Beyond *p*-values in the evaluation of brain-computer interfaces

To statistically evaluate the performance of brain-computer interfaces (BCIs), researchers usually rely on null hypothesis significance testing (NHST), i.e. *p*-values. However, over-reliance on NHST is often identified as one of the causes of the recent reproducibility crisis in psychology and neuroscience. In this paper we propose Bayesian estimation as an alternative to NHST in the analysis of BCI performance data. For the three most common experimental designs in BCI research - which would usually be analyzed using a *t*-test, a linear regression, or an ANOVA - we develop hierarchical models and estimate their parameters using Bayesian inference. Furthermore, we show that the described models are special cases of the hierarchical generalized linear model (HGLM), which we propose as a general framework for the analysis of BCI performance. The HGLM framework allows the analysis of complex experimental designs with multiple levels of hierarchy (e.g. multiple sessions, multiple subjects, multiple groups) and can accommodate different types of non-normal data (e.g. classification accuracy), which are often analyzed under inappropriate assumptions with NHST. We demonstrate the effectiveness of the proposed models on three real datasets and show how the results obtained with Bayesian estimation can give a more nuanced insight into BCI performance data, compared to NHST. Therefore we believe that a wider adoption of the Bayesian estimation approach in BCI studies could bring about greater transparency in data analysis, allow accumulation of knowledge across studies, and reduce questionable practices such as "*p*-hacking". To achieve this goal, we provide all the data and code necessary to reproduce the presented results, allowing BCI researchers to use Bayesian estimation in their own work.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1828v1 | CC-BY 4.0 Open Access | rec: 7 Mar 2016, publ: 7 Mar 2016

Beyond *p*-values in the evaluation of brain-computer interfaces

Filip Melinscak* Luis Montesano*†
* Bit&Brain Technologies S.L., Paseo de Sagasta 19, 50008 Zaragoza, Spain
† University of Zaragoza, Aragon Institute of Engineering Research (I3A), I+D+i building, Mariano Esquillor s/n, 50018 Zaragoza, Spain

***Abstract*—To statistically evaluate the performance of brain-computer interfaces (BCIs), researchers usually rely on null hypothesis significance testing (NHST), i.e. *p*-values. However, over-reliance on NHST is often identified as one of the causes of the recent reproducibility crisis in psychology and neuroscience. In this paper we propose Bayesian estimation as an alternative to NHST in the analysis of BCI performance data. For the three most common experimental designs in BCI research – which would usually be analyzed using a *t*-test, a linear regression, or an ANOVA – we develop hierarchical models and estimate their parameters using Bayesian inference. Furthermore, we show that the described models are special cases of the hierarchical generalized linear model (HGLM), which we propose as a general framework for the analysis of BCI performance. The HGLM framework allows the analysis of complex experimental designs with multiple levels of hierarchy (e.g. multiple sessions, multiple subjects, multiple groups) and can accommodate different types of non-normal data (e.g. classification accuracy), which are often analyzed under inappropriate assumptions with NHST. We demonstrate the effectiveness of the proposed models on three real datasets and show how the results obtained with Bayesian estimation can give a more nuanced insight into BCI performance data, compared to NHST. Therefore we believe that a wider adoption of the Bayesian estimation approach in BCI studies could bring about greater transparency in data analysis, allow accumulation of knowledge across studies, and reduce questionable practices such as "p-hacking". To achieve this goal, we provide all the data and code necessary to reproduce the presented results, allowing BCI researchers to use Bayesian estimation in their own work.**

***Index Terms*—Brain-computer interface (BCI), classification accuracy, Bayesian inference, Bayesian estimation, null hypothesis significance testing (NHST), *p*-values, hierarchical models, generalized linear model (GLM).**

I. Introduction

A little more than a decade ago, John Ioannidis put forward a statistical argument with a controversial conclusion: most published research findings are false [1]. The main point of Ioannidis' argument was that the post-study probability of a statistically significant research finding being true is rarely above 50% when one takes into account all the relevant statistical factors. Although Ioannidis' claim was based on theoretical and simulation-based reasoning, it was corroborated on empirical grounds in two recent studies. First, Button et al. have estimated the median statistical power (i.e. probability of rejecting the null hypothesis when it is false) of neuroscientific studies to lie between 8% and 31%, based on empirical evidence from 49 meta-analyses [2]. Low statistical power is not only a concern because of the wasted resources, but also because the statistically significant results from low-powered studies have small probability of actually being true. Second, a recent study by the Open Science Collaboration has tried to estimate the reproducibility of psychological science [3]. This collaborative effort entailed replicating 100 experiments, mainly from the fields of social and cognitive psychology. Although 97% of original studies were statistically significant at the 5% significance level, only 36% of replications reached significance; moreover, the mean effect size of the replications was halved in magnitude with respect to originally reported effects. These results have prompted calls for reform and the current situation has been referred to as a "reproducibility crisis" or a "statistical crisis" in science [4].

Although research on brain-computer interfaces (BCIs) is often focused on the engineering challenges, much of experimental methodology and statistical practices have been inherited from fields such as psychology and neuroscience. Hence, it seems prudent to also consider the implications of the statistical crisis on BCI research. With the recent advances in BCI research, which have brought BCIs closer both to markets and clinics, the stakes that depend on the veracity of research claims have also risen. The need of more rigorous statistical treatment of BCI results has been recognized [5–7], but the literature on the topic is still scant, and the statistical validation is in practice often carried out mechanistically and under inappropriate assumptions.

One of the issues often identified as the crux of the statistical crisis in science is the heavy reliance on null hypothesis significance testing (NHST), i.e. *p*-values. The reliance on NHST has been widely criticized in the statistical literature, and it is beyond the scope of this paper to rehash all the arguments surrounding NHST (for some discussion see references [8–14]). One of the proposed solutions for the deficiencies of NHST is the so-called "Bayesian new statistics" [15]. This framework differs from NHST in two major ways: first, instead of hypothesis testing, the goal is estimation of model parameters with uncertainty; and second, instead of using frequentist inference, parameters are estimated using Bayesian inference.

In the area of BCI research and brain decoding studies, Bayesian methods have already shown promise in the analysis of classification results. Olivetti et al. applied Bayesian inference to test the hypothesis of a decoder performing at chance level in a population of users [16]. An important feature of this work is that the decoder performance is modeled in a

Corresponding author: Filip Melinscak (filip.melinscak@bitbrain.es)

hierarchical fashion, taking into account that the group level accuracy is derived from subject-level accuracies, which are in turn estimated on a finite sample of trials. In a similar manner Brodersen et al. proposed several Bayesian hierarchical models of classification performance, also in the context of brain decoding studies [17]. Their approach focused more on estimation than hypothesis testing, in line with the trends of "new statistics" previously outlined. Importantly, the hierarchical approach was contrasted with classical non-hierarchical approaches and shown to be superior, and the models were extended to the case of unbalanced class proportions.

Although the aforementioned works have demonstrated the effectiveness of the Bayesian approach to the evaluation of BCIs, Bayesian inference is still rarely used in practice. One possible reason, which we try to address in this paper, is that previous works have illustrated the Bayesian approach only for the most simple experimental design: testing a single BCI with a group of subjects (which would usually be analyzed using a *t*-test). In practice, however, BCI studies often utilize more complex experimental designs.

The main contribution of this paper is to bridge this apparent gap between developments in statistical methods and BCI research practice. We show that the three most common BCI experimental designs can be formulated within a hierarchical generalized linear model. The usual *t*-test, regression and ANOVA approach can be seen as special cases of the generalized linear model. We demonstrate the effectiveness of this approach on three previously published studies, corresponding to the three main BCI experimental designs, and show how the Bayesian estimation approach can lead to a more nuanced understanding of the obtained results.

The proposed approach is highly flexible and can easily accommodate even more complex experimental designs including multiple levels of hierarchy (e.g. multiple sessions per subject, multiple subjects per group, multiple groups per study), multiple experimental factors and multiple covariates of interest. Unlike in the classical approach, all the modeling assumptions are overtly stated, can be scrutinized, and easily changed if found unsatisfactory. Finally, the implementation of the three proposed models, together with the data and code that produced the results of this paper, are made openly available online at www.github.com/fmelinscak/bayesian-bci-performance.

II. Background

Most BCI studies involve answering questions of the following three types:

- "how well does a BCI perform?",
- "how is some independent variable of interest associated with BCI performance?",
- "how does performance of different BCI approaches compare?"

We will now consider how NHST answers these questions, what are some of the problems associated with this statistical approach, and what are the possible solutions. Additionally, we will illustrate the difference between NHST and Bayesian estimation on a simple example.

A. Problems with *p*-values in BCI research

The NHST in practice usually consists of three steps:

1) choosing an appropriate test statistic (implicitly, this correspond to assuming a data model and defining the null hypothesis),
2) computing the *p*-value,
3) rejecting the null hypothesis if the *p*-value is smaller than the predetermined significance level α (usually fixed at 5%).

Corresponding to the three most common BCI research questions, the null hypothesis usually takes on one of the following forms: (i) a BCI is operating at the chance level in the subject population; (ii) there is no association between an independent variable of interest (e.g. hours of sleep) and BCI performance; (iii) there is no difference in performance between multiple experimental or computational approaches (e.g. utilizing different stimuli or classifiers). These null hypotheses are usually tackled using the *t*-test, linear regression, or ANOVA, respectively.

We can now see the first problem of NHST in BCI research – most often we do not *a priori* believe the exact null hypotheses: BCIs rarely work exactly at chance level in the user population, there is usually some association between an independent variable and BCI performance, and multiple computational or experimental approaches will almost never yield the same performance. This has the worrying implication that we can always reject the null hypothesis as long as we collect enough data. A related problem is that a *p*-value gives us the probability of the data given the null hypothesis $P(\mathrm{data}|H_0)$, whereas we usually conduct experiments in order to assess the plausibility of hypotheses in the light of the observed data, i.e. to obtain the probability $P(H_0|\mathrm{data})$. Moreover, the *p*-value gives us no indication of the estimated effect size or uncertainty of the estimate, which is what we usually care about – for example, we usually want to know how well a BCI is performing and how certain we are in this estimate, rather than if the accuracy is strictly above chance level.

Another problematic aspect of *p*-values is their dependency on the unobserved data. Although *p*-values are often used for their supposed objectivity, they depend on the usually unstated and possibly unknowable intentions of the experimenter and the analyst – both the decision to stop collecting data and testing intentions affect *p*-values. For example, recomputing *p*-values after every subject has a 100% chance of eventually obtaining a significant result with a flexible sampling plan, even when the null hypothesis is exactly true. But even when the sampling plan is pre-specified and there is no problem of multiple comparisons (i.e. "*p*-hacking"), if data analysis choices are made contingent on the obtained data, or interim results, the *p*-values are no longer valid. This is known as the problem of researchers' degrees of freedom [18] or the problem of the "garden of forking paths" [19]. The problem of *p*-values' sensitivity to testing and stopping intentions is particularly relevant to BCI research where degrees of freedom in data analysis abound, choices of a computational approach are often contingent on interim results (e.g. choosing

a classifier based on grand averages of features), and the sampling plans are usually flexible.

And finally, but perhaps most importantly, the use of NHST leads to a black-and-white mode of scientific reasoning and to frequent misunderstanding of the results [20, 21]. On one hand, statistically significant effects are believed to be true, although they might be practically insignificant in size or we might have large uncertainty about the effect size; on the other hand, statistically non-significant results are discarded as being false, although they might stem from insufficient data rather than a lack of a practically significant effect. The problem gets compounded by the usual publication and reviewing practices, where the "$p < 0.05$" statement is often a necessary condition for a result to be accepted and published. This practice distorts the scientific record and litters it with statistically significant, but perhaps uncertain or inconsequential results, at the same time robbing us of negative, but perhaps fairly certain and practically relevant results [22–24].

B. Moving beyond NHST

One recent proposal to improve statistical practices and replace NHST has been termed "new statistics" [25]. The "new statistics" mostly involves recommendations of replacing NHST and p-values with the estimation of effect sizes and providing frequentist confidence intervals (CIs) for the estimated effects in order to quantify uncertainty. Although these methods are not new by themselves, their wide adoption by researchers would be a notable departure from the common practice. In our view, the most important aspect of "new statistics" is the rejection of the black-and-white thinking induced by the NHST. Instead of asking whether the effect is statistically significant, we can pose the more nuanced questions of how big the effect is and how uncertain we are of our estimate.

Although we believe that the adoption of "new statistics" in BCI research would be a step forward, adoption of confidence intervals instead of p-values would not solve all the problems associated with NHST. Since both p-values and CIs are based on the frequentist statistical methods they share some of the previously outlined problems. Most notably, frequentist CIs also depend on the possibly covert testing and stopping intentions of the analyst. Therefore all the problems related to the researchers' degrees of freedom or the "garden of forking paths" apply to the confidence intervals just as much as the p-values. Moreover, just like p-values, frequentist CIs are often misinterpreted by researchers [26, 27].

