Unmixed: Linear Mixed Models combined with Overlap Correction for M/EEG analyses. An Extension to the unfold Toolbox

Benedikt V. Ehinger (b.ehinger@donders.ru.nl)

The Donders Institute for Brain, Cognition and Behaviour
Radboud University, Nijmegen, Netherlands

Abstract:

Linear mixed models (LMMs) offer several benefits over traditional two-stage analysis methods common in EEG analysis: Higher power to detect effects, partial pooling with noisy data and the possibility to account for both subject and item effects. LMMs come at the price of increased computational cost, up to now making them incompatible to use in natural experiments that require time-resolved deconvolution methods of continuous EEG data. Here, I present *unmixed* an extension to the open source unfold-toolbox, allowing to fit LMMs and GAMMs to rERP (regression ERPs) using extended Wilkinson formulas. *Unmixed* supports mixed modelling of overlapping events and non-linear effects. It offers several different optimizers, Walds t-tests and likelihood ratio model comparison tests for statistical analysis, and Benjamini-Hochberg FDR for multiple comparison correction. This technique is promising for population where extensive data collection is not possible, e.g. infants or clinical populations.

Keywords: linear mixed modelling; regression ERPs; non-linear effects; EEG analysis; overlap correction

Introduction

Overlap and non-linear effects in EEG experiments Neuroimaging, and EEG in particular, is increasingly moving from simple designs towards complex situations, e.g. measuring during eye-movements as in reading (Alday, 2019), with continuous stimuli as in speech and movies, or with multi-modal stimulations as in navigation tasks (Ehinger et al., 2014). As a consequence, brain responses to consecutive events naturally overlap in time. Additionally, such designs often include parametric variables, for instance sound amplitude, stimulus contrast, walking speed or saccade amplitude. Commonly these variables are analyzed as linear effects on the ERP (or BOLD). However, many of these effects are actually non-linear and should be modelled as such (e.g. Dimigen & Ehinger, 2019; Rousselet, 2010; Tremblay & Newman, 2015).

Unfold – an open source toolbox for deconvolution and non-linear modeling Recently, we introduced open source software to analyze overlapping signals generated by subsequent events, and non-linear effects using spline regression (www.unfoldtoolbox.org, Ehinger & Dimigen, 2019). The toolbox uses linear models to calculate single subject estimates of deconvolved brain responses. In short, continuous EEG samples are modeled as the sum of temporally overlapping, but to-be-estimated ERPs (one for each condition/event). Such responses can be separated when the overlap between consecutive events differs. This is most often the case, e.g. due to differing reaction times, fixation durations or experimentally jittered inter-trial-intervals. This approach was recently described to be difficult to combine with linear mixed modeling due to computing resources (Ehinger & Dimigen, 2019; Sassenhagen, 2019).

Averaging, two-stage statistics and mixed models Due to the noisy nature of measurement, most, if not all, EEG studies need some kind of repetition of measurements within subjects to gain statistical power via averaging. Subsequently, analyses need to take such non-independences into account. Classically, a two-stage approach is used, where a single value per subject is calculated (through either averaging or regression) and subsequently these values are analyzed using paired t-tests or rANOVAs. Critically, this does not allow for uncertainties on the subject level to propagate to the group level.

Complementing this classical approach, linear mixed-effects modelling (LMM, e.g. Gelman & Hill, 2007) is becoming a popular statistical choice for hierarchical (longitudinal) data. It offers more specification flexibility compared to traditional rANOVAs and allow to control for data-dependencies within a subject. They also allow to include continuous predictors, propagate uncertainty between levels and allow for different number of trials (Baayen, Davidson, & Bates, 2008).

In this short paper I introduce a toolbox to use LMMs compatible with overlap correction and non-linear modeling of continuous EEG data.

Unmixed toolbox

The section below is written in a tutorial style describing the toolbox. Mathematical details on mixed models can be found in e.g. Gelman & Hill (2007). The functions of the toolbox have unit-tests to verify their correctness. The toolbox can is publicly available at https://github.com/unfoldtoolbox/unmixed and continuously developed.

