An empirically-driven guide on using Bayes Factors for M/EEG decoding

— Source link ↗

Lina Teichmann, Denise Moerel, Chris I. Baker, Tijl Grootswagers

Institutions: Macquarie University, University of Sydney

Related papers:

- Bayes Factors for Mixed Models

- Bayes Factor Model Comparisons Across Parameter Values for Mixed Models

- Illustrating Bayesian Evaluation of Informative Hypotheses for Regression Models

- Empirical Bayes shrinkage and false discovery rate estimation, allowing for unwanted variation.

- CorShrink: Empirical Bayes shrinkage estimation of correlations, with applications

An empirically-driven guide on using Bayes Factors for M/EEG decoding

Lina Teichmann[1][*], Denise Moerel[2], Chris Baker[1], Tijl Grootswagers[3]

[1] Laboratory of Brain and Cognition, National Institute of Mental Health, Bethesda, MD, USA

[2] Department of Cognitive Science, Macquarie University, Sydney, Australia

[3] The MARCS Institute for Brain, Behaviour & Development, Western Sydney University, Sydney, Australia

*Corresponding author: lina.teichmann@nih.gov

Abstract

Bayes Factors can be used to provide quantifiable evidence for contrasting hypotheses and have thus become increasingly popular in cognitive science. However, Bayes Factors are rarely used to statistically assess the results of neuroimaging experiments. Here, we provide an empirically-driven guide on implementing Bayes Factors for time-series neural decoding results. Using real and simulated Magnetoencephalography (MEG) data, we examine how parameters such as the shape of the prior and data size affect Bayes Factors. Additionally, we discuss benefits Bayes Factors bring to analysing multivariate pattern analysis data and show how using Bayes Factors can be used instead or in addition to traditional frequentist approaches.

32 1. Introduction

33 Bayes Factors provide an attractive alternative to the more traditional null hypothesis statistical
34 testing (NHST) framework. In particular, the use of Bayes Factors allows us to differentiate
35 between the amount of evidence for one theory over another in an intuitive way and sample
36 data without a strict sampling plan (Keysers et al., 2020; Wagenmakers et al., 2018). The
37 newfound popularity of Bayes Factors in cognitive science has not yet extended into cognitive
38 neuroscience, partly because there are no standard implementations. Here, we will provide a
39 data-driven guide on how Bayes Factors can be used in cognitive neuroscience, using an
40 example multivariate classification analysis of Magnetoencephalography (MEG) data.

41

42 Multivariate classification analyses have become a standard tool in analysing time-series
43 neuroimaging data (Carlson et al., 2019; Contini et al., 2017; Grootswagers et al., 2017;
44 Pantazis, 2020). To apply classification to time-series neural data, activation patterns are
45 extracted across MEG sensors and classification analyses are used to test whether a given
46 stimulus feature is represented in the neural data (see Figure 1 for an example). Across
47 participants, we can then test whether there is information in the signal by statistically
48 assessing whether classification accuracy is above chance. Under the null hypothesis the
49 sample mean equals chance decoding and under the alternative hypothesis the sample mean
50 is larger than chance decoding. Currently, NHST and p-values are the de-facto method of
51 choice when assessing whether decoding is above chance. However, recent studies have
52 started using Bayes Factors to assess classification accuracies (Grootswagers, Robinson, &
53 Carlson, 2019b; Grootswagers, Robinson, Shatek, et al., 2019; e.g., Grootswagers et al.,
54 2021; Kaiser et al., 2018; Mai et al., 2019; Proklova et al., 2019; Robinson et al., 2019, 2021).
55 In this paper, we focus on how Bayes Factors can be used to assess whether classification
56 accuracy is above-chance or at-chance. The Bayesian framework of hypothesis testing
57 involves directly comparing the predictions of two hypotheses (Jeffreys, 1939, 1935). Bayes
58 Factors describe the probability of one hypothesis over the other given the observed data. In
59 the multivariate pattern analysis (MVPA) context, this means we would use Bayes Factors to
60 test the probability of above-chance classification versus at-chance classification given the
61 classification results across participants at each timepoint.

62

63 The Bayesian approach brings several advantages (Dienes, 2011, 2014, 2016b; Keysers et
64 al., 2020; Morey et al., 2016; Wagenmakers et al., 2018). First, when calculating Bayes
65 Factors, two hypotheses are tested simultaneously. For time-series classification analyses, it
66 allows us to contrast evidence for above-chance versus at-chance decoding directly. In

67 addition, Bayes Factors are a measure of strength of evidence for one hypothesis versus
68 another which means we can *directly* assess how much evidence we have for above-chance
69 versus at-chance decoding at a given timepoint. This makes the interpretation of statistical
70 results more intuitive, as multiple Bayes Factors can be compared directly with larger numbers
71 reflecting more evidence. Another advantage is that Bayes Factors can be calculated
72 iteratively while more data are being collected and that testing can be stopped when there is
73 a sufficient amount of evidence. Such stopping-rules could be accompanied by a pre-specified
74 acquisition plan and potentially an (informal) preregistration via portals such as the Open
75 Science Framework (Foster & Deardorff, 2017). Using the data to determine when enough
76 evidence has been collected is particularly relevant for neuroimaging experiments, as it might
77 significantly reduce research costs and reduce the risk of having underpowered studies. Thus,
78 using a Bayesian approach to statistically assess time-series classification results can be
79 beneficial both from a theoretical as well as an economical standpoint and might ease the
80 ability to interpret and communicate scientific findings.
81
82 While there are clear advantages to using Bayes Factors for time-series decoding studies,
83 incorporating Bayes Factors into existing decoding pipelines may seem daunting. The goal of
84 the current paper is to present an empirically-driven guide to using Bayes Factors for
85 assessing time-series neuroimaging classification results. We present a practical example
86 based on a previously published time-series decoding study (Teichmann et al., 2019) and will
87 present results from simulations to show the influence of certain parameters on Bayes Factors.
88 We make use of the established Bayes Factor R package (Morey et al., 2015) to calculate the
89 Bayes Factors but provide sample codes along with this paper showing how to access the
90 Bayes   Factor   R   package   via   Matlab   and   Python
91 (https://github.com/LinaTeichmann1/BFF_repo). We also show how the Bayes Factors in our
92 example compare to p-values. Based on empirical evidence, we will give recommendations
93 for Bayesian analysis applied to M/EEG classification results. The aim of this paper is to
94 provide a broad introduction to Bayes Factors from a viewpoint of time-series neuroimaging
95 decoding. We aim to do so without going into the technical or mathematical detail, and instead
96 provide pointers to relevant literature on the specifics.
97

