Mixed-effects inference on classification performance in group studies

Kay H Brodersen[1,2] · Justin R Chumbley[2] · Christoph Mathys[2] · Jean Daunizeau[2] · Cheng Soon Ong[1] · Joachim M Buhmann[1] · Klaas E Stephan[2,3]

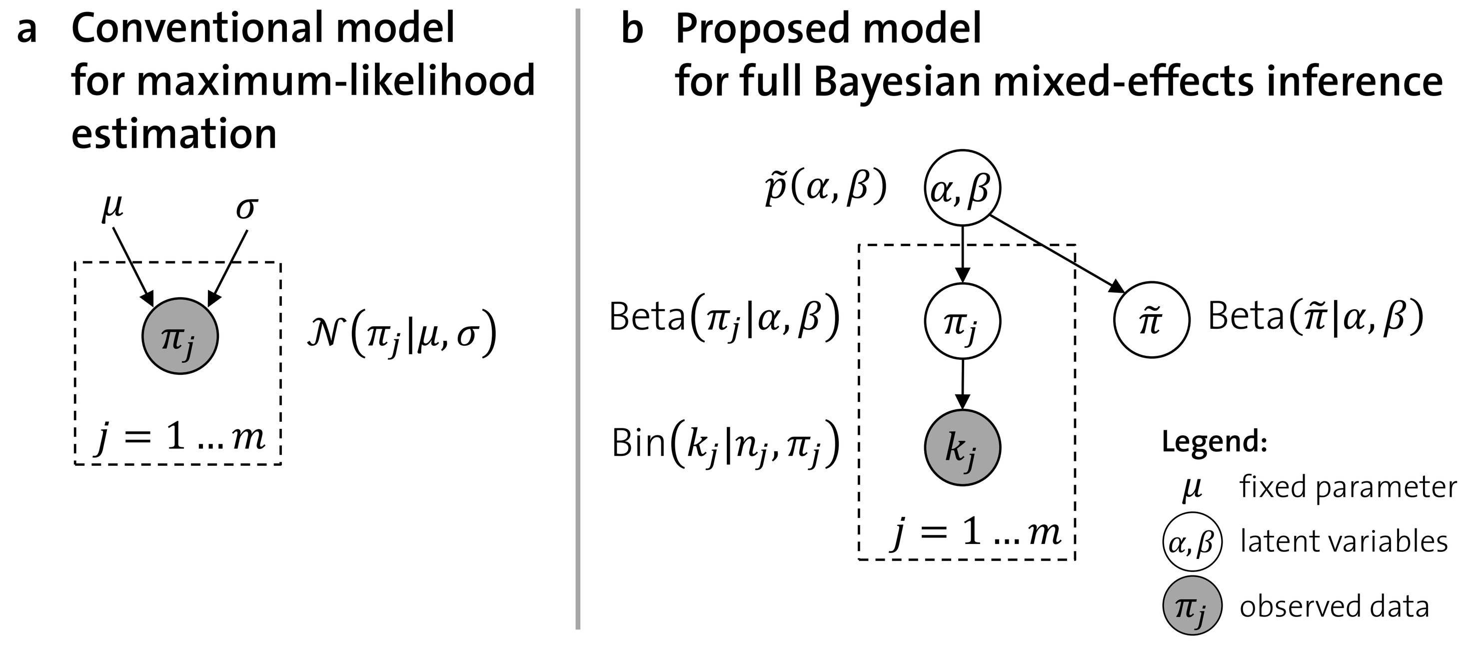
[1] Department of Computer Science, ETH Zurich, Switzerland [2] Laboratory for Social and Neural Systems Research, Department of Economics, University of Zurich, Switzerland [3] Wellcome Trust Centre for Neuroimaging, University College London, United Kingdom

1 Summary

- In neuroimaging, multivariate classification algorithms can be used to predict cognitive or pathophysiological states from measurements of distributed brain activity [1].
- The most common way of reporting how much information can be decoded from a particular observation of brain activity is based on a t-test on subject-specific sample classification accuracies.
- In certain situations, this conventional heuristic provides a reasonable approximation to fully Bayesian inference. However, there are three scenarios in which it may yield misleading results.
- Here, we introduce mixed-effects inference for classification in group studies. Our approach (i) is fully Bayesian, (ii) accounts for both within-subjects uncertainty and between-subjects variability, and (iii) is easily extensible to performance measures other than the accuracy.

