standard practice is provided by the massive univariate approach, especially when conducted to respect inter-individual differences and using robust parametric procedures (as implemented in Pernet, Chauveau, et al., 2011). More modern still, applying (possibly non-linear) multilevel models can further increase the power, flexibility, and robustness in modeling EEG data. Model-based approaches represent the most challenging, but rewarding ways of data analysis because they directly link to theory. Regardless of whether one uses machine learning to classify based on single-trials; multilevel, model-based, or other approaches trying to illuminate cognitive processes on a single-trial basis (Makeig, Debener, Onton, & Delorme, 2004; Quiroga, Atienza, Cantero, & Jongsma, 2007; Rey, Ahmadi, & Quiroga, 2015), increasing the signal to noise ratio in single trials is a formidable, but necessary to overcome challenge.

Based on this brief overview, future work might **a)** look into how model-based approaches to neuroergonomics can inform theoretical debate or increase predictive accuracy, **b)** scrutinize the regression-based approach along similar lines as Monti

(2011) did for fMRI research, **c)** extend the latter by using regularized estimators, **d)** demonstrate how mixed effects or multilevel models decompose the variance of the data in meaningful ways and improve estimation and prediction accuracy, **e)** review assumptions of GAMMs and their appropriateness for EEG data analysis, **f)** compare the massive univariate approach with GAMMs and judge their respective informativeness, **g)** provide an overview on how the univariate methods discussed here fail to map interesting functional multivariate connections between brain sites, **h)** evaluate the suitability of recently developed Bayesian methods for ERP data analysis **i)** apply the single-trial framework to workload in order to track potentially interesting dynamics on a finer scale, and **j)** compare novel single-trial denoising techniques for specific applications (i.e., an update of De Vos et al., 2012).

In sum, the analysis of EEG data is a complex endeavour, providing many pitfalls and thus ample opportunity for learning and exploration.

Technical Appendix

Convolution

Think about a function or signal, f, in the time-domain. Visually, convolution is the process of taking another function g called kernel, shifting it to the left, and sliding it across f.

Concretely, reflect or “swap” $g(x) \Rightarrow g(-x)$, and add a time-offset so it can slide across the x-axis, $g(-x+t)$. Then convolution means computing a weighted, sliding average between the original signal and the kernel

$$(f \star g)(t) = \int_{-\infty}^{\infty} f(x)g(t-x)\mathrm{d}t \tag{1}$$

If g is a symmetric function, this reduces to the cross-correlation. Without proof, convolution in the time-domain corresponds to multiplication in the frequency domain

$$\mathcal{F}(f \star g) = \mathcal{F}(f) \cdot \mathcal{F}(g) \tag{2}$$

where $\mathcal{F}$ stands for the Fourier transform. Because of the speed of the fast Fourier transform, convolution is commonly computed by

$$(f \star g) = \mathcal{F}^{-1}(\mathcal{F}(f) \cdot \mathcal{F}(g)) \tag{3}$$

Fourier transform

The Fourier transform is a linear transformation decomposing any signal into sine waves of varying frequencies. It matches the signal with complex exponentials (sine and cosine waves) of different frequencies. There are four different types, depending on whether the signal is continuous or discrete and periodic or non-periodic (see Freeman & Quiroga, 2012, ch. 2). Here I focus on discrete non-periodic signals. Noting the expansion

$$e^{ix} = \sum_{k=0}^{\infty} \frac{(ix)^k}{k!} \tag{4}$$
$$= 1 + ix - \frac{x^2}{2!} - i\frac{x^3}{3!} + \frac{x^4}{4!} + i\frac{x^5}{5!} \dots \tag{5}$$
$$= \sum_{k=0}^{\infty} \frac{x^{2k}}{2k!} + i \cdot \sum_{k=1}^{\infty} \frac{x^{2k+1}}{(2k+1)!} \tag{6}$$
$$= cos(x) + i \cdot sin(x) \tag{7}$$

where $i = \sqrt{-1}$ allows us to write sine and cosine waves compactly as

$$Ae^{i(2\pi ft+\phi)} = A[cos(2\pi ft + \phi) + i \cdot sin(2\pi ft + \phi)] \tag{8}$$

where A is the square root of the power (amplitude), f is the frequency, t is the time, and ϕ is the phase offset. Let $x[n]$, $n = 1, \dots, N$, be the digital signal derived by sampling the analog or continuous signal at equally spaced time intervals Δt; stated differently, the sampling frequency is $f_s = \frac{1}{\Delta t}$ and the length of the signal is $T = N \cdot \Delta t$.

The (discrete) Fourier transform decomposes any signal into a weighted sum of sine and cosine waves with frequencies $k \in \{0, 1, \dots, N-1\}$; the (complex) coefficients $X[k]$ are given by

$$X[k] = \sum_{n=0}^{N-1} x[n] \cdot e^{-i(2\pi kn/N)} \tag{9}$$

The inverse Fourier transform reconstructs the signal and is given by

$$x[n] = \frac{1}{N} \sum_{n=0}^{N-1} X[k] \cdot e^{i(2\pi kn/N)} \tag{10}$$

For more, see Freeman and Quiroga (2012, ch. 2) and Downey (2016, ch. 7).

Information theory

Information theory is a branch of mathematics that goes back to Claude Shannon and the 1950s. How surprising is an event? Shannon defined an event's surprisal as

$$h[X] = \log \frac{1}{p(x)} = -\log p(x) \tag{11}$$

which when averaged over the event's distribution is termed *entropy*

$$H[X] = -\mathbb{E}[\log p(x)] = -\sum_{x \in X} p(x) \log p(x) \tag{12}$$

Entropy is measured in bits, which is a measure of information; one bit encodes as much as information as provided by a yes/no question. With continuous variables, we generalize the notion to *differential entropy*[9].

Using this framework, we can introduce a notion of similarity between two probability distributions, called the *Kullback-Leibler* divergence; which is not symmetric and thus not a real distance. It is defined as

$$D_{KL}[p||q] = \sum_{x \in X} p(x) \log \frac{1}{q(x)} \mathrm{d}p(x) \tag{13}$$

The most popular measure of *linear dependence* is the correlation coefficient; generalizing the notion of dependence, *mutual information* can be used.

$$I[X:Y] = D_{KL}[p(x,y)||p(x)p(y)] = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \mathrm{d}x\mathrm{d}y \tag{14}$$

For a brief history on the application of information theory to neuroscience, see Dimitrov, Lazar, and Victor (2011); for a review and a new method for efficiently calculating mutual information, see Ince et al. (2016); for an excellent more general introduction, see Stone (2015).