An alternative to frequentist methods, and a possible solution to some of the problems with NHST, are Bayesian methods. One important distinction between frequentist and Bayesian inference is that Bayesian inference is insensitive to the stopping and testing intentions. The estimation approach of the "new statistics", but in a Bayesian framework, has recently been proposed under the name "Bayesian new statistics" [15]. This proposal argues that Bayesian methods are more apt at achieving the goals of "new statistics", namely building a cumulative body of knowledge based on estimating effect sizes. At the high level, the proposed Bayesian estimation approach can be summarized in the following steps, partly analogous to NHST:

1) hypothesizing a probabilistic model of the data (i.e. describing the dependence of the data on the model parameters and the prior information about the parameters),
2) estimating the model parameters conditional on the observed data using the Bayes' rule (i.e. computing the posterior probability distribution of the parameters),
3) communicating the inference results (i.e. the posterior distribution) using numerical and graphical summaries.

C. NHST vs. Bayesian estimation: a simple illustration

Since BCI literature is dominated by NHST, and Bayesian estimation is not yet a common practice in BCI research, we will now compare the two approaches on a simple example. We will use a common setup for both methods, assuming that we have experimentally obtained a random, independent sample $d = \{y_i | i = 1, \ldots, N\}$, where i indexes individual observations of a continuous random variable y, and N is the sample size. We have generated one such dataset ($N = 14$) using random normal numbers with mean 1 and standard deviation 3; the dataset is shown in Figure 1.A. Let us suppose that the goal of the experiment is to characterize the mean of the population from which the sample has been drawn.

In both NHST and Bayesian estimation, the first step is to hypothesize a model that could describe the data generating mechanism. In this example, the data generating mechanism is known but we will model the data as being normally distributed with unknown mean and variance parameters, i.e. $y \sim \mathrm{Normal}(\mu, \sigma^2)$. The model can also be represented graphically, by a directed acyclic graph (DAG), as shown in Figure 1.B.

In the NHST framework, the statistical question that might correspond to the substantive goal of characterizing the mean of the population is "does the mean μ differ significantly from 0?" An appropriate statistical test of this null hypothesis, under the given model assumptions, would be the t-test. In the given example the value of the t-statistic is 1.09 and the corresponding p-value is 0.29. Therefore, we would not reject the null hypothesis that the mean μ equals 0, at the usual 0.05 significance level.

In contrast, Bayesian estimation answers the question "what are the plausible values of the population mean μ?" The question is answered by the posterior distribution $p(\mu, \sigma|d)$, which provides the plausibility of all parameter values, conditional on the data. The posterior can be obtained by applying the Bayes' rule, i.e. combining the observed data d, the assumed model of the data (in the form of a likelihood function $p(d|\mu, \sigma)$), and the prior knowledge (in the form of a prior distribution $p(\mu, \sigma)$). The full posterior for the given example is shown in Figure 1.C, and it contains all the information about the parameters that is provided by the data, but also by the prior (unlike in NHST). Since the main question in the given example relates only to the mean parameter μ, we can summarize the full posterior $p(\mu, \sigma|d)$ with the marginal posterior $p(\mu|d)$ shown in Figure 1.D (for comparison with

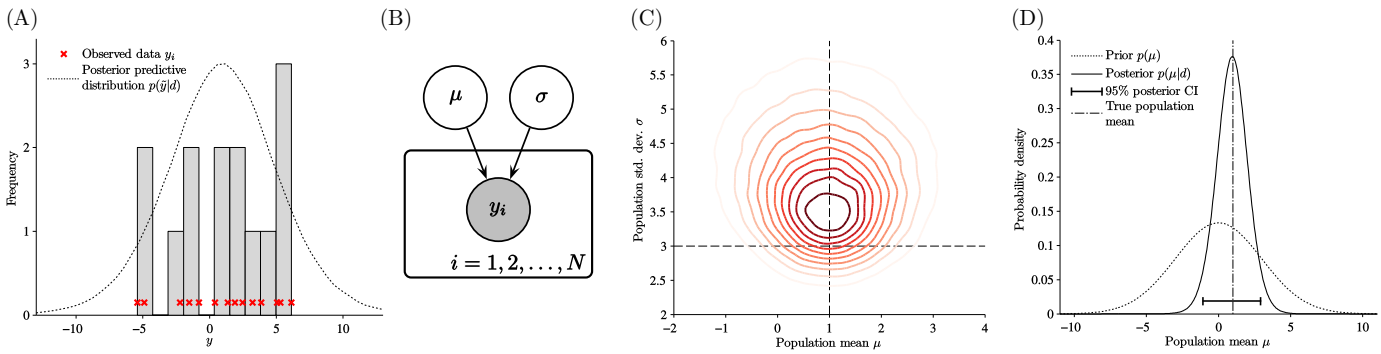

Fig. 1: (A) The generated dataset (red crosses and the histogram) and the corresponding posterior predictive distribution for future data (the density is arbitrarily scaled). (B) Diagram of the model for normal data in the form of a DAG. Arrows indicate dependency, shading indicates observed variables, and plate notation indicates repetition. (C) The full posterior distribution of the population mean μ and the population standard deviation σ. The dashed lines indicate true values of the parameters. (D) The marginal posterior and prior distribution of the population mean μ, together with indicated posterior 95% CI and the true parameter value.

the posterior, Figure 1.D also shows the marginal prior $p(\mu)$ that was used in the analysis). The marginal can further be numerically summarized, e.g. by its median (0.951) and 95% CI ([-1.12, 2.97]). Additionally, it is also possible to estimate future data $\tilde{y}$ using the posterior predictive distribution $p(\tilde{y}|d)$, which can be derived from the posterior. The posterior predictive distribution is shown in Figure 1.A, and comparing it to the histogram of the observed data constitutes a check of the model fit (i.e. a posterior predictive check).

We can now compare conclusions drawn from NHST and Bayesian estimation on the given dataset. Whereas NHST falsely fails to reject the null hypothesis that the population mean is 0, Bayesian estimation provides us with a more nuanced view: it shows we have a large uncertainty about the population mean (due to the small sample size), and that plausible values of the population mean span a wide interval that includes 0, but also a range of both large negative and positive values. Moreover, the posterior 95% CI includes the true value of μ and the posterior $p(\mu|d)$ is peaked around the true value. A more thorough account of the inference procedure in both the NHST and Bayesian estimation frameworks is given in the Appendix A.

III. Methods and Materials

A more detailed description of the Bayesian estimation approach, as we have used it in this paper, consists of the following steps:

1) define the relevant data d obtained from an experiment,
2) formulate a model for the data in the form of a likelihood $p(d|\theta)$ and state underlying assumptions,
3) formulate a prior for the model parameters $p(\theta)$ and motivate the choice,
4) use Bayes' rule to infer the posterior $p(\theta|d)$ (e.g. via Markov chain Monte Carlo simulation),
5) provide numerical and graphical summaries of the posterior and interpret them,
6) evaluate the model using a posterior predictive check: compare the posterior predictive distribution $p(\tilde{d}|d)$ with the observed data d.

It should be noted that the outlined process is iterative: if the model is found unsatisfactory in evaluation, it can be modified accordingly and the process is repeated. Furthermore, we would like to point out that this process applies to situations where the experiment has already been conducted and the data collected. Although this is a common situation in practice, it is often possible to consider the model that is going to be used to analyze the data before conducting an experiment. With the model formulated before the experiment, simulated data can be used to judge if the experimental design is adequate to answer research questions of interest, and modify the design if necessary. Lastly, the outlined process is not meant to cover all possible elements of an analysis, but rather provide a rough guideline. Therefore some important tools – such as model comparison, sample size planning, sensitivity analysis, etc. – have been omitted from the described framework, but are touched upon in the Discussion section.

We now illustrate the outlined Bayesian estimation approach on the three most common experimental designs in BCI research, listed here in the order of increasing complexity:

- performance of a single BCI in a group of subjects (Model 1),
- association between a subject-specific variable and BCI performance (Model 2),
- comparison of different BCI approaches in a within-subject design (Model 3).

Subsequently, we show that these three models are special cases of the hierarchical generalized linear model, which is proposed as an encompassing model for the analysis of BCI performance.

A. Model 1: performance of a single BCI in a group of subjects

A common question in BCI research, especially when introducing a novel computational or experimental approach, is "how well does a BCI approach perform in a particular population of subjects?" To answer the question a simple experimental design is used: the performance of the BCI is recorded for a sample of subjects, with multiple trials per

subject. The goal of statistical inference is then to estimate the mean and the variance of BCI performance in the population from which the subjects were recruited.

We will first assume that the data d from the experiment has been recorded as a list of pairs $d = \{(y_i, T_i)|i = 1, \ldots, N_S\}$, where i is the index of the subject, y_i is the number of successful trials, and T_i is the total number of trials. The model for this experimental design is shown in Figure 2, and we now examine the assumptions behind the model and the interpretation of its parameters.

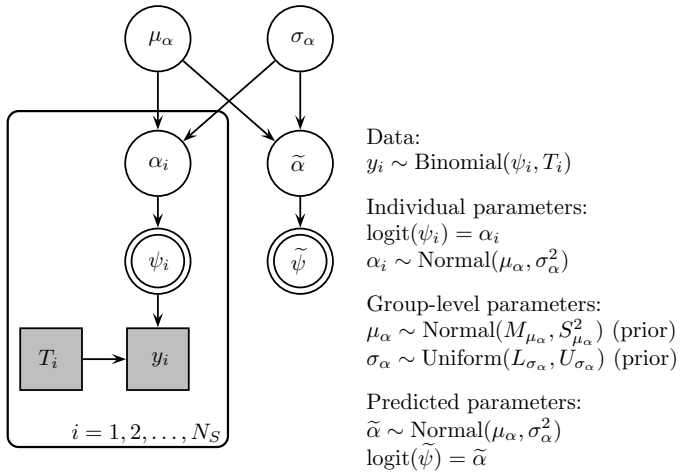

Fig. 2: Diagram and specification of Model 1 (performance of a single BCI in a group of subjects, together with predicted parameters). See the caption of Figure 1 for the interpretation of the diagram elements. Additionally, square nodes denote discrete variables and doubly outlined nodes are deterministically dependent on their parents. See the main text for the interpretation of variables.

If we assume that each of the T_i trials is an independent binary random variable (which indicates success or failure of the BCI), then the total subject-wise number of successful trials y_i can be modeled as a binomial random variable with the probability of success ψ_i (i.e. individual accuracy).

Next, we would like to model the subject-wise performance as being a sample from a population, which could in turn be described with a normal distribution; however, the individual accuracies are measured on the probability scale (on the interval [0, 1]) and the normal distribution is supported over the whole real line. To overcome this discrepancy we can transform individual accuracies ψ_i from the probability scale, to individual accuracies α_i on the log-odds scale using the logit function:

$$\alpha = \text{logit}(\psi) = \log O(\psi) = \log \frac{\psi}{1 - \psi}, \qquad (1)$$

where $O(\psi)$ are the odds corresponding to probability ψ. E.g. this transformation will map probabilities 0, 0.5, and 1 to log-odds of $-\infty$, 0, and $+\infty$, respectively.

The individual performance on the logit scale α_i can now be modeled as a sample from the normally distributed group-level performance, with mean parameter μ_α and between-subject variance parameter σ_α^2. We might also be interested in interpreting the group-level mean accuracy μ_α on the probability scale; in this case we can use the inverse of the logit function, i.e. the logistic function:

$$\mu_\psi = \text{logit}^{-1}(\mu_\alpha) = \frac{1}{1 + \exp(-\mu_\alpha)}, \qquad (2)$$

where μ_ψ is the group-level accuracy on the probability scale. Although the probability scale might be more common in practice (and thus more intuitive), we would argue that the log-odds scale has an important advantage in interpretation. Consider the following two cases: (i) increase of accuracy from 51% to 52%, and (ii) increase of accuracy from 98% to 99%. Although both cases represent a unit increase in probability, the first increase would usually be practically negligible, whereas the same increase in the second case could be of significant practical value because it halves the frequency of errors. In contrast, the corresponding improvements on the log-odds scale – 0.04 and 0.7, respectively – more closely reflect the practical importance of the accuracy increase.

