Diss. ETH No. 20687

Generative embedding and variational Bayesian inference for multivariate time series

by

Kay Henning Brodersen

MSc in Information Systems, University of Muenster

MSc in Neuroscience, University of Oxford

born 9 December 1982

A thesis submitted for the degree of

Doctor of Sciences

ETH Zurich

accepted on the recommendation of

Professor Joachim M. Buhmann, examiner, ETH Zurich

Professor Klaas E. Stephan, co-examiner, ETH Zurich

Professor Zoubin Ghahramani, co-examiner, University of Cambridge

October 2012

Abstract

Multivariate time series can be modelled using differential equations that describe how the components of an underlying dynamical system interact in time. A challenging domain of application is neuroscience, where *dynamic causal models* have been increasingly used to shed light on the mechanisms behind multivariate time series of brain activity acquired in the healthy and the diseased human brain. This thesis introduces an approach to translating such models into clinical applications which we refer to as *generative embedding*. Our approach exploits the notion that a mechanistically interpretable description of a system may provide more useful insights than the observed time series themselves. Conceptually, we begin by developing a *model-based classification* approach that is based on the combination of a generative model and a discriminative classifier. We show that this approach may lead to significantly more accurate diagnostic classifications and deeper mechanistic insights than previous schemes. Using a classifier on hierarchical data, as we do here, requires us to revisit conventional approaches to performance evaluation. We introduce novel Bayesian *fixed-effects* and *mixed-effects* models for inference on classification performance that correctly account for distinct sources of uncertainty to appropriately constrain posterior inferences. We propose to replace conventional classification accuracies by *balanced accuracies* whenever the data are not perfectly balanced themselves. We demonstrate the properties of these models using stochastic approximate inference based on Markov chain Monte Carlo. We then derive a computationally highly efficient deterministic *variational Bayes* approximation. Complementary to its use in classification, generative embedding may enable the discovery of mechanistically interpretable subgroups that were not known *a priori*. We develop a *model-based clustering* approach which we use to dissect a group of patients diagnosed with schizophrenia into subgroups with clinical validity. In summary, this thesis explores generative embedding and variational Bayesian inference to establish the conceptual, statistical, and computational foundations for utilizing model-based classification and clustering approaches in a clinical context. We envisage that future applications of our approach will enable the formulation of novel mechanistic hypotheses that decompose groups of patients with similar symptoms into pathophysiologically distinct subgroups.

Contents

List of Figures

Contents in detail

Chapter 1

Introduction

Multivariate time series can be modelled using differential equations that describe how the elements of an underlying dynamical system interact in time. One novel, highly challenging, and increasingly promising domain of application is clinical neuroscience, where dynamic models can be used to describe physiological mechanisms underlying multivariate time series of brain activity in the healthy and the diseased human brain.

This thesis introduces a novel approach to translating such models into clinical applications which we refer to as *generative embedding*. Our approach exploits the notion that, in order to understand a system and extract useful information from it, a mechanistically interpretable description may prove much more useful than the observed time series themselves. Our approach is multivariate and thus allows us to utilize information jointly encoded by multiple features of the system.

In brief, the central theme of this thesis is to (i) propose concrete implementations of model-based classification and clustering on the basis of generative embedding, (ii) develop a theory around the statistical evaluation of the obtained classification algorithms, and (iii) demonstrate the utility of the proposed approach in the context of functional neuroimaging data.

1.1 Statistical approach and applications

Recent years have seen a substantial increase in the use of functional neuroimaging for investigating healthy brain function and examining its pathophysiological deviations. The most popular type of analysis is *statistical*

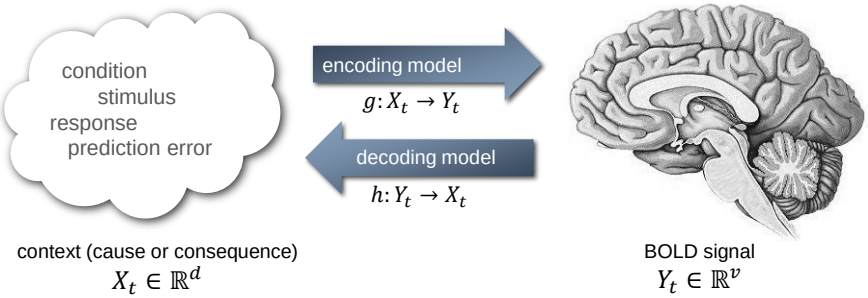

Figure 1.1: Encoding vs. decoding. An *encoding* model describes the conditional density of brain activity given a set of contextual variables, such as sensory inputs, behavioural responses, or states of a computational model of cognition. A *decoding* model adopts the inverse view: it describes the conditional density of a contextual variable in terms of brain activity. In order to establish the existence of a statistical relationship between context and brain activity, the direction of inference is not important. Decoding models, however, are more suitable when we aim to exploit such a relationship to afford predictions about a clinical variable in an individual subject.

parametric mapping (SPM; Friston *et al.*, 1995), a mass-univariate encoding model of functional magnetic resonance imaging (fMRI) data in which the statistical relationship between experimental variables and haemodynamic measurements of neural activity is examined independently for every voxel in the brain (Figure 1.1). One could use SPM, for example, to create a map showing in which parts of the brain activity levels differ significantly between patients and healthy controls.

While this approach has led to fundamental insights about functional abnormalities in psychiatric and neurological disorders, its scope is limited in two ways. First, since univariate models are insensitive to spatially distributed patterns of neural activity, they may fail to detect subtle, distributed differences between patients and healthy controls that are not expressed as local peaks or clusters of activity (Koutsouleris *et al.*, 2009). Second, while *encoding* models such as SPM are excellent for describing regional differences in brain activity across clinical groups, they are less well suited for clinical decision making, where the challenge is to predict the disease state of an *individual* subject from measured brain activity (Figure 1.1).

An alternative approach is offered by multivariate *decoding* methods. Unlike mass-univariate encoding models, these methods predict an experi-

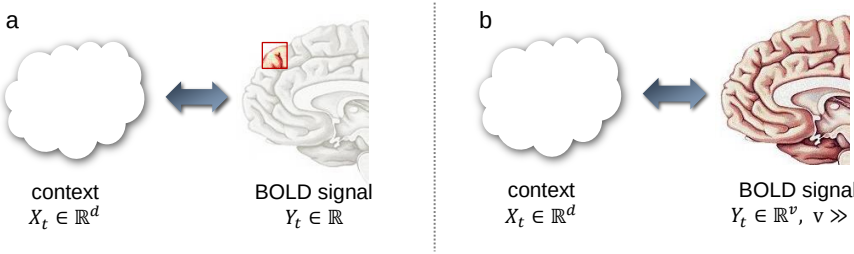

Figure 1.2: Univariate vs. multivariate models I. (a) In neuroimaging, a model that considers brain activity in an individual voxel is referred to as a *univariate* (or univoxel) model, and a model that is applied separately to each voxel in turn as a *mass-univariate* model. (b) A *multivariate* (or multivoxel) model, in contrast, considers several or all voxels simultaneously. Since most models relate several independent variables to one dependent variable, univariate models tend to come in the form *encoding* models, whereas multivariate models are typically, though not always, *decoding* models.

mental variable (e.g., a trial-specific condition, or a subject-specific disease state) from the activity pattern across voxels (Figures 1.2 and 1.3; see Norman *et al.*, 2006; Haynes and Rees, 2006; O'Toole *et al.*, 2007; Friston *et al.*, 2008; Pereira *et al.*, 2009, for reviews). The technique often rests upon the application of algorithms for pattern classification to neuroimaging data. A classifier is first trained on data from a set of subjects (or trials) with known labels (e.g., disease state A vs. B). It is then tested on new subjects that were not seen during training. Successful above-chance classification performance provides evidence that information about a particular brain state can indeed be decoded from the acquired data (see Fu *et al.*, 2008; Shen *et al.*, 2010; Wang *et al.*, 2010, for examples).[1]

Using multivariate *decoding* models instead of mass-univariate *encoding* models has interesting potential for clinical practice, particularly for diseases that are difficult to diagnose. Consequently, much work is currently being invested in constructing classifiers that can predict the diagnosis of individual subjects from structural or functional brain data (Ford *et al.*, 2003; Fan *et al.*, 2007; Fu *et al.*, 2008; Fan *et al.*, 2008a; Klöppel *et al.*,

[1]Throughout this thesis, the term 'above-chance classification' refers to a classification result whose estimate of generalization ability is significantly above the chance level. This implies, in particular, that the significance of an accuracy estimate (e.g., 85%) can only be judged in relation to the underlying number of test cases. See Chapters 3 and 4 for a detailed treatment.

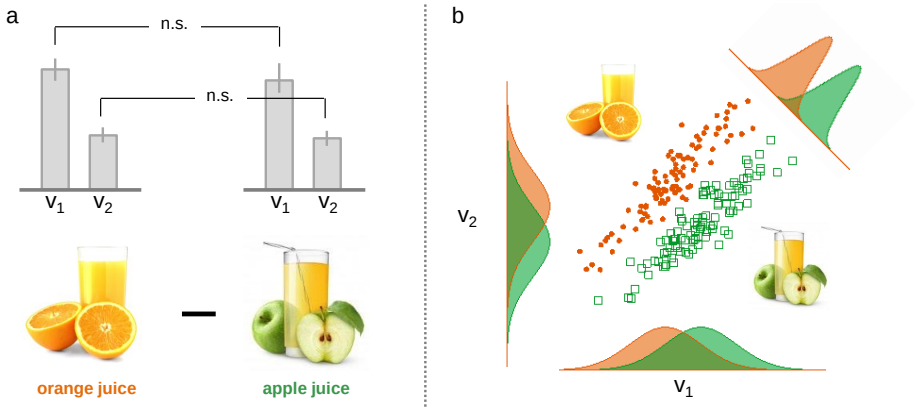

Figure 1.3: Univariate vs. multivariate models II. (a) Individual variables (or voxels) may not reveal significant differences in activity across conditions when treated independently. (b) When considering voxels jointly, clear separability of conditions may emerge. This is why multivariate models can be more powerful than univariate models.

2008, 2009; Shen *et al.*, 2010; Klöppel *et al.*, 2012). Historically, these efforts date back to positron emission tomography (PET) studies in the early 1990s (O'Toole *et al.*, 2007). Today, attempts of using multivariate classifiers for subject-by-subject diagnosis largely focus on MRI and fMRI data (Ford *et al.*, 2003; Fu *et al.*, 2008; Fan *et al.*, 2007, 2008b). In this broad domain, the present thesis attempts to solve the problem of clinical applicability in a very different fashion than conventional methods have done. It is motivated by two current challenges, as described next.

Current challenges. Despite their increasing popularity, two challenges critically limit the practical applicability of current classification and clustering methods for functional neuroimaging data. These are, as discussed below, (i) the high dimensionality of the data and (ii) the lack of interpretability of conventional solutions.

First, classifying or clustering subjects directly in voxel space is difficult. This is because functional neuroimaging datasets typically exhibit a very low signal-to-noise ratio; they are obtained in an extremely high-dimensional measurement space; and they are characterized by a severe mismatch between the large number of voxels and the small number of available subjects. In the case of fMRI, for instance, a whole-brain scan

may easily contain more than 100 000 voxels, whereas the number of experimental repetitions (i.e., trials or subjects) is usually on the order of tens. This is not an issue for procedures that rely, for instance, on Bayesian model inversion and therefore do not overfit. However, an extreme mismatch between features and data points does create problems for many models that explicitly or implicitly rely on parameter estimation rather than inversion. These models typically require carefully designed algorithms for reducing the dimensionality of the feature space without averaging out informative activity. The underlying challenge is the problem of *reconstruction* (to afford interpretability) or *feature selection* (to avoid overfitting).

Since an exhaustive search of the entire space of feature subsets is statistically unwarranted and computationally intractable, various heuristics have been proposed. One common approach, for example, is to simply include only those voxels whose activity, when considered by itself, significantly differs between classes within the training set (Cox and Savoy, 2003). This type of *univariate* feature selection is computationally efficient but fails to find voxels that only reveal information when considered as an ensemble (for an example of an intermediate strategy, see Brodersen, Wiech, Lomakina *et al.*, *in preparation*). Another method, termed searchlight analysis, finds those voxels whose local environment allows for above-chance classification (Kriegeskorte *et al.*, 2006). Unlike the first approach, searchlight feature selection is *multivariate*, but it fails to detect more widely distributed sets of voxels that jointly encode information about the variable of interest.

Further strategies include: preselecting voxels based on an anatomical mask (Haynes and Rees, 2005; Kamitani and Tong, 2005) or a separate functional localizer (Cox and Savoy, 2003; Serences and Boynton, 2007); spatial subsampling (Davatzikos *et al.*, 2005); finding informative voxels using univariate models (Fu *et al.*, 2008; Ford *et al.*, 2003; Fan *et al.*, 2007) or locally multivariate searchlight methods (Kriegeskorte *et al.*, 2006; Haynes *et al.*, 2007); and unsupervised dimensionality reduction (Shen *et al.*, 2010; Mourao-Miranda *et al.*, 2005). Some have attempted to account for the inherent spatial structure of the feature space (Kriegeskorte *et al.*, 2006; Soon *et al.*, 2009; Grosenick *et al.*, 2009) or use voxel-wise models to infer a particular stimulus identity (Kay *et al.*, 2008; Mitchell *et al.*, 2008; Formisano *et al.*, 2008). Others have adopted a regularization perspective, e.g., using automatic relevance determination (ARD; Yamashita *et al.*, 2008) or sparsity constraints (Grosenick *et al.*, 2008; van Gerven *et al.*, 2009). Finally, those submissions that performed best in the Pittsburgh Brain Activity Interpretation Competition (PBAIC 2007) highlighted the utility of kernel

ridge regression (Chu *et al.*, 2010) and relevance vector regression (Chu *et al.*, 2010; Valente *et al.*, 2010).

The common assumption underlying all of these approaches is that interesting variations of the data with regard to the class variable are confined to a manifold that populates a latent space of much lower dimensionality than the measurement space. However, most of these methods are only loosely constrained by rules of biological plausibility. As a result, they may easily lead to rather arbitrary subsets of selected voxels: deemed informative by the classifier, but next to impossible to interpret physiologically.

This is the second challenge in conventional classification methods for clinical applications: the interpretation of their results. Classification algorithms *per se* yield predictions and can be used to establish a statistical relationship between (multivariate) neural activity and a (univariate) variable of interest. The ability to make predictions is indeed the primary goal in fields concerned with the design of brain-machine interfaces (Sitaram *et al.*, 2007), algorithms for lie detection (Davatzikos *et al.*, 2005; Kozel *et al.*, 2005; Bles and Haynes, 2008; Krajbich *et al.*, 2009), or black-box tools for clinical diagnostics (Klöppel *et al.*, 2012). We argue, however, as others have done before (cf. Friston *et al.*, 2008), that cognitive and clinical neuroscience should not merely be aimed at maximizing prediction accuracy. Rather, a more important goal is to make inferences on structure-function mappings in the brain and to generate hypotheses for drug development and treatment. High prediction accuracies may serve as an accompanying measure of the amount of information that can be extracted from neural activity, but should not represent the only goal (Figure 1.4).

Most classification studies to date attempt to draw conclusions from overall prediction accuracies (Mitchell *et al.*, 2003; Ford *et al.*, 2003), the spatial deployment of informative voxels (Kamitani and Tong, 2005, 2006; Haynes and Rees, 2005; Hampton and O'Doherty, 2007; Kriegeskorte *et al.*, 2007; Grosenick *et al.*, 2008; Hassabis *et al.*, 2009; Howard *et al.*, 2009), the temporal evolution of discriminative information (Polyn *et al.*, 2005; Grosenick *et al.*, 2008; Bode and Haynes, 2009; Harrison and Tong, 2009; Soon *et al.*, 2009), or patterns of undirected regional correlations (Craddock *et al.*, 2009).

These approaches may support discriminative decisions; they may allow for the construction of information maps showing discriminative features; but they are blind to the neuronal mechanisms (such as effective connectivity or synaptic plasticity) that underlie discriminability of brain or disease states. Mechanistic conclusions that relate to biologically meaningful enti-

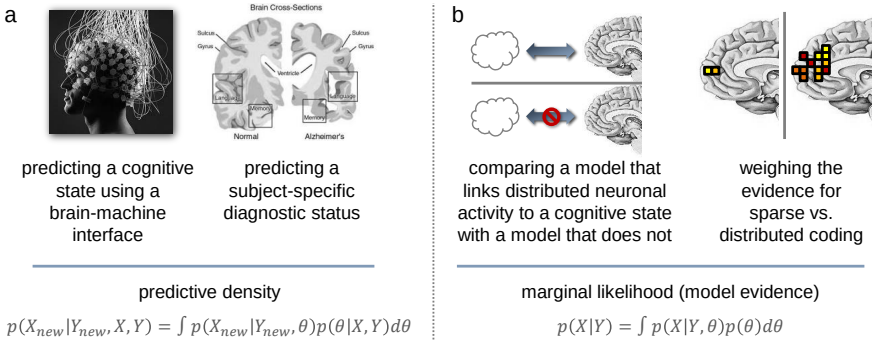

Figure 1.4: Prediction vs. inference. (a) The goal of *prediction* is to maximize predictive accuracy, which is the primary aim in applications such as brain-machine interfaces and automated black-box diagnostics. (b) The goal of *inference* is to compare the evidence for different models of structure-function mappings in the brain. Here, we emphasize that the second goal should not be neglected even when using classification algorithms that intrinsically focus on the first.

ties such as brain connectivity or synaptic plasticity are hard to draw. In other words: while some conventional classification studies have achieved impressive diagnostic accuracy (cf. Klöppel *et al.*, 2012), their results have not improved our mechanistic understanding of disease processes.

In summary, classification algorithms and their underlying decoding models have been increasingly used to infer cognitive or clinical brain states from measures of brain activity. The practicality of current classifiers, however, is restricted by two major challenges. First, due to the high data dimensionality and low sample size, algorithms struggle to separate informative from uninformative features, resulting in poor generalization performance. Second, popular classification methods, applied to voxel-based feature spaces, rarely afford mechanistic interpretability.

Generative embedding. This thesis describes a model-based analysis approach, based on the idea of generative embedding, that may provide a solution to the two challenges outlined above. In brief, our approach incorporates neurobiologically interpretable generative models into discriminative classification and clustering algorithms to provide a potential foundation for long-term utility in clinical practice. The specific implementation of generative embedding proposed in this thesis consists of six conceptual steps which we summarize below (see Figure 1.5).

The analysis begins, in step 1, by extracting time series of measurements from regions of interest. In step 2, the data are explained in terms of a generative model. Model inversion can be viewed as a mapping $\mathcal{X} \to \mathcal{M}_\Theta$ that projects an example $x \in \mathcal{X}$ from data space onto a multivariate probability distribution in a parametric family $\mathcal{M}_\Theta$. Crucially, the model is designed to accommodate observations gathered from all classes, and therefore, when being inverted, it remains oblivious to the class a given example stems from.

In step 3, a probability kernel $k_\mathcal{M} : \mathcal{M}_\Theta \times \mathcal{M}_\Theta$ is constructed that represents a similarity measure between two inverted DCMs. This step can be split up into an initial mapping $\mathcal{M}_\Theta \to \mathbb{R}^d$ followed by a vectorial kernel $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$. The kernel implies a model-based feature space, or so-called generative score space, that yields a comprehensive statistical summary representation of every subject. In the illustrative participant shown in Figure 1.5, the effective influence of region A on region B as well as the self-connection of region B were particularly strong. This kernel is used in step 4 for classification or clustering. For example, in the presence of known external labels, one could employ a support vector machine to distinguish between patients and healthy controls.

In step 5, the performance of the algorithm is evaluated with respect to known labels, using validation measures such as classification accuracy or clustering purity. The model-based feature space can, in a final step 6, be investigated to examine which model parameters jointly contributed most to the distinction between subgroups. In the example in Figure 1.5, the influence of A on B and C were jointly most informative in distinguishing between the two groups.

Dynamic causal models as generative models. The specific approach to creating generative models of brain activity time series in this thesis is based on dynamic causal modelling (DCM; Friston *et al.*, 2003). DCM views the brain as a nonlinear dynamical system of interconnected neuronal populations whose directed connection strengths may be modulated by external perturbations (i.e., experimental conditions) or endogenous activity. Specifically, DCM describes how the dynamics within interconnected populations of neurons evolve over time and how their (causal) interactions change as a function of external inputs. DCM was originally introduced for fMRI data (Friston *et al.*, 2003) but has subsequently been implemented for a variety of measurement types, such as event-related potentials or spectral densities obtained from electrophysiological measurements (David *et al.*, 2006; Kiebel

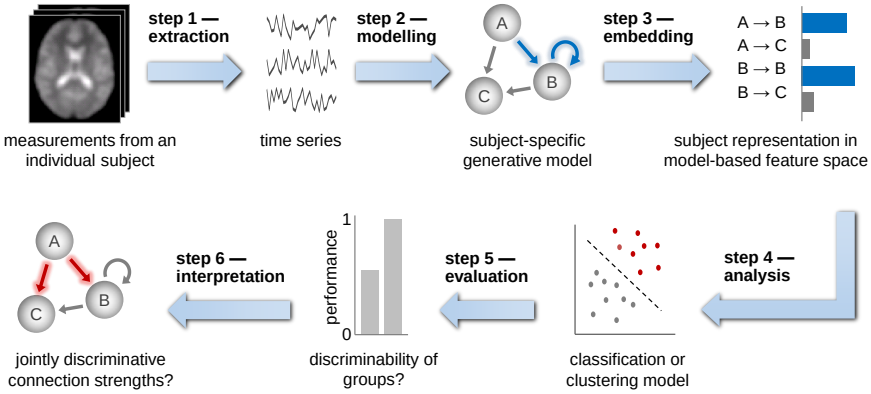

Figure 1.5: Generative embedding and model-based analyses. As described in the main text, this schematic illustrates the six conceptual steps by which generative embedding enables model-based analyses.

et al., 2009; Moran *et al.*, 2009). In this thesis, we will use DCM to map high-dimensional time series (i.e., fMRI or electrophysiological recordings) onto low-dimensional vectors of parameter estimates.

