Journal of Machine Learning Research 13 (2012) 3133-3176 Submitted 1/12; Revised 5/12; Published 10/12

Bayesian Mixed-Effects Inference on Classification Performance in Hierarchical Data Sets

Kay H. Brodersen[*] BRODERSEN@BIOMED.EE.ETHZ.CH
Christoph Mathys[†] CHMATHYS@ETHZ.CH
Translational Neuromodeling Unit (TNU)
Institute for Biomedical Engineering
ETH Zurich & University of Zurich
Wilfriedstrasse 6, 8032 Zurich, Switzerland

Justin R. Chumbley JUSTIN.CHUMBLEY@ECON.UZH.CH
Laboratory for Social and Neural Systems Research (SNS)
Department of Economics
University of Zurich
Bluemlisalpstrasse 10, 8006 Zurich, Switzerland

Jean Daunizeau[‡] J.DAUNIZEAU@FIL.ION.UCL.AC.UK
Translational Neuromodeling Unit (TNU)
Institute for Biomedical Engineering
ETH Zurich & University of Zurich
Wilfriedstrasse 6, 8032 Zurich, Switzerland

Cheng Soon Ong CHENGSOON.ONG@UNIMELB.EDU.AU
Joachim M. Buhmann JBUHMANN@INF.ETHZ.CH
Machine Learning Laboratory
Department of Computer Science
ETH Zurich
Universitaetstrasse 6, 8092 Zurich, Switzerland

Klaas E. Stephan[§] STEPHAN@BIOMED.EE.ETHZ.CH
Translational Neuromodeling Unit (TNU)
Institute for Biomedical Engineering
ETH Zurich & University of Zurich
Wilfriedstrasse 6, 8032 Zurich, Switzerland

∗. Additional affiliations: Machine Learning Laboratory, Department of Computer Science, ETH Zurich, Universitaetstrasse 6, 8092 Zurich, Switzerland; and Laboratory for Social and Neural Systems Research (SNS), Department of Economics, University of Zurich, Bluemlisalpstrasse 10, 8006 Zurich, Switzerland.

†. Additional affiliation: Laboratory for Social and Neural Systems Research (SNS), Department of Economics, University of Zurich, Bluemlisalpstrasse 10, 8006 Zurich, Switzerland.

‡. Additional affiliation: Laboratory for Social and Neural Systems Research (SNS), Department of Economics, University of Zurich, Bluemlisalpstrasse 10, 8006 Zurich, Switzerland.

§. Additional affiliations: Laboratory for Social and Neural Systems Research (SNS), Department of Economics, University of Zurich, Bluemlisalpstrasse 10, 8006 Zurich, Switzerland; Wellcome Trust Centre for Neuroimaging, University College London, 12 Queen Square, London, WC1N 3BG, United Kingdom.

Abstract

Classification algorithms are frequently used on data with a natural hierarchical structure. For instance, classifiers are often trained and tested on trial-wise measurements, separately for each subject within a group. One important question is how classification outcomes observed in individual subjects can be generalized to the population from which the group was sampled. To address this question, this paper introduces novel statistical models that are guided by three desiderata. First, all models explicitly respect the hierarchical nature of the data, that is, they are mixed-effects models that simultaneously account for within-subjects (fixed-effects) and across-subjects (random-effects) variance components. Second, maximum-likelihood estimation is replaced by full Bayesian inference in order to enable natural regularization of the estimation problem and to afford conclusions in terms of posterior probability statements. Third, inference on classification accuracy is complemented by inference on the balanced accuracy, which avoids inflated accuracy estimates for imbalanced data sets. We introduce hierarchical models that satisfy these criteria and demonstrate their advantages over conventional methods using MCMC implementations for model inversion and model selection on both synthetic and empirical data. We envisage that our approach will improve the sensitivity and validity of statistical inference in future hierarchical classification studies.

Keywords: beta-binomial, normal-binomial, balanced accuracy, Bayesian inference, group studies

1. Introduction

Classification algorithms are frequently applied to data whose underlying structure is hierarchical. One example is the domain of brain-machine interfaces, where classifiers are used to decode intended actions from trial-wise measurements of neuronal activity in individual subjects (Sitaram et al., 2008). Another example is spam detection, where a classifier is trained separately for each user to predict content classes from high-dimensional document signatures (Cormack, 2008). A third example is the field of neuroimaging, where classifiers are used to relate subject-specific multivariate measures of brain activity to a particular cognitive or perceptual state (Cox and Savoy, 2003). In all of these scenarios, the data have a two-level structure: they comprise n experimental trials (or e-mails, or brain scans) collected from each member of a group of m subjects (or users, or patients). For each subject, the classifier is trained and tested on separate partitions of the trial-wise data. This gives rise to a set of true labels and a set of predicted labels, separately for each subject within the group. The typical question of interest for studies as those described above is: What is the accuracy of the classifier in the general population from which the subjects were sampled? This paper is concerned with such group-level inference on classification accuracy for hierarchically structured data.

In contrast to a large literature on evaluating classification performance in non-hierarchical contexts (see Langford, 2005, for a review), relatively little attention has been devoted to evaluating classification algorithms in hierarchical (i.e., group) settings (Goldstein, 2010; Olivetti et al., 2012). Rather than treating classification outcomes obtained in different subjects as samples from the same distribution, a hierarchical setting requires us to account for the fact that each subject itself has been sampled from a heterogeneous population (Beckmann et al., 2003; Friston et al., 2005). Thus, any approach to evaluating classification performance should account for two independent sources of uncertainty: *fixed-effects* variance (i.e., within-subjects variability) that results from uncertainty about the true classification accuracy in any given subject; and *random-effects variance* (i.e., between-subjects variability) that results from the distribution of true accuracies in the population from which

subjects were drawn. Taking into account both types of uncertainty requires *mixed-effects* inference. This is a central theme of the models discussed in this paper.

There are several commonly used approaches to performance evaluation in hierarchical classification studies.[1] One approach rests on the *pooled sample accuracy*, that is, the number of correctly predicted trials divided by the number of trials in total, across all subjects. Statistical significance can then be assessed using a simple binomial test that is based on the likelihood of obtaining the observed number of correct trials by chance (Langford, 2005). The second commonly used method considers the sample accuracy obtained in each individual subject. The method then (explicitly or implicitly) performs a one-tailed t-test across subjects to assess whether the true accuracy is greater than expected by chance (e.g., Harrison and Tong, 2009; Krajbich et al., 2009; Knops et al., 2009; Schurger et al., 2010).

Both of these commonly used methods suffer from limitations. First of all, they neglect the hierarchical nature of the experiment. The first method represents a fixed-effects approach and disregards variability across subjects. The second method considers random effects, but does not explicitly model the uncertainty associated with subject-specific accuracies. Moreover, both methods use maximum-likelihood estimation which has a tendency to underestimate the variance of the distribution and thus may show suboptimal predictive performance in relation to unseen data (i.e., overfitting; cf. Bishop, 2007, pp. 27–28, 147). Finally, both above methods assess performance in terms of *accuracy*, which may lead to inflated estimates for imbalanced data sets and thus to false conclusions about the significance with which the algorithm has performed better than chance (Chawla et al., 2002; Japkowicz and Stephen, 2002; Akbani et al., 2004; Wood et al., 2007; Zhang and Lee, 2008; Demirci et al., 2008; Brodersen et al., 2010a).

This paper introduces hierarchical models which implement full Bayesian mixed-effects analyses of classification performance that can flexibly deal with different performance measures.[2] These models overcome the limitations of the ritualized approaches described above: First, the models introduced here explicitly represent the hierarchical structure of the data, simultaneously accounting for fixed-effects and random-effects variance components. Second, maximum-likelihood estimation is replaced by a Bayesian framework which enables regularized estimation and model selection with conclusions in terms of posterior probability statements (Gelman et al., 2003). Third, our approach permits inference on both the accuracy and the balanced accuracy, a performance measure that avoids bias when working with imbalanced data sets (Brodersen et al., 2010a).

The paper is organized as follows. Section 2 describes both existing and novel models for inferring the accuracy and balanced accuracy of classification algorithms in the context of hierarchical data sets. Section 3 provides a set of illustrative applications of these models on both synthetic and empirical data. Section 4 reviews the key characteristics of these models and discusses their role in future classification studies.

2. Theory

In a hierarchical setting, a classifier predicts the class label of each of n trials, separately for each subject from a group. Here, we deal with the most common situation, that is, binary classification,

1. This paper focuses on parametric models for performance evaluation. Nonparametric methods are not considered in detail here.
2. All models discussed in this paper have been implemented in MATLAB and can be downloaded from: `http://mloss.org/software/view/407/`.

where class labels are taken from $\{-1,+1\}$, denoted as 'positive' and 'negative' trials. (The extension to a multiclass setting is described in the Discussion.) Typically, the algorithm is trained and tested on separate partitions of the data, resulting in $k \in \{0 \ldots n\}$ correct and $n-k$ incorrect predictions. This procedure is repeated for each subject j within a group of size m.

This setting raises three principal questions. First, what is the classification accuracy at the group level? This is addressed by inference on the mean classification accuracy in the population from which subjects were drawn. Second, what is the classification accuracy in each individual subject? Addressing this question by considering each subject in turn is possible but potentially wasteful, since within-subject inference may benefit from across-subject inference (Efron and Morris, 1971). Third, which of several classification algorithms is best? This question can be answered by estimating how well an algorithm's classification performance generalizes to new data. In particular, we wish to predict how well a trial-wise classifier will perform 'out of sample', that is, on trials from an unseen subject drawn from the same population as the one underlying the presently studied group.

This section considers different models for answering these questions. To keep the paper self-contained, we begin by briefly reviewing the well-known beta-binomial model (Pearson, 1925; Skellam, 1948; Lee and Sabavala, 1987). This introduces most of the concepts we require for subsequently introducing two new models designed to support hierarchical Bayesian inference: the twofold beta-binomial model and the bivariate normal-binomial model.

2.1 Inference on the Accuracy Using the Beta-Binomial Model

A classification algorithm, applied to n trials from a single subject, produces a sequence of classification outcomes $y_1, \ldots, y_n$ which are either correct (1) or incorrect (0). Analyses of these outcomes are typically based on the assumption that, on any given trial independently, the classifier makes a correct prediction with probability $0 \leq \pi \leq 1$, and an incorrect one with probability $1-\pi$. Thus, conditional on π, outcomes are given as a series of independent and identically distributed (i.i.d.) Bernoulli trials,

$$p(y_i \mid \pi) = \mathrm{Bern}(y_i \mid \pi) = \pi^{y_i}(1-\pi)^{1-y_i} \quad \forall i = 1 \ldots n.$$

The i.i.d. assumption derives from the assumption that the observations in the test set are i.i.d. themselves. This assumption is not always made in the context of cross-validation, but is easily justified when the data are only split once, without any cross-validation (cf. Discussion).

2.1.1 THE BETA-BINOMIAL MODEL

The i.i.d. assumption about individual classification outcomes allows us to summarize a sequence of outcomes in terms of the number of correctly predicted trials, k, and the total number of test trials, n. Thus, classification outcomes are converted into a random variable $k = \sum_{i=1}^{n} y_i$ which represents the number of successes over n trials. Since the sum of several Bernoulli variables follows a binomial distribution, the number of successes is given by:

$$p(k \mid \pi, n) = \mathrm{Bin}(k \mid \pi, n) = \binom{n}{k} \pi^k (1-\pi)^{n-k} \tag{1}$$

In this setting, Bayesian inference differs from classical maximum-likelihood estimation in that it assesses the plausibility of all possible values of π before and after observing actual data, rather than

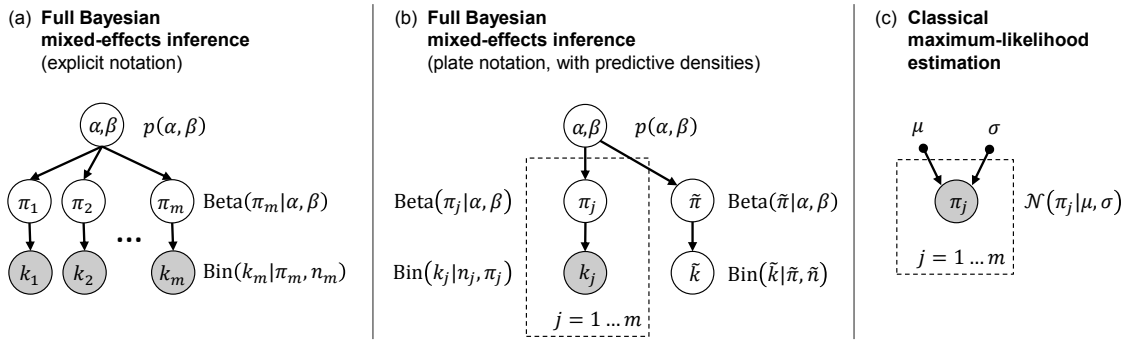

Figure 1: Models for inference on classification accuracies. This illustration shows graphical representations of different models for classical and Bayesian inference on classification accuracies, as discussed in Sections 2.1 and 2.2. Blank circles correspond to latent variables, filled circles represent observed data.

viewing π as a fixed parameter which is to be estimated. (Note that n depends on the experimental design and is not subject to inference.) It is precisely this problem that formed the basis of the first Bayesian analyses published by Bayes and Price (1763) and Laplace (1774). A natural choice for the prior distribution $p(\pi)$ is the Beta distribution,

$$p(\pi \mid \alpha_0, \beta_0) = \text{Beta}(\pi \mid \alpha_0, \beta_0) = \frac{\Gamma(\alpha_0 + \beta_0)}{\Gamma(\alpha_0)\,\Gamma(\beta_0)} \pi^{\alpha_0 - 1} (1 - \pi)^{\beta_0 - 1}, \tag{2}$$

where $\alpha_0, \beta_0 > 0$ are hyperparameters, and the Gamma function $\Gamma(\cdot)$ is required for normalization. Multiplying (1) with (2) gives rise to an overdispersed form of the binomial distribution known as the beta-binomial model (Figure 1; Pearson, 1925; Skellam, 1948; Lee and Sabavala, 1987).

In the absence of prior knowledge about π, we use a *noninformative* prior by setting $\alpha_0 = \beta_0 = 1$, which turns the Beta distribution into a uniform distribution over the $[0, 1]$ interval. The hyperparameters α_0 and β_0 can be interpreted as virtual prior counts of $\alpha_0 - 1$ correct and $\beta_0 - 1$ incorrect trials. Thus, a uniform prior corresponds to zero virtual prior observations of either kind.[3]

Because the Beta prior in (2) is a *conjugate* prior for the binomial likelihood in (1), the posterior distribution $p(\pi \mid k)$ has the same functional form as the prior,

$$p(\pi \mid k) = \text{Beta}(\pi \mid \alpha_n, \beta_n), \tag{3}$$

with updated observation counts $\alpha_n = \alpha_0 + k$ and $\beta_n = \beta_0 + n - k$.

In our context, classification is carried out separately for each subject within a group, hence the available data are k_j out of n_j correct predictions for each subject $j = 1 \ldots m$. One might be tempted to concatenate these data, form group summaries $k = \sum_{j=1}^{m} k_j$ and $n = \sum_{j=1}^{m} n_j$, and proceed to inference on π. However, this would treat π as a fixed effect in the population and disregard how the data were generated. For example, when there are many heterogeneous subjects with few trials each and a single subject with many trials, the data from this single subject would unduly dominate the

3. For a discussion of alternative priors, see Gustafsson et al. (2010).

inference at the group level. Put differently, concatenation falsely assumes zero between-subjects variability.

This limitation is resolved by explicitly modelling both within-subjects (fixed-effects) and between-subjects (random-effects) variance components in a hierarchical model comprising two levels. At the level of individual subjects, for each subject j, the number of correctly classified trials k_j can be modelled as

$$p(k_j \mid \pi_j, n_j) = \mathrm{Bin}(k_j \mid \pi_j, n_j) = \binom{n_j}{k_j} \pi_j^{k_j} (1-\pi_j)^{n_j - k_j}, \tag{4}$$

where n_j is the total number of trials in subject j, and π_j represents the fixed but unknown accuracy that the classification algorithm achieves on that subject. (Note that our notation will suppress n_j unless this introduces ambiguity.) At the group level, the model must account for variability across subjects. This is achieved by modelling subject-wise accuracies as drawn from a population distribution described by a Beta density,

$$p(\pi_j \mid \alpha, \beta) = \mathrm{Beta}(\pi_j \mid \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \pi_j^{\alpha-1} (1-\pi_j)^{\beta-1}, \tag{5}$$

such that α and β characterize the population as a whole. This step is formally identical with the Beta prior placed on the accuracy in (2) which represents uncertainty about π before observing the outcome k. Equation (5) states that uncertainty about any particular subject is best quantified by our knowledge about variability in the population, that is, the distribution of π_j over subjects (which, as described below, can be learnt from the data). Formally, a particular subject's π_j is drawn from a population characterized by α and β: subject-specific accuracies are assumed to be i.i.d., conditional on the population parameters α and β.

To describe our uncertainty about the population parameters, we use a diffuse prior on α and β which ensures that the posterior will be dominated by the data. One option would be to assign uniform densities to both the prior expected accuracy $\alpha/(\alpha+\beta)$ and the prior virtual sample size $\alpha+\beta$, using logistic and logarithmic transformations to put each on a $(-\infty, \infty)$ scale; but this prior would lead to an improper posterior density (Gelman et al., 2003). An alternative is to put a uniform density on the prior expected accuracy $\alpha/(\alpha+\beta)$ and the inverse root of the virtual sample size $(\alpha+\beta)^{-1/2}$ (Gelman et al., 2003). This combination corresponds to the prior

$$\tilde{p}(\alpha, \beta) \propto (\alpha+\beta)^{-5/2} \tag{6}$$

on the natural scale. However, although this prior leads to a proper posterior density, it is improper itself (as indicated by the tilde) and thus prevents computation of the model evidence, that is, the marginal likelihood of the data given the model, which will later become important for model comparison. We resolve this limitation by using a proper (i.e., integrable and normalized) variant,

$$p(\alpha, \beta) = \frac{3}{4} (\alpha+\beta+1)^{-5/2} \tag{7}$$

which represents a special case of the generalization of (6) proposed by Everson and Bradlow (2002). This prior can be rewritten in an unnormalized, reparameterized form as

$$\tilde{p}\left(\ln\left(\frac{\alpha}{\beta}\right), \ln(\alpha+\beta)\right) = \alpha\beta(\alpha+\beta+1)^{-5/2},$$

which will be useful in the context of model inversion (Gelman et al., 2003). Two equivalent graphical representations of this model (using the formalism of Bayesian networks; Jensen and Nielsen, 2007) are shown in Figures 1a and 1b.

