NeuroImage 76 (2013) 345–361

Contents lists available at SciVerse ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/ynimg

Variational Bayesian mixed-effects inference for classification studies

Kay H. Brodersen [a,b,c,*], Jean Daunizeau [c,d], Christoph Mathys [a,c], Justin R. Chumbley [c], Joachim M. Buhmann [b], Klaas E. Stephan [a,c,e]

[a] Translational Neuromodeling Unit (TNU), Institute for Biomedical Engineering, University of Zurich & ETH Zurich, Switzerland
[b] Machine Learning Laboratory, Department of Computer Science, ETH Zurich, Switzerland
[c] Laboratory for Social and Neural Systems Research (SNS), Department of Economics, University of Zurich, Switzerland
[d] Institut du Cerveau et de la Moelle Épinière (ICM), Hôpital Pitié Salpêtrière, Paris, France
[e] Wellcome Trust Centre for Neuroimaging, University College London, UK

ARTICLE INFO

Article history:
Accepted 9 March 2013
Available online 16 March 2013

Keywords:
Variational Bayes
Fixed effects
Random effects
Normal-binomial
Balanced accuracy
Bayesian inference
Group studies

ABSTRACT

Multivariate classification algorithms are powerful tools for predicting cognitive or pathophysiological states from neuroimaging data. Assessing the utility of a classifier in application domains such as cognitive neuroscience, brain–computer interfaces, or clinical diagnostics necessitates inference on classification performance at more than one level, i.e., both in individual subjects and in the population from which these subjects were sampled. Such inference requires models that explicitly account for both fixed-effects (within-subjects) and random-effects (between-subjects) variance components. While models of this sort are standard in mass-univariate analyses of fMRI data, they have not yet received much attention in multivariate classification studies of neuroimaging data, presumably because of the high computational costs they entail. This paper extends a recently developed hierarchical model for mixed-effects inference in multivariate classification studies and introduces an efficient variational Bayes approach to inference. Using both synthetic and empirical fMRI data, we show that this approach is equally simple to use as, yet more powerful than, a conventional *t*-test on subject-specific sample accuracies, and computationally much more efficient than previous sampling algorithms and permutation tests. Our approach is independent of the type of underlying classifier and thus widely applicable. The present framework may help establish mixed-effects inference as a future standard for classification group analyses.

Introduction

Multivariate classification algorithms have emerged from the field of machine learning as powerful tools for predicting cognitive or pathophysiological states from neuroimaging data (Haynes and Rees, 2006). Classifiers are based on decoding models that differ in two ways from conventional mass-univariate encoding analyses based on the general linear model (GLM; Friston et al., 1995). First, multivariate approaches explicitly account for dependencies among voxels. Second, they reverse the direction of inference, predicting a contextual variable from brain activity (decoding) rather than the other way around (encoding). There are three related areas of application in which these two characteristics have sparked most interest.

In cognitive neuroscience, and in particular neuroimaging, classifiers have been employed to decode subject-specific cognitive or perceptual states from multivariate measures of brain activity, such as those obtained by fMRI (Brodersen et al., 2012b; Cox and Savoy, 2003; Haynes and Rees, 2006; Norman et al., 2006; Tong and Pratte, 2012). A second area is the design of brain–machine interfaces which aim at decoding subjective cognitive states (e.g., intentions or decisions) from trial-wise measurements of neuronal activity in individual subjects (Blankertz et al., 2011; Sitaram et al., 2008). A third important domain concerns clinical applications that explore the utility of multivariate decoding approaches for diagnostic purposes (Davatzikos et al., 2008; Klöppel et al., 2008, 2012; Marquand et al., 2010). Recently, decoding models have also been integrated with biophysical models of brain function, such as dynamic causal models (Friston et al., 2003), to afford mechanistically interpretable classifications (Brodersen et al., 2011a,b).

Many applications of multivariate classification operate on data with a two-level hierarchical structure. Consider, for example, a study in which a classification algorithm is used to decode from fMRI data whether a subject chose option A or B on each of *n* experimental repetitions or trials. This analysis gives rise to *n* estimated labels (representing which choice the classifier predicted on each trial) and *n* true labels (indicating which option was truly chosen). Comparing predicted to true labels yields a sequence of classification *outcomes* (indicating for each trial whether the prediction was correct or incorrect). Repeating this analysis for each member of a group of *m* subjects yields the typical two-level structure (*m* subjects times *n* trials each) that is illustrated in Fig. 1; for a concrete example see Figs. 7a,e. A two-level structure underlies virtually all trial-by-trial decoding studies (see, among many others,

* Corresponding author at: Translational Neuromodeling Unit (TNU), Institute for Biomedical Engineering, University of Zurich & ETH Zurich, Wilfriedstrasse 6, CH 8032 Zurich, Switzerland.
E-mail address: brodersen@biomed.ee.ethz.ch (K.H. Brodersen).

1053-8119/$ – see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.neuroimage.2013.03.008

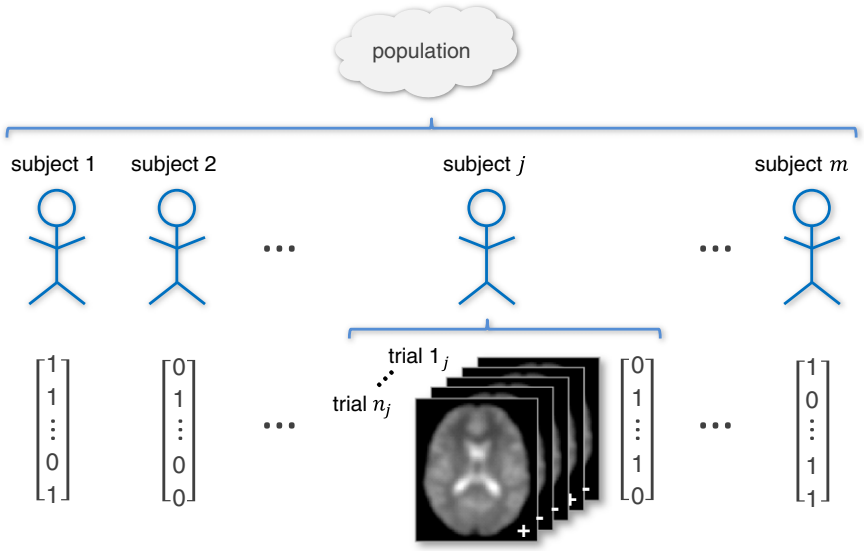

Fig. 1. Overview of the outcomes generated by a classification group study. In a trial-by-trial classification analysis, a classifier is trained and tested, separately for each subject, to predict a binary label (+ or −) from trial-wise correlates of brain activity. This constitutes a hierarchical design. The first level concerns trial-wise classification outcomes (where 1 and 0 represent correctly and incorrectly classified trials) that are drawn from latent subject-specific classification accuracies. The second level concerns subject-specific accuracies themselves, which are drawn from a population distribution. When evaluating the performance of a classification algorithm, we are interested in inference on subject-specific accuracies and on the population accuracy itself.

Brodersen et al., 2012b; Chadwick et al., 2010; Harrison and Tong, 2009; Johnson et al., 2009; Krajbich et al., 2009). The same two-level structure often applies to subject-by-subject classification studies (e.g., decoding a diagnostic state or predicting a clinical outcome), especially when subjects are partitioned into groups that are analyzed separately.

A hierarchical (or multilevel) design of this sort gives rise to the questions of what we can infer about the accuracy of the classifier in individual subjects, and what about the accuracy in the population from which the subjects were sampled. Any approach to answering these questions must provide a means of (i) *estimation* (e.g., of the accuracy itself as well as an appropriate interval that describes our uncertainty about the accuracy); and (ii) *testing* (e.g., whether the accuracy is significantly above chance). This paper is concerned with such subject-level and group-level inferences on classification accuracy for multilevel data.

The statistical evaluation of classification performance in non-hierarchical (e.g., single-subject) applications of classification has been discussed extensively in the literature (Brodersen et al., 2010a; Langford, 2005; Lemm et al., 2011; Pereira and Botvinick, 2011; Pereira et al., 2009). By contrast, relatively little attention has thus far been devoted to evaluating classification algorithms in hierarchical (i.e., group) settings (Goldstein, 2010; Olivetti et al., 2012). This is unfortunate since the field would benefit from a broadly accepted standard.

Such a standard approach to evaluating classification performance in a hierarchical setting should account for two independent sources of variability: *fixed-effects* (i.e., within-subjects) variance that results from uncertainty about the true classification accuracy in any given subject; and *random-effects* variance (i.e., between-subjects variability) that reflects the distribution of true accuracies in the population from which subjects were sampled. This distinction is crucial because classification outcomes obtained in different subjects cannot be treated as samples from the same distribution; in a hierarchical setting, each subject itself has been sampled from a population with an unknown intrinsic heterogeneity (Beckmann et al., 2003; Friston et al., 2005). Models that explicitly separate both sources of uncertainty are known as *mixed-effects* models. They are the objects of interest in this paper.

Contemporary approaches to performance evaluation in classification group studies fall into several groups.[1] One approach rests on the *pooled sample accuracy*, i.e., the number of correctly predicted trials, summed across all subjects, divided by the overall number of trials. The statistical significance of the pooled sample accuracy can be assessed using a simple classical binomial test (assuming the standard case of binary classification) that is based on the likelihood of obtaining the observed number of correct trials (or more) by chance (Langford, 2005). A less frequent variant of this analysis uses the *average sample accuracy* instead of the pooled sample accuracy (Clithero et al., 2011).

A second approach, more commonly used, is to consider *subject-specific sample accuracies* and estimate their distribution in the population. This method typically (explicitly or implicitly) uses a classical one-tailed *t*-test across subjects to assess whether the population mean accuracy is greater than what would be expected by chance (e.g., Harrison and Tong, 2009; Knops et al., 2009; Krajbich et al., 2009; Schurger et al., 2010).

In the case of single-subject studies, the first method (i.e., a binomial test on the pooled sample accuracy) is an appropriate approach. However, there are three reasons why neither method is optimal for group studies. Firstly, both of the above methods neglect the hierarchical nature of the experiment. The first method (based on the pooled sample accuracy) represents a fixed-effects approach and disregards variability across subjects. This leads to overly optimistic inferences and provides results that are only representative for the specific sample of subjects studied, not for the population they were drawn from. The second method (*t*-test on sample accuracies) does consider random effects; but it neither explicitly models the uncertainty associated with subject-specific accuracies, nor does it account for violations of homoscedasticity (i.e., the differences in variance of the data between subjects).

[1] This paper focuses on *parametric* models for performance evaluation. While *non-parametric* methods are available (e.g., based on permutation tests), these methods can be very time-consuming in hierarchical settings and are not considered in detail here (see e.g. Hassabis et al., 2009; Just et al., 2010; Pereira and Botvinick, 2011; Pereira et al., 2009; Stelzer et al., 2013).

The second limitation of the above methods is rooted in their distributional assumptions. In the standard case of binary classification, it is reasonable to assume individual classification outcomes to follow binomial distributions (justifying the binomial test in single-subject studies). However, it is not well founded to assume that sample accuracies follow a Gaussian distribution (which, in this particular case, is the implicit assumption of a classical *t*-test on sample accuracies). This is because a Gaussian has infinite support, which means it inevitably places probability mass on values below 0% and above 100% (for an alternative, see Dixon, 2008).

A third problem, albeit not an intrinsic characteristic of the above methods, is their typical focus on classification accuracy, which is known to be a poor indicator of performance when classes are not perfectly balanced. Specifically, a classifier trained on an imbalanced dataset may acquire a bias in favor of the majority class, resulting in an overoptimistic accuracy. This motivates the use of an alternative performance measure, the *balanced accuracy*, which removes this bias from performance evaluation.

We recently proposed a solution to the three above limitations using Bayesian hierarchical models for mixed-effects inference on classification performance. In particular, we introduced the *beta-binomial* model and the *normal-binomial* model for inferring on both accuracies and balanced accuracies (Brodersen et al., 2012a). Both models use a fully Bayesian framework for mixed-effects inference, are based on natural distributional assumptions, and enable more accurate inferences than the two conventional approaches described earlier. The models are independent of the type of underlying classifier, which makes them widely applicable.

The practical utility of our models, however, has been limited by the high computational complexity of the underlying Markov chain Monte Carlo (MCMC) sampling algorithms required for model inversion (i.e., the process of passing from a prior to a posterior distribution over model parameters, given the data). MCMC is asymptotically exact; but it is also exceedingly slow, especially when performing inference in a voxel-by-voxel fashion, as is common, for example, in 'searchlight' approaches (Kriegeskorte et al., 2006; Nandy and Cordes, 2003).

In this paper, we present a variational Bayes (VB) algorithm to overcome this critical limitation.[2] Our approach has three main features. First, we present a mixed-effects model that explicitly respects the hierarchical structure of the data. Second, the model can be equally used for inference on the accuracy and the balanced accuracy. Third, our novel variational inference scheme dramatically reduces the computational complexity (i.e., runtime) compared to our previous sampling approach based on MCMC.

The paper is organized as follows. In the Theory section, we present variations of our recently developed normal-binomial model for mixed-effects inference (Brodersen et al., 2012a). These are the *univariate normal-binomial* model (for inference on the *accuracy*) and the *twofold normal-binomial* model (for inference on the *balanced accuracy*).[3] We then describe a novel VB algorithm for model inversion and compare it to an MCMC sampler. In the Applications section, we provide a set of illustrative results on both synthetic data and empirical fMRI measurements. Finally, in the Discussion, we review the key characteristics of our approach, compare it to similar models in other analysis domains, and discuss its role in future classification studies.

Theory

In a hierarchical setting, a classifier is typically used to predict a class label for each trial, where trials are further structured into sets, for instance because they were recorded from different subjects. The most common situation is binary classification, where class labels are taken from $\{+1,-1\}$, denoting 'positive' and 'negative' trials, respectively. Less common, but equally amenable to the approach presented in this paper, are multiclass settings in which trials fall into more than two classes (see Discussion).

The above situation raises three principal questions (cf. Brodersen et al., 2012a). First, can one obtain successful classification at the group level? This requires statistical inference on the mean classification accuracy in the population from which subjects were drawn. Second, do the subject-wise data permit classification in each individual? Considering each subject in isolation is statistically short-sighted, since subject-specific inference may benefit from simultaneous across-subject inference (Efron and Morris, 1971). Third, which of several possible classification algorithms should be chosen? This is typically answered by evaluating how well an algorithm's performance generalizes (to unseen data). In a Bayesian framework, this expected performance is given by the posterior predictive density of classification performance. The present section describes a variational Bayes (VB) approach to answering these questions (Fig. 2).

The univariate normal-binomial model for inference on the accuracy

Within each subject, classification outcomes can be summarized in terms of the number of correctly predicted trials, k, and the total number of trials, n. It is important to note that this summary is independent of the type of underlying classifier. This means that the model can be applied regardless of whether classification results were obtained using, for instance, logistic regression, nearest-neighbor classification, a support vector machine, or a Gaussian process classifier. Under the assumption that trial-specific predictions are conditionally independent, k follows a binomial distribution,

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

where π represents the latent (unobservable) accuracy of the classifier, $0 \le \pi \le 1$. Thus, in a group study, where the classifier has been trained and tested separately in each subject, the available data are k_j and n_j for each subject $j = 1\ldots m$.