Advantages over conventional methods. Generative embedding offers three advantages over conventional analysis methods. First, it rests upon a principled and biologically informed way of creating a feature space. As a result, it may provide more accurate predictions by exploiting discriminative information encoded in 'hidden' physiological quantities such as synaptic connection strengths (e.g., Section 5.5.5).

Another advantage is that results can be interpreted in the context of a mechanistic model. Thus, the approach affords mechanistic interpretability of clinical classification and clustering solutions (e.g., Section 5.5.6 or Section 6.4).

The third advantage of generative embedding is that it may supplement evidence-based model-selection approaches, such as Bayesian model selection (BMS), in two ways: (i) it enables model-based decoding when discriminability of trials or subjects is not afforded by differences in model structure but only by differences in patterns of parameter estimates under the same model structure; and (ii) it enables structural model selection in cases where BMS is not applicable (see p. 195 in Chapter 7).

Long-term application: dissecting psychiatric conditions. Neurological and psychiatric spectrum disorders are typically defined in terms of particular *symptom* sets, despite increasing evidence that the same symptom may be caused by very different *pathologies*. Pathophysiological discovery, classification, and effective treatment of such disorders will increasingly require a mechanistic understanding of inter-individual differences and clinical tools for making accurate diagnostic inferences in individual patients.

In contrast to previous classification studies based on descriptive measures, which typically do not afford pathophysiological insights, generative embedding may enable the discovery of mechanistically interpretable subgroups that are defined in terms of hidden physiological quantities such as synaptic connection strengths. We argue that this can be achieved using a combination of two primary analysis types, as introduced below.

Analysis type 1: model-based classification. In model-based classification, a generative model is combined with a discriminative classifier. A preview of the sort of results that can be obtained from this analysis type are shown in Figure 1.6. In this example, we distinguished a group of stroke patients with moderate aphasia from a group of healthy controls, using a DCM of fMRI activity recorded during a passive speech-listening task. Generative embedding achieved a near-perfect balanced classification accuracy of 98% (sensitivity 100%, specificity 96%) and significantly outperformed conventional activation-based and correlation-based methods. This example demonstrates how disease states can be detected with very high accuracy and, as we will see in Section 5.5, be interpreted mechanistically in terms of abnormalities in connectivity.

Using classification algorithms on hierarchical datasets, as we do here, requires us to revisit traditional approaches to performance evaluation. We propose novel *fixed-effects* and *mixed-effects* models for inference on classification performance. We propose to replace conventional classification accuracies by *balanced accuracies* whenever the data are not perfectly balanced themselves. We illustrate the properties of these models using a stochastic approximation based on Markov chain Monte Carlo. We then derive a computationally more efficient deterministic approximation using variational Bayes.

Application type 2: model-based clustering. Classification analyses, as described above, may provide evidence that we can relate patterns of model features to a known external variable, such as a disease state.

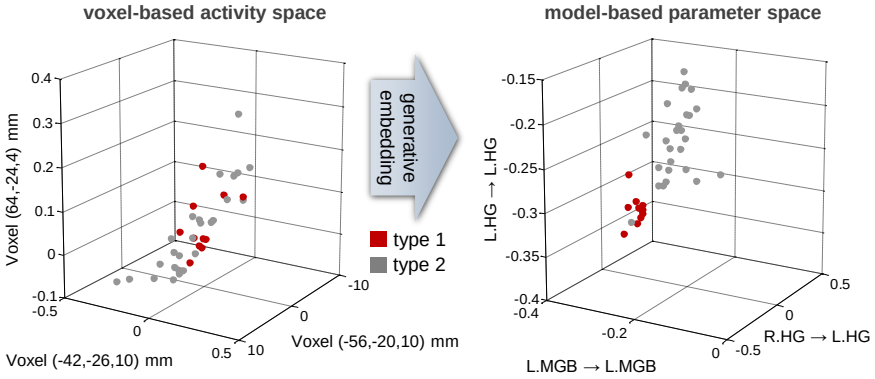

Figure 1.6: Detecting a remote lesion using generative embedding. As detailed in Section 5.5, we analysed fMRI data from healthy volunteers and stroke patients with moderate aphasia. We found that subject-specific directed connection strengths among cortical regions involved in speech processing contained sufficiently rich discriminative information to enable accurate predictions of the diagnostic category (healthy or aphasic) of a previously unseen individual. Compared to a feature space of voxel-wise activity (left), the induction of a generative score space (right) facilitates both separability and interpretability (cf. Section 5.23).

Initially, however, these labels are absent in precisely those applications in which one may expect our approach to unfold its greatest utility: in the domain of psychiatric spectrum disorders. These disorders are typically diagnosed in terms of symptoms. Clearly, following a questionnaire that leads to a diagnosis will always remain simpler and cheaper than acquiring fMRI data followed by the application of a pattern-recognition algorithm. Put differently, the costly reproduction of a potentially flawed label is of no use for practical clinical applications.

We must therefore go one step further and complement classification schemes by unsupervised analyses (Figure 1.7). What groups would we *discover* if no external label information was present? Can we use model parameters to *generate hypotheses* about subtypes within a group of patients sharing the same symptoms?

It may seem ironic that, in order to validate any clustering solution, we must return to known external variables, e.g., a known disease subtype. Here, we will motivate the measure of *balanced purity* which we will use to assess how well a clustering solution agrees with known structure in the population.

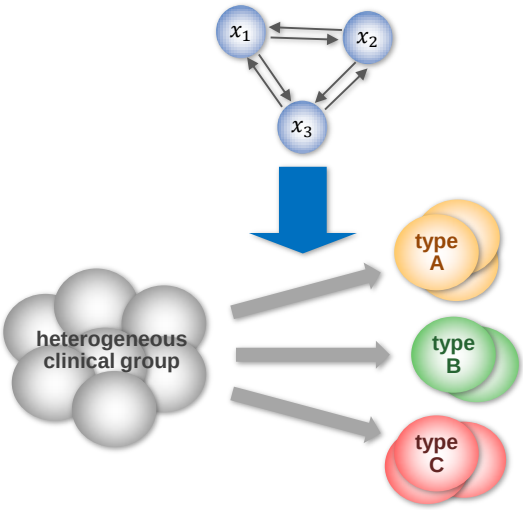

Figure 1.7: Long-term ambition. Psychiatry lacks pathophysiologically informed diagnostic classifications. This problem is particularly important in the domain of spectrum disorders. Generative embedding may help dissect such disorders into mechanistically defined subgroups. An initial proof of principle will be provided in Chapter 6.

We thus argue that model-based classification and model-based clustering must be used in conjunction to allow the utility of generative embedding for the clinic to unfold. Classification is used to demonstrate the feasibility of relating a pattern of connection strengths to an external variable. Clustering is then used to establish subgroups when such external variables are not available or flawed. Finally, if a clustering solution has been shown to relate to clinically relevant variables (e.g., treatment response or clinical outcome), we may return to classification and develop a diagnostic aid for clinical practice.

1.2 Goals of this thesis

In summary of the strategy outlined above, this thesis aims to develop and test the following research hypotheses:

- that model-based classification on the basis of generative embedding may provide more accurate discrimination than conventional classifi-

cation schemes;

- that model-based analysis results offer deeper mechanistic insights than conventional analyses (in the sense of providing an explanation in terms of components of a dynamical system and their causal interactions);

- that group classification analyses require mixed-effects inference and that this form of inference offers greater statistical sensitivity and higher estimation accuracy than fixed-effects or random-effects models;

- that model-based clustering may provide interpretable groupings that relate to relevant external variables such as clinical symptom scores.

1.3 Structure of this thesis

Chapter 2. We begin by describing the conceptual aspects of generative embedding using the example of dynamic causal models for neuroimaging data.

Chapter 3. A major application of generative embedding is model-based classification, an important aspect of which is performance evaluation. We highlight the flaws associated with contemporary ways of reporting classification performance. We argue that classification accuracies should be replaced by balanced accuracies for which we present a fully Bayesian framework.

Chapter 4. Classification algorithms in neuroimaging are typically employed in a group setting, which has important implications for their correct statistical evaluation. We begin by exposing the theory underlying fixed-effects, random-effects, and mixed-effects inference. We then propose several hierarchical Bayesian models for mixed-effects inference on classification accuracies and balanced accuracies. We illustrate the properties of these models using a stochastic approximation based on Markov chain Monte Carlo (MCMC). We then derive a computationally more efficient deterministic approximation using variational Bayes (VB).

Chapter 5. Using DCM as a generative model and a support vector machine as a discriminative classifier, we illustrate the utility of model-based classification using three datasets. In the first two studies, we use a generative model of local field potentials (LFP) in rodents to decode the trial-wise identity of a sensory stimulus from activity in somatosensory and auditory cortex. In the third study, we infer the presence or absence of a remote lesion from healthy brain regions, using fMRI in healthy participants and stroke patients with aphasia. In brief, we will see that (i) generative embedding yields a near-perfect classification accuracy, (ii) significantly outperforms conventional 'gold standard' activation-based and correlation-based classification schemes, and (iii) affords a novel mechanistic interpretation of the differences between aphasic patients and healthy controls during processing of speech.

Chapter 6. Finally, we turn to generative embedding and its use in the domain of unsupervised learning. We show how a procedure for model-based clustering enables the discovery of subgroups that are defined in terms of 'hidden' physiological quantities such as synaptic connection strengths. We envisage that future applications of the approach proposed in this thesis may become relevant for generating novel mechanistic hypotheses for clinical applications, by decomposing groups of patients with similar symptoms into pathophysiologically distinct subgroups.

1.4 Original contributions

- **Generative embedding for dynamic causal models.** We establish a generative embedding approach for use with dynamic causal models of brain function.

- **Balanced accuracy.** Classification accuracy is a poor measure of classification performance when the data are imbalanced. We propose that the balanced accuracy is a better indicator as it removes the bias that may arise when a classifier is trained and tested on an imbalanced dataset.

- **Mixed-effects inference on classification performance.** We develop fully Bayesian approaches for classification group studies that account for both fixed-effects (within-subjects) and random-effects

(between-subjects) variance components, thus affording mixed-effects inference.

- **Variational Bayesian approximate inference.** We derive a computationally highly efficient variational approximation to mixed-effects inference that is based on interpretable update equations.

- **Model-based classification.** We demonstrate the utility of generative embedding by classifying trial-wise and subject-wise states on the basis of parameter estimates.

- **Model-based clustering.** We finally illustrate how the submission of parameter estimates to a clustering algorithm makes it possible to generate novel hypotheses about mechanistically defined subgroups of a disease.

1.5 Publications

The research underlying this thesis has been published, or is in the process of being published, as listed below.

First-author peer-reviewed papers

1. Brodersen, K. H., Lin, Z., Gupta, A., Deserno, L., Penny, W. D., Schlagenhauf, F., Buhmann, J. M., and Stephan, K. E. (*in preparation*). Model-based clustering.

2. Brodersen, K. H., Daunizeau, J., Mathys, C., Chumbley, J. R., Buhmann, J. M., and Stephan, K. E. (*under review*). Variational Bayesian mixed-effects inference for classification group studies.

3. Brodersen, K. H., Mathys, C., Chumbley, J. R., Daunizeau, J., Ong, C. S., Buhmann, J. M., and Stephan, K. E. (*in press*). Mixed-effects inference on classification performance in hierarchical datasets. *Journal of Machine Learning Research*.

4. Brodersen, K. H., Wiech, K., Lomakina, E. I., Lin, C.-S., Buhmann, J. M., Bingel, U., Ploner, M., Stephan, K. E., and Tracey, I. (2012). Decoding the perception of pain from fMRI using multivariate pattern analysis. *NeuroImage*, 63, 1162–1170.

5. Brodersen, K. H., Schofield, T. M., Leff, A. P., Ong, C. S., Lomakina, E. I., Buhmann, J. M., and Stephan, K. E. (2011). Generative embedding for model-based classification of fMRI data. *PLoS Computational Biology*, 7(6), e1002079.

6. Brodersen, K. H., Haiss, F., Ong, C. S., Jung, F., Tittgemeyer, M., Buhmann, J. M., Weber, B., and Stephan, K. E. (2011). Model-based feature construction for multivariate decoding. *NeuroImage*, 56(2), 601–615.

7. Brodersen, K. H., Ong, C. S., Stephan, K. E., and Buhmann, J. M. (2010). The balanced accuracy and its posterior distribution. *Proceedings of the 20th International Conference on Pattern Recognition*, pages 3121–3124. IEEE Computer Society.

8. Brodersen, K. H., Ong, C. S., Stephan, K. E., and Buhmann, J. M. (2010). The binormal assumption on precision-recall curves. *Proceedings of the 20th International Conference on Pattern Recognition*, pages 4263–4266. IEEE Computer Society.

First-author peer-reviewed talks and presentations

9. Brodersen, K. H., Lin, Z., Gupta, A., Penny, W. D., Leff, A. P., Chehreghani, M. H., Busetto, A.-G., Buhmann, J. M., and Stephan, K. E. (2012). Model-based clustering using generative embedding. Oral presentation at *Human Brain Mapping* 2012, Beijing, China. Awarded with a Trainee Abstract Award.

10. Brodersen, K. H., Daunizeau, J., Mathys, C., Chumbley, J. R., Buhmann, J. M., and Stephan, K. E. (2012). Variational Bayesian mixed-effects inference for classification studies. Presented at *Human Brain Mapping* 2012, Beijing, China. Awarded with a Trainee Abstract Award.

11. Brodersen, K. H., Gupta, A., Lin, Z., Lomakina, E. I., Buhmann, J. M., and Stephan, K. E. (2011). Clustering biological systems using generative embedding. Presented at SystemsX.ch 2011, Basel. Awarded with a Best Posters recognition.

12. Brodersen, K. H., Schofield, T. M., Leff, A. P., Ong, C. S., Lomakina, E. I., Buhmann, J. M., and Stephan, K. E. (2011). Generative embedding for model-based classification of fMRI data. Oral presentation at *Human Brain Mapping*, Quebec City, Canada. Awarded with a Trainee Abstract Award.

13. Brodersen, K. H., Chumbley, J. R., Mathys, C., Daunizeau, J., Buhmann, J. M., and Stephan, K. E. (2011). Mixed-effects inference on classification performance in group studies. Interactive session at *Human Brain Mapping* 2011, Quebec City, Canada. Awarded with a Trainee Abstract Award.

14. Brodersen, K. H., Lin, C.-S., Lomakina, E. I., Stephan, K. E., Wiech, K., and Tracey, I. (2011). Multivariate decoding of perceptual decisions about pain. Interactive session at *Human Brain Mapping* 2011, Quebec City, Canada.

15. Brodersen, K. H., Hunt, L. T., Lomakina, E. I., Rushworth, M. F. S., Behrens, T. E. J. (2011). Orbitofrontal cortex distributes reinforcement to the decision that caused it. Oral presentation at *Human Brain Mapping*, Quebec City, Canada.

16. Brodersen, K. H., Hunt, L. T., Lomakina, E. I., Rushworth, M. F. S., Behrens, T. E. J. (2011). The amygdala becomes reward-sensitive when an outcome cannot be assigned to the correct decision. Oral presentation at *Human Brain Mapping*, Quebec City, Canada.

17. Brodersen, K. H., Wiech, K., Lin, C.-S., and Tracey, I. (2010). Threat-dependent modulation of anterior insula connectivity predicts pain. Oral presentation at *Human Brain Mapping*, Barcelona, Spain.

18. Brodersen, K. H., Haiss, F., Ong, C. S., Jung, F., Allen, P., Tittgemeyer, M., Buhmann, J. M., McGuire, P., Weber, B., and Stephan, K. E. (2010). Model-based multivariate decoding and model selection. Presented at *Human Brain Mapping*, Barcelona, Spain. Awarded with a Trainee Abstract Award.

Chapter 2

Generative embedding and dynamic causal modelling

Generative embedding constitutes an approach to model-guided dimensionality reduction. It exploits the idea that both the performance and interpretability of classification and clustering approaches may benefit considerably from the incorporation of available prior knowledge about the process generating the observed data (see Shawe-Taylor and Cristianini, 2004, for an overview). Dynamic causal modeling offers one way of achieving this.

2.1 Generative embedding

Generative embedding rests on two components: a generative model for principled selection of mechanistically interpretable features on the one hand; and a discriminative method for classification or clustering on the other (see Figure 1.5 on p. 23). This chapter describes how this idea may become useful in clinical neuroimaging. For corresponding publications, see Brodersen *et al.* (2011a) and Brodersen *et al.* (2011b).[1]

[1]The term *generative embedding* is sometimes used to denote a particular model-induced feature space, or so-called generative score space, in which case the associated line of research is said to be concerned with generative embeddings. Here, we will use the term in singular form to denote the process of using a generative model to project

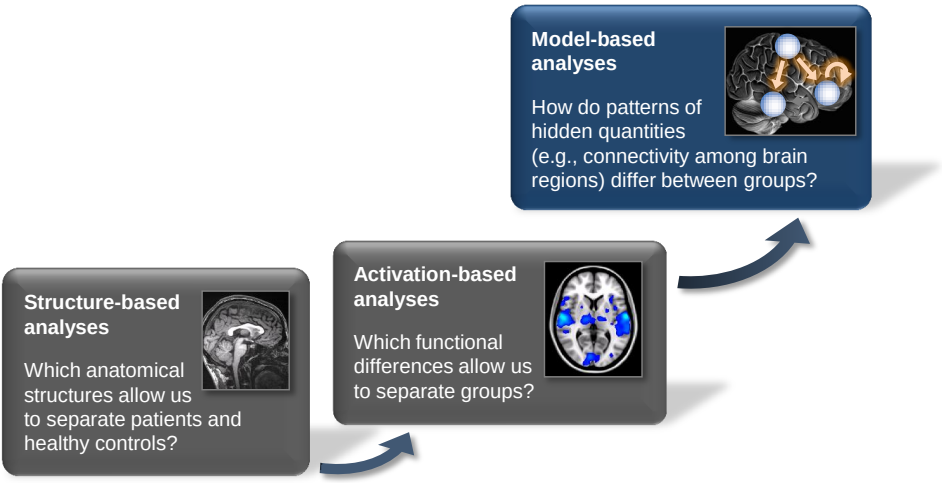

Figure 2.1: Analyses by data representation. Generative embedding can be viewed as an alternative to previous analysis approaches that were either based on *structural* or *functional* data. Generative embedding may be based on the same data as these previous approaches (here: functional data), but it recruits a model to account for the latent mechanisms by which the data were generated. In the case of neuroimaging, such models may, for instance, comprise variables describing *connectivity* among elements of a dynamical system.

Most current applications of classification algorithms in neuroimaging begin by embedding measured data in a d-dimensional Euclidean space $\mathbb{R}^d$. In fMRI, for example, a subject can be represented by a vector of d features, each of which corresponds to the signal measured in a particular voxel at a particular point in time. This approach makes it possible to use any learning algorithm that expects vectorial input; but it ignores the spatio-temporal structure of the data as well as the process that generated them. This limitation has motivated the search for kernel methods that provide a more natural way of measuring the similarity between the functional datasets of two subjects.

Generative models have proven powerful in explaining how observed data are caused by an underlying (neuronal) system.[2] One example in neu-

the data into a generative score space, rather than using the term to denote the space itself.

[2]Unlike discriminative models, generative models describe the joint density of the observed data and the parameters of a postulated generative process, rather than just

roimaging is *dynamic causal modelling* (DCM; Friston *et al.*, 2003). DCM aims to model observed time series by a system of parameterized differential equations with Gaussian observation noise. This approach enables statistical inference on physiological quantities that are not directly observable with current methods, such as directed interregional coupling strengths and their modulation, e.g., by synaptic gating (Stephan *et al.*, 2008).

From a pathophysiological perspective, disturbances of synaptic plasticity and neuromodulation are at the heart of psychiatric spectrum diseases such as schizophrenia (Stephan *et al.*, 2009b) or depression (Castren, 2005). It is therefore likely that classification and clustering of disease states could benefit from exploiting estimates of these quantities.