The last step before applying the Bayes' rule is to define the prior distributions of the top-level model parameters μ_α and σ_α. For the group-level mean μ_α we use a vague normal prior on the logit scale with mean $M_{\mu_\alpha} = 0$ and standard deviation $S_{\mu_\alpha} = \sqrt{2}$. This choice of a prior corresponds to a fairly uniform distribution on the probability scale, indicating the lack of strong prior information [28, p. 85]. For the variance between subjects we use a uniform prior over the standard deviation σ_α, with a lower bound $L_{\sigma_\alpha} = 0$ and a relatively large upper bound $U_{\sigma_\alpha} = 10$, again indicating the lack of prior information, and letting the data to drive the inference (for other choices consult refs. [29–31]).

Since we are often interested not only in the average performance and variance in the population, but also in predicting the performance of future subjects, we define predicted performance of a new subject $\widetilde{\alpha}$ on the logit scale, or equivalently $\widetilde{\psi}$ on the probability scale. The distribution of predicted performance reflects our posterior uncertainty about both the population-level mean and variance, given the data that we have observed in the experiment.

Example dataset for Model 1: To illustrate the analysis with Model 1, we chose the study of Power et al. [32]. This study investigated whether it is possible to implement a NIRS-based BCI for binary communication by differentiating cognitive tasks of mental arithmetic and music imagery. Each of the 10 healthy subjects participated in three experimental sessions, with each session consisting of 17 trials of mental arithmetic, and 17 trials of music imagery: in total there were 102 trials for each subject, with balanced class proportions (hence, the chance level was 50%). The BCI was tested using 5-fold cross-validation, and the paper describing the study provides the accuracy obtained in cross-validation (averaged across folds) for each subject, with the trials from all the sessions aggregated together. The exact number of trials that were correctly classified is not provided for each subject, and therefore we have obtained the approximate number of correct trials by multiplying the reported subject-wise accuracy with the total number of trials, and rounding to the nearest integer.

B. Model 2: association between a subject-specific variable and BCI performance

Another frequent question in BCI research is "how is some subject-specific variable associated with BCI performance?" For example, we might be interested in the association between the hours of sleep a subject has had, and the BCI performance he obtained. The experimental design used to answer this question is essentially the same as the one used with Model 1, but now the value of the subject-specific variable also has to be recorded.

The data obtained from such an experiment can be represented as a list of triples $d = \{(y_i, T_i, x_i)|i = 1, \ldots, N_S\}$. The i, y_i and T_i have the same meaning as in Model 1 and the x_i represents the recorded value of the continuous, subject-specific variable of interest. It is useful to transform the values of the covariate x_i to z-scores z_i by subtracting the sample mean $\bar{x}$, and standardizing with the sample standard deviation s_x – as we will see shortly, this leads to more meaningful model parameters. The model we propose for this type of data is specified in Figure 3.

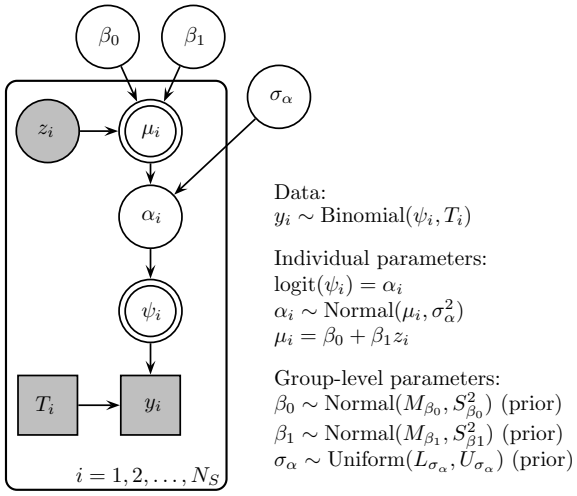

Fig. 3: Diagram and specification of Model 2 (association between a subject-specific variable and BCI performance). See Figure 2 for notation and the main text for the interpretation of variables.

The main change in Model 2, relative to Model 1, is that the subject-specific logit accuracies α_i are now not drawn from a single normal distribution, but rather from a normal distribution whose mean μ_i depends linearly on the value of the covariate z_i. The parameters of this linear association are the intercept β_0 and the slope β_1. Since we are using z-scores of the covariate, we can interpret β_0 as the expected logit accuracy μ for the average value of the covariate x (i.e. when z is zero), and β_1 as the expected increase in logit accuracy obtained when the covariate x increases for one standard deviation (i.e. unit increase in z)[1].

[1]Had we not standardized the covariate x, the intercept β_0 would be interpreted as the expected logit accuracy μ when the value of the covariate x was zero, and the slope β_1 would be interpreted as the change in the expected μ for a unit increase in x. In many cases the zero value for the covariate x might not be meaningful. Moreover, standardizing x leads to scale invariance, allowing for easier modeling of the slope β_1.

Although the log-odds scale is mathematically convenient in allowing us to fit a linear additive model, the parameter interpretation on this scale may not be so intuitive. One way to obtain more interpretable results is to use the odds scale – a linear additive model on the log-odds scale will give a multiplicative model on the odds scale. For example, let us consider the expected odds of success $O(\psi)$ for known parameters β_0, β_1, and a known value of the covariate z:

$$\begin{aligned} E[O(\psi)|z, \beta_0, \beta_1] &= E[\exp(\alpha)|z, \beta_0, \beta_1], \\ &= \exp(\beta_0)\exp(\beta_1 z), \end{aligned}$$

where we have used eqn. (1) to relate logit accuracy α with odds of success $O(\psi)$, and the specification of the model in Figure 3 to compute the expectation. In this formulation we can interpret $\exp(\beta_0)$ as the baseline odds and $\exp(\beta_1)$ as the factor by which the baseline odds are multiplied for a unit increase in the covariate z.

The interpretation of the variance parameter σ_α also changes relative to the same parameter in Model 1: σ_α no longer represents the overall between-subject variance, but rather the between-subject variance observed when we account for the the covariate x (i.e. the variance unexplained by the covariate).

The priors for the top-level parameters β_0, β_1, and σ_α are again relatively vague, expressing the lack of prior information or the intention to let the data determine the inferences. For the intercept β_0 we use the same vague prior as for the group-level mean μ_α of Model 1. For the slope β_1 we use a "skeptical" normal prior, with mean $M_{\beta_1} = 0$ (indicating lack of prior information on the direction of the effect), but with a large standard deviation $S_{\beta_1} = 5$, allowing the inferred effect to have a large size, if such an inference is supported by the data (see refs. [31, 33] for more discussion of priors in logistic regression). For the unexplained variance parameter σ_α we use the same prior as in Model 1.

Although the predicted accuracy $\widetilde{\alpha}$ is not specified in Figure 3 for the sake of simplicity, it is obtained similarly as it was in Model 1, with a minor addition – it is necessary to specify all the values of the covariate x for which we wish to predict the accuracy.

Example dataset for Model 2: To illustrate the analysis with Model 2, we chose the study of Blankertz et al. [34]. This study investigated if there is an association between the spectral power of resting state EEG in the alpha band over the motor cortex, and the subsequent performance in operating a motor imagery BCI for binary selection. Each of the 80 healthy subjects participated in two phases of the experiment – a calibration phase and an online feedback phase. The calibration phase was used to train the BCI and the feedback phase was used to test it in a balanced, binary selection task (hence, the chance level was 50%). The feedback phase consisted of three runs, each with 100 trials. Out of the 100 trials in each feedback run, 20 were used for the adaptation of the BCI, and 80 were used to test it. Therefore, the maximum number of test trials per subject was 240, but some of the subjects did not complete all of the feedback runs. While the subject-wise values of the covariate (i.e. resting alpha power) and the accuracies are available in the paper describing

the study, the subject-wise numbers of trials are not given; however, the authors of the paper have kindly provided us with the numbers of trials upon request.

C. Model 3: comparison of different BCI approaches in a within-subject design

The third common question in BCI research that we consider in this paper is "how well do different BCI approaches work in a population of subjects?" Here under the "BCI approach" we denote both differences in the employed experimental paradigm (e.g. changing the set of used stimuli) and differences in the computational implementation of the BCI (e.g. changing the used classifier). We also constrain our attention to the within-subject (i.e. repeated measures) experimental designs, where each of the subjects uses the BCI in all the experimental conditions of interest. This is the most common setup in practice, especially for offline studies of different computational approaches. In such "computational experiments" there is usually no barrier to trying out all the approaches in each subject. We also limit the discussion to a study of a single discrete experimental factor, although the approach is general and can easily be extended to multiple factors (see subsection III-D "A unifying model for the analysis of BCI performance").

The data of a single-factor, within-subject BCI experiment can usually be represented as a list of tuples $d = \{(y_i, T_i, l_i, s_i)|i = 1, \ldots, N_O\}$. The y_i and T_i again have the same meaning as before, whereas $l_i \in \{1, \ldots, N_L\}$ is the level of the experimental factor (i.e. the experimental condition), $s_i \in \{1, \ldots, N_S\}$ is the index of the subject, and i is the index of the observation. While in Model 1 and 2 we did not record explicitly for which subject each observation was made, as each subject contributed only one observation, here we need to explicitly take into account which observations come from the same subject. This adds an additional level in the hierarchy of the model.

Model 3 (shown in Figure 4) shares most of its structure with Model 2, but some changes are necessary to accommodate multiple observations from the same subject. The predicted performance μ_i for a particular level of the factor l_i and subject s_i is modeled as a linear combination of the grand-average performance β_0, factor-level effect β_{1,l_i} and the subject-specific effect η_{s_i}. In this parametrization β_0 is the expected performance over all the levels of the experimental factor and all the subjects (i.e. grand-average). Parameters $\beta_{1,k}$ are the level-specific deviations from the grand-average (i.e. fixed effects). The random subject-specific effects η_j are modeled as normally distributed with mean zero and between-subject variance σ_η. The η_j effects represent the subject-specific deviations from the grand-average performance β_0, when averaging over all the levels of the factor. To enforce the interpretation of parameters $\beta_{1,k}$ and η_j as deviations from the grand-average β_0, it is necessary to constrain the two sums over these sets of parameters to zero (i.e. sum-to-zero or STZ constraints). The parameter σ_α is interpreted as the variance that has not been explained neither by the factor-specific effects, nor by subject-specific effects.

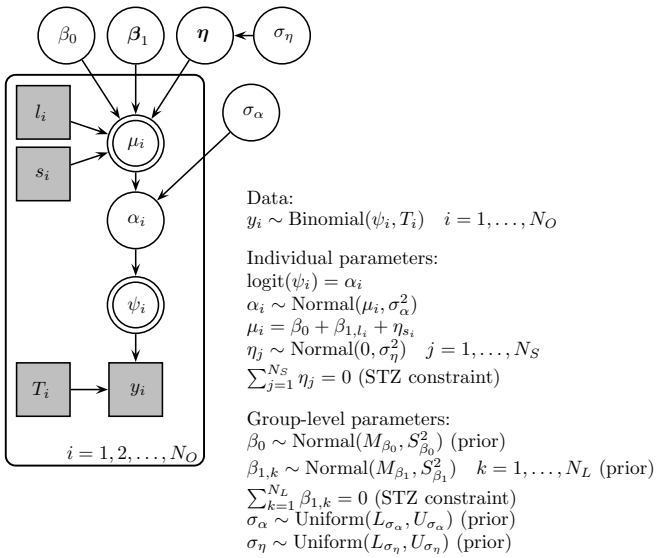

Fig. 4: Diagram and specification of Model 3 (comparison of different BCI approaches in a within-subject design). See Figure 2 for notation and the main text for the interpretation of variables.

For the top level parameters β_0, $\beta_{1,k}$, σ_η, and σ_α we again use vague priors. The forms and parameters of the priors are the same as in Model 2.

The predicted variables have again been omitted from the model in Figure 4 for the sake of simplicity. To predict the logit accuracy $\widetilde{\alpha}_k$ of a future subject for all the levels k of the experimental factor, we first define the predicted subject-specific effect $\widetilde{\eta}$ which depends on the inferred σ_η. Then we model the dependency of $\widetilde{\alpha}_k$ on the predicted effect $\widetilde{\eta}$ and the inferred top level parameters β_0, $\beta_{1,k}$, and σ_α.

Example dataset for Model 3: To illustrate the analysis with Model 3, we chose the study of Brunner et al. [35]. This study compared three EEG-based BCI approaches for binary selection: a motor-imagery paradigm based on the event-related desynchronization (ERD), a visual paradigm based on steady-state visual evoked potentials (SSVEP), and a hybrid paradigm combining motor imagery and visual stimulaton. Each of the 12 healthy subjects used all of the three BCI approaches in a binary selection task with balanced classes (chance accuracy was 50%). The experiment consisted of a calibration phase and an online feedback phase. The calibration phase was used to train the BCIs and the feedback phase was used to test them. Although the BCIs were also tested within the calibration phase using cross-validation, here we only consider the results from the feedback phase. The feedback phase consisted of three runs, one per each BCI approach, with 40 trials per run.