Principal Component Analysis

The goal of principal component analysis is to project the data X onto a new space $Y = PX$ that represents our data using an orthogonal basis with axes along the directions of maximal variance.

Note that a matrix' eigenvector does not change direction when multiplied by the matrix $A\vec{v} = \lambda\vec{v}$. Letting E be a matrix of eigenvectors, and D be a diagonal matrix of

[9] However, note that this makes little sense (only in relative comparisons), and information is defined on a discrete domain; after all, having an infinite amount of digits after the comma, a continuous number is able to transmit an infinite amount of information.

corresponding eigenvalues, we can write

$$AE = ED \tag{15}$$

$$A = EDE^{-1} = EDE^T \tag{16}$$

where we have assumed that A is a symmetric matrix, which implies that E is orthogonal.

$$\begin{aligned}\text{Cov}[Y] &= \text{Cov}[PX] &(17)\\ &= P\text{Cov}[X]P^T &(18)\\ &= E^T\text{Cov}[X]E &(19)\\ &= E^T(EDE^T)E = D &(20)\end{aligned}$$

Thus the eigenvalues of the covariance matrix of the data gives the variance in the projected space. We can reduce the dimensionality of the data by only keeping the first k columns of E, and then projecting $Y = E_k^T X$.

For a good tutorial on PCA, see Shlens (2014b).

Independent Component Analysis

While PCA merely decorrelates the data, ICA removes higher order dependencies such as mutual information and does not require an orthogonal basis. Concretely, the problem is one of blind source separation. We assume that the data arise from a linear mixture of sources, $x = As$, but we neither know A nor s – the problem is not identifiable. We want to find the *unmixing* matrix $W = A^{-1}$ such that $\hat{s} = Wx$. To achieve identifiability, we assume that $\text{Cov}[s] = \text{Cov}[Wx] = \mathbb{I}$. Via a singular value decomposition we see that $A^{-1} = (UDV^T)^{-1} = VD^{-1}U^T$.

Due to our assumption we see that $W = VD^{-1/2}E^T$ because then it holds that

$$\begin{aligned} \mathrm{Cov}[Wx] &= W(EDE^T)W^T && (21) \\ &= VD^{-1/2}E^T(EDE^T)ED^{-1/2}V^T = \mathbb{I} && (22) \end{aligned}$$

The matrix W does two operations; first, it *whitens* the data X such that $\mathrm{Cov}[X] = \mathbb{I}$. Denote the whitened data as $x_w = D^{-1/2}E^T x$ Second, it rotates the data. But how?

We choose V such that the rotated data minimizes the *multi-information* of the sources $\mathbf{s} = [s_1, \ldots, s_n]$. Multi-information generalizes the notion of mutual independence to n probability distributions; minimizing it means maximizing the independence between the probability distributions of the source signals.

$$\begin{aligned} I[\mathbf{s}] &= \int_{s\in\mathcal{S}} p(\mathbf{s}) \log \frac{p(\mathbf{s})}{\prod_s p(s)} \mathrm{d}\mathbf{s} && (23) \\ &= \int_{s\in\mathcal{S}} p(\mathbf{s}) \log p(\mathbf{s}) \mathrm{d}\mathbf{s} - \int_{s\in\mathcal{S}} p(\mathbf{s}) \sum_{i=1}^{n} \log p(s_i) \mathrm{d}\mathbf{s} && (24) \\ &= -h[\mathbf{s}] + \sum_{i=1}^{n} h[s_i] = \sum_{i=1}^{n} h[s_i] - h[\mathbf{s}] && (25) \\ &= \sum_{i=1}^{n} h[(Vx_w)_i] - h[Vx_w] = \sum_{i=1}^{n} h[s_i] - (h[x_w] + \log|V|) && (26) \end{aligned}$$

where the equivalence of $h[Vx_w] = h[x_w] + \log|V|$ is due to a change of variables. Since a rotation matrix is orthogonal, its log determinant is zero. Additionally we can drop $h[x_w]$ because it is a constant, leaving

$$V = \underset{V}{\mathrm{argmin}} \sum_{i=1}^{n} h[(Vx_w)_i] \qquad (27)$$

In total, ICA whitens the data and subsequently rotates it such that the marginal entropies of the source components are minimized. Another way to say this is that ICA tries to find directions that maximize the *non-gaussianity* of the source components, because the Gaussian distribution has the highest entropy (among all distributions with equal variance).

Note that ICA equals PCA for gaussian data because uncorrelatedness implies independence for gaussian random variables – there are no further higher-order dependencies fueling the mutual information machinery. For good tutorials on ICA, see Shlens (2014a) and Hyvärinen and Oja (2000).

Linear Discriminant Analysis

There are different ways to think about linear discriminant analysis. Here I motivate it as a Bayes classifier in a two class classification setting. Given the predictors, choose the class that has the highest posterior probability.

Note the law of total covariance

$$SS_{total} = SS_{within} + SS_{between} \tag{28}$$

$$\text{Cov}[X] = \mathbb{E}[\text{Cov}[X|J]] + \text{Cov}[\mathbb{E}[X|J]] \tag{29}$$

and Bayes' rule, which follows from the *sum rule* and *product rule* of probability

$$P(B) = \sum_A P(B|A)P(A) \tag{30}$$

$$P(A,B) = P(A|B)P(B) = P(B|A)P(A) \tag{31}$$

$$\Rightarrow \underbrace{P(A|B)}_{\text{posterior}} = \frac{\overbrace{P(B|A)}^{\text{likelihood}}\overbrace{P(A)}^{\text{prior}}}{\underbrace{\sum_A P(B|A)P(A)}_{\text{marginal likelihood}}} \tag{32}$$

Using class labels 0 and 1, denote the class conditional covariances as $C_0 = \text{Cov}[X|J=0]$ and $C_1 = \text{Cov}[X|J=1]$, the class conditional means as $\mu_0 = \mathbb{E}[X|J=0]$ and $\mu_1 = \mathbb{E}[X|J=1]$, and the priors for the respective classes as π_0 and π_1. Linear discriminant analysis assumes that $C_0 = C_1$, and that the predictors follow a multivariate normal distribution, $X_j \sim \text{N}(\mu_j, C_j)$ with $j \in \{0,1\}$. Recall its shape

$$f(x^d|\mu, C) = (2\pi)^{-d/2}|C|^{-1/2}exp\left(-\frac{1}{2}(x-\mu)^T C^{-1}(x-\mu)\right) \tag{33}$$

where d is the dimensionality of the features. Turning the Bayesian handle and classifying given a specific feature x results in

$$p(J=0|X=x) = \frac{p(x|J=0)\pi_0}{p(x|J=0)\pi_0 + p(x|J=1)\pi_1} \tag{34}$$

$$p(J=1|X=x) = \frac{p(x|J=1)\pi_1}{p(x|J=0)\pi_0 + p(x|J=1)\pi_1} \tag{35}$$

$$\Rightarrow \frac{p(J=0|X=x)}{p(J=1|X=x)} = \frac{p(x|J=0)\pi_0}{p(x|J=1)\pi_1} \tag{36}$$