One might be tempted to 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, using such a *pooled sample accuracy* would assume zero between-subjects variability. In other words, π would be treated as a *fixed effect* in the population. This approach would not permit inferences about the population; it would only allow for results to be reported as a case study (Friston et al., 1999).

Alternatively, one might summarize the data from each subject in terms of a subject-specific *sample accuracy*, k_j/n_j. One could then ask, using a one-tailed *t*-test, whether sample accuracies reflect a normal distribution with a mean greater than what would be expected by chance (Fig. 2a). This approach no longer treats accuracy as a fixed effect. However, it suffers from two other problems.

First, submitting subject-specific sample accuracies to a *t*-test assumes that accuracies, which are confined to the [0,1] interval, follow a normal distribution, which has infinite support. This may lead to non-interpretable results such as confidence intervals that include accuracies above 100% (or below 0%).[4]

Second, even if one were to overcome the above problem (e.g., using a logit transform), a *t*-test on sample accuracies neither explicitly accounts for within-subjects uncertainty nor for violations of homoscedasticity. This is because it uses sample accuracies as summary statistics without carrying forward the uncertainty associated with them (Mumford and

[2] The approach proposed in this paper has been implemented as open-source software for both MATLAB and R. The code can be downloaded from: http://www.translationalneuromodeling.org/software/.

[3] Note that the terms 'univariate' and 'twofold' are used to characterize the number and structure of model parameters in each subject; these differences are unrelated to the distinction between univariate and multivariate analyses.

[4] Nonparametric mixed-effects approaches make it possible to overcome this limitation; however, these are often computationally expensive and are not discussed in detail here.

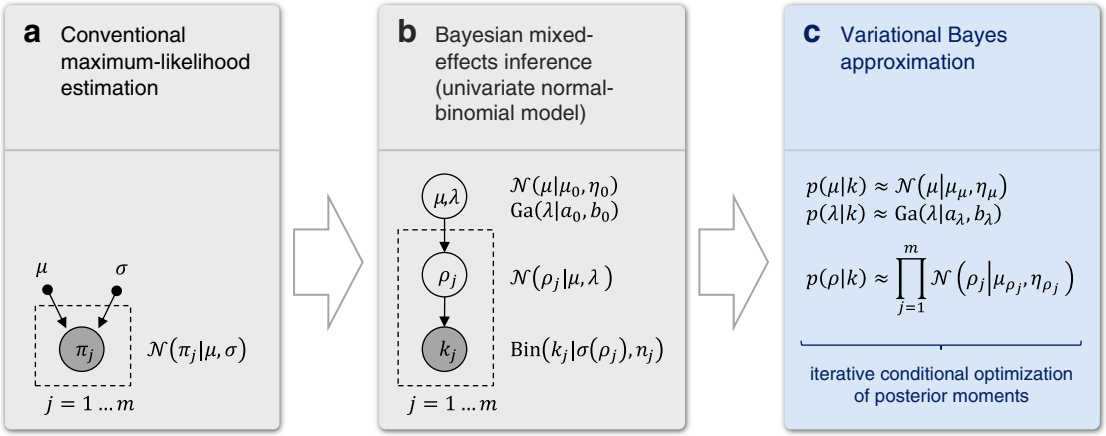

Fig. 2. Inference on classification accuracies. (a) Conventional maximum-likelihood estimation does not explicitly model within-subjects (fixed-effects) variance components and is based on an ill-justified normality assumption. It is therefore inadequate for the statistical evaluation of classification group studies. (b) The normal-binomial model respects the hierarchical structure of the study and makes natural distributional assumptions, thus enabling mixed-effects inference, which makes it suitable for group studies. The model uses the sigmoid transform $\sigma(\rho_j) := (1 + \exp(-\rho_j))^{-1}$ which turns log-odds with real support $(-\infty,\infty)$ into accuracies on the [0,1] interval. (b) Model inversion can be implemented efficiently using a variational Bayes approximation to the posterior densities of the model parameters (see Fig. 3 for details).

Nichols, 2009). For example, sample accuracies do not distinguish between an accuracy of 80% that was obtained as 80 correct out of 100 trials (i.e., an estimate with high confidence) and the same accuracy obtained as 8 out of 10 trials (i.e., an estimate with low confidence). Furthermore, no distinction regarding the confidence in the inference is being made between 80 correct out of 100 trials (i.e., high confidence) and 50 correct out of 100 trials (lower confidence, since the variance of a binomial distribution depends on its mean and becomes maximal at a mean of 0.5).

In order to explicitly capture both within-subjects (fixed-effects) and between-subjects (random-effects) variance components, we must instead use a hierarchical model in which separate levels account for different sources of variability (Fig. 2b). At the level of individual subjects, for each subject j, the number of correctly classified trials k_j is modeled as

$$p\left(k_j|\pi_j,n_j\right) = \mathrm{Bin}\left(k_j|\pi_j,n_j\right) \tag{2}$$

where π_j represents the latent classification accuracy in subject j.[5] Next, at the group level, we account for variability between subjects by modeling subject-specific accuracies as drawn from a population distribution. The *natural* parameter of the binomial density is $\ln\frac{\pi}{1-\pi}$. Thus, one possible parameterization is to assume accuracies to be logit-normally distributed and conditionally independent given the population parameters. In other words, each logit accuracy $\rho_j := \sigma^{-1}\left(\pi_j\right) := \ln\frac{\pi_j}{1-\pi_j}$ is drawn from a normal distribution. The inverse-sigmoid (or logit) transform $\sigma^{-1}(\pi_j)$ turns accuracies with support on the [0,1] interval into log-odds with support on the real line $(-\infty,+\infty)$. Thus,

$$p\left(\rho_j|\mu,\lambda\right) = N\left(\rho_j|\mu,\lambda\right) = \sqrt{\frac{\lambda}{2\pi}}\exp\left(-\frac{\lambda}{2}\left(\rho_j-\mu\right)^2\right) \tag{3}$$

where μ and λ represent the population mean and the population precision (i.e., inverse variance), respectively.

Since neuroimaging studies are typically confined to relatively small sample sizes, an adequate expression of our prior ignorance about the population parameters is critical (cf. Woolrich et al., 2004). We use a diffuse prior on μ and λ such that the posterior will be dominated by the data (for a validation of this prior, see Applications). A straightforward parameterization is to use independent conjugate densities:

$$p(\mu|\mu_0,\eta_0) = N(\mu|\mu_0,\eta_0) \tag{4}$$

$$p(\lambda|a_0,b_0) = \mathrm{Ga}(\lambda|a_0,b_0). \tag{5}$$

In the above densities, μ_0 and η_0 encode the prior mean and precision of the population mean, and a_0 and b_0 represent the shape and scale parameter,[6] respectively, that specify the prior distribution of the population precision (for an alternative, see Leonard, 1972). In summary, the univariate normal-binomial model uses a binomial distribution at the level of individual subjects and a logit-normal distribution at the group level (Fig. 2b).

In principle, inverting the above model immediately yields the desired posterior density over parameters,

$$p(\mu,\lambda,\rho|k) = \frac{\prod_{j=1}^{m}\left[\mathrm{Bin}\left(k_j|\sigma\left(\rho_j\right)\right)N\left(\rho_j\middle|\mu,\lambda\right)\right]N(\mu|\mu_0,\eta_0)\mathrm{Ga}(\lambda|a_0,b_0)}{p(k)}. \tag{6}$$

In practice, however, integrating the expression in the denominator of the above expression, which provides the normalization constant for the posterior density, is prohibitively difficult. We previously described a stochastic approximation based on MCMC algorithms; however, the practical use of these algorithms was limited by their considerable computational complexity (Brodersen et al., 2012a). Here, we propose to invert the above model using a deterministic VB approximation (Fig. 2c). This approximation is no longer asymptotically exact, but it conveys considerable computational advantages. The remainder of this section describes its derivation (see Fig. 3 for a summary).

Variational inference

The difficult problem of finding the exact posterior $p(\mu,\lambda,\rho|k)$ can be transformed into the easier problem of finding an approximate parametric posterior $q(\mu,\lambda,\rho|\delta)$ with moments (i.e., parameters) δ. (We will omit δ to simplify the notation.) Inference then reduces to finding

[5] From now on, we will omit n_j unless this introduces ambiguity.

[6] Under the Gamma parameterization used here, the prior expectation of λ is $\langle\lambda\rangle = a_0 b_0$.

a density q that minimizes a measure of dissimilarity between q and p. This can be achieved by maximizing the so-called negative free energy F of the model, a lower-bound approximation to the log model evidence, with respect to (the moments of) q. For details, see MacKay (1995), Attias (2000), Ghahramani and Beal (2001), Bishop et al. (2002), and Fox and Roberts (2012). Maximizing the negative free energy minimizes the Kullback–Leibler (KL) divergence between the approximate and the true posterior, q and p:

$$\mathrm{KL}[q||p] := \iiint q(\mu,\lambda,\rho)\ln\frac{q(\mu,\lambda,\rho)}{p(\mu,\lambda,\rho|k)}\mathrm{d}\mu\mathrm{d}\lambda\mathrm{d}\rho \tag{7}$$

$$= \iiint q(\mu,\lambda,\rho)\ln\frac{q(\mu,\lambda,\rho)}{p(k,\mu,\lambda,\rho)}\mathrm{d}\mu\mathrm{d}\lambda\mathrm{d}\rho + \ln p(k) \tag{8}$$

$$\Leftrightarrow \ln p(k) = \mathrm{KL}[q||p] + \underbrace{\left\langle \ln\frac{p(k,\mu,\lambda,\rho)}{q(\mu,\lambda,\rho)}\right\rangle_{q(\mu,\lambda,\rho)}}_{=:F(q,k)}. \tag{9}$$

This means that the log-model evidence $\ln p(k)$ can be expressed as the sum of (i) the KL-divergence between the approximate and the true posterior and (ii) the negative free energy $F(q,k)$. Because the KL-divergence cannot be negative, maximizing the negative free energy with respect to q minimizes the KL-divergence and thus results in an approximate posterior that is maximally similar to the true posterior. At the same time, maximizing the negative free energy provides a lower-bound approximation to the log-model evidence, which permits Bayesian model comparison (Bishop, 2007; Penny et al., 2004). In summary, maximizing the negative free energy $F(q,k)$ in Eq. (9) enables both inference on the posterior density over parameters and model comparison. In this paper, we are primarily interested in the posterior density.

In trying to maximize $F(q,k)$, variational calculus tells us that

$$\frac{\partial F(q,k)}{\partial q} = 0 \Rightarrow q(\mu,\lambda,\rho) \propto \exp[\underbrace{\ln p(k,\mu,\lambda,\rho)}_{\text{negative variational energy}}] \tag{10}$$

This means that the approximate posterior which maximizes the negative free energy is equal to the true posterior and thus proportional to the joint density over data and parameters[7] (with the normalization constant being given by the model evidence). In other words, the VB approach is complete in the sense that, in the absence of any other approximations, optimizing F with respect to q yields the exact posterior density and model evidence.

Mean-field approximation

To make the optimization on the l.h.s. in Eq. (10) tractable, we assume that the joint posterior over all model parameters factorizes into specific parts. Using one density for each variable,

$$q(\mu,\lambda,\rho) = q(\mu)q(\lambda)q(\rho) \tag{11}$$

the mean-field assumption turns the problem of maximizing $F(q,k)$ into the problem of deriving three expectations:

$$I_1(\mu) = \langle \ln\, p(k,\mu,\lambda,\rho)\rangle_{q(\lambda,\rho)} \tag{12}$$

$$I_2(\lambda) = \langle \ln\, p(k,\mu,\lambda,\rho)\rangle_{q(\mu,\rho)} \tag{13}$$

$$I_3(\rho) = \langle \ln\, p(k,\mu,\lambda,\rho)\rangle_{q(\mu,\lambda)}. \tag{14}$$

[7] The dependence of the joint probability in Eq. (10) on the prior (μ_0,η_0,a_0,b_0) has been omitted for brevity.

This transformation has several advantages over working with Eq. (10) directly: it makes it more likely that we can find the exact distributional form of a marginal approximate posterior (as will be the case for μ and λ); it may make the Laplace assumption more appropriate in those cases where we cannot identify a fixed form (as will be the case for ρ); and it often provides us with interpretable update equations (as will be the case, in particular, for μ and λ).

Parametric assumptions

Due to the structure of the model, the posteriors on the population parameters μ and λ are conditionally independent given the data. In addition, owing to the conjugacy of their priors, the posteriors on μ and λ follow the same distributions and do not require any additional parametric assumptions:

$$q(\mu) = N\left(\mu|\mu_\mu,\eta_\mu\right) \tag{15}$$

$$q(\lambda) = \mathrm{Ga}(\lambda|a_\lambda,b_\lambda). \tag{16}$$

Subject-specific (logit) accuracies $q \equiv (\rho_1,\ldots,\rho_m)$ are also conditionally independent given the data. This is a consequence of the fact that the posterior for each subject only depends on its Markov blanket, i.e., the subject's data and the population parameters (but not the other subject's logit accuracies). This can be seen from the fact that

$$q(\mu,\lambda,\rho) = q(\mu)\,q(\lambda)\,q(\rho) \tag{17}$$

$$= q(\mu)q(\lambda)\textstyle\prod_{j=1}^{m} q\left(\rho_j\right). \tag{18}$$

However, we do require a distributional assumption for the above subject-specific posteriors to make model inversion feasible. Here, we assume posterior subject-specific (logit) accuracies to be normally distributed:

$$q(\rho) = \textstyle\prod_{j=1}^{m} N\left(\rho_j|\mu_{\mu_j},\eta_{\rho_j}\right). \tag{19}$$

The conditional independence in Eq. (19) differs in a subtle but important way from the assumption of *unconditional* independence that is implicit in random-effects analyses on the basis of a *t*-test on subject-specific sample accuracies (see Introduction). In the case of such *t*-tests, estimation in each subject only ever uses data from that same subject. By contrast, the subject-specific posteriors in Eq. (20) borrow strength from *all* observations. This can be seen from the fact that the subject-specific posteriors $q(\rho)$ are computed with respect to the population posteriors $q(\mu)$ and $q(\lambda)$ which are themselves informed by observations from the entire group (see Eqs. (12)–(14)).