D. A unifying model for the analysis of BCI performance

In the previous sections we have described a general methodology based on Bayesian parameter estimation and presented three use cases for the arguably most typical experimental designs in BCI research. All three models can be derived from a common model, the hierarchical generalized linear model (HGLM) [36]. Using this HGLM framework it

is also possible to derive models that cover other experimental designs. We now describe the HGLM for BCI performance and provide some directions on how to extend its applicability.

An HGLM for classification accuracy can be described as:

$$y_i \sim \text{Binomial}(\psi_i, T_i), \tag{3}$$

$$\text{logit}(\psi_i) \sim \text{Normal}(\mu_i, \sigma^2), \tag{4}$$

$$\mu_i = \beta_0 + \sum_{k=1}^{K} \beta_{1,k} x_{i,k}, \tag{5}$$

where equation (3) models the observed outcomes, equation (4) models the unexplained variability of individual accuracies, and equation (5) models the expected logit accuracy based on a linear prediction from explanatory variables $x_{i,k}$.

Previously described models 1 and 2 are direct instances of the described HGLM for accuracy. We can obtain Model 1 by modifying the linear predictor of equation (5) to a simple intercept-only form, i.e. setting $\mu_i = \beta_0 = \mu_\alpha$, where μ_α is the group-level accuracy. Model 2 is obtained simply by using only one continuous explanatory variable in the linear predictor, i.e. setting $K = 1$. However, the linear predictor of equation (5) does not restrict us to continuous variables – discrete variables can also be included by using dummy encoded binary variables. This allows us to implement multi-factor ANOVA-like models with multiple discrete factors or ANCOVA-like models with a mix of continuous and discrete explanatory variables. Moreover, instead of just using simple main effects, we can also study interactions between explanatory variables by including interaction terms (obtained as products of explanatory variables).

In addition to including multiple continuous and discrete explanatory variables, HGLM can also be extended with extra levels of the hierarchy. We can see an example of this in Model 3. To obtain Model 3 from the HGLM we modify the linear predictor as follows:

$$\mu_i = \beta_{0,i} + \sum_{k=1}^{K} \beta_{1,k} x_{i,k}, \tag{6}$$

$$\beta_{0,i} = \beta_0 + \eta_{s_i}, \tag{7}$$

$$\eta_j \sim \text{Normal}(0, \sigma_\eta^2). \tag{8}$$

Here we have used the varying intercepts $\beta_{0,i}$ to model the nesting of repeated measures within subjects. The same pattern of expanding the model by additional levels of hierarchy can further be applied to analyze datasets with, for example, multiple sessions per subject, multiple groups of subjects per study (e.g. a control and a patient group), multiple studies in a meta-analysis, etc.

It is also worth to consider cases of multi-class classification and classification with unbalanced classes. In both situations the HGLM described in equations (3)-(5) can simply be applied to trials of each class separately, with y_i and T_i representing the number of correct trials and the total number of trials for one of the classes. To deal with class unbalanced problems, class-specific accuracies can be combined into balanced accuracy (i.e. accuracy averaged over classes) for which the chance level is always $1/C$, where C is the number of classes [37, 38]. If we wish to model also the covariation of accuracy for different classes, instead of using separate univariate models, we can use a multivariate HGLM by utilizing a multivariate normal distribution in equation (4) [17].

The HGLM can also be used to model different types of performance metrics. For example, if the full confusion matrices are available for all the subjects they can be modeled as multinomial outcomes in an HGLM (i.e. multinomial regression). In this case we can also obtain the Cohen's kappa coefficient [39, p. 65-67]. If the BCI is used to predict or decode continuous variables, the HGLM can be used by modeling the errors as normally distributed outcomes in the equation (3). Lastly, if we wanted to model count-based metrics (e.g. number of commands completed in a period of time) we could use the log-Poisson version of the HGLM. For a more thorough account of different modeling possibilities with the HGLM we refer the interested reader to the text by Ntzoufras [40].

E. Computational details of the inference procedure

To inspect the properties of the joint posterior distribution $p(\theta|d)$, we have obtained a random sample from it by using Markov chain Monte Carlo (MCMC) simulation [41]. For MCMC sampling we used the freely available WinBUGS software [42]. For each of the analyses we ran three parallel MCMC chains, recording 50000 samples per each chain, after discarding the first 50000 samples (burn-in period). For each of the parameters presented in the Results section we have verified that the effective sample size was at least 10000 samples (i.e. Monte Carlo standard error was below 1% of the standard deviation of the parameter). Furthermore, we have checked the convergence of the chains by visual inspection of the traces and by verifying that the Gelman-Rubin statistic was below 1.1, which is usually taken as a threshold to diagnose convergence issues [43, 44].

IV. Results

A. Results from Model 1 on the example dataset

For Model 1 we will inspect both the parameter estimates at the subject level and at the group level. Although group level parameters are usually of greater interest, as we want to generalize out of the sample of the subjects, subject-level inferences might also be of interest – for example, if a pilot study is performed with the intention of screening subjects for a future study. In Figure 5 we show the results of estimating the parameters of Model 1 on the example dataset of Power et al. In Figure 5.A the obtained marginal posterior distributions of subject-level accuracies ψ_i are summarized by their medians and 95% CIs. Comparing the posterior medians to sample accuracies, we can see the pooling (or shrinkage) effect of the hierarchical model, where we have assumed the subjects' accuracies come from a common normal distribution (on the logit scale). For each subject, its accuracy estimate is influenced by the estimates for all the other subjects. This is most evident in the subjects which are further from the group mean accuracy: for this subjects estimates are most strongly shrunk towards the group mean. In this way information is pooled

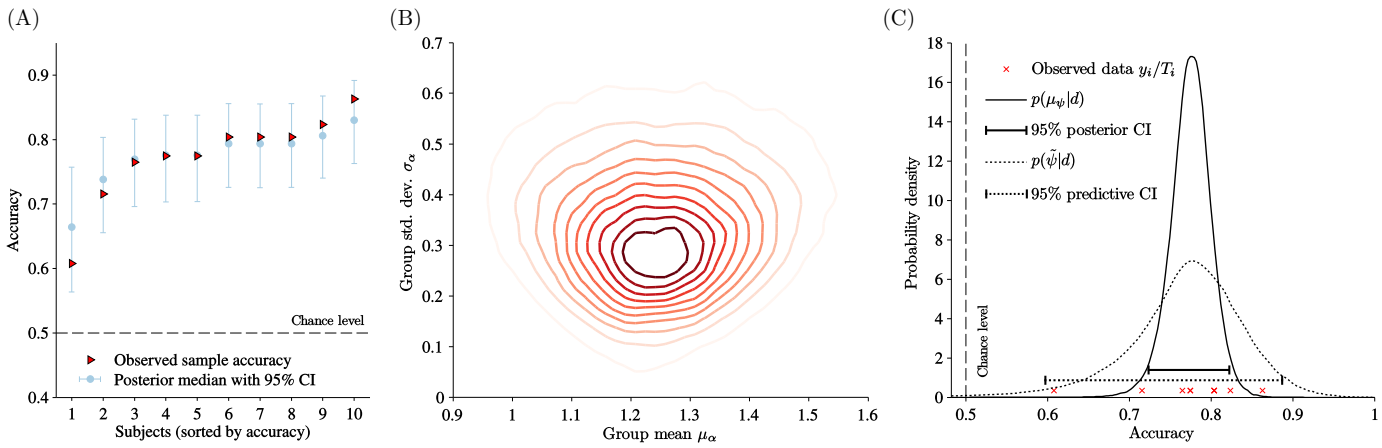

Fig. 5: First example dataset and the results of Model 1. (A) Subject-level inferences (posteriors) for the accuracy ψ_i on the probability scale, together with sample accuracies. (B) Group-level inference for the group mean accuracy μ_α and group accuracy SD σ_α on the logit scale. The contours are obtained using 2D kernel density estimation on the MCMC sample. (C) The posterior for the group mean accuracy μ_ψ on the probability scale, together with the posterior predictive distribution of accuracy $\tilde{\psi}$ on the probability scale and the observed sample accuracies (horizontal lines indicate 95% CI). The probability densities are obtained using kernel density estimation on the MCMC sample.

across subjects, and we avoid making extreme inferences based on noisy data, since shrinkage acts as a form of regularization.

Figure 5.B shows the results of inference at the group-level parameters, i.e. group mean accuracy, and the group SD of accuracy, both on the logit scale. From the joint posterior distribution depicted in the figure, we can clearly see which values of the parameters are jointly credible, and we can observe if there are correlations between the parameters in the posterior distribution. For example, in the given dataset we can see that for extreme credible values of mean accuracy μ_α, only large values of SD σ_α are plausible, whereas for central credible values of μ_α, a wider range of values for σ_α are credible.

Figure 5.C compares the observed sample accuracies (i.e. the data), the marginal posterior of the group mean accuracy μ_ψ (obtained by transforming μ_α to the probability scale, using the logistic function), and the posterior distribution of accuracy $\tilde{\psi}$ for future subjects (i.e. the posterior predictive distribution). Here we can see that the marginal distribution of mean accuracy is fairly narrow ($\mathrm{Mdn} = 0.776$, 95% CI: $[0.722, 0.822]$), mainly due to low inter-subject variation in performance. However, it is important to note that although the posterior of the mean is narrow, the posterior predictive distribution of the subject-wise accuracies is relatively wide-spread ($\mathrm{Mdn} = 0.775$, 95% CI: $[0.596, 0.891]$). This reflects the fact that the posterior predictive distribution takes into account both the mean and the variance of the subject population. Consequently, with an increasing sample size, the posterior distribution for the mean (or variance) would become increasingly peaked, whereas the posterior predictive distribution would stay relatively wide (unless the estimate for the variance decreased significantly with the new data).

With the MCMC sample of the posterior distribution, we can also answer other questions of interest. For example, 70% accuracy is often considered to be a lower bound for a BCI to be practically useful; we might therefore be interested in the probability that the mean group accuracy is above 70%. If we had the joint posterior in the analytical form, answering this question would require integrating all the variables except group-level mean accuracy out of the joint posterior, and then finding the area under the probability distribution for accuracies larger than 70%. However, since we have the MCMC sample from the posterior available, we can answer this question using Monte Carlo integration. Taking into account only the samples of accuracy μ_ψ corresponds to integrating out the other variables, and determining the proportion of samples of μ_ψ larger than 70% by simply counting them corresponds to integrating the marginal probability distribution of μ_ψ. In the example dataset, the posterior probability that group average accuracy exceeds 70% is $P(\mu_\psi > 0.7|d) = P(\mu_\alpha > 0.847|d) \approx 99.4\%$. From the posterior predictive distribution of future subject's accuracy $\tilde{\psi}$, we can find out also what is the probability that a future subject will obtain accuracy larger than 70%: $P(\tilde{\psi} > 0.7|d) \approx 85.8\%$.

As a qualitative check of the model, we can graphically compare the posterior predictive distribution of subject-wise accuracy with the observed subject-wise sample accuracies, and see if the observed data is credible given the model. In the presented case it seems that the model properly predicts (or "postdicts") the data from which it has been estimated, therefore not eliminating the model as a good description of the data.

B. Results from Model 2 on the example dataset

From Model 2 results, we will for brevity only look into the results at the group-level, although the subject-level parameter estimates are also available in the full joint posterior. In Figure 6 we show the dataset of Blankertz et al., as well as the results of inference based on Model 2. Figure 6.A shows values of the recorded covariate (alpha log-power) and the sample accuracies obtained by subjects. For reference, in this figure we also present the linear model fitted using ordinary

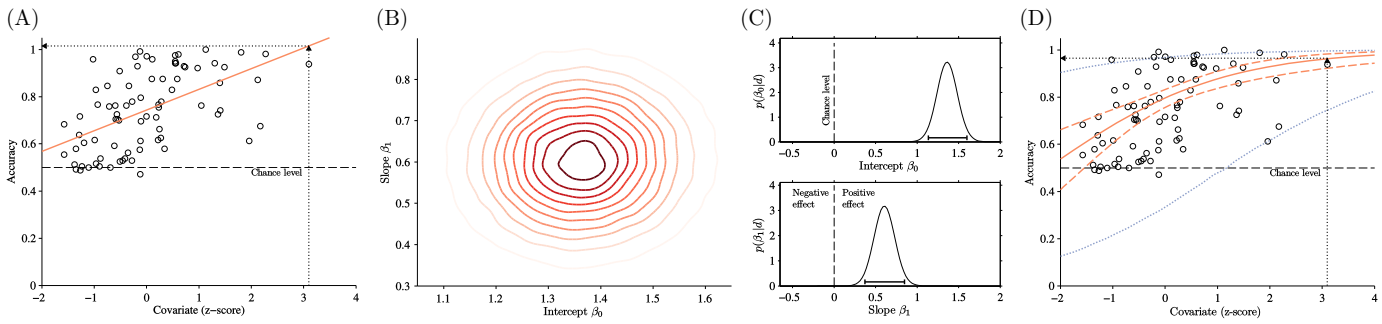

Fig. 6: Second example dataset and the results of Model 2. (A) The observed subject-wise values of the covariate (standardized) and the observed sample accuracies, with the ordinary least squares fit for reference. The dotted arrows show that the linear model can predict accuracies larger than 1 even for observed levels of the covariate. (B) Group-level inference for the intercept β_0 and slope β_1 on the logit scale. (C) The marginal posteriors for the intercept β_0 and β_1 (95% CI indicated). (D) The fitted logistic model on the probability scale. Orange lines show the posterior median (solid) and 95% CI for accuracy (dashed), and the dotted blue line indicates the 95% posterior predictive interval. All the predictions of the model are now constrained between 0 and 1, the natural boundaries for accuracy (as shown by the dotted arrows).

least squares procedure. The dotted arrows point out one of the problems with the linear, normal model – namely, the possibility of predicting accuracy larger than 100% (or smaller than 0%), even for the values of the covariate that are present in the dataset.