2.1.2 MODEL INVERSION

Inverting the beta-binomial model allows us to infer on (i) the posterior population mean accuracy, (ii) the subject-specific posterior accuracies, and (iii) the posterior predictive accuracy. We propose a numerical procedure for model inversion which is described in detail in Appendix A. Below, we restrict ourselves to a brief conceptual summary.

First, to obtain the posterior density over the population parameters α and β we need to evaluate

$$p(\alpha,\beta \mid k_{1:m}) = \frac{p(k_{1:m} \mid \alpha,\beta)\, p(\alpha,\beta)}{\iint p(k_{1:m} \mid \alpha,\beta)\, p(\alpha,\beta)\, \mathrm{d}\alpha\, \mathrm{d}\beta} \tag{8}$$

with $k_{1:m} := (k_1, k_2, \ldots, k_m)$. Under i.i.d. assumptions about subject-specific accuracies π_j we obtain the likelihood function

$$\begin{aligned} p(k_{1:m} \mid \alpha,\beta) &= \prod_{j=1}^{m} \int p(k_j \mid \pi_j)\, p(\pi_j \mid \alpha,\beta)\, \mathrm{d}\pi_j \\ &= \prod_{j=1}^{m} \mathrm{Bb}(k_j \mid \alpha,\beta), \end{aligned} \tag{9}$$

where $\mathrm{Bb}(\cdot)$ denotes the beta-binomial distribution. Since the integral on the right-hand side of (8) cannot be evaluated in closed form, we resort to a Markov chain Monte Carlo (MCMC) procedure. Specifically, we use a Metropolis algorithm (Metropolis and Ulam, 1949; Metropolis et al., 1953) to sample from the variables at the top level of the model and obtain a set $\{(\hat{\alpha}^{(\tau)}, \hat{\beta}^{(\tau)})\}$ for $\tau = 1 \ldots c$. This set allows us to obtain samples from the posterior population mean accuracy,

$$p\left(\frac{\alpha}{\alpha+\beta} \,\middle|\, k_{1:m}\right).$$

We can use these samples in various ways, for example, to obtain a point estimate of the population mean accuracy using the posterior mean,

$$\frac{1}{c}\sum_{\tau=1}^{c} \frac{\hat{\alpha}^{(\tau)}}{\hat{\alpha}^{(\tau)} + \hat{\beta}^{(\tau)}}.$$

We could also numerically evaluate the posterior probability that the mean classification accuracy in the population does not exceed chance,

$$p = \Pr\left(\frac{\alpha}{\alpha+\beta} \leq 0.5 \,\middle|\, k_{1:m}\right)$$

which can be viewed as a Bayesian analogue of a classical p-value. We shall refer to this quantity as the (posterior) *infraliminal probability* of the classifier. It lives on the same $[0,1]$ scale as a classical p-value, but has a much more intuitive (and less error-prone) interpretation: rather than denoting the probability of observing the data (or more extreme data) under the 'null hypothesis' of a chance

classifier (classical p-value), the infraliminal probability represents the (posterior) probability that the classifier operates at or below chance. We will revisit this aspect in the Discussion.

Finally, we could compute the posterior probability that the mean accuracy in one population is greater than in another,

$$p = \Pr\left(\frac{\alpha^{(1)}}{\alpha^{(1)} + \beta^{(1)}} > \frac{\alpha^{(2)}}{(\alpha^{(2)} + \beta^{(2)})} \,\middle|\, k_{1:m^{(1)}}, k_{1:m^{(2)}} \right).$$

The second question of interest concerns the classification accuracies in individual subjects. Specifically, we wish to infer on $p(\pi_j \mid k_{1:m})$ to characterize our posterior uncertainty about the true classification accuracy in subject j. Given a pair of samples $(\alpha^{(\tau)}, \beta^{(\tau)})$, we can obtain samples from subject-specific posteriors simply by drawing from

$$\mathrm{Beta}\left(\pi_j^{(\tau)} \,\middle|\, \alpha^{(\tau)} + k_j,\ \beta^{(\tau)} + n_j - k_j \right).$$

Because samples for α and β are influenced by data $k_1 \ldots k_m$ from the entire group, so are the samples for π_j. In other words, each subject's individual posterior accuracy is informed by what we have learned about the group as a whole, an effect known as *shrinking to the population*. It ensures that each subject's posterior mean lies between its sample accuracy and the group mean. Subjects with fewer trials will exert a smaller effect on the group and shrink more, while subjects with more trials will have a larger influence on the group and shrink less.

The third question of interest is how one classifier compares to another. To address this, we must assess how well the observed performance generalizes across subjects. In this case, we are typically less interested in the average effect in the group but more in the effect that a new subject from the same population would display, as this estimate takes into account both the population mean and the population variance. The expected performance is expressed by the posterior predictive density,

$$p(\tilde{\pi} \mid k_{1:m}),$$

in which $\tilde{\pi}$ denotes the classification accuracy in a new subject drawn from the same population as the existing group of subjects with latent accuracies $\pi_1, \ldots, \pi_m$ (cf. Figure 1b).[4] Samples for this density can easily be obtained using the samples $\alpha^{(\tau)}$ and $\beta^{(\tau)}$ from the posterior population mean.[5]

The computational complexity of a full Bayesian approach can be diminished by resorting to an empirical Bayes approximation (Deely and Lindley, 1981). This approach, however, is not without criticism (Robert, 2007). Here, we will keep our treatment fully Bayesian.

2.2 Inference on the Balanced Accuracy Using the Twofold Beta-Binomial Model

A well-known phenomenon in binary classification is that a training set consisting of different numbers of representatives from either class may result in a classifier that is biased towards the majority class. When applied to a test set that is similarly imbalanced, this classifier yields an optimistic accuracy estimate. In an extreme case, the classifier might assign every single test case to the majority

4. The term 'posterior predictive density' is sometimes exclusively used for densities over variables that are unobserved but are observable in principle. Here, we use the term to refer to the posterior density of any unobserved variable, whether observable in principle (such as $\tilde{k}$) or not (such as $\tilde{\pi}$).

5. If data were indeed obtained from a new subject (represented in terms of $\tilde{k}$ correct predictions in $\tilde{n}$ trials), then $p(\tilde{\pi} \mid k_{1:m}, n_{1:m})$ would be used as a prior to compute the posterior $p(\tilde{\pi} \mid \tilde{k}, \tilde{n}, k_{1:m}, n_{1:m})$.

class, thereby achieving an accuracy equal to the proportion of test cases belonging to the majority class.

In previous literature (Chawla et al., 2002; Japkowicz and Stephen, 2002; Akbani et al., 2004; Wood et al., 2007; Zhang and Lee, 2008; Demirci et al., 2008; Brodersen et al., 2010a), this has motivated, amongst other strategies, the use of a different performance measure: the *balanced accuracy*, defined as the arithmetic mean of sensitivity and specificity, or the average accuracy obtained on either class,

$$\phi = \frac{1}{2}\left(\pi^+ + \pi^-\right). \tag{10}$$

where π^+ and π^- denote classification accuracies on positive and negative trials, respectively. If the classifier performs equally well on either class, this term reduces to the conventional accuracy (i.e., the number of correct predictions divided by the total number of predictions). In contrast, if the conventional accuracy is above chance *only* because the classifier takes advantage of an imbalanced test set, then the balanced accuracy, as appropriate, will drop to chance. We can evaluate the balanced accuracy in a hierarchical setting by extending the beta-binomial model, as described next.

2.2.1 THE TWOFOLD BETA-BINOMIAL MODEL

One way of inferring on the balanced accuracy is to duplicate the beta-binomial model and apply it separately to the two classes (Figure 2a). In other words, we consider the number of correctly predicted positive trials k^+ and the number of correctly predicted negative trials k^-, and express our uncertainty about ϕ (10) before and after observing k^+ and k^-. In a single-subject setting, as in (2), we can place separate noninformative Beta priors on π^+ and π^-,

$$\begin{aligned} p(\pi^+ \mid \alpha_0^+, \beta_0^+) &= \text{Beta}(\pi^+ \mid \alpha_0^+, \beta_0^+), \\ p(\pi^- \mid \alpha_0^-, \beta_0^-) &= \text{Beta}(\pi^- \mid \alpha_0^-, \beta_0^-), \end{aligned} \tag{11}$$

where $\alpha_0^+ = \beta_0^+ = \alpha_0^- = \beta_0^- = 1$. Inference on class-specific accuracies π^+ and π^- could be done in exactly the same way as discussed in the previous section. Here, however, we are primarily interested in the posterior density of the balanced accuracy,

$$p(\phi \mid k^+, k^-) = p\left(\frac{1}{2}(\pi^+ + \pi^-) \;\middle|\; k^+, k^-\right).$$

The balanced accuracy is thus a new random variable defined via two existing random variables from our model, π^+ and π^-. Even in a single-subject setting, a closed form for its posterior distribution is not available, and so we must resort to a numerical approximation (Brodersen et al., 2010a). For this, we first note that the distribution of the sum of the two class-specific accuracies, $s := \pi^+ + \pi^-$, is the convolution of the distributions for π^+ and π^-,

$$p\left(s \mid \alpha_n^+, \beta_n^+, \alpha_n^-, \beta_n^-\right) = \int_0^s p_{\pi^+}(s - z \mid \alpha_n^+, \beta_n^+)\; p_{\pi^-}(z \mid \alpha_n^-, \beta_n^-)\; \mathrm{d}z,$$

where the subscripts of the posterior distributions $p_{\pi^+}(\cdot)$ and $p_{\pi^-}(\cdot)$ serve to remove ambiguity. We can now obtain the posterior distribution of the balanced accuracy by replacing the sum of class-

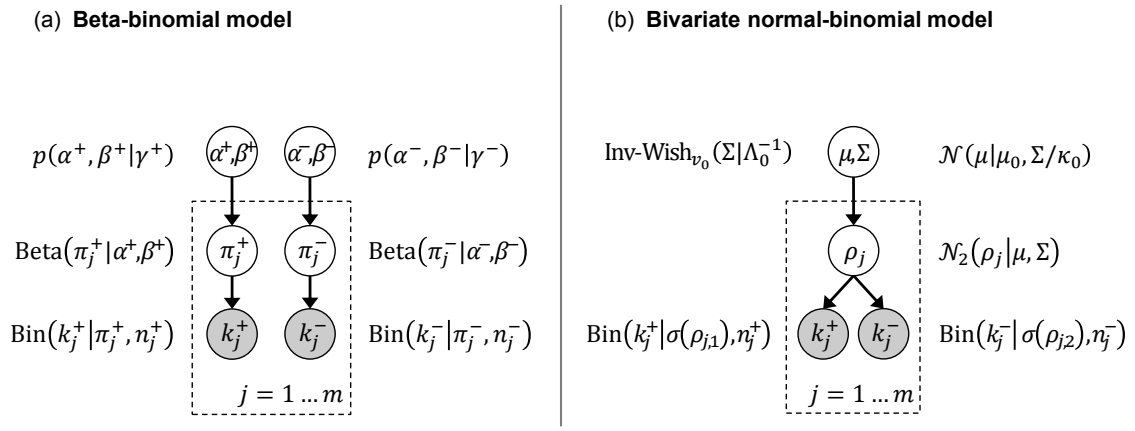

Figure 2: Models for inference on balanced classification accuracies. This figure shows two models for Bayesian mixed-effects inference on the balanced accuracy, as discussed in Sections 2.2 and 2.3. The models are based upon different assumptions and parameterizations and can be compared by Bayesian model comparison.

specific accuracies by their arithmetic mean,

$$\begin{aligned} p(\phi \mid \alpha_n^+, \beta_n^+, \alpha_n^-, \beta_n^-) &= \int_0^{2\phi} p_{\pi^+}(2\phi - z \mid \alpha_n^+, \beta_n^+)\; p_{\pi^-}(z \mid \alpha_n^-, \beta_n^-)\; \mathrm{d}z \\ &= \int_0^{2\phi} \mathrm{Beta}(2\phi - z \mid \alpha_n^+, \beta_n^+)\; \mathrm{Beta}(z \mid \alpha_n^-, \beta_n^-)\; \mathrm{d}z. \end{aligned}$$

This expression can be approximated by a simple one-dimensional grid integration over the $[0, 1]$ interval. In the same way, we can obtain approximations to the posterior mean, the posterior mode, or a posterior probability interval.

In a group setting, one can expand the above model in precisely the same way as for the simpler case of the classification accuracy in Section 2.1. Specifically, we define diffuse priors on the class-specific population parameters α^+ and β^+ as well as α^- and β^-, in analogy to (7). A graphical representation of this model is shown in Figure 2a.

2.2.2 MODEL INVERSION

Given that the twofold beta-binomial model consists of two independent instances of the simple beta-binomial model considered in Section 2.1 (Figure 1b), statistical inference follows the same approach as described previously (see Section 3.3 for an application). For instance, we can obtain the posterior population parameters, $p(\alpha^+, \beta^+ \mid k_{1:m}^+)$ and $p(\alpha^-, \beta^- \mid k_{1:m}^-)$ using the same sampling procedure as summarized in Section 2.1, except that we are now applying the procedure twice. The two sets of samples can then be averaged in a pairwise fashion to obtain samples from the posterior mean balanced accuracy in the population,

$$p\left(\phi \mid k_{1:m}^+, k_{1:m}^-\right),$$

where we have defined

$$\phi := \frac{1}{2}\left(\frac{\alpha^+}{\alpha^+ + \beta^+} + \frac{\alpha^-}{\alpha^- + \beta^-}\right).$$

Similarly, we can average pairs of posterior samples from π_j^+ and π_j^- to obtain samples from the posterior densities of subject-specific balanced accuracies,

$$p\left(\phi_j \mid k_{1:m}^+, k_{1:m}^-\right).$$

Using the same idea, we can obtain samples from the posterior predictive density of the balanced accuracy that can be expected in a new subject from the same population,

$$p\left(\tilde{\phi} \mid k_{1:m}^+, k_{1:m}^-\right).$$

2.3 Inference on the Balanced Accuracy Using the Bivariate Normal-Binomial Model

In the previous section, we saw that the twofold beta-binomial model enables mixed-effects inference on the balanced accuracy. However, it may not always be optimal to treat accuracies on positive and negative trials separately (cf. Leonard, 1972). That is, if π^+ and π^- were related in some way, the model should reflect this. For example, one could imagine a group study in which some subjects exhibit a more favourable signal-to-noise ratio than others, leading to well-separated classes. In this case, an unbiased classifier yields high accuracies on either class in some subjects and lower accuracies in others, inducing a positive correlation between class-specific accuracies. On the other hand, within each subject, any classification algorithm faces a trade-off between performing better on one class at the expense of the other class. Thus, any variability in setting this threshold leads to negatively correlated class-specific accuracies, an argument that is formally related to receiver-operating characteristics. Moreover, if the degree of class imbalance in the data varies between subjects, classifiers might be biased in different ways, again leading to negatively correlated accuracies.

In summary, π^+ and π^- may not always be independent. We therefore turn to an alternative model for mixed-effects inference on the balanced accuracy that embraces potential dependencies between class-specific accuracies (Figure 2b).

2.3.1 THE BIVARIATE NORMAL-BINOMIAL MODEL

The bivariate normal-binomial model no longer assumes that π^+ and π^- are drawn from separate populations. Instead, we use a bivariate population density whose covariance structure defines the form and extent of the dependency between π^+ and π^-.

For this combined prior, we use a bivariate normal density. Because this density has infinite support, we do not define it on the accuracies themselves but on their log odds. In this way, each subject j is associated with a two-dimensional vector of class-specific accuracies,

$$\rho_j = \begin{pmatrix}\rho_j^+ \\ \rho_j^-\end{pmatrix} = \begin{pmatrix}\sigma^{-1}(\pi_j^+) \\ \sigma^{-1}(\pi_j^-)\end{pmatrix} \in \mathbb{R}^2,$$

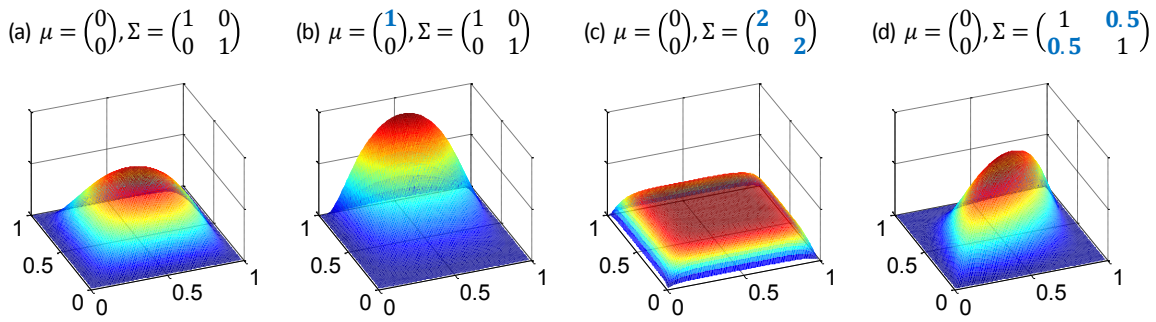

Figure 3: Distributions of class-specific accuracies in the bivariate normal-binomial model. In the bivariate normal-binomial model (Section 2.3), class-specific accuracies are assumed to follow a bivariate logit-normal distribution. This figure illustrates the flexibility of this distribution. Specifically, (a) the standard parameterization is compared to a distribution with (b) an increased accuracy on one class but not the other, (c) an increased population heterogeneity, and (d) a correlation between class-specific accuracies. The x- and y-axis represent the accuracies on positive and negative trials, respectively.

where $\sigma^{-1}(\pi) := \ln\pi - \ln(1-\pi)$ represents the logit (or inverse-logistic) transform. Conversely, class-specific accuracies can be recovered using

$$\pi_j = \begin{pmatrix} \pi_j^+ \\ \pi_j^- \end{pmatrix} = \begin{pmatrix} \sigma(\rho_j^+) \\ \sigma(\rho_j^-) \end{pmatrix} \in [0,1]^2,$$

where $\sigma(\rho) := 1/(1+\exp(-\rho))$ denotes the sigmoid (or logistic) transform. Thus, we can replace the two independent Beta distributions for π^+ and π^- in (11) by a single bivariate Gaussian prior,

$$p(\rho_j \mid \mu, \Sigma) = \mathcal{N}_2(\rho_j \mid \mu, \Sigma), \tag{12}$$

in which $\mu \in \mathbb{R}^2$ represents the population mean and $\Sigma \in \mathbb{R}^{2\times 2}$ encodes the covariance structure between accuracies on positive and negative trials. The resulting density on $\pi \in \mathbb{R}^2$ is a bivariate logit-normal distribution (Figure 3).