3

98
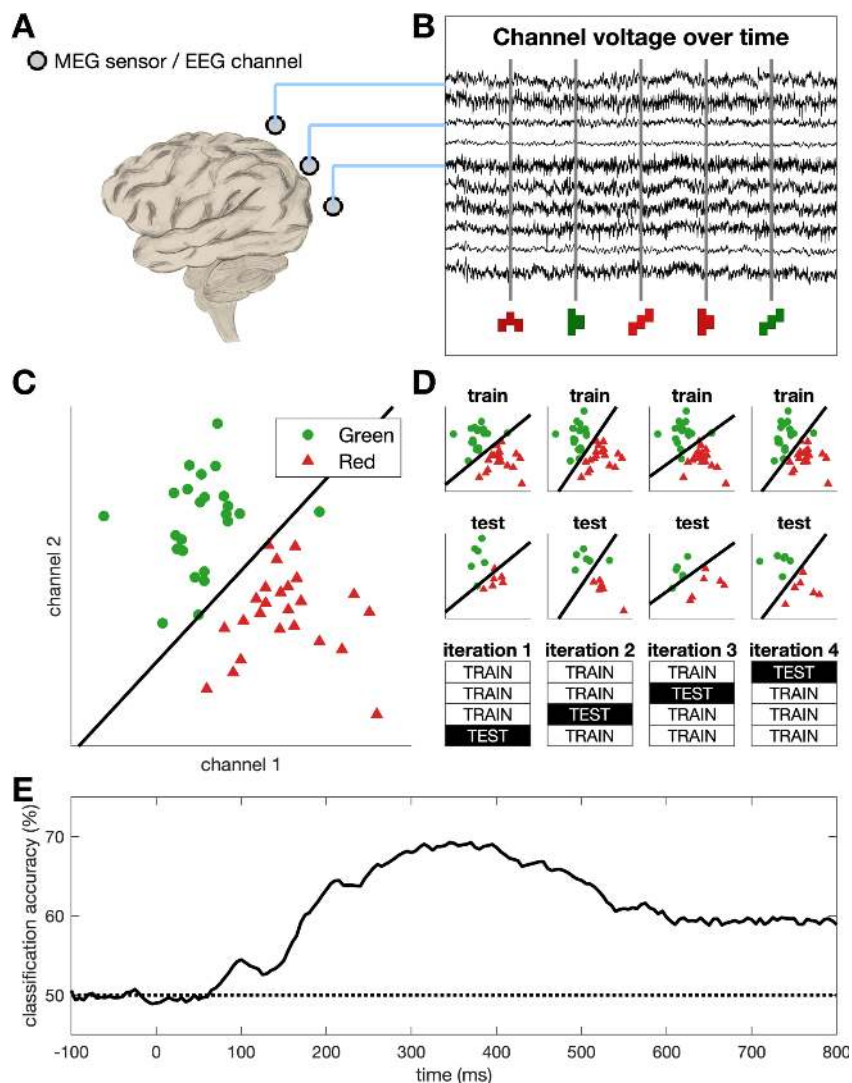
Figure 1. **Overview of MVPA for time-series neural data with simulated data.** (A) Example MEG sensors / EEG channels. (B) Simulated time-series neuroimaging data for a few sensors/channels. Vertical lines show stimulus onsets with example stimuli plotted below. Data is first epoched from -100 to 800 ms relative to stimulus onset, resulting in multiple time-series chunks associated with seeing a red or a green shape. (C) Using the epoched data, we can extract the sensor/channel activation pattern across the different sensors/channels (only 2 displayed for simplicity) for every trial at every timepoint. Then a classifier (black line) is trained to differentiate between the activation patterns evoked by red and green trials. (D) Example of a 4-fold cross validation where the classifier is trained on three quarters of the data and tested on the left-out quarter. This process is repeated at every timepoint. (E) We can calculate how often the classifier accurately predicts the colour of the stimulus at each timepoint by averaging across all testing folds. Theoretical chance level is 50% as there are two conditions in the simulated data (red and green). During the period before stimulus onset, we expect decoding to be at chance, and thus the baseline period can serve as a sanity check.

113

2.   Methods & Results

2.1 Example dataset & inferences based of Bayes Factors

The aim of the current paper is to show how to use Bayes Factors when assessing time-series neuroimaging classification results and test what effect different analysis parameters have on the results. We have used a practical example of previously published MEG data (Teichmann et al., 2019), which we re-analysed using Bayes Factors. In the original experiment, eighteen participants viewed coloured shapes and grayscale objects in separate blocks while the neural signal was recorded using MEG. Here, we only considered the coloured shape trials ("real colour blocks", 1600 trials in total). Identical shapes were coloured in red or green and were shown for 100 ms followed by an inter-stimulus-interval of 800-1100 ms. The data was epoched from -100 ms to 800 ms (200 Hz resolution) relative to stimulus onset and a linear classifier was used to differentiate between the neural responses evoked by red and green shapes. A 5-fold cross-validation was used with the classifier being trained on 80% of the data and tested on the remaining 20%. This classification analysis resulted in decoding accuracies over time for each participant. In the original study, permutation tests and cluster-corrected p-values were used to assess decoding accuracies as implemented in CoSMoMVPA (Oosterhof et al., 2016). Here, we calculated Bayes Factors instead and examined how parameter changes affected the results.

When running statistical tests on classification results, we are interested in whether decoding accuracy is above-chance at each timepoint. To test this, we can use permutation tests to establish whether there is enough evidence to reject H_0 which states that decoding is equal to chance. If there is enough evidence we can reject H_0 and conclude that decoding is different from chance. Given that below-chance decoding accuracies are not meaningful, we usually are interested only in above-chance decoding (directional hypothesis). In contrast to the frequentist approach, Bayes Factors quantify how much the plausibility of two hypotheses changes, given the data (see e.g., Ly et al., 2016). Here, we ran a Bayesian t-test of Bayes Factor R package (Morey et al., 2015) at each timepoint, testing whether the data is more consistent with H_a (decoding is larger than chance) over H_0 (decoding is equal to chance). The resulting Bayes Factors center around 1 with numbers smaller than 1 representing evidence for H_0 and numbers larger than 1 representing evidence for H_a. In contrast to p-values, Bayes Factors are directly interpretable and comparable (cf. Keysers et al., 2020; Morey et al., 2016; Wagenmakers et al., 2016). That is a Bayes Factor of 10 means that it is 10 times more likely the data came from H_a as opposed to H_0. Similarly a Bayes Factor of 1/10 means that it is 10 times more likely the data came from H_0 as opposed to H_a. Thus, in the

5

149 context of time-series decoding, Bayes Factors allow us to directly assess whether and how

150 much evidence there is at a given timepoint for the alternative over the null hypothesis and

151 *vice versa* (Figure 2C).

