Using cluster-based permutation tests to estimate MEG/EEG onsets: how bad is it?

Guillaume A. Rousselet

School of Psychology and Neuroscience, College of Medical, Veterinary and Life Sciences, University of Glasgow, 62 Hillhead Street, G12 8QB, Glasgow, UK

Guillaume.Rousselet@glasgow.ac.uk

https://orcid.org/0000-0003-0006-8729

Abstract

Localising effects in space, time and other dimensions is a fundamental goal of magneto- and electro-encephalography research. A popular exploratory approach applies mass-univariate statistics that are corrected for multiple comparisons using cluster inferences, by pooling together neighbouring effects. However, these cluster-based methods have an important limitation: because the inference is at the cluster level, there is no explicit localisation, and one must be cautious about interpreting when and where effects start and end. Sassenhagen & Draschkow (2019) provided an important reminder of this limitation. They also reported results from a simulation, suggesting that estimated onsets are both positively biased and underestimate real onsets in a large proportion of simulated experiments. However, the simulation did not provide results from other methods as a comparison. Using a new simulation, here I demonstrate that cluster-based inferences perform well relative to other standard methods that provide *p* values at every testing point: correcting for multiple comparisons using the false discovery rate adjustment leads to higher underestimation, while a maximum statistic leads to higher overestimation. Researchers using cluster-based inferences will report more accurate onset estimates in the long run than if they use these alternative methods. It is also worth keeping in mind that all common approaches to infer effect onsets rely on a statistical fallacy, because they never explicitly test the implied interaction between statistically significant and non-significant points. Better approaches to onset estimation can be implemented, for instance using change point detection algorithms.

KEYWORDS

Cluster inference, MEG, EEG, permutation, onset estimation, Monte-Carlo simulation, correction for multiple comparisons, family-wise error rate, false discovery rate

1 | Introduction

To localise effects in magnetoencephalography (MEG) and electroencephalography (EEG), one must apply a threshold to statistical maps. A popular frequentist approach consists in deriving a univariate threshold at every testing point, followed by a correction for multiple comparisons, often with the goal to control the family-wise error rate (FWER), the probability to make at least one type I error (false positive). A popular way to achieve this type of error control is to use cluster-based inferences, a family of methods that aim to correct for multiple comparisons while preserving high power when effects are distributed over time, space and other dimensions (Fields & Kuperberg, 2020; Frossard & Renaud, 2022; Maris & Oostenveld, 2007; Pernet et al., 2015; Rosenblatt et al., 2018; Rousselet et al., 2014; Smith & Nichols, 2009). These methods capitalise on the correlation of effects across measurements to control the family-wise error rate (FWER, probability of at least one false positive): neighbouring effects that pass a threshold are lumped together, and such combinations of effects are then compared to a reference distribution, typically a null distribution established using a permutation or a bootstrap approach (Pernet et al., 2015). These techniques are particularly popular in fMRI, EEG and MEG research (for instance, as of November 2023, according to Google Scholar, Smith & Nichols (2009) has been cited 4962 times, Maris & Oostenveld (2007) 7088 times). Cluster-based methods also have the potential to be applied to a wider array of problems in neuroscience (Dugué et al., 2011; Rousselet, 2018).

However, cluster-based methods are not without issues. Most notably, Sassenhagen & Draschkow (2019) reminded the community about limitations in the interpretation of cluster-based inferences and suggested guidelines on how to report such results. This useful reminder has been heard, as demonstrated by the 322 citations of the article, according to Google Scholar. Essentially, the main problem with cluster-based methods is that they provide *p* values at the cluster level only, or even for a collection of clusters for hierarchical/meta-analysis methods such as TFCE (Ehinger, 2019; Smith & Nichols, 2009). There are no *p* values at individual testing points, and no error control for onsets. Consequently, cluster-based methods do not provide direct inferences about the localisation of effects. But does that mean we cannot use cluster-based methods to localise, for instance to estimate effect onsets?

To address this question, Sassenhagen & Draschkow (2019) reported results from a simulation, which suggests that onset estimation using cluster-based inferences is very noisy. They used a truncated Gaussian embedded in a sequence of zeros as a template, to define a real onset to be estimated. Then, they compared 50 trials of noise only, to 50 trials of noise + template, which were compared at every testing point using *F* statistics. Correction for multiple comparisons was achieved by using cluster-sum statistics: neighbouring *F* values that passed a threshold were summed, and these cluster-sums were compared to a permutation (null) distribution of maximum cluster-sums to determine statistical significance. For each simulation iteration, the first statistically significant point was declared the effect onset. The distribution of estimated onsets across 10,000 simulation iterations revealed two important patterns: estimated onsets were positively biased, with the distribution mode higher than the real onset; the distribution was asymmetric, with a thicker left

tail than right tail. Sassenhagen & Draschkow (2019) reported that "on >20% of runs, the effect onset was estimated too early; divergence of 40 ms or more were found at >10% of runs." No bias estimate was reported.

So, how bad are these results for cluster-sum inferences? We don't know because no other method was used to estimate onsets, so we cannot put the results in perspective. Here I address the comparison problem by contrasting cluster-based inferences with inferences from two other popular methods, the false discovery rate correction for multiple comparisons (FDR, Benjamini & Hochberg, 1995; T. Nichols & Hayasaka, 2003) and the maximum statistics correction for multiple comparisons (MAX, Holmes et al., 1996; T. E. Nichols & Holmes, 2002). FDR methods provide weak control of the FWER by controlling the number of false positives among the positive tests; when there is no effect, an FDR correction is equivalent to controlling the FWER; when effects are present, it is more powerful than controlling the FWER (Benjamini & Hochberg, 1995; Fields & Kuperberg, 2020). This increased power comes at a cost, though, because when the global null is not true, the FDR correction does not guarantee control over the FWER for subsets of hypotheses, which means that FDR corrected statistical maps cannot be used to localise effects, similarly to cluster-based inferences (T. Nichols & Hayasaka, 2003). In contrast, the MAX correction achieves strong control of the FWER by comparing each statistic, say a *t* or *F* value, to a distribution of maximum statistics obtained under the null. MAX corrected statistical maps can be used to localise effects, at the cost of lower power. The FDR and MAX methods, like cluster-based methods, do not explicitly estimate onsets, but are nevertheless routinely used to localise effects.

Other methods have been proposed, for instance defining the onset as the time at which a certain proportion of the peak amplitude has been reached (Fahrenfort, 2020; Liesefeld, 2018). However, such methods require that an effect has already been detected, and they depend on potentially subjective peak identification, which, from experience, can be problematic when making inferences in individual participants. Fractional methods are also guaranteed, by definition, to lead to positively biased onset estimates. So these methods are not considered here but maybe they have some merit.

Another recent method combines cluster inferences while delivering *p* values at every testing point, thus solving the main issue of cluster-based inferences as implemented so far (Frossard & Renaud, 2022). But this approach, like previous cluster-based methods, and alternative ways to correct for multiple comparisons such as FDR or MAX, do not solve a fundamental issue that tends to be ignored: whether a *p* value is available at each testing point or not is irrelevant, as none of the methods routinely used to localise MEG and EEG effects explicitly assess onsets. That's because onsets are second order properties of the data. As such, they require to establish the lack of effects in some areas, the presence of effects in others, and a test of the interaction. However, none of the cluster-based methods, or any other standard methods used to localise in EEG/MEG/fMRI and other fields, explicitly test for the interaction. The interaction is an inference performed essentially by eyeballing the results. So localisation attempts typically rely on an interaction fallacy,

a general problem in neuroscience (Nieuwenhuis et al., 2011) and other fields (Gelman & Stern, 2006).