Let $p_0 := p(J=0|X=x)$ and $p_1 := p(J=0|X=x)$. Simplifying the problem by taking logs and using $C_0 = C_1$ yields

$$\log\frac{p_0}{p_1} = \log\left((2\pi)^{-d/2}|C|^{-1/2}\right) - \frac{1}{2}(x-\mu_0)^T C^{-1}(x-\mu_0) \tag{37}$$

$$- \log\left((2\pi)^{-d/2}|C|^{-1/2}\right) + \frac{1}{2}(x-\mu_1)^T C^{-1}(x-\mu_1) + \log\frac{\pi_0}{\pi_1} \tag{38}$$

$$= -\frac{1}{2}(x-\mu_0)^T C^{-1}(x-\mu_0) + \frac{1}{2}(x-\mu_1)^T C^{-1}(x-\mu_1) + \log\frac{\pi_0}{\pi_1} \tag{39}$$

$$= x^T C^{-1}u_0 - \frac{1}{2}u_0^T C^{-1}u_0^T - x^T C^{-1}u_1 + \frac{1}{2}u_1^T C^{-1}u_1^T + \log\frac{\pi_0}{\pi_1} \tag{40}$$

$$= x^T C^{-1}(u_0 - u_1) - \frac{1}{2}(u_0 + u_1)^T C^{-1}(u_0 - u_1) + \log\frac{\pi_0}{\pi_1} \tag{41}$$

a function linear in x. If we were to assume unequal covariances across classes, the function would be quadratic in x, and the resulting classifier would be called Quadratic Discriminant Analysis (QDA). Assuming that all features are independent, $p(x_1,\ldots,x_N) = \prod_{i=1}^N p(x_i)$, would result in the naive Bayes classifier.

The most challenging objective with discriminant analysis is to accurately estimate the covariance; commonly, this is done using the unbiased empirical covariance estimator $\hat{\Sigma} = \frac{1}{N-1}\sum_{i=1}^N (x_i - \hat{\mu})(x_i - \hat{\mu})^T$ with $\mu = \frac{1}{n}\sum_{i=1}^N x_i$. However, with a high-dimensional features space as in EEG, this estimation has high variance. For this reason, one regularizes the estimator by shrinking it towards the scalar matrix

$$\hat{\Sigma}_r = (1-\gamma)\hat{\Sigma} + \gamma\hat{\sigma}^2\mathbb{I} \tag{42}$$

with tuning parameter $\gamma \in [0, 1]$ and $\hat{\sigma}^2 = \text{trace}(\hat{\Sigma})/d$. $\gamma = 0$ results in classical linear discriminant analysis; for more, see Blankertz et al. (2011, pp. 818).

Regularization

This provides a slightly different look at regularization. What is the maximum likelihood estimate for θ in a coin flip experiment where we observe $k = 3$ heads in $n = 3$ flips? It is $\theta = k/n = 1$, which is ridiculous. Using Bayesian inference, the problem setup is

$$p(\theta|k, n) = \frac{p(k|\theta, n)p(\theta)}{\int_\theta p(k|\theta, n)p(\theta)\mathrm{d}\theta} \tag{43}$$

Specifying $\theta \sim \text{Beta}(a = 1, b = 1)$ as prior distribution corresponds to a uniform prior over θ. Since the beta distribution is conjugate for the binomial likelihood (our coin toss experiment), we arrive at posterior distribution by simple arithmetic $\theta|k, n \sim \text{Beta}(a + k, b + n - k)$, with a posterior mean of $\theta_{MAP} = \frac{a}{a+b} = .8$, a value less ridiculous. This process of *taming one's estimators* is called *regularization*, and has a natural Bayesian interpretation. For more on Bayesian inference, see Jackman (2009), McElreath (2016) and Etz, Gronau, Dablander, Edelsbrunner, and Baribault (invited submission).

Linear regression

Linear regression is the most important idea in supervised learning, the branch of machine learning that tries to learn a function for predicting an outcome given certain inputs, $y = f(x) + \epsilon$. Linear regression restricts $f(x)$ to be a linear function in the coefficients β, $f(x) = \sum_{i=1}^{n} \langle x_i, \beta \rangle = X\beta$

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} x_{11} & x_{12} & \ldots & x_{1p} \\ x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & \vdots & \ddots & x_{2p} \\ x_{n1} & x_{n2} & \ldots & x_{np} \end{pmatrix} \cdot \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} \tag{44}$$

Linear regression models the outcome conditional on the predictors as a (multivariate) gaussian, $y|X, \beta \sim \mathcal{N}(X\beta, \sigma^2\mathbb{I})$, which can be written as $y = X\beta + \epsilon$ where $\epsilon \sim \mathcal{N}(0, \sigma^2)$. In addition to this *heteroscedasticity* requirement, linear regression assumes *weak exogeneity*, that is, the predictors are not a random variable but fixed, i.e. $\mathbb{E}(X) = X$.

Using a principled loss function based on information theory called *cross entropy* reduces to the familiar sum of squares error function in case of linear regression

$$\begin{aligned} L(\beta) &= \sum_{i=1}^{n} \log \mathcal{N}(y_i, \sigma^2\mathbb{I}|\langle x_i, \beta\rangle) \propto \sum_{i=1}^{n} (y_i - \langle x_i, \beta\rangle)^2 && (45) \\ &= (y - X\beta)^T (y - X\beta) && (46) \end{aligned}$$

Minimizing this expression over β yields a simple closed form solution

$$\begin{aligned} & \frac{\partial}{\partial\beta}(y - X\beta)^T(y - X\beta) \overset{!}{=} 0 && (47) \\ \Leftrightarrow & \frac{\partial}{\partial\beta}\left(y^T y - 2X^T\beta^T y + \beta^T X^T X\beta\right) \overset{!}{=} 0 && (48) \\ \Leftrightarrow & \, 0 - 2X^T X\beta + 2X^T y = 0 && (49) \\ \Leftrightarrow & \, 2X^T X\beta = 2X^T y && (50) \\ \Leftrightarrow & \, \beta = (X^T X)^{-1} X^T y && (51) \end{aligned}$$

