

Precise temporal localisation of M/EEG effects with Bayesian generalised additive multilevel models

Ladislav Nalborczyk¹ and Paul Bürkner²

¹Aix Marseille Univ, CNRS, LPL

²TU Dortmund University, Department of Statistics

Abstract

Time-resolved electrophysiological measurements such as those offered by magneto- or electro-encephalography (M/EEG) provide a unique window onto neural activity underlying cognitive processes. Typically, researchers are interested in testing whether and when such measures differ across conditions and/or groups. The conventional approach consists in conducting mass-univariate statistics through time followed by some form of multiplicity correction (e.g., FDR, FWER) or cluster-based inference. However, while allowing efficient error-rates control, cluster-based methods have an important downside: they shift the focus of inference from the timepoint to the cluster level, thus preventing any conclusion to be made about the precise onset or offset of effects (e.g., differences across conditions or groups). Here, we introduce a *model-based* approach for analysing M/EEG timeseries such as ERPs or decoding performance through time, and their differences across conditions or groups. This approach relies on Bayesian generalised additive multilevel models and outputs the posterior probability of the effect being above 0 (or above chance) at every timestep, while naturally taking into account the temporal dependencies and between-subject variability present in such data. Using both simulated and actual M/EEG data, we show that the proposed approach largely outperforms conventional methods in estimating the onset and offset of M/EEG effects, producing more precise and more reliable estimates. We provide an R package implementing the approach and illustrate how to integrate it into M/EEG statistical pipelines in MNE-Python.

Keywords: EEG, MEG, generalised additive models, mixed-effects models, multilevel models, Bayesian statistics, brms

Table of contents


Introduction

3

Introduction

3

Ladislav Nalborczyk  <https://orcid.org/0000-0002-7419-9855>

Paul Bürkner  <https://orcid.org/0000-0001-5765-8995>

The authors have no conflicts of interest to disclose.

Correspondence concerning this article should be addressed to Ladislav Nalborczyk, Aix Marseille Univ, CNRS, LPL, 5 avenue Pasteur, 13100 Aix-en-Provence, France, email: ladislav.nalborczyk@cnrs.fr

11	Problem statement	3
12	Previous work	3
13	Bayesian regression modelling	3
14	Generalised additive models	3
15	Bayesian generalised additive multilevel models	4
16	Objectives	4
17	Methods	4
18	M/EEG data simulation	4
19	Model fitting	5
20	Posterior probability of difference above 0	6
21	Multilevel modelling using ERP summary statistics	7
22	Error properties of the proposed approach	9
23	Comparing the identified onsets/offsets to other approaches	10
24	Simulation study	11
25	Application to actual MEG data	11
26	Results	12
27	Simulation results (bias and variance)	12
28	Application to actual MEG data (reliability)	13
29	Discussion	15
30	Summary of the proposed approach	15
31	Increasing potential usage	15
32	Limitations and future directions	15
33	Conclusions	15
34	Data and code availability	16
35	Packages	16
36	References	17
37	Application to 2D time-resolved decoding results (cross-temporal generalisation)	22
38	Mathematical formulation of the bivariate GAM	24
39	Integration with MNE-Python	25

Precise temporal localisation of M/EEG effects with Bayesian generalised additive multilevel models

Introduction

Here are some useful references to be discussed (Combrisson & Jerbi, 2015; Ehinger & Dimigen, 2019; Frossard & Renaud, 2021, 2022; Gramfort, 2013; Hayasaka, 2003; Luck & Gaspelin, 2017; Maris & Oostenveld, 2007; E. J. Pedersen et al., 2019; Pernet et al., 2015)... See also Maris (2011) and Rosenblatt et al. (2018) (history of cluster-based approaches)... Clusters failures (Eklund et al., 2016)...

Problem statement

Description of cluster-based approaches (see Sassenhagen & Draschkow, 2019)... As pointed by the original authors themselves (Maris & Oostenveld, 2007), “there is a conflict between this interest in localized effects and our choice for a global null hypothesis: by controlling the FA [false alarm] rate under this global null hypothesis, one cannot quantify the uncertainty in the spatiotemporal localization of the effect” (Maris & Oostenveld, 2007; Sassenhagen & Draschkow, 2019)... In contrast, the proposed approach, based on Bayesian generalised additive multilevel models, allows quantifying the probability of effects at the level of timesteps, voxels, sensors (etc), while naturally taking into account spatiotemporal dependencies.

Previous work

Recent example of GLM for EEG (Fischer & Ullsperger, 2013; Wüllhorst et al., 2025)... See also (Hauk et al., 2006; Rousselet et al., 2008)... Example of two-stage regression analysis (i.e., individual-level then group-level, Dunagan et al., 2024)...

See also the rERP framework (N. J. Smith & Kutas, 2014a, 2014b) and Tremblay & Newman (2014)...

From Dimigen & Ehinger (2021): Recently, spline regression has been applied to ERPs (e.g., Hendrix et al., 2017; Kryuchkova et al., 2012)... GAMMs for EEG data (Abugaber et al., 2023; Meulman et al., 2015, 2023)...

Disentangling overlapping processes (Skukies et al., 2024; Skukies & Ehinger, 2021)... Weighting single trials (Pernet, 2022)... The LIMO toolbox (Pernet et al., 2011)...

Recently, Teichmann (2022) provided a detailed tutorial on using Bayes factors (BFs) to analyse the 1D or 2D output from MVPA, that is, for testing, at every timestep, whether decoding performance is above chance level. However, this approach provides timeseries of BFs that ignores temporal dependencies..

Bayesian regression modelling

Short intro/recap about Bayesian (linear and generalised) regression models...

Generalised additive models

See for instance these tutorials (Sóskuthy, 2017; Winter & Wieling, 2016) or application to phonetic data (Sóskuthy, 2021; Wieling, 2018) or this introduction (Baayen & Linke, 2020) or these reference books (Hastie & Tibshirani, 2017; Wood, 2017a)... application to pupillometry (Rij et al., 2019)... GAMLSS for neuroimaging data (Dinga et al., 2021)... Modelling autocorrelation in GAMMs + EEG example (Baayen et al., 2018)...

In generalised additive models (GAMs), the functional relationship between predictors and response variable is decomposed into a sum of low-dimensional non-parametric functions. A typical GAM has the following form:

$$y_i \sim \text{EF}(\mu_i, \phi)$$

$$g(\mu_i) = \underbrace{\mathbf{A}_i \gamma}_{\text{parametric part}} + \underbrace{\sum_{j=1}^J f_j(x_{ij})}_{\text{non-parametric part}}$$

where $y_i \sim \text{EF}(\mu_i, \phi)$ denotes that the observations y_i are distributed as some member of the exponential family of distributions (e.g., Gaussian, Gamma, Beta, Poisson) with mean μ_i and scale parameter ϕ ; $g(\cdot)$ is the link function, \mathbf{A}_i is the i th row of a known parametric model matrix, γ is a vector of parameters for the parametric terms (to be estimated), f_j is a smooth function of covariate x_j (to be estimated as well). The smooth functions f_j are represented in the model via penalised splines basis expansions of the covariates, that are a weighted sum of basis functions:

$$f_j(x_{ij}) = \sum_{k=1}^K \beta_{jk} b_{jk}(x_{ij})$$

where β_{jk} is the weight (coefficient) associated with the k th basis function $b_{jk}()$ evaluated at the covariate value x_{ij} for the j th smooth function f_j . Splines coefficients are penalised (usually through the squared of the smooth functions' second derivative) in a way that can be interpreted as a prior on the “wiggleness” of the function...

Terminology: *splines* are functions composed of simpler functions. These simpler functions are basis functions (e.g., a cubic polynomial) and the set of basis functions is a *basis*. Each basis function gets its coefficient and the resultant spline is the sum of these weighted basis functions.

Bayesian generalised additive multilevel models

Now describe the Bayesian GAMM (Miller, 2025)... Proper inclusion of varying/random effects in the model specification protects against overly wiggly curves (Baayen & Linke, 2020)... Generalising to scale and shape or “distributional GAMs” (Rigby & Stasinopoulos, 2005; Umlauf et al., 2018) and applied to neuroimaging data (Dinga et al., 2021)...

Instead of averaging, obtain the smooth ERP signal from multilevel GAM... less susceptible to outliers (Meulman et al., 2023)...

Objectives

Focusing on identifying onset and offset of effects (as assessed by ERP differences or decoding performance)... Assessing the performance of a model-based approach (i.e., Bayesian GAMMs) to conventional methods (multiplicity corrections or cluster-based permutation)...

Methods

M/EEG data simulation

Following the approach of Sassenhagen & Draschkow (2019) and Rousselet (2025), we simulated EEG data stemming from two conditions, one with noise only, and the other with noise + signal. As in previous studies, the noise was generated by superimposing 50 sinusoids at different frequencies, following an EEG-like spectrum (see details and code in Yeung et al., 2004). As in Rousselet (2025), the signal was generated from truncated Gaussian with an objective onset at 160 ms, a peak at 250 ms, and an offset at 342 ms. We simulated this signal for 250 timesteps between 0 and 0.5s, akin to a 500 Hz sampling rate. We simulated such data for a group of 20 participants with 50 trials per participant and condition (Figure 1).