More explicit estimation of onsets could be achieved by using methods that go beyond mass-univariate tests and consider time-courses explicitly. For instance, there is a vast family of algorithms aimed at detecting change points in time-series that could be used to estimate MEG/EEG onsets (Aminikhanghahi & Cook, 2017; Chen & Gupta, 2013; Lindeløv, 2020, 2023; Zhao et al., 2013). Here, for comparison with cluster-based, FDR and MAX methods, I used a basic change point detection method that relies on differences in mean and variance, as implemented in the *changepoint* package (Killick et al., 2022; Killick & Eckley, 2014) in R (R Core Team, 2021). Unlike cluster-based, FDR and MAX methods, the change point method was applied directly to the time-course of *t* values, so it relies on a very different type of inference. Let's compare the distributions of onsets obtained using these different methods in a simulation.

2 | Methods

Following the approach used by Sassenhagen & Draschkow (2019), I compared two conditions, one with noise only, one with noise + signal. 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). Examples of 4 noise-only trials are presented in Figure 1A. On each trial, the noise was normalised to have a variance of 1. Increasing the variance to 1.5 affected the Monte-Carlo summary statistics of all methods, but did not affect the conclusion of the simulations. The signal was a truncated Gaussian defining an objective onset at 160 ms, a maximum at 250 ms, and an offset at 342 ms (Figure 1B). The sampling rate was 500 Hz (a sample every 2 ms).

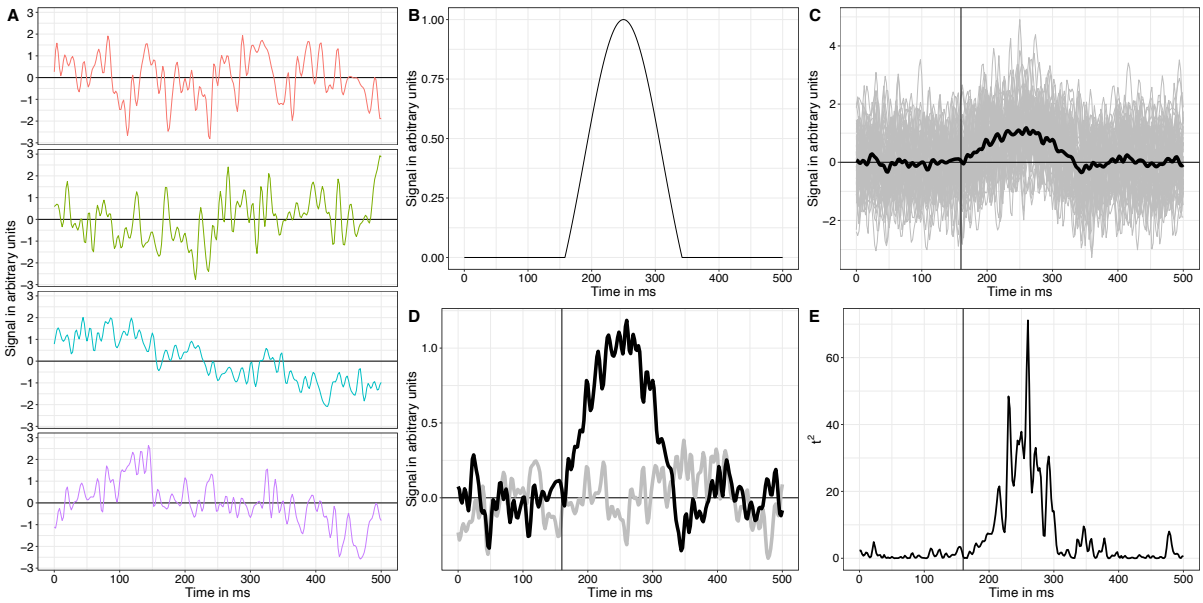


Figure 1. Data generation process. (A) Examples of 4 noise trials. (B) Truncated Gaussian used as signal. (C) Example of 50 trials of noise + signal (grey), with the superimposed mean (black). (D)

Example of averages of 50 trials for the noise only condition (grey) and the noise + signal condition (black). (E) Time-course of squared *t* values from a series of t-tests for two independent samples with unequal variances.

Each Monte-Carlo simulation consisted of 10,000 iterations, starting with the creation of 50 trials of the noise only condition, and 50 trials of the noise + signal condition (Figure 1C, D). The two conditions were then compared at every time point using a t-test for two independent samples with unequal variances. This approach is very similar to that of Sassenhagen & Draschkow (2019), except that I simulated only one electrode, with cluster inferences only in the time domain. There is no need to simulate extra dimensions (electrodes, time-frequencies), because the argument against cluster-based inferences is not specific to one dimension. Also, I used synthetic noise instead of using data from one participant.

Null distributions were estimated using a permutation with 2,000 iterations (Maris & Oostenveld, 2007; Pernet et al., 2015), which were used to compute univariate *p* values and cluster forming thresholds. Four methods for multiple comparison corrections were considered. (1) A cluster-based inference was implemented using a cluster-sum statistic of squared *t* values, using my own R version of the implementation in the *LIMO EEG* toolbox (Pernet et al., 2011, 2015). I will refer to this method as cluster-sum or method CS in the results. (2) The original false discovery rate (FDR) correction (Benjamini & Hochberg, 1995) was applied using the *p.adjust* function in R. (3) The maximum statistics (MAX) correction (Holmes et al., 1996; T. E. Nichols & Holmes, 2002) was applied by using the (1-alpha) quantile of the permutation distribution of maximum t^2 values. (4) Change point detection was applied to the time-course of t^2 values using the *cpt.meanvar* function, with method *BinSeg*, in the *changepoint* R package (Killick et al., 2022; Killick & Eckley, 2014). This algorithm looked for abrupt differences in mean and variance (Chen & Gupta, 2013). The number of expected change points was set to 2 and the first of these two change points was recorded as the effect onset. Other parameters were left to their default values. Using more than 2 expected change points would lead to higher variability in onset estimation, thus this approach provides a good benchmark about what can be achieved with a simple change detection algorithm.

Alpha was set to the arbitrary value of 0.05. For the cluster-sum, FDR and MAX methods, I discarded the first cluster of significant points if it included zero, as this is unrealistic (Bieniek et al., 2016). Including onsets at zero ms was only detrimental to the FDR method.

Figure 2 shows the onsets estimated by each method, using the data from the example in Figure 1. In grey is the permutation distribution of squared *t* values. The solid vertical line indicates the true onset. The dotted vertical line marks the estimated onset. The horizontal dashed lines are permutation thresholds.