Speaking in terms of EEG data, we can generalize this to do regression not just at a single point in time, but across t time points:

$$\begin{pmatrix} y_{11} & y_{12} & \dots & y_{1t} \\ y_{21} & y_{22} & \dots & y_{2t} \\ \vdots & \vdots & \ddots & y_{2t} \\ y_{n1} & y_{n2} & \dots & y_{nt} \end{pmatrix} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & x_{2p} \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix} \cdot \begin{pmatrix} \beta_{11} & \beta_{12} & \dots & \beta_{1t} \\ \beta_{21} & \beta_{22} & \dots & \beta_{2t} \\ \vdots & \vdots & \ddots & \beta_{2p} \\ \beta_{p1} & \beta_{p2} & \dots & \beta_{pt} \end{pmatrix} \qquad (52)$$

Similarly, we adjust the loss function to this multivariate case

$$L(\mathbf{B}) = \sum_{i=1}^{n}\sum_{t=1}^{T}(y_{it} - \langle x_{i\cdot}, \beta_{\cdot t}\rangle)^2 \tag{53}$$

$$= \mathbf{tr}[(\mathbf{Y} - \mathbf{XB})^T(\mathbf{Y} - \mathbf{XB})] \tag{54}$$

which leads to an equally concise solution

$$\hat{\mathbf{B}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \tag{55}$$

Since t-tests and ANOVA and all that are just variants of linear regression with dummy coding, framing them as a regression problem provides more flexible modeling (see main text).

In general, the *maximum-likelihood estimate* is prone to overfit the data, which can be avoided by introducing a penalty on the size of the regression weights (except the coefficient that encodes the intercept).

$$\hat{\beta} = \underset{\beta}{\text{argmin}}\left(\sum_{i=1}^{n}(y_i - \beta_0 - \langle x_i, \beta\rangle)^2\right) \quad \ldots \text{ least squares} \tag{56}$$

$$\hat{\beta} = \underset{\beta}{\text{argmin}}\left(\sum_{i=1}^{n}(y_i - \beta_0 - \langle x_i, \beta\rangle)^2 + \lambda\sum_{j=1}^{p}\beta_j^2\right) \quad \ldots \text{ ridge} \tag{57}$$

$$\hat{\beta} = \underset{\beta}{\text{argmin}}\left(\sum_{i=1}^{n}(y_i - \beta_0 - \langle x_i, \beta\rangle)^2 + \lambda\sum_{j=1}^{p}|\beta_j|\right) \quad \ldots \text{ lasso} \tag{58}$$

λ is a hyper-parameter that is often chosen by cross-validation. While ridge regression shrinks coefficients towards each other, the lasso shrinks coefficients to zero, and thus is extremely powerful in high-dimensional (see Hastie, Tibshirani, & Friedman, 2009, ch. 3). Ridge regression has a closed-form solution, $\hat{\beta} = (X^TX + \lambda\mathbb{I})^{-1}X^Ty$, and was initially proposed so that estimation works even when X is not full-rank. The lasso does not have a closed form solution but is still a convex optimization problem (Tibshirani, 1996).

Note that, from a Bayesian perspective, the ridge solution is given by the *maximum a posteriori* estimate using a normal prior, while the lasso estimate results when choosing a Laplace prior (cf. Park & Casella, 2008).

Linear mixed effects models

There is quite some naming controversy around mixed effects models (for example, see here). The general idea is, in addition to having coefficients that are common across groups, to also include coefficients that vary across groups. Take a simple regression model where N participants indexed by j responded to I trials indexed by i in which words of different lengths were presented, and reaction time measured.

We can distinguish three possible estimation techniques, either a) averaging over all participants and computing a single mean, $y_{ij} \sim \mathcal{N}(\beta_0, \sigma^2)$, termed *pooling*, or b) computing a single mean for each participant seperately, $y_{ij} \sim \mathcal{N}(\beta_{0j}, \sigma^2)$, termed *no pooling*, or, finally, c) estimate a mean for each participant, but regularize these estimates with a prior, $y_{ij} \sim \mathcal{N}(\beta_{0j}, \sigma^2)$, where $\beta_{0j} \sim \mathcal{N}(\mu, \omega^2)$, termed *partial pooling* or hierarchical model. In the latter model, μ denotes the grand participant average, and ω^2 denotes the between group variance. In the hierarchical model, we have specified a model for parameters – the participant means – in addition to our model for the data. We can denote the former as level 2, and the latter as level 1. Estimating both levels within the same model is what I have called "proper" hierarchical modeling in the main text, which yields more accurate estimation because of the regularizing prior over the participant means (Gelman, 2012; Efron & Morris, 1977).

Extending this hierarchical model into a *multilevel* model entails adding predictors at the 2 level. Assume our participants responded to words of different length, and we are interested in how word saliency influences reaction time. Adding reaction time as a *fixed effect* would result in the following model

$$y_{ij} \sim \mathcal{N}(\beta_{0j} + x\beta_1, \sigma^2) \tag{59}$$

$$\beta_{0j} \sim \mathcal{N}(\beta_0, \omega^2) \tag{60}$$

where we still allow the participants mean to vary, but not the effect of word saliency. Extending this to have word saliency as a *random effect* would result in

$$y_{ij} \sim \mathcal{N}(\beta_{0j} + x\beta_{1j}, \sigma^2) \tag{61}$$

$$\mathbf{b}_j \sim \text{MvN}(\mathbf{b}, \Sigma) \tag{62}$$

where $\mathbf{b}_j = (\beta_{0j}, \beta_{1j})$ and Σ denotes the covariance matrix, specifying possible dependencies, i.e. correlation, between mean reaction time and effect of word saliency.

Another terminology, which is usually associated with the classical estimation of these models, extends the usual regression formula by writing the conditional (multivariate) distribution of Y given the realized random effects $\mathcal{B} = \mathbf{b}$

$$Y|\mathbf{b} \sim \mathcal{N}(X\beta + Z\mathbf{b}, \mathbb{I}\sigma^2) \tag{63}$$

where β denote effects that are fixed across units and $\mathbf{b}$ denote effects that vary across unit. X is now the design matrix for the fixed effects, while Z is the design matrix for the random effects.