Derivation of variational densities

For each mean-field part in Eq. (11), the variational density $q(\cdot)$ can be obtained by evaluating the variational energy $I(\cdot)$, as described next. The first variational energy concerns the posterior density over the population mean μ. It is given by

$$I_1(\mu) = \langle \ln p(k,\mu,\lambda,\rho)\rangle_{q(\lambda,\rho)} \tag{20}$$

$$= \langle \ln p(k|\rho)\rangle_{q(\lambda,\rho)} + \langle \ln p(\rho|\mu,\lambda)\rangle_{q(\lambda,\rho)} + \langle \ln p(\mu,\lambda)\rangle_{q(\lambda,\rho)} \tag{21}$$

$$= \textstyle\sum_{j=1}^{m}\left\langle \ln N\left(\rho_j|\mu,\lambda\right)\right\rangle_{q(\lambda,\rho)} + \left\langle \ln\left(N(\mu|\mu_0,\eta_0)\,\mathrm{Ga}(\lambda|a_0,b_0)\right)\right\rangle_{q(\lambda,\rho)} + c \tag{22}$$

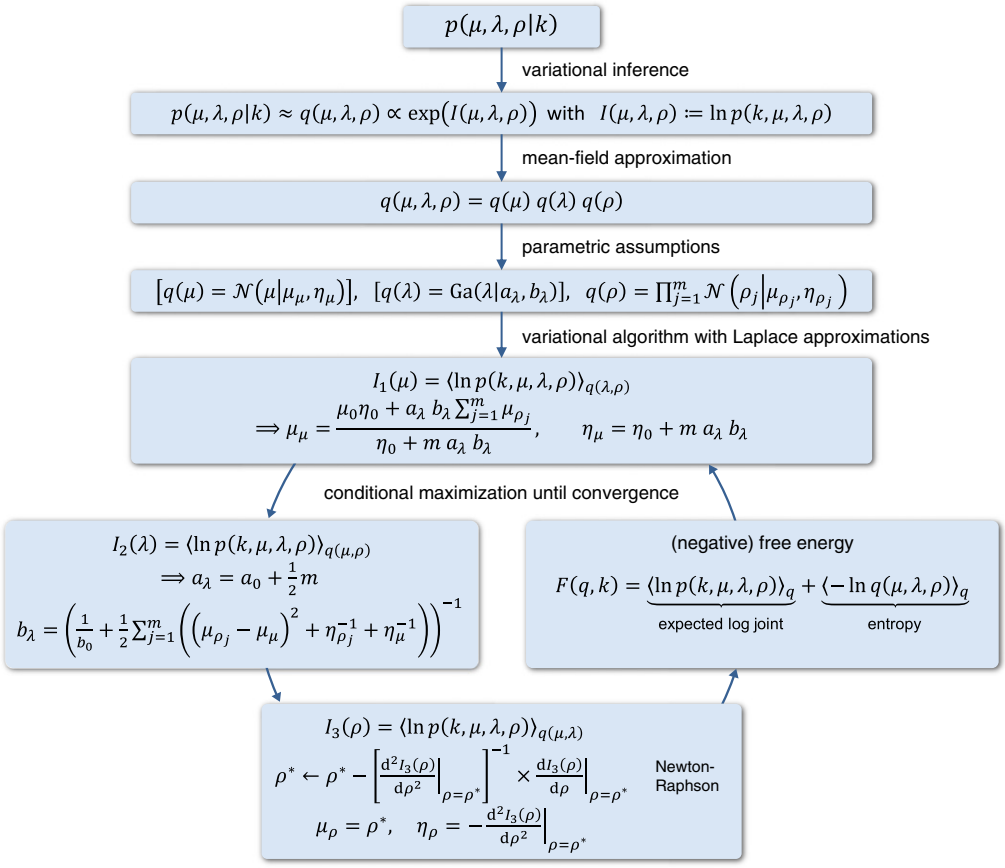

Fig. 3. Variational inversion of the univariate normal-binomial model. This schematic summarizes the individual steps involved in the variational approach to the inversion of the univariate normal-binomial model, as described in the main text (see Theory).

$$= \sum_{j=1}^{m}\left\langle \frac{1}{2}\ln\lambda - \frac{1}{2}\ln 2\pi - \frac{\lambda}{2}\left(\rho_j - \mu\right)^2\right\rangle_{q(\lambda,\rho)} + \left\langle \frac{1}{2}\ln\eta_0 - \frac{\eta_0}{2}(\mu-\mu_0)^2\right\rangle_{q(\lambda,\rho)} + c \tag{23}$$

$$= \sum_{j=1}^{m} -\frac{1}{2}\left\langle \lambda\rho_j^2 - 2\lambda\rho_j\mu + \lambda\mu^2\right\rangle_{q(\lambda,\rho)} - \frac{\eta_0}{2}(\mu-\mu_0)^2 + c \tag{24}$$

$$= -\frac{1}{2}\sum_{j=1}^{m}\left[-2\,\mu\mu_{\rho_j} + \mu^2\right]a_\lambda b_\lambda + \mu\,\eta_0\left(\mu_0 - \frac{1}{2}\mu\right) + c \tag{25}$$

$$= \mu\, a_\lambda\, b_\lambda\left(-\frac{1}{2}m\mu + \sum_{j=1}^{m}\mu_{\rho_j}\right) + \mu\,\eta_0\left(\mu_0 - \frac{1}{2}\mu\right) + c \tag{26}$$

where the symbol c is used for any expression that is constant with respect to μ.

In principle, we could proceed by optimizing the sufficient statistics of the approximate posterior. Instead, we only optimize the mean and equate the variance to the *observed information*, i.e., the negative curvature at the mode. This procedure is known as the *Laplace approximation* (or normal approximation) and implies that the negative free energy is a function simply of the posterior means (as opposed to a function of the posterior means and covariances). It is a local, rather than a global, optimization solution.

Conveniently, the Laplace approximation is typically more accurate for the conditional posterior (of one parameter given the others) than for the full posterior (of all parameters). In addition, it is computationally efficient (see Discussion) and often gives rise to interpretable update equations (see below).

Setting the first derivative to zero yields an analytical expression for the maximum,

$$\frac{\mathrm{d}I_1(\mu)}{\mathrm{d}\mu} = -\mu(\eta_0 + ma_\lambda b_\lambda) + \mu_0\eta_0 + a_\lambda b_\lambda\sum_{j=1}^{m}\mu_{\rho_j} = 0 \tag{27}$$

$$\Rightarrow \mu^* = \frac{\mu_0\eta_0 + a_\lambda b_\lambda\sum_{j=1}^{m}\mu_{\rho_j}}{\eta_0 + ma_\lambda b_\lambda}. \tag{28}$$

Having found the mode of the approximate posterior, we can use a second-order Taylor expansion to obtain closed-form approximations for its moments:

$$\mu_\mu = \mu^* \quad \text{and} \tag{29}$$

$$\eta_\mu = -\frac{\mathrm{d}I_1^2(\mu)}{\mathrm{d}\mu^2}\Big|_{\mu=\mu^*} = \eta_0 + ma_\lambda b_\lambda. \tag{30}$$

Thus, the posterior density of the population mean logit accuracy under our mean-field and Gaussian approximations is $N(\mu|\mu_\mu,\eta_\mu)$.

The use of a Laplace approximation, as we do here, often leads to interpretable update equations. In Eq. (30), for example, we can see that the posterior precision of the population mean (η_μ) is simply the sum of the prior precision (η_0) and the mean of the posterior population precision ($a_\lambda b_\lambda$), correctly weighted by the number of subjects m.

Based on the above approximation for the posterior logit accuracy, we can see that the posterior mean accuracy itself, $\xi := \sigma(\mu)$, is logit-normally distributed and can be expressed in closed form,

$$\mathrm{logit}\,N\left(\xi|\mu_\mu,\eta_\mu\right) = \frac{1}{\xi(1-\xi)}\sqrt{\frac{\eta_\mu}{2\pi}}\exp\left(-\frac{\eta_\mu}{2}\left(\sigma^{-1}(\xi) - \mu_\mu\right)^2\right) \tag{31}$$

where μ_μ and η_μ represent the posterior mean and precision, respectively, of the population mean logit accuracy.

The second variational energy concerns the population precision λ and is given by

$$I_2(\lambda) = \langle \ln p(k,\mu,\lambda,\rho)\rangle_{q(\mu,\rho)} \tag{32}$$

$$= \frac{m}{2}\ln\lambda - \frac{\lambda}{2}\sum_{j=1}^{m}\left(\left(\mu_{\rho_j}-\mu_\mu\right)^2 + \eta_{\rho_j}^{-1} + \eta_\mu^{-1}\right) + (a_0-1)\ln\lambda - \frac{\lambda}{b_0} + c \tag{33}$$

where c represents a term that is constant with respect to λ. The above expression already has the form of a log-Gamma distribution with parameters

$$a_\lambda = a_0 + \frac{1}{2}m \quad \text{and} \tag{34}$$

$$b_\lambda = \left(\frac{1}{b_0} + \frac{1}{2}\sum_{j=1}^{m}\left(\left(\mu_{\rho_j}-\mu_\mu\right)^2 + \eta_{\rho_j}^{-1} + \eta_\mu^{-1}\right)\right)^{-1}. \tag{35}$$

From this we can see that the shape parameter a_λ is a weighted sum of prior shape a_0 and data m. When viewing the second parameter as a 'rate' coefficient b_λ^{-1} (as opposed to a shape coefficient b_λ), it becomes clear that the posterior rate really is a weighted sum of: the prior rate (b_0^{-1}); the dispersion of subject-specific means; their variances ($\eta_{\rho_j}^{-1}$); and our uncertainty about the population mean (η_μ^{-1}).

The variational energy of the third partition concerns the model parameters representing subject-specific latent accuracies. This energy is given by

$$I_3(\rho) = \langle \ln p(k,\mu,\lambda,\rho)\rangle_{q(\mu,\lambda)} \tag{36}$$

$$= \sum_{j=1}^{m}\left(k_j \ln\sigma\left(\rho_j\right) + \left(n_j-k_j\right)\ln\left(1-\sigma\left(\rho_j\right)\right) - \frac{1}{2}a_\lambda b_\lambda\left(\rho_j-\mu_\mu\right)^2\right) + c. \tag{37}$$

Since an analytical expression for the maximum of this energy does not exist, we resort to an iterative Newton–Raphson scheme based on a quadratic Taylor-series approximation to the variational energy $I_3(\rho)$. For this, we begin by considering the Jacobian

$$\left(\frac{\mathrm{d}I_3(\rho)}{\mathrm{d}\rho}\right)_j = \frac{\partial I_3(\rho)}{\partial\rho_j} = k_j - n_j\sigma\left(\rho_j\right) + a_\lambda b_\lambda\left(\mu_\mu-\rho\right) \tag{38}$$

and the Hessian

$$\left(\frac{\mathrm{d}^2 I_3(\rho)}{\mathrm{d}\rho^2}\right)_{jk} = \frac{\partial^2 I_3(\rho)}{\partial\rho_j\partial\rho_k} = -\delta_{jk}\left(n_j\sigma\left(\rho_j\right)\left(1-\sigma\left(\rho_j\right)\right) + a_\lambda b_\lambda\right) \tag{39}$$

where the Kronecker delta operator δ_{jk} is 1 if $j = k$ and 0 otherwise. As noted before, the absence of off-diagonal elements in the Hessian is not based on an assumption of conditional independence of subject-specific posteriors; it is a consequence of the mean-field separation in Eq. (11). Each GN iteration performs the update

$$\rho^* \leftarrow \rho^* - \left[\frac{\mathrm{d}^2 I_3(\rho)}{\mathrm{d}\rho^2}\bigg|_{\rho=\rho^*}\right]^{-1} \times \frac{\mathrm{d}I_3(\rho)}{\mathrm{d}\rho}\bigg|_{\rho=\rho^*} \tag{40}$$

until the vector ρ^* converges, i.e., $\|\rho^*_{\text{current}} - \rho^*_{\text{previous}}\|^2 < 10^{-3}$. Using this maximum, we can use a second-order Taylor expansion (i.e., the Laplace approximation) to set the moments of the approximate posterior:

$$\mu_\rho = \rho^* \quad \text{and} \tag{41}$$

$$\eta_\rho = -\frac{\mathrm{d}^2 I_3(\rho)}{\mathrm{d}\rho^2}\bigg|_{\rho=\rho^*}. \tag{42}$$

Variational algorithm and free energy

The expressions for the three variational energies depend on one another. This circularity can be resolved by iterating over the expressions sequentially and updating the moments of each approximate marginal given the current moments of the other marginals. This approach of conditional maximization (or stepwise ascent) maximizes the (negative) free energy $F \equiv F(q,k)$ and leads to approximate marginals that are maximally similar to the exact marginals.

The free energy itself can be expressed as the sum of the expected log-joint density (over the data and the model parameters) and the Shannon entropy of the approximate posterior:

$$F = \underbrace{\langle \ln p(k,\mu,\lambda,\rho)\rangle_q}_{\text{expected log joint}} + \underbrace{\langle -\ln q(\mu,\lambda,\rho)\rangle_q}_{\text{entropy } H[q]}. \tag{43}$$

We begin by considering the expectation of the log joint w.r.t. the variational posterior:

$$\langle \ln p(k,\mu,\lambda,\rho)\rangle_q = \sum_{j=1}^{m}\Big\langle \underbrace{\Big\langle\Big\langle \ln \mathrm{Bin}\left(k_j|\sigma\left(\rho_j\right)\right) + \ln N\left(\rho_j|\mu,\lambda\right)\Big\rangle_{q(\mu)}\Big\rangle_{q(\lambda)}}_{\equiv I(\rho_j)}\Big\rangle_{q(\rho_j)} + \langle \ln N(\mu|\mu_0,\eta_{|0})\rangle_q + \langle \ln \mathrm{Ga}(\lambda|a_0,b_0)\rangle_q. \tag{44}$$

The above expression contains the variational energy of ρ_j,

$$I\left(\rho_j\right) = \ln \mathrm{Bin}\left(k_j|\sigma\left(\rho_j\right)\right) + \frac{1}{2}(\psi(a_\lambda) + \ln b_\lambda) - \frac{1}{2}\ln 2\pi - \frac{1}{2}a_\lambda b_\lambda\left(\left(\rho_j-\mu_\mu\right)^2 + \eta_\mu^{-1}\right) \tag{45}$$

where $\psi(\cdot)$ is the digamma function. $I(\rho_j)$ is the only term in Eq. (44) whose expectation [w.r.t. $q(\rho_j)$] cannot be derived analytically. Under the Laplace approximation, however, it is replaced by a second-order Taylor expansion around the variational posterior mode μ_{ρ_j},

$$I\left(\rho_j\right) \approx I\left(\mu_{\rho_j}\right) + I'\left(\mu_{\rho_j}\right)\left(\rho_j-\mu_{\rho_j}\right) + \frac{1}{2}I''\left(\mu_{\rho_j}\right)\left(\rho_j-\mu_{\rho_j}\right)^2. \tag{46}$$