Generative embedding represents a special case of using *generative kernels*, such as the *P-kernel* (Haussler, 1999) or the *Fisher kernel* (Jaakkola and Haussler, 1999). Generative kernels have been fruitfully exploited in a range of applications (Bicego *et al.*, 2004; Jebara *et al.*, 2004; Hein and Bousquet, 2005; Cuturi *et al.*, 2006; Bosch *et al.*, 2006, 2008; Bicego *et al.*, 2009b; Smith and Niranjan, 2000; Holub *et al.*, 2005; Jaakkola *et al.*, 1999; Bicego *et al.*, 2009a; Hofmann, 2000) and define an active area of research (Minka, 2005; Lasserre *et al.*, 2006; Perina *et al.*, 2010; Martins *et al.*, 2010).

Generative kernels are functions that define a similarity metric for observed examples using a generative model. In the special case of generative embedding, a generative kernel is used to transfer the inverted models into a vectorial feature space in which an appropriate similarity metric is defined. This feature space, which we refer to as a *generative score space*, embodies a model-guided dimensionality reduction of the observed data. The kernel defined in this space could be a simple inner product of feature vectors, or it could be based on any other higher-order function, as long as it is positive definite (Mercer, 1909). Thus, while all generative kernels (e.g., the Fisher kernel) are based on an implicit feature space, generative embedding makes this feature space explicit. This exposition has the advantage that subsequent analysis results become interpretable, dimension by dimension, in relation to the underlying feature space.

Using a kernel for classification and clustering confers important advantages with respect to the general applicability of our approach. In particular, it is possible to define a problem-specific kernel and combine it with a general-purpose algorithm for discrimination. This makes our approach modular and easily applicable, e.g., to different acquisition modalities.

---

the conditional density of the data given the parameters.

2.2 Dynamic causal modelling

Dynamic causal modelling is a modelling approach designed to estimate activity and effective connectivity in a network of interconnected populations of neurons. DCM regards the brain as a nonlinear dynamical system of interconnected nodes and an experiment as a designed perturbation of the system's dynamics (Friston *et al.*, 2003). Its goal is to provide a mechanistic explanation of observed measures of brain activity. While the mathematical formulation of DCM varies across measurement types, common mechanisms modelled by all DCMs[3] include synaptic connection strengths and their experimentally induced modulation (Stephan *et al.*, 2008; David *et al.*, 2006; Chen *et al.*, 2008; Moran *et al.*, 2009; Daunizeau *et al.*, 2009). Such experimental manipulations enter the model in two different ways: they can elicit responses through direct influences on specific regions (e.g., sensory inputs), or they can modulate the strength of coupling among regions (e.g., task demands or learning).

Regardless of data modality, dynamic causal models are generally hierarchical, comprising two model layers (Stephan *et al.*, 2007b). The first layer is a model of the dynamics among interacting neuronal populations in the context of experimental perturbations. The second layer is a modality-specific forward model that translates source activity into measurable observations. It is the neuronal model that is typically of primary interest (Figure 2.2).

DCM strives for neurobiological interpretability of its parameters; all model constituents mimic neurobiological mechanisms and hence have an explicit neuronal interpretation. In particular, the neural-mass model embodied by DCM is largely based on the mechanistic model of cortical columns originally proposed by Jansen and Rit (1995) and further refined in follow-up studies (David and Friston, 2003; David *et al.*, 2006; Moran *et al.*, 2009). DCM parameters represent, for example, synaptic weights and their context-specific modulation. In the case of electrophysiological data, the model describes even more fine-grained processes such as spike-frequency adaptation or conduction delays. In this regard, DCM fundamentally departs from previous approaches, such as multivariate autoregressive models, that either characterized experimental effects in a purely phenomenological fashion or were only loosely coupled with biophysical mechanisms.

For a given set of recorded data, estimating the parameters of a dynamic causal model means inferring what neural causes have most likely given rise

[3]Following standard practice, we use the term DCM to refer both to a specific dynamic causal model and to dynamic causal modelling as a method.

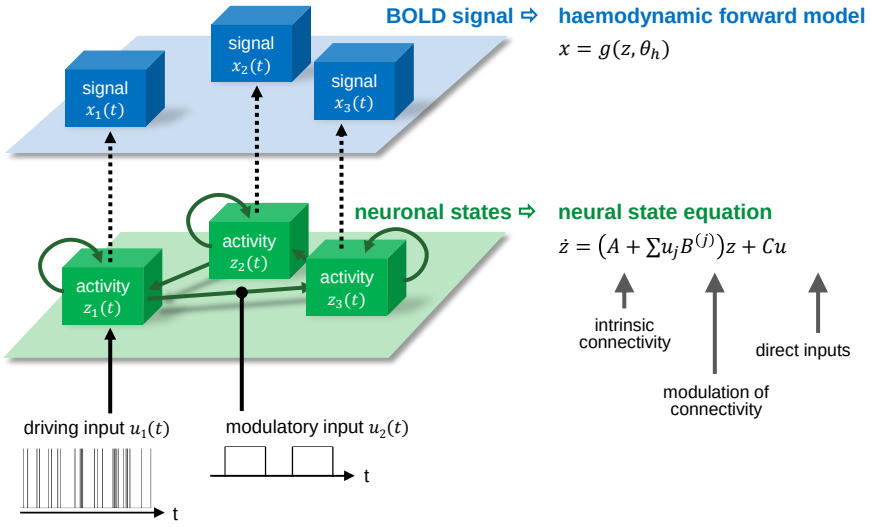

Figure 2.2: Dynamic causal modelling. DCM regards the brain as a nonlinear dynamical system of interconnected nodes, and an experiment as a designed perturbation of the system's dynamics. The model consists of a neuronal model as well as a forward model that describes how activity at the neuronal level translates into observed signals. DCM strives for a mechanistic explanation of experimental measures of brain activity. This makes it an attractive model for generative embedding. (Figure adapted from slides by K.E. Stephan.)

to the observed responses, conditional on the model. Such models can be applied to a single population of neurons, e.g., a cortical column, to make inferences about neurophysiological processes such as amplitudes of postsynaptic responses or spike-frequency adaptation (Moran *et al.*, 2008). More frequently, however, models are used to investigate the effective connectivity among remote regions and how it changes with experimental context (e.g., Garrido *et al.*, 2008; Stephan *et al.*, 2008). In this thesis, we will use DCM in both ways, applying it to different datasets, ranging from single-site recording from the somatosensory barrel cortex two a two-electrode recording from the auditory cortex.

DCM uses a fully Bayesian approach to parameter estimation, with empirical priors for the haemodynamic parameters and conservative shrinkage priors for the coupling parameters (Friston, 2002; Friston *et al.*, 2003) all of which can be updated in light of new experimental evidence (cf. Stephan

et al., 2007a). Given a model m, combining the prior density over the parameters $p(\theta \mid m)$ with the likelihood function $p(x \mid \theta, m)$ yields the posterior density $p(\theta \mid x, m)$. This inversion can be carried out efficiently by maximizing a variational free-energy bound to the log model evidence, $\ln p(x \mid m)$, under Gaussian assumptions about the posterior (the Laplace assumption; see Friston *et al.*, 2007, for details).[4] Model inversion can be viewed as a mapping $\mathcal{X} \to \mathcal{M}_\Theta$ that projects a given example $x \in \mathcal{X}$ (i.e., data from a single subject) onto a multivariate probability distribution $p(\theta \mid x)$ in a parametric family $\mathcal{M}_\Theta$. Importantly, model inversion proceeds in an unsupervised fashion, i.e., in ignorance of external labels y that might later be used in the context of classification or clustering.[5]

While model selection is an important theme in DCM (Stephan *et al.*, 2010), in this thesis we are not primarily concerned with the question of which of several alternative DCMs may be optimal for explaining the data or for classifying subjects; these issues can be addressed using Bayesian evidence methods (Stephan *et al.*, 2009a; Penny *et al.*, 2004) or by applying cross-validation to the classifications suggested by each of the models, respectively. However, an important issue is that model specification cannot be treated in isolation from its subsequent use for classification or clustering. Specifically, some procedures for selecting time series can lead to biased estimation of classification accuracy or purity. We will therefore provide a detailed assessment of different strategies for time-series selection in DCM-based generative embedding and highlight those procedures which safeguard against obtaining optimistic estimates of classification performance.

2.3 An embedding for electrophysiology

This thesis proposes two concrete approaches to generative embedding for neuroscientific datasets. The first concerns direct invasive electrophysiological recordings in animals and will be developed in this section. Details can be found in other publications (David and Friston, 2003; Kiebel *et al.*, 2009; Moran *et al.*, 2008, 2009). However, in order to keep the present thesis self-contained, a brief summary of the main modelling principles is presented in the following subsections.

[4] A similar strategy will be used for a variational Bayesian approach to inference on classification performance; see Chapter 4.

[5] The literature on DCM has adopted the convention of denoting the hidden states by x and the data by y. Here, in order to keep the notation consistent with the literature on classification, we use z for the hidden states, x for the data, and y for external labels.

Neural-mass model

The neural-mass model in DCM describes a set of n neuronal populations as a system of interacting elements, and it models their dynamics in the context of experimental perturbations. At each time point t, the state of the system is expressed by a vector $z(t) \in \mathbb{R}^n$. The evolution of the system over time is described by a set of ordinary differential equations that evolve the state vector and use a Taylor approximation to account for small conduction delays among spatially separate populations. The equations specify the rate of change of activity in each region [i.e., of each element in $z(t)$] as a function of three variables: the current state $z(t)$ itself; the strength of experimental inputs $u(t)$ (e.g., sensory stimulation); and a set of time-invariant neuronal parameters θ_n. Thus, in general terms, the dynamics of the model are given by an n-valued function

$$\frac{\mathrm{d}z(t)}{\mathrm{d}t} = f\left(z(t), u(t), \theta_n\right). \tag{2.3.1}$$

Within the framework of DCM, each of the n regions is modelled as a microcircuit whose properties are derived from the biophysical model of cortical columns proposed by Jansen and Rit (1995). Specifically, each region is assumed to comprise three subpopulations of neurons whose voltages and currents constitute the state vector $z^{(k)} \in \mathbb{R}^9$ of a region k (for a description of the individual components see next paragraph). These populations comprise pyramidal cells, excitatory interneurons, and inhibitory interneurons. The connectivity within a column or region is modelled by intrinsic excitatory and inhibitory connections. Connections between remote neuronal populations are excitatory and target specific neuronal populations, depending on their relative hierarchical position, resulting in lateral, forward, and backward connections (Felleman and van Essen, 1991). Experimentally controlled sensory inputs affect the granular layer and are modelled as a mixture of one fast, event-related and various slow, temporally dispersed components of activity.

Region-specific constants and parameters comprise (i) time constants G of the intrinsic connections, (ii) time constants and maximum amplitudes of excitatory/inhibitory postsynaptic responses (T_e/T_i, H_e/H_i), and (iii) input parameters which specify the delay and dispersion of inputs arriving in the granular layer. Two additional sets of parameters control connections between regions: (iv) extrinsic connection parameters, which specify the specific coupling strengths between any two regions; and (v) conduction delays, which characterize the temporal properties of these connections,

and whose effect can be approximated without the need to reformulate the system in terms of delay differential equations.

Forward model for LFPs

The forward model within DCM describes how (latent) neuronal activity in individual regions generates (observed) measurements. Compared to the relatively complex forward models used for fMRI or EEG, the forward model for LFPs is simpler, requiring only a single (gain) parameter for approximating the spatial propagation of electrical fields in cortex (Moran *et al.*, 2009).

In most applications of dynamic causal modelling, one or several candidate models are fitted to all data from each experimental condition (e.g., by concatenating the averages of all trials from all conditions and providing modulatory inputs that allow for changes in connection strength across conditions). In the context of generative embedding for LFP data, by contrast, we are fitting the model in a true trial-by-trial fashion. It is therefore critical that the model is not aware of the category a given trial was taken from. Instead, its inherent biophysical parameters should be able to reflect class differences by themselves.

2.4 An embedding for fMRI

In addition to an embedding for the trial-by-trial classification of electrophysiological recordings, we propose a DCM-based embedding for subject-by-subject classification of fMRI data.

Neuronal model for fMRI

In the classical bilinear DCM formulation (Friston *et al.*, 2003) as implemented in the software package SPM8/DCM10, the neuronal model is given by

$$\frac{\mathrm{d}z(t)}{\mathrm{d}t} = f\left(z(t), \theta_n, u(t)\right) \tag{2.4.1}$$

$$= \left(A + \sum_{j=1}^{J} u_j(t) B^{(j)}\right) z(t) + Cu(t) \tag{2.4.2}$$

where $z(t)$ represents the latent neuronal state at time t, A is a matrix of endogenous connection strengths, $B^{(j)}$ represents the additive change of these connection strengths induced by modulatory input u_j, and C denotes the strengths of direct (driving) inputs $j = 1 \dots J$. These neuronal parameters $\theta_n = (A, B^{(1)}, \dots, B^{(J)}, C)$ are rate constants with units s^{-1}.

Forward model for fMRI

The second layer of DCM for fMRI is a biophysically motivated forward model that describes how a given neuronal state translates into a measurement:

$$p\left(x(t) \mid z, \theta_h, \sigma\right) = \mathcal{N}\left(x(t) \,\middle|\, g\left(z, t, \theta_h\right), \sigma^2\right) \tag{2.4.3}$$

The forward model rests upon a nonlinear operator $g(\cdot)$ that links a time series of latent neuronal states $z(t)$ to a predicted blood oxygen level dependent (BOLD) signal $x(t)$ via changes in vasodilation, blood flow, blood volume, and deoxyhaemoglobin content (see Stephan *et al.*, 2007a, for details). The model has haemodynamic parameters θ_h and a Gaussian measurement error with variance σ^2.

The haemodynamic parameters primarily serve to account for variations in neurovascular coupling across regions and subjects and are typically not of primary scientific interest. In addition, the haemodynamic parameters exhibit strong inter-dependencies and thus high posterior (co)variances (Stephan *et al.*, 2007a), which makes it difficult to establish the distinct contribution afforded by each parameter. For these reasons, the model-induced feature spaces in this thesis will be based exclusively on the neuronal parameters θ_n, now simply referred to as θ.

2.5 Constructing the kernel

Given a collection of inverted generative models, the kernel defines the similarity metric under which independently inverted models are to be compared to one another for the purposes of classification or clustering.

In generative embedding, the choice of an appropriate kernel depends on the definition of the generative score space. A straightforward way to create a Euclidean vector space from an inverted DCM is to consider the posterior means or maximum a posteriori (MAP) estimates of model parameters of interest (e.g., parameters encoding synaptic connection strengths).

More formally, we could define a mapping $\mathcal{M}_\Theta \to \mathbb{R}^d$ that extracts a subset of posterior means $\hat{\mu}$ from the posterior distribution $p(\theta \mid x, m)$. This simple d-dimensional vector space expresses discriminative information encoded in the connection strengths *between* regions, as opposed to activity levels *within* these regions. Alternatively, we could incorporate elements of the posterior covariance matrix into the vector space. This step would be beneficial if class differences were revealed by the precision with which connection strengths can be estimated from the data.

In principle, once a generative score space has been created, any conventional vectorial kernel

$$k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R} \tag{2.5.1}$$

can be used to compare two posteriors over model parameters. The simplest one is the linear kernel,

$$k(x_i, x_j) := x_i^{\mathrm{T}} x_j, \tag{2.5.2}$$

representing the inner product between two vectors x_i and x_j without any prior feature transformation.

Nonlinear kernels, such as quadratic, polynomial or radial basis function kernels, transform the generative score space, which makes it possible to consider quadratic (or higher-order) class boundaries and therefore account for possible interactions between features. Nonlinear kernels, however, have several disadvantages for generative embedding. As the complexity of the kernel increases, so does the risk of overfitting. Furthermore, feature weights are easiest to interpret in relation to the underlying model when they do not undergo further transformation. In the case of model-based classification, in particular, we will see that the contribution of a particular feature (i.e., model parameter) to the success of the classifier can be understood as the degree to which the neuronal mechanism represented by that parameter aids classification. A simple linear kernel will therefore be our preferred choice.

In summary, we define a mapping $\mathcal{M}_\Theta \to \mathbb{R}^d$ from a subject-specific posterior distribution of model parameters $p(\theta \mid x, m)$ to a feature vector $\hat{\mu}$. We then use a linear kernel $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ for this model-based feature space. Together, these two steps define a probability kernel $k_{\mathcal{M}} : \mathcal{M}_\Theta \times \mathcal{M}_\Theta \to \mathbb{R}$ that represents a similarity metric between two inverted models and allows for mechanistic interpretations of how group membership of different trials or subjects is encoded by spatiotemporal LFP or fMRI data.

Chapter 3

Fixed-effects inference on classification performance

Assessing the utility of model-based classification critically necessitates a measure of classification performance. There are two reasons for this requirement. Firstly, the practical clinical value of model-based classification depends on the degree to which unseen examples (e.g., subjects) can be identified with their correct class labels (e.g., disease states; Figure 3.1). Secondly, model-based classification can be used to compare competing models in terms of their discriminative capacity in relation to an external class label.

Whenever there is no specific need to impose different costs on different types of misclassification, the overall *accuracy* is of primary interest. In these cases, a commonly adopted approach for summarizing cross-validation results is to report the average sample accuracy (or average sample error) across all folds. However, measuring performance in this way has two shortcomings.

The first is that the approach does not entail meaningful confidence intervals of a true underlying quantity. In particular, computing the standard error of the mean across all folds enforces symmetric limits and may lead to confidence intervals of accuracy including values above 100%. The second weakness in considering the average accuracy is that it may give a misleading idea about generalization performance in situations where a biased classifier is tested on an imbalanced dataset. Under these conditions, a naïve evaluation of the average accuracy (i.e., one that does not take into

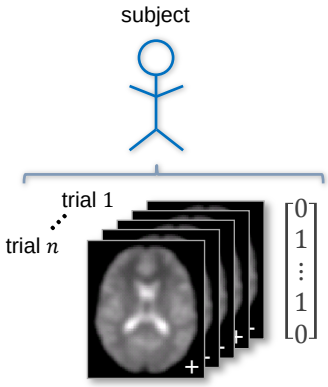

Figure 3.1: Trial-by-trial classification. When classifying individual trials in a given subject, classification outcomes can be represented as 1's and 0's, corresponding to correctly and incorrectly classified trials, respectively. Statistical inference concerns two important questions: what is the generalization performance of the classifier; and how does its performance compare to a competing classifier.

account the degree of class imbalance) may lead to false conclusions about the significance with which an algorithm has performed better than chance.

In this chapter, we demonstrate how both shortcomings can be overcome by replacing the average sample accuracy by the *posterior balanced accuracy* for which we derive a conditional density using a fully Bayesian approach. We begin by considering a simple Bayesian approach to evaluating classification accuracy (Section 3.1). We then extend this approach to inference on the balanced accuracy, which is a more natural measure of performance when the data are not fully balanced (Section 3.2). For corresponding publications, see Brodersen *et al.* (2010a,b) and Brodersen *et al.* (*in press*).

3.1 Inference on the accuracy

In a binary classification analysis, let n be the number of examples underlying a leave-m-out cross-validation scheme with K folds. This setting gives rise to two types of hypothesis that we wish to test. First, is a classification algorithm operating at the level of guessing, or is its generalization accuracy significantly above chance? Second, more generally, does a classification al-

gorithm significantly outperform an alternative algorithm? Both questions require statistical inference on a measure of generalizability.

3.1.1 Classical inference for a single subject

A common way of computing an estimate of generalizability begins by summing the number of correctly labelled test cases, k, across all cross-validation folds,

$$k = \sum_{i=1}^{K} r_i, \tag{3.1.1}$$

where $r_i \in \{0, \ldots, m\}$. The *sample accuracy*[1] can then be reported as the fraction $\frac{k}{n}$.

One way of estimating the significance of the sample accuracy is by considering the standard error of the mean, $\hat{\sigma}/\sqrt{K-1}$, where $\hat{\sigma}$ is the empirical standard deviation of $\frac{r_i}{m}$, observed across all cross-validation folds $i = 1 \ldots K$. This quantity, however, is dependent on arbitrary design choices such as m, the number of test cases in each cross-validation fold, and, worse still, may easily lead to error bars including values above 100%.

A very different route can be taken by invoking an explicit statistical model of classification outcomes. One well-known possibility is to regard each test case as an independent Bernoulli experiment and compare the obtained sample accuracy to the level that must be reached by an above-chance learning algorithm. In classical inference one could obtain a maximum-likelihood (ML) estimate for the true accuracy π of the classifier. This estimate is

$$\hat{\pi}_{\mathrm{ML}} = \arg\max_{\pi} \mathrm{Bin}(k \mid \pi, n) = \frac{k}{n}, \tag{3.1.2}$$

which corresponds exactly to the sample accuracy itself and hence provides an alternative interpretation for it.

Point estimates by themselves are of course not sufficient to assess statistical significance. In order to determine, for instance, whether a given classification outcome is the result of an algorithm that operates significantly

[1] Since *classification error* $= 1 -$ *classification accuracy*, the sample error could be reported instead; this correspondence pertains to all other accuracy-related quantities discussed throughout this thesis.

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

$$p = 1 - \mathcal{F}_{\text{Bin}}(k \mid 0.5), \tag{3.1.3}$$

where $\mathcal{F}_{\text{Bin}}(k \mid 0.5)$ is the cumulative distribution function of the binomial distribution with $\pi = 0.5$. While this approach does offer a measure of classical statistical significance, it does not associate the accuracy with a level of precision. As a result, it does not yet provide, for example, an immediate way of comparing two algorithms both of which have been found to operate above chance. More generally, maximum-likelihood estimation risks overfitting, and it does not explicitly account for prior or posterior uncertainty about classification performance.