2 Inference on the accuracy

Decoding studies typically proceed by training and testing a classifier on trial-wise data, using cross-validation. For each subject, this procedure results in a set of true versus predicted labels. Here, we compare different ways of analysing classifier performance, using the formalism of Bayesian networks.



Symbols: m = number of subjects π_j = latent accuracy (in subject j) n_j = number of trials k_j = number of correctly classified trials α,β = noninformative prior on population accuracy $\tilde{\pi}$ = predictive accuracy in a new subject.

a | The most common way of assessing the generalization performance of the classifier considers the sample mean of the accuracy and its standard error across subjects. However, this approach is limited in several ways (Box 5).

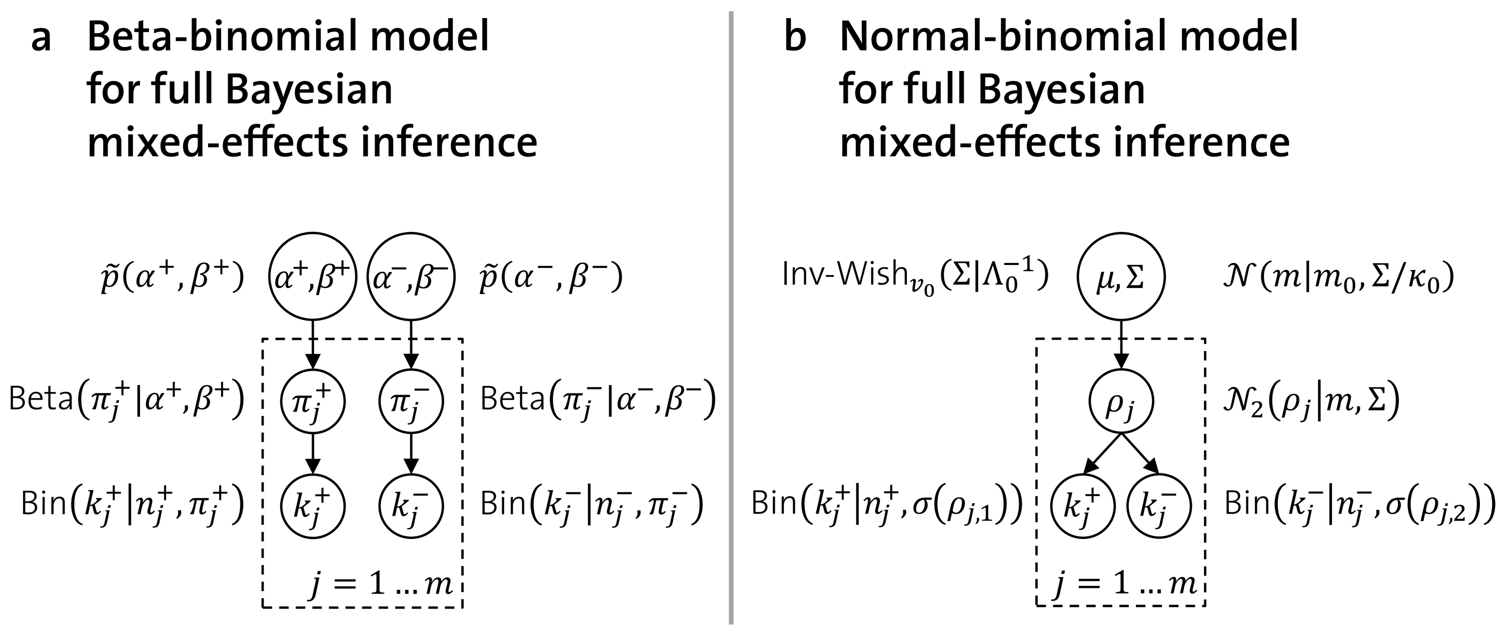
b | We introduce an approach that overcomes these limitations by accounting for both fixed-effects and random-effects components of uncertainty, with full Bayesian inference implemented using Markov Chain Monte Carlo (MCMC).