In analogy with the prior placed on α and β in Section 2.1, we now specify a prior for the population parameters μ and Σ. Specifically, we seek a diffuse prior that induces a weakly informative bivariate distribution over $[0,1]\times[0,1]$. We begin by considering the family of conjugate priors for (μ, Σ), that is, the bivariate normal-inverse-Wishart distribution,

$$\begin{aligned} &p(\mu, \Sigma \mid \mu_0, \kappa_0, \Lambda_0, \nu_0) \\ &\propto |\Sigma|^{-\left(\frac{\nu_0}{2}+2\right)} \exp\left(-\frac{1}{2}\mathrm{tr}(\Lambda_0\Sigma^{-1}) - \frac{\kappa_0}{2}(\mu-\mu_0)^{\mathrm{T}}\Sigma^{-1}(\mu-\mu_0)\right). \end{aligned}$$

In this distribution, the population hyperparameters Λ_0 and ν_0 specify the scale matrix and the degrees of freedom, while the parameters μ_0 and κ_0 represent the prior mean and the number of prior measurements on the Σ scale, respectively (Gelman et al., 2003). A more convenient representation

can be obtained by factorizing the density into

$$p(\Sigma \mid \Lambda_0, \nu_0) = \text{Inv-Wishart}_{\nu_0}(\Sigma \mid \Lambda_0^{-1}) \quad \text{and}$$
$$p(\mu \mid \Sigma, \mu_0, \kappa_0) = \mathcal{N}_2(\mu \mid \mu_0, \Sigma/\kappa_0).$$

In order to illustrate the flexibility offered by the bivariate normal density on ρ, we derive $p(\pi \mid \mu, \Sigma)$ in closed form (Appendix B) and then compute the bivariate density on a two-dimensional grid (Figure 3).

For the purpose of specifying a prior, we seek hyperparameters μ_0, κ_0, Λ_0, and ν_0 that induce a diffuse bivariate distribution over π. This can be achieved using

$$\mu_0 = (0,0)^{\mathrm{T}},\ \kappa_0 = 1,\ \Lambda_0 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}^{-1},\ \nu_0 = 5.$$

2.3.2 MODEL INVERSION

In contrast to the twofold beta-binomial model discussed in the previous section, the bivariate normal-binomial model makes it difficult to sample from the posterior densities over model parameters using a Metropolis implementation. In order to sample from $p(\mu, \Sigma \mid k^+_{1:m}, k^-_{1:m})$, we would have to evaluate the likelihood $p(k^+_{1:m}, k^-_{1:m} \mid \mu, \Sigma)$; this would require us to integrate out π^+ and π^-, which is difficult.

A simpler strategy is to design a Gibbs sampler (Geman and Geman, 1984) to draw from the joint posterior $p(\rho_{1:m}, \mu, \Sigma \mid k^+_{1:m}, k^-_{1:m})$, from which we can derive samples for the conditional posteriors $p(\rho_{1:m} \mid k^+_{1:m}, k^-_{1:m})$ and $p(\mu, \Sigma \mid k^+_{1:m}, k^-_{1:m})$. In contrast to a Metropolis scheme, Gibbs sampling requires only full conditionals, that is, distributions of one latent variable conditioned on all other variables in the model (Gelfand and Smith, 1990). Whenever a full conditional is not available, we can sample from it using a Metropolis step. Thus, we combine a Gibbs skeleton with interleaved Metropolis steps to sample from the posterior $p(\rho_{1:m}, \mu, \Sigma \mid k^+_{1:m}, k^-_{1:m})$. See Section 3.3 for an application.

First, population parameter estimates can be obtained by sampling from the posterior density $p(\mu, \Sigma \mid k^+_{1:m}, k^-_{1:m})$ using a Metropolis-Hastings approach. Second, subject-specific accuracies are estimated by first sampling from $p(\rho_j \mid k^+_{1:m}, k^-_{1:m})$ and then applying a sigmoid transform to obtain samples from the posterior density over subject-specific balanced accuracies, $p(\phi_j \mid k^+_{1:m}, k^-_{1:m})$. Finally, the predictive density $p(\tilde{\phi} \mid k^+_{1:m}, k^-_{1:m})$ can be obtained using an ancestral-sampling step on the basis of $\mu^{(\tau)}$ and $\Sigma^{(\tau)}$ followed by a sigmoid transform. As before, we use the obtained samples in all three cases to compute approximate posterior probability intervals or Bayesian p-values. A detailed description of this algorithm can be found in Appendix C.

2.4 Bayesian Model Selection

While the twofold beta-binomial model assumes independent class-specific accuracies, the bivariate normal-binomial model relaxes this assumption and allows for correlations between accuracies. This raises two questions. First, given a particular data set, which model is best at explaining observed classification outcomes? And second, can we combine the two models to obtain posterior inferences that integrate out uncertainty about which model is best? Both questions can be answered using the marginal likelihood, or model evidence, that is, the probability of the data given the model,

after integrating out the parameters:

$$p(k^+_{1:m}, k^-_{1:m} \mid M) = \int p(k^+_{1:m}, k^-_{1:m} \mid \theta)\; p(\theta \mid M)\; \mathrm{d}\theta$$

Here, θ serves as a placeholder for all model parameters and $p(\theta \mid M)$ represents its prior distribution under a given model M. Under a flat prior over models, Bayes' theorem indicates that the model with the highest evidence has the highest posterior probability given the data:

$$p(M \mid k^+_{1:m}, k^-_{1:m}) \propto p(k^+_{1:m}, k^-_{1:m} \mid M)$$

In practice, the model evidence is usually replaced by the log model evidence, which is monotonically related but numerically advantageous.

Concerning the first model described in this section, the twofold beta-binomial model M_{bb}, the log model evidence is given by

$$\begin{aligned}
&\ln p(k^+_{1:m}, k^-_{1:m} \mid M_{bb}) \\
&= \ln \int p(k^+_{1:m} \mid \pi^+_{1:m})\, p(\pi^+_{1:m})\; \mathrm{d}\pi^+_{1:m} + \ln \int p(k^-_{1:m} \mid \pi^-_{1:m})\, p(\pi^-_{1:m})\; \mathrm{d}\pi^-_{1:m} && (13) \\
&= \ln \left\langle \prod_{j=1}^{m} p(k^+_j \mid \pi^+_j) \right\rangle_{\pi^+_{1:m}} + \ln \left\langle \prod_{j=1}^{m} p(k^-_j \mid \pi^-_j) \right\rangle_{\pi^-_{1:m}} && (14)
\end{aligned}$$

where we have omitted the conditional dependence on M_{bb} in (13) and (14).[6] The expression can be approximated by

$$\approx \ln \frac{1}{c} \sum_{\tau=1}^{c} \prod_{j=1}^{m} \mathrm{Bin}\left(k^+_j \;\middle|\; \pi^{+(\tau)}_j\right) + \ln \frac{1}{c} \sum_{\tau=1}^{c} \prod_{j=1}^{m} \mathrm{Bin}\left(k^-_j \;\middle|\; \pi^{-(\tau)}_j\right),$$

where $\pi^{+(\tau)}_j$ and $\pi^{-(\tau)}_j$ represent independent samples from the prior distribution over subject-specific accuracies. They can be obtained using ancestral sampling, starting from the prior over α and β, as given in (7).

In the case of the bivariate normal-binomial model M_{nb}, the model evidence no longer sums over model partitions as in (13), and so the approximation is derived differently,

$$\begin{aligned}
&\ln p(k^+_{1:m}, k^-_{1:m} \mid M_{nb}) \\
&= \ln \int p\left(k^+_{1:m}, k^-_{1:m} \mid \rho_{1:m}\right)\, p\left(\rho_{1:m} \mid M_{nb}\right)\, \mathrm{d}\rho_{1:m} && (15) \\
&\approx \ln \frac{1}{c} \sum_{\tau=1}^{c} \prod_{j=1}^{m} \mathrm{Bin}\left(k^+_j \;\middle|\; \sigma\left(\rho^{(\tau,1)}_j\right)\right) \mathrm{Bin}\left(k^-_j \;\middle|\; \sigma\left(\rho^{(\tau,2)}_j\right)\right), && (16)
\end{aligned}$$

for which we provide additional details in Appendix C (24). Having computed the model evidences, one can proceed to Bayesian model selection (BMS) by evaluating the log Bayes factor,

$$\ln BF_{bb,nb} = \ln p(k^+_{1:m}, k^-_{1:m} \mid M_{bb}) - \ln p(k^+_{1:m}, k^-_{1:m} \mid M_{nb}), \qquad (17)$$

6. One could also express the model evidence in terms of an expectation with respect to $p(\alpha, \beta \mid M_{bb})$.

representing the evidence in favour of the beta-binomial over the normal-binomial model. By convention, a log Bayes factor greater than 3 is considered strong evidence in favour of one model over another, whereas a log Bayes factor greater than 5 is referred to as very strong evidence (Kass and Raftery, 1995). The best model can then be used for posterior inferences on the mean accuracy in the population or the predictive accuracy in a new subject from the new population.

The second option to make use of the model evidences of competing models is Bayesian model averaging (Cooper and Herskovits, 1992; Madigan and Raftery, 1994; Madigan et al., 1996). Under this view, we do not commit to a particular model but average the predictions made by all of them, weighted by their respective posteriors. In this way, we obtain a mixture expression for the posterior of the mean accuracy in the population,

$$p\left(\phi \mid k^+_{1:m}, k^-_{1:m}\right) = \sum_M p\left(\phi \mid k^+_{1:m}, k^-_{1:m}, M\right)\ p\left(M \mid k^+_{1:m}, k^-_{1:m}\right).$$

Similarly, we can obtain the posterior predictive distribution of the balanced accuracy in a new subject from the same population,

$$p\left(\tilde{\phi} \mid k^+_{1:m}, k^-_{1:m}\right) = \sum_M p\left(\tilde{\phi} \mid k^+_{1:m}, k^-_{1:m}, M\right)\ p\left(M \mid k^+_{1:m}, k^-_{1:m}\right).$$

The computational complexity of the above stochastic approximations is considerable, and so it can sometimes be useful to resort to a deterministic approximation instead, such as variational Bayes (see Discussion). While we do not consider this approach in detail here, it does provide a helpful perspective on interpreting the model evidence. Specifically, the model evidence can be approximated by a variational lower bound, the negative free-energy $\mathcal{F}$. In the case of the beta-binomial model for instance, this quantity can be written as

$$\mathcal{F} = \left\langle \ln p\left(k_{1:m} \mid \alpha, \beta, \pi_{1:m}\right)\right\rangle_q - \mathrm{KL}\left[q(\alpha, \beta, \pi_{1:m}) \parallel p(\alpha, \beta, \pi_{1:m})\right].$$

The first term is the log-likelihood of the data expected under the approximate posterior $q(\alpha, \beta, \pi_{1:m})$; it represents the goodness of fit (or accuracy) of the model. The second term is the Kullback-Leibler divergence between the approximate posterior and the prior; it represents the complexity of the model. This complexity term increases with the number of parameters, their inverse prior covariances, and with the deviation of the posterior from the prior that is necessary to fit the data. Thus, the free-energy approximation shows that the model evidence incorporates a trade-off between explaining the observed data (i.e., goodness of fit) and remaining consistent with our prior (i.e., simplicity or negative complexity). In other words, the model evidence encodes how well a model strikes the balance between explaining the data and remaining simple (Pitt and Myung, 2002; Beal, 2003; Stephan et al., 2009).

Classical approaches differ from the Bayesian framework presented above in several ways. For a comparison between classical and Bayesian inference, see Appendix D.

3. Applications

This section illustrates the practical utility of the Bayesian models discussed in the previous section and compares them to inference obtained through classical (frequentist) statistics. We begin by simulating classification outcomes to highlight the key features of Bayesian mixed-effects inference (Sections 3.1 and 3.2). We then contrast inference on accuracies with inference on balanced accuracies (Section 3.3). Finally, we illustrate the application of our approach to synthetic data (Section 3.4) as well as empirical data obtained from an imaging experiment (Section 3.5).

3.1 Inference on the Population Mean and the Predictive Accuracy

In a first experiment, we simulated classification outcomes for a group of 20 subjects with 100 trials each. Outcomes were generated using the beta-binomial model with a population mean of 0.75 and a population variance of 0.02 (i.e., $\alpha \approx 6.28$ and $\beta \approx 2.09$, corresponding to a population standard deviation of 0.141; Figure 4).

Raw data, that is, the number of correct predictions within each subject, are shown in Figure 4a. Their empirical sample accuracies are shown in Figure 4b, along with the ground-truth density of the population accuracy. Inverting the beta-binomial model, using the MCMC procedure of Section 2.1 (Figure 4c), and examining the posterior distribution over the population mean accuracy showed that more than 99.9% of its mass was above 50%, in agreement with the fact that the true population mean was above chance (Figure 4d).

We also used this simulation to illustrate the differences between a Bayesian mixed-effects central 95% posterior probability interval, a fixed-effects probability interval, and a random-effects confidence interval (Figure 4e). All three schemes arrive at the same conclusion with respect to the population mean being above chance. However, while the random-effects interval (red) is very similar to the proposed mixed-effects interval (black), the fixed-effects interval (yellow) displays too small a variance as it disregards the important between-subjects variability.

We finally considered the predictive posterior distribution over the accuracy that would be observed if we were to acquire data from a new subject (Figure 4f). This posterior did not allow for the conclusion that, with a probability larger than 0.95, the accuracy in a new subject would be above chance. This result is driven by the large heterogeneity in the population, inducing a dispersed predictive density. Importantly, the dispersion of the predictive density would not vanish even in the limit of an infinite number of subjects. This is in contrast to the dispersion of the posterior over the population mean, which becomes more and more precise with an increasing amount of data.

Inference was based on 100 000 samples, generated using 8 parallel chains. We used several standard approaches to convergence evaluation. In particular, we considered trace plots for visual inspection of mixing behaviour and convergence to the target distributions. In addition, we monitored the average ratio of within-chain variance to between-chain variance, which was bigger than 0.995. In other words, the variances of samples within and between chains were practically indistinguishable. The Metropolis rejection rate was 0.475, thus ensuring an appropriate balance between exploration (of regions with a lower density) and exploitation (of regions with a higher density). Finally, we assessed the uncertainty inherent in MCMC-based quantities such as log Bayes factors by computing standard deviations across repetitions, which led us to use 10^5 or 10^6 samples for each computation (see Section 3.3). All subsequent applications were based on the same algorithmic settings.

In frequentist inference, a common way of representing the statistical properties of a test is to estimate the probability of rejecting the null hypothesis at a fixed threshold (e.g., 0.05) under different regimes of ground truth, which leads to the concept of *power curves*. Here, we adopted this frequentist perspective to illustrate the properties of Bayesian mixed-effects inference on classification performance (Figure 5).