Casting our previous example in this terminology, and assuming that $\mathcal{B} \sim \text{MvN}(0, \Sigma)$, our design matrices would be

$$X = \begin{pmatrix} 1 & x_{12} \\ \vdots & \vdots \\ 1 & x_{p2} \end{pmatrix}, Z = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \tag{64}$$

where $x_{12}, \ldots, x_{p2}$ denotes the word saliency. Doing the computation becomes clear that the varying or random effects offset the common effects on a participant by participant basis

$$y_{ij} = \begin{pmatrix} 1 & x_{12} \\ \vdots & \vdots \\ 1 & x_{p2} \end{pmatrix} \cdot \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} b_{0j} \\ b_{1j} \end{pmatrix} \tag{65}$$

$$y_{ij} = \beta_0 + x\beta_1 + b_{0j} + b_{1j} \tag{66}$$

These models provide a powerful means of data analysis. For an overview and implementation on how to estimate these using classical statistics, see Bates et al. (2014); for a practical tutorial in the same spirit, see Baayen et al. (2008). For a tutorial on how to estimate these models using Bayesian tools, see Sorensen and Vasishth (2015).

Splines and General additive models

One can model non-linear functions in linear regression by including non-linear transforms of the predictors, relaxing some of the rigidity linearity implies. However, these transformations are *global*, and it might be hard to find the right transformations. Splines are piecewise polynomial functions that are fit to the data and allow increased flexibility. They are connected via knots, but because it is difficult to select how many knots are needed, i.e. how many seperate polynomial functions are needed, *smoothing splines* avoid the knot selection problem by regularizing the "wiggliness" of the function.

Concretely, we optimize the following penalized residual sum of squares expression

$$PRSS = \sum_{i=1}^{N} (y_i - f(x_i))^2 + \lambda \int \{f''(t)\}^2 \mathrm{d}t \tag{67}$$

where the first term measures closeness to the data and the second term penalizes the curvature of the function. If $\lambda = 0$, then we can fit any function to the data, while when $\lambda = 0$, this reduces to least squares (see Hastie et al., 2009, pp. 151). While defined on an infinite dimensional function space comprised of all functions that have a squared second derivative, it turns out that this has a unique minimizer, but showing that is beyond this paper (and, currently, me).

For excellent, hands-on explanations see Kim Larsen's blog post and Austin Rochford's blog post. Jacolien van Rij provides an overview on modeling time series data with GAMMs.

Reader's appendix

This short appendix details things I have found useful in writing this paper. For easy reading, I liked Cohen (2015) for some EEG background, and Lyons and Fugal (2014) for a signal processing starter. Having never taken a class on how to analyze neural time series data, I have found Mike Cohen's book to be an excellent guide (Cohen, 2014a); he also has videos on his webpage, see here. The first few chapters in Luck (2014) provided me with necessary background I lacked on ERP modeling, while the last few chapters detailed the standard approach to statistical analysis; as prices for textbooks go, it is also very cheap. For more on signal processing, I enjoyed parts of Downey (2016) and the beautiful visualizations here; more specifically tuned to EEG, I recommend the first chapters of Freeman and Quiroga (2012). Writing this paper made me realize just how spellbound researchers are by the countless MATLAB toolboxes[10], and that Python is the only real contender in the space of neuroscience and signal processing. Because using open, general purpose tools is important, I recommend McKinney (2012) and the MNE package (Gramfort et al., 2013, 2014).

Further reading:

- Makeig, S., Debener, S., Onton, J., & Delorme, A. (2004). Mining event-related brain dynamics. *Trends in cognitive sciences*, *8*(5), 204–210. doi:10.1016/j.tics.2004.03.008
- Herrmann, C. S., Rach, S., Vosskuhl, J., & Strüber, D. (2014). Time–frequency analysis of event-related potentials: a brief tutorial. *Brain topography*, *27*(4), 438–450. doi:10.1007/s10548-013-0327-5
- Smith, N. J. & Kutas, M. (2015a). Regression-based estimation of ERP waveforms: I. The rERP framework. *Psychophysiology*, *52*(2), 157–168. doi:10.1111/psyp.12317

[10] Yes, I realize that figure 1 was made with MATLAB.

References

Ahmadi, M. & Quiroga, R. Q. (2013). Automatic denoising of single-trial evoked potentials. *NeuroImage*, *66*, 672–680. doi:10.1016/j.neuroimage.2012.10.062

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. doi:10.1016/j.jml.2007.12.005

Bagiella, E., Sloan, R. P., & Heitjan, D. F. (2000). Mixed-effects models in psychophysiology. *Psychophysiology*, *37*(01), 13–20. doi:10.1111/1469-8986.3710013

Başar, E., Başar-Eroglu, C., Karakaş, S., & Schürmann, M. (2001). Gamma, alpha, delta, and theta oscillations govern cognitive processes. *International Journal of Psychophysiology*, *39*(2), 241–248. doi:10.1016/S0167-8760(00)00145-8

Bastiaansen, M., Mazaheri, A., & Jensen, O. (2012). Beyond ERPs: oscillatory neuronal dynamics. In *The oxford handbook of event-related potential components* (pp. 31–50). Oxford University Press.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. *arXiv preprint arXiv:1406.5823*.

Berger, J. O. & Delampady, M. (1987). Testing precise hypotheses. *Statistical Science*, 317–335.

Bigdely-Shamlo, N., Mullen, T., Kothe, C., Su, K.-M., & Robbins, K. A. (2015). The PREP pipeline: standardized preprocessing for large-scale EEG analysis. *Frontiers in Neuroinformatics*, *9*. doi:10.3389/fninf.2015.00016

Blankertz, B., Lemm, S., Treder, M., Haufe, S., & Müller, K.-R. (2011). Single-trial analysis and classification of ERP components—a tutorial. *NeuroImage*, *56*(2), 814–825. doi:10.1016/j.neuroimage.2010.06.048

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. In *Engineering in Medicine and Biology Society (EMBC), 2013*

35th Annual International Conference of the IEEE (pp. 1680–1683). IEEE. doi:10.1109/EMBC.2013.6609841

Buzsáki, G., Anastassiou, C. A., & Koch, C. (2012). The origin of extracellular fields and currents—EEG, ECoG, LFP and spikes. *Nature reviews neuroscience*, *13*(6), 407–420. doi:10.1038/nrn3241

Buzsáki, G., Logothetis, N., & Singer, W. (2013). Scaling brain size, keeping timing: evolutionary preservation of brain rhythms. *Neuron*, *80*(3), 751–764. doi:10.1016/j.neuron.2013.10.002

Carp, J. (2012). The secret lives of experiments: methods reporting in the fMRI literature. *NeuroImage*, *63*(1), 289–300. doi:10.1016/j.neuroimage.2012.07.004

Chambers, C. D. (2013). Registered reports: a new publishing initiative at Cortex. *Cortex*, *49*(3), 609–610. doi:10.1016/j.cortex.2012.12.016

Cohen, M. X. (2014a). *Analyzing neural time series data: theory and practice.* MIT Press.