Continuous EEGs with events

Specify model & generate design matrix
um_designmat()

Time-expand design matrix
um_timeexpandDesignmat()

Solve regression model (y=Xb+Zu+e)
uf_mmfit()

Extract betas & random effects
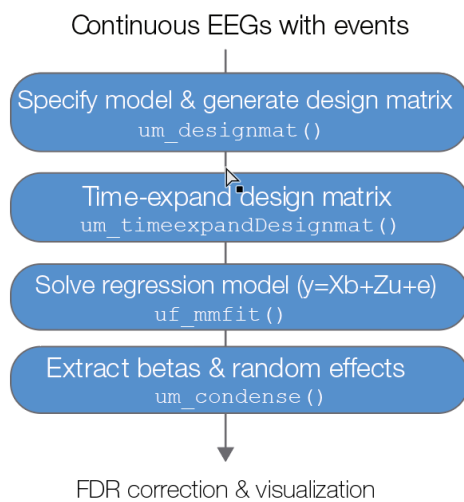um_condense()

FDR correction & visualization

Figure 1: Flowchart of the *unmixed* toolbox. The functions are analogous to the *unfold* toolbox

Model set up In the unmixed toolbox, *um_designmat* allows to specify a separate linear mixed model for each event type. The *um_designmat* function constructs the design matrix X and prepares the random effect design matrices.

In the case of mixed models, the specification of the design matrix is made by following the extended Wilkinson formulation (lme4, matlab, statsmodels). As an example, take the following formula: *rERP ~ A + cat(B) + spl(C,5) + (1+A||subject) + (1|item)* In this specification we would model the following parameters. *Fixed effects*: A will be modeled as a continuous variable, cat(B) as a categorical effect (with automatic expansion to the number of levels, e.g. using effect coding), and C as a non-linear effect using five b-splines. *Random Effects*: We have two grouping variables, one indicating repeated measurements

within subjects (multiple trials per subject) and one indicating repeated items (one item is shown multiple times throughout all experiments). In addition, we allow the intercept and the continuous variable to vary between subjects. That is, we do not assume each subject to be affected by A in the same way. Finally, we assume that the correlation between the random coefficients 1 ("intercept") and A is equal to 0 (as indicated by the || notation).

Time expansion for overlap correction The time expansion of the design matrix X to the continuous EEG design matrix X_{dc} is performed in *um_timeexpandDesignmat*. For this, the time-limits of the events (which define the resulting rERP-length) need to be specified. The toolbox will construct one such matrix for the fixed effects and one matrix for each of the random groupings. The random effects structure I use can be understood in the following lme4 syntax:

$$Y \sim 1 + A_0 + \cdots + A + (1 + A_0|sub_0) + \cdots + (1 + A_\tau|sub_\tau)$$

where the one represents an intercept, τ is the number of timelags and the brackets indicate three random effects within one of the random grouping variables (intercept, slope and correlation). Therefore, we have *n* random effect groupings to be estimated by $\sum_{i=1}^{N_{sub}} N_{samples,i}$ datapoints.