In summary, the practical simplicity of maximum likelihood is offset by its conceptual limitations. These limitations can be resolved by turning to a fully Bayesian approach, as described next.

3.1.2 The beta-binomial model

Rather than averaging the outcomes obtained on different cross-validation folds or computing an ML estimate of classification accuracy, we now turn to a more flexible Bayesian treatment.

Model

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

$$p(y_i \mid \pi) = \text{Bern}(y_i \mid \pi) \tag{3.1.4}$$

$$= \pi^{y_i}(1 - \pi)^{1-y_i} \quad \forall i = 1 \ldots n. \tag{3.1.5}$$

The i.i.d. assumption derives from the assumption that the observations in the test set are i.i.d. themselves.[2] It allows us to summarize a sequence of outcomes in terms of the number of correctly predicted trials, $k = \sum_{i=1}^{n} y_i$, and the total number of test trials, n. Since the sum of several Bernoulli variables follows a binomial distribution, the number of successes is distributed as:

$$p(k \mid \pi, n) = \mathrm{Bin}(k \mid \pi, n) \tag{3.1.6}$$

$$= \binom{n}{k} \pi^k (1-\pi)^{n-k} \tag{3.1.7}$$

In this setting, Bayesian inference differs from classical maximum-likelihood estimation in that it assesses the plausibility of all possible values of π before and after observing actual data, rather than viewing π as a fixed parameter that is to be estimated.[3] Incidentally, it is precisely this problem that formed the basis of the first Bayesian analyses published by Bayes and Price (1763) and Laplace (1774).

A natural choice for the prior distribution $p(\pi)$ is the Beta distribution,

$$p(\pi \mid \alpha_0, \beta_0) = \mathrm{Beta}(\pi \mid \alpha_0, \beta_0) \tag{3.1.8}$$

$$= \frac{\Gamma(\alpha_0 + \beta_0)}{\Gamma(\alpha_0)\,\Gamma(\beta_0)} \pi^{\alpha_0 - 1} (1-\pi)^{\beta_0 - 1}, \tag{3.1.9}$$

where $\alpha_0, \beta_0 > 0$ are hyperparameters, and the Gamma function $\Gamma(\cdot)$ is required for normalization. Multiplying (3.1.7) with (3.1.9) (and integrating out π) gives rise to an overdispersed form of the binomial distribution known as the beta-binomial distribution (Figure 3.2; Pearson, 1925; Skellam, 1948; Lee and Sabavala, 1987).

In the absence of prior knowledge about π, we use a flat prior by setting $\alpha_0 = \beta_0 = 1$, which turns the Beta distribution into a uniform distribution over the $[0, 1]$ interval. The hyperparameters α_0 and β_0 can be interpreted as virtual prior counts of $\alpha_0 - 1$ correct and $\beta_0 - 1$ incorrect trials. Thus, a uniform prior is sometimes interpreted as 'zero virtual prior observations' of either kind, although this interpretation has its limits as there is no absolute scale on the number of virtual points.[4]

[2] This assumption is not always made in the context of cross-validation, but is easily justified when the data are only split once, without any cross-validation (see discussion in Section 4.8 on p. 119).

[3] Note that n depends on the experimental design and is not subject to inference.

[4] For a discussion of alternative priors, see Gustafsson *et al.* (2010).

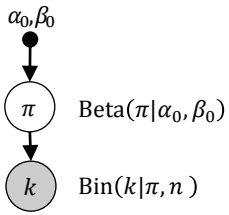

Figure 3.2: Beta-binomial model for Bayesian fixed-effects inference on classification accuracy. Because individual classification outcomes can be viewed as Bernoulli outcomes, the sum of correctly classified examples follows a Binomial distribution. Our uncertainty about its parameter π is described by a Beta density. Blank circles correspond to latent variables, filled circles represent observed data.

Inference

Rather than finding a point estimate $\hat{\pi}_{\mathrm{ML}}$ of true classification accuracy, the above model allows us to compute a full posterior distribution by multiplying the prior with the likelihood,

$$p(\pi \mid k, \alpha_0, \beta_0) = \frac{p(k \mid \pi) p(\pi \mid \alpha_0, \beta_0)}{p(k \mid \alpha_0, \beta_0)}, \tag{3.1.10}$$

where integration over π provides the normalization constant

$$p(k \mid \alpha_0, \beta_0) = \int_0^1 p(k \mid \pi)\, p(\pi \mid \alpha_0, \beta_0)\, \mathrm{d}\pi \tag{3.1.11}$$

$$= \mathrm{Bb}(k \mid \alpha, \beta). \tag{3.1.12}$$

The beta-binomial distribution in (3.1.12) relates an observation k directly to the hyperparameters α_0 and β_0 by marginalizing over the intermediate variable π.

Thanks to conjugacy, the posterior over π has the same functional form as the prior,

$$p(\pi \mid k, \alpha_0, \beta_0) = \mathrm{Beta}(\pi \mid \alpha_0 + k, \beta_0 + n - k), \tag{3.1.13}$$

which can be rewritten as

$$p(\pi \mid \alpha_n, \beta_n) = \mathrm{Beta}(\pi \mid \alpha_n, \beta_n). \tag{3.1.14}$$

This shows that the above inference can be interpreted as an update step in which the virtual prior counts α_0 and β_0 have been incremented by the number of observed successes and failures, respectively, leading to new counts $\alpha_n = \alpha_0 + k$ and $\beta_n = \beta_0 + n - k$. As the current posterior turns into the next prior, a sequential update scheme evolves that is just one of the beneficial consequences of conjugacy in Bayesian inference.

One of the principal advantages of the Bayesian approach is the flexibility with which posterior inferences can be summarized. For instance, given the posterior distribution $p(\pi \mid k, \alpha_0, \beta_0)$ with hyperparameters $\alpha_0 = \beta_0 = 1$, one could report a point estimate of the classification accuracy by giving the posterior expectation

$$\mathbb{E}[\pi \mid k, \alpha_0, \beta_0] = \frac{k+1}{n+2}. \tag{3.1.15}$$

Alternatively, one could report the mode of the distribution, that is, the *maximum a posteriori* (MAP) estimate

$$\arg\max_k \, p(\pi \mid k, \alpha_0, \beta_0) = \frac{k}{n}. \tag{3.1.16}$$

The above shows that the sample accuracy k/n can now be reinterpreted as the MAP estimate of the accuracy under a noninformative prior.[5] These two examples show how classical and Bayesian inference may lead to similar conclusions at the surface; the Bayesian approach, however, extends more easily to more complex problems.

Rather than picking a point from the posterior, it is often more informative to summarize the distribution in terms of a posterior interval or *credible interval*. For example, one could give a central 95% posterior interval of the classification accuracy as

$$\left[B^{-1}_{0.025}(\alpha_0 + k, \beta_0 + n - k); \quad B^{-1}_{0.975}(\alpha_0 + k, \beta_0 + n - k)\right], \tag{3.1.17}$$

where $B^{-1}_p(\cdot)$ denotes the inverse cumulative density function of the Beta distribution at point p.

The most important question for classification analyses in neuroimaging is whether the true classification accuracy is greater than chance. A Bayesian analogue of a classical p-value can be obtained as

$$\Pr(\pi < 0.5 \mid k, \alpha_0, \beta_0) = B(\alpha_0 + k, \beta_0 + n - k), \tag{3.1.18}$$

[5] Alternatively, the sample accuracy could be interpreted as the posterior expectation under an improper prior with $\alpha_0 = \beta_0 = 0$.

where $B(\cdot)$ is the cumulative density function of the Beta distribution. We refer to this probability as the (posterior) *infraliminal probability* of the classifier. It represents the subjective posterior probability of the classification accuracy being smaller than the performance expected under chance.[6]

In addition to inference on latent variables such as the classification accuracy π, a Bayesian approach also enables inference on quantities that are *observable* but not yet *observed*. Of particular interest is the posterior predictive distribution over $\tilde{k}$ the number of correct predictions in a future sequence of $\tilde{n}$ trials, which can now be computed without resorting to a plug-in estimate for π. While the prior predictive density is conditional only on the hyperparameters α_0 and β_0,

$$p(\tilde{k} \mid \alpha_0, \beta_0) \tag{3.1.19}$$

$$= \mathrm{Bb}(\tilde{k} \mid \alpha_0, \beta_0) \tag{3.1.20}$$

$$= \frac{\Gamma(\tilde{n}+1)}{\Gamma(\tilde{k}+1)\,\Gamma(\tilde{n}-\tilde{k}+1)} \frac{\Gamma(\alpha_0+k)\,\Gamma(\beta_0+\tilde{n}-\tilde{k})}{\Gamma(\alpha_0+\beta_0+\tilde{n})} \frac{\Gamma(\alpha_0+\beta_0)}{\Gamma(\alpha_0)\Gamma(\beta_0)}, \tag{3.1.21}$$

the posterior predictive density is also conditional on the data k,

$$p\left(\tilde{k} \mid k, \alpha_0, \beta_0\right) \tag{3.1.22}$$

$$= \int p\left(\tilde{k} \mid \pi\right) p(\pi \mid k, \alpha_0, \beta_0) \mathrm{d}\pi \tag{3.1.23}$$

$$= \binom{\tilde{n}}{\tilde{k}} \frac{\Gamma(\alpha_n+\beta_n)\,\Gamma(\alpha_n+\tilde{k})\,\Gamma(\beta_n+\tilde{n}-\tilde{k})}{\Gamma(\alpha_n)\,\Gamma(\beta_n)\,\Gamma(\alpha_n+\beta_n+\tilde{n})}, \tag{3.1.24}$$

where we have used the same definitions for α_n and β_n as in (3.1.14). One common way of making use of the posterior predictive distribution is for the purpose of model validation. For instance, one could compute the probability that the data $\tilde{k}$ obtained from a replication of the experiment with $\tilde{n} = n$, could be more extreme than the observed data. More importantly still, the posterior predictive distribution is the main distribution of interest when comparing two classifiers to one another in a hierarchical setting (see Chapter 4).

The approach described here is conceptually well-established, although it leaves an important question unanswered: the question of which performance measure is most suitable when asking whether a (model-based) classification algorithm has picked up a statistical relationship between data

[6] An alternative to using the infraliminal probability is Bayesian model comparison. See p. 121 in Section 4.8 for a discussion.

features and class labels. In the next section, we will therefore show how an important practical weakness of the above approach can be overcome by considering a different performance measure.

3.2 Inference on the balanced accuracy

A well-known phenomenon in binary classification is that a training set consisting of different numbers of representatives from either class may result in a classifier that is biased towards the majority class. When applied to a test set that is similarly imbalanced, this classifier yields an optimistic accuracy estimate. In an extreme case, the classifier might assign every single test case to the majority class (which is the optimal strategy given no information about the classes other than their frequencies in the training set), thereby achieving an accuracy equal to the proportion of test cases belonging to the majority class.

Previous studies have examined different ways of addressing this problem (see Akbani *et al.*, 2004; Chawla *et al.*, 2002; Japkowicz and Stephen, 2002). One strategy, for example, is to restore balance on the training set by undersampling the large class or by oversampling the small class. Another strategy is to modify the costs of misclassification in such a way that no bias is acquired. However, while these methods may under some circumstances prevent a classifier from becoming biased, they do not provide generic safeguards against reporting an optimistic accuracy estimate.

Another solution 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 performance would have to include an explicit baseline level, which makes the comparison of accuracies across studies, datasets, or classifiers involved and tedious.

The above considerations motivate the use of a different performance measure: the *balanced accuracy*, defined as the arithmetic mean of sensitivity and specificity, or the average accuracy obtained on either class,

$$\phi := \frac{1}{2}\left(\pi^+ + \pi^-\right). \tag{3.2.1}$$

where π^+ and π^- denote classification accuracies on positive and negative

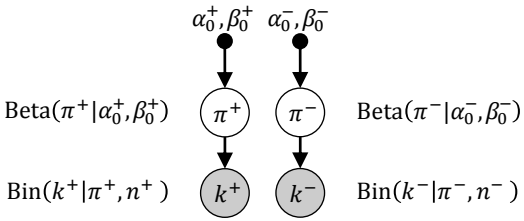

Figure 3.3: Twofold beta-binomial model for Bayesian fixed-effects inference on the balanced classification accuracy. Blank circles correspond to latent variables, filled circles represent observed data.

trials, respectively.[7] If the classifier performs equally well on either class, this term reduces to the conventional accuracy (i.e., the number of correct predictions divided by the total number of predictions). In contrast, if the conventional accuracy is above chance *only* because the classifier takes advantage of an imbalanced test set, then the balanced accuracy, as appropriate, will drop to chance.[8] We can evaluate the balanced accuracy in a hierarchical setting by extending the beta-binomial model, as described next.

Model

Treating the balanced accuracy as a random variable allows us to reason about its prior and posterior distribution. Depending on whether the true label of a given test case is positive or negative, we regard a prediction as a draw either (i) from a bucket of 'true positive' or 'false negative' balls, or (ii) from a bucket of 'true negative' or 'false positive' balls. Put differently, under the same distributional assumptions as made above, we are effectively applying the beta-binomial model separately to the results on truly positive and truly negative examples. We are hence interested in the probability density of $\phi = \frac{1}{2}(\pi^+ + \pi^-)$, where π^+ and π^- are random variables specifying the accuracy on positive and negative examples, respectively.

[7] Unlike the measure described in Velez *et al.* (2007), the balanced accuracy is symmetric about the type of class. If desired, this symmetry assumption can be dropped, yielding $c \times \frac{k^+}{n^+} + (1-c) \times \frac{k^-}{n^-}$, where $c \in [0, 1]$ is the cost associated with the misclassification of a positive example.

[8] Additional details on the conceptual difference between accuracies and balanced accuracies will be discussed in Section 4.8.

Inference

A closed form for the distribution of ϕ is not available, and so we resort to a numerical approximation. For this, we first note that the distribution of the sum of the two class-specific accuracies, $s := \pi^+ + \pi^-$, is the convolution of the distributions for π^+ and π^-,

$$p\left(s \mid \alpha_n^+, \beta_n^+, \alpha_n^-, \beta_n^-\right) \tag{3.2.2}$$

$$= \int_0^s p_{\pi^+}(s - z \mid \alpha_n^+, \beta_n^+)\, p_{\pi^-}(z \mid \alpha_n^-, \beta_n^-)\, \mathrm{d}z, \tag{3.2.3}$$

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

$$p(\phi \mid \alpha_n^+, \beta_n^+, \alpha_n^-, \beta_n^-) \tag{3.2.4}$$

$$= \int_0^{2\phi} p_{\pi^+}(2\phi - z \mid \alpha_n^+, \beta_n^+)\, p_{\pi^-}(z \mid \alpha_n^-, \beta_n^-)\, \mathrm{d}z \tag{3.2.5}$$

$$= \int_0^{2\phi} \mathrm{Beta}(2\phi - z \mid \alpha_n^+, \beta_n^+)\, \mathrm{Beta}(z \mid \alpha_n^-, \beta_n^-)\, \mathrm{d}z. \tag{3.2.6}$$

Thus, assuming a flat prior for the true balanced accuracy, we can report cross-validation results by describing the posterior distribution of the balanced accuracy. Note that the mean (mode) of the distribution of the balanced accuracy does not necessarily equal the mean of the means (modes) of the separate accuracy distributions for positive and negative examples. There are no analytical forms for the mean, the mode, or a posterior probability interval. However, we can compute numerical approximations.

We will postpone further details on the sort of inferences one might typically want to report to Chapter 4, in which we will discuss a more general (hierarchical) approach to performance evaluation. We therefore restrict ourselves here to a small set of simulations to illustrate the key features of our approach in contrast to classical inference.

3.2.1 Applications

Before turning to empirical data (Section 4.7.4), we will begin by demonstrating the key properties of a Bayesian approach to inference on balanced

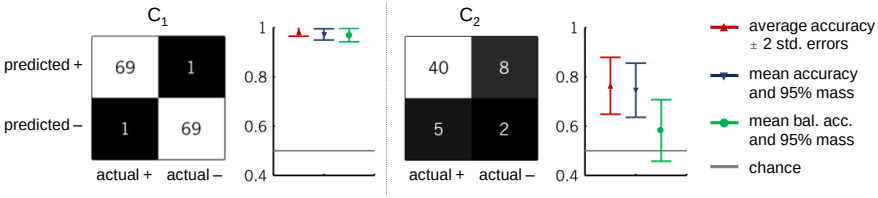

Figure 3.4: Comparison of accuracy measures. Based on two illustrative confusion matrices C_1 and C_2, the first example shows how the conventional average accuracy (red) may imply a confidence interval that includes values above 100%. The second example shows how accuracies, unlike balanced accuracies (green), falsely suggest above-chance generalizability in the case of a biased classifier that has taken advantage of an imbalanced test set.

accuracies using a small set of hand-crafted examples. As a result of training and testing two independent classifiers on different datasets, let C_1 and C_2 be the confusion matrices of the respective results, summed across all cross-validation folds. We wish to compare the average accuracy (along with standard errors) to the posterior accuracy and the posterior balanced accuracy (Figure 3.4).

In the first example, the test set is perfectly balanced (70 positive vs. 70 negative examples). As a result, the differences between the three numbers are not substantial. However, the simulation does illustrate how an interval of 2 standard errors around the average accuracy (i.e., the common approximation to a 95% confidence interval) includes values above 100% (Figure 3.4, left box, red interval). In contrast, the probability intervals of the posterior accuracy and balanced accuracy show the desired asymmetry (blue and green intervals).

In the second example, both the average accuracy and the mean of the posterior accuracy seem to indicate strong classification performance (Figure 3.4, right box, red and blue intervals). The balanced accuracy, by contrast (green interval), reveals that in this simulation the test set was imbalanced (45 positive vs. 10 negative examples) and, in addition, the classifier had acquired a bias towards the large class (48 positive vs. 7 negative predictions). Accuracy measures on their own fail to detect this situation and give the false impression of above-chance generalizability.

The difference between accuracies and balanced accuracies is further illustrated in Figure 3.5. Based on the confusion matrix C_2, the two plots show all of the statistics mentioned in Sections 3.1 and 3.2 superimposed on the central 95% probability interval of the respective posterior distributions.

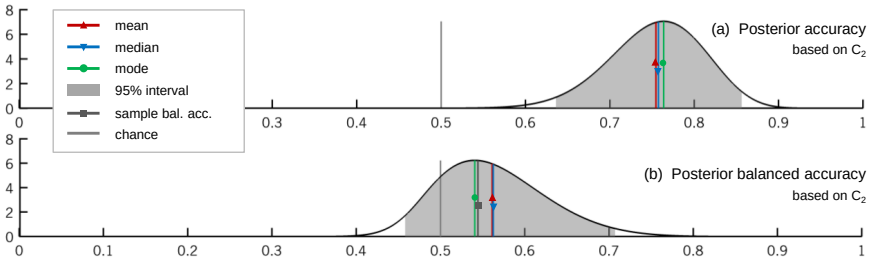

Figure 3.5: Comparison between the posterior distribution of the accuracy and the balanced accuracy. Posterior densities are based on the confusion matrix C_2 depicted in Figure 3.4.

The figure also contains the *sample* balanced accuracy, computed as the mean of the modes of the accuracies on positive and negative examples. The simulation shows how a biased classifier applied to an imbalanced test set leads to a hugely optimistic estimate of generalizability when measured in terms of the accuracy rather than the balanced accuracy.

3.3 Discussion

In binary classification, confusion matrices form the basis of a multitude of informative measures of generalizability. Yet, it is still common to average accuracies across cross-validation folds. This approach neither supports meaningful confidence intervals; nor does it provide safeguards against a biased classifier that has taken advantage of an imbalanced test set. The first limitation can be overcome by the well-known approach of considering the full posterior distribution of the accuracy instead of a point estimate (Bishop, 2007, pp. 68–74); and the second one by the idea of replacing conventional accuracies by balanced accuracies.

Throughout this chapter, we have made no distinction between individual classifiers on the one hand and classification algorithms on the other, since the idea of considering the posterior distribution of the balanced accuracy applies to either. The only way in which the two cases differ is whether we look at the confusion matrix that results from a single train/test cycle (yielding the posterior of the balanced accuracy of an individual classifier) or whether we sum the confusion matrices across all cross-validation folds (leading to the posterior of the algorithm as a whole). In most practical ap-

plications, it is the generalizability of the algorithm that will be of primary interest. The approach can therefore be used for any number of underlying cross-validation folds: it solely requires the overall confusion matrix, as obtained by summing individual confusion matrices across all folds.

An interesting generalization is the notion of balancing not only class labels themselves but also other variables that correlate with class labels. This reweighting is important, for instance, in the case of a test set with balanced class labels in which another binary variable, closely correlated with class labels, is imbalanced. A biased classifier could then falsely suggest high generalizability while, in fact, it has learnt to separate examples according to the additional variable rather than according to the original class labels. It could be instrumental to investigate (i) how resampling and cost-modification techniques could be used to efficiently deal with several criteria, and (ii) whether (multiply) balanced accuracies might again prove useful in reporting generalizability in a way that is safeguarded against optimistic accuracy estimates.