This allows us to approximate the expectation of $I(\rho_j)$ by

$$\left\langle I\left(\rho_j\right)\right\rangle_{q(\rho_j)} \approx \underbrace{\left\langle I\left(\mu_{\rho_j}\right)\right\rangle_{q(\rho_j)}}_{I\left(\mu_{\rho_j}\right)} + I'\left(\mu_{\rho_j}\right)\underbrace{\left\langle \rho_j-\mu_{\rho_j}\right\rangle_{q(\rho_j)}}_{0} \tag{47}$$

$$+\frac{1}{2}\underbrace{I''\left(\mu_{\rho_j}\right)}_{-\eta_{\rho_j}}\underbrace{\left\langle\left(\rho_j-\mu_{\rho_j}\right)^2\right\rangle_{q(\rho_j)}}_{\eta_{\rho_j}^{-1}} = I\left(\mu_{\rho_j}\right) - \frac{1}{2} \tag{48}$$

where the equality $I''\left(\mu_{\rho_j}\right) = -\eta_{\rho_j}$ follows directly from Eq. (42). Hence, the expected log joint is:

$$\langle \ln p(k,\mu,\lambda,\rho)\rangle_q \approx \overbrace{\frac{1}{2}\ln\frac{\eta_0}{2\pi} - \frac{\eta_0}{2}\left(\left(\mu_\mu-\mu_0\right)^2 + \eta_\mu^{-1}\right)}^{\langle \ln N(\mu|\mu_0,\eta_0)\rangle_q} \overbrace{- \ln\Gamma(a_0) - a_0\ln b_0 + (a_0-1)(\psi(a_\lambda) + \ln b_\lambda) - \frac{a_\lambda b_\lambda}{b_0}}^{\langle \ln \mathrm{Ga}(\lambda|a_0,b_0)\rangle_q} + \sum_{j=1}^{m}\Big[\ln \mathrm{Bin}\left(k_j|\sigma\left(\mu_{\rho_j}\right)\right) + \frac{1}{2}(\psi(a_\lambda) + \ln b_\lambda) - \frac{1}{2}\ln 2\pi - \frac{1}{2}a_\lambda b_\lambda\left(\left(\mu_{\rho_j}-\mu_\mu\right)^2 + \eta_\mu^{-1}\right) - \frac{1}{2}\Big] \tag{49}$$

The second term of the free energy in Eq. (43) is the entropy $H[q]$ of the variational posterior:

$$\langle -\ln q(\mu,\lambda,\rho)\rangle_q = \overbrace{\frac{1}{2}\ln\frac{2\pi e}{\eta_\mu}}^{H\left[N(\mu|\mu_\mu,\eta_\mu)\right]} + \sum_{j=1}^{m}\overbrace{\frac{1}{2}\ln\frac{2\pi e}{\eta_{\rho_j}}}^{H\left[N\left(\rho_j|\mu_{\rho_j},\eta_{\rho_j}\right)\right]} + \overbrace{a_\lambda + \ln b_\lambda + \ln\Gamma(a_\lambda) + (1-a_\lambda)\psi(a_\lambda)}^{H[\mathrm{Ga}(\lambda|a_\lambda,b_\lambda)]} \tag{50}$$

Substituting Eqs. (49) and (50) into Eq. (43) yields an expression for the free energy,

$$\begin{aligned} F \approx\ & \frac{1}{2}\ln\frac{\eta_0}{\eta_\mu} - \frac{\eta_0}{2}\left(\left(\mu_\mu-\mu_0\right)^2+\eta_\mu^{-1}\right) + a_\lambda - a_0\ln b_0 + \ln\frac{\Gamma(a_\lambda)}{\Gamma(a_0)} \\ & - a_\lambda b_\lambda\left(\frac{1}{b_0}+\frac{m}{2\eta_\mu}\right) + \left(a_0+\frac{m}{2}\right)\ln b_\lambda + \left(a_0-a_\lambda+\frac{m}{2}\right)\psi(a_\lambda) \\ & + \frac{1}{2} + \sum_{j=1}^{m}\left[\ln \mathrm{Bin}\left(k_j\middle|\sigma\left(\mu_{\rho_j}\right)\right) - \frac{1}{2}a_\lambda b_\lambda\left(\mu_{\rho_j}-\mu_\mu\right)^2 - \frac{1}{2}\ln\eta_{\rho_j}\right]. \end{aligned} \tag{51}$$

The availability of the above approximation to the free energy leads to a straightforward variational algorithm. The algorithm is initialized by setting the moments of all approximate posteriors to the moments of their respective priors. It terminates when

$$F_{\text{current}} - F_{\text{previous}} < 10^{-3} \tag{52}$$

i.e., when the free energy has converged. This criterion typically leads to the same inference as a criterion based on the parameter estimates themselves, e.g.,

$$\left\|\theta_{\text{current}} - \theta_{\text{previous}}\right\|^2 < 10^{-3} \tag{53}$$

where convergence of $\theta \equiv \left(\mu_\mu, \eta_\mu, a_\lambda, b_\lambda, \mu_{\rho_1}, \ldots, \mu_{\rho_m}, \eta_{\rho_1}, \ldots, \eta_{\rho_m}\right)$ is expressed through a bound on their (squared) ℓ_2-norm. However, computing (an approximation to) the free energy itself has the additional advantage that it provides an approximation to the log model evidence (see Eqs. (9) and (43)), which permits Bayesian model selection (for an example, see Brodersen et al., 2012a).

MCMC sampling

The variational Bayes scheme presented above is computationally highly efficient; it typically converges after just a few iterations. However, its results are only exact to the extent to which its distributional assumptions are justified. To validate these assumptions, we compared VB to an asymptotically exact stochastic approach, i.e., Markov chain Monte Carlo (MCMC), which is computationally much more expensive than variational Bayes but exact in the limit of infinite runtime.

In the Supplemental Material, we describe a Gibbs sampler for inverting the univariate normal-binomial model introduced above. This algorithm is analogous to the one we previously introduced for the inversion of the bivariate normal-binomial model in Brodersen et al. (2012a). It proceeds by cycling over model parameters, drawing samples from their full-conditional distributions, until the desired number of samples (e.g., 10^6) has been generated (see Supplemental Material).

Unlike VB, which was based on a mean-field assumption, the posterior obtained through MCMC retains any potential conditional dependencies among the model parameters. The algorithm is computationally burdensome; but it can be used to validate the distributional assumptions underlying variational Bayes (see Applications).

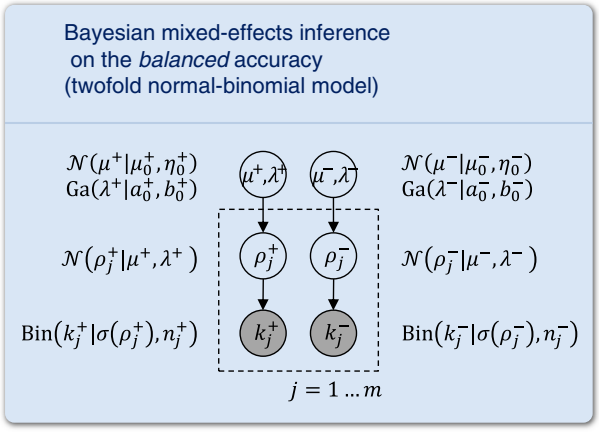

Fig. 4. Inference on balanced accuracies. The univariate normal-binomial model (Fig. 2) can be easily extended to enable inference on the balanced accuracy. Specifically, the model is inverted separately for classification outcomes obtained on positive and negative trials. The resulting posteriors are then recombined (see main text).

The twofold normal-binomial model for inference on the balanced accuracy

Seemingly strong classification accuracies can be trivially obtained on datasets consisting of different numbers of representatives from either class. For instance, a classifier might assign every example to the majority class and thus achieve an accuracy equal to the proportion of test cases belonging to the majority class. Thus, the use of classification accuracy as a performance measure may easily lead to optimistic inferences (Akbani et al., 2004; Brodersen et al., 2010a, 2012a; Chawla et al., 2002; Japkowicz and Stephen, 2002).

This has motivated 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,

$$\varphi := \frac{1}{2}\left(\pi^+ + \pi^-\right) \tag{54}$$

where $\pi^+ := \sigma(\mu^+)$ and $\pi^- := \sigma(\mu^-)$ denote the (population) classification accuracies on positive and negative trials, respectively.[8] The balanced accuracy reduces to the conventional accuracy whenever the classifier performed equally well on either class; and it drops to chance when the classifier performed well purely because it exploited an existing class imbalance. We will revisit the conceptual differences between accuracies and balanced accuracies in the Discussion. In this section, we show how the univariate normal-binomial model presented above can be easily extended to allow for inference on the balanced accuracy.

We have previously explored different ways of constructing models for inference on the balanced accuracy (Brodersen et al., 2012a). Here, we infer on the balanced accuracy by duplicating our generative model for accuracies and applying it separately to data from the two classes. This constitutes the *twofold* normal-binomial model (Fig. 4).

To infer on the balanced accuracy, we separately consider the number of correctly classified positive trials k_j^+ and the number of correctly predicted negative trials k_j^- for each subject $j = 1\ldots m$. We next describe the true accuracies within each subject as π_j^+ and π_j^-. The population parameters μ^+, λ^+ and μ^-, λ^- then represent the population accuracies on positive and negative trials, respectively.

Inverting the model proceeds by inverting its two parts independently. However, in contrast to the inversion of the *univariate*

[8] The extension to multiclass problems is considered in the Discussion.

normal-binomial model, we are no longer interested in the posterior densities over the population mean accuracies μ^+ and μ^- themselves. Rather, we wish to obtain the posterior density of the balanced accuracy,

$$p\left(\phi|k^+,k^-\right) = p\left(\frac{1}{2}\left(\sigma\left(\mu^+\right)+\sigma(\mu^-)\right)\middle|k^+,k^-\right). \tag{55}$$

Unlike the population mean accuracy (Eq. (29)), which was logit-normally distributed, the posterior mean of the population *balanced* accuracy can no longer be expressed in closed form. The same applies to subject-specific posterior balanced accuracies. We therefore approximate the respective integrals by (one-dimensional) numerical integration. If we were interested in the *sum* of the two class-specific accuracies, $s := \sigma(\mu^+) + \sigma(\mu^-)$, we would consider the convolution of the distributions for $\sigma(\mu^+)$ and $\sigma(\mu^-)$,

$$p\left(s\middle|k^+,k^-\right) = \int_0^s p_{\sigma(\mu^+)}\left(s-z\middle|k^+\right)p_{\sigma(\mu^-)}(z|k^-)dz \tag{56}$$

where $p_{\sigma(\mu^+)}$ and $p_{\sigma(\mu^-)}$ represent the individual posterior distributions of the population accuracy on positive and negative trials, respectively. In the same spirit, the modified convolution

$$p\left(\phi|k^+,k^-\right) = \int_0^{2\varphi} p_{\sigma(\mu^+)}\left(2\phi-z|k^+\right)p_{\sigma(\mu^-)}(z|k^-)dz \tag{57}$$

yields the posterior distribution of the *arithmetic mean* of two class-specific accuracies, i.e., the balanced accuracy.

Applications

This section illustrates the sort of inferences that can be made using VB in a classification study of a group of subjects. We begin by considering synthetic classification outcomes to evaluate the consistency of our approach and illustrate its link to classical fixed-effects and random-effects analyses. We then apply our approach to empirical fMRI data obtained from a trial-by-trial classification analysis.

Application to synthetic data

We examined the statistical properties of our approach in two typical settings: (i) a larger simulated group of subjects with many trials each; and (ii) a small group of subjects with few trials each, including missing trials. Before we turn to the results of these simulations, we will pick one simulated dataset from either setting to illustrate inferences supported by our model (Fig. 5).

The first synthetic setting is based on a group of 30 subjects with 200 trials each (i.e., 100 trials in each class). Outcomes were generated using the univariate normal-binomial model with a population mean (logit accuracy) of $\mu = 1.1$ (corresponding to a population mean accuracy of 71%) and a relatively high logit population precision of $\lambda = 4$ (corresponding to a population accuracy standard deviation of 9.3%; Fig. 5a). MCMC results were based on 100,000 samples, obtained from 8 parallel chains (see Supplemental Material).

In inverting the model, the parameter of primary interest is μ, the (logit) population mean accuracy. Our simulation showed a typical result in which the posterior distribution of the population mean was sharply peaked around the true value, with its shape virtually indistinguishable from the corresponding MCMC result (Fig. 5b). In practice, a good way of summarizing the posterior is to report a central 95% posterior probability interval (or Bayesian credible interval). Although this interval is conceptually different from a classical (frequentist) 95% confidence interval, in this particular case the two intervals agreed very closely (Fig. 5c), which is typical in the context of a large sample size. In contrast, fixed-effects intervals were overconfident when based on the pooled sample accuracy and underconfident when based on the average sample accuracy (Fig. 5c).

Another informative way of summarizing the posterior population mean is to report the posterior probability mass p that is below chance (e.g., 0.5 for binary classification). We refer to this probability as the (posterior) *infraliminal probability* of the classifier (cf. Brodersen et al., 2012a). Compared with a classical p-value, it has a deceptively similar, but more natural, interpretation. Rather than representing the frequency of observing the observed outcome (or a more extreme outcome) under the 'null' hypothesis of a classifier operating at or below chance (classical p-value), the infraliminal probability represents our posterior belief that the classifier does not perform better than chance. In the above simulation, we obtained $p \approx 10^{-10}$.

We next considered the true *subject-specific* accuracies and compared them (i) with conventional sample accuracies and (ii) with VB posterior means (Fig. 5e). This comparison highlighted one of the principal features of hierarchical models, that is, their *shrinkage* effect. Because of the limited numbers of trials, sample accuracies exhibited a larger variance than ground truth; accordingly, the posterior means, which were informed by data from the entire group, appropriately compensated for this effect by shrinking to the group mean. This effect is also known as *regression to the mean* and dates back to works as early as Galton's law of 'regression towards mediocrity' (Galton, 1886). It is obtained naturally in a hierarchical model and, as we will see below, leads to systematically more accurate posterior inferences at the single-subject level.

We repeated the above analysis on a sample dataset from a second simulation setting. This setting was designed to represent the example of a small group with varying numbers of trials across subjects.[9] Such a scenario is important to consider because it occurs in real-world applications whenever the number of trials eligible for subsequent classification is not entirely under experimental control. Varying numbers of trials also occur, for example, in clinical diagnostics of diseases like epilepsy where one may have different numbers of observations per patient. Classification outcomes were generated using the univariate normal-binomial model with a population mean logit accuracy of $\mu = 2.2$ and a low logit population precision of $\lambda = 1$; the corresponding population mean accuracy was 87%, with a population standard deviation of 11.2% (Fig. 5f).

Comparing the resulting posteriors (Figs. 5g–j) to those obtained on the first dataset, several differences are worth noting. Concerning the population parameters (Figs. 5g,i), all estimates remained in close agreement with ground truth; at the same time, minor discrepancies began to arise between variational and MCMC approximations, with the variational results slightly too precise (Figs. 5g,i). This can be seen best from the credible intervals (Fig. 5h, black). By comparison, an example of inappropriate inference can be seen in the frequentist confidence interval for the population accuracy, which does not only exhibit an optimistic shift towards higher performance but also includes accuracies above 100% (Fig. 5h, red).