152

153
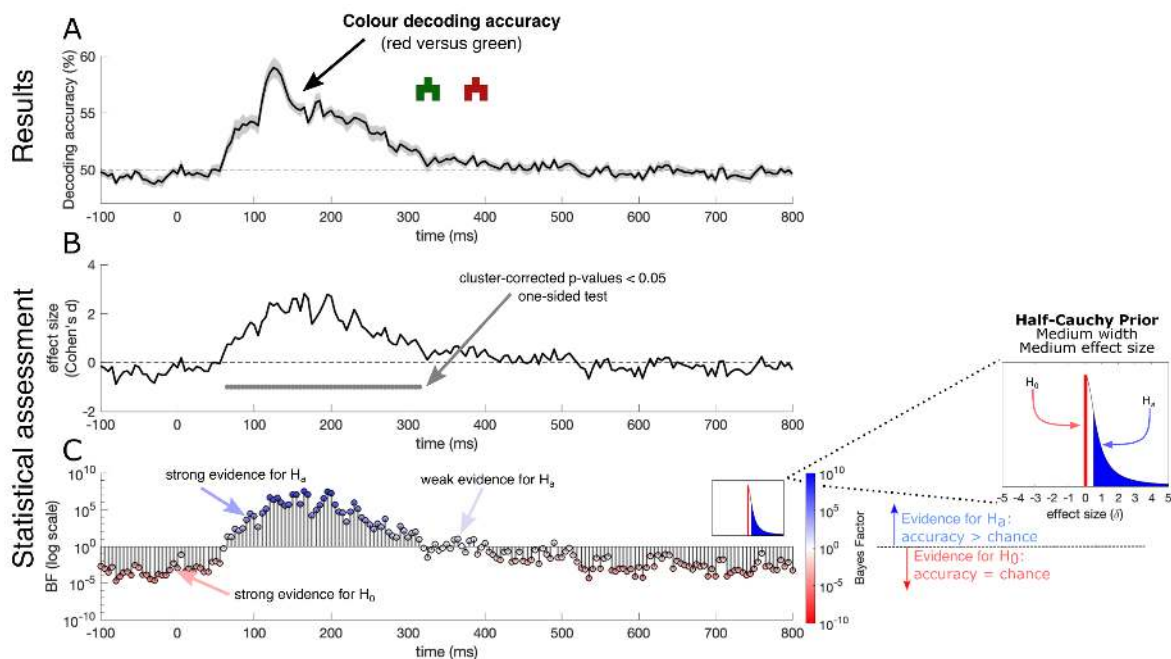
Figure 2. **Decoding results of our practical example dataset with statistical**
assessments. (A) Colour decoding over time (black line). The dashed line shows theoretical
chance decoding (50%). The grey shaded area represents the standard error across
participants. (B) Effect size over time with the cluster-corrected p-values at each timepoint
printed below in grey. (C) Bayes Factors over time for this dataset on a logarithmic scale. Blue,
upwards pointing stems indicate evidence for above-chance decoding and red, downwards
pointing stems show evidence for at-chance decoding at every timepoint. We used a hybrid
one-sided model comparing evidence for above-chance decoding versus a point-nil at $\delta = 0$
(no effect). For the alternative hypothesis, we used a half-cauchy prior with medium width (r =
0.707) covering an interval from $\delta = 0.5$ to $\delta = \infty$. The half-cauchy prior assumes that small
effect sizes are more likely than large ones but the addition of the interval deems very small
effects $\delta < 0.5$ as irrelevant. During the baseline period (i.e., before stimulus onset), the Bayes
Factors strongly support the null hypothesis, confirming the sanity check expectation.

167

168 <u>2.2 Adjusting the prior range to account observed chance decoding</u>

169 Bayes Factors represent the plausibility that the data emerged from one hypothesis compared

170 to another. In the example dataset, the two hypotheses are that decoding is at chance (i.e.,

171 H_0, no colour information present) or that decoding is above chance (i.e., H_a, colour

172 information present). To deal with the fact that observed decoding can be different than the

173 theoretical chance level, we can adjust the prior range of the alternative hypothesis to allow

174 for small effects under the null hypothesis (Rouder et al., 2009). The prior range (called "null

175 interval" in the R package) is defined in standardized effect sizes and consists of a lower and

176 upper bound. To incorporate the differences between observed and theoretical chance level,

6

177   we can define a range of relevant effect sizes for the alternative hypothesis, for example, from

178   δ = 0.5 to δ = ∞. To determine which values are reasonable as the lower bound of this interval,

179   we changed the prior range systematically and examined the effect on the resulting Bayes

180   Factors (Figure 3). We found that smaller lower bounds at δ = 0 and δ = 0.2 resulted in weaker

181   evidence supporting the null hypothesis than ranges starting at δ = 0.5 and δ = 0.8. The range

182   did not have a large effect on timepoints with strong evidence for H_a. The effect of changing

183   the prior range is larger for the null hypothesis than the alternative as chance decoding is not

184   exactly 50% but distributed around chance. Changing the lower bound of the prior range

185   means that the effects that are just larger than δ = 0 can support the null hypothesis. Thus,

186   the results here demonstrate that we can compensate for the differences between theoretical

187   and observed chance by adjusting the prior range and effectively considering small effect sizes

188   as evidence for the null hypothesis rather than the alternative.

189
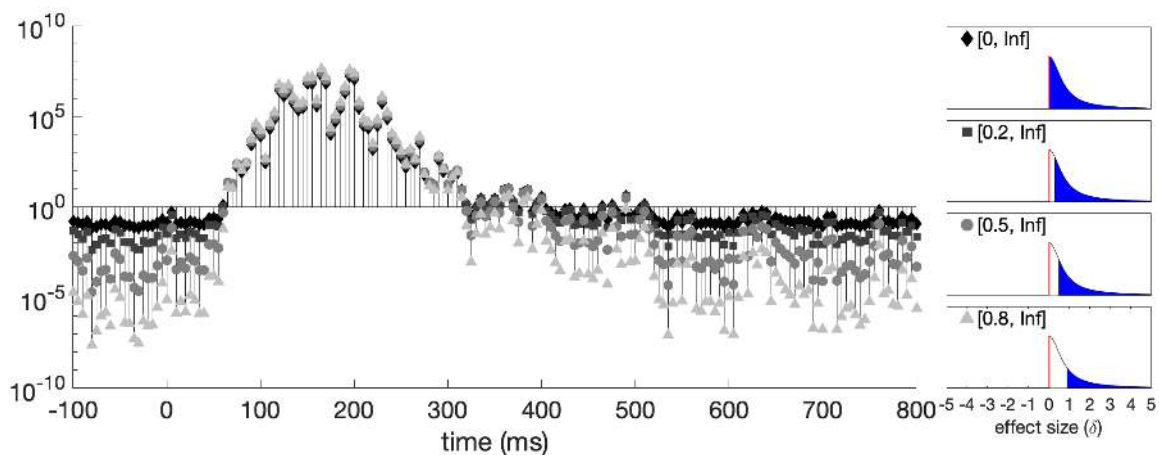
190

191   **Figure 3. The effect of changing the prior range (null interval) on Bayes Factors in our**
192   **example data.** Intervals starting at larger effect sizes led to more timepoints showing
193   conclusive evidence for H_0. This is due to the fact that theoretical and observed chance levels
194   are not the same. The panels on the right show the prior distributions with the different null
195   intervals.

196

197   To further examine what a reasonable lower bound of the prior range is, we looked at effect

198   sizes observed during the baseline window (before stimulus onset) in a selection of our

199   previous studies (Grootswagers et al., 2021; Grootswagers, Robinson, & Carlson, 2019a;

200   Moerel, Grootswagers, et al., 2021; Moerel, Rich, et al., 2021; Teichmann et al., 2018, 2020).

201   Using the baseline window allows us to quantify the difference between theoretical and

202   observed chance, as we do not expect any meaningful effects before stimulus onset (e.g.,

203   stimulus colour is not decodable before the stimulus is presented). Thus, the baseline period

204   can effectively tell us which effect sizes can be expected by chance. Across our selection of

205   previous studies, we found an average maximum effect size of δ = 0.39 before stimulus onset

7

206   and an average maximum effect size of δ = 1.91 after stimulus onset (Figure 4). This survey

207   shows that effect sizes as large as δ = 0.5 can be observed when when no meaningful

208   information is in the signal. Thus, this supports the conclusions from the example dataset

209   showing that prior ranges with a lower bound of δ = 0.5 may be a sensible choice when using