3 Inference on the balanced accuracy

The accuracy can be a misleading performance measure when a biased classifier is tested on an imbalanced dataset. The *balanced accuracy* removes this bias [2]. It is defined as the mean between sensitivity and specificity:

$$\frac{1}{2}\left(\frac{TP}{TP+FN}+\frac{TN}{TN+FP}\right)$$

We propose two models for mixed-effects inference on the balanced accuracy. We can decide between them using Bayesian model selection.



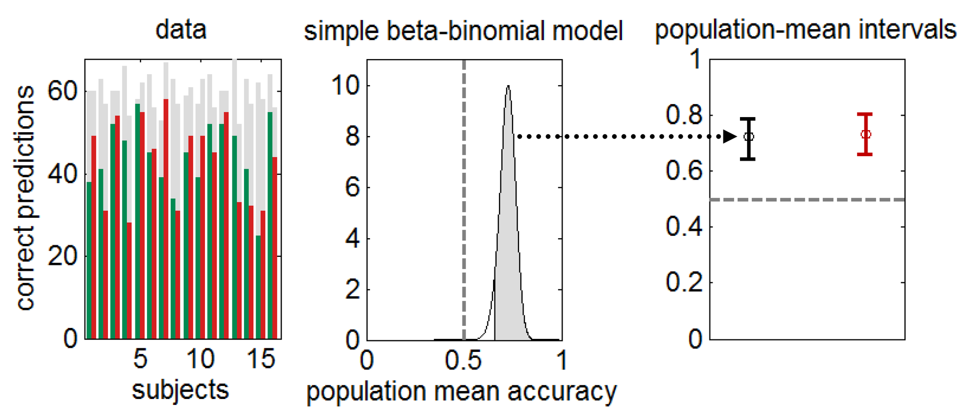
New symbols: π_j^+, π_j^- = latent accuracy on positive and negative trials, respectively (in subject j) $n_j^+, n_j^-, k_j^+, k_j^-$ defined accordingly $\rho_j \in \mathbb{R}^2$ = combined bivariate variable for the latent accuracy (in logit space) on positive and negative trials μ,Σ = mean and covariance matrix of a noninformative prior on π_j.

a | The first model assumes class-specific accuracies to be independent.

b | The second model captures dependencies between class-specific accuracies.

4 Results on empirical fMRI data

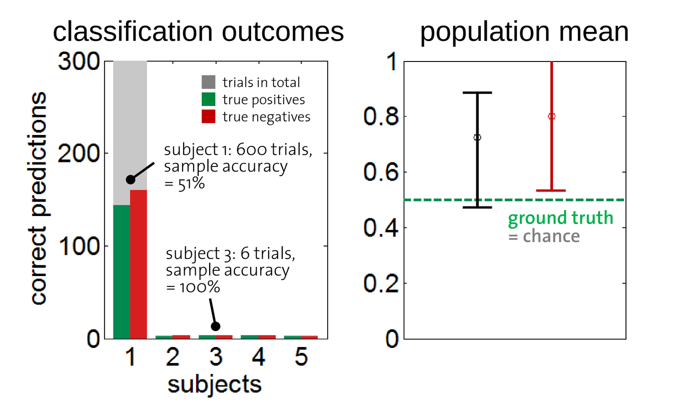
- We analysed fMRI data from 16 subjects engaged in a decision-making task with 120 trials [3]. We used a linear SVM to predict which choice had been indicated on a given trial, as indicated by the left or right index finger.
- We analysed the performance of the classifier using the model in Box 2.