Another typical consequence of a small dataset with variable trial numbers can be seen in the shrinkage of subject-specific inferences (Fig. 5j). In comparison to the first setting, there are fewer trials per subject, and so the shrinkage effect is stronger. In addition, subjects with fewer trials (red) are shrunk more than those with more trials (blue). Thus, the order between sample accuracies and posterior means has changed, as indicated by crossing black lines. Restoring the correct order of subjects can become important, for example, when one wishes to relate subject-specific accuracies to independent subject-specific characteristics, such as behavioral, demographic, or genetic information.

The primary advantage of VB over sampling algorithms is its computational efficiency. To illustrate this, we examined the computational load required to invert the normal-binomial model on the dataset

[9] Note that the heteroscedasticity in this dataset results both from the fact that subjects have different numbers of trials and from their different sample accuracies.

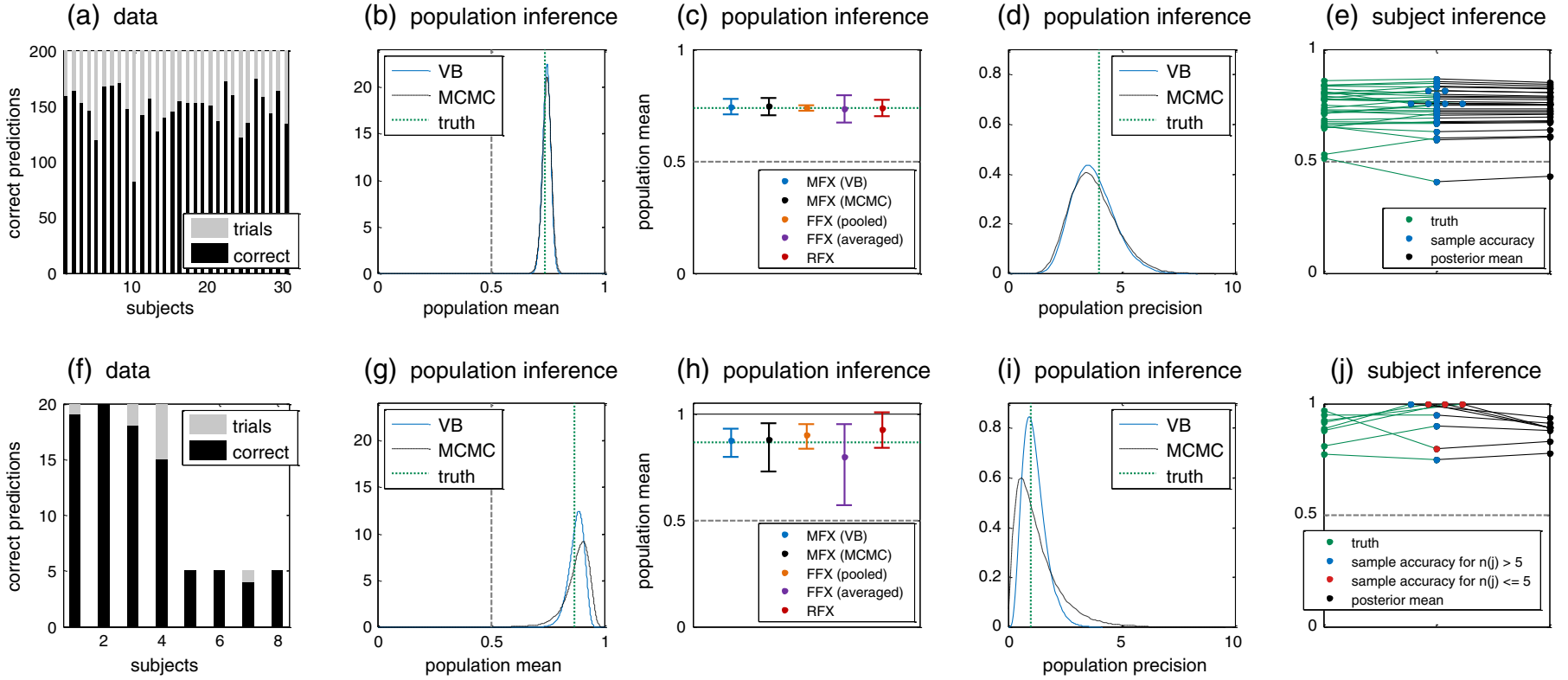

Fig. 5. Application to simulated data. Two simple synthetic datasets illustrate the sort of inferences that can be obtained using a mixed-effects model. (a) Simulated data, showing the number of trials in each subject (gray) and the number of correct predictions (black). (b) Resulting posterior density of the population mean accuracy when using variational Bayes or MCMC. (c) Posterior densities can be summarized in terms of central 95% posterior intervals. Here, the two Bayesian intervals (blue/black) are compared with a frequentist random-effects 95% confidence interval and with fixed-effects intervals based on the pooled and the averaged sample accuracy. (d) Posterior densities of the population precision (inverse variance). (e) The benefits of a mixed-effects approach in subject-specific inference can be visualized (cf. Brodersen et al., 2012a) by contrasting the increase in dispersion (as we move from ground truth to sample accuracies) with the corresponding decrease in dispersion (as we move from sample accuracies to posterior means). This effect is a consequence of the hierarchical structure of the model, and it yields better estimates of ground truth (cf. Figs. 7d,h). Notably, shrinking may change the order of subjects (when sorted by accuracy) since its extent depends on the subject-specific (first-level) posterior uncertainty. Note that the x-axis does not represent any quantity by itself but simply serves to space out the three groups of data points (ground truth, samples accuracies, and posterior means). Overlapping sample accuracies are additionally scattered horizontally for better visibility. (f–j) Same plots as in the top row, but based on a different simulation setting with a much smaller number of subjects and a smaller and more heterogeneous number of trials in each subject. The smaller size of the dataset enhances the merits of mixed-effects inference over conventional approaches and increases the shrinkage effect in subject-specific accuracies.

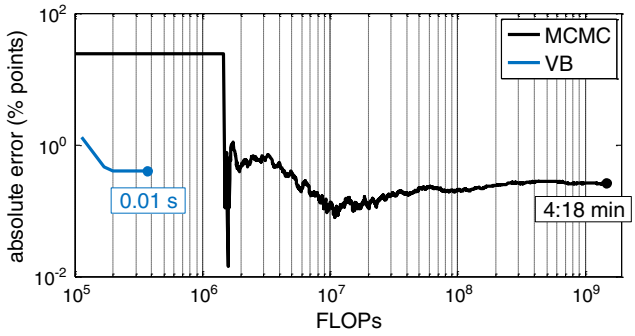

Fig. 6. Estimation error and computational complexity. VB and MCMC differ in the way estimation error and computational complexity are traded off. The plot shows estimation error in terms of the absolute difference of the posterior mean of the population mean accuracy in percentage points (y-axis). Computational complexity is shown in terms of the number of floating point operations (FLOPs) consumed. VB converged after 370,000 FLOPs (iterative update $< 10^{-6}$) to a posterior mean of the population mean accuracy of 73.5%. Given a true population mean of 73.9%, the estimation error of VB was −0.4 percentage points. In contrast, MCMC used up 1.47×10^9 FLOPs to draw 10,000 samples (excluding 100 burn-in samples). Its posterior mean estimate was 73.6%, implying an error of −0.26 percentage points. Thus, while MCMC ultimately achieved a marginally lower error, VB was computationally more efficient by more than 3 orders of magnitude. It should be noted that the plot uses log–log axes for readability; the difference between the two algorithms would be visually even more striking on a linear scale.

shown in Fig. 5a. Rather than measuring computation time (which is platform-dependent), we considered the number of floating-point operations (FLOPs), which we related to the absolute error of the inferred posterior mean of the mean population accuracy (in percentage points; Fig. 6). We found that MCMC used 4000 times more arithmetic operations to achieve an estimate that was better than VB by no more than 0.13 percentage points.

Application to a larger number of simulations

Moving beyond the single case examined above, we replicated our analysis many times while varying the true population mean accuracy between 0.5 and 0.9. For each point, we ran 200 simulations. This allowed us to examine the properties of our approach from a frequentist perspective (Fig. 7).

In the first setting (Fig. 7, top row), each simulation was based on synthetic classification outcomes from 30 subjects with 200 trials each, as described in the previous section. One instance of these simulations is shown as an example (Fig. 7a); all subsequent plots are based on 200 independent datasets generated in the same way.

We began by asking, in each simulation, whether the population mean accuracy was above chance (0.5). We answered this question by computing *p*-values using the following five methods: (i) fixed-effects inference based on a binomial test on the pooled sample accuracy (orange); (ii) fixed-effects inference based on a binomial test on the average sample accuracy (violet); (iii) mixed-effects inference using VB (solid black); (iv) mixed-effects inference using an MCMC sampler with 100,000 samples (dotted black); and (v) random-effects inference using a *t*-test on subject-specific sample accuracies (red).

An important aspect of inferential conclusions (whether frequentist or Bayesian under a diffuse prior) is their validity with respect to a given test size. For example, when using a test size of $\alpha = 0.05$, we expect the test statistic to be at or beyond the corresponding critical value for the 'null' hypothesis (of the classification accuracy to be at or below the level of chance) in precisely 5% of all simulations. We thus plotted the empirical *specificity*, i.e., the fraction of false rejections, as a function of test size (Fig. 7b). For any method to be a valid test, *p*-values should be uniformly distributed on the [0, 1] interval under the 'null'; thus, the empirical cumulative distribution function should approximate the main diagonal.

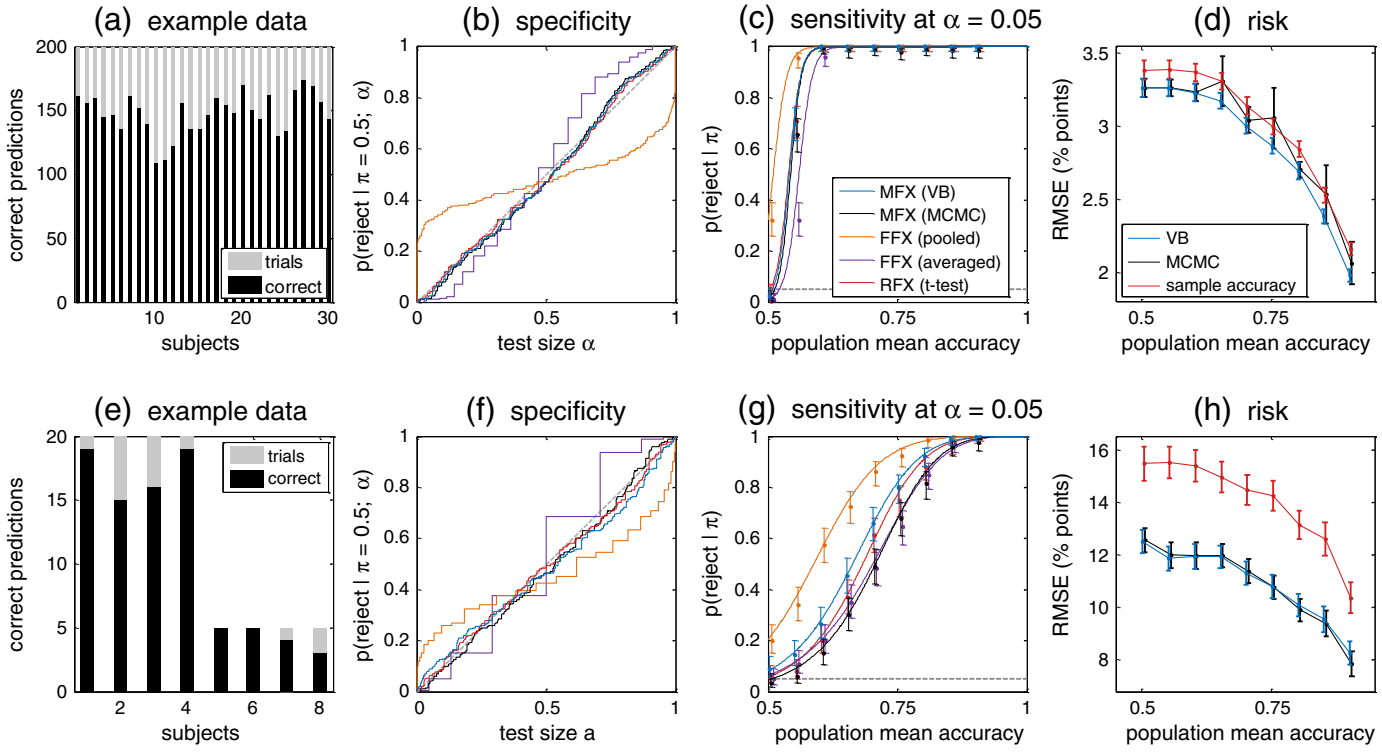

Fig. 7. Application to a larger number of simulations. (a) One example of 200 simulations of synthetic classification outcomes (generated using the same model as in Fig. 5a). (b) Specificity of competing methods for testing whether the population mean accuracy is greater than chance, given a true population mean of 0.5. (c) Power curve, testing whether the population mean accuracy is greater than chance, given different true population mean accuracies. (d) Comparison of accuracy of subject-specific estimates, using different inference methods. (e) Example of a smaller dataset (sampled from the same model as in Fig. 5f). (f–h) Same analyses as above, but based on smaller experiments.

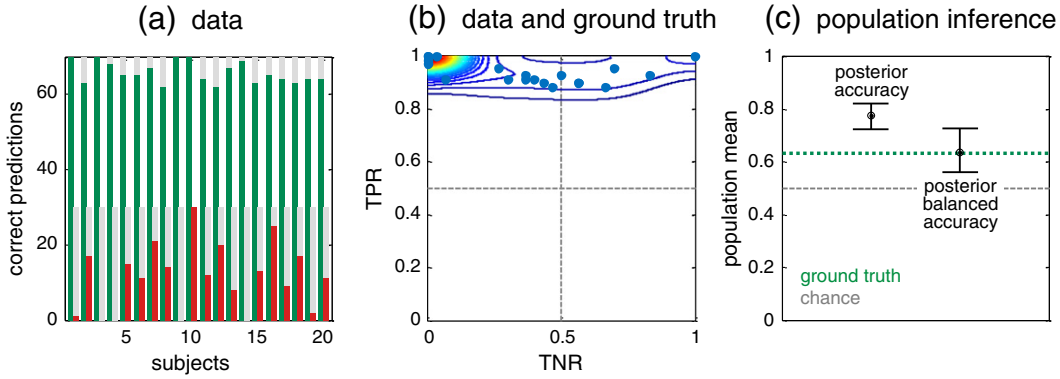

Fig. 8. Imbalanced data and the balanced accuracy. (a) In analogy with Fig. 7a, the panel shows a set of classification outcomes obtained by applying a linear support vector machine (SVM) to synthetic data, using 5-fold cross-validation. Individual bars represent, for each subject, the number of correctly classified positive (green) and negative (red) trials, as well as the respective total number of trials (gray). (b) Sample accuracies on positive (true positive rate, TPR) and negative classes (true negative rate, TNR), based on the classification outcomes shown in (a). The underlying true population distribution is shown in terms of a bivariate Gaussian kernel density estimate (contour lines). Sample accuracies can be thought of as being drawn from this two-dimensional density. The plot shows that the population accuracy is high on positive trials and low on negative trials; the imbalance in the data has led the SVM to acquire a bias in favor of the majority class. (c) As an example of an inference that can be obtained using the approach presented in this paper, the last panel shows central 95% posterior probability intervals of the population mean accuracy and the balanced accuracy. The plot shows that inference on the accuracy is misleading as it must be interpreted in relation to an implicit baseline that is different from 0.5. By contrast, the balanced accuracy interval provides a sharply peaked estimate of the true balanced accuracy; its baseline is 0.5.