Fitting the model The model is fitted using the *um_mmfit* function. Internally, the MatLab statistics toolbox is used to fit the linear mixed model. Different optimizers are. MatLab offers (and recommends) a quasi-Newton solver. It also offers solving using fminunc and fminsearch. In addition, *unmixed* offers to use the bobyqa solver (Ulrich Römer, 2019), which currently seems to be one of the most efficient optimizer for mixed models. It is still unclear whether the MatLab statistics toolbox implementation of mixed models benefits from the bobyqa solver as much as the implementations in e.g. lme4 (R) or Mixed.Models (Julia).
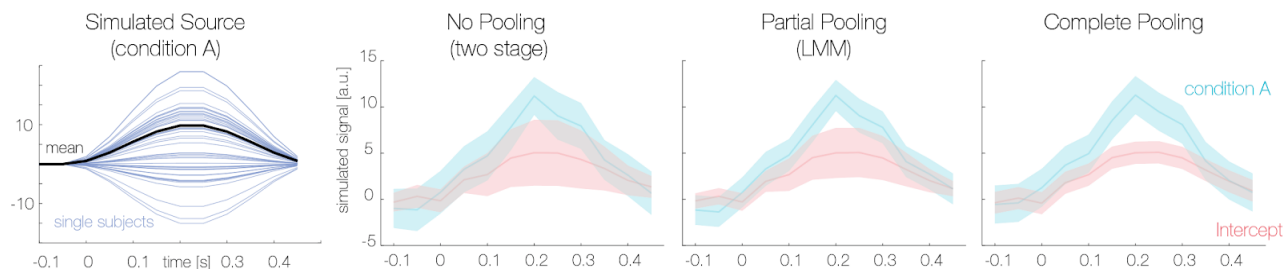
Figure 2. Simulation results. The first panel shows initial simulated sources for each subject and mean, without overlap (hanning window) and without any noise (note the different scale). The second panel shows per-subject linear model fit. The third panel shows the combined LMM over all data. The fourth panel shows one large linear model without taking into account subject effects (and thus with tight uncertainty intervals). The ribbons show 95% confidence intervals around the mean. The benefit of LMMs over the two stage model are readily visible.

Application to Simulated Data

Simulation In the following, I will present a toy simulation. The data were simulated with a 20Hz sampling rate. I simulated 50 subjects, where each subject saw ten different stimuli randomly chosen (with repetitions) for on total ten events in an overlapping regime based on a uniform distribution of inter-event distances. I use a hanning window over 0.5s as the signal. I added an intercept (5µV + noise over subjects with sd = 10µV) and a main effect (10µV + sd = 2µV). Random white white noise with amplitude 20µV (+ sd = 10µV) was added as well. The resulting signal was projected to the scalp from five randomly placed sources in the brain.

Analysis scripts The data were analyzed using the unmixed toolbox. The following script contains the complete analysis

```
cfgDesign = struct()
cfgDesign.inputGroupingName='subject';
cfgDesign.eventtypes= 'sim';
cfgDesign.formula= 'rERP~1+cat(A)+
                          (1+cat(A)|subject)';
cfgDesign.codingschema = 'effects';

EEG = um_designmat(input,cfgDesign);

EEG= um_timeexpandDesignmat(EEG,'timelimits',
                          [-0.1,0.5]);

EEG = um_mmfit(EEG,input,'channel',1,
    'optimizer','bobyqa','covariance',Diagonal);

umresult = um_condense(EEG);
```

Results

In Figure 2A we see one of the simulated ERPs without overlap. In Figure 2B-D we see the resulting rERP from a two-stage procedure (no pooling) from the mixed model (partial pooling) and from a standard linear regression (complete pooling). Due to the linear nature of the model, the fixed effect estimates are identical for all models, but the uncertainty (here 95% confidence intervals) is quite different. The width of the confidence intervals of the mixed models are in between the no pooling and the complete pooling estimates. This is to be expected by the LMM.
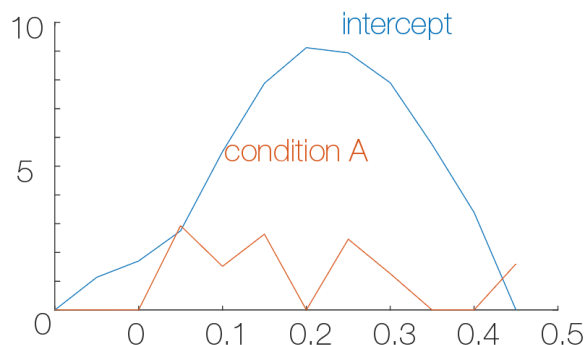

Figure 3: Random effect variance over time. The intercept random effect was simulated with a hanning window with amplitude 10, condition A with an amplitude of 1.