We computed the average of the ERP difference (Figure 2)...

Figure 1

Averaged (mean) simulated EEG activity in two conditions with 50 trials each, for a group of 20 participants. The error band represents the mean plus/minus 1 standard error of the mean.



Model fitting

We then fitted a Bayesian GAM using the `brms` package (Bürkner, 2017, 2018; Nalborczyk et al., 2019). We used the default priors in `brms` (i.e., weakly informative priors). We ran eight Markov Chain Monte-Carlo (MCMC) to approximate the posterior distribution, including each 5000 iterations and a warmup of 2000 iterations, yielding a total of $8 \times (5000 - 2000) = 24000$ posterior samples to use for inference. Posterior convergence was assessed examining trace plots as well as the Gelman–Rubin statistic \hat{R} . The `brms` package uses the same syntax as the R package `mgcv` v 1.9-1 (Wood, 2017b) for specifying smooth effects. Figure 4 shows the predictions of this model together with the raw data.

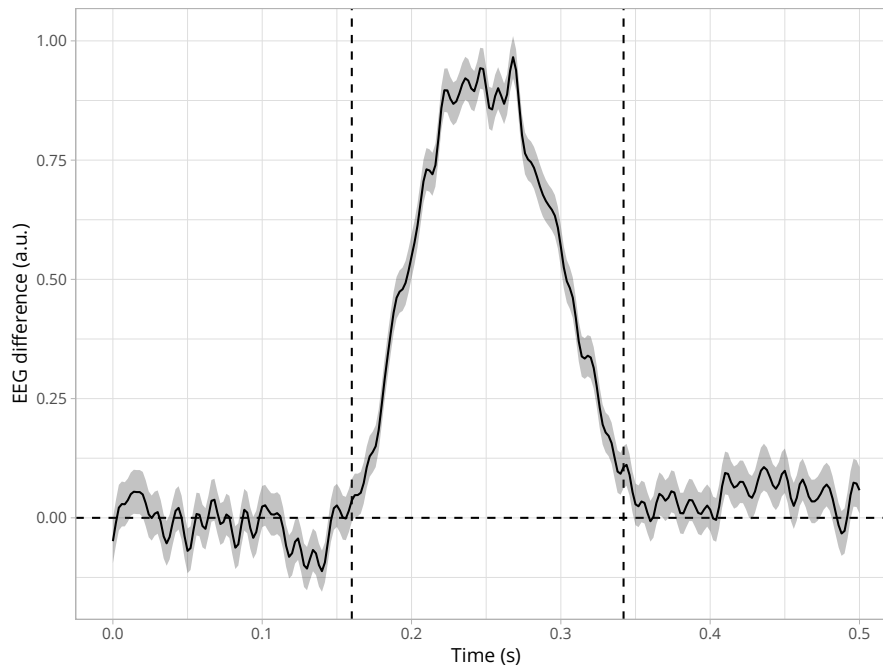
```
# averaging across participants
ppt_df <- raw_df %>%
  group_by(participant, condition, time) %>%
  summarise(eeg = mean(eeg) ) %>%
  ungroup()

# defining a contrast for condition
contrasts(ppt_df$condition) <- c(-0.5, 0.5)

# fitting the GAM
gam <- brm(
```

Figure 2

Group-level average difference between conditions (mean \pm standard error of the mean). The ‘true’ onset and offset are indicated by the vertical dashed lines.



```
# cubic regression splines with k-1 basis functions
eeg ~ condition + s(time, bs = "cr", k = 20, by = condition),
data = ppt_df,
family = gaussian(),
warmup = 2000,
iter = 5000,
chains = 8,
cores = 8,
file = "models/gam.rds"
)
```

Posterior probability of difference above 0

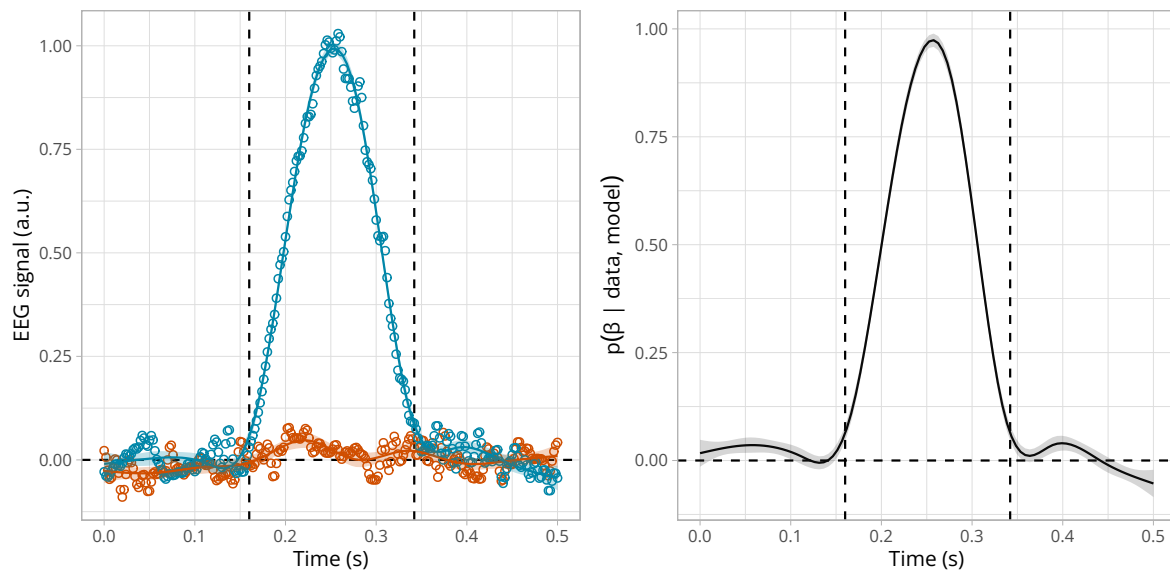
We then plot the posterior predictions together with the posterior estimate of the slope for `condition` at each timestep (Figure 3).

We then compute the posterior probability of the slope for `condition` being above $0 + \epsilon$ (Figure 4), with $\epsilon := 0.05$, which can be interpreted as the smallest effect size of interest.

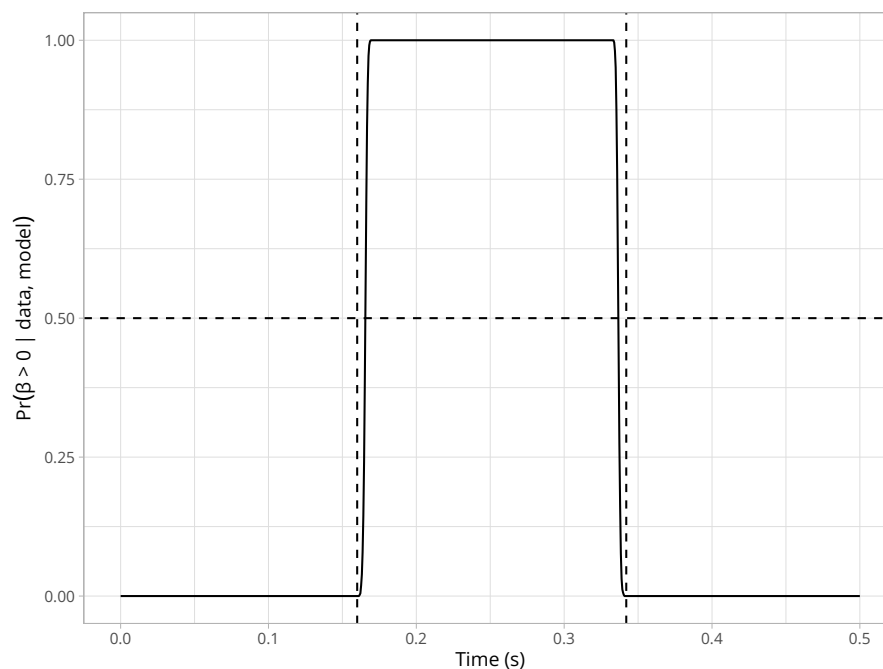
We can also express this as the ratio of posterior probabilities (i.e., $p/(1-p)$) and visualise the timecourse of this ratio superimposed with the conventional thresholds on evidence ratios (Figure 5).

Figure 3

Posterior estimate of the ERP in each condition (left) or directly for the difference of ERPs (right) according to the GAM.

**Figure 4**

Posterior probability of the ERP difference (slope) being above 0 according to the GAM.