As can be seen from Fig. 7b, the first method violates this requirement (fixed-effects analysis, orange). It pools the data across all subjects; as a result, above-chance performance is concluded too frequently at small test sizes and not concluded frequently enough at larger test sizes. In other words, a binomial test on the pooled sample accuracy provides invalid inference on the population mean accuracy.

A second important property of inference schemes is their *sensitivity* or statistical *power* (Fig. 7c). An *ideal* test (falsely) rejects the null with a probability of α when the null is true, and always (correctly) rejects the null when it is false. In the presence of observation noise, such a test is only guaranteed to exist in the limit of an infinite amount of data. Thus, given a finite dataset, we can compare the power of different inference methods by examining how quickly their rejection rates rise once the null is no longer true. Using a test size of $\alpha = 0.05$, we carried out 200 simulations for each level of true population mean accuracy (0.5, 0.6, …, 0.9) and plotted empirical rejection rates. The figure shows, as expected, that a Binomial test on the pooled sample accuracy is an invalid test, in the sense that it rejects the null hypothesis too frequently when it is true. This effect will become even clearer when using a smaller dataset (see below).[10]

Finally, we examined the performance of our VB algorithm for estimating subject-specific accuracies (Fig. 7d). We compared three estimators: (i) posterior means of $\sigma(\rho_j)$ using VB; (ii) posterior means $\sigma(\rho_j)$ using MCMC; and (iii) sample accuracies, i.e., individual maximum-likelihood estimates. The figure shows that posterior estimates based on a mixed-effects model led to a slightly smaller estimation error than sample accuracies. This effect was small in this scenario but became substantial when considering a smaller dataset, as described next.

In the second setting (Fig. 7, bottom row), we carried out the same analyses as above, but based on small datasets of just 8 subjects with different numbers of trials (Fig. 7e). Regarding test specificity, as before, we found fixed-effects inference to yield highly over-optimistic inferences at low test sizes (Fig. 7f).

The same picture emerged when considering sensitivities (Fig. 7g). Fixed-effects inference on the pooled sample accuracy yielded over-confident results; it systematically rejected the null hypothesis too easily. A conventional *t*-test on subject-specific sample accuracies provided a valid test, with no more false positives under the null than prescribed by the test size (red). However, it was outperformed by a mixed-effects approach (black), whose rejection probability rises more quickly when the null is no longer true, thus offering greater statistical power than the *t*-test.

Finally, in this setting of a small group size and few trials, subject-specific inference benefitted substantially from a mixed-effects model (Fig. 7h). This is due to the fact that subject-specific posteriors are informed by data from the entire group, whereas sample accuracies are only based on the data from an individual subject.

Accuracies versus balanced accuracies

As described above, the classification accuracy of an algorithm (obtained on an independent test set or through cross-validation) can be a misleading measure of generalization ability when the underlying data are not perfectly balanced. To resolve this problem, we use a straightforward extension of our model, the twofold normal-binomial model (Fig. 4), that enables inference on balanced accuracies. To illustrate the differences between the two quantities, we revisited, using our new VB algorithm, an analysis from a previous study in which we had generated an imbalanced synthetic dataset and used a linear support vector machine (SVM) for classification (Fig. 8; for details, see Brodersen et al., 2012a).

We observed that, as expected, the class imbalance caused the classifier to acquire a bias in favor 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 (Fig. 8a). The bias is reflected accordingly by the estimated bivariate density of class-specific classification accuracies, in which the majority class consistently performed well whereas the accuracy on the minority class varied strongly, covering virtually the entire [0, 1] range (Fig. 8b). In this setting, we found that the twofold normal-binomial model of the balanced accuracy provided an excellent estimate of the true balanced accuracy under which the data had been generated (dotted green line in Fig. 8c). In stark contrast, using the single normal-binomial model to infer on the population accuracy resulted in estimates that were considerably too optimistic and therefore misleading.

Application to fMRI data

To demonstrate the practical applicability of our VB method for mixed-effects inference, we analyzed data from an fMRI experiment involving 16 volunteers who participated in a simple decision-making task (Fig. 9). During the experiment, subjects had to choose, on each trial, between two options that were presented on the screen. Decisions

[10] The above simulation could also be used for a power analysis to assess what population mean accuracy would be required to reach a particular probability of obtaining a positive (above-chance) finding.

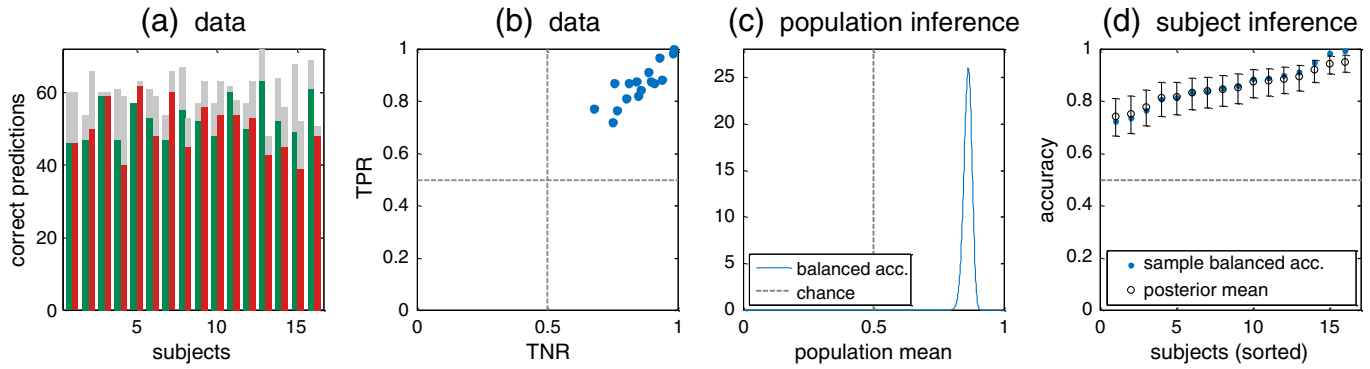

Fig. 9. Application to empirical fMRI data: overall classification performance. (a) Classification outcomes obtained by applying a linear SVM to trial-wise fMRI data from a decision-making task. (b) Posterior population mean accuracy, inferred on using variational Bayes. (c) Posterior population precision. (d) Subject-specific posterior inferences. The plot contrasts sample accuracies with central 95% posterior probability intervals. In this case, the shrinkage effect (discrepancy between blue dots and black circles) is diminished by the large number of trials per subject.

were indicated by button press (left/right index finger). Details on the underlying experimental design, data acquisition, and preprocessing can be found elsewhere (Behrens et al., 2007). Here, we aimed to decode (i.e., classify) from fMRI measurements which option had been chosen on each trial. Because different choices were associated with different buttons, we expected to find 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 entered a linear support vector machine (SVM), as implemented by Chang and Lin (2011), that was trained and tested using 5-fold cross-validation. Comparing predicted to actual choices resulted in 120 classification outcomes for each of the 16 subjects (Fig. 9a).

Using the univariate normal-binomial model for inference on the population mean accuracy, we obtained clear evidence (infraliminal probability $p < 0.001$) that the classifier was operating above chance (Fig. 9b). The variational posterior population mean balanced accuracy (posterior mean 73.7%; Fig. 9c) agreed closely with an MCMC-based posterior (73.5%; not shown). Inference on subject-specific balanced accuracies yielded fairly precise posterior intervals whose shrinkage to the population, due to the large number of trials per subject, was only small (Fig. 9d).

The overall computation time for the above VB inferences was approximately 7 ms on a 2.53 GHz Intel Xeon E5540 processor. This speedup in comparison to previous MCMC algorithms makes it feasible to construct whole-brain maps of above-chance accuracies. We illustrate this using a searchlight classification analysis (Kriegeskorte et al., 2006; Nandy and Cordes, 2003). In this analysis, we passed a sphere (radius 6 mm) across the brain. At each location, we trained and tested a linear SVM using 5-fold cross-validation. We associated the voxel at the center of the current sphere with the number of correct predictions (i.e., the vector $k_{1:16} \in \mathbf{N}^{16}$). We then used our VB algorithm to compute a whole-brain posterior accuracy map (PAM; Fig. 10). Comprising 220,000 voxels, the map took no more than 7 min 18 s to complete. The map shows the posterior population mean accuracy in voxels with an infraliminal probability of less than 0.001. Thus, it

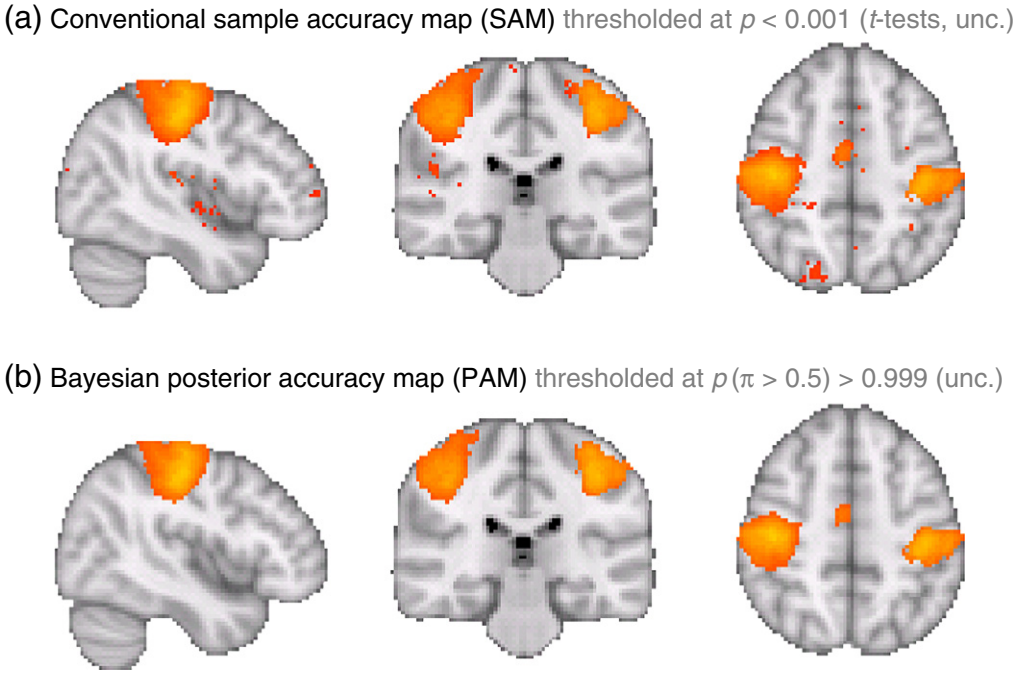

Fig. 10. Application to empirical fMRI data: posterior accuracy map. (a) A conventional sample accuracy map (SAM) highlights regions in which a one-tailed t-test on subject-specific sample accuracies yielded $p < 0.001$ (uncorrected). (b) Using the VB algorithm presented in this paper, we can instead create a posterior accuracy map (PAM), which highlights those regions in which the posterior accuracy of the classification algorithm operating above chance is greater than 99.9%.

highlights regions with a posterior probability of the classifier operating above chance at the group level that is at least 99.9%.

For comparison, we contrast this result with a conventional sample accuracy map (SAM), thresholded at $p < 0.001$ (uncorrected). While the results are, overall, rather similar, the SAM shows several scattered small clusters and isolated voxels in white matter and non-motor regions that the PAM does not display.

Discussion

In this paper, we have introduced a VB algorithm for highly efficient inversion of a hierarchical Bayesian normal-binomial model which enables full mixed-effects inference on classification accuracy in group studies. Owing to its hierarchical structure, the model reflects both within-subjects and between-subjects variance components and exploits the available group data to optimally constrain inference in individual subjects. The ensuing shrinkage effects yield more accurate subject-specific estimates than those obtained through non-hierarchical models. The proposed model follows a natural parameterization and can be inverted in a fully Bayesian fashion. It is independent of the type of underlying classifier, and it supports inference both on the accuracy and on the balanced accuracy; the latter is the preferred performance measure when the data are not perfectly balanced.

In previous work, we have successfully established sampling (MCMC) approaches to Bayesian mixed-effects inference on classification accuracy at the group level (Brodersen et al., 2012a). Extending this work, the critical contribution of the present paper is the derivation and validation of a VB method for model inversion. Our new approach drastically reduces the computational complexity of previous sampling schemes (by more than 3 orders of magnitude; cf. Fig. 6) while maintaining comparable accuracy.

The parameterization used in the present manuscript differs slightly from that introduced in our previous implementations (Brodersen et al., 2012a). Specifically, we are now modeling the population distribution of subject-specific accuracies using a logit-normal density rather than a beta density. Neither is generally superior to the other; they simply represent (minimally) different assumptions about the distribution of classifier performance across subjects. However, the logit-normal density is a more natural candidate distribution when deriving a Laplace approximation, as we do here, since it implies closed-form, interpretable update equations and since it enables a straightforward approximation to the free energy (cf. Theory section). Should the issue which distribution is optimal for a given dataset at hand become a question of interest for a particular application, one can weigh the evidence for different parameterizations by means of Bayesian model selection, using the code provided in our toolbox, as shown in Brodersen et al. (2012a).

In addition to their excessive runtime, MCMC approaches to model inversion come with a range of practical challenges, such as: how to select the number of required samples; how to check for convergence, or even guarantee it; how long to design the burn-in period; how to choose the proposal distribution in Metropolis steps; how many chains to run in parallel; and how to design overdispersed initial parameter densities. By contrast, deterministic approximations such as VB involve fewer practical engineering considerations. Rather, they are based on a set of distributional assumptions that can be captured in a simple graphical model (cf. Figs. 2 and 4). While not a specific theme of this paper, it is worth reiterating that the free-energy estimate provided by VB represents an approximation to the log evidence of the model (cf. Eqs. (9) and (43)), making it easy to compare alternative distributional assumptions.

Thus, compared to previous MCMC implementations of mixed-effects inference, the present paper is fundamentally based on an idea that has been at the heart of many recent innovations in the statistical analysis of neuroimaging data: the idea that minor reductions in statistical accuracy are warranted in return for a major increase in computational efficiency.