Specifying a true population mean of 0.5 and variance of 0.001 (standard deviation 0.0316), we generated classification outcomes, in the same way as above, for a synthetic group of 20 subjects with 100 trials each. Inverting the beta-binomial model, we inferred whether the population mean was above chance by requiring more than 95% of the posterior probability mass of the population

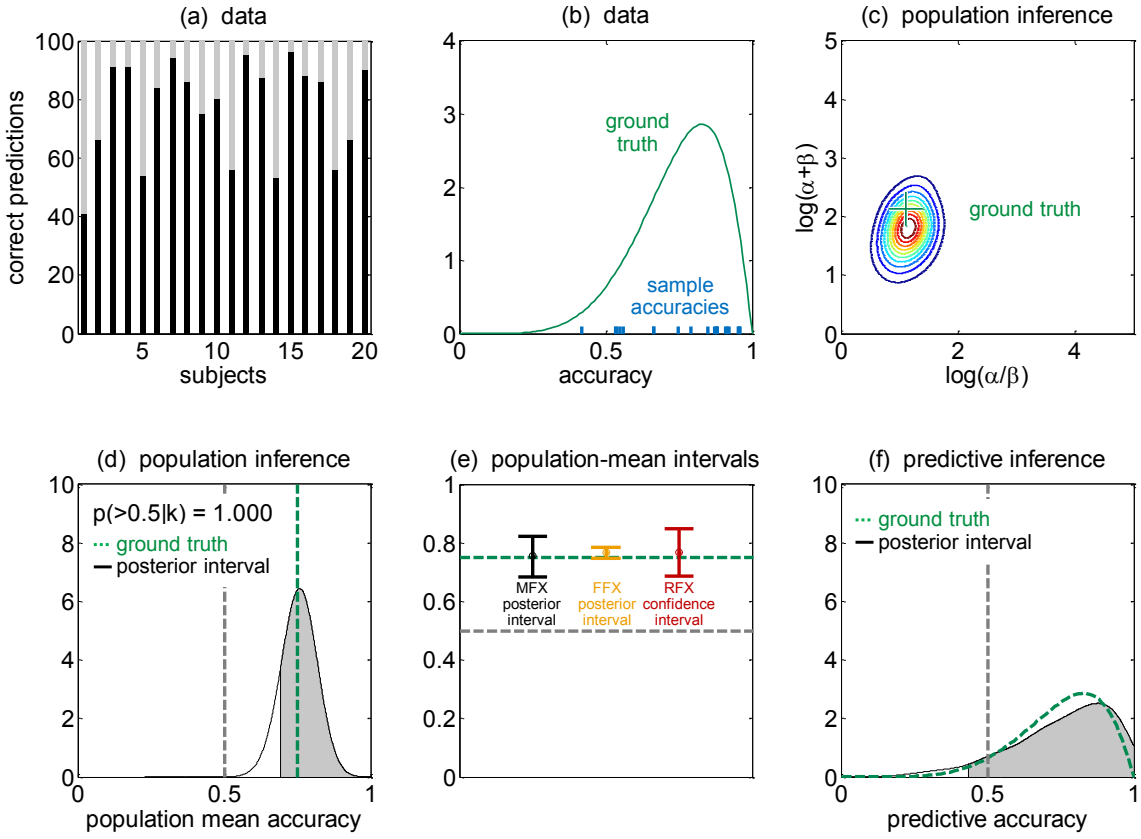

Figure 4: Inference on the population mean and the predictive accuracy. (a) Classification outcomes were generated for 20 subjects using the beta-binomial model. Each subject is fully characterized by the number of correctly classified trials (black) out of a given set of 100 trials (grey). (b) Empirical sample accuracies (blue) and their underlying population distribution (green). (c) Inverting the beta-binomial model yields samples from the posterior distribution over the population parameters, visualized using a nonparametric (bivariate Gaussian kernel) density estimate (contour lines). (d) The posterior about the population mean accuracy, plotted using a kernel density estimator (black), is sharply peaked around the true population mean (green). The upper 95% of the probability mass are shaded (grey). Because the lower bound of the shaded area is greater than 0.5, the population mean can be concluded to be above chance. (e) While the central 95% posterior interval (black) and the classical 95% confidence interval (red) look similar, the two intervals are conceptually very different. The fixed-effects interval (orange) is overly optimistic as it disregards between-subjects variability. (f) The posterior predictive distribution over $\tilde{\pi}$ represents the posterior belief of the accuracy expected in a new subject (black). Its dispersion reflects irreducible population heterogeneity.

mean to be greater than 0.5, that is, by requiring an infraliminal probability of less than 5%. We repeated this process 1 000 times and counted how many times the population mean was deemed

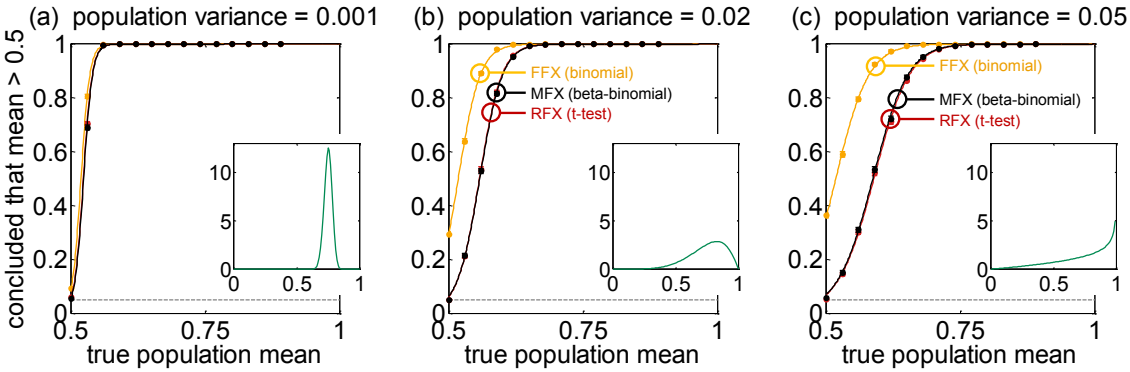

Figure 5: Inference on the population mean under varying population heterogeneity. The figure shows Bayesian estimates of the frequentist probability of above-chance classification performance, as a function of the true population mean, separately for three different level of population heterogeneity (a,b,c). Each data point is based on 1 000 simulations, each of which used 10 000 samples from every subject-specific posterior to make a decision. The figure shows that, in this setting, frequentist inference based on *t*-tests (red) agrees with Bayesian inference based on the beta-binomial model (black). By contrast, a fixed-effects approach (orange) offers invalid population inference as it disregards between-subjects variability; at a true population mean of 0.5, the hypothesis of chance-level performance is rejected more frequently than prescribed by the test size. Each data point is plotted in terms of the fraction of above-chance conclusions and a 95% central posterior interval, based on a Beta model with a flat prior. Points are joined by a sigmoidal function that was constrained to start at 0 and end at 1, with two remaining degrees of freedom. Where the true population mean exceeds 0.5, the graphs reflect the empirical sensitivity of the inference scheme. Its empirical specificity corresponds to the vertical distance between the graphs and 1 at the point where the population mean is 0.5. Insets show the distribution of the true underlying population accuracy (green) for a population mean accuracy of 0.75.

greater than chance. We then varied the true population mean and plotted the fraction of decisions for an above-chance classifier as a function of population mean (Figure 5a). At a population mean of 0.5, the vertical distance between the data points and 1 represents the empirical specificity of the test (which was designed to be $1-\alpha = 0.95$). At population means above 0.5, the data points show the empirical sensitivity of the test, which grows rapidly with increasing population mean. In this setting, the inferences that one would obtain by a frequentist *t*-test (red) are in excellent agreement with those afforded by the proposed beta-binomial model (black). Since the population variance was chosen to be very low in this initial simulation, the inferences afforded by a fixed-effects analysis (yellow) prove very similar as well; but this changes drastically when increasing the population variance to more realistic levels, as described below.

One important issue in empirical studies is the heterogeneity of the population. We studied the effects of population variance by repeating the above simulations with different variances (Figures 5b,c). As expected, an increase in population variance reduced statistical sensitivity. For

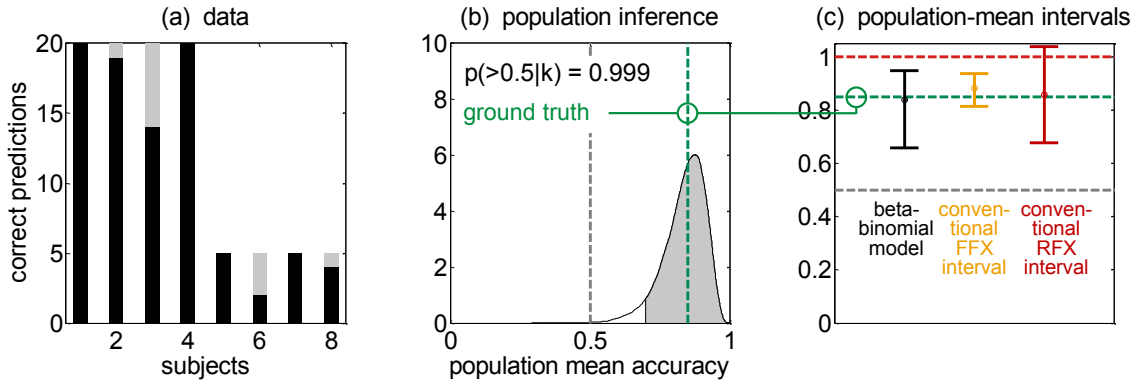

Figure 6: Inadequate inferences provided by fixed-effects and random-effects models. (a) The simulation underlying this figure represents the case of a small heteroscedastic group with varying numbers of trials across subjects. Classification outcomes were generated in the same way as in the simulation underlying Figure 5a. (b) The (mixed-effects) posterior density of the population mean (black) provides a good estimate of ground truth (green). (c) A central 95% posterior probability interval, based on the density shown in (b), comfortably includes ground truth. By contrast, a fixed-effects interval (orange) is overoptimistic as it disregards between-subjects variability. The corresponding random-effects confidence interval (red) is similar to the mixed-effects interval but lacks asymmetry, thus inappropriately including accuracies above 100% (red dashed line).

example, given a fairly homogeneous population with a true population mean accuracy of 60% and a variance of 0.001 (standard deviation 0.0316; Figure 5a), we can expect to correctly infer above-chance performance in more than 99.99% of all cases. By contrast, given a more heterogeneous population with a variance of 0.05 (standard deviation ≈ 0.22), the fraction of correct conclusions drops to 61%; in all other cases we would fail to recognize that the classifier was performing better than chance.

The above simulations show that a fixed-effects analysis (yellow) becomes an invalid procedure to infer on the population mean when the population variance is non-negligible. In more than the prescribed 5% of simulations with a true population mean of 0.5, the procedure concluded that the population mean was above chance. This is because a fixed-effects analysis yields too small variances on the population mean and therefore too easily makes above-chance conclusions.

All above simulations were based on a group of 20 subjects with 100 trials each. This emulated a setting as it frequently occurs in practice, for example, in neuroimaging data analyses. We repeated the same analysis as above on a sample data set from a second simulation setting (Figure 6). This setting was designed to represent the example of a small heterogeneous group with varying numbers of trials across subjects. Specifically, we generated data for 8 subjects, half of which had 20 trials, and half of which had only 5 trials. Classification outcomes were generated using the beta-binomial model with a population mean of 0.85 and a population variance of 0.02 (corresponding to a population standard deviation of 0.14; Figure 6a).

The example shows that the proposed beta-binomial model yields a posterior density with the necessary asymmetry; it comfortably includes the true population mean (Figure 6b). By contrast,

the fixed-effects probability interval (based on a Beta density) is overly optimistic. Finally, the random-effects confidence interval is similar to the mixed-effects interval but lacks the necessary asymmetry, including accuracies above 100% (Figure 6c).

3.2 Inference on Subject-Specific Accuracies

In the Bayesian models of this paper, classification accuracies of individual subjects are represented by a set of latent variables $\pi_1, \ldots, \pi_m$. A consequence of hierarchical Bayesian inference is that such subject-specific variables are informed by data from the entire group. Effectively, they are *shrunk* to the group mean, where the amount of shrinkage depends on the subject-specific posterior uncertainty.

To illustrate this, we generated synthetic classification outcomes and computed subject-specific posterior inferences (Figure 7). This simulation was based on 45 subjects overall; 40 subjects were characterized by a relatively moderate number of trials ($n = 20$) while 5 subjects had even fewer trials ($n = 5$). The population accuracy had a mean of 0.8 and a variance of 0.01 (standard deviation 0.1). Using this data set, we computed subject-specific central 95% posterior probability intervals and sorted them in ascending order by subject-specific sample accuracy (Figure 7a). The plot shows that, in each subject, the posterior mode (black) represents a compromise between the observed sample accuracy (blue) and the population mean (0.8). This compromise in turn provides a better estimate of ground truth (green) than sample accuracies by themselves. This effect demonstrates a key difference between the two quantities: subject-specific posteriors are informed by data from the entire group, whereas sample accuracies are based on the data from an individual subject.

Another way of demonstrating the shrinkage effect is by illustrating the transition from ground truth to sample accuracies (with its increase in dispersion) and from sample accuracies to posterior means (with its decrease in dispersion). This shows how the high variability in sample accuracies is reduced, informed by what has been learned about the population (Figure 7b). Notably, because the amount of shrinking depends on each subject's posterior uncertainty, the shrinking effect may modify the order of subjects, as indicated by crossing lines. Here, subjects with only 5 trials were shrunk more than subjects with 20 trials.

In a next step, we examined power curves, systematically changing the true population accuracy and repeating the above simulation 1 000 times (Figure 7c). Within a given simulation, we concluded that a subject-specific accuracy was above chance if more than 95% of its posterior probability mass was above 0.5. We binned subjects across all simulations into groups of similar accuracies and plotted the fraction of above-chance decisions against these true accuracies, contrasting the Bayesian model with a conventional t-test. As shown in Figure 7c, t-tests falsely detected above-chance subject-specific accuracies in about 5% of the cases, in agreement with the intended test size. By contrast, our Bayesian scheme was considerably more sensitive and detected above-chance accuracy in subjects whose true accuracy was within a small bin around 0.5. This reflected the fact that the Bayesian procedure incorporated what had been learned about the population when deciding on individual subjects. That is, a population mean well above chance (here: 0.8) made it more likely that individual subjects performed above chance as well, even in the presence of a low sample accuracy.

In addition to enabling decisions that take into account information about the group, the posterior distributions of subject-specific accuracies also yield more precise point estimates. To illustrate this, we simulated 1 000 data sets in the same way as above. Within each simulation, we compared

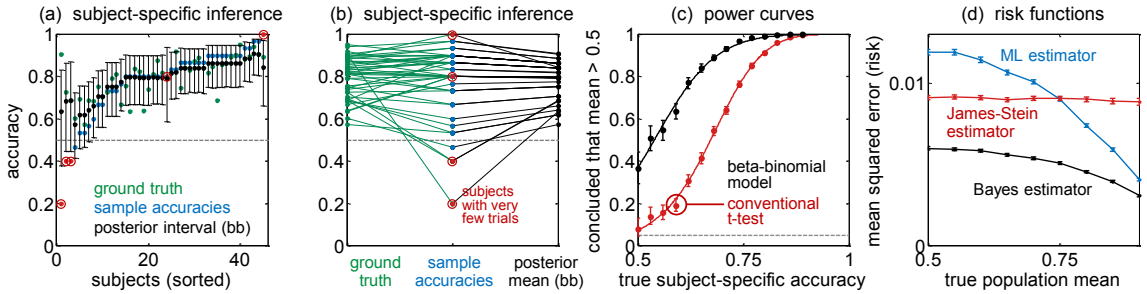

Figure 7: Inference on subject-specific accuracies. (a) Classification outcomes were generated for a synthetic heterogeneous group of 45 subjects (40 subjects with 20 trials each, 5 subjects with 5 trials each). All data were generated using the beta-binomial model (population mean 0.8, standard deviation 0.1). The figure shows subject-specific posterior means and central 95% posterior probability intervals (black), sample accuracies (blue if based on 20 trials, red if based on 5 trials), and true subject-specific accuracies (green) as a function of subject index, sorted in ascending order by sample accuracy. Whenever a subject's sample accuracy is very low or very high in relation to the remaining group, the Bayesian posterior interval is shrunk to the population. (b) Another way of visualizing the shrinking effect is to contrast the increase in dispersion (as we move from ground truth to sample accuracies) with the decrease in dispersion (as we move from sample accuracies to posterior means) enforced by the hierarchical model. Shrinking changes the order of subjects (when sorted by posterior mean as opposed to by sample accuracy) as the amount of shrinking depends on the subject-specific (first-level) posterior uncertainty. Subjects with just 5 trials (red) are shrunk more than subjects with 20 trials (blue). (c) Based on 1 000 simulations, the plot shows the fraction of simulations in which a subject's accuracy was concluded to be above chance, based on a Bayesian posterior interval (black) or a frequentist t-test (red). In contrast to classical inference, the Bayesian procedure incorporates a desirable shift towards the population in making decisions about individual group members. (d) Across the same 1 000 simulations, a Bayes estimator, based on the posterior means of subject-specific accuracies (black), was superior to both a classical ML estimator (blue) and a James-Stein estimator (red).

three different ways of obtaining an estimator for each subject's accuracy: (i) a Bayes estimator (posterior mean of the subject-specific accuracy); (ii) a maximum-likelihood estimator (sample accuracy); and (iii) a James-Stein estimator, with a similar shrinkage effect as the Bayes estimator but less explicit distributional assumptions (Figure 7d). For each estimator, we computed the mean squared error (or risk) across all subjects, averaged across all simulations. We then repeated this process for different population means. We found that the James-Stein estimator outperformed the ML estimator for low accuracies. However, both estimators were dominated by (i.e., inferior to) the Bayes estimator which provided the lowest risk throughout.

It is important to keep in mind that the above simulations are based on synthetic classification outcomes which fulfil the assumptions of the normal-binomial model by design, in particular the assumption of logit-normally distributed subject-specific accuracies and the assumption of condi-

tional independence given the population parameters. For empirical data, these assumptions may not always hold and so posterior inferences, including the shrinkage effect, may be suboptimal. This highlights the importance of model checking when using the models presented in this paper in practical applications.

3.3 Inference on the Balanced Accuracy

The balanced accuracy is a more useful performance measure than the accuracy, especially when a classifier was trained on an imbalanced test set and may thus exhibit bias. In order to illustrate the relative utility of these two measures in our Bayesian models, we simulated an imbalanced data set, composed of 20 subjects with 100 trials each, where each subject had between 70 and 90 positive trials (drawn from a uniform distribution) and between 10 and 30 negative trials.

An initial simulation specified a high population accuracy on the positive class and a low accuracy on the negative class, with equal variance in both (Figure 8a,b). These accuracies were chosen such that the classifier would perform at chance on a hypothetical balanced sample. This allowed us to mimic the commonly observed situation in which a classifier takes advantage of the imbalance in the data and preferably predicts the majority class. We independently inverted three competing models: (i) the beta-binomial model to infer on the classification accuracy; and the (ii) twofold beta-binomial and (iii) bivariate normal-binomial models to infer on the balanced accuracy. As expected, the beta-binomial model falsely suggested high evidence for above-chance classification. In contrast, the twofold beta-binomial and normal-binomial models correctly indicated the absence of a statistical relation between data and class labels (Figure 8c).

These characteristics were confirmed across a large set of simulations. As expected, inference on the accuracy falsely concluded above-chance performance, especially in the presence of a significant degree of class imbalance. By contrast, inference on the balanced accuracy did not incorrectly reject the hypothesis of the classifier operating at the level of chance more often than prescribed by the test size (Figure 8d).