Figure 6.B presents the joint posterior distribution of the group-level parameters which are usually of main interest in studies of this type: intercept β_0 and slope β_1. The joint posterior shows that the slope and the intercept parameter estimates are not correlated for the given dataset. In Figure 6.C we can inspect the marginal posterior distributions of the intercept and the slope. With the posterior of the intercept we may again wish to answer questions such as: "what is the probability that the group level accuracy is above 70%, when controlling for the covariate z?" This can be answered with the probability $P(\text{logit}^{-1}(\beta_0 + \beta_1 z) > 0.7|d, z = 0) = P(\text{logit}^{-1}(\beta_0) > 0.7|d) = P(\beta_0 > 0.847|d) \approx 100\%$. We can also summarize the marginal posterior of the intercept on the logit scale with its median (1.36), and its 95% CI ([1.13, 1.60]). However, usually the slope β_1 is of greater interest, as it tells us the strength of the association between the covariate and the accuracy. In the given dataset we can determine with high level of certainty that higher alpha power has a positive association with accuracy ($P(\beta_1 > 0|d) \approx 100\%$), and that the effect is quite large ($\text{Mdn} = 0.606$, 95% CI: $[0.372, 0.844]$).

The model fit (i.e. posterior median of accuracy for a given value of the covariate), the point-wise confidence intervals, and the point-wise prediction intervals are shown in Figure 6.D. It is instructive to compare this model fit to the linear model fit in Figure 6.A. As we can see, the linear model predicts that the subject with the highest value of alpha power will have accuracy above 1, whereas the logistic model correctly constrains the predicted accuracies within the [0, 1] interval, due to the employed logit link function.

Again, we can asses the model qualitatively, by comparing the posterior predictive intervals with the observed data in Figure 6.D. For the given dataset we can see that most observed data points lie within the predictive interval; however, for lower values of the covariate the model predicts a larger proportion of accuracies below 0.5 chance level than we observe in the data. This is due to the fact that classifiers used in BCIs rarely perform below chance level, but in the Model 2 this prior information is not explicitly used. In future modeling efforts one might want to use this knowledge to explicitly constrain the predicted accuracies to the [0.5, 1] interval. As it is not impossible that in some studies we might want the model to predict below chance accuracies (e.g. if the data generating process is adversarial), for generality we have not pursued the direction of constraining the model only to above chance accuracies.

C. *Results from Model 3 on the example dataset*

With Model 3, we again look only at the group level results, although the inference procedure also provides us with the subject-level parameter estimates (in the full joint posterior). Figure 7 shows the dataset of Brunner et al. [35] and the results obtained from using Model 3 with this dataset. In Figure 7.A we can see the sample accuracies recorded for each of the subject with the three proposed BCI approaches – ERD, SSVEP, and hybrid. The sample accuracies recorded within the same subject are connected to indicate the within-subject nature of the experimental design.

The inferred posteriors of accuracy for different approaches are shown in Figure 7.B with violin plots. To obtain the inferred approach-specific accuracy, it is necessary to sum the grand average parameter β_0 (common to all the levels of the factor) and the approach-specific parameter $\beta_{1,k}$, where k indicates the level of the factor (in this dataset $k \in 1, 2, 3$, and corresponds to ERD, SSVEP, and hybrid approaches, respectively). This yields the accuracy on the logit scale, so we need to apply the inverse-logit mapping to obtain accuracies on the probability scale; i.e. the marginal distributions of interest are $p(\text{logit}^{-1}(\beta_0 + \beta_{1,k})|d)$. From the marginal posteriors we can see that the hybrid approach was the best performing one ($\text{Mdn} = 0.978$, 95% CI: $[0.946, 0.993]$), followed by SSVEP ($\text{Mdn} = 0.971$, 95% CI: $[0.929, 0.991]$), and ERD ($\text{Mdn} = 0.792$, 95% CI: $[0.622, 0.897]$).

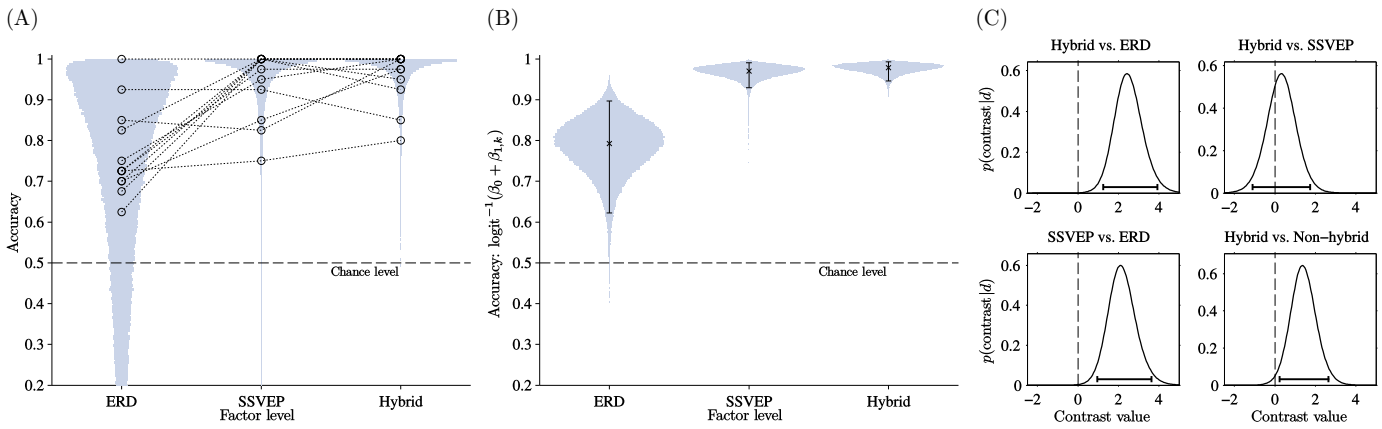

Fig. 7: Third example dataset and the results of Model 3. (A) The observed subject-wise sample accuracies for all three experimental conditions (within-subject data points are connected). Overlaid on observed accuracies are the posterior predictive distributions for the three levels of the factor. (B) Group-level posterior probability distributions for the average accuracy achieved with each approach, on the proability scale (median and 95% CI indicated). (C) Probability distribution for differences (contrasts) between different BCI approaches (on the logit scale). Horizontal lines indicate the 95% CI.

However, the main questions of interest in studies of this type pertain to the differences between different approaches. We explore the differences between approaches in Figure 7.C. For example, if we were interested in the difference between the hybrid and ERD approach we would take the estimate of the difference $\beta_{1,3} - \beta_{1,1}$ (the distribution for this difference is shown in upper left panel of Figure 7.C). More generally we can construct contrasts that encode the questions of interest in vector form $\mathbf{c} = [c_1, \ldots, c_{N_L}]^\intercal$ (to be a contrast, the elements of $\mathbf{c}$ have to sum up to zero). The contrast value estimate is then obtained with the dot product $\mathbf{c}^\intercal \beta_1$. For the aformentioned difference between the hybrid and the ERD approach, the contrast is $\mathbf{c} = [-1, 0, 1]^\intercal$. We can also use contrasts that combine multiple approaches; e.g., if we wanted to know if the hybrid approach is better than non-hybrid approaches (average of SSVEP and ERD), we would use the contrast $\mathbf{c} = [-0.5, -0.5, 1]^\intercal$ (this contrast is shown in lower right panel of Figure 7.C).

From the analysis of contrasts we can answer the main question of the original study directly – whether hybrid approach outperforms non-hybrid approaches – by calculating the probability $P(\text{Hybrid} > \text{Non} - \text{hybrid}) = P(\text{Hybrid} > \text{ERD\&SSVEP}) = P(\beta_{1,3} > 0.5\beta_{1,1} + 0.5\beta_{1,2}) \approx 99.0\%$. Therefore, we have strong evidence that the hybrid approach outperforms non-hybrid approaches. Perhaps a more important question is how much better the hybrid approach is than non-hybrid approaches: the median of this difference (on the logit scale) is 1.39, with a 95% CI [0.240, 2.69]. Here it is important to note that the CI spans both small effects, as well as large positive ones. This indicates that although we have strong evidence that the hybrid approach is better than non-hybrid approaches, we have a large degree of uncertainty about how much better it is. Similar calculations can be made for other contrasts in Figure 7.C, but we omit them for brevity.

Lastly, with the fitted model we can again inspect the posterior predictive distributions for different levels of the factor and compare these distributions with the observed data. This comparison has been made in Figure 7.A. Again, we can see that all the observed data is plausible under the fitted model (all the points fall within the 95% prediction intervals, not shown here to avoid clutter). However, in the ERD condition the model predicts a substantial probability of below-chance accuracies, similar to results of the Model 2. Again the problem could be tackled by constricting the model to above-chance accuracies.

V. Discussion

A. What have we gained from rejecting NHST?

In this paper we have proposed an alternative to NHST for statistical validation of BCI results: Bayesian estimation with the hierarchical generalized linear model. While we have motivated the use of these methods on theoretical considerations from statistics and empirical findings from other disciplines, we can now directly compare hierarchical Bayesian estimation with NHST on analyses of real BCI results.

Performance of a single BCI in a group of subjects (Model 1): In the dataset of Power et al. [32] NHST analysis can proceed at two levels: single-subject level and group level. At the single-subject level, we can test for each subject if the obtained number of correct trials is above chance level using the binomial test. In the given dataset binomial tests would reject the null hypothesis of chance level performance for all of the subjects. However, implicit in the tests is the assumption that the subject-wise number of trials was fixed before the experiment, which may often not be the case in practice (e.g. when artifactual trials are rejected).

Another option at this point is to apply one of the multiple comparison corrections (e.g. Bonferroni correction) to the family of subject-wise tests, in order to ensure that the type I error rate is preserved at the level α. In the given dataset, after Bonferroni correction one subject-wise test would not be considered significant anymore. It is worth noting that the multiple comparison correction at the subject-level has an undesirable property: assuming that the subject-wise accuracies

are samples from a population (i.e. they have a fixed mean), all the subject-wise tests will be non-significant if a large enough sample is used. In this case, using a large sample is detrimental to inference, counter to intuition and desired behavior of the procedure.

At the group level, we would usually first summarize subject-wise data by the sample accuracy, and assume that these sample accuracies are drawn from a normal group-level distribution; then we can use a right-tail t-test to test the null hypothesis that the group-level mean is equal or smaller than chance level. In the given dataset the p-value for this null hypothesis is smaller than 0.001, and thus we can reject the null hypothesis of chance level operation.

Beyond previously outlined issues with NHST, there are two additional issues with this particular procedure. First, assuming that sample accuracies are normally distributed is not appropriate – sample accuracies are bounded between 0 and 1, whereas the normal distribution is unbounded. This modeling error will be more pronounced for high group-level accuracies, where the data has a larger negative skew – as a consequence, the group-level mean will be underestimated in this case. Second, by summarizing the subject-wise data with sample accuracies, information is lost because we have ignored the hierarchical nature of the experiment (i.e. that the trials are nested within subjects). In effect, all the variance in the data is assigned to between-subject variance, instead of decomposing it into a within-subject and between-subject component; therefore, the between-subject variance is going to be overestimated.