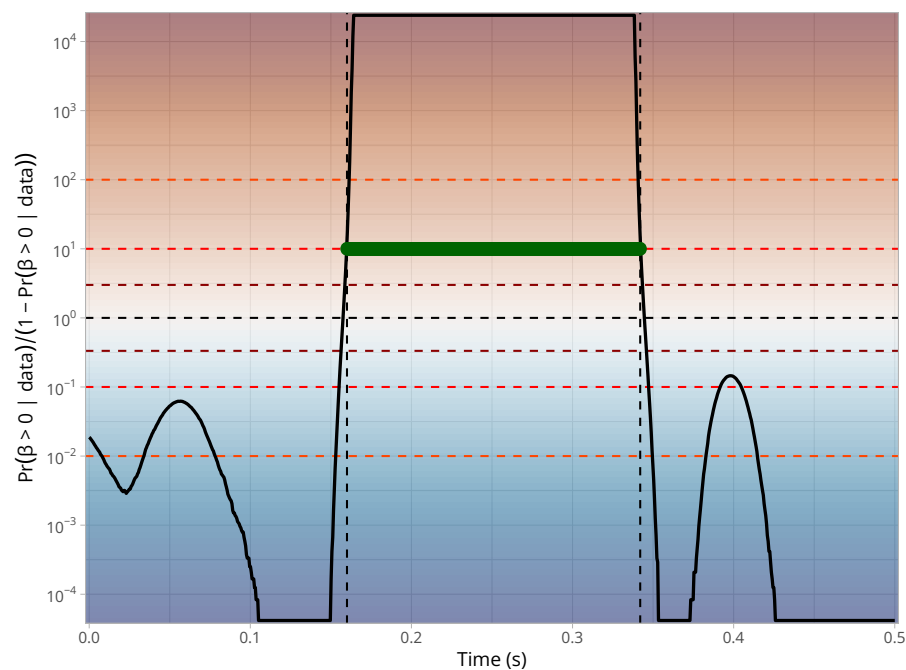


96 Multilevel modelling using ERP summary statistics

97 Next we fit a hierarchical/multilevel GAM using summary statistics of ERPs (mean and
 98 SD) at the participant level (similar to what is done in meta-analysis). Although it is possible
 99 to fit a GAMM at the single-trial level, we present a resource-efficient version of the model that
 100 is fitted directly on the by-participant summary statistics (mean and SD)...

Figure 5

Ratio of posterior probability according to the GAM (on a log10 scale). Timesteps above threshold (10) are highlighted in green.



```
# averaging across participants
summary_df <- raw_df %>%
  summarise(
    eeg_mean = mean(eeg),
    eeg_sd = sd(eeg),
    .by = c(participant, condition, time)
  )

# defining a contrast for condition
contrasts(summary_df$condition) <- c(-0.5, 0.5)

# fitting the GAM
meta_gam <- brm(
  # using by-participant SD of ERPs across trials
  eeg_mean | se(eeg_sd) ~
    condition + s(time, bs = "cr", k = 20, by = condition) +
    (1 | participant),
  data = summary_df,
  family = gaussian(),
  warmup = 2000,
  iter = 5000,
  chains = 8,
  cores = 8,
  file = "models/meta_gam.rds"
)
```


Error properties of the proposed approach

We then computed the difference between the true and estimated onset/offset of the ERP difference (error $:= |\hat{\theta} - \theta|$), according to various **eps** and **threshold** values. Remember that the signal is generated from a truncated Gaussian with an objective onset at 160 ms, a maximum at 250 ms, and an offset at 342 ms. Figure 6 shows that the hierarchical GAM can *exactly* recover the true onset and offset values, given some reasonable choice of **eps** and **threshold** values (e.g., a threshold of 20).

	eps	threshold	estimated_onset	estimated_offset	error_onset	error_offset
1	0.00	13	0.16	0.34	0	0.002
2	0.00	14	0.16	0.34	0	0.002
3	0.00	15	0.16	0.34	0	0.002
4	0.00	16	0.16	0.34	0	0.002
5	0.00	17	0.16	0.34	0	0.002
6	0.01	10	0.16	0.34	0	0.002

Figure 6

Error function of onset (left) and offset (right) estimation according to various **eps** and **threshold** values (according to the hierarchical GAM). Minimum error values are indicated by red crosses.



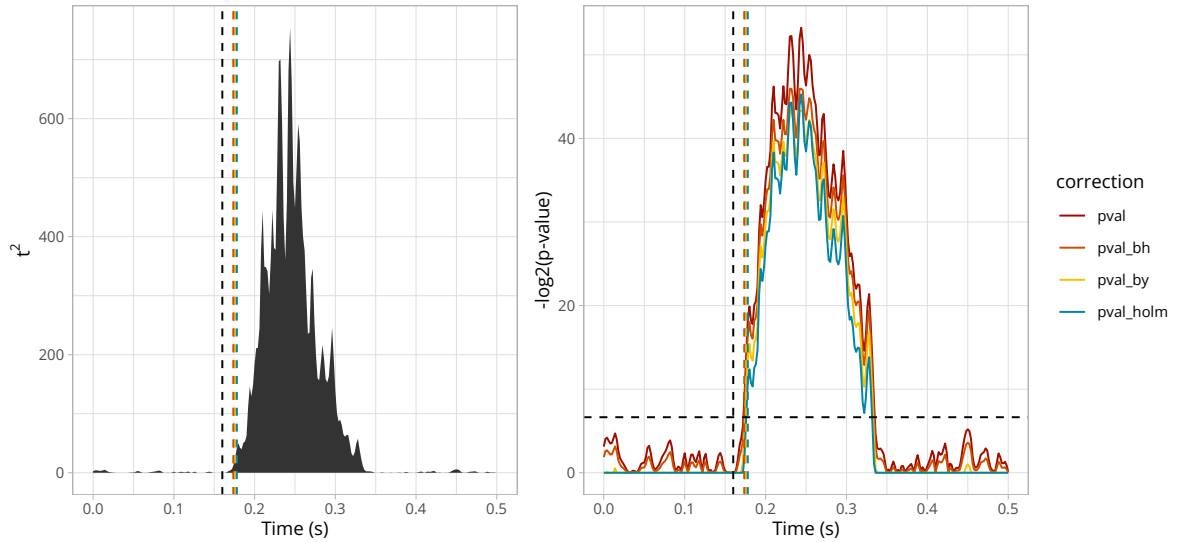
Comparing the identified onsets/offsets to other approaches

We compared the ability of the GAMM to correctly estimate the onset and offset of the ERP difference to widely used methods. First, we conducted mass-univariate t-tests (thus treating each timestep independently) and identified the onset and offset of the ERP difference as the first and last values crossing an arbitrary significance threshold ($\alpha = 0.01$). We then followed the same approach but after applying different forms of multiplicity correction the p-values. We compared two methods that control the false discovery rate (FDR) (i.e., BH95, [Benjamini & Hochberg, 1995](#); and BY01, [Benjamini & Yekutieli, 2001](#)), one method that controls the familywise error rate (FWER) (i.e., Holm–Bonferroni method, [Holm, 1979](#)), and two cluster-based methods (permutation with a single threshold and TFCE, [S. Smith & Nichols, 2009](#)). The BH95, BY01, and Holm corrections were applied to the p-values using the `p.adjust()` function in R. The cluster-based inference was implemented using a cluster-sum statistic of squared t-values, as implemented in MNE-Python ([Gramfort, 2013](#)), called via the R package `reticulate` v 1.35.0 ([Ushey et al., 2024](#)). We also compared these estimates to the onset and offset as estimated using the binary segmentation algorithm, as implemented in the R package `changept` v 2.2.4 ([Killick et al., 2022a](#)), and applied directly to the squared t-values (as in [Rousselet, 2025](#)). For visualisation and interpretability purposes, we converted p-values to s-values, which can be interpreted as bits of surprising information, assuming a null effect ([Greenland, 2019](#)) (Figure 7).

Figure 7

Timecourse of squared t-values and s-values, with true onset (black dashed line) and onsets identified using the raw (uncorrected) p-values or the corrected p-values (BH, BY, Holm).

True onset: 0.16s, Uncorrected onset: 0.174s, BH onset: 0.174s, BY onset: 0.176s, Holm onset: 0.178s



Simulation study

Onset/offset estimation methods were assessed using the median absolute error (MAE) and variance of 10.000 simulated datasets...