One important limitation remains: so far, we have considered data from a single subject only, disregarding the variability that there might exist in the population. In the next chapter, we will therefore examine how the Bayesian approach described here can be extended to a mixed-effects analysis that accounts for both within-subjects and between-subjects variability in classification performance.

Chapter 4

Mixed-effects inference on classification performance

Classification algorithms are frequently used on data with a natural *hierarchical* structure. For instance, classifiers are often trained and tested on trial-wise measurements, separately for each subject within a group. The classification analyses considered in the previous chapter, by contrast, were 'flat,' in the sense that they did not reflect a hierarchical structure.

An important question in hierarchical analyses is how classification outcomes observed in individual subjects can be generalized to the population from which the group was sampled. To address this question, this chapter introduces novel statistical models that are guided by three desiderata. First, all models explicitly respect the hierarchical nature of the data, that is, they are *mixed-effects* models that simultaneously account for *within-subjects* (fixed-effects) and *across-subjects* (random-effects) variance components. Second, maximum-likelihood estimation is replaced by full Bayesian inference in order to enable natural regularization of the estimation problem and to afford conclusions in terms of posterior probability statements. Third, inference on classification accuracy is complemented by inference on the balanced accuracy, which avoids inflated accuracy estimates for imbalanced datasets.

We introduce hierarchical models that satisfy these criteria and demonstrate their advantages over conventional methods using both Markov chain Monte Carlo (MCMC) and variational Bayes (VB) implementations for model inversion and model selection. We illustrate the strengths of these

methods using both synthetic and empirical fMRI data. Corresponding publications are Brodersen, Mathys *et al.* (*in press*) for model inversion using MCMC; and Brodersen, Daunizeau *et al.* (*under review*) for a subsequent variational Bayesian treatment.

4.1 Hierarchical analyses and mixed-effects inference

Classification algorithms are frequently applied to data whose underlying structure is hierarchical (as defined in the next paragraph; Figure 4.1). One example is the domain of brain-machine interfaces, where classifiers are used to decode intentions and decisions from trial-wise measurements of neuronal activity in individual subjects (Sitaram *et al.*, 2008; Blankertz *et al.*, 2011). Another example is spam detection, where a classifier is trained separately for each user to predict content classes from high-dimensional document signatures (Cormack, 2008). A third example is the field of neuroimaging, where classifiers are used to relate subject-specific multivariate measures of brain activity to a particular cognitive or perceptual state (Cox and Savoy, 2003; Haynes and Rees, 2006; Norman *et al.*, 2006; Tong and Pratte, 2012).

In all of these scenarios, the data have a two-level structure: they comprise n experimental trials (or e-mails, or brain scans) collected from each member of a group of m subjects (or users, or patients). For each subject, the classifier is trained and tested on separate partitions of trial-specific data. This procedure gives rise to a set of true labels and a set of predicted labels, separately for each subject within the group. The typical question of interest for studies as those described above is: what is the accuracy of the classifier in the general population from which the subjects were sampled? This chapter is concerned with such group-level inference on classification accuracy for hierarchically structured data.

In contrast to a large literature on evaluating classification performance in non-hierarchical applications of classification (see Langford, 2005, for a review), relatively little attention has been devoted to evaluating classification algorithms in hierarchical (i.e., group) settings (but see Goldstein, 2010; Olivetti *et al.*, 2012).

This is unfortunate since a broadly accepted standard would be highly beneficial. Rather than treating classification outcomes obtained in different subjects as samples from the same distribution, a hierarchical setting requires us to account for the fact that each subject itself has been sam-

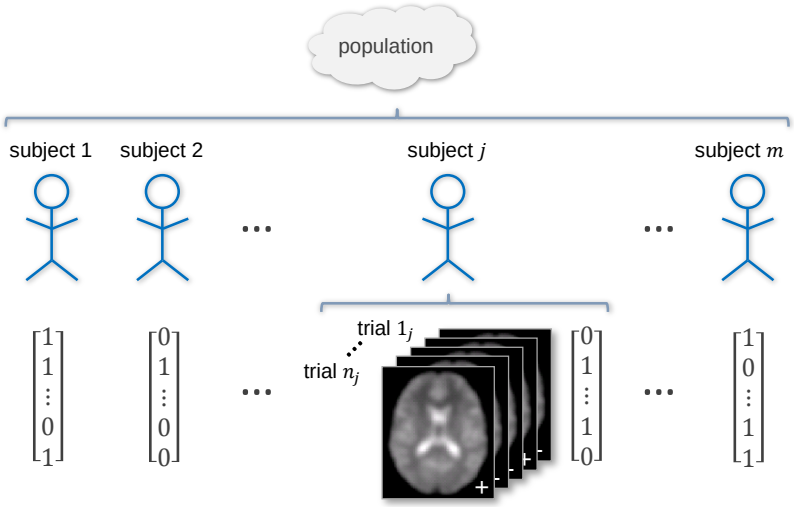

Figure 4.1: Hierarchical trial-by-trial classification. When classifying trials in individual subjects from a group, we must account for uncertainty at the within-subjects and the between-subjects level. This can be achieved using a hierarchical model which we will invert using two different approaches to approximate Bayesian inference: Markov chain Monte Carlo and variational Bayes.

pled from a heterogeneous population (Beckmann *et al.*, 2003; Friston *et al.*, 2005).

Thus, a standard approach to evaluating classification performance should account for two independent sources of uncertainty: *fixed-effects* variance (i.e., within-subjects variability) that results from uncertainty about the true classification accuracy in any given subject; and *random-effects* variance (i.e., between-subjects variability) that results from the distribution of true accuracies in the population from which subjects were drawn. Taking into account both types of uncertainty requires *mixed-effects* inference. This is a central theme of the models discussed in this chapter.

Conventional approaches. There are several common approaches to performance evaluation in hierarchical classification studies (Figure 4.2).[1] One approach rests on the *pooled sample accuracy*, i.e., the number of cor-

[1]This thesis focuses on parametric models for performance evaluation. Nonparametric methods, in particular permutation tests, are not considered in detail here.

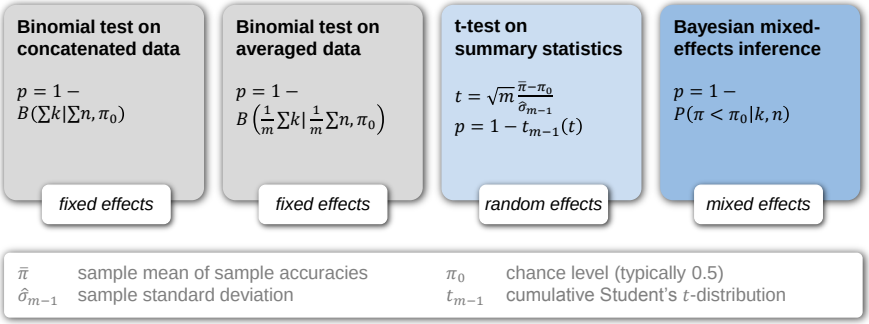

Figure 4.2: Approaches to inference on classification performance. Different approaches to inference can be distinguished on the basis of the sources of uncertainty they account for. Fixed-effects and random-effects approaches are suboptimal for group studies since they disregard between-subjects or within-subjects uncertainty, respectively. Mixed-effects models, by contrast, account for both sources of uncertainty and appropriately constrain inferences at the subject and group level.

rectly predicted trials divided by the number of trials in total, across all subjects. 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 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; Krajbich *et al.*, 2009; Knops *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 (although see Chapter 3 for a more flexible Bayesian treatment). 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 severely overoptimistic inferences. 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 their inherent violations of homoscedasticity (i.e., the differences in variance of the data between subjects).

The second limitation inherent in the above methods is rooted in their distributional assumptions. It is ill-justified to assume sample accuracies to 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 problematic characteristic of (though not intrinsic to) the above conventional methods is a consequence of their 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 favour of the majority class, resulting in an overoptimistic accuracy (Chawla *et al.*, 2002; Japkowicz and Stephen, 2002; Akbani *et al.*, 2004; Wood *et al.*, 2007; Zhang and Lee, 2008; Demirci *et al.*, 2008; Brodersen *et al.*, 2010a). This motivates the use of an alternative performance measure, the *balanced accuracy*, which removes this bias from performance evaluation.

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

It is worth highlighting that the above considerations correspond closely to those that drove the development of mixed-effects models and Bayesian estimation approaches in other domains of analysis. One example are mass-univariate fMRI analyses based on the general linear model (GLM), where

[2]All models discussed in this chapter have been implemented in MATLAB and can be downloaded from: http://mloss.org/software/view/407/. An R package is currently in preparation.

early fixed-effects models were soon replaced by random-effects and full mixed-effects approaches that have become a standard in the field (Holmes and Friston, 1998; Friston *et al.*, 1999; Beckmann *et al.*, 2003; Woolrich *et al.*, 2004; Friston *et al.*, 2005; Mumford and Nichols, 2009).

Another example are group analyses on the basis of DCM (Friston *et al.*, 2003), where fixed-effects inference has been supplemented by random-effects inference that is more appropriate when different models are optimal in characterizing different subjects in a group (Stephan *et al.*, 2009a). The present study addresses the same issues, but in a new context, that is, in group analyses based on trial-by-trial classification. We will revisit this point in Section 4.8 on p. 122.

Overview. This chapter is organized as follows. We begin by briefly reviewing classical approaches to inference in two-level designs (Section 4.2). We then describe both existing and novel models for inferring on the accuracy and the balanced accuracy using stochastic approximate inference, i.e., sampling (Sections 4.3 and 4.4). Following this, we describe models for which we derive a computationally more efficient variational Bayes approximation (Sections 4.5 and 4.6). Illustrative applications of these models on both synthetic and empirical data are provided throughout the chapter (and in particular in Section 4.7). Finally, we review the key characteristics of our models and their inversion strategies and discuss their role in future classification studies (Section 4.8).

4.2 Classical inference in a group study

In a hierarchical setting, group-level inference frequently proceeds by applying a one-sample, one-tailed t-test to subject-specific sample accuracies.[3] This test evaluates the null hypothesis that subject-specific accuracies are drawn from a distribution with a mean at chance level, using the t-statistic

$$\sqrt{m}\frac{\bar{\pi} - \pi_0}{\hat{\sigma}_{m-1}} \sim t_{m-1}, \tag{4.2.1}$$

[3]This chapter focuses on those classical procedures that are widely used in application domains such as neuroimaging and brain-machine interfaces. However, it is worth noting that alternative maximum-likelihood procedures exist that eschew the normality assumption implicit in a classical t-test (e.g., Dixon, 2008). We will revisit this point in Section 4.8.

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

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

$$\left[\bar{\pi} \pm t_{0.025,m-1} \times \frac{\hat{\sigma}_{m-1}}{\sqrt{m}}\right], \tag{4.2.2}$$

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

The differences between the classical procedure and the full Bayesian approach discussed earlier can be best understood by considering their respective assumptions. The distributional assumption underlying both the t-statistic in (4.2.1) and the confidence interval in (4.2.2) is that the sample mean of the subject-wise accuracies, under the null hypothesis, is normally distributed,

$$\bar{\pi} \sim \mathcal{N}\left(\bar{\pi} \,\middle|\, \mu, \frac{\sigma}{\sqrt{m}}\right), \tag{4.2.3}$$

where the population standard deviation σ has been estimated by the sample standard deviation $\hat{\sigma}_{m-1}$. The corresponding graphical model is shown in Figure 4.3a.

As touched upon in Section 4.1, this analysis is popular but suffers from two faults that are remedied by our Bayesian treatment.[4] First, accuracies

[4]For a classical mixed-effects approach, see Section 4.8.

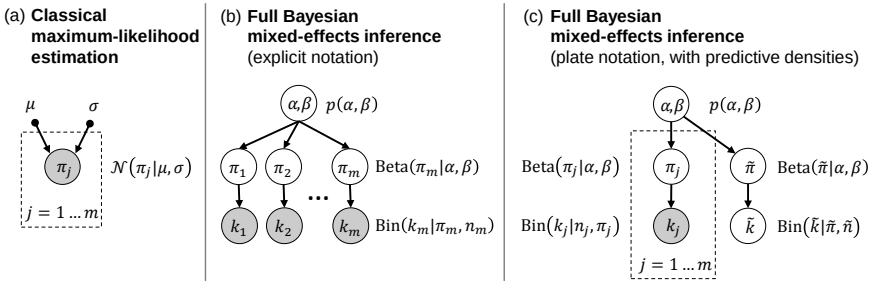

Figure 4.3: Models for inference on classification accuracies. This illustration shows graphical representations of different models for classical and Bayesian inference on classification accuracies, as discussed in Sections 4.3.1 and 4.4.1. Blank circles correspond to latent variables, filled circles represent observed data.

are confined to the $[0, 1]$ interval, but are modelled by a normal distribution with infinite support. Consequently, error bars based on confidence intervals (4.2.2) may well include values above 1 (see Figure 4.6c for an example). By contrast, the Beta distribution used in the Bayesian approach has the desired $[0, 1]$ support and thus represents a more natural candidate distribution.

Second, a t-test on sample accuracies as in (4.2.3) neither explicitly accounts for within-subjects uncertainty nor for violations of homoscedasticity. This is because it uses sample accuracies as summary statistics, treating them as infinitely precise observations, without carrying forward the uncertainty associated with them (cf. Mumford and Nichols, 2009). For example, no distinction is made 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). In fact, 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 summary, classifier performance cannot be observed directly; it must be inferred. While the classical model above does allow for inference on random-effects (between-subjects) variability, it does not explicitly account for fixed-effects (within-subject) uncertainty. This uncertainty is only taken into account indirectly by its influence on the variance of the observed sam-

ple accuracies.

With regard to subject-specific accuracies, one might be tempted to use $\hat{\pi}_j = k_j/n_j$ as individual estimates. However, in contrast to Bayesian inference on subject-specific accuracies (see Section 4.3.1), individual sample accuracies do not take into account the moderating influence provided by knowledge about the group (i.e., 'shrinkage'). An effectively similar outcome is found in classical inference using the James-Stein estimator (James and Stein, 1961, see Appendix A1.2).

4.3 Stochastic Bayesian inference on the accuracy

As described above, in a hierarchical setting, a classifier predicts the class label of each of n trials, separately for each subject from a group. This setting raises three principal questions. First, what is the classification accuracy at the group level? This is addressed by inference on the mean classification accuracy in the population from which subjects were drawn. Second, what is the classification accuracy in each individual subject? Addressing this question by considering each subject in turn is possible but potentially wasteful, since within-subject inference may benefit from across-subject inference (Efron and Morris, 1971). Third, which of several classification algorithms is best? This question can be answered by estimating how well an algorithm's classification performance generalizes to new data.

This section considers different models for answering these questions. To keep the chapter self-contained, we initially review the well-known beta-binomial model (Pearson, 1925; Skellam, 1948; Lee and Sabavala, 1987). We also describe the normal-binomial model, which will later prove useful for our variational treatment. This introduces most of the concepts we require for subsequently introducing two new models designed to support hierarchical Bayesian inference on the balanced accuracy: the twofold beta-binomial model and the bivariate normal-binomial model.

4.3.1 The beta-binomial model

In a hierarchical classification study, classification is carried out separately for each subject within a group, hence the available data are k_j out of n_j correct predictions for each subject $j = 1 \ldots m$. At the level of individual subjects, for each subject j, the number of correctly classified trials k_j can

be modelled as

$$p(k_j \mid \pi_j, n_j) = \mathrm{Bin}(k_j \mid \pi_j, n_j) \tag{4.3.1}$$

$$= \binom{n_j}{k_j} \pi_j^{k_j} (1-\pi_j)^{n_j - k_j}, \tag{4.3.2}$$

where n_j is the total number of trials in subject j, and π_j represents the fixed but unknown accuracy that the classification algorithm achieves on that subject.[5] At the group level, the model must account for variability across subjects. This is achieved by modelling subject-wise accuracies as drawn from a population distribution described by a Beta density,

$$p(\pi_j \mid \alpha, \beta) = \mathrm{Beta}(\pi_j \mid \alpha, \beta) \tag{4.3.3}$$

$$= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \pi_j^{\alpha-1} (1-\pi_j)^{\beta-1}, \tag{4.3.4}$$

such that α and β characterize the population as a whole. This step is formally identical with the Beta prior placed on the accuracy in (3.1.9) on p. 47 where it represented our subjective uncertainty about π before observing the outcome k. Here, (4.3.4) states that uncertainty about any particular subject is best quantified by our knowledge about variability in the population, i.e., the distribution of π_j over subjects (which, as described below, can be learnt from the data). Formally, subject-specific accuracies are assumed to be i.i.d., conditional on the population parameters α and β.

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

$$p(\alpha, \beta) \propto (\alpha+\beta)^{-5/2} \tag{4.3.5}$$

on the natural scale. However, although this prior leads to a proper posterior density, it is improper itself and thus prevents computation of the model

[5] Note that our notation will suppress n_j unless this introduces ambiguity.

evidence, i.e., the marginal likelihood of the data given the model, which will later become important for model comparison. We resolve this limitation by using a proper (i.e., integrable and normalized) variant,

$$p(\alpha, \beta) = \frac{3}{4}(\alpha + \beta + 1)^{-5/2} \tag{4.3.6}$$

which represents a special case of the generalization of (4.3.5) proposed by Everson and Bradlow (2002). The positive constant (here: +1) ensures integrability. This prior can be rewritten in an unnormalized (as indicated by a tilde), reparameterized form as

$$\tilde{p}\left(\ln\left(\frac{\alpha}{\beta}\right), \ln(\alpha + \beta)\right) = \alpha\beta(\alpha + \beta + 1)^{-5/2}, \tag{4.3.7}$$

which will be useful in the context of model inversion (Gelman *et al.*, 2003). Two equivalent graphical representations of this model (using the formalism of Bayesian networks; Jensen and Nielsen, 2007) are shown in Figures 4.3a and 4.3b.

4.3.2 Stochastic approximate inference

Inverting the beta-binomial model allows us to infer on (i) the posterior population mean accuracy, (ii) the subject-specific posterior accuracies, and (iii) the posterior predictive accuracy. We propose a numerical procedure for model inversion (Appendix A.1) which we briefly summarize below.

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

$$p(\alpha, \beta \mid k_{1:m}) = \frac{p(k_{1:m} \mid \alpha, \beta)\, p(\alpha, \beta)}{\iint p(k_{1:m} \mid \alpha, \beta)\, p(\alpha, \beta)\, \mathrm{d}\alpha\, \mathrm{d}\beta} \tag{4.3.8}$$

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

$$p(k_{1:m} \mid \alpha, \beta) = \prod_{j=1}^{m} \int p(k_j \mid \pi_j)\, p(\pi_j \mid \alpha, \beta)\ \mathrm{d}\pi_j \tag{4.3.9}$$

$$= \prod_{j=1}^{m} \mathrm{Bb}(k_j \mid \alpha, \beta), \tag{4.3.10}$$

where $\mathrm{Bb}(\cdot)$ denotes the beta-binomial distribution. Since the integral on the right-hand side of (4.3.8) cannot be evaluated in closed form, we resort to a Markov chain Monte Carlo (MCMC) procedure. Specifically, we use a Metropolis algorithm (Metropolis and Ulam, 1949; Metropolis *et al.*, 1953) to sample from the variables at the top level of the model and obtain a set $\{(\alpha^{(\tau)}, \beta^{(\tau)})\}$ for $\tau = 1 \ldots c$. This set allows us to obtain samples from the posterior population mean accuracy,

$$p\left(\frac{\alpha}{\alpha+\beta} \;\middle|\; k_{1:m}\right). \tag{4.3.11}$$

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

$$\frac{1}{c}\sum_{\tau=1}^{c}\frac{\alpha^{(\tau)}}{\alpha^{(\tau)}+\beta^{(\tau)}}. \tag{4.3.12}$$

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

$$p = P\left(\frac{\alpha}{\alpha+\beta} \leq 0.5 \;\middle|\; k_{1:m}\right) \tag{4.3.13}$$

which can be viewed as a Bayesian analogue of a classical p-value. We refer to this quantity as the (posterior) *infraliminal probability* of the classifier. It lives on the same $[0, 1]$ scale as a classical p-value, but has a more intuitive (and less error-prone) interpretation: rather than denoting the (frequentist) probability of observing the data (or more extreme data) under the 'null hypothesis' of a chance classifier (classical p-value), the infraliminal probability represents the (posterior) belief that the classifier operates at or below chance. We will revisit this aspect in Section 4.8.