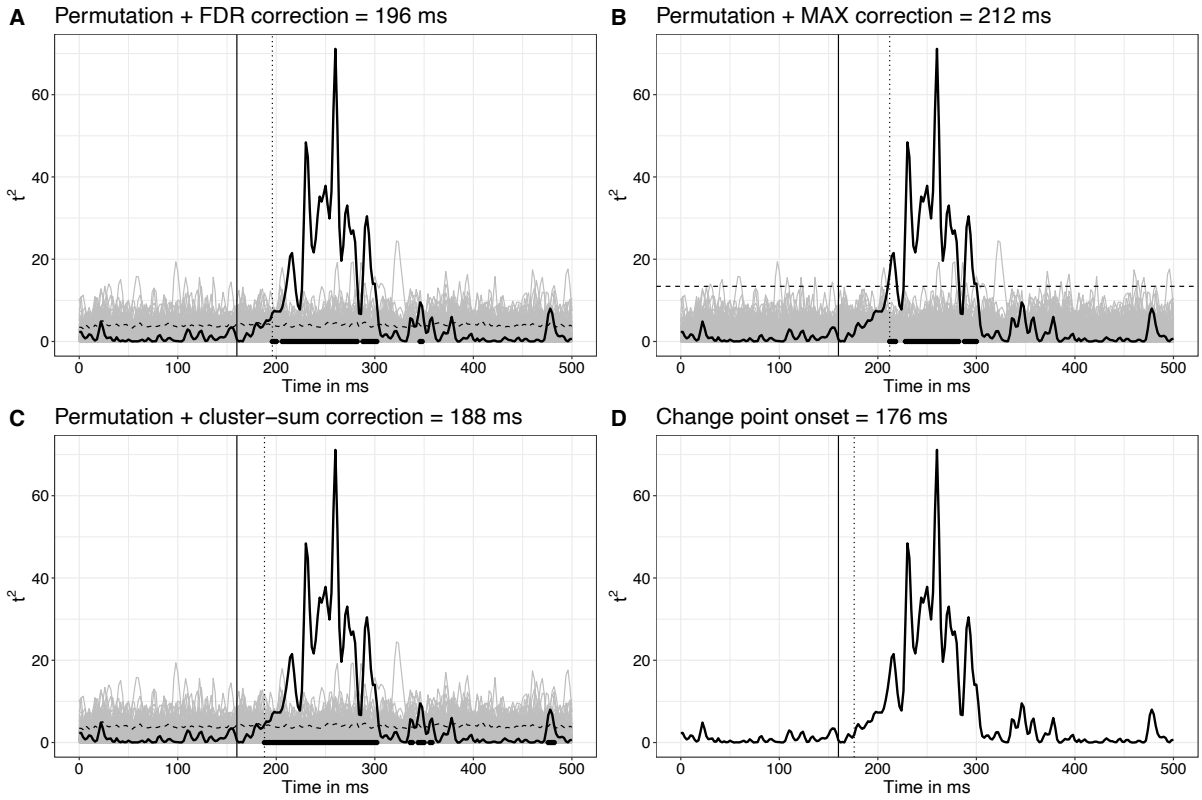


Figure 2. Examples of onset inferences. The black vertical lines indicate the true onset (160 ms). The dotted vertical lines mark the estimated onsets. Each panel shows the same time-course of t^2 values in black. The superimposed grey lines are the 2,000 t^2 values from the permutation. Dashed horizontal lines mark the 1-alpha quantiles of the permutation distributions. (A) The permutation distributions were used to compute a *p* value at each time point, before applying the FDR correction. (B) For each permutation, the maximum value was saved, and the 1-alpha quantile of the distribution of maxima is illustrated as a straight dashed line. (C) The dashed line is the same as in panel A. Neighbouring points above these univariate thresholds were clustered, and for each cluster the sum of the t^2 values was assessed against the 1-alpha quantile of the permutation distribution of maximum cluster-sums. (D) The change point algorithm was applied directly to the time-course of t^2 values, without considering the permutation distribution.

To assess the accuracy of group onset estimation, I ran a simulation like the one already described, but now with 20 participants. In each simulation iteration, each participant was assigned a random onset between 150 to 170 ms, drawn from a uniform distribution with 2 ms steps. Once the onsets were estimated for every participant, group onsets were quantified as the median onset across participants, separately for each method.

Onset estimation methods were assessed using these summary statistics of the Monte-Carlo sampling distributions of onsets: bias, mean absolute error and variance (Aminikhanghahi & Cook, 2017; Morris et al., 2019). Following Sassenhagen & Draschkow (2019), I also report the proportion of estimated onsets that occurred before the true onset (proportion too early). Bias was quantified by subtracting the true onset from the median of each sampling distribution. Because the distributions are asymmetric, the median provides a better reflection of the typical onset than the mean (Rousselet & Wilcox, 2020).

All the figures and simulations presented here can be reproduced using R code available on GitHub: https://github.com/GRousselet/onsetsim. Similar results were obtained using a Matlab version of the same analyses (The MathWorks Inc., 2021). Only the R results are presented here. I also performed simulations using less complex 1/f type of noise (Stevens, 2009). Using pink noise gave very similar results to the ones reported here, but using white noise substantially altered the behaviour of the FDR method – see full simulations on GitHub. This doesn't affect the main conclusion of the simulations: there is nothing wrong about using cluster-based inferences to estimate onsets.

3 | Results

After 10,000 iterations, for our four methods, we get the distributions of estimated onsets in Figure 3. Each distribution has a shape similar to that reported by Sassenhagen & Draschkow (2019): overall a positive bias, with the modes shifted to right of the true onset (black vertical line), and some amount of under-estimation, which is particularly visible for FDR.

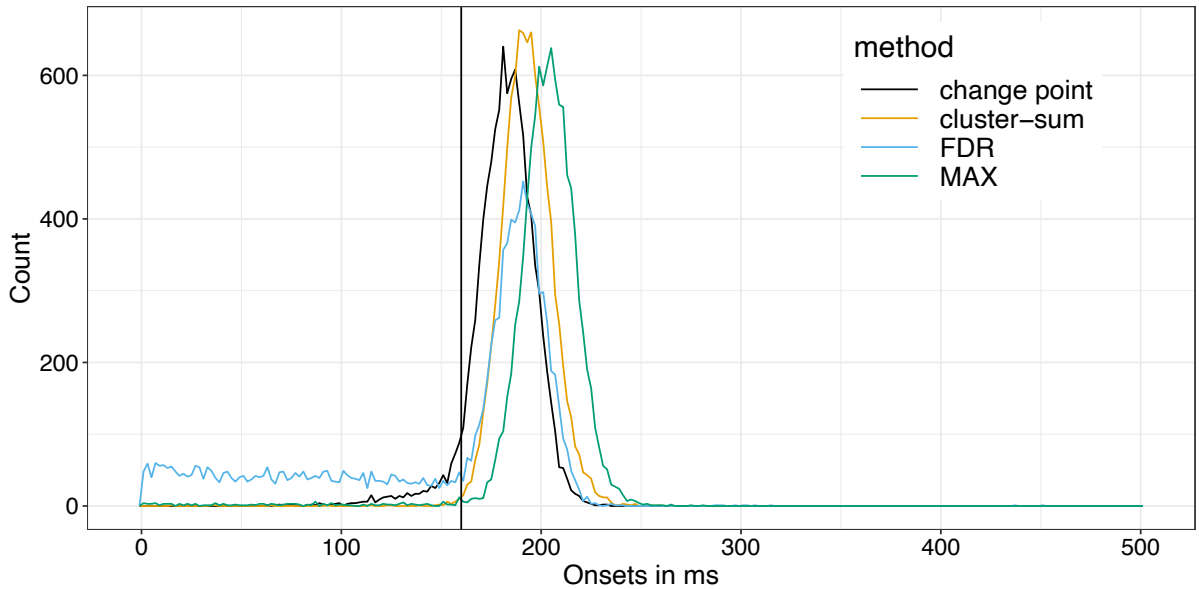


Figure 3. Sampling distributions of estimated onsets. The vertical black line indicates the true onset (160 ms).

We can quantify the results by considering summary statistics of these distributions: bias, mean absolute error (MAE), variance, and proportion too early (before the true onset). These summary statistics are reported in Table 1.

Statistic	FDR	MAX	CS	CP
Bias	22 ms	44 ms	34 ms	24 ms
MAE	48.6 ms	45.8 ms	34 ms	24.8 ms
Variance	3777 ms^2	447 ms^2	166 ms^2	263 ms^2
% too early	32.3 %	1.5 %	0.3 %	5.6 %

Table 1. Summary statistics of the sampling distributions presented in Figure 3. FDR = false discovery rate; MAX = maximum statistics; CS = cluster-sum; CP = change point. MAE = mean absolute error.

Confirming our visual inspection of the distributions, all methods are positively biased: the least biased were FDR and change point, the most biased was MAX; cluster-sum fell in between. The high bias of MAX is not surprising, as it is by definition conservative, the price to pay for strong FWER control. The lower bias of FDR is however deceptive: the mode of the onset distribution (192 ms) is similar to that of cluster-sum (190 ms), but the FDR distribution also has a very thick left tail. As a result, FDR is associated with the largest mean absolute error (MAE) and the largest variance. In contrast, cluster-sum had the second lowest MAE, and the lowest variance among the four methods. In comparison, FDR had a variance 23 times larger than cluster-sum. As Sassenhagen & Draschkow (2019) did, if we consider the proportion of onsets earlier than the true onset, again, there is nothing wrong with the cluster-sum approach: all other methods performed worst.