We compared the two models for inference on the balanced accuracy by means of Bayesian model comparison. Using 10^6 samples with Equation (17), we obtained a log Bayes factor of 33.2 in favour of the twofold beta-binomial model (i.e., under a flat prior over models, the posterior belief in the beta-binomial model is greater than 99.99%). This result represents very strong evidence (Kass and Raftery, 1995) that the beta-binomial model provided a better explanation of the synthetic classification outcomes than the normal-binomial model. This finding is plausible since no correlation structure among class-specific accuracies was imposed in the simulation; thus, in this case, the beta-binomial model is a better model than the more complex normal-binomial model.

To assess the sampling-induced uncertainty about this result, we repeated the computation of the log Bayes factor 100 times. We obtained a sample standard deviation of 8.0, that is, the uncertainty was small in relation to the overall strength of evidence. By comparison, when using only 10^3 samples instead of 10^6, the standard deviation increased to 25.5. We used 10^6 samples for all subsequent analyses.

We repeated the main analysis above 1 000 times and plotted the fraction of above-chance conclusions against the degree of class imbalance. Note that the resulting curve is not a power curve in the traditional sense, as its independent variable is not the true (balanced) accuracy but the accuracy on positive trials, that is, an indicator of the degree of class imbalance. Figure 8d shows that the simple beta-binomial model provides progressively misleading conclusions with class imbalance at

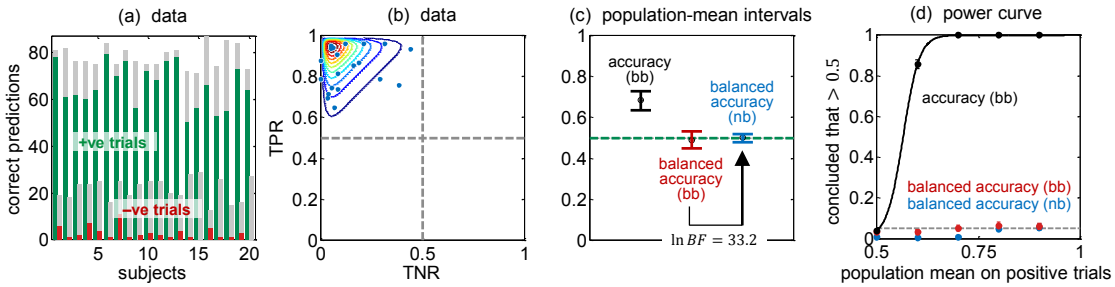

Figure 8: Inference on the balanced accuracy. (a) The simulation underlying this figure mimics an imbalanced data set which has led the classification algorithm to acquire a bias in favour of the majority class. The plot shows, for each subject, the number of correctly classified positive (green) and negative (red) trials, as well as the respective total number of trials (grey). (b) Visualizing sample accuracies separately for the two classes gives rise to a two-dimensional plot, in which the true positive rate on the y-axis and the true negative rate on the x-axis represent the accuracies on positive and negative trials, respectively. The underlying true population distribution is represented by a bivariate Gaussian kernel density estimate (contour lines). The plot shows that the population accuracy is high on positive trials and low on negative trials. (c) Central 95% posterior probability intervals based on three models: the simple beta-binomial model for inference on the population accuracy; and the twofold beta-binomial model as well as the bivariate normal-binomial model for inference on the balanced accuracy. The true mean balanced accuracy in the population is at chance (green). It is accurately estimated by models inferring on the balanced accuracy (red, blue). Bayesian model selection yielded very strong evidence (Kass and Raftery, 1995) in favour of the normal-binomial model (posterior model probability = 97.7%). (d) Probability of falsely detecting above-chance performance, using different inference schemes. The true balanced accuracy is 0.5. The x-axis represents the degree of class imbalance.

the group level (cf. Figure 5). In contrast, both schemes for inference on the balanced accuracy made above-chance conclusions in less than 5% of the simulations, as intended by their test size.

All models considered in this paper are based on diffuse priors designed in such a way that posterior inferences are clearly dominated by the data. However, one might ask to what extent such inferences depend on the exact form of the prior. To examine this dependence, we carried out a sensitivity analysis in which we considered the infraliminal probability of the posterior population mean as a function of prior moments (Figure 9). We found that inferences were extremely robust, in the sense that the influence of the prior moments on the resulting posterior densities was negligible in relation to the variance resulting from the fact that we are using a (stochastic) approximate inference method for model inversion. In particular, varying the constant (originally: 1) in Equation (7) for the beta-binomial prior left the infraliminal probability of the posterior accuracy unaffected (Figure 9a,b). Similarly, varying μ_0, κ_0, or ν_0 in the normal-binomial model had practically no influence on the infraliminal probability of the posterior balanced accuracy (Figure 9c,d,e).

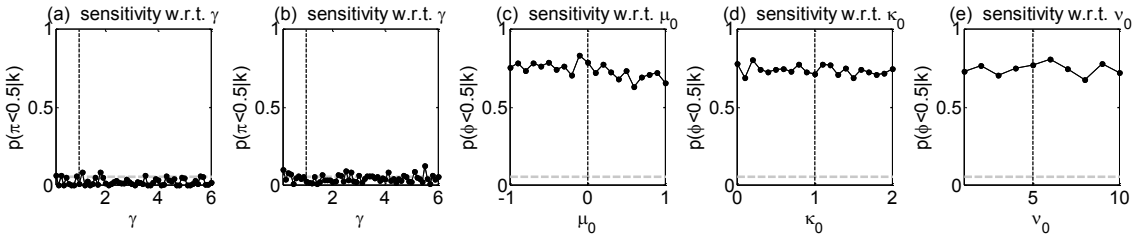

Figure 9: Sensitivity analysis. This figure illustrates the dependence of posterior inferences on the exact form of the priors proposed in this paper. Each graph shows the infraliminal probability of the population mean accuracy (i.e., the posterior probability mass below 0.5) as a function of a particular parameter of the prior distribution used for model inversion. (a,b) Same data sets as shown those shown in Figures 4a and 6a, but with a slightly lower population mean of 0.7. Inferences on the population accuracy are based on the beta-binomial model. (c,d,e) Same data set as shown in Figure 8a. Inferences on the population balanced accuracy are based on the bivariate normal-binomial model.

3.4 Application to Synthetic Data

All experiments described so far were based on classification outcomes sampled from the beta-binomial or normal-binomial model. This ensured, by construction, that the distributional assumptions underlying the models were fulfilled. To illustrate the generic applicability of our approach when its assumptions are not satisfied by construction, we applied models for mixed-effects inference to classification outcomes obtained on synthetic data features for a group of 20 subjects with 100 trials each (Figure 10). In addition to probing the models' robustness with regard to distributional assumptions, this allows one to examine what correlations between class-specific accuracies may be observed in practice.

Synthetic data were generated using a two-level sampling approach that reflected the hierarchical nature of group studies. We specified a population distribution, sampled subject-specific means and variances from it, and then used these to generate trial-specific feature vectors. In a first simulation (Figure 10, top row), we generated 50 positive trials and 50 negative trials for each individual subject j by drawing one-dimensional feature vectors from two normal distributions, $\mathcal{N}(x_{ij} \mid \mu_j^+, \sigma_j)$ and $\mathcal{N}(x_{ij} \mid \mu_j^-, \sigma_j)$, respectively. The moments of these subject-specific distributions, in turn, were drawn from a population distribution, using $\mathcal{N}(\mu_j^+ \mid \frac{1}{2}, \frac{1}{2})$ and $\mu_j^- = -\mu_j^+$ for the means, and $\mathrm{Ga}^{-1}(\sigma_j \mid 10, \frac{1}{10})$ for the variance. The normal distribution and the inverse Gamma distribution are conjugate priors for the mean and variance of a univariate normal distribution and, thus, represent natural choices for the population distribution.

To obtain classification outcomes, separately for each subject, we trained and tested a linear support vector machine (SVM), as implemented by Chang and Lin (2011), using 5-fold cross-validation. Classification outcomes are shown in Figure 10a, in which the numbers of correctly classified trials are illustrated separately for the two classes and for each subject. The same data are represented in terms of sample accuracies in Figure 10b (blue dots). To illustrate ground truth, we repeated the above procedure (of generating synthetic data and applying an SVM) 1 000 times and added a contour plot of the resulting distribution of sample accuracies in the same figure. This dis-

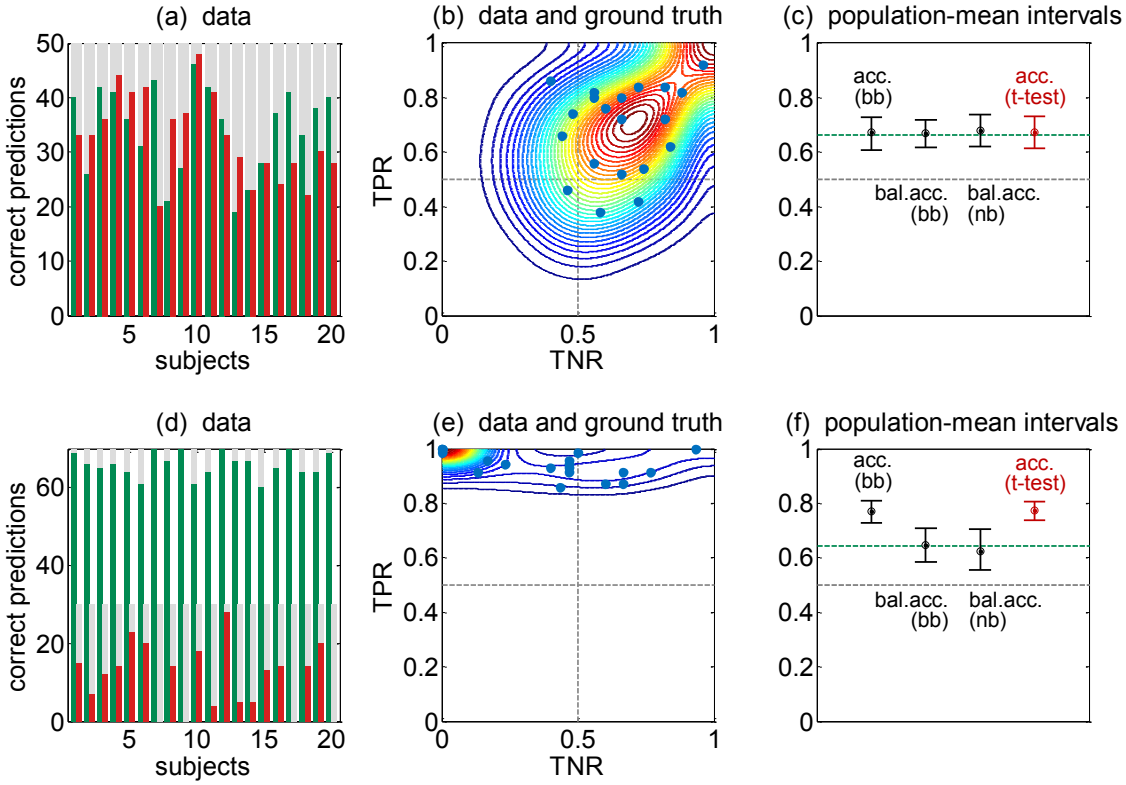

Figure 10: Application to synthetic data. (a) Classification outcomes obtained by applying a linear support vector machine to synthetic data, using 5-fold cross-validation. (b) Sample accuracies on positive (TPR) and negative classes (TNR) show the positive correlation between class-specific accuracies (blue). The underlying population distribution is represented by a bivariate Gaussian kernel density estimate (contour lines). (c) Central 95% posterior probability intervals, resulting from inversion of the beta-binomial model for inference on the population mean accuracy as well as the twofold beta-binomial model (*bb*) and the bivariate normal-binomial model (*nb*) for inference on the population mean balanced accuracy (all black). A frequentist 95% confidence interval (red) is shown for comparison. Bayesian model selection yielded very strong evidence (Kass and Raftery, 1995) in favour of the normal-binomial model (posterior model probability = 99.99%). (d) A second simulation was based on a synthetic heterogeneous group with varying numbers of trials. (e) In this case, the classifier acquires a strong bias in favour of the majority class. (f) As a result, inference on the accuracy is misleading; the balanced accuracy is a much better performance indicator, whether based on the beta-binomial (*bb*) or normal-binomial model (*nb*).

tribution was symmetric with regard to class-specific accuracies while these accuracies themselves were strongly positively correlated, as one would expect from a linear classifier tested on perfectly balanced data sets.

We applied all three models discussed in this paper for inference: the beta-binomial model for inference on the accuracy (Section 2.1), and the twofold beta-binomial and normal-binomial model for inference on the balanced accuracy (Sections 2.2 and 2.3). Central 95% posterior probability intervals about the population mean are shown in Figure 10c, along with a frequentist confidence interval of the population mean accuracy. All four approaches provided precise intervals around the true population mean. Comparing the two competing models for inference on the balanced accuracy, we obtained a log Bayes factor of 22.1 in favour of the twofold beta-binomial model (posterior model probability $> 99.99\%$; standard deviation of log Bayes factor across computations ≈ 5.1), representing very strong evidence (Kass and Raftery, 1995) that this model provided a better explanation of the data (i.e., a better balance between fit and complexity) than the bivariate normal-binomial model. This finding makes sense in light of the posterior covariance matrix of the normal-binomial model (cf. Figure 2b). Its off-diagonal elements (accounting for the potential dependency between class-specific accuracies) did not only have a very small mean ($\Sigma_{12} = \Sigma_{21} = 0.19$); they were also associated with considerable posterior uncertainty (95% credible interval $[-0.01, 0.44]$). In other words, the small additional fit provided by the off-diagonal elements was outweighed by the additional model complexity incurred.

We repeated the above analysis with a subtle but important modification: instead of using perfectly balanced data (50 positive and 50 negative trials), we created imbalanced synthetic data using 70 positive and 30 negative trials per subject. All other details of the analysis remained unchanged (Figure 10, bottom row). We observed that, as expected, the class imbalance caused the classifier to acquire a bias in favour of the majority class. This can be seen from the raw classification outcomes in which many more positive trials (green) than negative trials (red) were classified correctly, relative to their respective prevalence in the data (grey; Figure 10d). The bias is reflected accordingly by the estimated bivariate density of class-specific classification accuracies, in which the majority class consistently performs well whereas the accuracy on the minority class varies strongly (Figure 10e). In this setting, we found that both the twofold beta-binomial model and the normal-binomial model provided excellent estimates of the true balanced accuracy (Figure 10f; log Bayes factor in favour of the beta-binomial model: 47.3; standard deviation 11.3). In stark contrast, using the single beta-binomial model or a conventional mean of sample accuracies to infer on the population accuracy (as opposed to balanced accuracy) resulted in estimates that were overly optimistic and therefore misleading.

3.5 Application to Empirical Data

In order to illustrate the practical application of the approaches discussed in this paper, we analysed a neuroimaging data set, obtained by functional magnetic resonance imaging (fMRI). In neuroimaging, classifiers are often used as part of decoding models designed to infer a perceptual or cognitive state from brain activity, typically on a trial-by-trial basis, but across a group of subjects. The interpretation of the ensuing results critically relies on the validity of the models used for inference on classification performance.

Here, we analysed data from an fMRI experiment involving 16 volunteers designed to study the cognitive processes underlying decision making. During the experiment, subjects had to choose, on each trial, between two alternative options. Choices were indicated by button press (left/right index finger). Over the course of the experiment, subjects learned, by trial and error, the reward probabilities of these two options. Details on experimental design, data acquisition, and preprocessing can

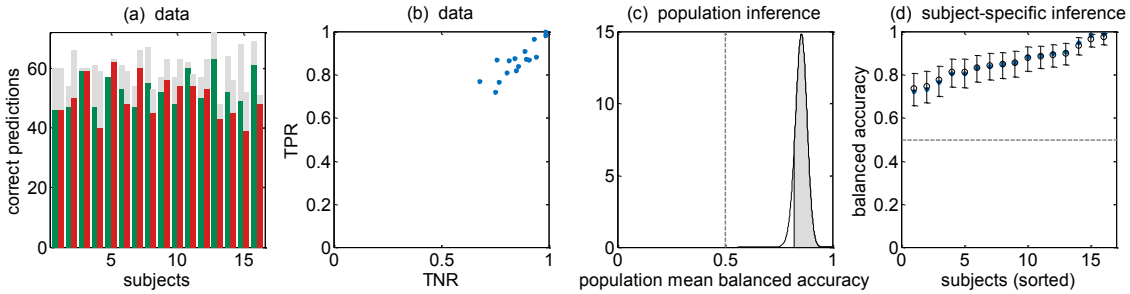

Figure 11: Application to empirical data. (a) Classification outcomes obtained by applying a linear SVM to trial-wise fMRI data from a decision-making task. (b) Replotting classification outcomes in terms of sample accuracies on positive (TPR) and negative trials (TNR) reveals the positive correlation between class-specific accuracies. (c) In this data set, when inferring on the balanced accuracy, the bivariate normal-binomial model has a higher evidence (marginal likelihood) than the twofold beta-binomial model. Inverting the former model, which captures potential dependencies between class-specific accuracies, yields a posterior distribution over the population mean balanced accuracy (black) which shows that the classifier is performing above chance. (d) The same model can be used to obtain subject-specific posterior inferences. The plot contrasts sample accuracies (blue) with central 95% posterior probability intervals (black), which avoid overfitting by shrinking to the population mean.

be found elsewhere (Behrens et al., 2007). Here, we predicted from the fMRI data, on a trial-by-trial basis, which option had been chosen. Because choices were indicated by button presses, we expected highly discriminative activity in the primary motor cortex.

Separately for each subject, a general linear model (Friston et al., 1995) was used to create a set of parameter images representing trial-specific estimates of evoked brain activity in each volume element. These images were used for subsequent classification. We trained a linear support vector machine (SVM) using 5-fold cross-validation. Comparing predicted to actual choices resulted in 120 classification outcomes for each of the 16 subjects. These data were used for inference on the classification accuracy using the beta-binomial model (Figure 11).