210   Bayes Factors to examine time-series M/EEG decoding results.

211
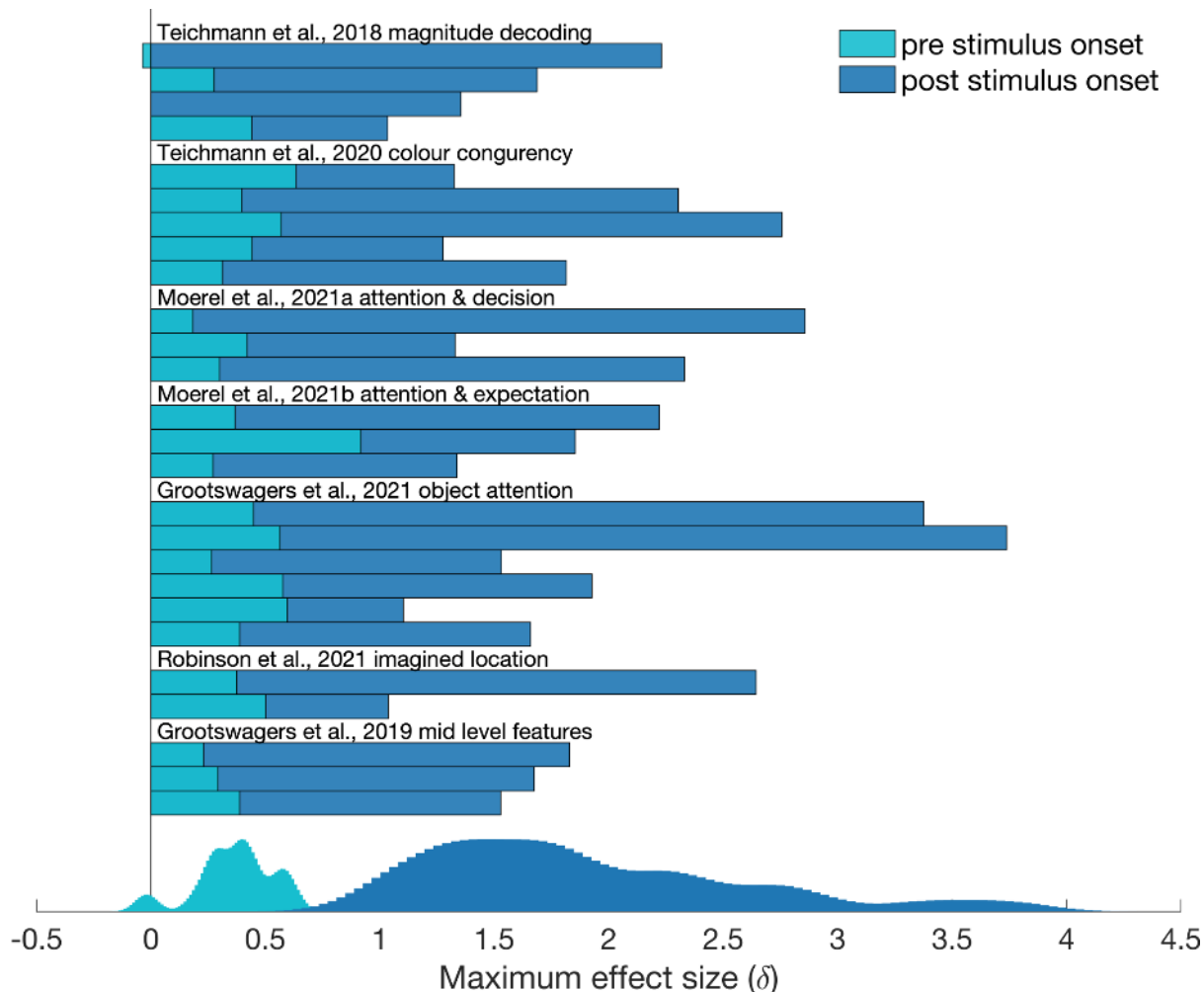
212

213   **Figure 4. Estimated maximum effect sizes during baseline and after stimulus onset for**
214   **prior studies using visual stimuli.** These estimations show that a reasonable range for H_a
215   would start at δ = 0.5 or above, as during baseline decoding accuracies corresponding to
216   standardized effect sizes as high as δ = 0.5 were observed.

217

218   2.3 Changing the prior width to capture different effect sizes

219   Another feature that can be changed in the Bayesian t-test is the width of the half-Cauchy

220   distribution (referred to as r-value in the Bayes Factor Package). Small r-values create a

221   narrower, sharply peaking distribution, whereas larger values make the distribution wider with

222   a prolonged peak. Standard prior widths incorporated in the Bayes Factor R package are

223   medium (r = 0.707), wide (r = 1), and ultrawide (r = 1.414). Keeping the prior range consistent

224   ([0.5, Inf]) while using the three prior widths implemented into the R Bayes Factor Package

225   (medium = 0.707; wide = 1; ultrawide = 1.414). We found that changing the width of the

226   Cauchy prior did not have a pronounced effect on the Bayes Factors (Figure 5). In our specific

227   example, this is probably the case because the effect sizes quickly rose to $\delta > 2$ (Figure 2b)

228   which means that the subtle differences between the different prior widths do not have a

229   substantial effect on the likelihood of the data arising from H_a over H_0. Thus, using the default

230   prior width (r = 0.707) for the decoding context seems like a reasonable choice.

231

232
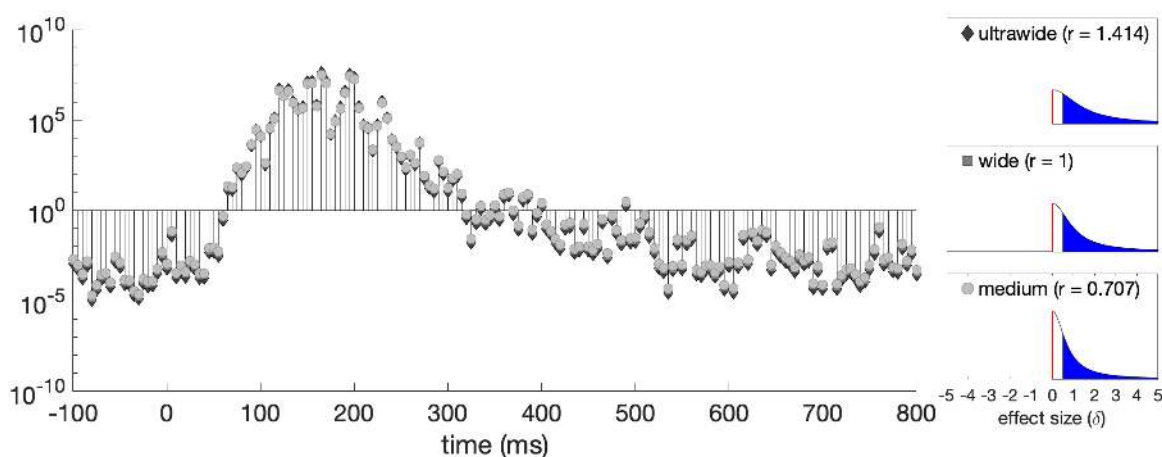
Figure 5. **Bayes Factors over time for the example data set when the prior width is changed.** The width of the prior had no pronounced effect on the Bayes Factors we calculated. The panels on the right show the prior distributions with the different widths.

236

237   2.4 The effect of data size on statistical inferences

238   In a lot of cases, there are financial and time limits on how many participants can be tested

239   and for how long. To obtain an estimate of how much data is needed to draw conclusions and

240   avoid ending up with underpowered studies, we used the example dataset and reduced the

241   data size for analysis. As classification analyses are usually run at the subject level but

242   statistical assessment is run at the group level, we tested how changing data size both by trial

243   numbers and participant numbers influences Bayes Factors in the time-series decoding

244   context (Figure 6). In the original example dataset, the classifier was trained on 1408 trials

245   and tested on 352 trials (5-fold cross-validation). There were five different shapes in the red

246   and the green condition (160 repetitions for each coloured shape) and the cross-validation

9

247   schema was based on leaving all trials of one shape out for testing. Statistical inferences were

248   drawn on the group level which contained data from 18 participants. To examine the effect of

249   data size (and effectively noise level) on the Bayes Factor calculations, we re-ran the analysis

250   reducing the data size first by subsampling from the trials each participant completed, retaining

251   1200 (75%), 800 (50%), 400 (25%), and 160 (10%) trials. We cross-validated in the same way

252   as in the original paper, with the only difference being how many trials of each shape were

253   included. In addition, we subsampled from the whole group, retaining data from the first 6, 12,

254   or all 18 participants and re-ran the statistical analysis. We then compared the results from

255   the reduced-size colour datasets using Bayes Factors and cluster-corrected p-values[1].