These results are not specific to a sample size of 50 trials per condition, as demonstrated in Figure 4, which presents results for sample sizes ranging from 20 to 150 trials. Bias can be reduced considerably by using larger sample sizes, but the MAX method remains overly conservative (Figure 4A). Mean absolute error, variance and proportion of early onsets all improve with increasing sample size too, except for FDR. Again, cluster-sum performed well: despite relatively high bias (still 18 ms for n=150), it has lower MAE than FDR and MAX, and the lowest variance and proportion of too early onsets of all methods. The change point approach looks very promising, with relatively low bias, MAE and variance, but higher under-estimation that MAX and cluster-sum.

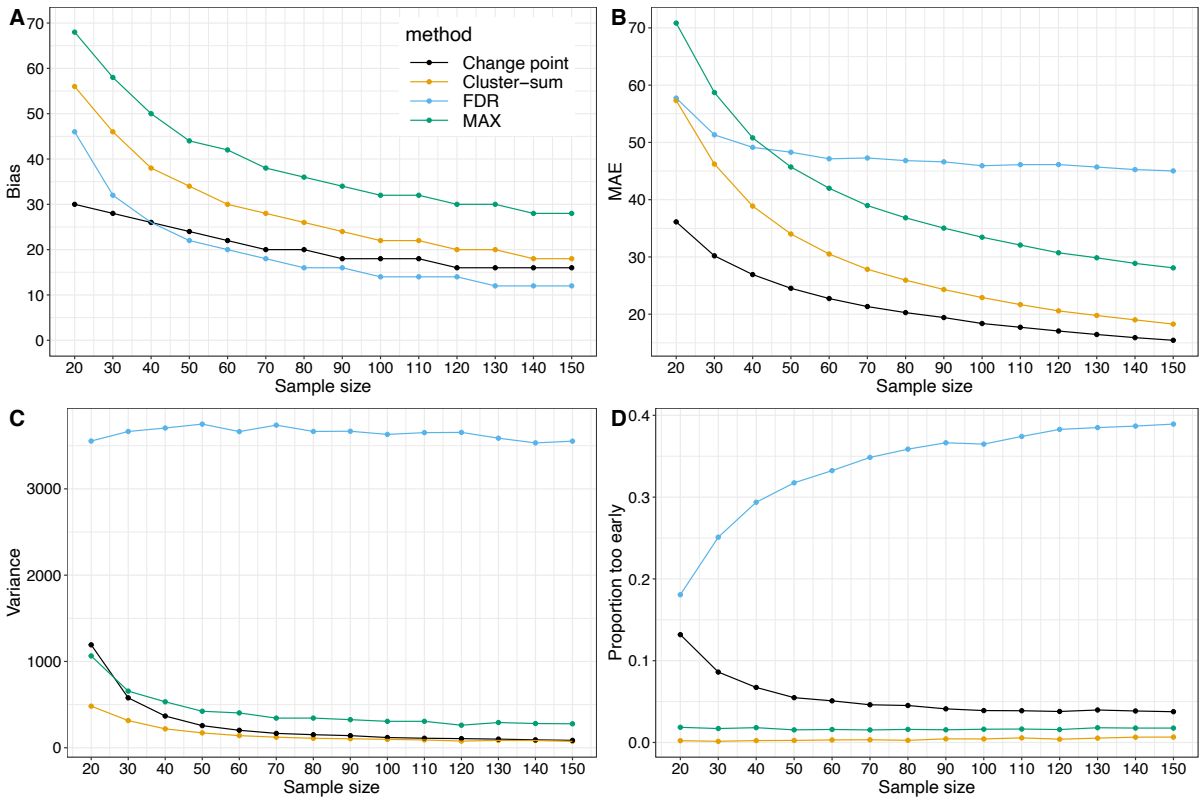


Figure 4. Summary statistics as a function of sample size.

So far we have considered results from one simulated participant. What happens in a more realistic situation in which we try to estimate onsets from 20 participants instead of one? To find out, let's assume that each participant has a random onset between 150 and 170 ms, and 50 trials per condition. In each simulation iteration, after estimating the onset in each of the 20 participants, I used the median onset as the group estimate. This leads to the sampling distributions of estimated group onsets in Figure 5.

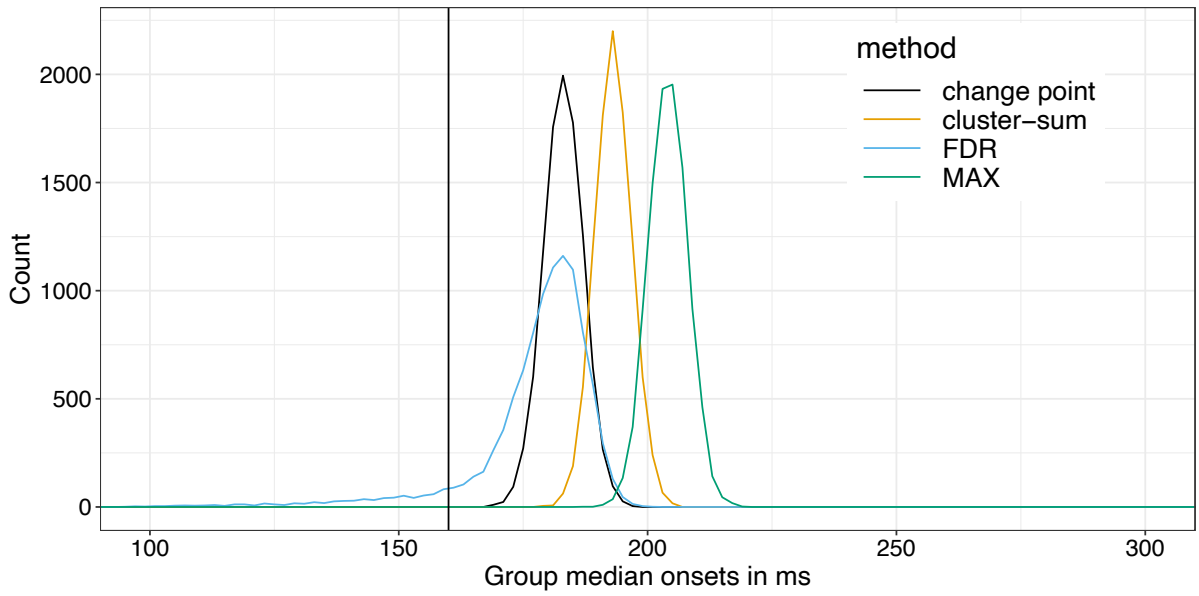


Figure 5. Sampling distributions of estimated group onsets. The vertical black line indicates the true onset (160 ms).

The group results clearly distinguish the four methods, with, in increasing order of bias and MAE: FDR, change point, cluster-sum and MAX (Table 2). Variance is again dramatically larger for FDR than for the other methods, which do not differ much from each other. FDR is also the only method that produces group under-estimation. For the other methods, combining onsets, using the median across participants, eliminates under-estimation.

Statistic	**FDR**	**MAX**	**CS**	**CP**
Bias	21	45	34	24
MAE	20.6	44.7	33.6	23.6
Variance	166	15	14	16
% too early	6.9 %	0 %	0 %	0 %

Table 2. Summary statistics of the sampling distributions presented in Figure 5. FDR = false discovery rate; MAX = maximum statistics; CS = cluster-sum; CP = change point. MAE = mean absolute error.

4 | Discussion

Onsets are inherently noisy because they correspond to the testing points with the lowest signal to noise ratio. As a result, we can expect most onset estimation methods to produce positively biased and variable results. But both estimation bias and variability can be reduced by increasing the number of trials and pooling estimates across participants. And based on the present simulation results, a cluster-sum approach performs well relative to other standard methods, FDR and MAX. In fact, if the goal is to minimise bias and variability, cluster-sum provides a good compromise between the two methods. But considering the group simulation results, the change point approach performs even better, even though I used a relatively simple algorithm. It remains to be determined what can be achieved using modern methods, for instance Bayesian approaches that allow hierarchical pooling of estimates (Lindeløv, 2020).