As can be seen from raw classification outcomes, class-specific accuracies seemed to be positively correlated (Figures 11a,b), in a similar way as for the synthetic data considered above. Thus, we used both the twofold beta-binomial model and the bivariate normal-binomial model for inference. Bayesian model comparison yielded a log Bayes factor of 12.5 in favour of the beta-binomial model (standard deviation across computations ≈ 4.69), suggesting that the additional complexity of the normal-binomial model may not have balanced its higher flexibility in explaining the correlations between class-specific accuracies. Using the beta-binomial model for inference on the population mean balanced accuracy, we obtained very strong evidence (infraliminal probability $p < 0.001$) that the classifier was operating above chance (Figure 11c).

Inference on subject-specific accuracies yielded fairly precise posterior intervals (Figure 11d). The shrinkage effect in these subject-specific accuracies was rather small: the average absolute difference between sample accuracies and posterior means amounted to 1.39 percentage points.

Even the biggest observed shift among all subjects was no more than 3.05 percentage points (from a sample accuracy of 99.2% down to a posterior mean of 96.2%). This minor impact of shrinkage is expected given the relatively small number of subjects (16) and the relatively large number of trials per subject (120).

4. Discussion

Canonical classification algorithms are frequently used on multilevel or hierarchically structured data sets, where a classifier is trained and tested for each subject within a group. This paper showed how the evaluation of classification performance in this setting may benefit from mixed-effects models that explicitly capture the hierarchical structure of the data. We organize the following discussion around the three principal features of this approach.

4.1 Replacing Fixed-Effects by Mixed-Effects Models

The primary contribution of this paper is the introduction and analysis of several models for Bayesian mixed-effects inference for group-level classification studies. To capture the two key sources of variation in hierarchical data sets, we simultaneously account for fixed-effects (within-subjects) and random-effects (across-subjects) variance components. This idea departs from previous models which are widely used for classification studies but ignore within- or between-subjects variability. Fixed-effects models make inappropriate assumptions and yield overconfident inference. Conversely, random-effects models treat subject-specific sample accuracies as observed, rather than inferred, and thus omit uncertainty associated with such sample accuracies.

The mixed-effects models considered in this paper ensure that known dependencies between inferences on subject-specific accuracies are accommodated within an internally consistent representation of the data. Specifically, the posterior distribution of the accuracy of one subject is partially influenced by the data from all other subjects, correctly weighted by their respective posterior precisions (see Section 3.2). Thus, the available group data are exploited to constrain individual inference appropriately. Non-hierarchical models, by contrast, risk being under-parameterized or over-parameterized. For example, pooling classification outcomes across subjects and modelling them as being drawn from a single distribution corresponds to an under-parameterized model whose single parameter (i.e., the latent accuracy) is insufficient to explain any population variability. Conversely, replicating the single-subject model in Equations (1)–(3) for each subject leads to an over-parameterized model with $2n$ parameters that is likely to overfit the data and generalize poorly. Hierarchical models overcome this problem in a natural way: they regularize the inversion problem by incorporating the structural dependencies that are assumed to govern the observed data.

An important aspect to keep in mind is that shrinkage is a posterior inference, and as such is conditional on the model. A corollary of this is that shrinkage is suboptimal when the hierarchical model structure represents an unreasonable assumption. This highlights the importance of model checking as an integral part of statistical inference. In particular, researchers applying the models proposed in this paper are advised to check whether the hierarchical structure of the models can be defended on substantive grounds. For example, in an experiment where each subject was either assigned to a treatment group or a control group, it may no longer be justified to treat their accuracies as conditionally independent and identically distributed given a single vector of population parameters; instead, it might be more appropriate to analyse the two subgroups separately (or augment the present models by a third level).

In those situations where a hierarchical structure is justified, we are not aware of alternatives that are superior to shrinkage. One possibility is to use *no pooling* of information across subjects, leading to a set of isolated subject-specific sample accuracies. Another possibility is *complete pooling*, leading to a single group mean accuracy. Between these two extremes lie the weighted estimates provided by a hierarchical model. Its shrinkage effect ensures that information from different sources is weighted correctly and incorporated into the posterior density of each model parameter.

Shrinkage is not a consequence of the Bayesian perspective adopted in this paper. It is a fundamental aspect of statistical dependencies in hierarchical structures which has been known for more than a century, dating back to work as early as Galton's law of 'regression towards mediocrity' (Galton, 1886). It is perfectly possible to obtain shrinkage through classical inference where it has undergone considerable scrutiny; one of the best-known examples is the James-Stein estimator (Appendix E) whose beneficial effect on estimation precision has long been recognized in frequentist statistics. For early contributions to the extensive literature on shrinkage effects, see Stein (1956), James and Stein (1961), and Efron and Morris (1971, 1972). For typical applications in other fields of science, see the many examples described by Gelman et al. (2003).

The hierarchical models presented in this paper are motivated by two-level designs that distinguish between inference at the subject level and inference at the group level. However, it should be noted that these models can easily be extended to accommodate multi-level studies. For example, in order to model classification performance in different task conditions or in different sessions, one could introduce separate latent accuracies $\pi_j^a, \pi_j^b, \ldots$, all of which are drawn from a common subject-specific accuracy π_j. In this way, one would explicitly model task- or session-specific accuracies to be conditionally independent from one another given an overall subject-specific effect π_j, and conditionally independent from other subjects given the population parameters. This example shows that additional relationships between portions of the acquired data can be naturally expressed in a hierarchical model to appropriately constrain inferences.

Mixed-effects models are not only useful when *evaluating* a classifier but also when *designing* it. For instance, Schelldorfer et al. (2011) proposed a linear mixed-effects model for classification that accounts for different sources of variation in the data. The model has been shown to improve classification performance in the domain of brain-computing interfaces (Fazli et al., 2011).

4.2 Replacing Frequentist by Bayesian Inference

The second feature of our approach is to provide Bayesian alternatives to the frequentist procedures that have been dominating classification group studies so far. Although these two schools share commonalities, there are primarily deep conceptual differences. Frequentist approaches consider the distribution of an estimator as a function of the unknown true parameter value and view probabilities as long-term frequencies; estimation yields point estimates and confidence intervals, while inference takes the form of statements on the probability of estimator values under a 'null hypothesis.' Bayesian methods, by contrast, consider the subjective belief about the parameter, before and after having observed actual data, drawing on probability theory to optimally quantify inferential uncertainty.

An additional aspect of Bayesian approaches is that one can evaluate different models by comparing their respective model evidences. This corresponds to inference about model structure as defined by the model's priors. For example, in Section 2.4 we showed how alternative *a priori* as-

sumptions about the population covariance of class-specific accuracies can be evaluated, relative to the priors of the models, using Bayesian model selection.

Bayesian inference in hierarchical models is typically analytically intractable, which is why we resort to approximate inference, for example by using stochastic approximation schemes based on MCMC methods. While computationally less efficient than deterministic approximations (e.g., variational Bayes, VB), these are easy to implement, avoid additional distributional assumptions, and are asymptotically exact. This paper exclusively relied on MCMC for model inversion. In future work, we will also provide VB algorithms for inverting models of the sort presented in this paper (see below).

It is worth noting that classical inference does not necessarily have to assume the form currently prevalent in the evaluation of hierarchical classification studies. For example, as noted by one of our reviewers, the t-test that is presently used by the large majority of classification analyses could be replaced by a classical mixed-effects model. This would require two things. Firstly, the definition of a decision statistic, for example, the fraction of correctly classified trials, pooled across subjects, or more simply, a hierarchical model such as the beta-binomial model, but estimated using maximum-likelihood estimation (for an example using logistic regression, see Dixon, 2008). Secondly, an inference scheme: under the null hypothesis that the classifiers perform at chance, the number of correctly/incorrectly classified trials can be swapped across subjects; this would provide a permutation mechanism to test the significance of the decision statistic.

An advantage of the above frequentist scheme would be that it no longer requires an assumption common to all other approaches considered in this paper: the assumption that trial-wise classification outcomes y_i are conditionally independent and identically distributed (i.i.d.) given a subject-specific accuracy π. This is typically justified by assuming that, in a classification analysis, test observations are i.i.d. themselves, conditional on the parameters of the latent process that generated the data. The situation is less clear in a cross-validation setting, where, strictly speaking, classification outcomes are no longer independent of one another (Kohavi, 1995; Wickenberg-Bolin et al., 2006; Gustafsson et al., 2010). Because violations of i.i.d. assumptions lead to conservative inference when controlling false positive rates, the i.i.d. assumption has generally not been a major concern in the literature; however, it remains a relevant topic, and further research into the ensuing statistical bias and its adequate correction is required. In the present paper, we used 5-fold cross-validation. If trial-by-trial dependence is an issue, then one possibility is to resort to a single-split (or hold-out) scheme, by training on one half of the data, and testing on the other (see Breiman and Spector, 1992, for details).

4.3 Replacing the Accuracy by the Balanced Accuracy

The third feature of our approach is its flexibility with regard to performance measures. While it is common to compare algorithms with regard to their accuracy, the limitations of this metric are well-known. For example, when a classifier is tested on an imbalanced data set, the accuracy may be inflated and lead to false conclusions about the classifier's performance. There are different potential solutions to this problem (Akbani et al., 2004; Chawla et al., 2002; Japkowicz and Stephen, 2002). One can, for example, restore balance by undersampling the large class or by oversampling the small class, or modify the costs of misclassification (Zhang and Lee, 2008). A more generic safeguard is to replace the accuracy with the balanced accuracy, defined as the arithmetic mean of

the class-specific accuracies. Unlike the measure described by Velez et al. (2007), the balanced accuracy is symmetric with respect to the type of class.[7]

Notably, the balanced accuracy is not confined to binary classification but can easily be generalized to K classes, by defining the balanced accuracy as the arithmetic mean of all K class-specific accuracies. For the twofold beta-binomial model, one could then replace π^+ and π^- by $\pi^{(1)}, \pi^{(2)}, \ldots, \pi^{(K)}$, whereas for the normal-binomial model, the bivariate normal distribution would be replaced by a K-dimensional normal distribution.

Using the example of the balanced accuracy, we have described how hierarchical models enable Bayesian inference on performance measures other than the accuracy. Future examples might include functional measures such as the receiver-operating characteristic (ROC) or the precision-recall curve (cf. Brodersen et al., 2010b). We also demonstrated that there may be multiple plausible models *a priori*. In this case, Bayesian model selection can be used to decide between competing models. Alternatively, Bayesian model averaging produces predictions which account for posterior model uncertainty. This approach can be adopted with any other performance measure of interest.[8]

The choice of a versatile yet convenient parameterization of the distributions for class-specific accuracies π^+ and π^- has been a recurring theme in the literature. Whereas early treatments adopted an empirical Bayes approach (e.g., Albert, 1984; Good, 1956; Griffin and Krutchkoff, 1971), the more recent literature has discussed various fully hierarchical approaches (see Agresti and Hitchcock, 2005, for an overview). For instance, Leonard (1972) proposed to replace independent Beta priors on each element of π, such as those in (2.2), by independent normal priors on each element of $\text{logit}(\pi)$. While this is analytically convenient, it requires independence assumptions in relation to the elements of π. This limitation was addressed by Berry and Christensen (1979), who placed a Dirichlet process prior on the elements of π. A related approach was proposed by Albert and Gupta (1983), who placed Beta priors on the components of π such that their degree of correlation could be controlled by a common hyperparameter. As mentioned above, a principled way of evaluating such different propositions rests upon Bayesian model comparison (MacKay, 1992; Madigan and York, 1997; Penny et al., 2004), which we illustrate by deciding between alternative parameterizations for inference on the balanced accuracy.

A similar approach to the one discussed in this article has been suggested by Olivetti et al. (2012), who carry out inference on the population mean accuracy by comparing two beta-binomial models: one with a population mean prior at 0.5 (i.e., chance), and one with a uniform prior on the interval $[0.5, 1]$. Inference then takes the form of model selection, resulting in a Bayes factor and its conventional interpretation (Kass and Raftery, 1995). Our approach differs from the above work in four ways: (i) in addition to classification accuracy, we consider the balanced accuracy, which is a more useful performance measure whenever the data are not perfectly balanced, and for which we offer different parameterizations that can be optimized using Bayesian model selection; (ii) we explicitly frame our approach in terms of fixed-effects (FFX), random-effects (RFX), and mixed-effects (MFX) inference, and we provide the respective graphical models; (iii) we emphasize the use of uninformative priors on the interval $[0, 1]$ to obtain unbiased posterior estimates, which allows us to use infraliminal probabilities for inference; (iv) finally, we provide extensive simulation results

7. If desired, this symmetry assumption can be dropped by introducing class-specific misclassification costs.

8. It should be noted that, in this context, model selection is carried out to ask which model best explains observed classification outcomes. This is different from asking what sort of model (i.e., classification algorithm) might be best at classifying the data in the first place.

that demonstrate the differences between FFX, RFX, and MFX approaches, shrinkage effects, and reduced estimation risks.

4.4 Summary of Present Results and Conclusions

To examine the properties of our approach and demonstrate its practical applicability, we reported several applications of the different models to synthetic and empirical data. Our results illustrated the characteristic features of our approach: (i) posterior densities as opposed to point estimates of parameters; (ii) the ability to compare alternative (non-nested) models; (iii) the 'shrinking-to-the-population' effect that regularizes estimates of classification performance in individual subjects (Figure 7b); (iv) increased sensitivity (Figure 7c); (v) more precise parameter estimates (Figure 7d); (vi) avoidance of classifier bias for imbalanced data sets using the balanced accuracy (Figure 8).

One practical limitation of our approach lies in the high computational complexity of our current inversion methods. In particular, our MCMC algorithms lack guarantees about convergence rates. Our algorithms also include heuristics regarding the number of burn-in samples, the precision of the overdispersed initial distributions and the proposal densities, and regarding the number of chains run in parallel. To address these issues, we are currently preparing a variational Bayesian approach that may offer computationally highly efficient model inversion.

We hope that the models for Bayesian mixed-effects analyses introduced in this paper will find widespread use, improving the sensitivity and validity of future classification studies at the group level. To facilitate the use of our approach, an open-source MATLAB implementation of all models discussed in this paper is available for download (`http://mloss.org/software/view/407/`).

Acknowledgments

This research was supported by the University Research Priority Program 'Foundations of Human Social Behaviour' at the University of Zurich (KHB, KES), by the SystemsX.ch project 'Neurochoice' (KHB, KES), and by the NCCR 'Neural Plasticity' (KES). The authors wish to thank Thomas Nichols, Timothy E.J. Behrens, Mark W. Woolrich, Adrian Groves, Ged Ridgway, and Anne Broger for insightful discussions during the course of this research.

Appendix A. Inversion of the Beta-Binomial Model

The algorithm is initialized by drawing initial values for $\alpha^{(0)}$ and $\beta^{(0)}$ from an overdispersed starting distribution. We represent these as

$$\omega^{(0)} = \left(\ln\left(\frac{\alpha^{(0)}}{\beta^{(0)}}\right), \ln\left(\alpha^{(0)} + \beta^{(0)}\right) \right)^{\mathrm{T}}.$$

This coordinate transformation makes sampling more efficient (Gelman et al., 2003). Subsequently, on each iteration τ, a new candidate ω^* is drawn from a symmetric proposal distribution

$$q_\tau\left(\omega^* \,\middle|\, \omega^{(\tau-1)}\right) = \mathcal{N}_2\left(\omega^* \,\middle|\, \omega^{(\tau-1)}, \begin{pmatrix} 1/8 & 0 \\ 0 & 1/8 \end{pmatrix}\right).$$

This candidate sample ω^* is accepted with probability

$$\min\left\{1,\ \frac{p(k_{1:m} \mid \alpha^*, \beta^*)\, p(\alpha^*, \beta^*)}{p(k_{1:m} \mid \alpha^{(\tau-1)}, \beta^{(\tau-1)})\, p(\alpha^{(\tau-1)}, \beta^{(\tau-1)})}\right\}$$
$$= \min\left\{1,\ \exp\left(\sum_{j=1}^{m} f(\alpha^*, \beta^*, k_j) - f(\alpha^{(\tau-1)}, \beta^{(\tau-1)}, k_j)\right)\right\}$$

where (7) and (9) (main text) were used in defining

$$f(\alpha, \beta, k) := \ln \mathrm{Bb}(k \mid \alpha, \beta) + \ln p(\alpha, \beta).$$

In order to assess whether the mean classification performance achieved in the population is above chance, we must evaluate our posterior knowledge about the population parameters α and β. Specifically, inference on $\alpha/(\alpha+\beta)$ serves to assess the mean accuracy achieved in the population. For example, its posterior expectation represents a point estimate that minimizes a squared-error loss function,

$$\mathbb{E}\left[\frac{\alpha}{\alpha+\beta} \;\middle|\; k_{1:m}\right] \approx \frac{1}{c}\sum_{\tau=1}^{c} \frac{\alpha^{(\tau)}}{\alpha^{(\tau)}+\beta^{(\tau)}}.$$

Another informative measure is the posterior probability that the mean classification accuracy in the population does not exceed chance,

$$p = \Pr\left(\frac{\alpha}{\alpha+\beta} \le 0.5 \;\middle|\; k_{1:m}\right) \approx \#\left\{\frac{\alpha^{(\tau)}}{(\alpha^{(\tau)}+\beta^{(\tau)}} \le 0.5\right\},$$

which we refer to as the (posterior) infraliminal probability of the classifier. The symbol $\#\{\cdot\}$ denotes a count of samples.