256

257   Overall, our analyses highlight that we need to have a large enough number of trials and a

258   large enough number of participants to draw firm conclusions about our time-resolved

259   decoding results. Testing more participants resulted in stronger evidence for H_a and H_0, with

260   fewer timepoints in the inconclusive range (Bayes Factors) and more significant above-chance

261   decoding timepoints (p-values). Similarly, running the classification with more trials, led to

262   more timepoints with large Bayes Factors supporting H_a and more above-chance decoding

263   timepoints. However, one of the key advantages of using Bayes Factors instead of p-values

264   is that we can potentially obtain a good idea of how many trials are needed even if we run a

265   pilot experiment with a limited number of participants. A reasonable strategy would be to

266   overpower the subject-level data (i.e., number of trials) for the pilot sample and then sub-

267   sample to explore how many trials are needed. In our example, we can see that the amount

268   of evidence for H_a at peak decoding is not sufficient when we only use 160 trials (10% of the

269   original sample), regardless of the number of subjects. Increasing the trials to 400 or 800 (25%

270   or 50% of the original sample) leads to similar conclusions as using all 1600 trials. As Bayesian

271   statistics allow for sequential sampling, we could collect data from more participants until a

272   criterion is reached. The data here suggest that insufficient data at the subject-level ultimately

273   leads to inconclusive evidence, highlighting that a large number of trials is just as, if not more

274   important, than large numbers of participants.

275

---

[1] In comparison to the original paper, we did not use trial label permutations. Instead, we performed sign-flip permutations (which reduces the computational time) as implemented in CoSMoMVPA to generate the null distribution.
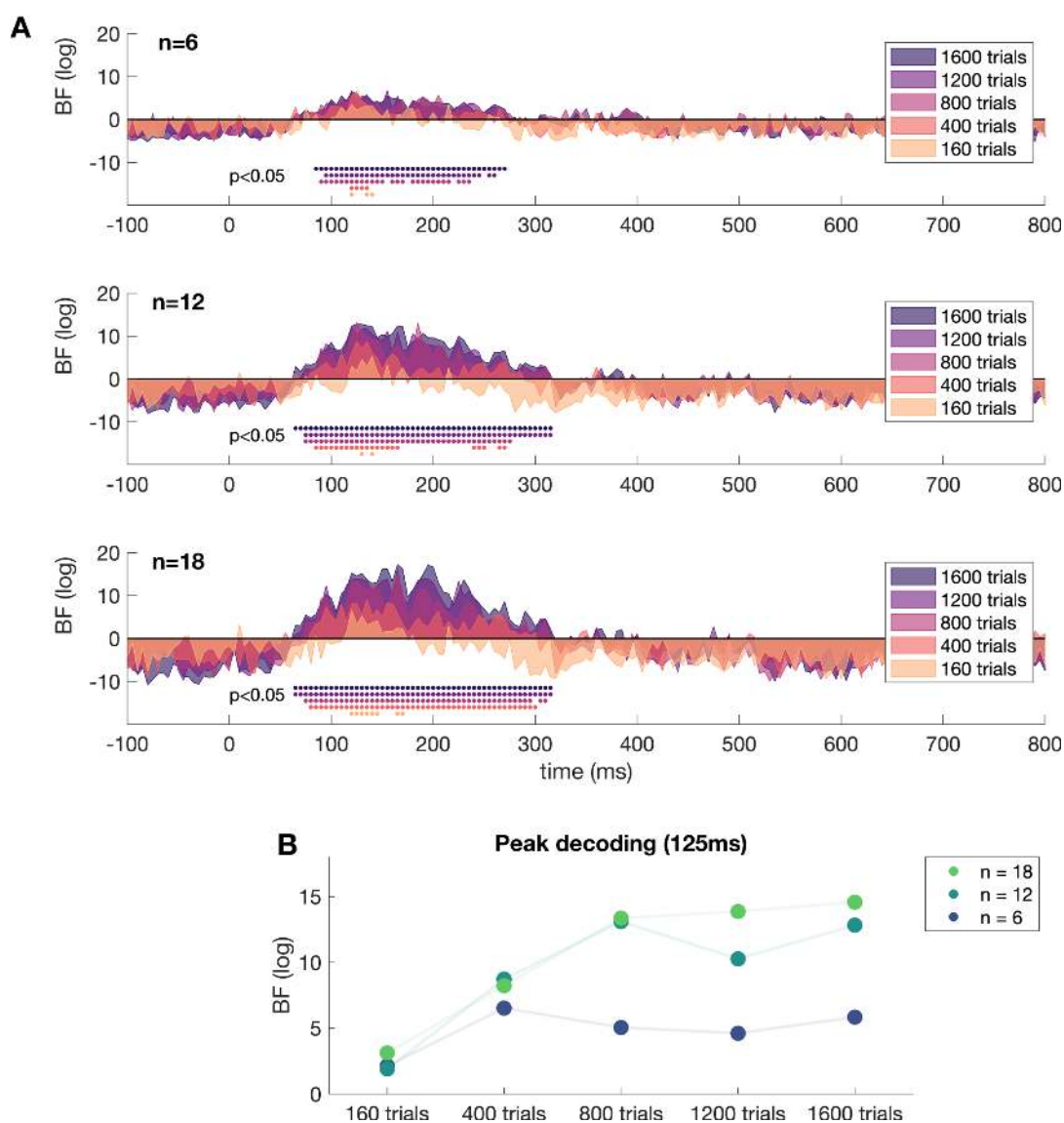
Figure 6. Results of colour MEG decoding, using a limited number of trials and participant data to simulate a piloting scenario. (A) The first three plots show Bayes Factors over time along with cluster-corrected p-values. The colour in all plots reflects the number of trials used to train and test the classifier. (B) Compares Bayes Factors at peak decoding (125ms) for the different data sizes.

11

285   In addition to manipulating data size, we also simulated larger datasets with fixed effect sizes

286   between $\delta = 0$ and $\delta = 1$ and examined the interaction of sample size with different prior ranges

287   (Figure 7). We simulated 1000 datasets with specific effect sizes for each sample size and

288   calculated the Bayes Factors. We then calculated the median Bayes Factor for each sample-

289   and effect size combination to show how prior range choices interact with the possibility of

290   finding evidence for effects of different sizes. Specifically, we compared a prior range of 0.5 to

291   infinity (Figure 7A) to a prior range of zero to infinity (Figure 7B). When specifying the prior

292   range to 0.5 to infinity (Figure 7A), our results show that particularly small effect sizes lead to

293   substantial evidence for H_0 faster, while particularly large effect sizes lead to substantial

294   evidence for H_a faster. In these cases, large sample sizes were not needed to draw solid

295   conclusions. In contrast, if the effect size fell in between the specified ranges for the prior of

296   H_a and H_0 (i.e., between 0 and 0.5), we found that small sample sizes in particular tended to

297   result in inconclusive Bayes Factors neither supporting H_a or H_0. However, if the sample size