Whether change point detection algorithms can replace cluster-based inferences to estimate onsets will require more simulations tackling a variety of situations. Even if other methods can do better, the simulation results presented here do not demonstrate any alarming behaviour of cluster-based onset estimation that would require researchers to stop using them. So were Sassenhagen & Draschkow (2019) wrong to remind us that cluster-based tests do not establish the statistical significance of effect latency or location? No, but we need to recognise that this statement applies to all default MEG/EEG techniques: whether a *p* value, effect size or posterior distribution is available at every time point or not, no default technique explicitly localises, let alone estimates onsets. Whether correction for multiple comparison is achieved using cluster-based statistics, FDR, MAX or any related method, even the new cluster depth tests (Frossard & Renaud, 2022), onset

estimation relies on a statistical fallacy, because the implied interaction is never explicitly tested. Does it mean that non-sensical results are obtained? Simulations are an excellent way to answer that question, letting users decide on a method depending on their priorities, in the long run, for instance minimising bias or variance. Certainly, based on the current simulations, it seems that cluster-based inferences can provide more accurate results than some alternative methods. So, even if cluster-based methods do not afford explicit inferences about onsets, they can nevertheless be used to estimate them. Methods can be used for purposes other than initially intended as long as we can demonstrate appropriate performance.

When estimating onsets, several steps can be taken to improve and assess the precision of the results. To extend my simple simulations, one could use algorithms that generate realistic time-series of various types of effects (Barzegaran et al., 2019; Julius Vetter et al., 2023). When dealing with real data, instead of estimating onsets at the group level, it is better practice to measure onsets in each participant and to report group distributions (Bieniek et al., 2016). It is also very informative to assess the test-retest reliability of the onsets, either using the same data (via cross-validation), or data from a different EEG session. For instance, test-retest assessment of face ERP onsets suggest that the individual variability in onsets is relatively stable across sessions, indicating that individual differences reflect more than measurement noise. Further, computational consistency can be assessed by reporting results from more than one onset estimation method. Finally, it is worth considering other methodological aspects that can affect onsets. The most important factor is probably filtering, with non-causal high-pass filters introducing potentially severe signal distortions (Rousselet, 2012; van Driel et al., 2021; Widmann et al., 2015; Widmann & Schröger, 2012). Instead, causal filters can be very effective at attenuating low temporal frequencies while avoiding backpropagation of effects, both of which can distort onsets. Using data spatially deblurred could also improve onset estimation, for instance by working with independent components or current source densities (Delorme & Makeig, 2004; Tenke & Kayser, 2012). On the stimulus front, screen luminance can strongly influence onsets and should be considered when comparing results between studies (Bieniek et al., 2013).

Conflict of interest

I don't have a conflict of interest, other than having reported cluster-based inferences in several publications.

Data and code availability

The R and Matlab code necessary to reproduce the simulations is available on GitHub: https://github.com/GRousselet/onsetsim. The repository also contains simulation results and code to reproduce the figures in this article, without running the simulations.

References

Aminikhanghahi, S., & Cook, D. J. (2017). A survey of methods for time series change point detection. *Knowledge and Information Systems*, *51*(2), 339–367. https://doi.org/10.1007/s10115-016-0987-z

Barzegaran, E., Bosse, S., Kohler, P. J., & Norcia, A. M. (2019). EEGSourceSim: A framework for realistic simulation of EEG scalp data using MRI-based forward models and biologically plausible signals and noise. *Journal of Neuroscience Methods*, *328*, 108377. https://doi.org/10.1016/j.jneumeth.2019.108377

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 (Methodological)*, *57*(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Bieniek, M. M., Bennett, P. J., Sekuler, A. B., & Rousselet, G. A. (2016). A robust and representative lower bound on object processing speed in humans. *European Journal of Neuroscience*, *44*(2), 1804–1814. https://doi.org/10.1111/ejn.13100

Bieniek, M. M., Frei, L., & Rousselet, G. (2013). Early ERPs to faces: Aging, luminance, and individual differences. *Frontiers in Psychology*, *4*. https://www.frontiersin.org/articles/10.3389/fpsyg.2013.00268

Chen, J., & Gupta, A. K. (2013). *Parametric Statistical Change Point Analysis*. Springer Science & Business Media.

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. https://doi.org/10.1016/j.jneumeth.2003.10.009

Dugué, L., Marque, P., & VanRullen, R. (2011). Transcranial Magnetic Stimulation Reveals Attentional Feedback to Area V1 during Serial Visual Search. *PLOS ONE*, *6*(5), e19712. https://doi.org/10.1371/journal.pone.0019712

Ehinger, B. (2019). Threshold Free Cluster Enhancement Explained. *Benedikt Ehinger's Blog*. https://benediktehinger.de/blog/science/threshold-free-cluster-enhancement-explained

Fahrenfort, J. J. (2020). Multivariate Methods to Track the Spatiotemporal Profile of Feature-Based Attentional Selection Using EEG. In S. Pollmann (Ed.), *Spatial Learning and Attention Guidance* (pp. 129–156). Springer US. https://doi.org/10.1007/7657_2019_26

Fields, E. C., & Kuperberg, G. R. (2020). Having your cake and eating it too: Flexibility and power with mass univariate statistics for ERP data. *Psychophysiology*, *57*(2), e13468. https://doi.org/10.1111/psyp.13468

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

Gelman, A., & Stern, H. (2006). The Difference Between "Significant" and "Not Significant" is not Itself Statistically Significant. *The American Statistician*, *60*(4), 328–331. https://doi.org/10.1198/000313006X152649

Holmes, A. P., Blair, R. C., Watson, J. D., & Ford, I. (1996). Nonparametric analysis of statistic images from functional mapping experiments. *Journal of Cerebral Blood Flow and Metabolism: Official Journal of the International Society of Cerebral Blood Flow and Metabolism*, *16*(1), 7–22. https://doi.org/10.1097/00004647-199601000-00002

Julius Vetter, Jakob H. Macke, & Richard Gao. (2023). Generating realistic neurophysiological time series with denoising diffusion probabilistic models. *bioRxiv*, 2023.08.23.554148. https://doi.org/10.1101/2023.08.23.554148

Killick, R., & Eckley, I. A. (2014). changepoint: An R Package for Changepoint Analysis. *Journal of Statistical Software*, *58*, 1–19. https://doi.org/10.18637/jss.v058.i03

Killick, R., Haynes, K., & Eckley, I. A. (2022). *changepoint: An R package for changepoint analysis* (R package version 2.2.4) [Computer software]. https://CRAN.R-project.org/package=changepoint

Liesefeld, H. R. (2018). Estimating the Timing of Cognitive Operations With MEG/EEG Latency Measures: A Primer, a Brief Tutorial, and an Implementation of Various Methods. *Frontiers in Neuroscience*, *12*. https://www.frontiersin.org/articles/10.3389/fnins.2018.00765

Lindeløv, J. K. (2020). mcp: An R Package for Regression With Multiple Change Points. *OSF Preprints*. https://doi.org/10.31219/osf.io/fzqxv

Lindeløv, J. K. (2023). *An overview of change point packages in R*. Mcp: Regression with Multiple Change Points. https://lindeloev.github.io/mcp/articles/packages.html

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

Morris, T. P., White, I. R., & Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, *38*(11), 2074–2102. https://doi.org/10.1002/sim.8086

Nichols, T. E., & Holmes, A. P. (2002). Nonparametric permutation tests for functional neuroimaging: A primer with examples. *Human Brain Mapping*, *15*(1), 1–25. https://doi.org/10.1002/hbm.1058