When we are interested in the classification accuracies of individual subjects, we wish to infer on $p(\pi_j \mid k_{1:m})$. This expression fully characterizes our posterior uncertainty about the true classification accuracy in subject j. Given a pair of samples $\alpha^{(\tau)}, \beta^{(\tau)}$, we can obtain samples from these posterior distributions simply by drawing from

$$\mathrm{Beta}\left(\pi_j^{(\tau)} \;\middle|\; \alpha^{(\tau)} + k_j,\ \beta^{(\tau)} + n_j - k_j\right).$$

This can be derived by relating the full conditional $p(\pi_j \mid \alpha, \beta, \pi_{1:j-1, j+1:m}, k_{1:m})$ to the closed-form posterior in (3) (see main text; cf. Gelman et al., 2003).

In order to infer on the performance that may be expected in a new subject from the same population, we are interested in the posterior predictive density,

$$p(\tilde{\pi} \mid k_{1:m}),$$

in which $\tilde{\pi}$ denotes the classification accuracy in a new subject drawn from the same population as the existing group of subjects with latent accuracies $\pi_1, \ldots, \pi_m$.[9] Unlike the posterior on $\alpha/(\alpha+\beta)$,

9. As noted before, the term 'posterior predictive density' is sometimes exclusively used for densities over variables that are unobserved but observable in principle. Here, we use the term to refer to the posterior density of any unobserved variable, whether observable in principle (such as $\tilde{k}$) or not (such as $\tilde{\pi}$).

the posterior predictive density on $\tilde{\pi}$ reflects both the mean and the variance of the performance achieved in the population.[10]

In order to derive an expression for the posterior predictive distribution in closed form, one would need to integrate out the population parameters α and β,

$$p(\tilde{\pi} \mid k_{1:m}) = \iint p(\tilde{\pi} \mid \alpha, \beta)\, p(\alpha, \beta \mid k_{1:m})\, \mathrm{d}\alpha\, \mathrm{d}\beta,$$

which is analytically intractable. However, the integral shows that values can be drawn from the posterior predictive density on $\tilde{\pi}$ using a single ancestral-sampling step. Specifically, within each iteration τ, the current samples $\alpha^{(\tau)}$ and $\beta^{(\tau)}$ can be used to obtain a new sample $\tilde{\pi}^{(\tau)}$ by drawing from

$$\mathrm{Beta}\left(\tilde{\pi}^{(\tau)} \;\middle|\; \alpha^{(\tau)}, \beta^{(\tau)}\right).$$

Once a number of samples from $p(\tilde{\pi} \mid k_{1:m})$ have been obtained, summarizing posterior inferences becomes straightforward, for example, by reporting

$$p(\tilde{\pi} \leq 0.5) \approx \#\{\pi^{(\tau)} \leq 0.5\},$$

which represents the probability that the classifier, when applied to a new subject from the same population, will not perform better than chance.

Appendix B. Bivariate Normal Prior

In order to illustrate the flexibility offered by the bivariate Normal density on ρ, we derive $p(\pi \mid \mu, \Sigma)$ in closed form and then compute the bivariate density on a two-dimensional grid. We begin by noting that

$$p_\pi(\pi \mid \mu, \Sigma) = p_\rho(\sigma^{-1}(\pi) \mid \mu, \Sigma) \left| \frac{d\sigma}{d\rho} \right|^{-1},$$

where we have added indices to p_π and p_ρ to disambiguate between the two densities, and where σ^{-1} denotes the logit transform. The Jacobian is

$$\frac{d\sigma}{d\rho} = \begin{pmatrix} \sigma'(\rho_1) & 0 \\ 0 & \sigma'(\rho_2) \end{pmatrix},$$

in which σ' represents the first derivative of the sigmoid transform. From this, we obtain the inverse determinant of the Jacobian as

$$\left| \frac{d\sigma}{d\rho} \right|^{-1} = \frac{1}{\sigma'(\rho_1)\,\sigma'(\rho_2)}.$$

Thus, the conditional bivariate density $p_\pi(\pi \mid \mu, \Sigma)$ is given by

$$p_\pi(\pi \mid \mu, \Sigma) = \mathcal{N}_2\left(\sigma^{-1}(\pi) \;\middle|\; \mu, \Sigma\right) \frac{1}{\sigma'(\sigma^{-1}(\pi_1))\,\sigma'(\sigma^{-1}(\pi_2)))}$$

10. If data were indeed obtained from a new subject (represented in terms of $\tilde{k}$ correct predictions in $\tilde{n}$ trials), then $p(\tilde{\pi} \mid k_{1:m}, n_{1:m})$ would be used as a prior to compute the posterior $p(\tilde{\pi} \mid \tilde{k}, \tilde{n}, k_{1:m}, n_{1:m})$.

where $\sigma^{-1}(\pi) := (\sigma^{-1}(\pi_1), \sigma^{-1}(\pi_2))^{\mathrm{T}}$. When evaluating this density on a $[0,1] \times [0,1]$ grid, the normalization constant is no longer needed, so that we can use the simpler expression

$$p_\pi(\pi \mid \mu, \Sigma) \propto \frac{1}{\pi_1 \pi_2 (1-\pi_1)(1-\pi_2)} \exp\left\{ -\frac{1}{2} \left(\sigma^{-1}(\pi) - \mu\right)^{\mathrm{T}} \Sigma^{-1} \left(\sigma^{-1}(\pi) - \mu\right) \right\},$$

where we have used the fact that $\sigma'(x) = \sigma(x)(1-\sigma(x))$. This derivation allows us to illustrate the degrees of freedom of our family of prior distributions over μ and Σ.

Appendix C. Inversion of the Bivariate Normal-Binomial Model

The algorithm is initialized by drawing initial values for $\mu^{(0)}$, $\Sigma^{(0)}$, and $\rho_1^{(0)}, \ldots, \rho_m^{(0)}$ from overdispersed starting distributions. On each iteration $\tau = 1 \ldots c$, we then update one variable after another, by sampling from the full conditional distribution of one variable given the current values of all others.[11] We begin by finding a new sample $(\mu, \Sigma)^{(\tau)}$, which can be implemented in a two-step procedure (Gelman et al., 2003). We first set

$$\begin{aligned}
\kappa_m &= \kappa_0 + m \\
\nu_m &= \nu_0 + m \\
\mu_m &= \frac{\kappa_0}{\kappa_m}\mu_0 + \frac{m}{\kappa_m}\bar{\rho}^{(\tau-1)} \\
S &= \Sigma_{j=1}^{m} \left(\rho_j^{(\tau-1)} - \bar{\rho}^{(\tau-1)}\right)\left(\rho_j^{(\tau-1)} - \rho^{(\tau-1)}\right)^{\mathrm{T}} \\
\Lambda_m &= \Lambda_0 + S + \frac{\kappa_0 m}{\kappa_m}\left(\bar{\rho}^{(\tau-1)} - \mu_0\right)\left(\bar{\rho}^{(\tau-1)} - \mu_0\right)^{\mathrm{T}},
\end{aligned}$$

where $\bar{\rho}^{(\tau-1)} = \frac{1}{m}\sum_{j=1}^{m} \rho^{(\tau-1)}$, to draw

$$\Sigma^{(\tau)} \sim \text{Inv-Wishart}_{\nu_m}\left(\Sigma^{(\tau)} \,\middle|\, \Lambda_m^{-1}\right).$$

We then complete the first step by drawing

$$\mu^{(\tau)} \sim \mathcal{N}_2\left(\mu^{(\tau)} \,\middle|\, \mu_m, \Sigma^{(\tau)}/\kappa_m\right),$$

which we can use to obtain samples from the posterior mean balanced accuracy using

$$\phi^{(\tau)} := \frac{1}{2}\left(\mu_1^{(\tau)} + \mu_2^{(\tau)}\right).$$

Next, we update the bivariate variables $\rho_1, \ldots, \rho_m$. For each subject j, we wish to draw from the full conditional distribution

$$\begin{aligned}
&p\left(\rho_j^{(\tau)} \,\middle|\, k_{1:m}^{+}, k_{1:m}^{-}, \rho_{1:j-1}^{(\tau)}, \rho_{j+1:m}^{(\tau-1)}, \mu^{(\tau)}, \Sigma^{(\tau)}\right) \\
&= p\left(\rho_j^{(\tau)} \,\middle|\, k_j^{+}, k_j^{-}, \mu^{(\tau)}, \Sigma^{(\tau)}\right),
\end{aligned} \tag{18}$$

11. Here, we define one iteration as an update of all latent variables. Alternatively, one might update only one variable (or a subset of variables) per iteration, chosen randomly or systematically, as long as each variable is updated periodically.

which we have simplified by omitting all variables that are not part of the Markov blanket of ρ_j (cf. Figure 2b). Because we cannot sample from this distribution directly, we generate a candidate from a symmetric proxy distribution

$$q(\rho_j^*) = \mathcal{N}_2\left(\rho_j^* \,\middle|\, \rho_j^{(\tau-1)}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}^{\mathrm{T}}\right),$$

and then construct a Metropolis acceptance test. For this, we note that

$$\begin{aligned}
&p\left(\rho_j^* \,\middle|\, k_j^+, k_j^-, \mu^{(\tau)}, \Sigma^{(\tau)}\right) \\
&\propto \tilde{p}\left(\rho_j^* \,\middle|\, k_j^+, k_j^-, \mu^{(\tau)}, \Sigma^{(\tau)}\right) && (19)\\
&= p\left(k_j^+, k_j^- \,\middle|\, \rho_j^*, \mu^{(\tau)}, \Sigma^{(\tau)}\right)\; p\left(\rho_j^* \,\middle|\, \mu^{(\tau)}, \Sigma^{(\tau)}\right) && (20)\\
&= p\left(k_j^+, k_j^- \,\middle|\, \rho_j^*\right)\; p\left(\rho_j^* \,\middle|\, \mu^{(\tau)}, \Sigma^{(\tau)}\right) && (21)\\
&= p\left(k_j^+ \,\middle|\, \rho_{j,1}^*\right)\; p\left(k_j^- \,\middle|\, \rho_{j,2}^*\right)\; p\left(\rho_j^* \,\middle|\, \mu^{(\tau)}, \Sigma^{(\tau)}\right) && (22)\\
&= \mathrm{Bin}\left(k_j^+ \,\middle|\, \sigma(\rho_{j,1}^*)\right) \mathrm{Bin}\left(k_j^- \,\middle|\, \sigma(\rho_{j,2}^*)\right) \mathcal{N}_2\left(\rho_j^* \,\middle|\, \mu^{(\tau)}, \Sigma^{(\tau)}\right), && (23)
\end{aligned}$$

where (19) places our focus on the unnormalized density, (20) uses Bayes' theorem, (21) is based on the Markov blanket, (22) exploits the conditional independence of class-specific outcomes k_j^+ and k_j^-, and (23) relies on the model assumptions introduced in (4) and (12) (main text). We can use this result to accept the candidate sample ρ_j^* with probability

$$\min\{1, \exp(r)\},$$

where

$$\begin{aligned}
r &= \ln \frac{\tilde{p}\left(\rho_j^* \,\middle|\, k_j^+, k_j^-, \mu^{(\tau)}, \Sigma^{(\tau)}\right)}{\tilde{p}\left(\rho_j^{(\tau-1)} \,\middle|\, k_j^+, k_j^-, \mu^{(\tau)}, \Sigma^{(\tau)}\right)} \\
&= \ln \mathrm{Bin}\left(k_j^+ \,\middle|\, \sigma\left(\rho_{j,1}^*\right)\right) + \ln \mathrm{Bin}\left(k_j^- \,\middle|\, \sigma\left(\rho_{j,2}^*\right)\right) + \ln \mathcal{N}_2\left(\rho_j^* \,\middle|\, \mu^{(\tau)}, \Sigma^{(\tau)}\right) \\
&\quad - \ln \mathrm{Bin}\left(k_j^+ \,\middle|\, \sigma\left(\rho_{j,1}^{(\tau-1)}\right)\right) - \ln \mathrm{Bin}\left(k_j^- \,\middle|\, \sigma\left(\rho_{j,2}^{(\tau-1)}\right)\right) - \ln \mathcal{N}_2\left(\rho_j^{(\tau-1)} \,\middle|\, \mu^{(\tau)}, \Sigma^{(\tau)}\right).
\end{aligned}$$

We can now obtain samples from the posterior densities $p(\pi_j \mid k_{1:m}^+, k_{1:m}^-)$ for each subject j simply by sigmoid-transforming the current sample,

$$\pi_j^{(\tau)} = \sigma\left(\rho_j^{(\tau)}\right).$$

Based on this, we can obtain samples from the subject-specific balanced accuracies by setting

$$\phi_j^{(\tau)} := \frac{1}{2}\left(\pi_{j,1}^{(\tau)} + \pi_{j,2}^{(\tau)}\right).$$

Apart from using $\mu^{(\tau)}$ and $\Sigma^{(\tau)}$ to obtain samples from the posterior distributions over ρ_j, we can further use the two vectors to draw samples from the posterior predictive distribution $p(\tilde{\pi}^+_{1:m}, k^-_{1:m})$. For this we first draw

$$\tilde{\rho}^{(\tau)} \sim \mathcal{N}_2\left(\tilde{\rho}^{(\tau)} \,\middle|\, \mu^{(\tau)}, \Sigma^{(\tau)}\right),$$

and then obtain the desired sample using

$$\tilde{\pi}^{(\tau)} = \sigma\left(\tilde{\rho}^{(\tau)}\right),$$

from which we can obtain samples from the posterior predictive balanced accuracy using

$$\tilde{\phi}^{(\tau)} := \frac{1}{2}\left(\tilde{\pi}^{(\tau)}_1 + \tilde{\pi}^{(\tau)}_2\right).$$

In all above cases, we can use the obtained samples to compute approximate posterior probability intervals or Bayesian p-values.

The approximate expression for the model evidence in (16) can be obtained as follows:

$$\begin{aligned}
&\ln p(k^+_{1:m}, k^-_{1:m} \mid M_{nb}) \qquad (24)\\
&= \ln \int p(k^+_{1:m}, k^-_{1:m} \mid \rho_{1:m})\, \mathrm{d}\rho_{1:m}\\
&= \ln \left\langle p(k^+_{1:m}, k^-_{1:m} \mid \rho_{1:m}) \right\rangle_{\rho_{1:m}}\\
&= \ln \left\langle \prod_j^m p(k^+_j, k^-_j \mid \rho_j) \right\rangle_{\rho_{1:m}}\\
&= \ln \left\langle \prod_j^m p(k^+_j \mid \rho^{(1)}_j)\; p(k^-_j \mid \rho^{(2)}_j) \right\rangle_{\rho_{1:m}}\\
&\approx \ln \frac{1}{c} \sum_{\tau=1}^{c} \prod_j^m p\left(k^+_j \mid \rho^{(\tau,1)}_j\right) p\left(k^-_j \mid \rho^{(\tau,2)}_j\right)\\
&= \ln \frac{1}{c} \sum_{\tau=1}^{c} \prod_j^m \mathrm{Bin}\left(k^+_j \,\middle|\, \sigma\left(\rho^{(\tau,1)}_j\right)\right) \mathrm{Bin}\left(k^-_j \,\middle|\, \sigma\left(\rho^{(\tau,2)}_j\right)\right)
\end{aligned}$$

Appendix D. Comparison to Classical Inference

In a maximum-likelihood (ML) setting, one typically aims to obtain a point estimate for π, the true accuracy of the classifier under the binomial model.

D.1 Classical Inference for a Single Subject

In the case of a single-subject setting, the ML estimate for π is

$$\hat{\pi}_{\mathrm{ML}} = \arg\max_\pi \mathrm{Bin}(k \mid \pi, n) = \frac{k}{n},$$

which corresponds to the *sample accuracy*, that is, the number of correctly classified trials divided by the total number of trials.

Classical inference in the binomial model proceeds by asking how probable the observed value (or greater values) of the estimator are, assuming that the true accuracy π is at chance. This tests the null hypothesis $H_0 : \pi = 0.5$, yielding a p-value,

$$p = 1 - \mathcal{F}_{\text{Bin}}(k \mid 0.5),$$

where $\mathcal{F}_{\text{Bin}}(k \mid 0.5)$ is the cumulative distribution function of the binomial distribution with $\pi = 0.5$.

The practical simplicity of maximum likelihood is offset by its conceptual limitations. Specifically, using the sample accuracy k/n to estimate the true accuracy π risks overfitting. Furthermore, a point estimate for π ignores both (prior and posterior) uncertainty about classification performance.

D.2 Classical Inference in a Group Study

In a hierarchical setting, group-level inference frequently proceeds by applying a one-sample, one-tailed t-test to subject-specific sample accuracies.[12] This tests the null hypothesis that subject-specific accuracies are drawn from a distribution with a mean at chance level, using the t-statistic

$$\sqrt{m}\frac{\bar{\pi} - \pi_0}{\hat{\sigma}_{m-1}} \sim t_{m-1}, \tag{25}$$

where $\bar{\pi}$ and $\hat{\sigma}_{m-1}$ are the sample mean and sample standard deviation of subject-specific sample accuracies, π_0 is the accuracy at chance (e.g., 0.5 for binary classification), and t_{m-1} is Student's t-distribution on $m-1$ degrees of freedom.

Additionally, it is common practice to indicate the uncertainty about the population mean of the classification accuracy by reporting the 95% confidence interval

$$\left[\bar{\pi} \pm t_{0.025,m-1} \times \frac{\hat{\sigma}_{m-1}}{\sqrt{m}}\right], \tag{26}$$

where $t_{0.025,m-1}$ is a quantile from the t-distribution. It is worth emphasizing that this confidence interval has a merely illustrative purpose. This is because a central interval corresponds to a two-tailed test, whereas the t-test above is one-tailed. Since it is based on Gaussian assumptions, a one-tailed confidence interval would include the entire real line up to $+\infty$. Thus, a (two-sided) confidence-interval test actually has a false positive rate of $\alpha/2 = 0.025$. Similarly, under the null distribution, the 95% confidence interval will lie entirely below 0.5 in 2.5% of the cases. In a classical framework, one would have to call this 'significant,' in the sense of the classifier operating below chance. However, this is not the hypothesis one would typically want to test. Rather, it is more desirable to formulate a one-tailed test. In a Bayesian setting, this can be achieved by quantifying the (posterior) probability that the true accuracy is above chance.