298   increased, the confidence that small effects were "real" also increased and therefore resulted

299   in stronger confidence supporting one of the hypotheses. Importantly, however, large sample

300   sizes did not automatically lead to an interpretable Bayes Factor if the effect was truly in

301   between the specified prior ranges of H_a and H_0, indicating that sampling strategy had no

302   effect on Bayes Factors. Consistent with our results for the example data, the simulations also

303   showed that changing the range of the prior has a strong effect on finding substantial evidence

304   for H_0. If the prior range for the alternative is specified to start at zero (Figure 7B), it was almost

305   impossible to find any evidence for H_0, even if the effect size was truly zero. Thus, the

306   simulations show that defining the prior range with a gap between effects expected under H_0

307   and H_a is critical and that more data leads to larger Bayes Factors, but only if there is a true
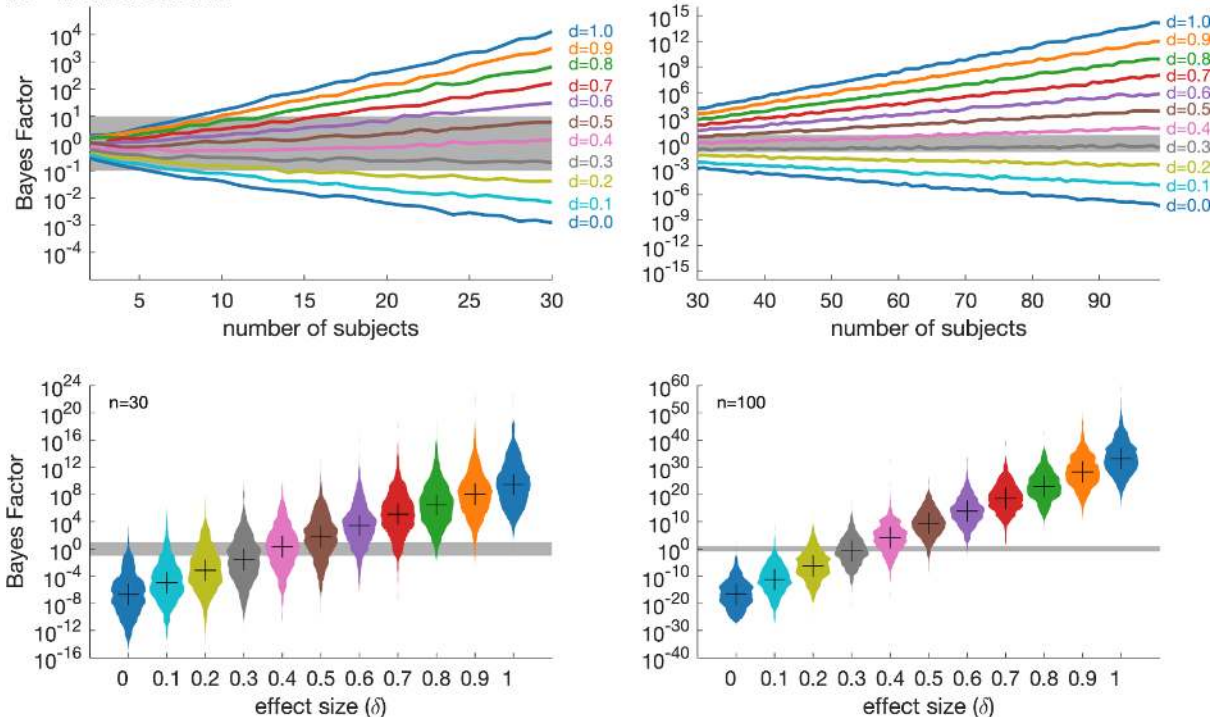
308   underlying effect.

309

310

311

312

12
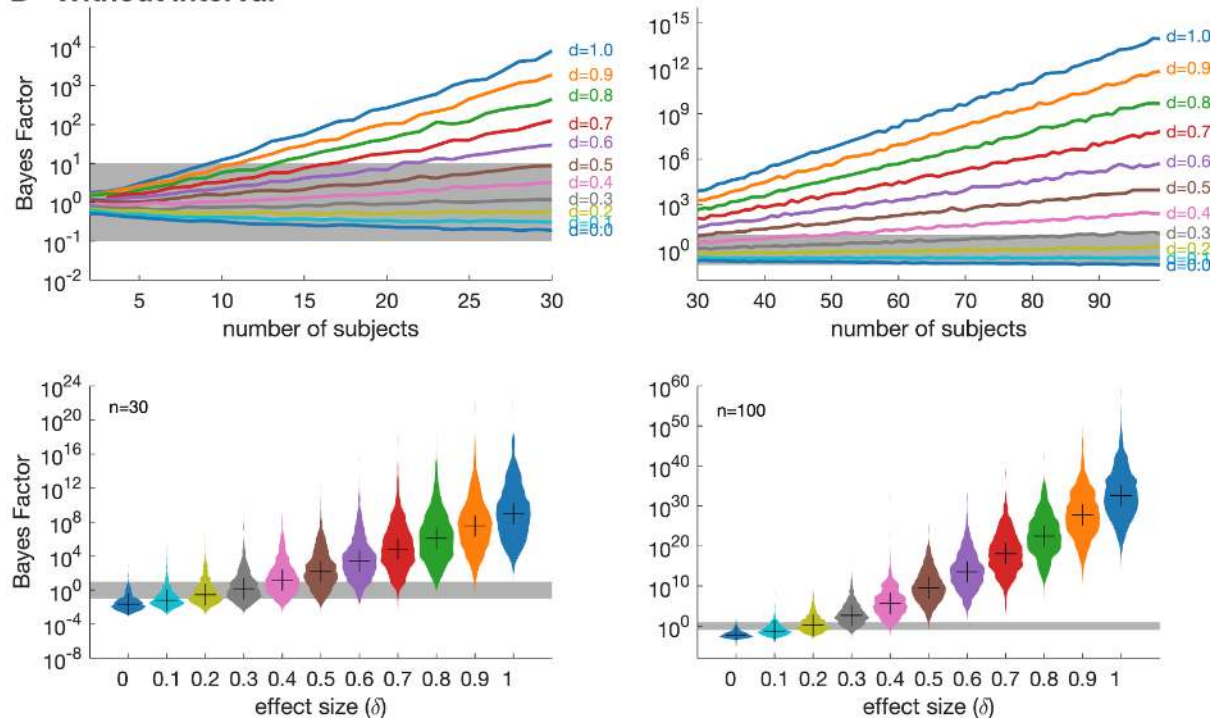
Figure 7. Simulated data varying effect sizes and numbers of participants. (A) Bayes Factors obtained using a half-cauchy with an interval [0.5 Inf]. (B) Bayes Factors obtained using a half-cauchy without an interval. The first and third rows show the median Bayes Factors of 1000 simulations as a function of the number of participants. The second and fourth rows show the distribution (violin plots) of the 1000 simulations at varying effect sizes using n=30 (left panels) and n=100 (right panels), with the black cross indicating the median. Note the different scales on the y-axis between panels.