Cohen, M. X. (2014b). Comparison of different spatial transformations applied to EEG data: A case study of error processing. *International Journal of Psychophysiology*, 245–257. doi:doi:10.1016/j.ijpsycho.2014.09.013

Cohen, M. X. (2015). *Cycles in mind: how brain rhythms control perception and action.* Sinc(x) press.

Cohen, M. X. & Gulbinaite, R. (2014). Five methodological challenges in cognitive electrophysiology. *NeuroImage*, *85*, 702–710. doi:10.1016/j.neuroimage.2013.08.010

Collura, T. F. (1993). History and evolution of electroencephalographic instruments and techniques. *Journal of Clinical Neurophysiology*, *10*(4), 476–504.

da Silva, F. L. (2013). EEG and MEG: relevance to neuroscience. *Neuron*, *80*(5), 1112–1128. doi:10.1016/j.neuron.2013.10.017

De Groot, A. (2014). The meaning of "significance" for different types of research [translated and annotated by Eric-Jan Wagenmakers, Denny Borsboom, Josine Verhagen, Rogier Kievit, Marjan Bakker, Angelique Cramer, Dora Matzke, Don

Mellenbergh, and Han LJ van der Maas]. *Acta Psychologica*, *148*, 188–194. doi:10.1016/j.actpsy.2014.02.001

De Vos, M., Thorne, J. D., Yovel, G., & Debener, S. (2012). Let's face it, from trial to trial: Comparing procedures for N170 single-trial estimation. *NeuroImage*, *63*(3), 1196–1202. doi:10.1016/j.neuroimage.2012.07.055

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.

Delorme, A., Palmer, J., Onton, J., Oostenveld, R., & Makeig, S. (2012). Independent EEG sources are dipolar. *PloS One*, *7*(2), e30135. doi:10.1371/journal.pone.0030135

Delorme, A., Sejnowski, T., & Makeig, S. (2007). Enhanced detection of artifacts in EEG data using higher-order statistics and independent component analysis. *NeuroImage*, *34*(4), 1443–1449. doi:10.1016/j.neuroimage.2006.11.004

Dien, J. (1998). Issues in the application of the average reference: Review, critiques, and recommendations. *Behavior Research Methods, Instruments, & Computers*, *30*(1), 34–43.

Dimitrov, A. G., Lazar, A. A., & Victor, J. D. (2011). Information theory in neuroscience. *Journal of Computational Neuroscience*, *30*(1), 1–5. doi:10.1007/s10827-011-0314-3

Downey, A. B. (2016). *Think DSP: Digital Signal Processing in Python*. Green Tea Press. Retrieved from https://github.com/AllenDowney/ThinkDSP

Efron, B. & Morris, C. (1977). Stein's paradox in statistics. *Scientific American*, *236*, 119–127.

Etz, A., Gronau, Q. F., Dablander, F., Edelsbrunner, P. A., & Baribault, B. (invited submission). How to become a Bayesian in eight easy steps. *Psychonomic Bulletin and Review*.

Forstmann, B. U. & Wagenmakers, E.-J. (2015). *An introduction to model-based cognitive neuroscience*. Springer.

Freeman, W. & Quiroga, R. Q. (2012). *Imaging brain function with EEG: Advanced temporal and spatial analysis of electroencephalographic signals*. Springer Science & Business Media.

Gaspar, C. M., Rousselet, G. A., & Pernet, C. R. (2011). Reliability of ERP and single-trial analyses. *NeuroImage*, *58*(2), 620–629. doi:10.1016/j.neuroimage.2011.06.052

Gelman, A. (2012). Multilevel (hierarchical) modeling: what it can and cannot do. *Technometrics*. doi:10.1198/004017005000000661

Gelman, A. & Loken, E. (2014). The statistical crisis in science. *American Scientist*, *102*(6), 460.

Genovese, C. R., Lazar, N. A., & Nichols, T. (2002). Thresholding of statistical maps in functional neuroimaging using the false discovery rate. *NeuroImage*, *15*(4), 870–878. doi:10.1006/nimg.2001.1037

Giraudet, L., Imbert, J.-P., Bérenger, M., Tremblay, S., & Causse, M. (2015). The neuroergonomic evaluation of human machine interface design in air traffic control using behavioral and EEG/ERP measures. *Behavioural brain research*, *294*, 246–253. doi:10.1016/j.bbr.2015.07.041

Goodin, P., Ciorciari, J., Baker, K., Carrey, A.-M., Harper, M., & Kaufman, J. (2012). A high-density EEG investigation into steady state binaural beat stimulation. *PloS One*, *7*(4), e34789. doi:10.1371/journal.pone.0034789

Gramfort, A., Luessi, M., Larson, E., Engemann, D. A., Strohmeier, D., Brodbeck, C., . . . Parkkonen, L., et al. (2013). MEG and EEG data analysis with MNE-Python. *Frontiers in Neuroscience*, *7*. doi:10.3389/fnins.2013.00267

Gramfort, A., Luessi, M., Larson, E., Engemann, D. A., Strohmeier, D., Brodbeck, C., . . . Hämäläinen, M. S. (2014). MNE software for processing MEG and EEG data. *NeuroImage*, *86*, 446–460. doi:10.1016/j.neuroimage.2013.10.027

Graps, A. (1995). An introduction to wavelets. *Computational Science & Engineering, IEEE*, *2*(2), 50–61. doi:10.1109/99.388960

Groppe, D. M., Urbach, T. P., & Kutas, M. (2011a). Mass univariate analysis of event-related brain potentials/fields I: A critical tutorial review. *Psychophysiology*, *48*(12), 1711–1725. doi:10.1111/j.1469-8986.2011.01273.x

Groppe, D. M., Urbach, T. P., & Kutas, M. (2011b). Mass univariate analysis of event-related brain potentials/fields II: Simulation studies. *Psychophysiology*, *48*(12), 1726–1737. doi:10.1111/j.1469-8986.2011.01272.x

Haegens, S., Cousijn, H., Wallis, G., Harrison, P. J., & Nobre, A. C. (2014). Inter-and intra-individual variability in alpha peak frequency. *NeuroImage*, *92*, 46–55. doi:10.1016/j.neuroimage.2014.01.049

Hastie, T. J., Tibshirani, R. J., & Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer.