Fundamentally, the differences between the classical procedure and the full Bayesian approach discussed earlier can best be understood by considering their respective assumptions. The distributional assumption underlying both the t-statistic in (25) and the confidence interval in (26) is that the sample mean of the subject-wise accuracies, under the null hypothesis, is normally distributed,

$$\bar{\pi} \sim \mathcal{N}\left(\mu, \frac{\sigma}{\sqrt{m}}\right), \tag{27}$$

12. It should be noted that the present manuscript focuses on those classical procedures that are widely used in application domains such as neuroimaging and brain-machine interfaces. However, it is worth noting that alternative maximum-likelihood procedures exist that eschew the normality assumption implicit in a classical t-test (e.g., Dixon, 2008, see also Discussion).

where the population standard deviation σ has been estimated by the sample standard deviation $\hat{\sigma}_{m-1}$. The corresponding graphical model is shown in Figure 1d.

This analysis is popular but suffers from two faults that are remedied by our Bayesian treatment. (For a classical mixed-effects approach, see Discussion.) First, accuracies are confined to the $[0,1]$ interval, but are modelled by a normal distribution with infinite support. Consequently, error bars based on confidence intervals (26) may well include values above 1 (see Figure 6c for an example). By contrast, the Beta distribution used in the Bayesian approach has the desired $[0,1]$ support and thus represents a more natural candidate distribution.[13]

Second, the random-effects group analysis under (27) does not acknowledge within-subject estimation uncertainty and only provides a summary-statistics approximation to full mixed-effects inference. More specifically, the model is based on subject-wise sample accuracies π_j as the units of observation, rather than using the number of correctly classified trials k to infer on the accuracy in each subject. Put differently, the model assumes that subject-wise accuracies have all been estimated with infinite precision. But the precision is finite, and it varies both with the number of observed trials n_j and with the sample accuracy k_j/n_j. (This can be seen from the expression for the variance of a Bernoulli variable, which is largest at the centre of its support.) In summary, classifier performance cannot be observed directly; it must be inferred. While the classical model above does allow for inference on random-effects (between-subjects) variability, it does not explicitly account for fixed-effects (within-subject) uncertainty. This uncertainty is only taken into account indirectly by its influence on the variance of the observed sample accuracies.

With regard to subject-specific accuracies, one might be tempted to use $\hat{\pi}_j = k_j/n_j$ as individual estimates. However, in contrast to Bayesian inference on subject-specific accuracies (see Section 2.1), individual sample accuracies do not take into account the moderating influence provided by knowledge about the group (i.e., 'shrinkage'). An effectively similar outcome is found in classical inference using the James-Stein estimator (James and Stein, 1961, see Appendix E). All of these conceptual differences can be illustrated best using synthetic and empirical data, as described in Section 3.

Appendix E. Classical Shrinkage Using the James-Stein Estimator

When inferring on subject-specific accuracies π_j, the beta-binomial model uses data from the entire group to inform inferences in individual subjects. Effectively, subject-specific posteriors are 'shrunk' to the population mean. This is in contrast to using sample accuracies $\hat{\pi} = k_j/n_j$ as individual estimates. In classical inference, a similar shrinkage effect can be achieved using the positive-part James-Stein estimator (James and Stein, 1961). It is given by

$$\hat{\pi}^{\mathrm{JS}}_{1:m} = (1-\xi)\bar{\hat{\pi}}_{1:m} + \xi\hat{\pi}_{1:m}$$
$$\xi = \left(1 - \frac{(m-2)\,\hat{\sigma}^2_m(\hat{\pi}_{1:m})}{\|\hat{\pi}_{1:m} - \bar{\hat{\pi}}_{1:m}\|^2_2}\right)^+$$

where $\hat{\pi}_{1:m} = (k_j/n_j)_{1:m}$ is a vector of sample accuracies, $\bar{\hat{\pi}}_{1:m}$ is its sample average, and $\hat{\sigma}^2_m$ denotes the population standard deviation. The weighing factor ξ balances the influence of the data ($\hat{\pi}_j$ for a given subject j) and the population ($\bar{\hat{\pi}}_{1:m}$) on the estimate.

13. A classical approach to obtaining more reasonable confidence intervals would be to apply a logit transform or a z-transform to sample accuracies and then compute confidence intervals in the space of log odds or z-scores.

References

A. Agresti and D. B. Hitchcock. Bayesian inference for categorical data analysis. *Statistical Methods and Applications*, 14(3):297–330, Dec. 2005. ISSN 1618-2510. doi: 10.1007/s10260-005-0121-y.

R. Akbani, S. Kwek, and N. Japkowicz. Applying support vector machines to imbalanced datasets. In *Machine Learning: ECML 2004*, pages 39–50. 2004.

J. H. Albert. Empirical Bayes estimation of a set of binomial probabilities. *Journal of Statistical Computation and Simulation*, 20(2):129–144, 1984. ISSN 0094-9655.

J. H. Albert and A. K. Gupta. Estimation in contingency tables using prior information. *Journal of the Royal Statistical Society. Series B (Methodological)*, 45(1):60–69, Jan. 1983. ISSN 00359246.

T. Bayes and R. Price. An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society*, 53:370–418, 1763. ISSN 0261-0523. doi: doi:10.1098/rstl.1763.0053.

M. Beal. *Variational Algorithms for Approximate Bayesian Inference*. Ph.D. thesis, University College London, United Kingdom, 2003.

C. F. Beckmann, M. Jenkinson, and S. M. Smith. General multilevel linear modeling for group analysis in fMRI. *NeuroImage*, 20(2):1052–1063, Oct. 2003. ISSN 1053-8119. doi: 10.1016/S1053-8119(03)00435-X.

T. E. J. Behrens, M. W. Woolrich, M. E. Walton, and M. F. S. Rushworth. Learning the value of information in an uncertain world. *Nature Neuroscience*, 10:1214–1221, 2007.

D. A. Berry and R. Christensen. Empirical Bayes estimation of a binomial parameter via mixtures of Dirichlet processes. *The Annals of Statistics*, 7(3):558–568, May 1979. ISSN 0090-5364. doi: 10.1214/aos/1176344677.

C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer New York., 2007.

L. Breiman and P. Spector. Submodel selection and evaluation in regression. The x-random case. *International Statistical Review/Revue Internationale de Statistique*, page 291–319, 1992.

K. H. Brodersen, C. S. Ong, K. E. Stephan, and J. M. Buhmann. The balanced accuracy and its posterior distribution. In *Proceedings of the 20th International Conference on Pattern Recognition*, pages 3121–3124. IEEE Computer Society, 2010a. ISBN 1051-4651. doi: 10.1109/ICPR.2010.764.

K. H. Brodersen, C. S. Ong, K. E. Stephan, and J. M. Buhmann. The binormal assumption on precision-recall curves. In *Proceedings of the 20th International Conference on Pattern Recognition*, pages 4263–4266. IEEE Computer Society, 2010b. ISBN 1051-4651. doi: 10.1109/ICPR.2010.1036.

C.-C. Chang and C.-J. Lin. LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3):27:1—27:27, 2011.

N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16(3):321–357, 2002.

G. F. Cooper and E. Herskovits. A Bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 9(4):309–347, Oct. 1992. ISSN 0885-6125. doi: 10.1007/BF00994110.

G. V. Cormack. Email spam filtering: a systematic review. *Foundations and Trends in Information Retrieval*, 1(4):335–455, Apr. 2008. ISSN 1554-0669. doi: 10.1561/1500000006.

D. D. Cox and R. L. Savoy. Functional magnetic resonance imaging (fMRI) "brain reading": Detecting and classifying distributed patterns of fMRI activity in human visual cortex. *NeuroImage*, 19(2):261–270, 2003. ISSN 1053-8119. doi: 10.1016/S1053-8119(03)00049-1.

J. J. Deely and D. V. Lindley. Bayes empirical Bayes. *Journal of the American Statistical Association*, 76(376):833–841, Dec. 1981. ISSN 01621459. doi: 10.2307/2287578.

O. Demirci, V. P. Clark, V. A. Magnotta, N. C. Andreasen, J. Lauriello, K. A. Kiehl, G. D. Pearlson, and V. D. Calhoun. A review of challenges in the use of fMRI for disease classification / characterization and a projection pursuit application from a multi-site fMRI schizophrenia study. *Brain Imaging and Behavior*, 2(3):207–226, Aug. 2008. ISSN 1931-7557. doi: 10.1007/s11682-008-9028-1.

P. Dixon. Models of accuracy in repeated-measures designs. *Journal of Memory and Language*, 59 (4):447–456, Nov. 2008. ISSN 0749-596X. doi: 10.1016/j.jml.2007.11.004.

B. Efron and C. Morris. Limiting the risk of Bayes and empirical Bayes estimators - part I: the Bayes case. *Journal of the American Statistical Association*, pages 807–815, 1971.

B. Efron and C. Morris. Limiting the risk of Bayes and empirical Bayes estimators – part II: the empirical Bayes case. *Journal of the American Statistical Association*, page 130–139, 1972.

P. J. Everson and E. T. Bradlow. Bayesian inference for the beta-binomial distribution via polynomial expansions. *Journal of Computational and Graphical Statistics*, 11(1):202–207, Mar. 2002. ISSN 10618600.

S. Fazli, M. Danoczy, J. Schelldorfer, and K.-R. Müller. L1-penalized linear mixed-effects models for high dimensional data with application to BCI. *NeuroImage*, 56(4):2100–2108, June 2011. ISSN 1053-8119. doi: 16/j.neuroimage.2011.03.061.

K. J. Friston, A. P. Holmes, K. J. Worsley, J. P. Poline, C. D. Frith, and R. S. J. Frackowiak. Statistical parametric maps in functional imaging: A general linear approach. *Human Brain Mapping*, 2(4):189–210, 1995.

K. J. Friston, K. E. Stephan, T. E. Lund, A. Morcom, and S. Kiebel. Mixed-effects and fMRI studies. *NeuroImage*, 24(1):244–252, Jan. 2005. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2004.08.055.

F. Galton. Regression towards mediocrity in hereditary stature. *The Journal of the Anthropological Institute of Great Britain and Ireland*, 15:246–263, 1886.

A. Gelfand and A. Smith. Sampling-based approaches to computing marginal densities. *Journal of the American Statistical Association*, 85(410):398–409, 1990.

A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis*. Chapman and Hall/CRC, 2 edition, July 2003. ISBN 9781584883883.

S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 2008, vol. 30, p. 712-727*, 6(1):721–741, 1984.

H. Goldstein. *Multilevel Statistical Models*, volume 847. Wiley, 2010.

I. J. Good. On the estimation of small frequencies in contingency tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 18(1):113–124, 1956. ISSN 0035-9246.

B. S. Griffin and R. G. Krutchkoff. An empirical Bayes estimator for P[success] in the binomial distribution. *The Indian Journal of Statistics, Series B*, 33(3/4):217–224, Dec. 1971. ISSN 05815738.

M. G. Gustafsson, M. Wallman, U. Wickenberg Bolin, H. Göransson, M. Fryknäs, C. R. Andersson, and A. Isaksson. Improving Bayesian credibility intervals for classifier error rates using maximum entropy empirical priors. *Artificial Intelligence in Medicine*, 49(2):93–104, June 2010. ISSN 0933-3657. doi: 10.1016/j.artmed.2010.02.004.

S. A. Harrison and F. Tong. Decoding reveals the contents of visual working memory in early visual areas. *Nature*, 458(7238):632–635, 2009. ISSN 0028-0836. doi: 10.1038/nature07832.

W. James and C. Stein. Estimation with quadratic loss. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*, page 361, 1961.

N. Japkowicz and S. Stephen. The class imbalance problem: A systematic study. *Intelligent Data Analysis*, 6(5):429–449, 2002.

F. V. Jensen and T. D. Nielsen. *Bayesian Networks and Decision Graphs*. Springer, 2007. ISBN 9780387682815.

R. E. Kass and A. E. Raftery. Bayes factors. *Journal of the American Statistical Association*, 90 (430):773–795, 1995. ISSN 01621459.

A. Knops, B. Thirion, E. M. Hubbard, V. Michel, and S. Dehaene. Recruitment of an area involved in eye movements during mental arithmetic. *Science (New York, N.Y.)*, 324(5934):1583–1585, May 2009. ISSN 1095-9203. doi: 10.1126/science.1171599.

R. Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *International Joint Conference on Artificial Intelligence*, volume 14, pages 1137–1145. Lawrence Erlbaum Associates Ltd., 1995.

I. Krajbich, C. Camerer, J. Ledyard, and A. Rangel. Using neural measures of economic value to solve the public goods free-rider problem. *Science (New York, N.Y.)*, 326(5952):596–599, Oct. 2009. ISSN 1095-9203. doi: 10.1126/science.1177302.

J. Langford. Tutorial on practical prediction theory for classification. *Journal of Machine Learning Research*, 6:273–306, 2005.

P. S. Laplace. *Memoire sur la probabilité des causes par les évènemens*. De l'Imprimerie Royale, 1774.

J. C. Lee and D. J. Sabavala. Bayesian estimation and prediction for the beta-binomial model. *Journal of Business & Economic Statistics*, 5(3):357–367, July 1987. ISSN 07350015. doi: 10.2307/1391611.

T. Leonard. Bayesian methods for binomial data. *Biometrika*, 59(3):581–589, Dec. 1972. doi: 10.1093/biomet/59.3.581.

D. J. C. MacKay. Bayesian interpolation. *Neural Computation*, 4:415–447, 1992.

D. Madigan and A. E. Raftery. Model selection and accounting for model uncertainty in graphical models using Occam's window. *Journal of the American Statistical Association*, 89(428):1535–1546, Dec. 1994. ISSN 0162-1459. doi: 10.2307/2291017.

D. Madigan and J. C. York. Bayesian methods for estimation of the size of a closed population. *Biometrika*, 84(1):19 –31, 1997. doi: 10.1093/biomet/84.1.19.

D. Madigan, A. E. Raftery, C. Volinsky, and J. Hoeting. Bayesian model averaging. In *Proceedings of the AAAI Workshop on Integrating Multiple Learned Models, Portland, OR*, pages 77–83, 1996.

N. Metropolis and S. Ulam. The Monte Carlo method. *Journal of American Statistical Association*, 44:335–341, 1949.

N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087, 1953. ISSN 00219606. doi: 10.1063/1.1699114.

E. Olivetti, S. Veeramachaneni, and E. Nowakowska. Bayesian hypothesis testing for pattern discrimination in brain decoding. *Pattern Recognition*, 45(6):2075–2084, June 2012. ISSN 0031-3203. doi: 10.1016/j.patcog.2011.04.025.

E. S. Pearson. Bayes' theorem, examined in the light of experimental sampling. *Biometrika*, 17 (3/4):388–442, 1925. ISSN 0006-3444.

W. D. Penny, K. E. Stephan, A. Mechelli, and K. J. Friston. Comparing dynamic causal models. *NeuroImage*, 22(3):1157–1172, 2004.

M. A. Pitt and I. J. Myung. When a good fit can be bad. *Trends in Cognitive Sciences*, 6(10), 2002.

C. P. Robert. *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation*. Springer, 2007. ISBN 9780387715988.

J. Schelldorfer, P. Bühlmann, and S. V. De Geer. Estimation for high-dimensional linear mixed-effects models using l1-penalization. *Scandinavian Journal of Statistics*, 38(2):197–214, 2011. ISSN 1467-9469. doi: 10.1111/j.1467-9469.2011.00740.x. URL `http://onlinelibrary.wiley.com/doi/10.1111/j.1467-9469.2011.00740.x/abstract`.

A. Schurger, F. Pereira, A. Treisman, and J. D. Cohen. Reproducibility distinguishes conscious from nonconscious neural representations. *Science*, 327(5961):97–99, Jan. 2010. doi: 10.1126/science.1180029.

R. Sitaram, N. Weiskopf, A. Caria, R. Veit, M. Erb, and N. Birbaumer. fMRI brain-computer interfaces: A tutorial on methods and applications. *Signal Processing Magazine, IEEE*, 25(1): 95–106, 2008. ISSN 1053-5888. doi: 10.1109/MSP.2008.4408446.

J. G. Skellam. A probability distribution derived from the binomial distribution by regarding the probability of success as variable between the sets of trials. *Journal of the Royal Statistical Society. Series B (Methodological)*, 10(2):257–261, 1948. ISSN 0035-9246.

C. Stein. Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proceedings of the Third Berkeley symposium on mathematical statistics and probability*, volume 1, page 197–206, 1956.

K. E. Stephan, W. D. Penny, J. Daunizeau, R. J. Moran, and K. J. Friston. Bayesian model selection for group studies. *NeuroImage*, 46(4):1004–1017, July 2009. ISSN 1095-9572. doi: 10.1016/j.neuroimage.2009.03.025.

D. R. Velez, B. C. White, A. A. Motsinger, W. S. Bush, M. D. Ritchie, S. M. Williams, and J. H. Moore. A balanced accuracy function for epistasis modeling in imbalanced datasets using multifactor dimensionality reduction. *Genetic Epidemiology*, 31(4):306–315, May 2007. ISSN 0741-0395. doi: 10.1002/gepi.20211.

U. Wickenberg-Bolin, H. Goransson, M. Fryknas, M. Gustafsson, and A. Isaksson. Improved variance estimation of classification performance via reduction of bias caused by small sample size. *BMC Bioinformatics*, 7(1):127, 2006. ISSN 1471-2105. doi: 10.1186/1471-2105-7-127.

I. A. Wood, P. M. Visscher, and K. L. Mengersen. Classification based upon gene expression data: bias and precision of error rates. *Bioinformatics*, 23(11):1363 –1370, June 2007. doi: 10.1093/bioinformatics/btm117.

D. Zhang and W. S. Lee. Learning classifiers without negative examples: A reduction approach. In *Third International Conference on Digital Information Management, 2008. ICDIM 2008*, pages 638 –643, 2008. doi: 10.1109/ICDIM.2008.4746761.