Nichols, T., & Hayasaka, S. (2003). Controlling the familywise error rate in functional neuroimaging: A comparative review. *Statistical Methods in Medical Research*, *12*(5), 419–446. https://doi.org/10.1191/0962280203sm341ra

Nieuwenhuis, S., Forstmann, B. U., & Wagenmakers, E.-J. (2011). Erroneous analyses of interactions in neuroscience: A problem of significance. *Nature Neuroscience*, *14*(9), 1105–1107. https://doi.org/10.1038/nn.2886

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*, e831409. 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. (2021). *R: A Language and Environment for Statistical Computing* [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/

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. (2012). Does Filtering Preclude Us from Studying ERP Time-Courses? *Frontiers in Psychology*, *3*. https://www.frontiersin.org/articles/10.3389/fpsyg.2012.00131

Rousselet, G. A. (2018). Cluster correction for multiple dependent comparisons. *Basic Statistics*. https://garstats.wordpress.com/2018/09/06/cluster/

Rousselet, G. A., Ince, R. A. A., van Rijsbergen, N. J., & Schyns, P. G. (2014). Eye coding mechanisms in early human face event-related potentials. *Journal of Vision*, *14*(13), 7. https://doi.org/10.1167/14.13.7

Rousselet, G. A., & Wilcox, R. R. (2020). Reaction Times and other Skewed Distributions: Problems with the Mean and the Median. *Meta-Psychology*, *4*. https://doi.org/10.15626/MP.2019.1630

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), e13335. https://doi.org/10.1111/psyp.13335

Smith, S. M., & Nichols, T. E. (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

Stevens, M. H. H. (2009). *A Primer of Ecology with R, 2nd printing*. Springer.

Tenke, C. E., & Kayser, J. (2012). Generator localization by current source density (CSD): Implications of volume conduction and field closure at intracranial and scalp resolutions. *Clinical Neurophysiology*, *123*(12), 2328–2345. https://doi.org/10.1016/j.clinph.2012.06.005

The MathWorks Inc. (2021). *MATLAB version: 9.10.0.1684407 (R2021a)* [Computer software]. The MathWorks Inc. https://www.mathworks.com

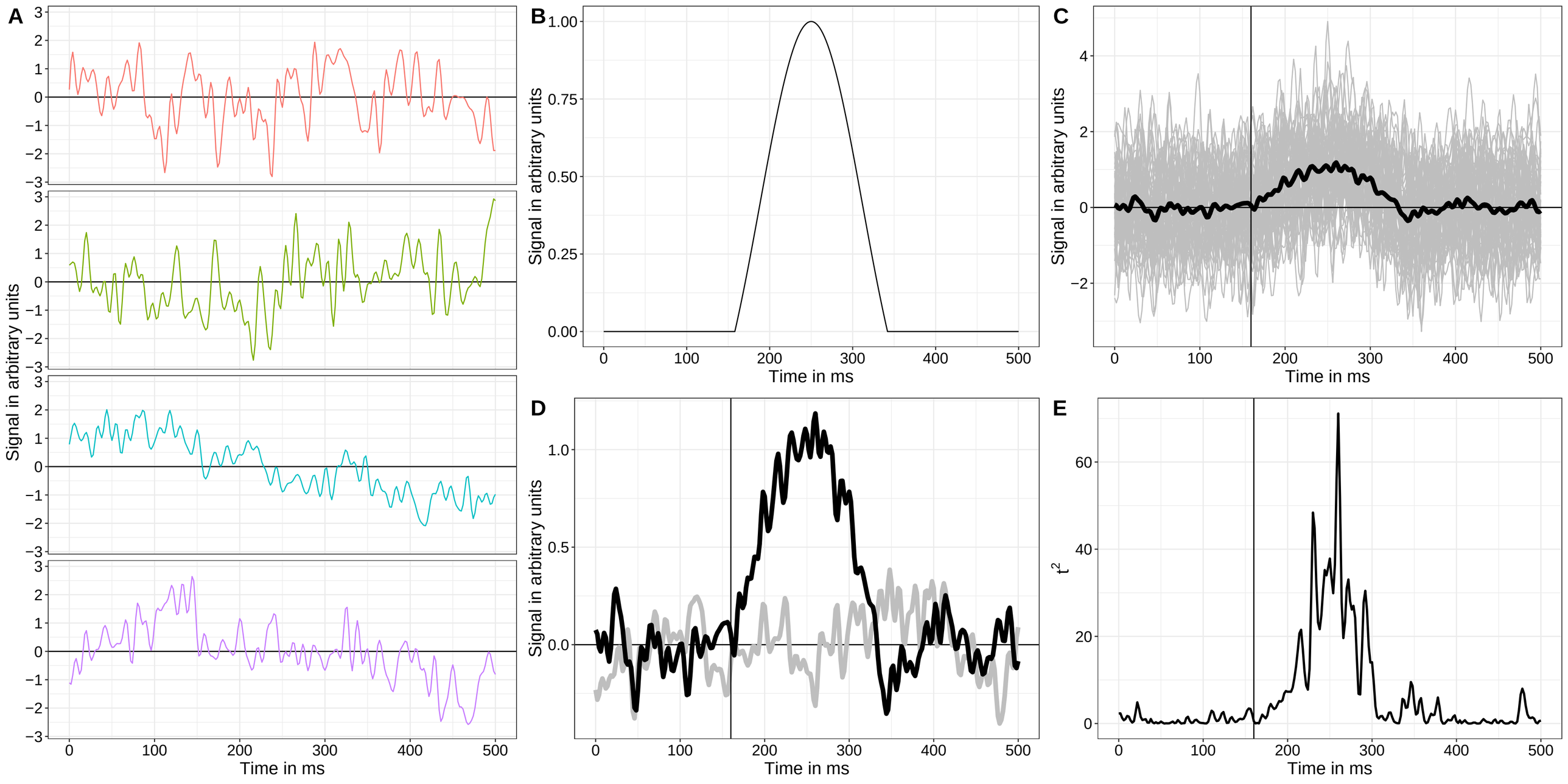
van Driel, J., Olivers, C. N. L., & Fahrenfort, J. J. (2021). High-pass filtering artifacts in multivariate classification of neural time series data. *Journal of Neuroscience Methods*, *352*, 109080. https://doi.org/10.1016/j.jneumeth.2021.109080