In certain scenarios such as this one, a conventional confidence interval of the mean accuracy (red) provides a good approximation to the Bayesian posterior probability interval (grey/black). However, this need not be the case (Box 4).

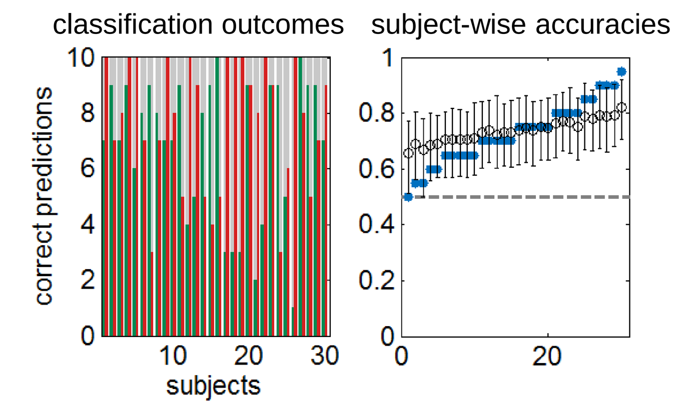
5 Results on synthetic data

Three synthetic datasets highlight the key advantages of Bayesian mixed-effects inference over conventional confidence intervals.



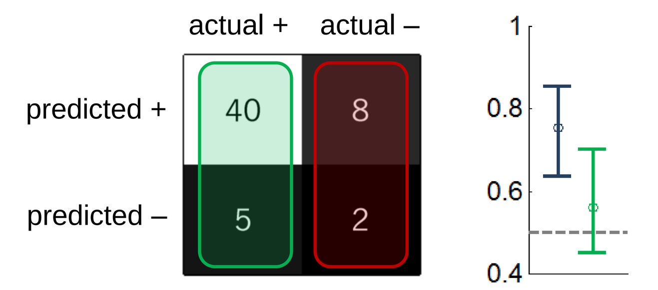
1 | Inference on the population mean

The proposed Bayesian posterior interval (**black**) removes the bias that may arise in conventional confidence intervals (red) when the group is heterogeneous and of limited size. In this example, the conventional confidence interval falsely suggests a significantly above-chance accuracy.



2 | Inference on individual accuracies

Unlike conventional sample accuracies (blue), the proposed Bayesian posterior means of individual accuracies (**black**) are informed by data from the entire group, which prevents overfitting. Individual posteriors are said to be 'shrinking to the population.'



3 | Inference on the balanced accuracy

The balanced accuracy (green) provides a more useful measure of classification performance than the accuracy (blue), especially in the context of imbalanced data. The example illustrates this in a single-subject setting.

6 Conclusions

- Bayesian mixed-effects inference for group studies provides three strengths over conventional confidence intervals and t-tests. (i) It explicitly models both within-subject and across-subjects uncertainty. (ii) Maximum-likelihood estimation is replaced by Bayesian inference on the posteriors, enabling regularization of the estimation problem, model selection, and conclusions in terms of probability statements. (iii) The approach can be used with various performance measures, such as the balanced accuracy.
- In certain situations, conventional heuristics approximate fully Bayesian inference to a reasonable degree. However, there are several scenarios in which conventional inference may give misleading results. We envisage that our approach will improve the precision and interpretability of statistical inference in future decoding studies.

Acknolwedgements

This study was funded by the University Research Priority Program 'Foundations of Human Social Behaviour' at the University of Zurich (KHB, KES), the SystemsX.ch project NEUROCHOICE (KHB, KES), and the NCCR 'Neural Plasticity' (KES).

References

1. Haynes, J. & Rees, G., 2006. Decoding mental states from brain activity in humans. *Nature Reviews Neuroscience*, 7(7), 523-534.
2. Brodersen, K.H. et al., 2010. The The balanced accuracy and its posterior distribution. *ICPR*, 3121-3124.
3. Behrens, T.E.J. et al., 2007. Learning the value of information in an uncertain world. *Nature Neuroscience*, 10, pp.1214-1221.