Discussion

321

322   Bayes Factors have seen a recent increase in popularity in cognitive science, as they can be
323   used to provide quantifiable evidence for contrasting hypotheses. However, their uptake has
324   to date been slow for neuroimaging experiments. To facilitate their adoption, we have provided
325   an empirically-driven guide on implementing Bayes Factors for time-series neuroimaging
326   decoding, using both real and simulated data. We showed that using Bayes Factors and
327   cluster-corrected p-values lead to similar results when statistically assessing time-series
328   neuroimaging decoding results. However, the key advantages of using Bayes Factors are the
329   ability to compare evidence for H_a with evidence for H_0 and having results that are quantifiable
330   (e.g., Dienes, 2014; Wagenmakers et al., 2016). Our results show that for time-series
331   decoding data, half-Cauchy priors with default width and an interval ranging from effect sizes
332   of 0.5 to infinity provide sensible results. We also show that even a small number of
333   participants can yield informative Bayes Factors, which can be useful for making decisions on
334   experimental design parameters (e.g., number of trials) during piloting stages of a study.

335

336   Our results showed that the overall conclusions derived from Bayes Factors and p-values
337   were quite similar, highlighting that theoretical considerations should be the deciding factor
338   when choosing a statistical approach to analyze neural time-series data. In the decoding
339   context, p-values afford a dichotomous decision of whether there is enough evidence to reject
340   the hypothesis that decoding is at chance at a given timepoint. Rejecting the null hypothesis
341   is decoupled from any prior beliefs or theories (Dienes, 2011) and is linked to an accepted
342   overall error rate such as α = 0.05. However, they allow us to test for the presence of an effect
343   at a given timepoint using widely accepted thresholds for evidence. While Bayes Factors can
344   in principle be thresholded to draw dichotomous conclusions, one of the added benefits of
345   Bayes Factors over p-values is the ability to quantify the evidence. Another useful benefit of
346   using Bayes Factors to analyse time-series decoding data is that Bayes Factors allow us to
347   accrue evidence for above-chance as well as at-chance decoding. For time-series analyses
348   in particular, this is a useful feature as the time period prior to stimulus onset can be considered
349   as a control period where we would expect evidence for the null hypothesis. Testing both
350   hypotheses simultaneously can also be a beneficial feature when the research question
351   involves hypotheses predicting certain time-periods without any information in the neural
352   signal (e.g., "X happens before Y" versus "Y happens before X"). Thus, depending on the
353   research question it may be clear which statistical approach suits the time-series decoding
354   analysis best. Otherwise, as overall conclusions do not differ, Bayes Factors and p-values can
355   be used in a complementary way to provide quantifiable evidence for and against the tested

356 hypotheses as well as definitive decisions (see also Lakens et al., 2020; van Dongen et al.,
357 2019; Wagenmakers et al., 2018).

358

359 Through our results, we provide an empirical, straightforward guide to help implement Bayes
360 Factors and demonstrate the extent of practical benefits when using Bayes Factors for time-
361 series neural decoding. Using a data-driven approach, we showed which analysis parameters
362 are most suitable for statistical assessment of time-series decoding data with Bayes Factors.
363 While the Bayes Factors in our example MEG decoding dataset were robust against changes
364 in the predefined width of the prior, defining the prior range so that there is a gap between H_a
365 and H_0 was critical for finding evidence for the H_0. This strong effect of the prior range on the
366 resulting Bayes Factors is particularly relevant in the decoding context, as classification
367 accuracies under the null are not symmetrically distributed around chance (cf. Allefeld et al.,
368 2016). Thus, a gap between H_0 and the lower bound of H_a ensures that small above-chance
369 classification accuracies are not treated as evidence for H_a. Furthermore, we systematically
370 varied dataset size and showed that using Bayes Factors for time-series decoding data is
371 particularly beneficial when there is limited, noisy data such as in a piloting scenario, as
372 quantifiable evidence for one hypothesis over another gives a stronger sense of whether it is
373 worth pursuing the research question with the piloted design, or make changes (e.g., modify
374 trial numbers or add/remove conditions). Finally, Bayes Factors can be calculated sequentially
375 while evidence accumulation is monitored to stop once a criterion is reached (Dienes, 2011;
376 Rouder, 2014), which can save resources and avoid underpowered studies (Wagenmakers et
377 al., 2018).

378

379 An open question is to what extent our parameter choices generalize to different paradigms,
380 analysis approaches, and modalities. The Bayes Factor parameters used here were optimized
381 for time-series decoding. It is in principle possible to use Bayes Factors in a similar way to
382 analyse other time-series data such as event related potentials, oscillations or regressions,
383 however, the Bayes Factor parameters might have to be adjusted. Similarly, the analysis
384 pipeline discussed here could be extended to other neural decoding modalities such as fMRI
385 (see e.g., Moerel, Rich, et al., 2021). Pilot data or analyses of previous data can be used to
386 examine how parameters have to be modified in order to get sensible results.

387

388 A final consideration is the multiple comparisons problem arising from statistically testing many
389 time points. When using Bayes Factors, as long as the evidence for each hypothesis is
390 interpreted at face value (and not thresholded for 'significance'), we do not need to control for
391 multiple comparisons (Dienes, 2011, 2016a; Świątkowski & Carrier, 2020). That is because
392 once we have established a prior and collected the data, we examine how much we have to

393 adjust our prior beliefs given the data and compare the adjustment required for both

394 hypotheses. This idea is not related to overall error rates and thus does not change if we

395 sample data sequentially or run multiple tests (Dienes, 2016a). If a research question strongly

396 depends on a dichotomous decision on multiple tests, then we advise to report corrected p-

397 values (for which correction methods are well established) alongside the Bayes Factors.

398

399 In conclusion, we have provided an empirically-driven guide on how to use and interpret Bayes

400 Factors for time-series neuroimaging decoding data. We show that Bayes Factors bring

401 several advantages to interpreting time-series decoding results such as quantifiable evidence

402 and an ability to compare evidence for above-chance with evidence for at-chance decoding.

403 We hope this guide, and the accompanying example code

404 (https://github.com/LinaTeichmann1/BFF_repo) can serve as a starting point to incorporate

405 Bayesian statistics to existing analysis pipelines.

406

References

408 Allefeld, C., Görgen, K., & Haynes, J.-D. (2016). Valid population inference for information-
409   based imaging: From the second-level t-test to prevalence inference. *Neuroimage*,
410   *141*, 378–392.

411 Carlson, T. A., Grootswagers, T., & Robinson, A. K. (2019). An introduction to time-resolved
412   decoding analysis for M/EEG. *ArXiv Preprint ArXiv:1905.04820*.

413 Contini, E. W., Wardle, S. G., & Carlson, T. A. (2017). Decoding the time-course of object
414   recognition in the human brain: From visual features to categorical decisions.
415   *Neuropsychologia*. https://doi.org/10.1016/j.neuropsychologia.2017.02.013

416 Dienes, Z. (2011). Bayesian versus orthodox statistics: Which side are you on? *Perspectives*
417   *on Psychological Science*, *6*(3), 274–290.

418 Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in*
419   *Psychology*, *5*, 781.

420 Dienes, Z. (2016a). How Bayes factors change scientific practice. *Journal of Mathematical*
421   *Psychology*, *72*, 78–89.

422 Dienes, Z. (2016b). How Bayes factors change scientific practice. *Journal of Mathematical*
423   *Psychology*, *72*, 78–89. https://doi.org/10.1016/j.jmp.2015.10.003

424 Foster, E. D., & Deardorff, A. (2017). Open Science Framework (OSF). *Journal of the*
425   *Medical Library Association : JMLA*, *105*(2), 203–206.
426   https://doi.org/10.5195/jmla.2017.88

427 Grootswagers, T., Robinson, A. K., & Carlson, T. A. (2019a). The representational dynamics
428   of visual objects in rapid serial visual processing streams. *NeuroImage*, *188*, 668–
429   679.