Let us now consider how does the hierarchical Bayesian estimation approach deal with the same dataset. Again, we are interested in both subject-level and group-level analysis; however, due to the hierarchical nature of Model 1, the inference is performed simultaneously at both levels. Analogous to subject-wise p-values, we can obtain posterior probabilities that the subject-wise accuracies are above chance level. In the given dataset for the subject with the lowest accuracy this probability is 99.93%, indicating high certainty that all of the subjects were performing above chance level. However, using the estimation approach, we can go beyond p-values by giving Bayesian CIs for each subject's accuracy, thus describing our uncertainty of individual estimates, due to the finite number of trials per subject (see Figure 5.A). Importantly, these posterior probabilities and confidence intervals are not conditional on sampling intentions and have a straightforward interpretation, unlike p-values and frequentist CIs. Moreover, the posterior probabilities do not need to be corrected for multiple comparisons since (i) the principal aim of Bayesian inference is coherence, rather than control of type I errors, and (ii) we have used a hierarchical model which shrinks individual accuracy estimates towards group-level accuracy, thus regularizing the inference [45].

At the group level, we can again provide the posterior probability that the the group-level mean is above chance level, and this probability in the given dataset is $\sim$100%. However, as we have pointed out earlier, a reasonably motivated BCI approach will rarely work at exactly chance level in a population of subjects, and thus the posterior probability of the group mean being over chance level is of limited value. Again, by using the estimation approach we are able to give more complete insight: we can provide the full posterior distribution over the group mean and inter-subject variance, and we can further summarize the posterior using point and interval estimates. For example, in the given dataset we can summarize the posterior by stating that the group-level mean accuracy is between 72.2% and 82.2% with 95% probability. Depending on the analyst's practical or research goals and peers' judgment, this estimate may or may not be sufficiently precise. In the latter case the Bayesian framework allows us to simply collect more data and update the posterior again using the Bayes' rule, still obtaining valid probabilities. In contrast, p-values and frequentist CIs would be invalidated by such additional data collection.

Additionally, in the proposed framework we can also predict the future data. For example, we might be interested what is the predicted accuracy for a new subject given the data we have observed in the experiment. We can obtain this information from the posterior predictive distribution over future data. In the given dataset the predicted accuracy for a new subject is between 59.6% and 89.1% with 95% probability. While the posterior estimates of parameters such as mean and variance can be made more precise by collecting more data, the predicted accuracy estimate will not necessarily become narrower with more data since it depends on inter-subject variability inherent to the BCI that is being tested. In the case that the prediction interval is too wide for practical purposes, the BCI approach itself should be modified to reduce the inter-subject variability in performance.

Whereas modeling assumptions are rarely verified when applying NHST in practice, in the proposed framework of Bayesian estimation we can use the posterior predictive check to assess if the assumptions of the model are justified. In the given dataset we can inspect the posterior predictive distribution of accuracy, and verify that the observed data does not deviate systematically from it.

Association between a subject-specific variable and BCI performance (Model 2): In the dataset of Blankertz et al. [34] NHST can again proceed at both single-subject and group level, but we will focus only on the group level, since the effect of a subject-specific covariate on accuracy can only be observed at this level. A typical NHST analysis for this experimental design would involve using linear regression to associate the covariate with accuracy and performing a t-test to determine if the slope of the association is significantly different than zero. In the given dataset the p-value obtained from the t-test is smaller than 0.001 and we can reject the null hypothesis that the slope is zero.

Apart from the aforementioned problems of disregarding the hierarchical nature of data and inappropriately assuming normally distributed data, there is an additional issue with assuming a linear dependency between a covariate and accuracy. The reason is again the fact that accuracy is bounded between 0 and 1 – this is opposed to the assumed linear relationship between the covariate and accuracy, which can predict accuracies smaller than 0 and larger than 1, even for values of the covariate present in the dataset (as seen in

Figure 6.A).

In contrast, the use of Bayesian estimation to fit the parameters of a hierarchical generalized linear model does not suffer any of the described problems. Using the appropriate link function in the generalized linear model (logistic link in this case) we obtain properly bounded predictions for all the values of the covariate. Moreover, we can again dispel black-and-white thinking by estimating the effect of the covariate on accuracy, instead of testing whether this effect is exactly zero[2]. In the given dataset we can estimate with 95% probability that the slope is between 0.372 and 0.844 on the log-odds scale, with the posterior median 0.606. In other words, we can expect that a subject with resting alpha power one standard deviation above the average will have an improvement between 1.45 and 2.33 times in the odds of correct decoding. Again, depending on practical considerations we can decide if this estimate is precise enough, and if it is not we can collect more data and update the posterior estimate appropriately.

As before, we can perform the posterior predictive check, predicting the expected accuracy for different values of the covariate. In the given dataset we can see that the model predicts a substantial proportion of subjects below chance level for low levels of alpha power (Figure 6.D), which is not observed in the actual data. This is a consequence of not modeling completely all the prior knowledge on the problem – in this concrete case, the model was not informed of the fact that classification accuracy will generally not be bellow the chance level. Hence, here the posterior predictive check reveals a systematic problem with the model which could then be resolved in a subsequent iteration of modeling by appropriately restricting the model. By just applying NHST without concern of the underlying assumptions, a discrepancy such as this one might easily go unnoticed.

Comparison of different BCI approaches in a within-subject design (Model 3): For the dataset of Brunner et al. we will again focus on the group-level analysis. In the NHST framework, a standard way to analyze the within-subject experimental design with discrete factors is to use repeated measures ANOVA. In the given dataset repeated measures ANOVA indicates that the effect of the employed BCI approach significantly reduces unexplained variance and the corresponding p-value is smaller than 0.001. Since the main hypothesis of the study is not that the used BCI approach affects accuracy (this is usually known *a priori*), but that the hybrid approach is better than the ERD-only and SSVEP-only approaches, additional pairwise *post hoc* tests would usually be conducted. Conducting pairwise t-tests (corrected using Bonferroni-Holm procedure) shows that the hybrid approach is significantly better than the ERD approach ($p = 0.0013$), but the difference between the hybrid approach and the SSVEP approach is not significant ($p = 0.457$).

In the framework of hierarchical Bayesian estimation we can readily obtain accuracy estimates both at the subject-level and for different approaches individually, but we will now proceed directly to the comparison of approaches, which will address the main question of the study. First, we can compute the posterior probability that the hybrid approach is better in pairwise comparisons with the ERD and SSVEP approaches: the probability that the hybrid approach is better than the ERD approach and the SSVEP approach is 99.9% and 68.0%, respectively. Moreover, we can also compare the hybrid approach with the non-hybrid approaches (average of the ERD and SSVEP estimates), and we obtain a probability of 99.0% that the hybrid approach is better. Whereas the *post hoc* tests in the NHST analysis suggest there is no improvement in using a hybrid approach over an SSVEP approach, Bayesian estimation suggest that there is a non-negligible probability that the hybrid approach is better.

However, since implementing a new BCI approach can be costly in terms of time, effort, money, and computational resources, it is not usually enough to show that the improvement is statistically significant, the improvement also needs to be practically significant. In other words, we also need to estimate the size of the improvement and indicate the precision of this estimate. Although the Bayesian analysis indicates that the hybrid approach is probably an improvement upon the ERD and SSVEP approaches, the size of this improvement is quite uncertain (see Figure 7.C). This is most apparent in the wide CI of the difference between the hybrid and SSVEP approaches, which spans from large negative effects up to large positive effects, with the posterior median of this difference being 0.319 (logit scale), i.e. the odds of successful decoding being 1.38 times bigger for the hybrid approach. This median improvement in odds would correspond to a relative decrease in error frequency of around 26%, with the SSVEP approach making an error approximately once in 34 trials and the hybrid approach making an error once in 46 trials. Although the difference between the hybrid and the SSVEP approach was deemed non-significant by NHST, and intuitively seemed small on the probability scale (see Figure 7.B), Bayesian estimation with the logistic model shows that the difference might be practically significant, although the data does not allow precise estimates.

Moreover, in the case of estimates with insufficient precision, the Bayesian framework provides simple guidance. A follow-up study could be conducted to collect more data, and the results of the present study could be used as a prior to obtain more precise estimates via the Bayes' rule – in this way knowledge can easily be accumulated across studies.

B. *Possible misgivings about Bayesian estimation*

One aspect of the proposed framework that might bother us, is the seemingly subjective nature of the employed Bayesian inference. One might argue: if the results of the inference depend on the prior, which should reflect subjective belief of the analyst, how can they be presented as scientifically objective? We can address this criticism from several viewpoints.

First, we should acknowledge that every statistical analysis (or scientific inquiry, for that matter) has subjective

[2]Since we usually test covariates which are likely to be related to accuracy based on prior substantive knowledge, testing this hypothesis is not very informative. Even if the slope is exactly zero, the estimation approach will give a narrow estimate around zero with enough data, providing the same conclusion. Alternatively, we can use Bayesian model comparison [46] between a model with the slope parameter and an intercept-only model.

elements [47]. The choice of hypotheses to test, the choice of data to collect, the form of the model to fit, the choice of the significance level to apply, etc., are all subjective choices, although usually based on some substantive knowledge. In this respect, choosing a subjective prior should not be more controversial than choosing a likelihood; therefore, we would not consider the frequentist approach more objective than the Bayesian approach. If objectivity is our concern, we might even prefer the Bayesian approach, where the subjective prior is overtly stated, to the frequentist approach, where the results depend on possibly covert sampling and testing intentions.

Second, there have been attempts to formulate objective priors [48]: these priors are chosen based on objective rules, and the results do not therefore depend on the analyst. However, there are also issues with objective priors: the posterior obtained from an objective prior might not be a proper probability distribution (integrating to unity) and, perhaps more importantly, the analyst still has to choose according to which rule to construct the prior (as multiple have been proposed).

Third, it is possible to choose a middle ground between fully subjective priors and non-informative objective priors, the so-called weakly informative priors. Here we interpret the prior as a way to supply some substantive information (for example, the scale of the data, or the expected magnitude of effects), but not enough to strongly influence our conclusions. In this way we can interpret the prior as a type of a regularization device, rather than expression of subjective belief. This is the approach we have mostly adhered to in the analyses presented in this paper.

Fourth, objectivity of the analyses should be ensured by proper peer review, which should also scrutinize the prior information that was included in the analysis. The chosen prior might seem too strong to a skeptical audience, and in that case might need to be weakened, but the reverse might also be true – the chosen prior could be too weak, relative to the information available from, for example, previous studies. In this case the prior becomes an asset, allowing us to accumulate knowledge across studies.

Finally, if there are multiple defensible priors, we can conduct a sensitivity analysis, observing how the posterior changes as a function of the prior. On the one hand, if different choices of priors lead to essentially same conclusions, we do not need to be overly concerned with the subjectivity of the analysis. On the other hand, if different reasonable priors lead to different conclusions, we might be better off admitting the lack of certainty in our conclusions, rather than stating one conclusion as being objectively preferable. Furthermore, if the data and model code are openly shared online, other researchers can draw their own conclusions based on their priors, and need not take the results of the original analysis at face value.

Another possible criticism of the proposed framework is the singular focus on parameter estimation. As pointed out by Morey et al. in the context of psychology, science needs both hypothesis testing and parameter estimation [49]. Their proposition is to use Bayesian hypothesis testing (also known as Bayesian model selection or comparison), alongside Bayesian parameter estimation. However, Bayesian hypothesis testing is not without critics (even among Bayesian inclined statisticians, e.g. see ref. [50]), mainly because of its strong sensitivity to the priors, which is not such a large concern for Bayesian parameter estimation. Although we agree with Morey et al. that science needs both hypothesis testing and parameter estimation in principle, in practice we consider the estimation approach more useful for the types of studies usually conducted in BCI research.

C. *Present limitations and future work*

One possible concern with the proposed models are violations of the underlying modeling assumptions. At the lowest level of the proposed models we assume that the trials are exchangeable (i.e. conditionally independent, given the subject's accuracy). We can find two possible reasons for this assumption to be violated. First, in BCIs the underlying data being classified has temporal structure and therefore the probability of correctly classifying a trial might be temporally correlated. Second, accuracy is often obtained using k-fold cross-validation. In this case exchangeability is also violated, as we would not judge the test trials to be exchangeable across folds. While some simulation-based studies have shown how cross-validated results violate the assumptions of binomial sampling [7, 51], to the best of our knowledge, a correction for this bias that could be integrated into a parametric model is not known. Although this is an issue worthy of further research, we would like to point out that the matter of violating assumptions is as applicable to the framework we have described, as it is to the usual NHST methods, which are also based on i.i.d. assumptions.