Application to actual MEG data

Next, we assessed the performance of the proposed approach on actual MEG data (decoding results of Nalborczyk et al., in preparation). We conducted time-resolved multivariate pattern analysis (MVPA), also known as decoding... As a result, we have a timecourse of decoding performance (ROC AUC), bounded between 0 and 1, for each participant (for a total of 32 participants). Now, we want to *test* whether the group-level average decoding accuracy is above chance (i.e., 0.5) at each timestep (Figure 8). We fitted a similar GAM as discussed previously, but we replaced the Normal likelihood function by a Beta one to account for the bounded nature of AUC values (between 0 and 1) (for a tutorial on Beta regression, see [Coretta & Bürkner, 2025](#)). Note that although we chose a basis dimension of $k = 50$, which seems appropriate for the present data, this choice should be adapted according to the properties of the modelled data (e.g., signal-to-noise ratio, prior low-pass filtering, sampling rate, etc) and should be assessed by the usual model checking tools (e.g., posterior predictive checks). We also define a region of practical equivalence (ROPE, [Kruschke & Liddell, 2017](#)), defined as the chance level plus the standard deviation of the (group-level average) decoding performance during the baseline period. This ensures that...¹

```
# fitting the Beta GAM
meg_decoding_gam <- brm(
  auc ~ s(time, bs = "cr", k = 50),
  data = decoding_df,
  family = Beta(),
  warmup = 2000,
  iter = 5000,
  chains = 4,
  cores = 4
)
```

We assessed the reliability of the proposed approach using a form of permutation-based split-half reliability (e.g., [Rosenblatt et al., 2018](#)), which consisted of the following steps:

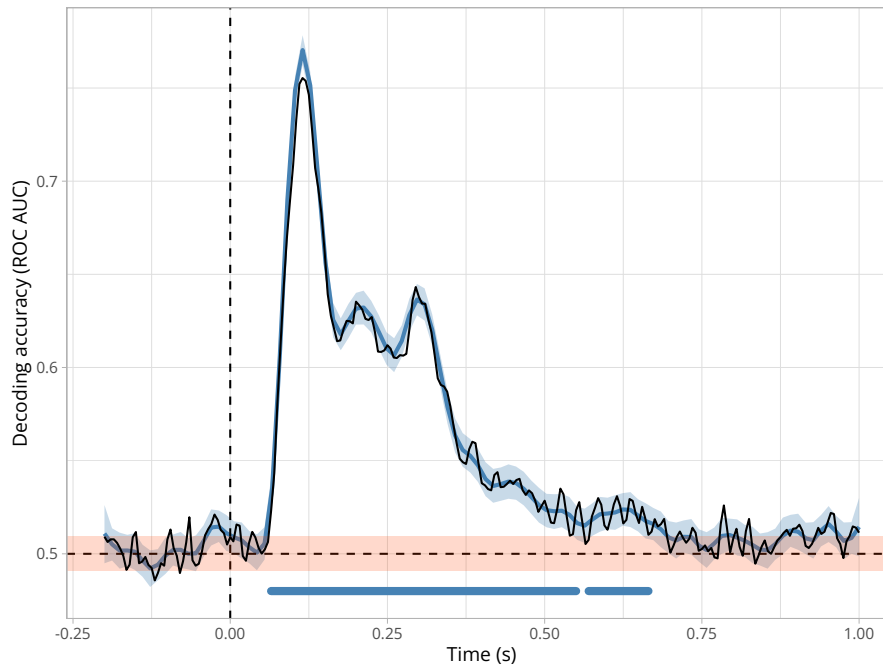
- Create many (e.g., 10 or 20) half train/test splits of the data
- For each fold, estimate the onset/offset on both splits using all methods
- Then summarise the distribution of onset/offset with the median and variance across folds

This will allow checking that the proposed approach produces reliable onset/offset estimates.

¹Could an alternative approach be to include a predictor for “baseline vs. after baseline”?

Figure 8

Group-level average decoding performance ($N=32$) superimposed with the GAM predictions (in blue) and the region of practical equivalence (ROPE, in orange) computed from the baseline period. The blue horizontal line indicates the timesteps at which the posterior probability ratio is equal to or greater than 20.



Results

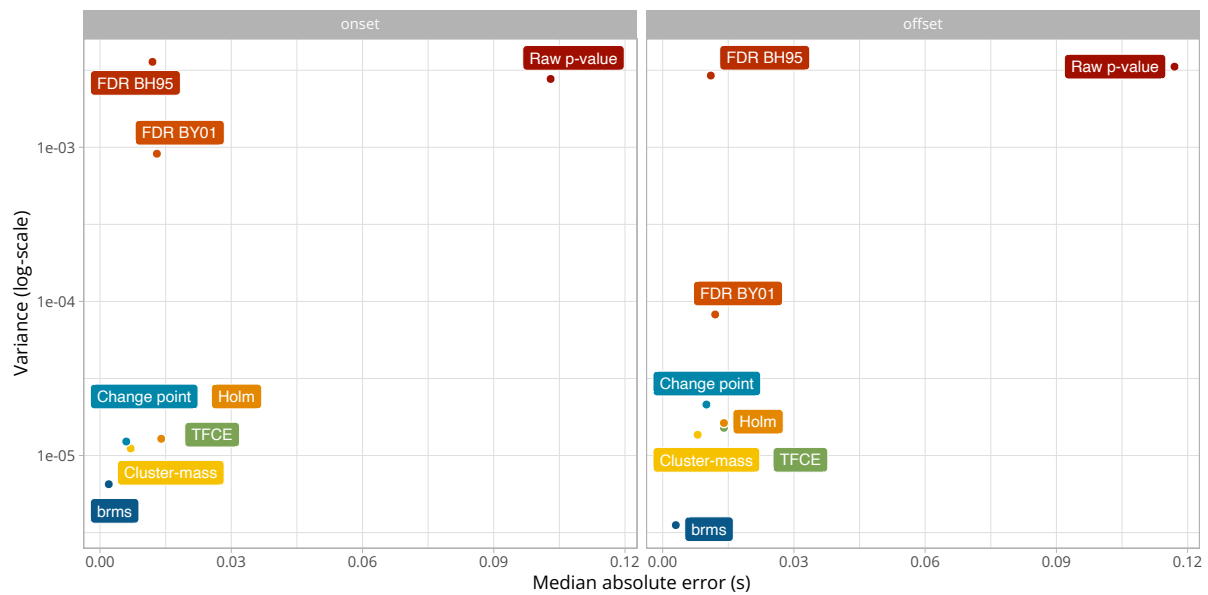
This section is divided in two parts. First, we present the results from the simulation study, assessing the bias and variance of each method when applied to simulated data in which the ground truth is known. Second, we present the results obtained when applying the different methods to actual MEG data (decoding performance through time), assessing the reliability of each method and the stability of its estimates.

Simulation results (bias and variance)

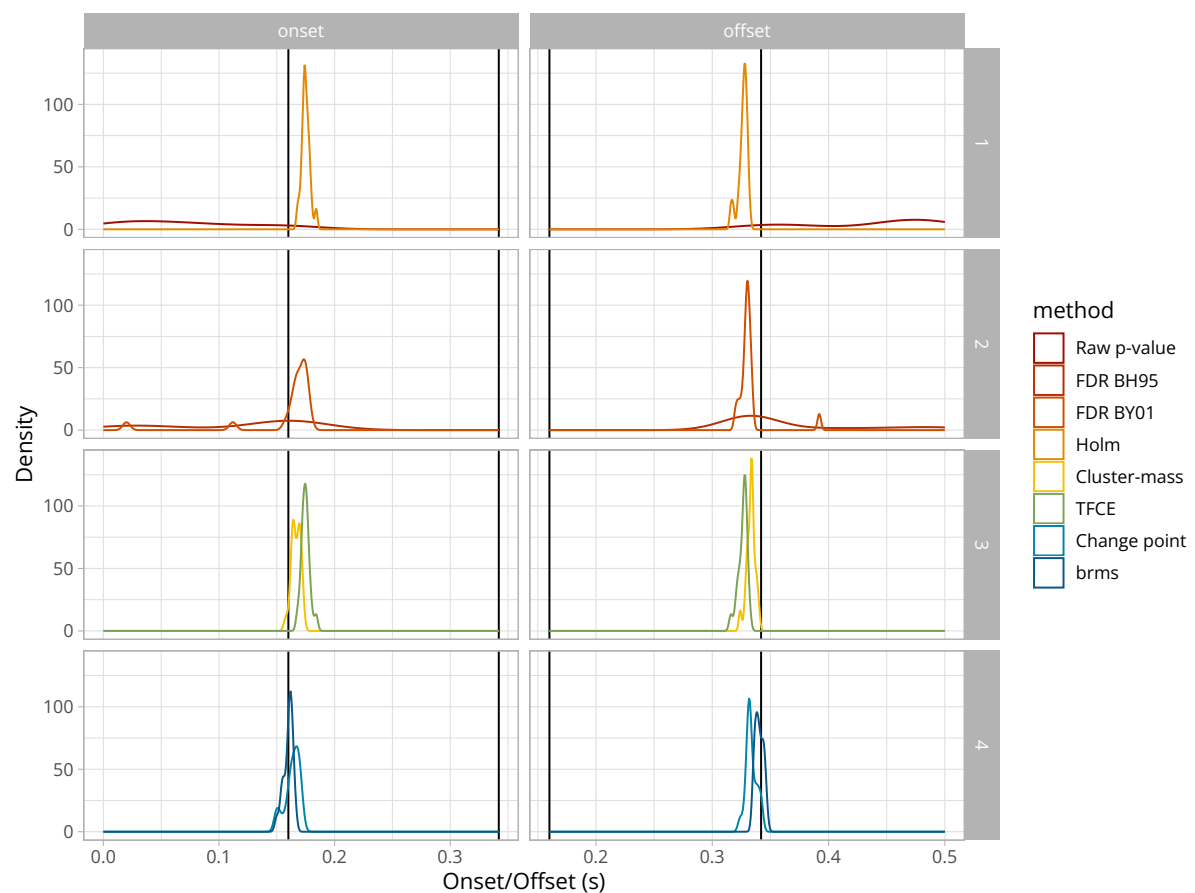
Figure 9 shows a summary of the simulation results, revealing that the proposed approach (**brms**) has the lowest MAE and variance for both the onset and offset estimates...