430 Grootswagers, T., Robinson, A. K., & Carlson, T. A. (2019b). The representational dynamics
431   of visual objects in rapid serial visual processing streams. *NeuroImage*, *188*, 668–
432   679. https://doi.org/10.1016/j.neuroimage.2018.12.046

433   Grootswagers, T., Robinson, A. K., Shatek, S. M., & Carlson, T. A. (2019). Untangling
434   featural and conceptual object representations. *NeuroImage*, *202*, 116083.
435   https://doi.org/10.1016/j.neuroimage.2019.116083
436   Grootswagers, T., Robinson, A. K., Shatek, S. M., & Carlson, T. A. (2021). The neural
437   dynamics underlying prioritisation of task-relevant information. *Neurons, Behavior,*
438   *Data Analysis, and Theory*, *5*(1), 1–17. https://doi.org/10.51628/001c.21174
439   Grootswagers, T., Wardle, S. G., & Carlson, T. A. (2017). Decoding Dynamic Brain Patterns
440   from Evoked Responses: A Tutorial on Multivariate Pattern Analysis Applied to Time
441   Series Neuroimaging Data. *Journal of Cognitive Neuroscience*, *29*(4), 677–697.
442   https://doi.org/10.1162/jocn_a_01068
443   Jeffreys, H. (1939). The Theory of Probability. *The Theory of Probability*.
444   Jeffreys, H. (1935). Some tests of significance, treated by the theory of probability.
445   *Mathematical Proceedings of the Cambridge Philosophical Society*, *31*(2), 203–222.
446   Kaiser, D., Moeskops, M. M., & Cichy, R. M. (2018). Typical retinotopic locations impact the
447   time course of object coding. *NeuroImage*.
448   https://doi.org/10.1016/j.neuroimage.2018.05.006
449   Keysers, C., Gazzola, V., & Wagenmakers, E.-J. (2020). Using Bayes factor hypothesis
450   testing in neuroscience to establish evidence of absence. *Nature Neuroscience*,
451   *23*(7), 788–799.
452   Lakens, D., McLatchie, N., Isager, P. M., Scheel, A. M., & Dienes, Z. (2020). Improving
453   inferences about null effects with Bayes factors and equivalence tests. *The Journals*
454   *of Gerontology: Series B*, *75*(1), 45–57.
455   Ly, A., Verhagen, J., & Wagenmakers, E.-J. (2016). Harold Jeffreys's default Bayes factor
456   hypothesis tests: Explanation, extension, and application in psychology. *Journal of*
457   *Mathematical Psychology*, *72*, 19–32.
458   Mai, A.-T., Grootswagers, T., & Carlson, T. A. (2019). In search of consciousness:
459   Examining the temporal dynamics of conscious visual perception using MEG time-
460   series data. *Neuropsychologia*, *129*, 310–317.
461   https://doi.org/10.1016/j.neuropsychologia.2019.04.015
462   Moerel, D., Grootswagers, T., Robinson, A. K., Shatek, S. M., Woolgar, A., Carlson, T. A., &
463   Rich, A. N. (2021). Undivided attention: The temporal effects of attention dissociated
464   from decision, memory, and expectation. *BioRxiv*, 2021.05.24.445376.
465   https://doi.org/10.1101/2021.05.24.445376
466   Moerel, D., Rich, A. N., & Woolgar, A. (2021). Selective attention and decision-making have
467   separable neural bases in space and time. *BioRxiv*, 2021.02.28.433294.
468   https://doi.org/10.1101/2021.02.28.433294
469   Morey, R. D., Romeijn, J.-W., & Rouder, J. N. (2016). The philosophy of Bayes factors and
470   the quantification of statistical evidence. *Journal of Mathematical Psychology*, *72*, 6–
471   18.
472   Morey, R. D., Rouder, J. N., Jamil, T., & Morey, M. R. D. (2015). Package 'bayesfactor.'
473   *URLh Http://Cran/r-Projectorg/Web/Packages/BayesFactor/BayesFactor Pdf i*
474   *(Accessed 1006 15)*.
475   Oosterhof, N. N., Connolly, A. C., & Haxby, J. V. (2016). CoSMoMVPA: Multi-modal
476   multivariate pattern analysis of neuroimaging data in Matlab/GNU Octave. *Frontiers*
477   *in Neuroinformatics*, *10*. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4956688/
478   Pantazis, D. (2020). Decoding cognitive function with magnetoencephalography. *Fifty Years*
479   *of Magnetoencephalography: Beginnings, Technical Advances, and Applications*, *19*,
480   278.

481 Proklova, D., Kaiser, D., & Peelen, M. V. (2019). MEG sensor patterns reflect perceptual but
482   not categorical similarity of animate and inanimate objects. *NeuroImage, 193*, 167–
483   177. https://doi.org/10.1016/j.neuroimage.2019.03.028

484 Robinson, A. K., Grootswagers, T., & Carlson, T. A. (2019). The influence of image masking
485   on object representations during rapid serial visual presentation. *NeuroImage, 197*,
486   224–231. https://doi.org/10.1016/j.neuroimage.2019.04.050

487 Robinson, A. K., Grootswagers, T., Shatek, S. M., Gerboni, J., Holcombe, A., & Carlson, T.
488   A. (2021). Overlapping neural representations for the position of visible and imagined
489   objects. *Neurons, Behavior, Data Analysis, and Theory, 4*(1), 1–28.
490   https://doi.org/10.51628/001c.19129

491 Rouder, J. N. (2014). Optional stopping: No problem for Bayesians. *Psychonomic Bulletin &*
492   *Review, 21*(2), 301–308.

493 Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests
494   for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review,*
495   *16*(2), 225–237.

496 Świątkowski, W., & Carrier, A. (2020). There is Nothing Magical about Bayesian Statistics:
497   An Introduction to Epistemic Probabilities in Data Analysis for Psychology Starters.
498   *Basic and Applied Social Psychology, 42*(6), 387–412.

499 Teichmann, L., Grootswagers, T., Carlson, T., & Rich, A. N. (2018). Decoding digits and dice
500   with magnetoencephalography: Evidence for a shared representation of magnitude.
501   *Journal of Cognitive Neuroscience, 30*(7), 999–1010.

502 Teichmann, L., Grootswagers, T., Carlson, T., & Rich, A. N. (2019). Seeing versus knowing:
503   The temporal dynamics of real and implied colour processing in the human brain.
504   *NeuroImage, 200*, 373.

505 Teichmann, L., Quek, G. L., Robinson, A. K., Grootswagers, T., Carlson, T. A., & Rich, A. N.
506   (2020). The influence of object-colour knowledge on emerging object representations
507   in the brain. *Journal of Neuroscience*.

508 van Dongen, N. N., van Doorn, J. B., Gronau, Q. F., van Ravenzwaaij, D., Hoekstra, R.,
509   Haucke, M. N., Lakens, D., Hennig, C., Morey, R. D., & Homer, S. (2019). Multiple
510   perspectives on inference for two simple statistical scenarios. *The American*
511   *Statistician, 73*(sup1), 328–339.

512 Wagenmakers, E.-J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Love, J., Selker, R.,
513   Gronau, Q. F., Šmíra, M., & Epskamp, S. (2018). Bayesian inference for psychology.
514   Part I: Theoretical advantages and practical ramifications. *Psychonomic Bulletin &*
515   *Review, 25*(1), 35–57.

516 Wagenmakers, E.-J., Morey, R. D., & Lee, M. D. (2016). Bayesian benefits for the pragmatic
517   researcher. *Current Directions in Psychological Science, 25*(3), 169–176.

518

519

520