Hawkins, G., Mittner, M., Boekel, W., Heathcote, A., & Forstmann, B. (2015). Toward a model-based cognitive neuroscience of mind wandering. *Neuroscience*, *310*, 290–305. doi:10.1016/j.neuroscience.2015.09.053

Herrmann, C. S., Rach, S., Vosskuhl, J., & Strüber, D. (2014). Time–frequency analysis of event-related potentials: a brief tutorial. *Brain topography*, *27*(4), 438–450. doi:10.1007/s10548-013-0327-5

Hu, L., Xiao, P., Zhang, Z., Mouraux, A., & Iannetti, G. D. (2014). Single-trial time–frequency analysis of electrocortical signals: Baseline correction and beyond. *NeuroImage*, *84*, 876–887. doi:10.1016/j.neuroimage.2013.09.055

Hyvärinen, A. & Oja, E. (2000). Independent component analysis: algorithms and applications. *Neural Networks*, *13*(4), 411–430.

Ince, R. A., Giordano, B. L., Kayser, C., Rousselet, G. A., Gross, J., & Schyns, P. G. (2016). A statistical framework for neuroimaging data analysis based on mutual information estimated via a Gaussian copula. *bioRxiv*, 1–53. doi:/10.1101/043745

Jackman, S. (2009). *Bayesian analysis for the social sciences*. John Wiley & Sons.

Jatoi, M. A., Kamel, N., Malik, A. S., Faye, I., & Begum, T. (2014). A survey of methods used for source localization using EEG signals. *Biomedical Signal Processing and Control*, *11*, 42–52. doi:10.1016/j.bspc.2014.01.009

Judd, C. M., Westfall, J., & Kenny, D. A. (2012). Treating stimuli as a random factor in social psychology: a new and comprehensive solution to a pervasive but largely ignored problem. *Journal of personality and social psychology*, *103*(1), 54. doi:10.1037/a0028347

Kappenman, E. S. & Luck, S. J. (2010). The effects of electrode impedance on data quality and statistical significance in ERP recordings. *Psychophysiology*, *47*(5), 888–904. doi:10.1111/j.1469-8986.2010.01009.x

King, M., Dablander, F., Jakob, L., Agan, M., Huber, F., Haslbeck, J., & Brecht, K. (2016). Registered reports for student research. *Journal of European Psychology Students*, *7*(1). doi:doi.org/10.5334/jeps.401

Kriegeskorte, N., Simmons, W. K., Bellgowan, P. S., & Baker, C. I. (2009). Circular analysis in systems neuroscience: the dangers of double dipping. *Nature Neuroscience*, *12*(5), 535–540. doi:10.1038/nn.2303

Lemm, S., Blankertz, B., Dickhaus, T., & Müller, K.-R. (2011). Introduction to machine learning for brain imaging. *NeuroImage*, *56*(2), 387–399. doi:10.1016/j.neuroimage.2010.11.004

Liang, F., Paulo, R., Molina, G., Clyde, M. A., & Berger, J. O. (2008). Mixtures of g priors for bayesian variable selection. *Journal of the American Statistical Association*. doi:10.1198/016214507000001337

Lotte, F., Congedo, M., Lécuyer, A., Lamarche, F., & Arnaldi, B. (2007). A review of classification algorithms for EEG-based brain–computer interfaces. *Journal of Neural Engineering*, *4*(2), R1.

Luck, S. J. (2014). *An introduction to the event-related potential technique*. MIT press.

Luck, S. J. & Gaspelin, N. (in press). How to Get Statistically Significant Effects in Any ERP Experiment (and Why You Shouldn't). *Psychophysiology*.

Luck, S. J. & Kappenman, E. S. (2011). *The oxford handbook of event-related potential components*. Oxford university press.

Lyons, R. G. & Fugal, D. L. (2014). *The essential guide to digital signal processing*. Pearson Education.

Makeig, S., Debener, S., Onton, J., & Delorme, A. (2004). Mining event-related brain dynamics. *Trends in cognitive sciences*, *8*(5), 204–210. doi:10.1016/j.tics.2004.03.008

Maris, E. & Oostenveld, R. (2007). Nonparametric statistical testing of EEG-and MEG-data. *Journal of Neuroscience Methods*, *164*(1), 177–190. doi:10.1016/j.jneumeth.2007.03.024

McElreath, R. (2016). *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. CRC Press.

McKinney, W. (2012). *Python for data analysis: Data wrangling with Pandas, NumPy, and IPython*. O'Reilly Media, Inc.

Meulman, N., Wieling, M., Sprenger, S. A., Stowe, L. A., & Schmid, M. S. (2015). Age effects in L2 grammar processing as revealed by ERPs and how (not) to study them. *PloS One*, *10*(12). doi:10.1371/journal.pone.0143328

Monti, M. M. (2011). Statistical analysis of fMRI time-series: A critical review of the GLM approach. *Frontiers in Human Neuroscience*, *5*, 28. doi:10.3389/fnhum.2011.00028

Mullen, T. R., Kothe, C. A., Chi, Y. M., Ojeda, A., Kerth, T., Makeig, S., . . . Cauwenberghs, G. (2015). Real-time neuroimaging and cognitive monitoring using wearable dry EEG. *Biomedical Engineering, IEEE Transactions on*, *62*(11), 2553–2567. doi:10.1109/TBME.2015.2481482

Müller, K.-R., Anderson, C. W., & Birch, G. E. (2003). Linear and nonlinear methods for brain-computer interfaces. *Neural Systems and Rehabilitation Engineering, IEEE Transactions on*, *11*(2), 165–169. doi:10.1109/TNSRE.2003.814484

Müller, K.-R., Tangermann, M., Dornhege, G., Krauledat, M., Curio, G., & Blankertz, B. (2008). Machine learning for real-time single-trial EEG-analysis: from brain–computer interfacing to mental state monitoring. *Journal of Neuroscience Methods*, *167*(1), 82–90. doi:10.1016/j.jneumeth.2007.09.022

Park, T. & Casella, G. (2008). The Bayesian lasso. *Journal of the American Statistical Association*, *103*(482), 681–686. doi:10.1198/016214508000000337

Pernet, C. R., Chauveau, N., Gaspar, C., & Rousselet, G. A. (2011). LIMO EEG: a toolbox for hierarchical LInear MOdeling of ElectroEncephaloGraphic data. *Computational Intelligence and Neuroscience*, *2011*, 3. doi:10.1155/2011/831409

Pernet, C. R., Sajda, P., & Rousselet, G. A. (2011). Single-trial analyses: why bother? *Frontiers in Psychology*, *2*. doi:10.3389/fpsyg.2011.00322

Plöchl, M., Ossandón, J. P., & König, P. (2012). Combining EEG and eye tracking: identification, characterization, and correction of eye movement artifacts in electroencephalographic data. *Frontiers in Human Neuroscience*, *6*. doi:10.3389/fnhum.2012.00278

Poline, J.-B. & Brett, M. (2012). The general linear model and fMRI: does love last forever? *NeuroImage*, *62*(2), 871–880. doi:10.1016/j.neuroimage.2012.01.133

Quiroga, R. Q., Atienza, M., Cantero, J., & Jongsma, M. (2007). What can we learn from single-trial event-related potentials? *Chaos Complex Lett. 2*, 345–363.