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

$$P\left(\frac{\alpha^{(1)}}{\alpha^{(1)}+\beta^{(1)}} > \frac{\alpha^{(2)}}{(\alpha^{(2)}+\beta^{(2)})} \;\middle|\; k_{1:m^{(1)}}, k_{1:m^{(2)}}\right). \tag{4.3.14}$$

The second question of interest concerns the classification accuracies in individual subjects. Specifically, we wish to infer on $p(\pi_j \mid k_{1:m})$ to characterize our posterior uncertainty about the true classification accuracy in subject

j. Given a pair of samples $(\alpha^{(\tau)}, \beta^{(\tau)})$, we can obtain samples from subject-specific posteriors simply by drawing from

$$\text{Beta}\left(\pi_j^{(\tau)} \;\middle|\; \alpha^{(\tau)} + k_j,\ \beta^{(\tau)} + n_j - k_j\right). \tag{4.3.15}$$

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

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

$$p(\tilde{\pi} \mid k_{1:m}), \tag{4.3.16}$$

in which $\tilde{\pi}$ denotes the classification accuracy in a new subject drawn from the same population as the existing group of subjects with latent accuracies $\pi_1, \ldots, \pi_m$ (cf. Figure 4.3b).[6] Samples for this density can be easily obtained using the samples $\alpha^{(\tau)}$ and $\beta^{(\tau)}$ from the posterior population mean.[7]

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

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

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

4.3.3 Applications

This section begins to illustrate the practical utility of the Bayesian approach discussed in the previous section and compares it to inference obtained through classical (frequentist) statistics.

Inference on the population mean and the predictive accuracy

In a first experiment, we simulated classification outcomes for a group of 20 subjects with 100 trials each (50 trials with a positive and 50 with a negative hidden true class label). Outcomes were generated using the beta-binomial model with a population mean of 0.8 and a population variance of 0.01 (i.e., $\alpha = 12$ and $\beta = 3$ for both positive and negative labels, corresponding to a population standard deviation of 0.1; Figure 4.4).

Raw data, i.e., the number of correct predictions within each subject, are shown in Figure 4.4a. Their empirical sample accuracies are shown in Figure 4.4b, along with the ground-truth density of the population accuracy. Inverting the beta-binomial model, using the MCMC procedure of Section 4.3.1 (Figure 4.4c), and examining the posterior distribution over the population mean accuracy showed that more than 99.9% of its mass was above 50%, in agreement with the fact that the true population mean was above chance (Figure 4.4d).

We also used this simulation to illustrate the differences between a Bayesian mixed-effects central 95% posterior probability interval, a fixed-effects probability interval, and a random-effects confidence interval (Figure 4.4e). All three schemes arrive at the same conclusion with respect to the population mean being above chance. However, while the random-effects interval (red) is very similar to the proposed mixed-effects interval (black), the fixed-effects interval (yellow) displays too small a variance as it disregards the important between-subjects variability.

We finally considered the predictive posterior distribution over the accuracy that would be observed if we were to acquire data from a new subject (Figure 4.4f). This posterior did not allow for the conclusion that, with a probability larger than 0.95, the accuracy in a new subject would be above chance. This result is driven by the large heterogeneity in the population, inducing a dispersed predictive density. Importantly, the dispersion of the predictive density would not vanish even in the limit of an infinite number of subjects. This is in contrast to the dispersion of the posterior over the population mean, which becomes more and more precise with an increasing amount of data.

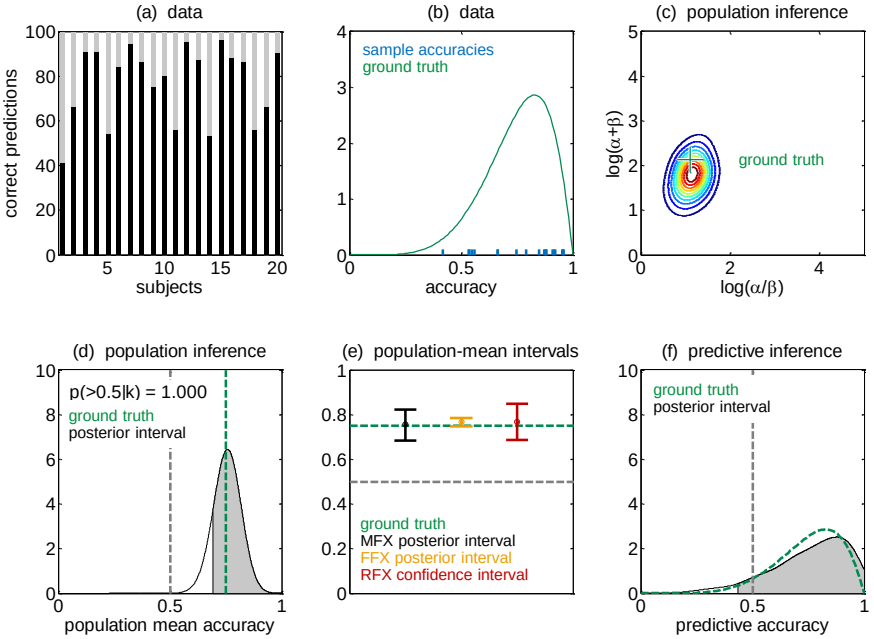

Figure 4.4: Inference on the population mean and the predictive accuracy. (a) Classification outcomes were generated for 20 subjects using the beta-binomial model. Each subject is fully characterized by the number of correctly classified trials (black) out of a given set of 100 trials (grey). (b) Empirical sample accuracies (blue) and their underlying population distribution (green). (c) Inverting the beta-binomial model yields samples from the posterior distribution over the population parameters, visualized using a nonparametric (bivariate Gaussian kernel) density estimate (contour lines). (d) The posterior about the population mean accuracy, plotted using a kernel density estimator (black), is sharply peaked around the true population mean (green). The upper 95% of the probability mass are shaded (grey). Because the lower bound of the shaded area is greater than 0.5, the population mean can be concluded to be above chance. (e) While the central 95% posterior interval (black) and the classical 95% confidence interval (red) look similar, the two intervals are conceptually very different. The fixed-effects interval (orange) is overly optimistic as it disregards between-subjects variability. (f) The posterior predictive distribution over $\tilde{\pi}$ represents the posterior belief of the accuracy expected in a new subject (black). Its dispersion reflects irreducible population heterogeneity.

Inference was based on 100 000 samples, generated using 8 parallel chains. We used several standard approaches to convergence evaluation. In particular, we considered trace plots for visual inspection of mixing behaviour

and convergence to the target distributions. In addition, we monitored the average ratio of within-chain variance to between-chain variance, which was bigger than 0.995. In other words, the variances of samples within and between chains were practically indistinguishable. The Metropolis rejection rate was 0.475, thus ensuring an appropriate balance between exploration (of regions with a lower density) and exploitation (of regions with a higher density). Finally, we assessed the uncertainty inherent in MCMC-based quantities such as log Bayes factors by computing standard deviations across repetitions, which led us to use 10^5 or 10^6 samples for each computation (see Section 4.4.6). All subsequent applications were based on the same algorithmic settings.

In frequentist inference, a common way of representing the statistical properties of a test is to estimate the probability of rejecting the null hypothesis at a fixed threshold (e.g., 0.05) under different regimes of ground truth, which leads to the concept of power curves. Here, we adopted this frequentist perspective to illustrate the properties of Bayesian mixed-effects inference on classification performance (Figure 4.5).

Specifying a true population mean of 0.5 and variance of 0.001 (standard deviation 0.0316), we generated classification outcomes, in the same way as above, for a synthetic group of 20 subjects with 100 trials each. Inverting the beta-binomial model, we inferred whether the population mean was above chance by requiring more than 95% of the posterior probability mass of the population mean to be greater than 0.5, that is, by requiring an infraliminal probability of less than 5%. We repeated this process 1 000 times and counted how many times the population mean was deemed greater than chance. We then varied the true population mean and plotted the fraction of decisions for an above-chance classifier as a function of population mean (Figure 4.5a). At a population mean of 0.5, the vertical distance between the data points and 1 represents the empirical specificity of the test (which was designed to be $1 - \alpha = 0.95$). At population means above 0.5, the data points show the empirical sensitivity of the test, which grows rapidly with increasing population mean. In this setting, the inferences that one would obtain by a frequentist t-test (red) are in excellent agreement with those afforded by the proposed beta-binomial model (black). Since the population variance was chosen to be very low in this initial simulation, the inferences afforded by a fixed-effects analysis (yellow) prove very similar as well; but this behaviour changes drastically when increasing the population variance to more realistic levels, as described below.

One important issue in empirical studies is the heterogeneity of the pop-

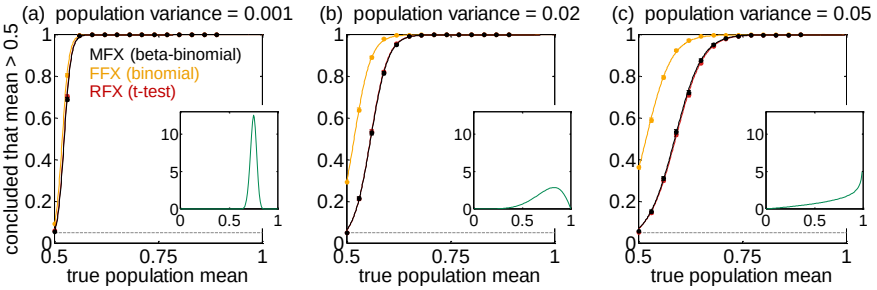

Figure 4.5: Inference on the population mean under varying population heterogeneity. The figure shows Bayesian estimates of the frequentist probability of above-chance classification performance, as a function of the true population mean, separately for three different level of population heterogeneity (a,b,c). Each data point is based on 1 000 simulations, each of which used 10 000 samples from every subject-specific posterior to make a decision. The figure shows that, in this setting, frequentist inference based on *t*-tests (red) agrees with Bayesian inference based on the beta-binomial model (black). By contrast, a fixed-effects approach (orange) offers invalid population inference as it disregards between-subjects variability; at a true population mean of 0.5, the hypothesis of chance-level performance is rejected more frequently than prescribed by the test size. Each data point is plotted in terms of the fraction of above-chance conclusions and a 95% central posterior interval, based on a Beta model with a flat prior. Points are joined by a sigmoidal function that was constrained to start at 0 and end at 1, with two remaining degrees of freedom. The insets show the distribution of the true underlying population accuracy (green) for a population mean accuracy of 0.8. Where the true population mean exceeds 0.5, the graphs reflect the empirical sensitivity of the inference scheme. Its empirical specificity corresponds to the vertical distance between the graphs and 1 at the point where the population mean is 0.5.

ulation. We studied the effects of population variance by repeating the above simulations with different variances (Figures 4.5b,c). As expected, an increase in population variance reduced statistical sensitivity. For example, given a fairly homogeneous population with a true population mean accuracy of 60% and a variance of 0.001 (standard deviation 0.0316; Figure 4.5a), we can expect to correctly infer above-chance performance in more than 99.99% of all cases. By contrast, given a more heterogeneous population with a variance of 0.05 (standard deviation $\approx$ 0.22), the fraction of correct conclusions drops to 61%; in all other cases we would fail to recognize that the classifier was performing better than chance.

The above simulations show that a fixed-effects analysis (yellow) becomes an invalid procedure to infer on the population mean when the population variance is non-negligible. In more than the prescribed 5% of sim-

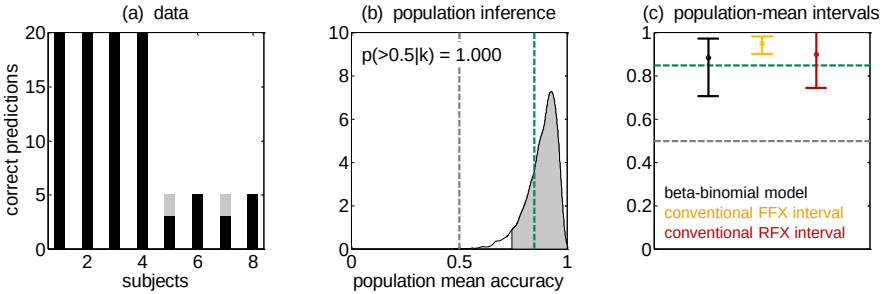

Figure 4.6: Inadequate inferences provided by fixed-effects and random-effects models. (a) The simulation underlying this figure represents the case of a small heteroscedastic group with varying numbers of trials across subjects. Classification outcomes were generated in the same way as in the simulation underlying Figure 4.5a. (b) The (mixed-effects) posterior density of the population mean (black) provides a good estimate of ground truth (green). (c) A central 95% posterior probability interval, based on the density shown in (b), comfortably includes ground truth. By contrast, a fixed-effects interval (orange) is overoptimistic as it disregards between-subjects variability. The corresponding random-effects confidence interval (red) is similar to the mixed-effects interval but lacks asymmetry, thus inappropriately including accuracies above 100%.

ulations with a true population mean of 0.5, the procedure concluded that the population mean was above chance. This is because a fixed-effects analysis yields too small variances on the population mean and therefore too easily makes above-chance conclusions.

All above simulations were based on a group of 20 subjects with 100 trials each, emulating a setting as it frequently occurs in practice, e.g., in neuroimaging data analyses. We repeated the same analysis as above on a sample dataset from a second simulation setting (Figure 4.6). This setting was designed to represent the example of a small heterogeneous group with varying numbers of trials across subjects. Specifically, we generated data for 8 subjects, half of which had 20 trials, and half of which had only 5 trials. Classification outcomes were generated using the beta-binomial model with a population mean of 0.85 and a population variance of 0.02 (corresponding to a population standard deviation of 0.14; Figure 4.6a).

The example shows that the proposed beta-binomial model yields a posterior density with the necessary asymmetry; it comfortably includes the true population mean (Figure 4.6b). By contrast, the fixed-effects probability interval (based on a Beta density) is overly optimistic. Finally, the random-effects confidence interval is similar to the mixed-effects interval

but lacks the necessary asymmetry, including accuracies above 100% (Figure 4.6c).

Inference on subject-specific accuracies

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

To illustrate this, we generated synthetic classification outcomes and computed subject-specific posterior inferences (Figure 4.7). This simulation was based on 45 subjects overall; 40 subjects were characterized by a relatively moderate number of trials ($n = 20$) while 5 subjects had even fewer trials ($n = 5$). The population accuracy had a mean of 0.8 and a variance of 0.01 (standard deviation 0.1). Using this dataset, we computed subject-specific central 95% posterior probability intervals and sorted them in ascending order by subject-specific sample accuracy (Figure 4.7a). The plot shows that, in each subject, the posterior mode (black) represents a compromise between the observed sample accuracy (blue) and the population mean (0.8). This compromise in turn provides a better estimate of ground truth (green) than sample accuracies by themselves. This effect demonstrates a key difference between the two types of inference: subject-specific posteriors are informed by data from the entire group, whereas sample accuracies are based on the data from an individual subject.

Another way of demonstrating the shrinkage effect is by illustrating the transition from ground truth to sample accuracies (with its increase in dispersion) and from sample accuracies to posterior means (with its decrease in dispersion). This shows how the high variability in sample accuracies is reduced, informed by what has been learned about the population (Figure 4.7b). Notably, because the amount of shrinking depends on each subject's posterior uncertainty, the shrinking effect may modify the order of subjects, as indicated by crossing lines. Here, subjects with only 5 trials were shrunk more than subjects with 20 trials.

In a next step, we examined power curves, systematically changing the true population accuracy and repeating the above simulation 1 000 times (Figure 4.7c). Within a given simulation, we concluded that a subject-specific accuracy was above chance if more than 95% of its posterior prob-

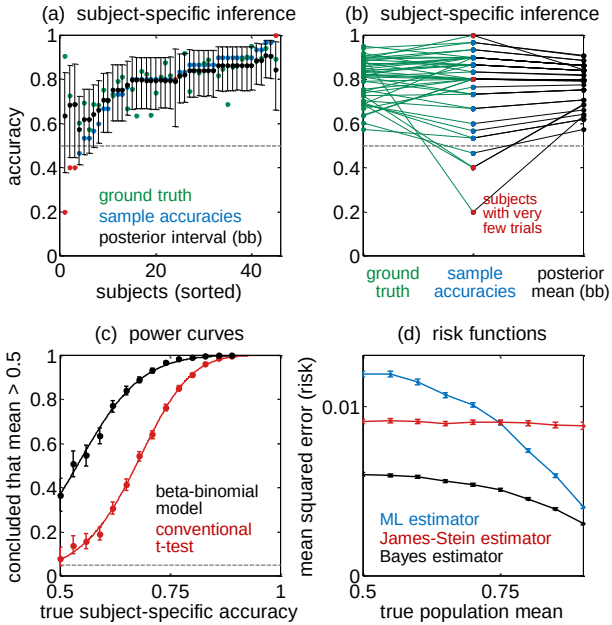

Figure 4.7: Inference on subject-specific accuracies. (a) Classification outcomes for a synthetic heterogeneous group of 45 subjects. The figure shows subject-specific posterior means and central 95% credible intervals (black), sample accuracies (blue if based on 20 trials, red if based on 5 trials), and true subject-specific accuracies (green) as a function of subject index, sorted in ascending order by sample accuracy. Due to the hierarchical model, Bayesian posterior intervals are shrunk to the population. (b) Alternative visualization of the shrinkage effect. Shrinking changes the order of subjects (when sorted by posterior mean as opposed to by sample accuracy) as the amount of shrinking depends on the subject-specific (first-level) posterior uncertainty. Subjects with just 5 trials (red) are shrunk more than subjects with 20 trials (blue). (c) Based on 1 000 simulations, the plot shows the fraction of simulations in which a subject's accuracy was concluded to be above chance, based on a Bayesian posterior interval (black) or a frequentist t-test (red). In contrast to classical inference, the Bayesian procedure incorporates a desirable shift towards the population in making decisions about individual group members. (d) Across the same 1 000 simulations, a Bayes estimator, based on the posterior means of subject-specific accuracies (black), was superior to both a classical ML estimator (blue) and a James-Stein estimator (red).

ability mass was above 0.5. We binned subjects across all simulations into groups of similar accuracies and plotted the fraction of above-chance decisions against these true accuracies, contrasting the Bayesian model with a

conventional t-test.

As shown in Figure 4.7c, t-tests falsely detected above-chance subject-specific accuracies in about 5% of the cases, in agreement with the intended test size. By contrast, our Bayesian scheme was considerably more sensitive and detected above-chance accuracy in subjects whose true accuracy was within a small bin around 0.5. This behaviour reflected the fact that the Bayesian procedure incorporated what had been learned about the population when deciding on individual subjects. That is, a population mean well above chance (here: 0.8) made it more likely that individual subjects performed above chance as well, even in the presence of a low sample accuracy.

In addition to enabling decisions that take into account information about the group, the posterior distributions of subject-specific accuracies also yield more precise point estimates. To illustrate this effect, we simulated 1 000 datasets in the same way as above. Within each simulation, we compared three different ways of obtaining an estimator for each subject's accuracy: (i) a Bayes estimator (posterior mean of the subject-specific accuracy); (ii) a maximum-likelihood estimator (sample accuracy); and (iii) a James-Stein estimator, with a similar shrinkage effect as the Bayes estimator but less explicit distributional assumptions (Figure 4.7d). For each estimator, we computed the mean squared error (or risk) across all subjects, averaged across all simulations. We then repeated this process for different population means. We found that the James-Stein estimator dominated the ML estimator for low accuracies. However, both estimators were inferior to the Bayes estimator which provided the lowest risk throughout.

4.4 Stochastic Bayesian inference on the balanced accuracy

The beta-binomial model discussed in the previous section enables inference on the accuracy, which is known to be a problematic indicator of performance when classes are not perfectly balanced. Here, we consider two models for inference on the *balanced accuracy*, a more suitable indicator of classification performance.

4.4.1 The twofold beta-binomial model

One way of inferring on the balanced accuracy ϕ is to duplicate the beta-binomial model and apply it separately to the two classes (Figure 4.8a). In

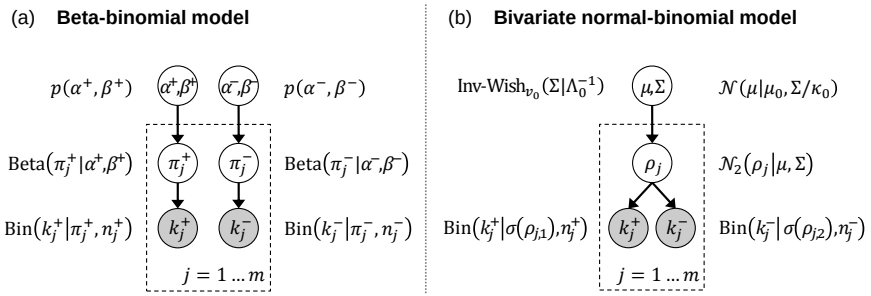

Figure 4.8: Models for inference on balanced classification accuracies. This figure shows two models for Bayesian mixed-effects inference on the balanced accuracy, as discussed in Sections 4.4.1 and 4.4.3. The models are based upon different assumptions and parameterizations and can be compared by Bayesian model comparison.

other words, we consider the number of correctly predicted positive trials k^+ and the number of correctly predicted negative trials k^-, and express our uncertainty about ϕ (3.2.1) before and after observing k^+ and k^-. In a single-subject setting, as in (3.1.9), we can place separate noninformative Beta priors on π^+ and π^-,

$$\begin{aligned} p(\pi^+ \mid \alpha_0^+, \beta_0^+) &= \text{Beta}(\pi^+ \mid \alpha_0^+, \beta_0^+), \\ p(\pi^- \mid \alpha_0^-, \beta_0^-) &= \text{Beta}(\pi^- \mid \alpha_0^-, \beta_0^-), \end{aligned} \tag{4.4.1}$$

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

$$p(\phi \mid k^+, k^-) = p\left(\frac{1}{2}(\pi^+ + \pi^-) \,\middle|\, k^+, k^-\right). \tag{4.4.2}$$

The balanced accuracy is thus a new random variable defined via two existing random variables from our model, π^+ and π^-. Even in a single-subject setting, a closed form for its posterior distribution is not available, and so we must resort to a numerical approximation (cf. Section 3.2 on p. 53).