Figure 9

Median absolute error and variance of onset and offset estimates for each method.

**Figure 10**

Distributions of onset and offset estimates for each method.



Application to actual MEG data (reliability)

Figure 11 shows the group-level average decoding performance through time with onset and offset estimates for each method. The `brms_full` method is similar to the `brms` method

except that the ROPE is defined on the entire dataset rather than on the split dataset. Overall, this figure shows that both the **Raw p-value** and **FDR BH95** methods were extremely lenient, considering that the decoding performance was above chance before the onset of the stimulus (false positive) and until the end of the trial. The **Change point** and **Cluster mass** methods were the most conservative methods, identifying a time window from approximately +60ms to +500ms. The **Holm**, **TFCE**, **brms**, and **brms_full** methods produced somewhat similar estimates of onset and offset, from approximately +60ms to +650ms.²

Figure 11

Group-level average decoding performance through time with onset and offset estimates for each method (data from Nalborczyk et al., in preparation).

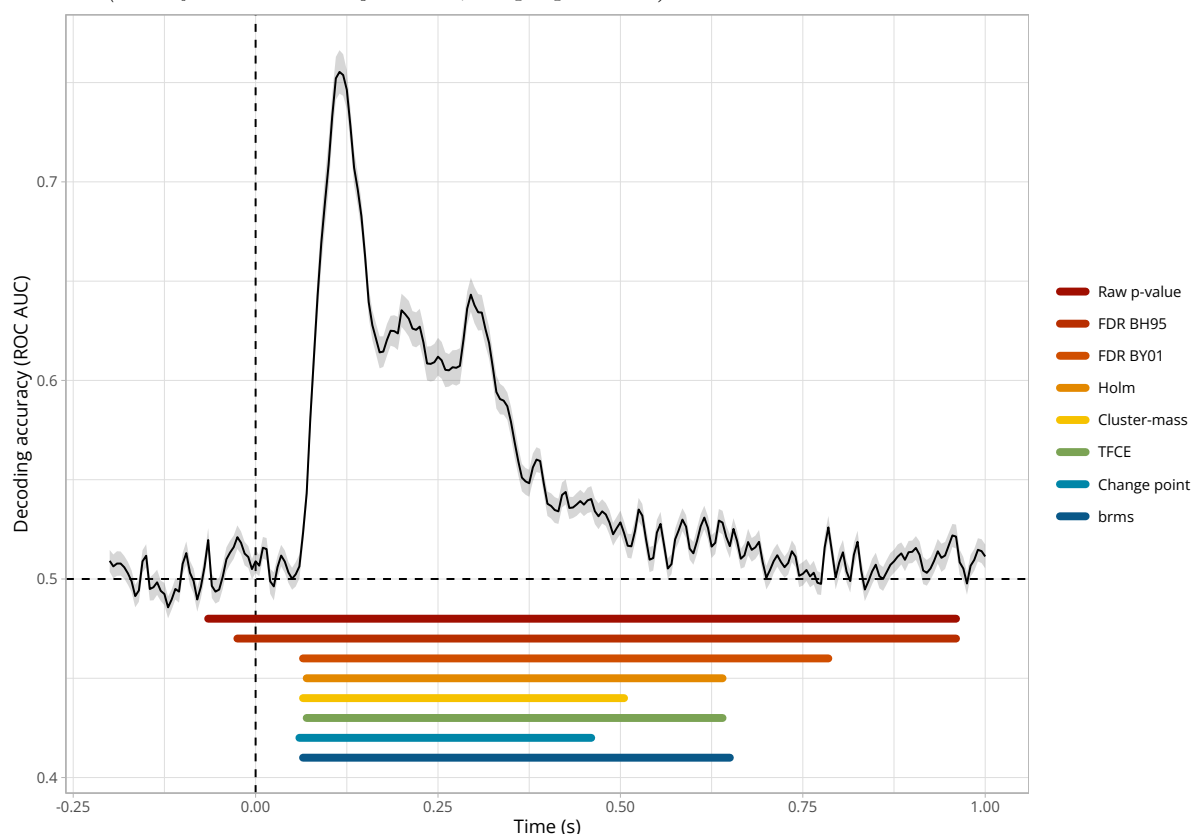
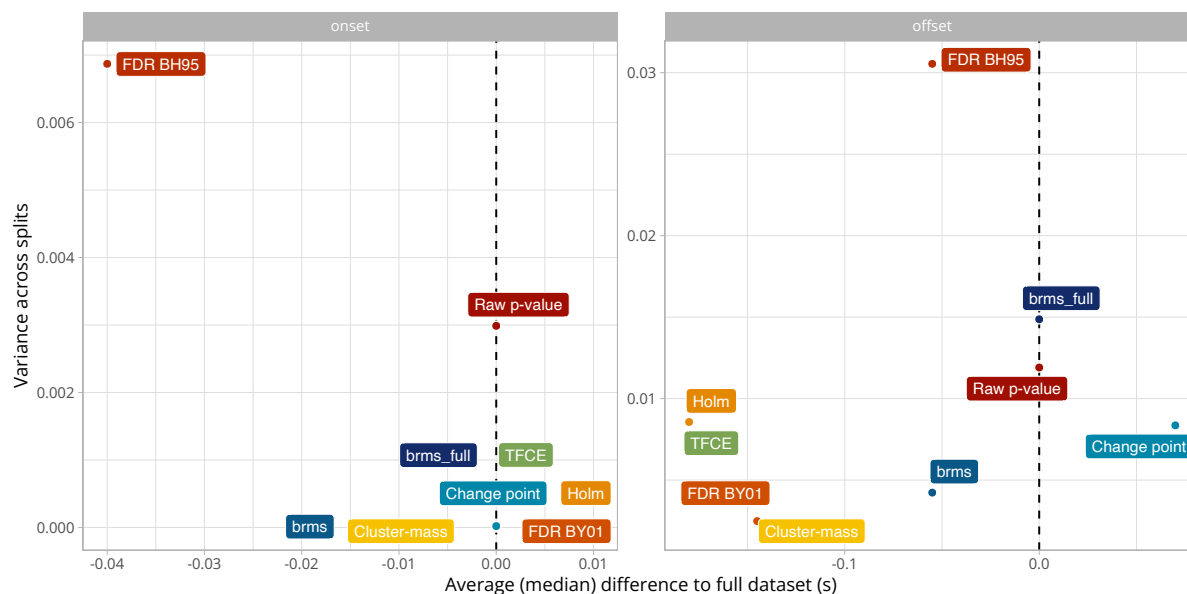


Figure 12 shows two properties of each method: i) the median difference between the onset and offset estimates from each data split and the onset and offset estimates from the full dataset (x-axis) and ii) and the variance of its onset and offset estimates across data splits (y-axis). This figure reveals that the **brms onset and offset** estimates on each split are the closest to the estimates from the full dataset on average (0ms difference for the onset estimate and 5ms difference for the offset estimate). The **Raw p-value** method has similar performance, but given the aberrant estimates it produces (cf. Figure 11), the result that it is consistent between data splits and the full dataset is not convincing on its own. Overall, the figure reveals that for all other methods, split datasets produce later onset estimates and earlier offset estimates (as compared to the estimates from the model fitted on the full dataset).

²It should be noted that although each method can produce several “clusters” of timesteps, we only considered the first (onset) and last (offset) timesteps identified by each method to compute the error (difference).

Figure 12

Median absolute error and variance of onset (left) and offset (right) estimates for each method.



Discussion

...

Summary of the proposed approach

Overall, before concluding on the onset/offset of effect based on the model, we need to ensure that the model provides a faithful description of the data-generating process (e.g., via posterior predictive checks etc)...

Increasing potential usage

Prepare a wrapper R package and show how to call it in Python and integrate it with MNE-Python (Gramfort, 2013) pipelines...

Limitations and future directions

As in previous simulation work (e.g., Rousselet et al., 2008; Sassenhagen & Draschkow, 2019), the present simulation results depend on various choices such as the specific cluster-forming algorithm and threshold, signal-to-noise ratio, negative impact of preprocessing steps (e.g., low-pass filter) on temporal resolution... note however, that the same caveats apply to all methods...

The error properties depend on the threshold parameter, a value of 10 or 20 seems to be a reasonable default, but the optimal threshold parameter can be adjusted using split-half reliability assessment...

Can be applied to any 1D timeseries (e.g., pupillometry, electromyography)... Extending the approach to spatiotemporal data (i.e., time + sensors)...

We kept the exemplary models simple, but can be extended by adding varying/random effects (intercept and slope) for item (e.g., word)... but also continuous predictors at the trial level?

Conclusions

...

Data and code availability

The simulation results as well as the R code to reproduce the simulations are available on GitHub: https://github.com/lnalborczyk/brms_meeg.