There are also several computational issues which need to be considered when using MCMC to estimate the parameters of the proposed models. First, although MCMC procedures are asymptotically exact, we cannot know with certainty that the chains have converged to their stationary distribution and that the samples we are using for inference are representative of the true posterior distribution. There is a number of heuristic diagnostics which can be used to detect the lack of convergence, but passing these diagnostics does not guarantee that the procedure has converged. Second, MCMC methods can also be computationally intensive, although this has not been a significant issue in the analyses conducted in this paper (MCMC sampling in all three example datasets was done in under a minute on a medium-grade computer). Moreover, the typical signal processing and machine learning pipelines used in BCI research to obtain subject-wise accuracies are orders of magnitude more time-demanding than the statistical analyses proposed here. Third, using MCMC we do not directly obtain the model evidence $p(y)$. Again, this has not been a significant issue in this paper as we have mainly been concerned with the estimation of model parameters, rather than model comparison where the model evidence plays a role. In the case that some of these issues turn out to be problematic in some practical situations, Bayesian inference might still be viable using approximate methods, such as variational Bayes. For example, a variational procedure has been developed by Brodersen et al. for the single group model of classification accuracy [52].

While we have mostly discussed statistical inference and parameter estimation, study planning is another important aspect of applied BCI research. A common practical issue when planning a study is determining an appropriate sample size for the desired experimental design. There are two general approaches in sample size determination, the "performance based" and the "utility based" approach, as termed by Wang and Gelfand [53]. Although the performance based approach also includes goals such as the classical statistical power (i.e. controlling type II error rate), from the "new statistics" perspective a more worthy goal is accuracy in parameter estimation (AIPE) [54]. An example of an AIPE goal would be to ensure a narrow confidence interval around the true value of the estimated parameter. The utility based approach is a more explicit application of decision theory to the sample size planning. In this approach it is necessary to define a utility function which expresses our valuation of the different outcomes of the study. Whichever way we stated the utility function, the solution to the optimal sample size is then obtained by maximizing the expected utility [55]. Whether we use the performance based or utility based approach for sample size planning, a general method to obtain the required sample size is to use Monte Carlo simulation, although this can be computationally demanding. Since inadequate sample sizes were identified as one of the reasons for the aforementioned "statistical crisis", particularly in neuroscience, we believe that sample size planning should be given careful thought in future work and replace the usual "rule of thumb" sample sizes.

VI. Conclusion

With the increasing applicability of BCIs to medical, research, and commercial domains, it is in our view the right time to give some serious thought to the statistical procedures used to make claims about the effectiveness of BCIs. Since BCI research is a relatively young discipline, taking the right methodological precautions now might go a long way in avoiding an embarrassing and costly reproducibility crisis further along the road, similar to the one that the related fields of psychology and neuroscience are experiencing now.

In this paper we have reviewed some of the problems of the usual NHST approach to the validation of BCIs and proposed an alternative framework. The proposed framework differs from "business as usual" in four distinct ways, listed here from most to least important, per our opinion: instead of hypothesis testing we conduct estimation of model parameters, instead of non-hierarchical we use hierarchical models, instead of frequentist we use Bayesian inference, and instead of a linear model of BCI performance we use a generalized linear model. The estimation approach dispels the black-and-white thinking induced by the NHST, hierarchical models allow us to flexibly fit data from complex experimental designs, Bayesian inference provides a principled method of reasoning about uncertainty in parameter estimates, and the generalized linear model allows us to analyze non-normal performance. Although the proposed framework is not itself a novelty, we extend it to typical experimental designs used in BCI research, demonstrate its effectiveness in three published datasets, and provide the accompanying code and data. In this way we believe we have reduced the gap between advances in statistical methods and BCI research practice.

Even though the proposed framework is in our opinion a step in the right direction, we also acknowledge that alternative approaches, such as frequentist estimation methods or Bayesian hypothesis testing, have their own merits. Whatever the "right approach" ultimately might be, BCI research practice will be improved by a more thorough look at the employed statistical methods and their wider discussion.

Acknowledgments

Authors would like to thank Andreea Sburlea, Martina Melinščak, and Carlos Escolano for valuable discussions and comments on the text, and to Benjamin Blankertz for providing additional information on one of the analyzed datasets. Authors also acknowledge funding by the European Commission through the FP7 Marie Curie Initial Training Network 289146, NETT: Neural Engineering Transformative Technologies, and the Horizon 2020 project MoreGrasp (H2020-ICT-2014-1 643955).

References

[1] J. P. A. Ioannidis. "Why Most Published Research Findings Are False". In: *PLoS Medicine* 2.8 (2005), e124.

[2] K. S. Button, J. P. A. Ioannidis, C. Mokrysz, B. A. Nosek, J. Flint, E. S. J. Robinson, and M. R. Munafò. "Power failure: why small sample size undermines the reliability of neuroscience". In: *Nature Reviews Neuroscience* 14.5 (2013), pp. 365–376.

[3] Open Science Collaboration. "Estimating the reproducibility of psychological science". In: *Science* 349.6251 (2015), aac4716.

[4] A. Gelman and E. Loken. "The Statistical Crisis in Science". In: *American Scientist* 102.6 (2014), p. 460.

[5] M. Billinger, I. Daly, V. Kaiser, J. Jin, B. Z. Allison, G. R. Müller-Putz, and C. Brunner. "Is It Significant? Guidelines for Reporting BCI Performance". In: *Towards Practical Brain-Computer Interfaces*. Springer Berlin Heidelberg, 2012, pp. 333–354.

[6] G. Müller-Putz, R. Scherer, C. Brunner, R. Leeb, and G. Pfurtscheller. "Better than random? A closer look on BCI results". In: *International Journal of Bioelectromagnetism* 10.1 (2008), pp. 52–55.

[7] E. Combrisson and K. Jerbi. "Exceeding chance level by chance: The caveat of theoretical chance levels in brain signal classification and statistical assessment of decoding accuracy." In: *Journal of neuroscience methods* 250 (2015), pp. 126–136.

[8] J. O. Berger and T. Sellke. "Testing a Point Null Hypothesis: The Irreconcilability of P Values and Evidence". In: *Journal of the American Statistical Association* 82.397 (1987), pp. 112–122.

[9] R. S. Nickerson. "Null hypothesis significance testing: A review of an old and continuing controversy." In: *Psychological Methods* 5.2 (2000), pp. 241–301.

[10] G. Gigerenzer. "Mindless statistics". In: *The Journal of Socio-Economics* 33.5 (2004), pp. 587–606.

[11] E.-J. Wagenmakers. "A practical solution to the pervasive problems of p values". In: *Psychonomic Bulletin & Review* 14.5 (2007), pp. 779–804.

[12] J. K. Kruschke. "Bayesian data analysis". In: *Wiley Interdisciplinary Reviews: Cognitive Science* 1.5 (2010), pp. 658–676.

[13] Z Dienes. "Bayesian Versus Orthodox Statistics: Which Side Are You On?" In: *Perspectives on Psychological Science* 6.3 (2011), pp. 274–290.

[14] R. B. Kline. *Beyond significance testing: Statistics reform in the behavioral sciences (2nd ed.)*. Washington: American Psychological Association, 2013.

[15] J. K. Kruschke and T. M. Liddell. "The Bayesian New Statistics: Two Historical Trends Converge". In: *SSRN Electronic Journal* (2015).

[16] E. Olivetti, S. Veeramachaneni, and E. Nowakowska. "Bayesian hypothesis testing for pattern discrimination in brain decoding". In: *Pattern Recognition* 45.6 (2012), pp. 2075–2084.

[17] K. H. Brodersen, C. Mathys, J. R. Chumbley, J. Daunizeau, C. S. Ong, J. M. Buhmann, and K. E. Stephan. "Bayesian Mixed-Effects Inference on Classification Performance in Hierarchical Data Sets". In: *The Journal of Machine Learning Research* 13 (2012), pp. 3133–3176.

[18] J. P. Simmons, L. D. Nelson, and U Simonsohn. "False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant". In: *Psychological Science* 22.11 (2011), pp. 1359–1366.

[19] A Gelman and E Loken. *The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis*. Tech. rep. Department of Statistics, Columbia University, 2013.

[20] R. Hoekstra, S. Finch, H. A. L. Kiers, and A. Johnson. "Probability as certainty: Dichotomous thinking and the misuse ofp values". In: *Psychonomic Bulletin & Review* 13.6 (2006), pp. 1033–1037.

[21] S. Goodman. "A Dirty Dozen: Twelve P-Value Misconceptions". In: *Seminars in Hematology* 45.3 (2008), pp. 135–140.

[22] E. J. Masicampo and D. R. Lalande. "A peculiar prevalence of p values just below .05". In: *The Quarterly Journal of Experimental Psychology* 65.11 (2012), pp. 2271–2279.

[23] C. J. Ferguson and M. T. Brannick. "Publication bias in psychological science: Prevalence, methods for identifying and controlling, and implications for the use of meta-analyses." In: *Psychological Methods* 17.1 (2012), pp. 120–128.

[24] J. P. Ioannidis, M. R. Munafò, P. Fusar-Poli, B. A. Nosek, and S. P. David. "Publication and other reporting biases in cognitive sciences: detection, prevalence, and prevention". In: *Trends in Cognitive Sciences* 18.5 (2014), pp. 235–241.

[25] G Cumming. "The New Statistics: Why and How". In: *Psychological Science* 25.1 (2014), pp. 7–29.

[26] R. Hoekstra, R. D. Morey, J. N. Rouder, and E.-J. Wagenmakers. "Robust misinterpretation of confidence intervals". In: *Psychonomic Bulletin & Review* 21.5 (2014), pp. 1157–1164.

[27] R. D. Morey, R. Hoekstra, J. N. Rouder, M. D. Lee, and E.-J. Wagenmakers. "The fallacy of placing confidence in confidence intervals". In: *Psychonomic Bulletin & Review* 23.1 (2016), pp. 103–123.

[28] D. Lunn, C. Jackson, N. Best, A. Thomas, and D. Spiegelhalter. *The BUGS book: A practical introduction to Bayesian analysis*. Boca Raton, FL: CRC Press, 2013.

[29] P. C. Lambert, A. J. Sutton, P. R. Burton, K. R. Abrams, and D. R. Jones. "How vague is vague? A simulation study of the impact of the use of vague prior distributions in MCMC using WinBUGS". In: *Statistics in Medicine* 24.15 (2005), pp. 2401–2428.

[30] A Gelman. "Prior distributions for variance parameters in hierarchical models(Comment on Article by Browne and Draper)". In: *Bayesian Analysis* 1 (2006), pp. 515–534.

[31] S. Van Dongen. "Prior specification in Bayesian statistics: Three cautionary tales". In: *Journal of Theoretical Biology* 242.1 (2006), pp. 90–100.

[32] S. D. Power, T. H. Falk, and T. Chau. "Classification of prefrontal activity due to mental arithmetic and music imagery using hidden Markov models and frequency domain near-infrared spectroscopy". In: *Journal of Neural Engineering* 7.2 (2010), p. 026002.

[33] A. Gelman, A. Jakulin, M. G. Pittau, and Y.-S. Su. "A weakly informative default prior distribution for logistic and other regression models". In: *The Annals of Applied Statistics* 2.4 (2008), pp. 1360–1383.

[34] B. Blankertz, C. Sannelli, S. Halder, E. M. Hammer, A. Kübler, K.-R. Müller, G. Curio, and T. Dickhaus. "Neurophysiological predictor of SMR-based BCI performance". In: *NeuroImage* 51.4 (2010), pp. 1303–1309.

[35] C Brunner, B. Z. Allison, C Altstätter, and C Neuper. "A comparison of three brain–computer interfaces based on event-related desynchronization, steady state visual evoked potentials, or a hybrid approach using both signals". In: *Journal of Neural Engineering* 8.2 (2011), p. 025010.

[36] Y Lee and J. Nelder. "Hierarchical generalized linear models". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 58.4 (1996), pp. 619–678.

[37] K. H. Brodersen, C. S. Ong, K. E. Stephan, and J. M. Buhmann. "The Balanced Accuracy and Its Posterior Distribution". In: *2010 20th International Conference on Pattern Recognition*. IEEE, 2010, pp. 3121–3124.

[38] H. Carrillo, K. H. Brodersen, and J. A. Castellanos. "Probabilistic performance evaluation for multiclass classification using the posterior balanced accuracy". In: *ROBOT2013: First Iberian Robotics Conference*. Vol. 252. Springer International Publishing, 2014, pp. 347–361.

[39] M. D. Lee and E.-J. Wagenmakers. *Bayesian Cognitive Modeling*. Cambridge: Cambridge University Press, 2013.

[40] I. Ntzoufras. *Bayesian Modeling Using WinBUGS*. Hoboken, NJ, USA: John Wiley & Sons, Inc., 2009.

[41] W. R. Gilks. "Markov Chain Monte Carlo". In: *Encyclopedia of Biostatistics*. Chichester, UK: John Wiley & Sons, Ltd, 2005.