In a group setting, we can expand the above model in precisely the same way as for the simpler case of the classification accuracy in Section 4.3.1.

Specifically, we define diffuse priors on the class-specific population parameters α^+ and β^+ as well as α^- and β^-, in analogy to (4.3.6). A graphical representation of this model is shown in Figure 4.8a.

4.4.2 Stochastic approximate inference

Given that the twofold beta-binomial model consists of two independent instances of the simple beta-binomial model considered in Section 4.3.1 (Figure 4.3b), statistical inference follows the same approach as described previously (see Section 4.4.6 for an application). For instance, we can obtain the posterior population parameters, $p(\alpha^+, \beta^+ \mid k^+_{1:m})$ and $p(\alpha^-, \beta^- \mid k^-_{1:m})$ using the same sampling procedure as summarized in Section 4.3.1, except that we are now applying the procedure twice. The two sets of samples can then be averaged in a pairwise fashion to obtain samples from the posterior mean balanced accuracy in the population,

$$p\left(\phi \mid k^+_{1:m}, k^-_{1:m}\right), \tag{4.4.3}$$

where we have defined

$$\phi := \frac{1}{2}\left(\frac{\alpha^+}{\alpha^+ + \beta^+} + \frac{\alpha^-}{\alpha^- + \beta^-}\right). \tag{4.4.4}$$

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

$$p\left(\phi_j \mid k^+_{1:m}, k^-_{1:m}\right). \tag{4.4.5}$$

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

$$p\left(\tilde{\phi} \mid k^+_{1:m}, k^-_{1:m}\right). \tag{4.4.6}$$

4.4.3 The bivariate normal-binomial model

In the previous section, we saw that the twofold beta-binomial model enables mixed-effects inference on the balanced accuracy. However, it may not always be optimal to treat accuracies on positive and negative trials

separately (cf. Leonard, 1972). That is, if π^+ and π^- were related in some way, the model should reflect this.

For example, one could imagine a group study in which some subjects exhibit a more favourable signal-to-noise ratio than others, leading to well-separated classes. In this case, an unbiased classifier yields high accuracies on either class in some subjects and lower accuracies in others, inducing a positive correlation between class-specific accuracies

On the other hand, within each subject, any classification algorithm faces a trade-off between performing better on one class at the expense of the other class. Thus, any variability in setting this threshold leads to negatively correlated class-specific accuracies, an argument that is formally related to receiver-operating characteristics. Moreover, if the degree of class imbalance in the data varies between subjects, classifiers might be biased in different ways, again leading to negatively correlated accuracies.

In summary, π^+ and π^- may not always be independent. We therefore turn to an alternative model for mixed-effects inference on the balanced accuracy that embraces potential dependencies between class-specific accuracies (Figure 4.8b).

The bivariate normal-binomial model no longer assumes that π^+ and π^- are drawn from separate populations. Instead, we use a bivariate population density whose covariance structure defines the form and extent of the dependency between π^+ and π^-. For this combined prior, we use a bivariate normal density. Because this density has infinite support, we do not define it on the accuracies themselves but on their log odds. In this way, each subject j is associated with a two-dimensional vector of class-specific accuracies,

$$\rho_j = \begin{pmatrix} \rho_j^+ \\ \rho_j^- \end{pmatrix} = \begin{pmatrix} \sigma^{-1}(\pi_j^+) \\ \sigma^{-1}(\pi_j^-) \end{pmatrix} \in \mathbb{R}^2, \tag{4.4.7}$$

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

$$\pi_j = \begin{pmatrix} \pi_j^+ \\ \pi_j^- \end{pmatrix} = \begin{pmatrix} \sigma(\rho_j^+) \\ \sigma(\rho_j^-) \end{pmatrix} \in [0,1]^2, \tag{4.4.8}$$

where $\sigma(\rho) := 1/(1+\exp(-\rho))$ denotes the sigmoid (or logistic) transform. Thus, we can replace the two independent Beta distributions for π^+ and π^- in (4.4.1) by a single bivariate Gaussian prior,

$$p(\rho_j \mid \mu, \Sigma) = \mathcal{N}_2(\rho_j \mid \mu, \Sigma), \tag{4.4.9}$$

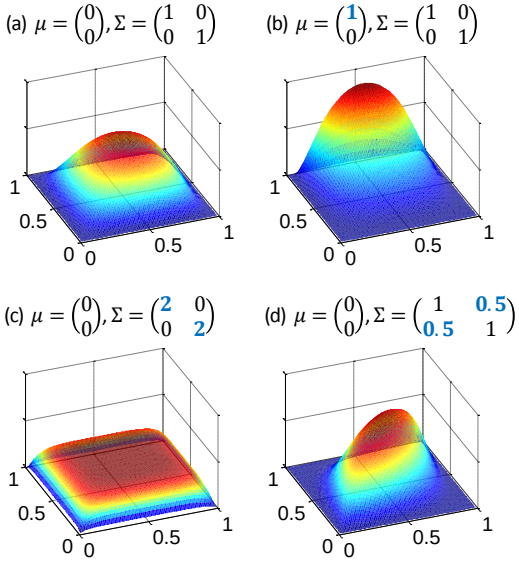

Figure 4.9: Bivariate densities of class-specific accuracies in the bivariate normal-binomial model. In the bivariate normal-binomial model (Section 4.4.3), class-specific accuracies are assumed to follow a bivariate logit-normal distribution. This figure illustrates the flexibility of this distribution. Specifically, (a) the standard parameterization is compared to a distribution with (b) an increased accuracy on one class but not the other, (c) an increased population heterogeneity, and (d) a correlation between class-specific accuracies. The x- and y-axis represent the accuracies on positive and negative trials, respectively.

in which $\mu \in \mathbb{R}^2$ represents the population mean and $\Sigma \in \mathbb{R}^{2\times 2}$ encodes the covariance structure between accuracies on positive and negative trials. The resulting density on $\pi \in \mathbb{R}^2$ is a bivariate logit-normal distribution (Figure 4.9).

In analogy with the prior placed on α and β in Section 4.3.1, we now specify a prior for the population parameters μ and Σ. Specifically, we seek a diffuse prior that induces a noninformative bivariate distribution over $[0, 1] \times [0, 1]$. We begin by considering the family of conjugate priors for

(μ, Σ), that is, the bivariate normal-inverse-Wishart distribution,

$$p(\mu, \Sigma \mid \mu_0, \kappa_0, \Lambda_0, \nu_0)$$
$$\propto |\Sigma|^{-\left(\frac{\nu_0}{2}+2\right)} \exp\left(-\frac{1}{2}\mathrm{tr}(\Lambda_0\Sigma^{-1}) - \frac{\kappa_0}{2}(\mu-\mu_0)^{\mathrm{T}}\Sigma^{-1}(\mu-\mu_0)\right). \quad (4.4.10)$$

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

$$p(\Sigma \mid \Lambda_0, \nu_0) = \text{Inv-Wishart}_{\nu_0}(\Sigma \mid \Lambda_0^{-1}) \quad \text{and} \quad (4.4.11)$$
$$p(\mu \mid \Sigma, \mu_0, \kappa_0) = \mathcal{N}_2(\mu \mid \mu_0, \Sigma/\kappa_0). \quad (4.4.12)$$

In order to illustrate the flexibility offered by the bivariate normal density on ρ, we derive $p(\pi \mid \mu, \Sigma)$ in closed form Appendix B.2 and then compute the bivariate density on a two-dimensional grid (Figure 4.9).

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

$$\mu_0 = (0,0)^{\mathrm{T}},\ \kappa_0 = 1,\ \Lambda_0 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}^{-1},\ \nu_0 = 5. \quad (4.4.13)$$

4.4.4 Stochastic approximate inference

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

A better strategy than Metropolis sampling is to use a Gibbs sampler (Geman and Geman, 1984) to draw from the joint posterior $p(\rho_{1:m}, \mu, \Sigma \mid k^+_{1:m}, k^-_{1:m})$, from which we can derive samples for the conditional posteriors $p(\rho_{1:m} \mid k^+_{1:m}, k^-_{1:m})$ and $p(\mu, \Sigma \mid k^+_{1:m}, k^-_{1:m})$. In contrast to a Metropolis scheme, Gibbs sampling requires only full conditionals, that is, distributions of one latent variable conditioned on all other variables in the model (Gelfand and Smith, 1990). Whenever a full conditional is not available, we

can sample from it using a Metropolis step. Thus, we combine a Gibbs skeleton with interleaved Metropolis steps to sample from the posterior $p(\rho_{1:m}, \mu, \Sigma \mid k^+_{1:m}, k^-_{1:m})$. See Section 4.4.6 for an application.

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

4.4.5 Bayesian model selection

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

$$p(k^+_{1:m}, k^-_{1:m} \mid M) = \int p(k^+_{1:m}, k^-_{1:m} \mid \theta)\; p(\theta \mid M)\; \mathrm{d}\theta \tag{4.4.14}$$

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

$$p(M \mid k^+_{1:m}, k^-_{1:m}) \propto p(k^+_{1:m}, k^-_{1:m} \mid M) \tag{4.4.15}$$

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

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

$$\ln p(k^+_{1:m}, k^-_{1:m} \mid M_{bb}) \tag{4.4.16}$$

$$= \ln \int p(k^+_{1:m} \mid \pi^+_{1:m})\, p(\pi^+_{1:m})\, \mathrm{d}\pi^+_{1:m}$$

$$+ \ln \int p(k^-_{1:m} \mid \pi^-_{1:m})\, p(\pi^-_{1:m})\, \mathrm{d}\pi^-_{1:m} \tag{4.4.17}$$

$$= \ln \left\langle \prod_{j=1}^{m} p(k^+_j \mid \pi^+_j) \right\rangle_{\pi^+_{1:m}} + \ln \left\langle \prod_{j=1}^{m} p(k^-_j \mid \pi^-_j) \right\rangle_{\pi^-_{1:m}} \tag{4.4.18}$$

where we have omitted the conditional dependence on M_{bb} in (4.4.17) and (4.4.18).[8] The expression can be approximated by

$$\approx \ln \frac{1}{c} \sum_{\tau=1}^{c} \prod_{j=1}^{m} \mathrm{Bin}\left(k^+_j \;\middle|\; \pi^{+(\tau)}_j\right) + \ln \frac{1}{c} \sum_{\tau=1}^{c} \prod_{j=1}^{m} \mathrm{Bin}\left(k^-_j \;\middle|\; \pi^{-(\tau)}_j\right), \tag{4.4.19}$$

where $\pi^{+(\tau)}_j$ and $\pi^{-(\tau)}_j$ represent independent samples from the prior distribution over subject-specific accuracies. They can be obtained using ancestral sampling, starting from the prior over α and β, as given in (4.3.6).

In the case of the bivariate normal-binomial model M_{nb}, the model evidence no longer sums over model partitions as in (4.4.17), and so the approximation is derived differently,

$$\ln p(k^+_{1:m}, k^-_{1:m} \mid M_{nb}) \tag{4.4.20}$$

$$= \ln \int p\left(k^+_{1:m}, k^-_{1:m} \;\middle|\; \rho_{1:m}\right)\, p\left(\rho_{1:m} \mid M_{nb}\right)\, \mathrm{d}\rho_{1:m} \tag{4.4.21}$$

$$\approx \ln \frac{1}{c} \sum_{\tau=1}^{c} \prod_{j=1}^{m} \mathrm{Bin}\left(k^+_j \;\middle|\; \sigma\left(\rho^{(\tau,1)}_j\right)\right)\, \mathrm{Bin}\left(k^-_j \;\middle|\; \sigma\left(\rho^{(\tau,2)}_j\right)\right), \tag{4.4.22}$$

for which we provide additional details in on p. 226 in Appendix B.1. Having computed the model evidences, one can proceed to Bayesian model selection (BMS) by evaluating the log Bayes factor,

$$\ln \mathrm{BF}_{bb,nb} = \ln p(k^+_{1:m}, k^-_{1:m} \mid M_{bb}) - \ln p(k^+_{1:m}, k^-_{1:m} \mid M_{nb}), \tag{4.4.23}$$

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

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

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

$$p\left(\phi \mid k^+_{1:m}, k^-_{1:m}\right) \tag{4.4.24}$$

$$= \sum_M p\left(\phi \mid k^+_{1:m}, k^-_{1:m}, M\right)\ p\left(M \mid k^+_{1:m}, k^-_{1:m}\right). \tag{4.4.25}$$

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

$$p\left(\tilde{\phi} \mid k^+_{1:m}, k^-_{1:m}\right) \tag{4.4.26}$$

$$= \sum_M p\left(\tilde{\phi} \mid k^+_{1:m}, k^-_{1:m}, M\right)\ p\left(M \mid k^+_{1:m}, k^-_{1:m}\right). \tag{4.4.27}$$

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

$$\begin{aligned}\mathcal{F} = &\left\langle \ln p\left(k_{1:m} \mid \alpha, \beta, \pi_{1:m}\right)\right\rangle_q \\ &- \mathrm{KL}\left[q(\alpha, \beta, \pi_{1:m}) \parallel p(\alpha, \beta, \pi_{1:m})\right].\end{aligned} \tag{4.4.28}$$

The first term is the log-likelihood of the data expected under the approximate posterior $q(\alpha, \beta, \pi_{1:m})$; it represents the goodness of fit (or accuracy) of the model. The second term is the Kullback-Leibler divergence between

the approximate posterior and the prior; it represents the complexity of the model. This complexity term increases with the number of parameters, their inverse prior covariances, and with the deviation of the posterior from the prior that is necessary to fit the data. Thus, the free-energy approximation shows that the model evidence incorporates a trade-off between explaining the observed data (i.e., goodness of fit) and remaining consistent with our prior (i.e., simplicity or negative complexity). In other words, the model evidence encodes how well a model strikes the balance between explaining the data and remaining simple (Pitt and Myung, 2002; Beal, 2003; Stephan *et al.*, 2009a).

4.4.6 Applications

Following the initial applications in Section 4.3.3, we now turn to simulations that contrast inference on accuracies with inference on balanced accuracies. Applications to empirical data will be deferred until Section 4.7 after having introduced a variational Bayes approximation.

Inference on the balanced accuracy

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

An initial simulation specified a high population accuracy on the positive class and a low accuracy on the negative class, with equal variance in both (Figure 4.10a,b). These accuracies were chosen such that the classifier would perform at chance on a hypothetical balanced sample. This allowed us to mimic the commonly observed situation in which a classifier takes advantage of the imbalance in the data and preferably predicts the majority class.

We independently inverted three competing models: (i) the beta-binomial model to infer on the classification accuracy; and the (ii) twofold beta-binomial and (iii) bivariate normal-binomial models to infer on the balanced accuracy. As expected, the beta-binomial model falsely suggested high evidence for above-chance classification. In contrast, the twofold beta-binomial and normal-binomial models correctly indicated the absence of a statistical

relation between data and class labels (Figure 4.10c).

These characteristics were confirmed across a large set of simulations. As expected, inference on the accuracy falsely concluded above-chance performance, especially in the presence of a significant degree of class imbalance. By contrast, inference on the balanced accuracy did not incorrectly reject the hypothesis of the classifier operating at the level of chance more often than prescribed by the test size (Figure 4.10d).

We compared the two models for inference on the balanced accuracy by means of Bayesian model comparison. Using 10^6 samples with Eqn. (4.4.23), we obtained a log Bayes factor of 33.2 in favour of the twofold beta-binomial model (i.e., under a flat prior over models, the posterior belief in the beta-binomial model is greater than 99.99%). This result represents very strong evidence (Kass and Raftery, 1995) that the beta-binomial model provided a better explanation of the synthetic classification outcomes than the normal-binomial model. This finding is plausible since no correlation structure among class-specific accuracies was imposed in the simulation; thus, in this case, the beta-binomial model is a better model than the more complex normal-binomial model.

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

We repeated the main analysis above 1 000 times and plotted the fraction of above-chance conclusions against the degree of class imbalance. Note that the resulting curve is not a power curve in the traditional sense, as its independent variable is not the true (balanced) accuracy but the accuracy on positive trials, i.e., an indicator of the degree of class imbalance. Figure 4.10d shows that the simple beta-binomial model provides progressively misleading conclusions with class imbalance at the group level (cf. Figure 4.5). In contrast, both schemes for inference on the balanced accuracy made above-chance conclusions in less than 5% of the simulations, as intended by their test size.

All models considered in this chapter are based on diffuse priors designed in such a way that posterior inferences are clearly dominated by the data. However, one might ask to what extent such inferences depend on the exact form of the prior. To examine this dependence, we carried out a sensitivity analysis in which we considered the infraliminal probability of the posterior

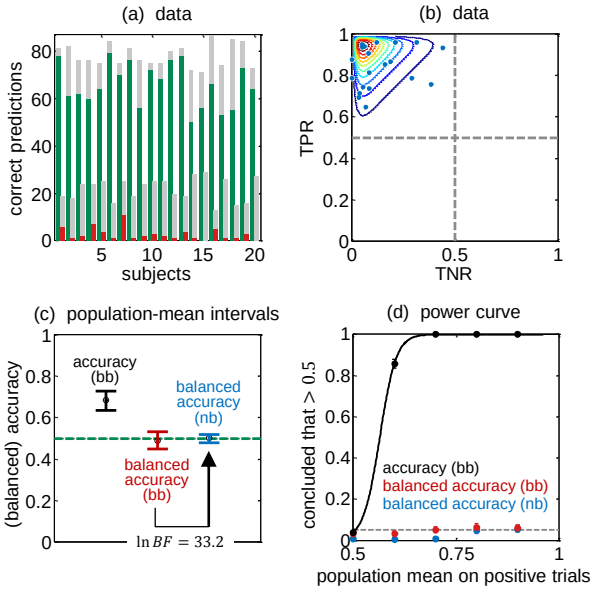

Figure 4.10: Inference on the balanced accuracy. (a) An imbalanced dataset which has led to a classification bias in favour of the majority class. The plot shows, for each subject, the number of correctly classified positive (green) and negative (red) trials and the respective total number of trials (grey). (b) Class-specific sample accuracies, with the true positive rate on the y-axis and the true negative rate on the x-axis. The underlying true population distribution is represented by a bivariate Gaussian kernel density estimate (contour lines), showing the bias of the classifier. (c) Central 95% posterior probability intervals based on the simple beta-binomial model for inference on the population accuracy as well as the twofold beta-binomial model and the bivariate normal-binomial model for inference on the balanced accuracy. The true mean balanced accuracy in the population is at chance (green). It is accurately estimated by models inferring on the balanced accuracy (red, blue). Bayesian model selection yielded very strong evidence in favour of the normal-binomial model (posterior model probability = 97.7%). (d) Probability of falsely detecting above-chance performance, using different inference schemes. The true balanced accuracy is 0.5. The x-axis represents the degree of class imbalance.

population mean as a function of prior moments (Figure 4.11).

We found that inferences were extremely robust, in the sense that the influence of the prior moments on the resulting posterior densities was negligible in relation to the variance resulting from the fact that we are using a (stochastic) approximate inference method for model inversion. In particu-

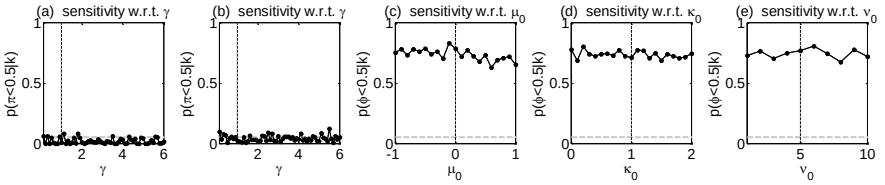

Figure 4.11: Sensitivity analysis. This figure illustrates the dependence of posterior inferences on the exact form of the priors proposed in this chapter. Each graph shows the infraliminal probability of the population mean accuracy (i.e., the posterior probability mass below 0.5) as a function of a particular parameter of the prior distribution used for model inversion. (a,b) Same datasets as shown those shown in Figures 4.4a and 4.6a, but with a slightly lower population mean of 0.7. Inferences on the population accuracy are based on the beta-binomial model. (c,d,e) Same dataset as shown in Figure 4.10a. Inferences on the population balanced accuracy are based on the bivariate normal-binomial model.

lar, varying the constant (originally: 1) in Eqn. (4.3.6) for the beta-binomial prior left the infraliminal probability of the posterior accuracy unaffected (Figure 4.11a,b). Similarly, varying μ_0, κ_0, or ν_0 in the normal-binomial model had practically no influence on the infraliminal probability of the posterior balanced accuracy (Figure 4.11c,d,e).

Application to synthetic data

All experiments described so far were based on classification outcomes sampled from the beta-binomial or normal-binomial model. This ensured, by construction, that the distributional assumptions underlying the models were fulfilled. To illustrate the generic applicability of our approach when its assumptions are not satisfied by construction, we applied models for mixed-effects inference to classification outcomes obtained on synthetic data features for a group of 20 subjects with 100 trials each (Figure 4.12). In addition to probing the models' robustness with regard to distributional assumptions, this allows one to examine what correlations between class-specific accuracies may be observed in practice.