Packages

We used R version 4.2.3 (R Core Team, 2023) and the following R packages: brms v. 2.22.0 (Bürkner, 2017, 2018, 2021), changepoint v. 2.2.4 (Killick et al., 2022b; Killick & Eckley, 2014), grateful v. 0.2.10 (Rodriguez-Sanchez & Jackson, 2023), knitr v. 1.45 (Xie, 2014, 2015, 2023), MetBrewer v. 0.2.0 (Mills, 2022), pakret v. 0.2.2 (Gallou, 2024), patchwork v. 1.2.0 (T. L. Pedersen, 2024), rmarkdown v. 2.29 (Allaire et al., 2024; Xie et al., 2018, 2020), scales v. 1.3.0 (Wickham et al., 2023), scico v. 1.5.0 (T. L. Pedersen & Crameri, 2023), tidybayes v. 3.0.6 (Kay, 2023), tidyverse v. 2.0.0 (Wickham et al., 2019).

References

- Abugaber, D., Finestrat, I., Luque, A., & Morgan-Short, K. (2023). Generalized additive mixed modeling of EEG supports dual-route accounts of morphosyntax in suggesting no word frequency effects on processing of regular grammatical forms. *Journal of Neurolinguistics*, 67, 101137. <https://doi.org/10.1016/j.jneuroling.2023.101137>
- Allaire, J., Xie, Y., Dervieux, C., McPherson, J., Luraschi, J., Ushey, K., Atkins, A., Wickham, H., Cheng, J., Chang, W., & Iannone, R. (2024). *rmarkdown: Dynamic documents for r*. <https://github.com/rstudio/rmarkdown>
- Baayen, R. H., & Linke, M. (2020). *Generalized Additive Mixed Models* (pp. 563–591). Springer International Publishing. https://doi.org/10.1007/978-3-030-46216-1_23
- Baayen, R. H., Rij, J. van, Cat, C. de, & Wood, S. (2018). *Autocorrelated errors in experimental data in the language sciences: Some solutions offered by generalized additive mixed models* (pp. 49–69). Springer International Publishing. https://doi.org/10.1007/978-3-319-69830-4_4
- Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 57(1), 289–300. <https://doi.org/10.1111/j.2517-6161.1995.tb02031.x>
- Benjamini, Y., & Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, 29(4). <https://doi.org/10.1214/aos/1013699998>
- Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, 80(1), 1–28. <https://doi.org/10.18637/jss.v080.i01>
- Bürkner, P.-C. (2018). Advanced Bayesian multilevel modeling with the R package brms. *The R Journal*, 10(1), 395–411. <https://doi.org/10.32614/RJ-2018-017>
- Bürkner, P.-C. (2021). Bayesian item response modeling in R with brms and Stan. *Journal of Statistical Software*, 100(5), 1–54. <https://doi.org/10.18637/jss.v100.i05>
- Combrisson, E., & Jerbi, K. (2015). Exceeding chance level by chance: The caveat of theoretical chance levels in brain signal classification and statistical assessment of decoding accuracy. *Journal of Neuroscience Methods*, 250, 126–136. <https://doi.org/10.1016/j.jneumeth.2015.01.010>
- Coretta, S., & Bürkner, P.-C. (2025). *Bayesian beta regressions with brms in r: A tutorial for phoneticians*. http://dx.doi.org/10.31219/osf.io/f9rqg_v1
- Dimigen, O., & Ehinger, B. V. (2021). Regression-based analysis of combined EEG and eye-tracking data: Theory and applications. *Journal of Vision*, 21(1), 3. <https://doi.org/10.1167/jov.21.1.3>
- Dinga, R., Frazz, C. J., Bayer, J. M. M., Kia, S. M., Beckmann, C. F., & Marquand, A. F. (2021). *Normative modeling of neuroimaging data using generalized additive models of location scale and shape*. <http://dx.doi.org/10.1101/2021.06.14.448106>
- Dunagan, D., Jordan, T., Hale, J. T., Pylkkänen, L., & Chacón, D. A. (2024). *Evaluating the timecourses of morpho-orthographic, lexical, and grammatical processing following rapid parallel visual presentation: An EEG investigation in english*. <http://dx.doi.org/10.1101/2024.04.10.588861>
- Ehinger, B. V., & Dimigen, O. (2019). Unfold: An integrated toolbox for overlap correction, non-linear modeling, and regression-based EEG analysis. *PeerJ*, 7, e7838. <https://doi.org/10.7717/peerj.7838>
- Eklund, A., Nichols, T. E., & Knutsson, H. (2016). Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. *Proceedings of the National Academy of Sciences*, 113(28), 7900–7905. <https://doi.org/10.1073/pnas.1602413113>
- Fischer, Adrian G., & Ullsperger, M. (2013). Real and Fictive Outcomes Are Processed Differently but Converge on a Common Adaptive Mechanism. *Neuron*, 79(6), 1243–1255.

- <https://doi.org/10.1016/j.neuron.2013.07.006>
- Frossard, J., & Renaud, O. (2021). Permutation Tests for Regression, ANOVA, and Comparison of Signals: The **permuco** Package. *Journal of Statistical Software*, 99(15). <https://doi.org/10.18637/jss.v099.i15>
- Frossard, J., & Renaud, O. (2022). The cluster depth tests: Toward point-wise strong control of the family-wise error rate in massively univariate tests with application to M/EEG. *NeuroImage*, 247, 118824. <https://doi.org/10.1016/j.neuroimage.2021.118824>
- Gallou, A. (2024). *pakret: Cite “R” packages on the fly in “R Markdown” and “Quarto”*. <https://CRAN.R-project.org/package=pakret>
- Gramfort, A. (2013). MEG and EEG data analysis with MNE-python. *Frontiers in Neuroscience*, 7. <https://doi.org/10.3389/fnins.2013.00267>
- Greenland, S. (2019). Valid P -Values Behave Exactly as They Should: Some Misleading Criticisms of P -Values and Their Resolution With S -Values. *The American Statistician*, 73(sup1), 106–114. <https://doi.org/10.1080/00031305.2018.1529625>
- Hastie, T. J., & Tibshirani, R. J. (2017). *Generalized Additive Models*. Routledge. <https://doi.org/10.1201/9780203753781>
- Hauk, O., Davis, M. H., Ford, M., Pulvermüller, F., & Marslen-Wilson, W. D. (2006). The time course of visual word recognition as revealed by linear regression analysis of ERP data. *NeuroImage*, 30(4), 1383–1400. <https://doi.org/10.1016/j.neuroimage.2005.11.048>
- Hayasaka, S. (2003). Validating cluster size inference: Random field and permutation methods. *NeuroImage*, 20(4), 2343–2356. <https://doi.org/10.1016/j.neuroimage.2003.08.003>
- Hendrix, P., Bolger, P., & Baayen, H. (2017). Distinct ERP signatures of word frequency, phrase frequency, and prototypicality in speech production. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 43(1), 128–149. <https://doi.org/10.1037/a0040332>
- Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2), 65–70. <http://www.jstor.org/stable/4615733>
- Kay, M. (2023). *tidybayes: Tidy data and geoms for Bayesian models*. <https://doi.org/10.5281/zenodo.1308151>
- Killick, R., & Eckley, I. A. (2014). changepoint: An R package for changepoint analysis. *Journal of Statistical Software*, 58(3), 1–19. <https://www.jstatsoft.org/article/view/v058i03>
- Killick, R., Haynes, K., & Eckley, I. A. (2022a). *changepoint: An R package for changepoint analysis*. <https://CRAN.R-project.org/package=changepoint>
- Killick, R., Haynes, K., & Eckley, I. A. (2022b). *changepoint: An R package for changepoint analysis*. <https://CRAN.R-project.org/package=changepoint>
- King, J.-R., & Dehaene, S. (2014). Characterizing the dynamics of mental representations: the temporal generalization method. *Trends in Cognitive Sciences*, 18(4), 203–210. <https://doi.org/10.1016/j.tics.2014.01.002>
- Kruschke, J. K., & Liddell, T. M. (2017). The Bayesian New Statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. *Psychonomic Bulletin & Review*, 25(1), 178–206. <https://doi.org/10.3758/s13423-016-1221-4>
- Kryuchkova, T., Tucker, B. V., Wurm, L. H., & Baayen, R. H. (2012). Danger and usefulness are detected early in auditory lexical processing: Evidence from electroencephalography. *Brain and Language*, 122(2), 81–91. <https://doi.org/10.1016/j.bandl.2012.05.005>
- Luck, S. J., & Gaspelin, N. (2017). How to get statistically significant effects in any ERP experiment (and why you shouldn’t). *Psychophysiology*, 54(1), 146–157. <https://doi.org/10.1111/psyp.12639>
- Maris, E. (2011). Statistical testing in electrophysiological studies. *Psychophysiology*, 49(4), 549–565. <https://doi.org/10.1111/j.1469-8986.2011.01320.x>
- Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neuroscience Methods*, 164(1), 177–190. <https://doi.org/10.1016/j.jneumeth.2007.03.024>