Advances in computing power might suggest that the importance of computational efficiency should become less critical over time; but neuroimaging has repeatedly experienced how new ideas radically increase demands on computation time and thus the importance of fast algorithms. One example is provided by large-scale analyses such as searchlight approaches (Kriegeskorte et al., 2006; Nandy and Cordes, 2003), in which we must potentially evaluate as many classification results as there are voxels in a whole-brain scan. The speed of our VB method makes it feasible to create a whole-brain map of posterior mean accuracies within a few minutes (Fig. 10). Ignoring the time taken by the classification algorithm itself, merely turning classification outcomes into posterior accuracies would have taken no less than 31 days when using an MCMC sampler with 30,000 samples for each voxel. By contrast, all computations were completed in less than 8 min when using variational Bayes, as we did in Fig. 10.

The conceptual differences between classical and Bayesian maps have been discussed extensively in the context of statistical parametric maps (SPM) and their Bayesian complements, i.e., posterior parametric maps (PPM; Friston et al., 2002; Friston and Penny, 2003). In brief, posterior accuracy maps (PAM) confer exactly the same advantages over sample accuracy maps (SAM) as PPMs over SPMs. This makes PAMs an attractive alternative to conventional (sample-accuracy) searchlight maps.

An important feature of our approach is its flexibility with regard to performance measures. While classification algorithms used to be evaluated primarily in terms of their accuracy, the limitations of this metric have long been known and are being increasingly addressed (Akbani et al., 2004; Chawla et al., 2002; Japkowicz and Stephen, 2002). For example, it has been suggested to restore balance by *undersampling* the larger class or by *oversampling* the smaller class. It is also possible to modify the costs of misclassification (Zhang and Lee, 2008) to prevent bias. A complementary, more generic safeguard is to replace the accuracy by the balanced accuracy, which removes the bias that may arise when a classifier is trained and tested on imbalanced data.

Fundamentally, accuracies and balanced accuracies address different scientific questions. Inference on the *accuracy* answers the question: what is the probability of making a correct prediction on a trial randomly drawn from a distribution with the same potential imbalance as that present in the current training set? Inference on the *balanced accuracy*, by contrast, answers the question: what is the probability of a correct prediction on a trial that is equally likely (a priori) to come from either class? To assess performance, this is what we are almost always interested in: the expected accuracy under a flat prior over classes.

The balanced accuracy is not confined to binary classification; it readily generalizes to K classes. Specifically, the twofold normal-binomial model becomes a K-fold model, and the balanced accuracy $\varphi = \frac{1}{K}\sum \pi^{(k)}$ is computed on the basis of a convolution of K random variables (cf. Eq. (54)).[11] Infraliminal probabilities are then determined w.r.t. the baseline level $1/K$. Even more generally, in those applications where one wishes to distinguish between different types of error (as, for example, in differential diagnostics where different misclassifications carry different costs), one could consider a weighted average of class-specific accuracies.

At first glance, another solution to dealing with imbalanced datasets would be to stick with the conventional accuracy but relate it to the correct baseline performance, i.e., the relative frequency of the majority class, rather than, e.g., 0.5 in the case of binary classification. The main weakness of this solution is that each and every report of classification

[11] It is worth remembering that, in the case of a multiclass setting, one would not necessarily replace the binomial distribution by a multinomial distribution. The bivariateness of the normal-binomial model refers to the fact that each classification outcome is either *correct* or *incorrect*, not to the number of classes.

performance would have to include an explicit baseline level, which would make the comparison of accuracies across studies, datasets, or classifiers slightly tedious. Future extensions of the approach presented in this paper might include functional performance measures such as the receiver-operating characteristic (ROC) or the precision-recall curve (Brodersen et al., 2010b).

Leaving classification studies aside for a moment, it is instructive to remember that mixed-effects inference and Bayesian estimation approaches have been successfully employed in other domains of neuroimaging data analysis (Fig. 11). One example are mass-univariate fMRI analyses based on the general linear model (GLM), where early fixed-effects models were soon replaced by random-effects and full mixed-effects approaches that have since become standards in the field (Beckmann et al., 2003; Friston et al., 1999, 2005; Holmes and Friston, 1998; Mumford and Nichols, 2009; Woolrich et al., 2004).

A parallel development in the domain of mass-univariate analyses has been the complementation of classical maximum-likelihood inference by Bayesian approaches (e.g., in the form of posterior probability maps; Friston and Penny, 2003). While maximum-likelihood schemes are concerned with the single most likely parameter value (i.e., mode), Bayesian inference aims for the full conditional density over possible parameter values given the data.

Another example concerns group-level model selection, e.g., in the context of dynamic causal modeling (DCM; Friston et al., 2003). Here, selecting an optimal model from a set of predefined alternatives initially rested on criteria for fixed-effects inference, such as the group Bayes factor (Stephan et al., 2007). This has subsequently been supplanted by random-effects inference that is more appropriate for typical applications in cognitive and clinical neuroscience when different mechanisms underlie measured data and thus different models are optimal across subjects (Stephan et al., 2009).

The present study addresses similar issues, but in a different context, that is, in classification group analyses. In both cases, an approximate but efficiently computable solution to a mixed-effects model (i.e., hierarchical VB) is preferable to an exact estimation of a non-hierarchical model (such as a *t*-test on sample accuracies) that disregards variability at the subject or group level. In other words: "An approximate answer to the right problem is worth a good deal more than an exact answer to an approximate problem" (John W. Tukey, 1915–2000).

It is worth noting that the model used in this paper is formally related to an earlier approach proposed by Leonard (1972). However, our choice of priors is motivated differently; we introduce a variational procedure for inference; and we use the normal-binomial model as a building block to construct larger models that can be used for inference on other performance measures, such as the balanced accuracy. Another related approach has been discussed by Olivetti et al. (2012), who carry out inference on the population mean accuracy by means of model selection between a null model and an above-chance model. For a more detailed discussion of these approaches, see Brodersen et al. (2012a).

One assumption common to all approaches considered in this paper, whether Bayesian or frequentist, is that trial-wise classification outcomes y_i are conditionally independent and identically distributed (i.i.d.) given a subject-specific accuracy π_j. This assumption implies *exchangeability*, which regards the joint distribution $p\left(y_{1_j},\ldots,y_{n_j}\right)$ as invariant to permutations of the indices $1_j \ldots n_j$. Exchangeability can be safely assumed whenever no information is conveyed by the trial indices themselves (see Gelman et al., 2003, for a discussion). The stronger i.i.d. postulate is justified by assuming that test observations are conditionally i.i.d. themselves. While this may not always hold in a cross-validation setting (Gustafsson et al., 2010; Kohavi, 1995; Pereira and Botvinick, 2011; Pereira et al., 2009; Wickenberg-Bolin et al., 2006), it is an appropriate assumption when adopting a single-split (or hold-out) scheme, by training on one half of the data and testing on the other (cf. discussion in Brodersen et al., 2012a).

An important aspect of the proposed model is that it can be applied regardless of what underlying classifier was used; its strengths result from the fact that it accounts for the hierarchical nature of classification outcomes observed at the group level. This suggests that one might want to use classifiers that account for the data hierarchy already at the stage of classification. Unlocking the potential benefits of this approach will be an interesting theme for future work (see Gopal et al., 2012, for an example).

Finally, it is worth noting that the regularization (shrinkage) of subject-specific posterior estimates by group-level estimates which our model conveys (cf. Figs. 5j, 9d) may be beneficial for a number of real-world applications. One example are studies where the number of observations (trials) per subject cannot be controlled experimentally. This is the case in all behavioral paradigms in which the number of trials eligible for classification depends (in part) on the subject's behavior. It is also the case in some clinical applications, e.g., in epilepsy or schizophrenia. In experiments involving patients suffering from these conditions, the occurrence of epileptic and hallucinatory events, respectively, cannot be controlled by the clinician during the period of investigation. In this case, the amount by which any one subject-specific estimate is shrunk towards the population mean is correctly scaled by the number of trials in that subject (Eq. (38)). The posterior population mean, in turn, is based on the sum of these subject-specific estimates (Eq. (28)) and thus also takes into account how many trials were obtained from each subject.

In summary, the VB approach proposed in this paper is as easy to use as a *t*-test, but conveys several advantages over contemporary

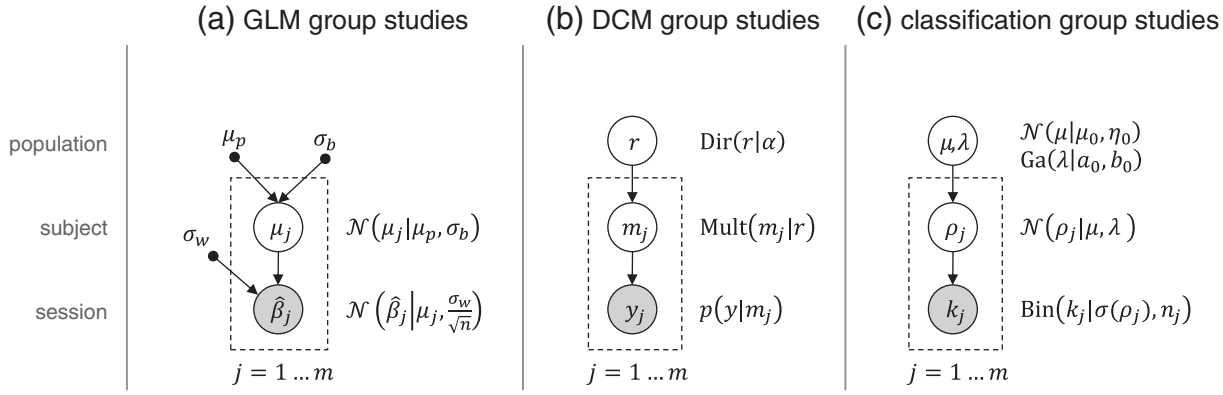

Fig. 11. Analogies between mixed-effects models in neuroimaging. (a) The first broadly adopted models for mixed-effects inference and Bayesian estimation in neuroimaging were developed for mass-univariate fMRI analyses based on the general linear model. The figure shows a graphical representation of the (random-effects) summary-statistics approximation to mixed-effects inference. (b) Mixed-effects models have subsequently also been developed for group studies based on dynamic causal modeling (DCM). (c) The present study addresses similar issues, but in a different context, that is, in group classification analyses.

fixed-effects and random-effects analyses. These advantages include: (i) posterior densities as opposed to point estimates of parameters; (ii) increased sensitivity (statistical power), i.e., a higher probability of detecting a positive result, especially with small sample sizes; (iii) a shrinking-to-the-population (or regression-to-the-mean) effect whose regularization leads to more precise subject-specific accuracy estimates; and (iv) posterior accuracy maps (PAM) which provide a mixed-effects alternative to conventional sample accuracy maps (SAM).

In order to facilitate its use and dissemination, the VB approach introduced in this paper has been implemented as open-source software for both MATLAB and R. The code is freely available for download (http://www.translationalneuromodeling.org/software/). With this software we hope to assist in improving the statistical sensitivity and interpretation of results in future classification group studies.

Acknowledgments

The authors wish to thank Tom Nichols, Tim Behrens, Mark Woolrich, Adrian Groves, Ged Ridgway, and Falk Lieder for insightful discussions, and additionally Tim Behrens for sharing fMRI data. This research was supported 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 René and Susanne Braginsky Foundation (KES).

Conflict of interest

The authors declare no conflict of interest.

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.neuroimage.2013.03.008.