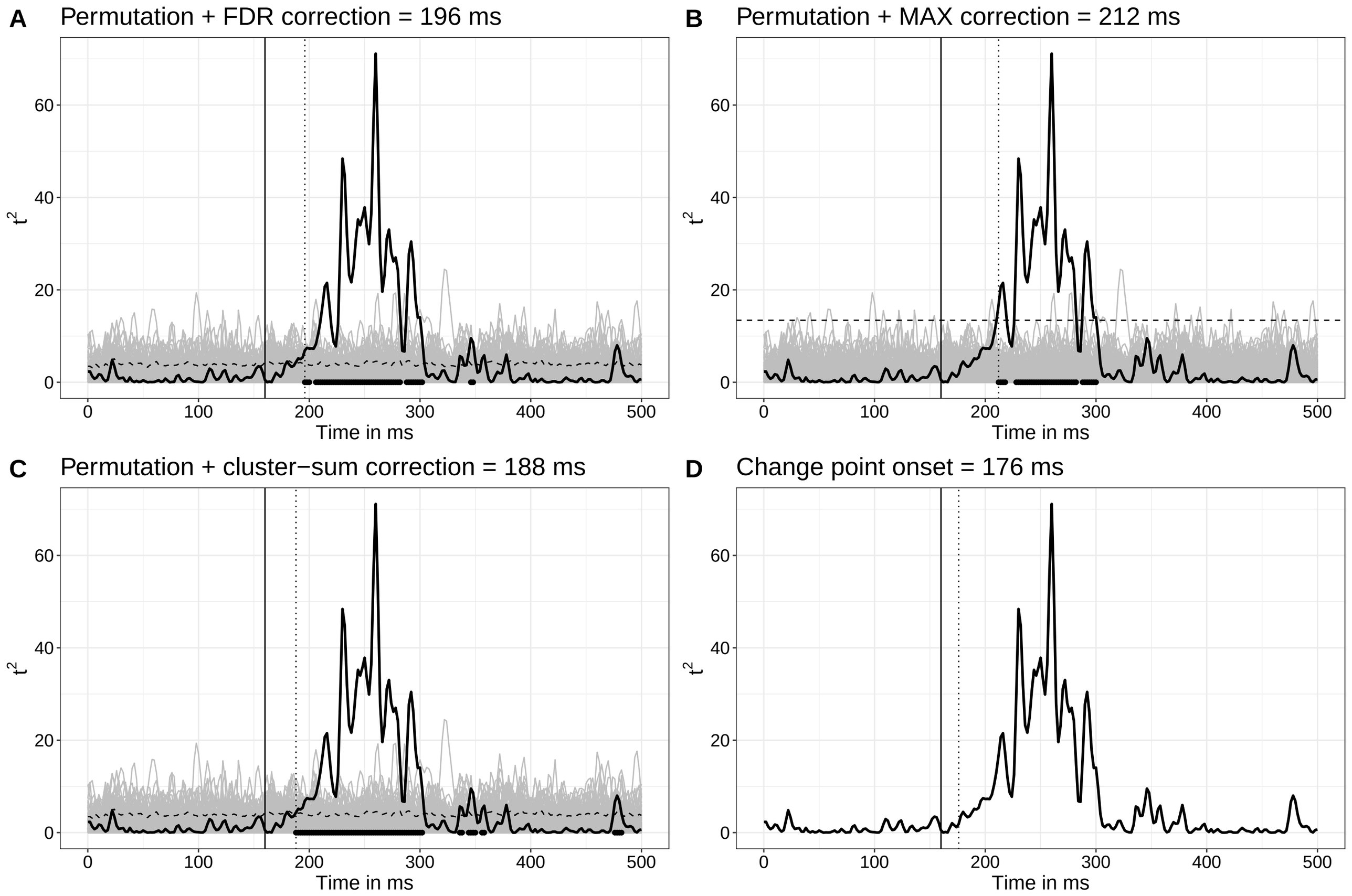
Widmann, A., & Schröger, E. (2012). Filter Effects and Filter Artifacts in the Analysis of Electrophysiological Data. *Frontiers in Psychology*, *3*. https://www.frontiersin.org/articles/10.3389/fpsyg.2012.00233

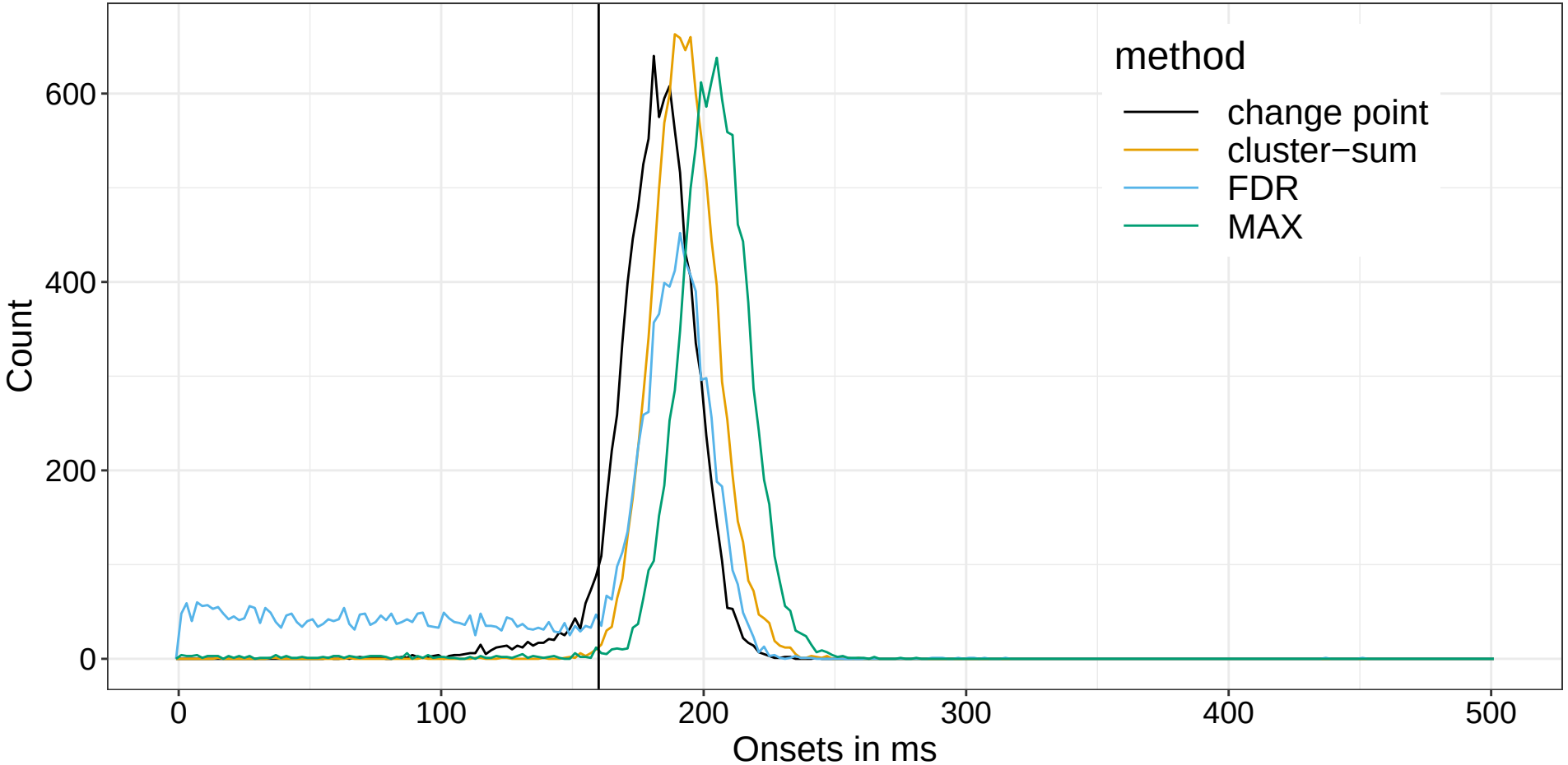
Widmann, A., Schröger, E., & Maess, B. (2015). Digital filter design for electrophysiological data – a practical approach. *Journal of Neuroscience Methods*, *250*, 34–46. https://doi.org/10.1016/j.jneumeth.2014.08.002