- Meulman, N., Sprenger, S. A., Schmid, M. S., & Wieling, M. (2023). GAM-based individual difference measures for L2 ERP studies. *Research Methods in Applied Linguistics*, 2(3), 100079. <https://doi.org/10.1016/j.rmal.2023.100079>
- 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), e0143328. <https://doi.org/10.1371/journal.pone.0143328>
- Miller, D. L. (2025). Bayesian views of generalized additive modelling. *Methods in Ecology and Evolution*. <https://doi.org/10.1111/2041-210x.14498>
- Mills, B. R. (2022). *MetBrewer: Color palettes inspired by works at the metropolitan museum of art*. <https://CRAN.R-project.org/package=MetBrewer>
- Nalborczyk, L., Batailler, C., Lœvenbruck, H., Vilain, A., & Bürkner, P.-C. (2019). An Introduction to Bayesian Multilevel Models Using brms: A Case Study of Gender Effects on Vowel Variability in Standard Indonesian. *Journal of Speech, Language, and Hearing Research*, 62(5), 1225–1242. https://doi.org/10.1044/2018_jslhr-s-18-0006
- Pedersen, E. J., Miller, D. L., Simpson, G. L., & Ross, N. (2019). Hierarchical generalized additive models in ecology: An introduction with mgcv. *PeerJ*, 7, e6876. <https://doi.org/10.7717/peerj.6876>
- Pedersen, T. L. (2024). *patchwork: The composer of plots*. <https://CRAN.R-project.org/package=patchwork>
- Pedersen, T. L., & Cramer, F. (2023). *scico: Colour palettes based on the scientific colour-maps*. <https://CRAN.R-project.org/package=scico>
- Pernet, C. R. (2022). Electroencephalography robust statistical linear modelling using a single weight per trial. *Aperture Neuro*, 2, 1–22. <https://doi.org/10.52294/apertureneuro.2022.2.seoo9435>
- 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, 1–11. <https://doi.org/10.1155/2011/831409>
- Pernet, C. R., Latinus, M., Nichols, T. E., & Rousselet, G. A. (2015). Cluster-based computational methods for mass univariate analyses of event-related brain potentials/fields: A simulation study. *Journal of Neuroscience Methods*, 250, 85–93. <https://doi.org/10.1016/j.jneumeth.2014.08.003>
- R Core Team. (2023). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. <https://www.R-project.org/>
- Rigby, R. A., & Stasinopoulos, D. M. (2005). Generalized Additive Models for Location, Scale and Shape. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 54(3), 507–554. <https://doi.org/10.1111/j.1467-9876.2005.00510.x>
- Rij, J. van, Hendriks, P., Rijn, H. van, Baayen, R. H., & Wood, S. N. (2019). Analyzing the Time Course of Pupillometric Data. *Trends in Hearing*, 23. <https://doi.org/10.1177/2331216519832483>
- Rodriguez-Sanchez, F., & Jackson, C. P. (2023). *grateful: Facilitate citation of r packages*. <https://pakillo.github.io/grateful/>
- Rosenblatt, J. D., Finos, L., Weeda, W. D., Solari, A., & Goeman, J. J. (2018). All-Resolutions Inference for brain imaging. *NeuroImage*, 181, 786–796. <https://doi.org/10.1016/j.neuroimage.2018.07.060>
- Rousselet, G. A. (2025). Using cluster-based permutation tests to estimate MEG/EEG onsets: How bad is it? *European Journal of Neuroscience*, 61(1), e16618. <https://doi.org/10.1111/ejn.16618>
- Rousselet, G. A., Pernet, C. R., Bennett, P. J., & Sekuler, A. B. (2008). Parametric study of EEG sensitivity to phase noise during face processing. *BMC Neuroscience*, 9(1). <https://doi.org/10.1186/1471-2202-9-98>
- Sassenhagen, J., & Draschkow, D. (2019). Cluster-based permutation tests of MEG/EEG data

- do not establish significance of effect latency or location. *Psychophysiology*, 56(6). <https://doi.org/10.1111/psyp.13335>
- Skukies, R., & Ehinger, B. (2021). Modelling event duration and overlap during EEG analysis. *Journal of Vision*, 21(9), 2037. <https://doi.org/10.1167/jov.21.9.2037>
- Skukies, R., Schepers, J., & Ehinger, B. (2024, December 9). *Brain responses vary in duration - modeling strategies and challenges*. <https://doi.org/10.1101/2024.12.05.626938>
- Smith, N. J., & Kutas, M. (2014a). Regression-based estimation of ERP waveforms: I. The rERP framework. *Psychophysiology*, 52(2), 157–168. <https://doi.org/10.1111/psyp.12317>
- Smith, N. J., & Kutas, M. (2014b). Regression-based estimation of ERP waveforms: II. Non-linear effects, overlap correction, and practical considerations. *Psychophysiology*, 52(2), 169–181. <https://doi.org/10.1111/psyp.12320>
- Smith, S., & Nichols, T. (2009). Threshold-free cluster enhancement: Addressing problems of smoothing, threshold dependence and localisation in cluster inference. *NeuroImage*, 44(1), 83–98. <https://doi.org/10.1016/j.neuroimage.2008.03.061>
- Sóskuthy, M. (2017). *Generalised additive mixed models for dynamic analysis in linguistics: A practical introduction*. <https://doi.org/10.48550/ARXIV.1703.05339>
- Sóskuthy, M. (2021). Evaluating generalised additive mixed modelling strategies for dynamic speech analysis. *Journal of Phonetics*, 84, 101017. <https://doi.org/10.1016/j.wocn.2020.101017>
- Teichmann, L. (2022). An empirically driven guide on using bayes factors for m/EEG decoding. *Aperture Neuro*, 2, 1–10. <https://doi.org/10.52294/apertureneuro.2022.2.maoc6465>
- Tremblay, A., & Newman, A. J. (2014). Modeling nonlinear relationships in ERP data using mixed-effects regression with R examples. *Psychophysiology*, 52(1), 124–139. <https://doi.org/10.1111/psyp.12299>
- Umlauf, N., Klein, N., & Zeileis, A. (2018). BAMLSS: Bayesian Additive Models for Location, Scale, and Shape (and Beyond). *Journal of Computational and Graphical Statistics*, 27(3), 612–627. <https://doi.org/10.1080/10618600.2017.1407325>
- Ushey, K., Allaire, J., & Tang, Y. (2024). *Reticulate: Interface to 'python'*. <https://CRAN.R-project.org/package=reticulate>
- Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., ... Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, 4(43), 1686. <https://doi.org/10.21105/joss.01686>
- Wickham, H., Pedersen, T. L., & Seidel, D. (2023). *scales: Scale functions for visualization*. <https://CRAN.R-project.org/package=scales>
- Wieling, M. (2018). Analyzing dynamic phonetic data using generalized additive mixed modeling: A tutorial focusing on articulatory differences between L1 and L2 speakers of English. *Journal of Phonetics*, 70, 86–116. <https://doi.org/10.1016/j.wocn.2018.03.002>
- Winter, B., & Wieling, M. (2016). How to analyze linguistic change using mixed models, Growth Curve Analysis and Generalized Additive Modeling. *Journal of Language Evolution*, 1(1), 7–18. <https://doi.org/10.1093/jole/lzv003>
- Wood, S. N. (2003). Thin Plate Regression Splines. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 65(1), 95–114. <https://doi.org/10.1111/1467-9868.00374>
- Wood, S. N. (2017a). *Generalized Additive Models*. Chapman; Hall/CRC. <https://doi.org/10.1201/9781315370279>
- Wood, S. N. (2017b). *Generalized additive models: An introduction with r* (2nd ed.). Chapman; Hall/CRC.
- Wüllhorst, V., Wüllhorst, R., Overmeyer, R., & Endrass, T. (2025). Comprehensive Analysis of Event-Related Potentials of Response Inhibition: The Role of Negative Urgency and Compulsivity. *Psychophysiology*, 62(2). <https://doi.org/10.1111/psyp.70000>

- Xie, Y. (2014). knitr: A comprehensive tool for reproducible research in R. In V. Stodden, F. Leisch, & R. D. Peng (Eds.), *Implementing reproducible computational research*. Chapman; Hall/CRC.
- Xie, Y. (2015). *Dynamic documents with R and knitr* (2nd ed.). Chapman; Hall/CRC. <https://yihui.org/knitr/>
- Xie, Y. (2023). *knitr: A general-purpose package for dynamic report generation in r*. <https://yihui.org/knitr/>
- Xie, Y., Allaire, J. J., & Golemund, G. (2018). *R markdown: The definitive guide*. Chapman; Hall/CRC. <https://bookdown.org/yihui/rmarkdown>
- Xie, Y., Dervieux, C., & Riederer, E. (2020). *R markdown cookbook*. Chapman; Hall/CRC. <https://bookdown.org/yihui/rmarkdown-cookbook>
- Yeung, N., Bogacz, R., Holroyd, C. B., & Cohen, J. D. (2004). Detection of synchronized oscillations in the electroencephalogram: An evaluation of methods. *Psychophysiology*, 41(6), 822–832. <https://doi.org/10.1111/j.1469-8986.2004.00239.x>