[42] D. J. Lunn, A. Thomas, N. Best, and D. Spiegelhalter. "WinBUGS - A Bayesian modelling framework: Concepts, structure, and extensibility". In: *Statistics and Computing* 10.4 (2000), pp. 325–337.

[43] A. Gelman and D. B. Rubin. "Inference from Iterative Simulation Using Multiple Sequences". In: *Statistical Science* 7.4 (1992), pp. 457–472.

[44] S. P. Brooks and A. Gelman. "General Methods for Monitoring Convergence of Iterative Simulations". In: *Journal of Computational and Graphical Statistics* 7.4 (1998), pp. 434–455.

[45] A. Gelman, J. Hill, and M. Yajima. "Why We (Usually) Don't Have to Worry About Multiple Comparisons". In: *Journal of Research on Educational Effectiveness* 5.2 (2012), pp. 189–211.

[46] J. Vandekerckhove, D. Matzke, and E.-J. Wagenmakers. "Model Comparison and the Principle of Parsimony". In: *The Oxford Handbook of Computational and Mathematical Psychology*. Ed. by J. R. Busemeyer, Z. Wang, J. T. Townsend, and A. Eidels. Oxford University Press, 2015, pp. 300–317.

[47] J. O. Berger and D. A. B. Berry. "Statistical Analysis and the Illusion of Objectivity". In: *American Scientist* 76.2 (1988), pp. 159–165.

[48] R. E. Kass and L. Wasserman. "The Selection of Prior Distributions by Formal Rules". In: *Journal of the American Statistical Association* 91.435 (1996), pp. 1343–1370.

[49] R. D. Morey, J. N. Rouder, J. Verhagen, and E.-J. Wagenmakers. "Why Hypothesis Tests Are Essential for Psychological Science: A Comment on Cumming (2014)". In: *Psychological Science* 25.6 (2014), pp. 1289–1290.

[50] A. Gelman and C. R. Shalizi. "Philosophy and the practice of Bayesian statistics". In: *British Journal of Mathematical and Statistical Psychology* 66.1 (2013), pp. 8–38.

[51] Q. Noirhomme, D. Lesenfants, F. Gomez, A. Soddu, J. Schrouff, G. Garraux, A. Luxen, C. Phillips, and S. Laureys. "Biased binomial assessment of cross-validated estimation of classification accuracies illustrated in diagnosis predictions". In: *NeuroImage: Clinical* 4 (2014), pp. 687–694.

[52] K. H. Brodersen, J. Daunizeau, C. Mathys, J. R. Chumbley, J. M. Buhmann, and K. E. Stephan. "Variational Bayesian mixed-effects inference for classification studies". In: *NeuroImage* 76.6 (2013), pp. 345–361.
[53] F. Wang and A. E. Gelfand. "A simulation-based approach to Bayesian sample size determination for performance under a given model and for separating models". In: *Statistical Science* 17.2 (2002), pp. 193–208.
[54] S. E. Maxwell, K. Kelley, and J. R. Rausch. "Sample Size Planning for Statistical Power and Accuracy in Parameter Estimation". In: *Annual Review of Psychology* 59.1 (2008), pp. 537–563.
[55] D. V. Lindley. "The choice of sample size". In: *Journal of the Royal Statistical Society: Series D (The Statistician)* 46.2 (1997), pp. 129–138.
[56] M Plummer. "JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling". In: *Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003)*. Vienna: Technische Universität Wien, 2003.
[57] A. Patil, D. Huard, and C. Fonnesbeck. "PyMC : Bayesian Stochastic Modelling in Python". In: *Journal of Statistical Software* 35.4 (2010), pp. 1–81.
[58] B Carpenter, A Gelman, and M Hoffman. "Stan: a probabilistic programming language". In: *Journal of Statistical Software* In press. (2015).

APPENDIX A
BASICS OF NHST AND BAYESIAN INFERENCE: THE SIMPLE ILLUSTRATION EXTENDED

Using the simple example outlined in subsection II-C we now provide the details of the NHST approach and the Bayesian estimation approach.

With the collected dataset and the defined model we first define a null hypothesis, e.g. $H_0 : \mu = 0$, which we then try to falsify. To do so, we must further define the measure of compatibility of the data with the null hypothesis, i.e. a test statistic $T(d)$, which is a function of the data. In the given example of normally distributed data with unknown mean and variance, a commonly used statistic is the t-statistic. Next, to determine if the observed data is improbable under the null hypothesis, it is necessary to obtain the null distribution $p(T(d^{\text{rep}})|H_0)$, i.e. the distribution of the test statistic T under repeated sampling, conditional on the null hypothesis being true. Here it is important to notice that the null distribution is conditional not only on H_0, but implicitly also on the sampling and testing intentions (e.g. whether N is predetermined). In the example the null distribution would be the Student's t-distribution with $N-1$ degrees of freedom, assuming that the sample size was fixed at N prior to the experiment. Finally, the discrepancy between the observed test statistic $T(d^*)$ and the null hypothesis is measured by the p-value, which is defined as the tail area of the null distribution:

$$p = P(T(d^{\text{rep}}) \geq T(d^*)|H_0). \tag{9}$$

Intuitively, we can interpret the p-value as the probability of obtaining data that is as extreme as, or more extreme, than the observed data, assuming the null hypothesis is true. As noted before, in the given dataset the value of the t-statistic is 1.09 and the p-value is 0.29; hence, we would not reject H_0 at the usual $\alpha = 0.05$ significance level.

Let us now compare NHST with Bayesian estimation. As stated, Bayesian estimation also starts with formulating a model, which is often represented as a directed acyclic graph (DAG). The model is formalized as a likelihood function $p(d|\theta)$, where θ represent all the parameters of the model. In the given example the likelihood function is

$$\begin{aligned} p(d|\theta) &= p(y_1, \ldots, y_N|\mu, \sigma) \\ &= \prod_{i=1}^{N} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-(y_i-\mu)^2}{2\sigma^2}\right), \end{aligned}$$

and the corresponding graphical model is shown in Figure 1.B.

However, unlike NHST, we additionally need to define the prior distribution $p(\theta)$ which formalizes information about the parameters of the model that is available before observing the data. Let us now additionally assume we *a priori* know that the mean μ is unlikely to be larger than 9 in magnitude, and the standard deviation σ cannot be larger than 10 – in this case we might use the following independent priors:

$$\begin{aligned} p(\theta) &= p(\mu, \sigma) = p(\mu)p(\sigma), \\ p(\mu) &= \text{Normal}(\mu; M_\mu, S_\mu^2), \\ p(\sigma) &= \text{Uniform}(\sigma; L_\sigma, U_\sigma), \\ M_\mu &= 0, S_\mu = 3, L_\sigma = 0, U_\sigma = 10, \end{aligned}$$

where M_μ, S_μ, L_σ, and U_σ are the hyper-parameters. The marginal prior for the population mean $p(\mu)$ is shown in Figure 1.D.

With the likelihood and the priors specified, we can proceed to the estimation of parameters conditional on the observed data. In contrast with NHST, the goal of statistical inference is now to answer questions such as "what are the plausible values of the model parameters θ given the observed data d?" In the Bayesian framework this question is answered by the posterior distribution $p(\theta|d)$. To obtain the posterior we use the Bayes' rule:

$$p(\theta|d) = \frac{p(d|\theta)p(\theta)}{p(d)} \propto p(d|\theta)p(\theta). \tag{10}$$

Since we will be concerned with the estimation of model parameters, rather than comparison of different models, the model evidence $p(d)$ (i.e. marginal likelihood) in the denominator of Bayes' rule (which does not depend on parameters θ) will not play a role, and can be considered just as a proportionality constant.

By inspecting the properties of the posterior distribution we can now interpret the results of the inference. For the example dataset the inferred joint posterior $p(\mu, \sigma|d)$ is shown in Figure 1.C. Since the posterior will generally be high dimensional and include parameters which may not be of interest (i.e. nuisance parameters), we will often want to obtain low dimensional probability distributions over particular parameters (i.e. marginal distributions). We can obtain the marginal posterior distribution for the parameter of interest θ_i as:

$$p(\theta_i|d) = \int p(\theta_i, \theta_{\setminus i}|d)\mathrm{d}\theta_{\setminus i}, \tag{11}$$

where $\theta_{\setminus i}$ is the set of all the parameters except θ_i. With the obtained marginals we can provide numerical summaries of

the inference, such as the expectation, median, mode, variance (standard deviation), or the 95% CI, as well as graphical summaries. For the example dataset, the marginal distribution $p(\mu|d)$ is shown in Figure 1.D, since the population mean is of main interest. We can also provide the numerical summaries of the posterior marginal: $\mathrm{Mdn} = 0.951$, 95% CI: [-1.12, 2.97][3].

Moreover, we can answer questions such as "what is the probability that μ is positive?" The answer is simply obtained by integrating $p(\mu|d)$ for the positive values of μ; in the given example $P(\mu > 0|d) = 83.0\%$. Comparing with NHST, which simply indicates that μ is not significantly different than 0, in Bayesian estimation we obtain richer information: the mean μ is positive with a high probability, but we are uncertain of its magnitude due to the small sample size (indicated by the large 95% CI).

Since the computation of the marginals and the computation of numerical summaries (e.g. expectation) involves integrals that are most often not analytically tractable, we resort to numerical approximations of the integrals using Monte Carlo (MC) integration. To perform the MC integration we need a sample from the probability distribution over which the integral is taken; however, the posterior distribution $p(\theta|d)$ (or equivalently $p(d|\theta)p(\theta)$) is usually too complex to be directly sampled from. Again, we can resort to a numerical solution – Markov chain Monte Carlo (MCMC) simulation – which provides the random sample from the posterior. While the computational part of the Bayesian estimation is more complex than NHST, there are multiple software packages that take a model specification in a formal language as input, and provide the user with an MCMC sample as output, removing the need to implement custom MCMC algorithms for a wide class of models [42, 56–58]. The details of the MCMC procedure we employed are given in the subsection III-E ("Computational details of the inference procedure").

With an MCMC sample $\{\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(T)}\}$ we can easily perform marginalization and computation of numerical summaries of the posterior. The sample of a marginal distribution $p(\theta_i|d)$ is obtained as $\{\theta_i^{(1)}, \theta_i^{(2)}, \ldots, \theta_i^{(T)}\}$, where nuisance parameters are simply ignored. Expectation of a function of a parameter θ_i can then be obtained using MC integration:

$$E[g(\theta_i)|d] \approx \frac{1}{T}\sum_{t=1}^{T} g(\theta_i^{(t)}), \tag{12}$$

where setting $g(\cdot)$ to identity yields the ordinary mean. To answer questions about the amount of probability mass within an interval $[l, u]$ we can also use MC approximation:

$$P(l < \theta_i < u|d) \approx \frac{1}{T}\sum_{t=1}^{T} \mathbb{I}[l < \theta_i^{(t)} < u], \tag{13}$$

where $\mathbb{I}[\cdot]$ is the indicator function, which gives 1 when its argument is true, and 0 otherwise. Similar calculations can be made for other types of inequalities. The calculation of the 95% CI and the median (as well as other percentiles) can be obtained by sorting the MCMC sample and taking the parameter values corresponding to appropriate ranks.

When doing Bayesian estimation, we may often be interested not only in the parameter estimates, but also in predicting the future data. Once the posterior distribution of model parameters has been inferred using the Bayes' rule, we can predict future data $\widetilde{d}$ using the posterior predictive distribution:

$$p(\widetilde{d}|d) = \int p(\widetilde{d}|\theta)p(\theta|d)\mathrm{d}\theta. \tag{14}$$

Here we take the top-down approach, with $p(\widetilde{d}|\theta)$ modeling the dependency of future data on the top-level parameters. In the example dataset we might be interested in the posterior predictive distribution of a new sample $\widetilde{y}$, which can be modeled in the same way as observed data: $\widetilde{y} \sim \mathrm{Normal}(\mu, \sigma)$. The posterior predictive distribution $p(\widetilde{y}|d)$ is shown in Figure 1.A. The computation of the posterior predictive distribution can again be achieved using MCMC simulation, and the summaries are obtained in an analogous way. As a check of the model fit, we can conduct a posterior predictive check, i.e. we can check (e.g. through graphical summaries such as Figure 1.A) if the posterior predictive distribution predicts data that is similar to the one we have observed. If we see systematic differences between the observed data and the predicted data, we might want to revisit the modeling assumptions and do another iteration of modeling and analysis.

[3]Here and elsewhere in text we use the equi-tailed 95% CI, for which 2.5% of probability mass is both below and above it. An alternative choice is to use the highest posterior density (HPD) 95% CI, which is the shortest CI that contains the specified probability mass.