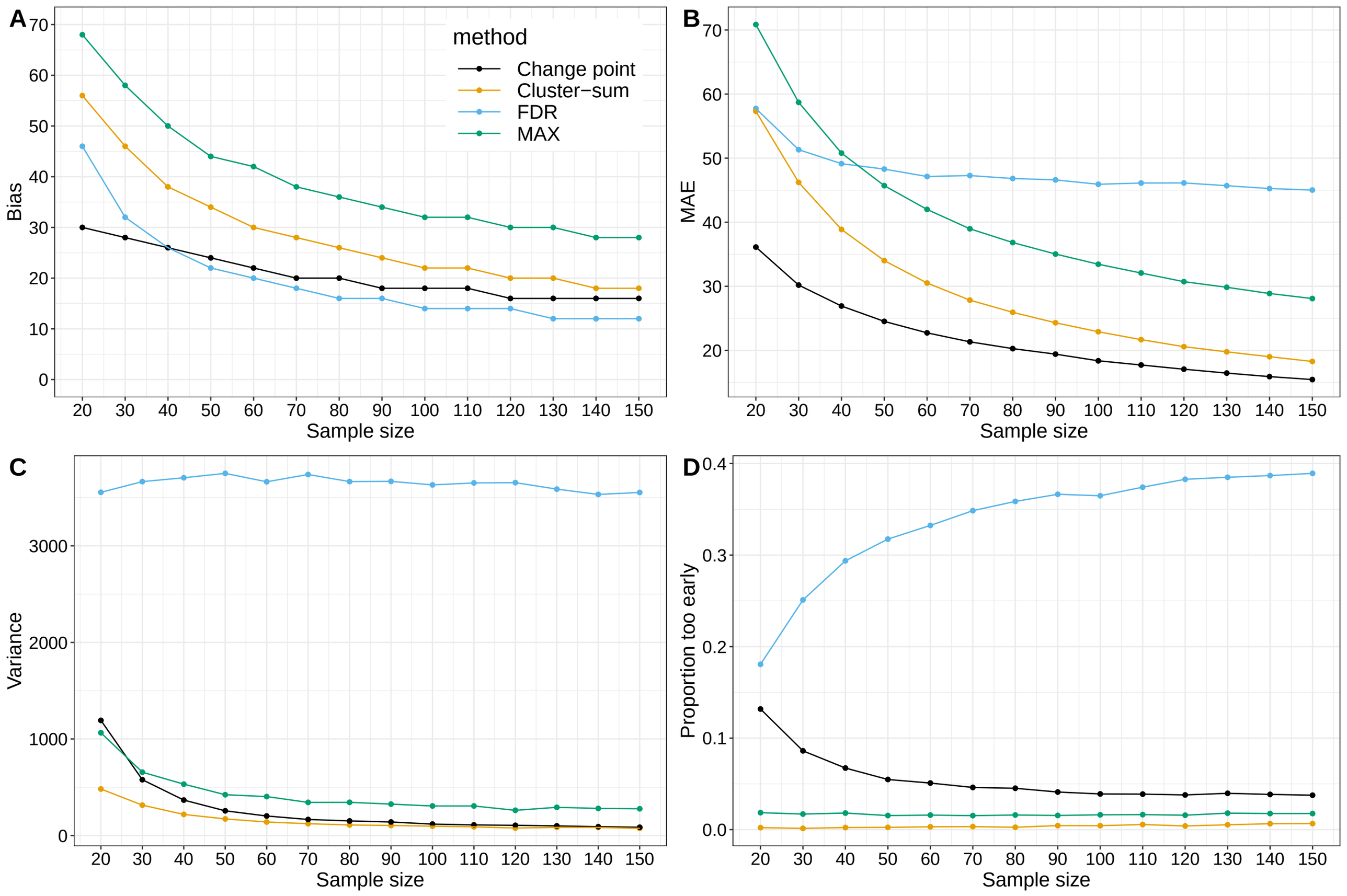
Yeung, N., Bogacz, R., Holroyd, C. B., & Cohen, J. D. (2004). Detection of synchronized oscillations in the electroencephalogram: An evaluation of methods. *Psychophysiology*, *41*(6), 822–832. https://doi.org/10.1111/j.1469-8986.2004.00239.x

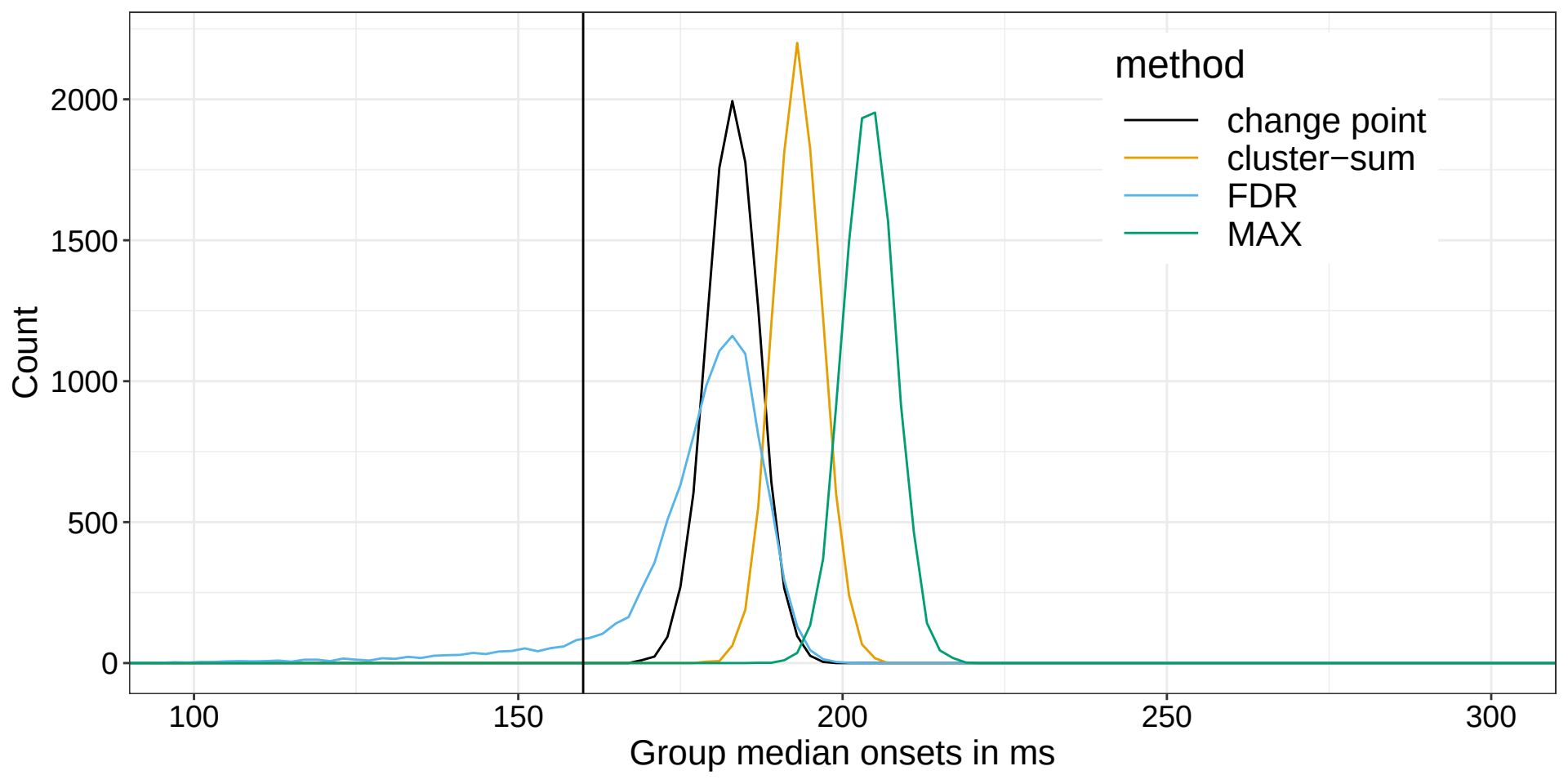
Zhao, K., Valle, D., Popescu, S., Zhang, X., & Mallick, B. (2013). Hyperspectral remote sensing of plant biochemistry using Bayesian model averaging with variable and band selection. *Remote Sensing of Environment*, *132*, 102–119. https://doi.org/10.1016/j.rse.2012.12.026


A
B
C
D
E
Signal in arbitrary units
Time in ms
t^2


A
Permutation + FDR correction = 196 ms
B
Permutation + MAX correction = 212 ms
C
Permutation + cluster−sum correction = 188 ms
D
Change point onset = 176 ms
t2
Time in ms
0
20
40
60
0
100
200
300
400
500


method
change point
cluster−sum
FDR
MAX
Count
600
400
200
0
0
100
200
300
400
500
Onsets in ms


A
B
C
D
method
Change point
Cluster−sum
FDR
MAX
Bias
MAE
Variance
Proportion too early
Sample size
Sample size
Sample size
Sample size


method
change point
cluster−sum
FDR
MAX
Count
2000
1500
1000
500
0
100
150
200
250
300
Group median onsets in ms