The time-resolved linear mixed model fitting also allows to plot estimated random effect variances over time as in Figure X. The underlying simulation variances (sd = 10 and 1) are partially recovered. Small variances Further improvements in the optimizer

During my simulations I, anecdotally, noticed that for other simulation regimes (e.g. many more trials and less subjects), the LMM quickly approaches the two-stage solution. I will continue investigating this by simulating more realistic EEG data using the SEREEGA toolbox (Krol et al., 2018) and apply the toolbox to real datasets.

Discussion

In this short paper, I presented a new open source toolbox to model continuous EEG data. It is still under development but it readily allows for overlap correction and non-linear effect estimation (not shown here) within linear mixed models.

Applications The applications of this toolbox are manifold and are born from the benefits of using overlap corrected mixed models over two-stage procedures. One advantage is partial pooling, which is beneficial for datasets where it is not possible to collect many repetitions. This is the case e.g. in clinical population or infant work. Especially the latter benefits from the overlap modelling, as it allows to collect more (but overlapping) data in shorter time, and then use the mixed model to partially pool the noisy estimates of all participants in one model. The other main benefit is accounting for item effects. Item effects are typically discussed in linguistics, as words tend to vary in many uncontrollable characteristics, even if the main differences, e.g. frequency, were carefully controlled. The "Language as fixed fallacy" (and their initial proposed F1/F2 ANOVA solution) describing this problem dates back nearly 50 years (Clark, 1973). There is no reason to believe item effects would not occur in EEG, therefore, in order to employ more natural experiments the overlap controlled mixed modeling needs to be used. Yet another, technical, application is to use the mixed model as a regularization tool. The mixed model can be used to circumvent the need to specify the number of splines for non-linear spline regression by including the splines as a random effect grouping variable (Wood, 2017).

Limitations There are some limitations in the implementation that I will improve over the next months. The main limitation is the optimizer speed and convergence. For realistic data, the model can take days to converge (as likelihood function evaluations can become quite slow with more data or more complex random effects structures). One solution I am working on is to leverage the much faster mixed model implementation of Julia.

Acknowledgments

References

Alday, P. M. (2019). *M/EEG analysis of naturalistic stories: a review from speech to language processing*. Language, Cognition and Neuroscience,

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.

Clark, H. H. (1973). *The language-as-fixed-effect fallacy: A critique of language statistics in psychological research.* Journal of Verbal Learning and Verbal Behavior

Dimigen, O., & Ehinger, B. V. (2019). *Analyzing combined EEG and eye-tracking data: The (non-) linear deconvolution framework.* in preparation.

Ehinger, B. V., & Dimigen, O. (2019). *Unfold: An integrated toolbox for overlap correction, non-linear modeling, and regression-based EEG analysis.* biorXiv

Ehinger, B. V., Fischer, P., Gert, A. L., Kaufhold, L., Weber, F., Pipa, G., & König, P. (2014). *Kinesthetic and vestibular information modulate alpha activity during spatial navigation: a mobile EEG study.* Frontiers in Human Neuroscience

Krol, L. R., Pawlitzki, J., Lotte, F., Gramann, K., & Zander, T. O. (2018). *SEREEGA: Simulating event-related EEG activity.* Journal of Neuroscience Methods,

Rousselet. (2010). *Healthy aging delays scalp EEG sensitivity to noise in a face discrimination task.* Frontiers in Psychology.

Sassenhagen, J. (2019). *How to analyse electrophysiological responses to naturalistic language with time-resolved multiple regression.* Language, Cognition and Neuroscience

Tremblay, A., & Newman, A. J. (2015). *Modeling nonlinear relationships in ERP data using mixed-effects regression with R examples: Modeling using mixed-effects regression.* Psychophysiology,

Wood, S. N. (2017). *Generalized additive models : an introduction with R.* Texts in statistical science