References

Akbani, R., Kwek, S., Japkowicz, N., 2004. Applying support vector machines to imbalanced datasets. Machine Learning: ECML 2004, pp. 39–50.
Attias, H., 2000. A variational Bayesian framework for graphical models. Adv. Neural Inf. Process. Syst. 12, 209–215.
Beckmann, C.F., Jenkinson, M., Smith, S.M., 2003. General multilevel linear modeling for group analysis in fMRI. Neuroimage 20, 1052–1063.
Behrens, T.E.J., Woolrich, M.W., Walton, M.E., Rushworth, M.F.S., 2007. Learning the value of information in an uncertain world. Nat. Neurosci. 10, 1214–1221.
Bishop, C.M., 2007. Pattern Recognition and Machine Learning. Springer, New York.
Bishop, C.M., Spiegelhalter, D., Winn, J., 2002. VIBES: a variational inference engine for Bayesian networks. Adv. Neural Inf. Process. Syst. 15, 777–784.
Blankertz, B., Lemm, S., Treder, M., Haufe, S., Müller, K.-R., 2011. Single-trial analysis and classification of ERP components—a tutorial. Neuroimage 15, 814–825.
Brodersen, K.H., Ong, C.S., Stephan, K.E., Buhmann, J.M., 2010a. The balanced accuracy and its posterior distribution. Proceedings of the 20th International Conference on Pattern Recognition. IEEE Computer Society, pp. 3121–3124.
Brodersen, K.H., Ong, C.S., Stephan, K.E., Buhmann, J.M., 2010b. The binormal assumption on precision-recall curves. Proceedings of the 20th International Conference on Pattern Recognition. IEEE Computer Society, pp. 4263–4266.
Brodersen, K.H., Haiss, F., Ong, C.S., Jung, F., Tittgemeyer, M., Buhmann, J.M., Weber, B., Stephan, K.E., 2011a. Model-based feature construction for multivariate decoding. Neuroimage 56, 601–615.
Brodersen, K.H., Schofield, T.M., Leff, A.P., Ong, C.S., Lomakina, E.I., Buhmann, J.M., Stephan, K.E., 2011b. Generative embedding for model-based classification of fMRI data. PLoS Comput. Biol. 7, e1002079.
Brodersen, K.H., Mathys, C., Chumbley, J.R., Daunizeau, J., Ong, C.S., Buhmann, J.M., Stephan, K.E., 2012a. Bayesian mixed-effects inference on classification performance in hierarchical data sets. J. Mach. Learn. Res. 13, 3133–3176.
Brodersen, K.H., Wiech, K., Lomakina, E.I., Lin, C., Buhmann, J.M., Bingel, U., Ploner, M., Stephan, K.E., Tracey, I., 2012b. Decoding the perception of pain from fMRI using multivariate pattern analysis. Neuroimage 63, 1162–1170.
Chadwick, M.J., Hassabis, D., Weiskopf, N., Maguire, E.A., 2010. Decoding individual episodic memory traces in the human hippocampus. Curr. Biol. 20, 544–547.
Chang, C.-C., Lin, C.-J., 2011. LIBSVM: a library for support vector machines. ACM Trans. Intell. Syst. Technol. 2, 27:1–27:27.
Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P., 2002. SMOTE: synthetic minority over-sampling technique. J. Artif. Intell. Res. 16, 321–357.
Clithero, J.A., Smith, D.V., Carter, R.M., Huettel, S.A., 2011. Within- and cross-participant classifiers reveal different neural coding of information. Neuroimage 56, 699–708.
Cox, D.D., Savoy, R.L., 2003. Functional magnetic resonance imaging (fMRI) "brain reading": detecting and classifying distributed patterns of fMRI activity in human visual cortex. Neuroimage 19, 261–270.
Davatzikos, C., Resnick, S.M., Wu, X., Parmpi, P., Clark, C.M., 2008. Individual patient diagnosis of AD and FTD via high-dimensional pattern classification of MRI. Neuroimage 41, 1220–1227.
Dixon, P., 2008. Models of accuracy in repeated-measures designs. J. Mem. Lang. 59, 447–456.
Efron, B., Morris, C., 1971. Limiting the risk of Bayes and empirical Bayes estimators—part I: the Bayes case. J. Am. Stat. Assoc. 807–815.
Fox, C.W., Roberts, S.J., 2012. A tutorial on variational Bayesian inference. Artif. Intell. Rev. 38, 85–95.
Friston, K.J., Penny, W., 2003. Posterior probability maps and {SPMs}. Neuroimage 19, 1240–1249.
Friston, K.J., Holmes, A.P., Worsley, K.J., Poline, J.P., Frith, C.D., Frackowiak, R.S.J., 1995. Statistical parametric maps in functional imaging: a general linear approach. Hum. Brain Mapp. 2, 189–210.
Friston, K.J., Holmes, A.P., Worsley, K.J., 1999. How many subjects constitute a study? Neuroimage 10, 1–5.
Friston, K.J., Penny, W., Phillips, C., Kiebel, S., Hinton, G., Ashburner, J., 2002. Classical and Bayesian inference in neuroimaging: theory. Neuroimage 16, 465–483.
Friston, K.J., Harrison, L., Penny, W., 2003. Dynamic causal modelling. Neuroimage 19, 1273–1302.
Friston, K.J., Stephan, K.E., Lund, T.E., Morcom, A., Kiebel, S., 2005. Mixed-effects and fMRI studies. Neuroimage 24, 244–252.
Galton, F., 1886. Regression towards mediocrity in hereditary stature. J. Anthropol. Inst. Great Brit. Ireland 15, 246–263.
Gelman, A., Carlin, J.B., Stern, H.S., Rubin, D.B., 2003. Bayesian Data Analysis, 2nd ed. Chapman and Hall/CRC.
Ghahramani, Z., Beal, M.J., 2001. Propagation algorithms for variational Bayesian learning. Adv. Neural Inf. Process. Syst. 507–513.
Goldstein, H., 2010. Multilevel Statistical Models. Wiley.
Gopal, S., Yang, Y., Bai, B., Niculescu-Mizil, A., 2012. Bayesian models for Large-scale Hierarchical Classification. Adv. Neural Inf. Process. Syst. 25, 2420–2428.
Gustafsson, M.G., Wallman, M., Wickenberg Bolin, U., Göransson, H., Fryknäs, M., Andersson, C.R., Isaksson, A., 2010. Improving Bayesian credibility intervals for classifier error rates using maximum entropy empirical priors. Artif. Intell. Med. 49, 93–104.
Harrison, S.A., Tong, F., 2009. Decoding reveals the contents of visual working memory in early visual areas. Nature 458, 632–635.
Hassabis, D., Chu, C., Rees, G., Weiskopf, N., Molyneux, P.D., Maguire, E.A., 2009. Decoding neuronal ensembles in the human hippocampus. Curr. Biol. 19, 546–554.
Haynes, J.-D., Rees, G., 2006. Decoding mental states from brain activity in humans. Nat. Rev. Neurosci. 7, 523–534.
Holmes, A.P., Friston, K.J., 1998. Generalisability, random effects and population inference. Fourth Int Conf on Functional Mapping of the Human Brain. Neuroimage 7, S754.
Japkowicz, N., Stephen, S., 2002. The class imbalance problem: a systematic study. Intell. Data Anal. 6, 429–449.
Johnson, J.D., McDuff, S.G.R., Rugg, M.D., Norman, K.A., 2009. Recollection, familiarity, and cortical reinstatement: a multivoxel pattern analysis. Neuron 63, 697–708.
Just, M.A., Cherkassky, V.L., Aryal, S., Mitchell, T.M., 2010. A neurosemantic theory of concrete noun representation based on the underlying brain codes. PLoS One 5, e8622.
Klöppel, S., Stonnington, C.M., Chu, C., Draganski, B., Scahill, R.I., Rohrer, J.D., Fox, N.C., Jack, C.R., Ashburner, J., Frackowiak, R.S.J., 2008. Automatic classification of MR scans in Alzheimer's disease. Brain 131, 681–689.
Klöppel, S., Abdulkadir, A., Jack Jr., C.R., Koutsouleris, N., Mourão-Miranda, J., Vemuri, P., 2012. Diagnostic neuroimaging across diseases. Neuroimage 61, 457–463.
Knops, A., Thirion, B., Hubbard, E.M., Michel, V., Dehaene, S., 2009. Recruitment of an area involved in eye movements during mental arithmetic. Science 324, 1583–1585.
Kohavi, R., 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. International Joint Conference on Artificial Intelligence. Lawrence Erlbaum Associates Ltd., pp. 1137–1145.
Krajbich, I., Camerer, C., Ledyard, J., Rangel, A., 2009. Using neural measures of economic value to solve the public goods free-rider problem. Science 326, 596–599.
Kriegeskorte, N., Goebel, R., Bandettini, P., 2006. Information-based functional brain mapping. Proc. Natl. Acad. Sci. U. S. A. 103, 3863–3868.
Langford, J., 2005. Tutorial on practical prediction theory for classification. J. Mach. Learn. Res. 6, 273–306.
Lemm, S., Blankertz, B., Dickhaus, T., Müller, K.-R., 2011. Introduction to machine learning for brain imaging. Neuroimage 56, 387–399.
Leonard, T., 1972. Bayesian methods for binomial data. Biometrika 59, 581–589.
MacKay, D.J.C., 1995. Ensemble learning and evidence maximization. Proc. NIPS (Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.54.4083&rep=rep1&type=pdf [Accessed January 4, 2013]).
Marquand, A., Howard, M., Brammer, M., Chu, C., Coen, S., Mourão-Miranda, J., 2010. Quantitative prediction of subjective pain intensity from whole-brain fMRI data using Gaussian processes. Neuroimage 49, 2178–2189.
Mumford, J.A., Nichols, T., 2009. Simple group fMRI modeling and inference. Neuroimage 47, 1469–1475.
Nandy, R.R., Cordes, D., 2003. Novel nonparametric approach to canonical correlation analysis with applications to low CNR functional MRI data. Magn. Reson. Med. 50, 354–365.
Norman, K.A., Polyn, S.M., Detre, G.J., Haxby, J.V., 2006. Beyond mind-reading: multi-voxel pattern analysis of fMRI data. Trends Cogn. Sci. 10, 424–430.
Olivetti, E., Veeramachaneni, S., Nowakowska, E., 2012. Bayesian hypothesis testing for pattern discrimination in brain decoding. Pattern Recognit. 45, 2075–2084.

Penny, W.D., Stephan, K.E., Mechelli, A., Friston, K.J., 2004. Comparing dynamic causal models. Neuroimage 22, 1157–1172.
Pereira, F., Botvinick, M., 2011. Information mapping with pattern classifiers: a comparative study. Neuroimage 56, 476–496.
Pereira, F., Mitchell, T., Botvinick, M., 2009. Machine learning classifiers and fMRI: a tutorial overview. Neuroimage 45, S199–S209.
Schurger, A., Pereira, F., Treisman, A., Cohen, J.D., 2010. Reproducibility distinguishes conscious from nonconscious neural representations. Science 327, 97–99.
Sitaram, R., Weiskopf, N., Caria, A., Veit, R., Erb, M., Birbaumer, N., 2008. fMRI brain–computer interfaces: a tutorial on methods and applications. IEEE Signal Process. Mag. 25, 95–106.
Stelzer, J., Chen, Y., Turner, R., 2013. Statistical inference and multiple testing correction in classification-based multi-voxel pattern analysis (MVPA): random permutations and cluster size control. Neuroimage 65, 69–82.
Stephan, K.E., Weiskopf, N., Drysdale, P.M., Robinson, P.A., Friston, K.J., 2007. Comparing hemodynamic models with DCM. Neuroimage 38, 387–401.
Stephan, K.E., Penny, W.D., Daunizeau, J., Moran, R.J., Friston, K.J., 2009. Bayesian model selection for group studies. Neuroimage 46, 1004–1017.
Tong, F., Pratte, M.S., 2012. Decoding patterns of human brain activity. Annu. Rev. Psychol. 63, 483–509.
Wickenberg-Bolin, U., Goransson, H., Fryknas, M., Gustafsson, M., Isaksson, A., 2006. Improved variance estimation of classification performance via reduction of bias caused by small sample size. BMC Bioinformatics 7, 127.
Woolrich, M.W., Behrens, T.E.J., Beckmann, C.F., Jenkinson, M., Smith, S.M., 2004. Multilevel linear modelling for FMRI group analysis using Bayesian inference. Neuroimage 21, 1732–1747.
Zhang, D., Lee, W.S., 2008. Learning classifiers without negative examples: A reduction approach. International Conference on Digital, Information Management (ICDIM).

SUPPLEMENTAL MATERIAL

Variational Bayesian mixed-effects inference for classification studies

Brodersen, Daunizeau, Mathys, Chumbley, Buhmann, and Stephan

MCMC algorithm

The variational Bayes scheme presented in the main text is computationally highly efficient. However, its results are only exact to the extent to which its distributional assumptions are justified. To validate these assumptions, we compared VB to an asymptotically exact stochastic approach. Markov chain Monte Carlo (MCMC) is one such approach; it is computationally much more expensive than variational Bayes but exact in the limit of infinite runtime.

Here, we describe a Gibbs sampler for inverting the univariate normal-binomial model introduced in the main text. This algorithm is analogous to the one we previously introduced for the inversion of the bivariate normal-binomial model in Brodersen et al. (2012) but uses the new parameterization introduced in the present manuscript. It proceeds by cycling over model parameters, drawing samples from their full-conditional distributions, until the desired number of samples (e.g., 10^6) has been generated.

The algorithm is initialized by drawing initial samples from overdispersed starting distributions:

$$\mu^{(0)} \leftarrow N\left(\mu^{(0)} \mid 0,\ 3\right) \tag{1}$$

$$\lambda^{(0)} \leftarrow G\left(\lambda^{(0)} \mid 1,\ \tfrac{1}{10}\right) \tag{2}$$

$$\rho^{(0)} \leftarrow N_m\left(\rho^{(0)} \mid 0,\ 3 \times I_{m \times m}\right) \tag{3}$$

(We use the notation $x \leftarrow p\left(x \mid \theta\right)$ to denote the process of sampling a new value from the probability distribution $p\left(x \mid \theta\right)$ [a distribution in x with parameters θ] and assigning it to the variable x.)

Next, to obtain a sample from the first posterior of interest, $p(\mu \mid k)$, we draw from the full-conditional[1] distribution $p(\mu \mid \lambda^{(\tau-1)}, \rho^{(\tau-1)})$. Since the Gaussian prior $p(\mu)$ is conjugate with

[1] In Gibbs sampling, the *full-conditional* distribution of a model parameter refers to the conditional posterior given the data and all model parameters other than the one under consideration.

respect to the likelihood $p(\rho_j \mid \mu, \lambda)$, the full-conditional posterior (i.e., is the distribution from which $\mu^{(\tau)}$ is sampled) is available in closed form,

$$\mu^{(\tau)} \leftarrow N\left(\mu^{(\tau)} \mid \frac{\eta_0}{\eta_0 + m\lambda^{(\tau-1)}} \mu_0 + \frac{m\lambda^{(\tau-1)}}{\eta_0 + m\lambda^{(\tau-1)}} \overline{\rho}^{(\tau-1)},\ \eta_0 + m\lambda^{(\tau-1)} \right). \tag{4}$$

In the above distribution, μ_0 and η_0 represent the prior population mean and precision, $\lambda^{(\tau-1)}$ is the latest sample from the population precision, and $\overline{\rho}^{(\tau-1)}$ is the sample average over the components of the latest samples from subject-specific accuracies. Thus, as is typical of Bayesian inference, both moments of the full-conditional distribution embody the balance between prior precision η_0 and data precision $m\lambda^{(\tau-1)}$.

Having drawn a sample from $p(\mu \mid k)$, we next turn to the problem of sampling from $p(\lambda \mid k)$. For this we consider the full-conditional distribution $p(\lambda \mid \mu^{(\tau)}, \rho^{(\tau-1)})$. As above, the choice of a conjugate prior yields a closed-form posterior,

$$\lambda^{(\tau)} \leftarrow \mathrm{Ga}\left(\lambda^{(\tau)} \mid a_0 + \frac{m}{2},\ b_0 + \frac{1}{2} \sum_{j=1}^{m} \left(\rho_j^{(\tau-1)} - \mu^{(\tau)} \right)^2 \right), \tag{5}$$

where $\rho_j^{(\tau-1)}$ represents the latest sample from the posterior accuracy in subject j, and where $\mu^{(\tau)}$ is the sample drawn in (4).

Finally, to sample from the subject-specific posteriors $p(\rho_j \mid k)$, we consider each subject's full-conditional distribution $p(\rho_j \mid \mu^{(\tau)}, \lambda^{(\tau)}, k_j)$ in turn. Since a closed-form expression is not available for these distributions, we embed a Metropolis-Hastings step into our Gibbs sampler. This step can be implemented by drawing a candidate sample from a (symmetric) proxy density

$$\rho_j^* \leftarrow N\left(\rho_j^* \mid \rho_j^{(\tau-1)},\ 2^2 \right), \tag{6}$$

where the choice of variance of the proxy density was guided empirically to balance exploration and exploitation of the resulting Markov chain (cf. Brodersen et al., 2012). The sample drawn in (6) is accepted as the next $\rho_j^{(\tau)}$ with probability

$$\min\left\{ 1, \frac{\mathrm{Bin}\left(k_j \mid \sigma(\rho_j^*), n_j \right) N\left(\rho_j^* \mid \mu^{(\tau)}, \lambda^{(\tau)} \right)}{\mathrm{Bin}\left(k_j \mid \sigma(\rho_j^{(\tau-1)}), n_j \right) N\left(\rho_j^{(\tau-1)} \mid \mu^{(\tau)}, \lambda^{(\tau)} \right)} \right\}. \tag{7}$$

Iterating over all three above steps yields a series of samples $(\mu^{(\tau)}, \lambda^{(\tau)}, \rho^{(\tau)})$ whose empirical joint distribution approaches the true posterior $p(\mu, \lambda, \rho \mid k)$ in the limit of an infinite number of samples.

Inferences presented in the paper were based on 100,000 samples, generated using 8 parallel chains (cf. Brodersen et al., 2012). Assessing convergence, we found that with these settings the average ratio of within-chain variance to between-chain variance was bigger than 0.99. In other words, the variances of samples within and between chains were practically indistinguishable. The Metropolis rejection rate was approx. 0.3, thus ensuring an appropriate balance between exploration (of regions with a lower density or other potential modes) and exploitation (of regions with a higher density).

Supplemental references

Brodersen KH, Mathys C, Chumbley JR, Daunizeau J, Ong CS, Buhmann JM, Stephan KE (2012) Bayesian mixed-effects inference on classification performance in hierarchical data sets. Journal of Machine Learning Research 13:3133–3176.