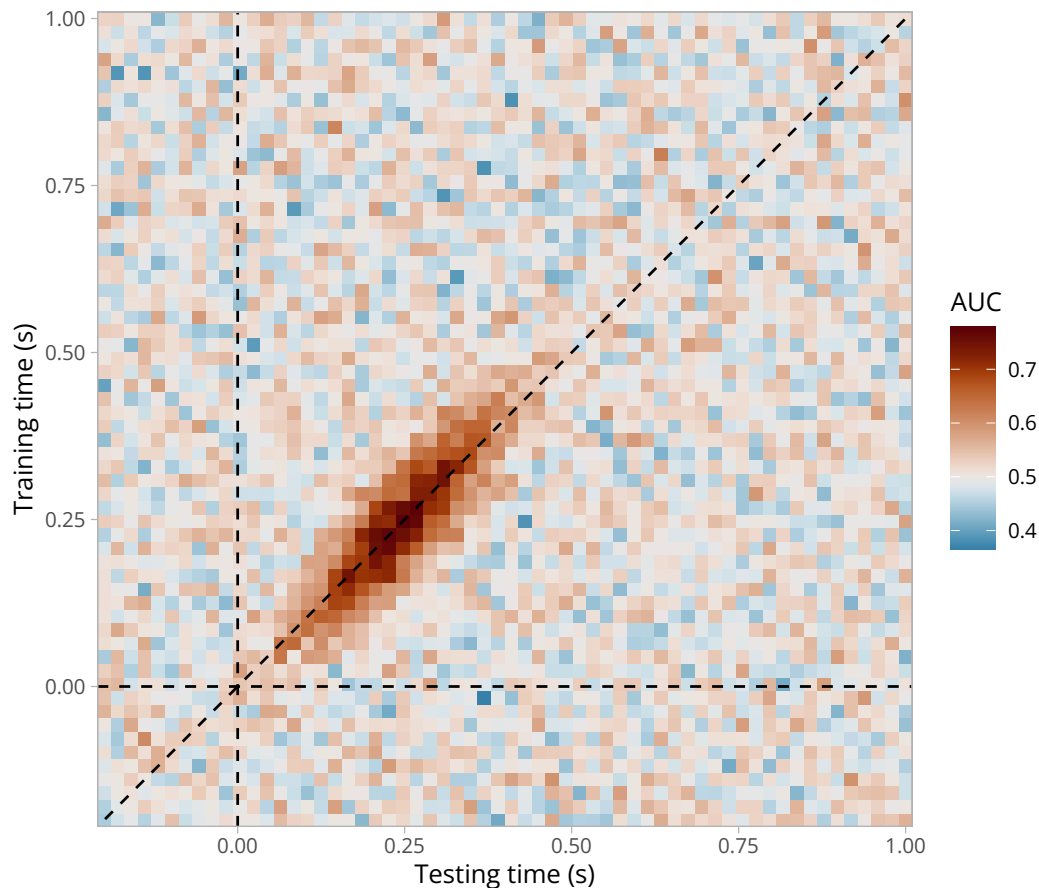
Appendix A

Application to 2D time-resolved decoding results (cross-temporal generalisation)

Assume we have M/EEG data and we have conducted cross-temporal generalisation analyses (King & Dehaene, 2014). As a result, we have a 2D matrix where each element contains the decoding accuracy (e.g., ROC AUC) of a classifier trained at timestep training_i and tested at timestep testing_j (Figure A1).

Figure A1

Exemplary (simulated) group-level average cross-temporal generalisation matrix of decoding performance (ROC AUC).



Now, we want to test whether and when decoding performance is above chance level (0.5 for a binary decoding task). These two models are computationally heavier to fit (more observations and 2D smooth functions)...

```
# fitting a GAM with two temporal dimensions
timegen_gam <- brm(
  # 2D thin-plate spline (tp)
  auc ~ t2(train_time, test_time, bs = "tp", k = 20),
  data = timegen_data,
  family = Beta(),
  iter = 5000,
  chains = 4,
  cores = 4,
  file = "models/timegen_gam_t2.rds"
)
```

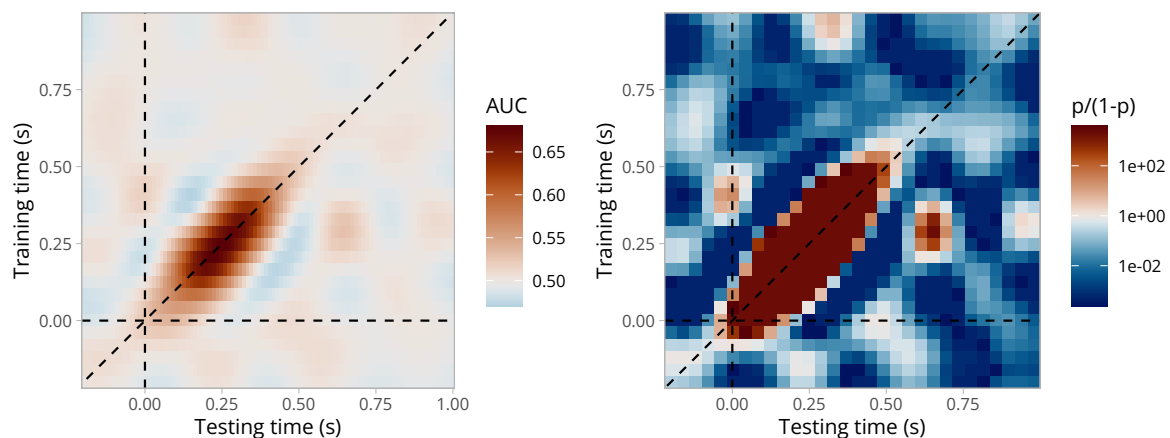
```

# fitting a GP with two temporal dimensions
# timegen_gp <- brm(
#   auc ~ gp(train_time, test_time, k = 20),
#   data = timegen_data,
#   family = Beta(),
#   control = list(adapt_delta = 0.95),
#   iter = 2000,
#   chains = 4,
#   cores = 4,
#   file = "models/timegen_gp.rds"
# )

```

Figure A2

Posterior probability of decoding accuracy being above chance level (2D GAM).



450 Could be extended to spatial and temporal dimensions with formulas such as `te(x, y,`
 451 `Time, d = c(2, 1))...`

Appendix B

Mathematical formulation of the bivariate GAM

452 To model cross-temporal generalisation matrices of decoding performance (ROC AUC), we
 453 extended the initial (decoding) GAM to take into account the bivariate temporal distribution
 454 of AUC values, thus producing naturally smoothed estimates (timecourses) of AUC values and
 455 posterior probabilities. This model can be written as follows:

$$\begin{aligned} \text{AUC}_i &\sim \text{Beta}(\mu_i, \phi) \\ g(\mu_i) &= f(\text{train}_i, \text{test}_i) \end{aligned}$$

456 where we assume that AUC values come from a Beta distribution with two parameters
 457 μ and ϕ . We can think of $f(\text{train}_i, \text{test}_i)$ as a surface (a smooth function of two variables) that
 458 we can model using a 2-dimensional splines. Let $\mathbf{s}_i = (\text{train}_i, \text{test}_i)$ be some pair of training and
 459 testing samples, and let $\mathbf{k}_m = (\text{train}_m, \text{test}_m)$ denote the m^{th} knot in the domain of train_i and
 460 test_i . We can then express the smooth function as:

$$f(\text{train}_i, \text{test}_i) = \alpha + \sum_{m=1}^M \beta_m b_m(\tilde{s}_i, \tilde{k}_m)$$

461 Note that $b_m(\cdot)$ is a basis function that maps $R \times R \rightarrow R$. A popular bivariate basis
 462 function uses *thin-plate splines* (Wood, 2003), which extend to $\mathbf{s}_i \in \mathbb{R}^d$ and ∂l_g penalties. These
 463 splines are designed to interpolate and approximate smooth surfaces over two dimensions (hence
 464 the “bivariate” term). For $d = 2$ dimensions and $l = 2$ (smoothness penalty involving second
 465 order derivative):

$$f(\tilde{s}_i) = \alpha + \beta_1 x_i + \beta_2 z_i + \sum_{m=1}^M \beta_{2+m} b_m(\tilde{s}_i, \tilde{k}_m)$$

466 using the the radial basis function given by:

$$b_m(\tilde{s}_i, \tilde{k}_m) = \left\| \tilde{s}_i - \tilde{k}_m \right\|^2 \log \left\| \tilde{s}_i - \tilde{k}_m \right\|$$

467 where $\|\mathbf{s}_i - \mathbf{k}_m\|$ is the Euclidean distance between the covariate \mathbf{s}_i and the knot location
 468 \mathbf{k}_m .

Appendix C

Integration with MNE-Python

469 Explain how to use the R package with MNE epochs...

```
# TO-DO: adding some code here...
```