Quiroga, R. Q. & Garcia, H. (2003). Single-trial event-related potentials with wavelet denoising. *Clinical Neurophysiology*, *114*(2), 376–390. doi:10.1016/S1388-2457(02)00365-6

Rey, H. G., Ahmadi, M., & Quiroga, R. Q. (2015). Single trial analysis of field potentials in perception, learning and memory. *Current opinion in neurobiology*, *31*, 148–155. doi:10.1016/j.conb.2014.10.009

Ries, A. J., Touryan, J., Vettel, J., McDowell, K., & Hairston, W. D. (2014). A comparison of electroencephalography signals acquired from conventional and mobile systems. *Journal of Neuroscience and Neuroengineering*, *3*(1), 10–20. doi:10.1166/jnsne.2014.1092

Roach, B. J. & Mathalon, D. H. (2008). Event-related EEG time-frequency analysis: an overview of measures and an analysis of early gamma band phase locking in schizophrenia. *Schizophrenia bulletin*, *34*(5), 907–926. doi:10.1093/schbul/sbn093

Rorden, C., Karnath, H.-O., & Bonilha, L. (2007). Improving lesion-symptom mapping. *Journal of Cognitive Neuroscience*, *19*(7), 1081–1088. doi:10.1162/jocn.2007.19.7.1081

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*(2), 225–237. doi:10.3758/PBR.16.2.225

Rousselet, G. A. & Pernet, C. R. (2012). Improving standards in brain-behavior correlation analyses. *Frontiers in Human Neuroscience*, *6*, 119. doi:10.3389/fnhum.2012.00119

Schalk, G., McFarland, D. J., Hinterberger, T., Birbaumer, N., & Wolpaw, J. R. (2004). BCI2000: a general-purpose brain-computer interface (BCI) system. *Biomedical Engineering, IEEE Transactions on*, *51*(6), 1034–1043. doi:10.1109/TBME.2004.827072

Scheer, M., Bülthoff, H. H., & Chuang, L. L. (2016). Steering Demands Diminish the Early-P3, Late-P3 and RON Components of the Event-Related Potential of Task-Irrelevant Environmental Sounds. *Frontiers in Human Neuroscience*, *10*. doi:10.3389/fnhum.2016.00073

Shlens, J. (2014a). A tutorial on independent component analysis. *arXiv preprint arXiv:1404.2986*.

Shlens, J. (2014b). A tutorial on principal component analysis. *arXiv preprint arXiv:1404.1100*.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 1359–1366. doi:10.1177/0956797611417632

Smith, N. J. & Kutas, M. (2015a). Regression-based estimation of ERP waveforms: I. The rERP framework. *Psychophysiology*, *52*(2), 157–168. doi:10.1111/psyp.12317

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. doi:10.1111/psyp.12320

Sorensen, T. & Vasishth, S. (2015). Bayesian linear mixed models using Stan: A tutorial for psychologists, linguists, and cognitive scientists. *arXiv preprint* *arXiv:1506.06201*. Retrieved from https://github.com/vasishth/BayesLMMTutorial

Stelzer, J., Lohmann, G., Mueller, K., Buschmann, T., & Turner, R. (2014). Deficient approaches to human neuroimaging. *Frontiers in Human Neuroscience*, *8*. doi:10.3389/fnhum.2014.00462

Stone, J. V. (2015). *Information theory: a tutorial introduction*. Sebtel Press.

Strang, G. (1994). Wavelets. *American Scientist*, *82*(3), 250–255.

Tanner, D., Morgan-Short, K., & Luck, S. J. (2015). How inappropriate high-pass filters can produce artifactual effects and incorrect conclusions in ERP studies of language and cognition. *Psychophysiology*, *52*(8), 997–1009. doi:10.1111/psyp.12437

Tibon, R. & Levy, D. A. (2015). Striking a balance: analyzing unbalanced event-related potential data. *Frontiers in Psychology*, *6*. doi:10.3389/fpsyg.2015.00555

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 267–288.

Treder, M. S., Porbadnigk, A. K., Avarvand, F. S., Müller, K.-R., & Blankertz, B. (2016). The LDA beamformer: Optimal estimation of ERP source time series using linear discriminant analysis. *NeuroImage*. doi:doi:10.1016/j.neuroimage.2016.01.019

Tremblay, A. & Newman, A. J. (2015). Modeling nonlinear relationships in ERP data using mixed-effects regression with R examples. *Psychophysiology*, *52*(1), 124–139. doi:10.1111/psyp.12299

Turner, B. M., Forstmann, B. U., Love, B. C., Palmeri, T. J., & Van Maanen, L. (2016). Approaches to analysis in model-based cognitive neuroscience. *Journal of Mathematical Psychology*. doi:10.1016/j.jmp.2016.01.001

Van Vliet, M., Chumerin, N., De Deyne, S., Wiersema, J. R., Fias, W., Storms, G., & Van Hulle, M. M. (2016). Single-trial ERP component analysis using a spatiotemporal LCMV Beamformer. *Biomedical Engineering, IEEE Transactions on*, *63*(1), 55–66. doi:10.1109/TBME.2015.2468588

Vul, E., Harris, C., Winkielman, P., & Pashler, H. (2009). Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition. *Perspectives on Psychological Science*, *4*(3), 274–290. doi:10.1111/j.1745-6924.2009.01125.x

Wacker, M. & Witte, H. (2013). Time-frequency techniques in biomedical signal analysis. *Methods of information in medicine*, *52*(4), 279–296. doi:10.3414/ME12-01-0083

Wilcox, R. R. (2012). *Introduction to robust estimation and hypothesis testing.* Academic Press.

Woldorff, M. G. (1993). Distortion of ERP averages due to overlap from temporally adjacent ERPs: analysis and correction. *Psychophysiology*, *30*(1), 98–119. doi:10.1111/j.1469-8986.1993.tb03209.x

Wood, S. (2006). *Generalized additive models: an introduction with R.* CRC press.