Synthetic data were generated using a two-level sampling approach that reflected the hierarchical nature of group studies. We specified a population distribution, sampled subject-specific means and variances from it, and then used these to generate trial-specific feature vectors.

In a first simulation (Figure 4.12, top row), we generated 50 positive trials and 50 negative trials for each individual subject j by drawing one-

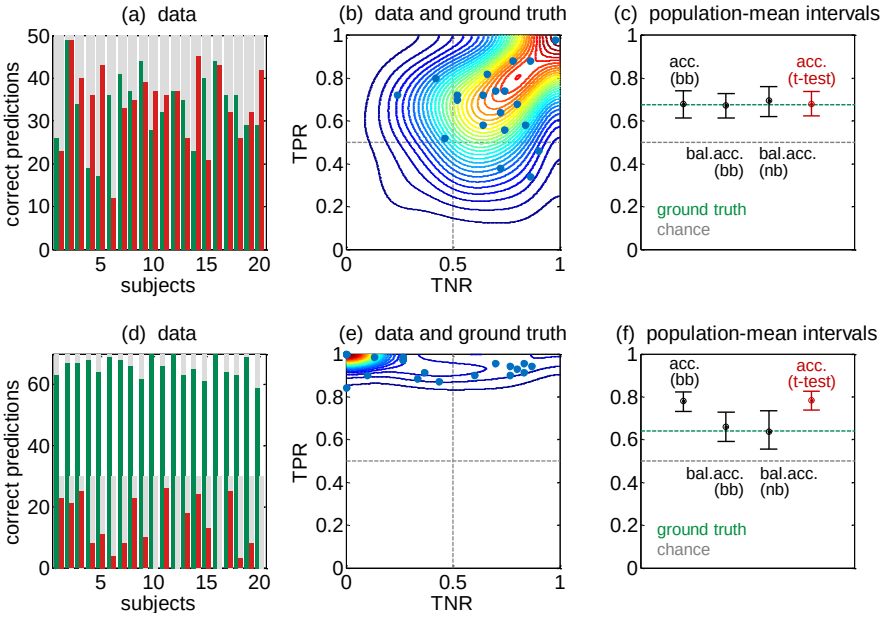

Figure 4.12: Application to synthetic data. (a) Classification outcomes obtained by applying a linear support vector machine to synthetic data, using leave-one-out cross-validation. (b) Sample accuracies on positive (TPR) and negative classes (TNR) show the positive correlation between class-specific accuracies (blue). The underlying population distribution is represented by a bivariate Gaussian kernel density estimate (contour lines). (c) Central 95% posterior probability intervals, resulting from inversion of the beta-binomial model for inference on the population mean accuracy as well as the twofold beta-binomial model (*bb*) and the bivariate normal-binomial model (*nb*) for inference on the population mean balanced accuracy (all black). A frequentist 95% confidence interval (red) is shown for comparison. Bayesian model selection yielded very strong evidence (Kass and Raftery, 1995) in favour of the normal-binomial model (posterior model probability = 99.99%). (d) A second simulation was based on a synthetic heterogeneous group with varying numbers of trials. (e) In this case, the classifier acquires a strong bias in favour of the majority class. (f) As a result, inference on the accuracy is misleading; the balanced accuracy is a much better performance indicator, whether based on the beta-binomial (*bb*) or normal-binomial model (*nb*).

dimensional feature vectors from two normal distributions, $\mathcal{N}(x_{ij} \mid \mu_j^+, \sigma_j)$ and $\mathcal{N}(x_{ij} \mid \mu_j^-, \sigma_j)$, respectively. The moments of these subject-specific distributions, in turn, were drawn from a population distribution, using $\mathcal{N}(\mu_j^+ \mid \frac{1}{2}, \frac{1}{2})$ and $\mu_j^- = -\mu_j^+$ for the means, and $\mathrm{Ga}^{-1}(\sigma_j \mid 10, \frac{1}{10})$ for the

variance. The normal distribution and the inverse Gamma distribution are conjugate priors for the mean and variance of a univariate normal distribution and, thus, represent natural choices for the population distribution.

To obtain classification outcomes, separately for each subject, we trained and tested a linear support vector machine (SVM; Chang and Lin, 2011), using leave-one-trial-out cross-validation. Classification outcomes are shown in Figure 4.12a, in which the numbers of correctly classified trials are illustrated separately for the two classes and for each subject. The same data are represented in terms of sample accuracies in Figure 4.12b (blue dots). To illustrate ground truth, we repeated the above procedure (of generating synthetic data and applying an SVM) 1 000 times and added a contour plot of the resulting distribution of sample accuracies in the same figure. This distribution was symmetric with regard to class-specific accuracies while these accuracies themselves were strongly positively correlated, as one would expect from a linear classifier tested on perfectly balanced datasets.

We applied all three models discussed in this chapter for inference: the beta-binomial model for inference on the accuracy (Section 4.3.1), and the twofold beta-binomial and normal-binomial model for inference on the balanced accuracy (Sections 4.4.1 and 4.4.3). Central 95% posterior probability intervals about the population mean are shown in Figure 4.12c, along with a frequentist confidence interval of the population mean accuracy. All four approaches provided precise intervals around the true population mean. Comparing the two competing models for inference on the balanced accuracy, we obtained a log Bayes factor of 22.7 in favour of the twofold beta-binomial model (posterior model probability $> 99.99\%$), representing very strong evidence (Kass and Raftery, 1995) that this model provided a better explanation of the data (i.e., a better balance between fit and complexity) than the bivariate normal-binomial model (standard deviation 8.8).

We repeated the above analysis with a subtle but important modification: instead of using perfectly balanced data (50 positive and 50 negative trials), we created imbalanced synthetic data using 70 positive and 30 negative trials per subject. All other details of the analysis remained unchanged (Figure 4.12, bottom row). We observed that, as expected, the class imbalance caused the classifier to acquire a bias in favour of the majority class. This can be seen from the raw classification outcomes in which many more positive trials (green) than negative trials (red) were classified correctly, relative to their respective prevalence in the data (grey; Figure 4.12d). The bias is reflected accordingly by the estimated bivariate density of class-

specific classification accuracies, in which the majority class consistently performs well whereas the accuracy on the minority class varies strongly (Figure 4.12e).

In this setting, we found that both the twofold beta-binomial model and the normal-binomial model provided excellent estimates of the true balanced accuracy (Figure 4.12f; log Bayes factor in favour of the beta-binomial model: 65.3; standard deviation 14.2). In stark contrast, using the single beta-binomial model or a conventional mean of sample accuracies to infer on the population accuracy (as opposed to balanced accuracy) resulted in estimates that were overly optimistic and therefore misleading.

4.4.7 Interim conclusions

Canonical classification algorithms are frequently used on multilevel or hierarchically structured datasets, where a classifier is trained and tested for each subject within a group. The first half of this chapter showed how the evaluation of classification performance may benefit from mixed-effects models that explicitly capture the hierarchical structure of the data.

Results on synthetic data have illustrated the characteristic features of our approach: (i) posterior densities as opposed to point estimates of parameters; (ii) the ability to compare alternative (even non-nested) models; (iii) the 'shrinking-to-the-population' effect that regularizes estimates of classification performance in individual subjects (Figure 4.7b); (iv) increased sensitivity (Figure 4.7c); (v) more precise parameter estimates (Figure 4.7d); and (vi) avoidance of classifier bias for imbalanced datasets using the balanced accuracy (Figure 4.10).

An important practical limitation of our approach lies in the high computational complexity of our current inversion methods. In particular, our MCMC algorithms lack guarantees about convergence rates. Our algorithms also include heuristics regarding the number of burn-in samples, the precision of the overdispersed initial distributions and the proposal densities, and regarding the number of chains run in parallel.

Some of these issues can be addressed using a Bayesian reformulation of MCMC that allows one, for instance, to take into account prior knowledge about the smoothness of the integrand (Rasmussen and Ghahramani, 2003). The problem of high computational complexity itself, however, can be best addressed by replacing stochastic inference by algorithms for *deterministic* approximate inference, as proposed in the remaining part of this chapter.

4.5 Variational Bayesian inference on the accuracy

"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

In the first half of this chapter, we proposed a solution to the limitations of conventional methods for assessing classification performance. Our approach was based on hierarchical models that naturally enabled mixed-effects inference in a fully Bayesian framework. The practical utility of these models, however, was limited by the high computational complexity of the underlying MCMC sampling algorithms for model estimation. 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 (Nandy and Cordes, 2003; Kriegeskorte *et al.*, 2006).

To overcome this limitation, we will devote the second part of this chapter to the development of a variational Bayes (VB) algorithm. Our approach is characterized by three features. First, the model described below is a mixed-effects model; it explicitly respects the hierarchical structure of the data by simultaneously accounting for fixed-effects and random-effects variance components. Second, the model can be equally used for inference on the accuracy and the balanced accuracy, which is a better performance indicator when the data are not perfectly balanced. Third, our variational approximate inference scheme dramatically reduces the computational complexity compared to sampling approaches.

4.5.1 The univariate normal-binomial model

As motivated in Section 4.3, we begin by modelling the number of correctly classified trials k_j in subject j as

$$p(k_j \mid \pi_j, n_j) = \mathrm{Bin}\left(k_j \mid \pi_j, n_j\right), \tag{4.5.1}$$

where π_j represents the latent classification accuracy in subject j.[9] In contrast to previous models, at the group level, we shall assume accuracies to

[9] We will omit n_j unless this introduces ambiguity.

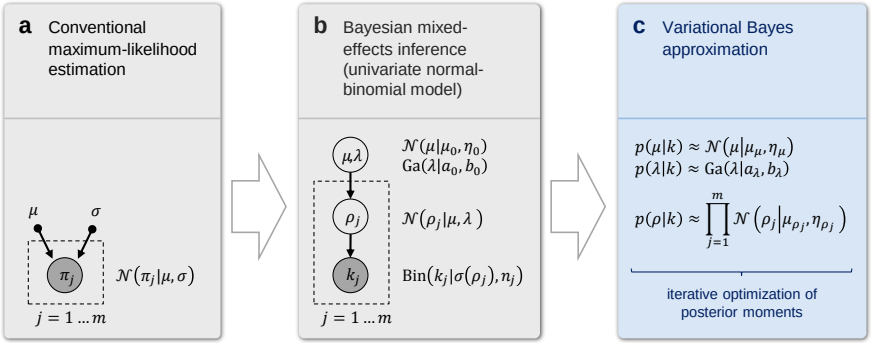

Figure 4.13: Inference on classification accuracies. (a) Conventional maximum-likelihood estimation does not explicitly model within-subjects (fixed-effects) variance components and is based on an ill-justified normality assumption. It is therefore inadequate for the statistical evaluation of classification 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. (c) Model inversion can be implemented efficiently using a variational approximation to the posterior densities of the model parameters (see Figure 4.14 for details).

be *logit-normally* distributed. In other words, each logit accuracy

$$\rho_j := \sigma^{-1}(\pi_j) := \ln \frac{\pi_j}{1-\pi_j} \tag{4.5.2}$$

is drawn from a normal distribution,

$$p(\rho_j \mid \mu, \lambda) = \mathcal{N}(\rho_j \mid \mu, \lambda) \tag{4.5.3}$$

$$= \frac{\lambda}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\lambda(\rho_j - \mu)^2\right), \tag{4.5.4}$$

where μ and λ represent the population mean and the population precision (i.e., inverse variance), respectively. 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 $(-\infty, +\infty)$.

Since neuroimaging studies are typically confined to small sample sizes, expressing prior ignorance about the population parameters is critical. 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 Section 4.7). A natural

parameterization is to use independent conjugate densities:

$$p(\mu \mid \mu_0, \lambda_0) = \mathcal{N}(\mu \mid \mu_0, \lambda_0) \tag{4.5.5}$$
$$p(\lambda \mid a_0, b_0) = \mathrm{Ga}(\lambda \mid a_0, b_0) \tag{4.5.6}$$

In the above densities, μ_0 and λ_0 encode the prior mean and precision of the population, and a_0 and b_0 represent the shape and scale parameters that specify the prior distribution of the population precision.[10] 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 (Figure 4.13b).[11]

In principle, given classification outcomes $k \equiv k_{1:m} \equiv (k_1, \ldots, k_m)$, inverting the above model immediately yields the desired posterior density over parameters,

$$p(\mu, \lambda, \rho \mid k) \tag{4.5.7}$$

$$= \frac{\prod_{j=1}^{m} \mathrm{Bin}(k_j|\sigma(\rho_j))\,\mathcal{N}(\rho_j|\mu,\lambda)\,\mathcal{N}(\mu|\mu_0,\lambda_0)\,\mathrm{Ga}(\lambda|a_0,b_0)}{\int\!\!\int\!\cdots\!\int \prod_{j=1}^{m} \mathrm{Bin}(k_j|\sigma(\rho_j))\,\mathcal{N}(\rho_j|\mu,\lambda)\,\mathcal{N}(\mu|\mu_0,\lambda_0)\,\mathrm{Ga}(\lambda|a_0,b_0)\,\mathrm{d}\rho_1 \cdots \mathrm{d}\rho_m\,\mathrm{d}\mu\,\mathrm{d}\lambda}.$$

In practice, however, computing the integral 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 practicality of these algorithms was limited by their considerable computational complexity. Here, we propose to invert the above model using a deterministic VB approximation (Figure 4.13c). This approximation is no longer asymptotically exact, but it conveys considerable computational advantages. The remainder of this section describes its derivation (see Figure 4.14 for a summary).

[10] For an alternative parameterization, see Leonard (1972).

[11] The *univariate* normal-binomial model described here is formally related to the previously proposed bivariate normal-binomial model for inference on the balanced accuracy (Section 4.4.3). The two models differ in the number of observed variables per subject; their differences are unrelated to the distinction between (mass-)univariate and multivariate analyses.

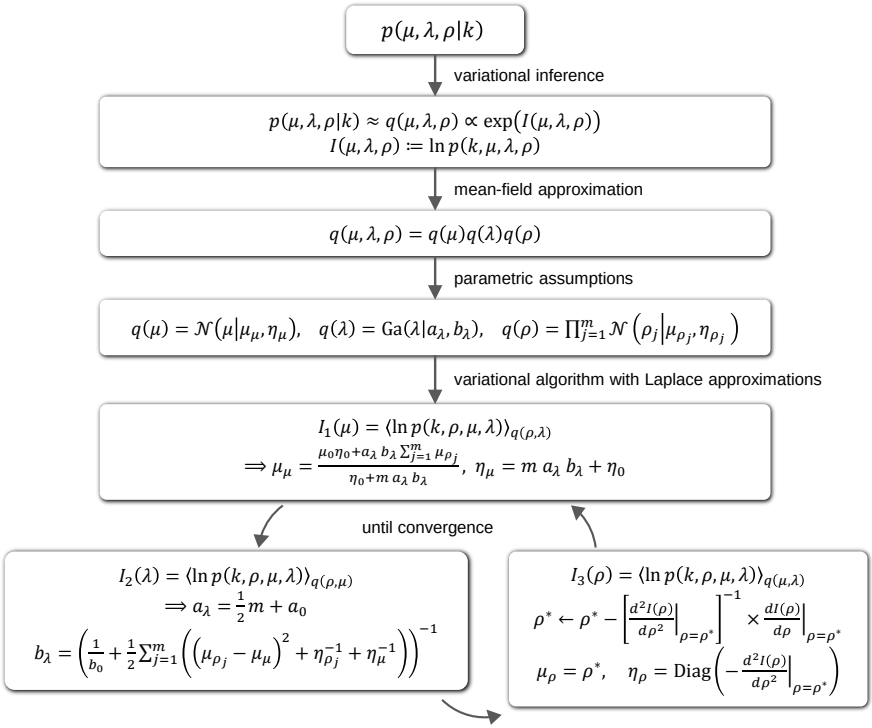

Figure 4.14: Variational inversion of the univariate normal-binomial model. Variational inference translates a difficult integration problem into an easier optimization problem. This schematic summarizes the individual steps involved in a variational approach to the inversion of the univariate normal-binomial model, as detailed in the main text.

4.5.2 Variational inference

The difficult problem of finding the exact posterior $p(\mu, \lambda, \rho \mid k)$ can be transformed into the easier problem of finding an approximate parametric posterior $q(\mu, \lambda, \rho \mid \delta)$ with moments (i.e., parameters) δ (which we will omit to simplify the notation). Inference then reduces to finding a density q that minimizes a measure of dissimilarity between q and p. This comparison can be achieved by maximizing the so-called *negative free-energy* of the model under the assumption of q as the posterior, given the data, by varying the moments of q. This is equivalent to minimizing the Kullback-Leibler

divergence (KL) between the approximate and true posteriors, q and p:

$$\mathrm{KL}[q \parallel p] := \iiint q(\mu, \lambda, \rho) \ln \frac{q(\mu, \lambda, \rho)}{p(\mu, \lambda, \rho \mid k)} \, \mathrm{d}\mu \, \mathrm{d}\lambda \, \mathrm{d}\rho \tag{4.5.8}$$

$$= \iiint q(\mu, \lambda, \rho) \ln \frac{q(\mu, \lambda, \rho)}{p(\mu, \lambda, \rho \mid k) \, p(k)} \mathrm{d}\mu \, \mathrm{d}\lambda \, \mathrm{d}\rho + \ln p(k) \tag{4.5.9}$$

$$\implies \ln p(k) = \mathrm{KL}[q \parallel p] \underbrace{- \left\langle \ln \frac{q(\mu, \lambda, \rho)}{p(k, \mu, \lambda, \rho)} \right\rangle_{q(\mu,\lambda,\rho)}}_{=: \mathcal{F}(q,k)} \tag{4.5.10}$$

Thus, the log model evidence $\ln p(k)$ is the sum of (i) the KL-divergence between the approximate and the true posterior and (ii) the negative free-energy $\mathcal{F}(q, k)$. 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. In addition, maximizing the negative free-energy means tightening the lower bound to the log model evidence which enables Bayesian model comparison.

In summary, maximizing the negative free-energy in (4.5.10) enables both inference on the posterior density over parameters and model comparison. Here, we are primarily interested in the posterior density.

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

$$\frac{\partial F(q, k)}{\partial q} = 0 \quad \implies \quad q(\mu, \lambda, \rho) \propto \exp\left(I(\mu, \lambda, \rho)\right). \tag{4.5.11}$$

The above says that the approximate posterior which maximizes the negative free-energy is proportional to the exponential of the negative variational energy, which is itself defined as

$$I(\mu, \lambda, \rho) := \ln p(k, \mu, \lambda, \rho). \tag{4.5.12}$$

Mean-field approximation

Rather than working with the negative variational energy in (4.5.12) itself, it is beneficial to assume that the joint posterior over model parameters factorizes into specific parts. Using one density for each variable constitutes the mean-field assumption

$$q(\mu, \lambda, \rho) = q(\mu) \, q(\lambda) \, q(\rho), \tag{4.5.13}$$

which turns the problem of maximizing $\mathcal{F}(q, k)$ into the problem of deriving three log expectations:

$$I_1(\mu) = \ln \langle p(k, \mu, \lambda, \rho) \rangle_{q(\lambda,\rho)} \tag{4.5.14}$$

$$I_2(\lambda) = \ln \langle p(k, \mu, \lambda, \rho) \rangle_{q(\mu,\rho)} \tag{4.5.15}$$

$$I_3(\rho) = \ln \langle p(k, \mu, \lambda, \rho) \rangle_{q(\mu,\lambda)} \tag{4.5.16}$$

Invoking a mean-field approximation in this way has several advantages over working with (4.5.12) directly: (i) 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); (ii) 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 (iii) it may provide us with interpretable update equations (as will be the case 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:

$$q(\mu) = \mathcal{N}(\mu \mid \mu_\mu, \eta_\mu) \tag{4.5.17}$$

$$q(\lambda) = \mathrm{Ga}(\lambda \mid a_\lambda, b_\lambda) \tag{4.5.18}$$

Subject-specific (logit) accuracies $\rho \equiv (\rho_1, \ldots, \rho_m)$ are also conditionally independent given the data. However, we require a distributional assumption for their posteriors to make model inversion feasible. Specifically, we assume posterior subject-specific (logit) accuracies to be normally distributed:

$$q(\rho) = \prod_{j=1}^{m} \mathcal{N}\left(\rho_j \mid \mu_{\rho_j}, \eta_{\rho_j}\right) \tag{4.5.19}$$

It should be noted that the above conditional independence is not an additional assumption but 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

$$q(\mu, \lambda, \rho) = q(\mu)\, q(\lambda)\, q(\rho) \tag{4.5.20}$$

$$= q(\mu)\, q(\lambda) \prod_{j=1}^{m} q(\rho_j). \tag{4.5.21}$$

The conditional independence in (4.5.19) 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. 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 (4.5.19) are informed by (or are borrowing strength from) observations from the entire group (by means of their effect on the population parameters). As will be seen in Section 4.7, the ensuing shrinkage effect is crucial for obtaining precise subject-specific estimates.

Derivation of variational densities

For each mean-field part in (4.5.13), the variational density $q(\cdot)$ can be obtained by evaluating the variational energy $I(\cdot)$, as described next.

First variational energy. The first variational energy concerns the posterior density over the population mean μ. It is given by

$$I_1(\mu) = \langle \ln p(k, \mu, \lambda, \rho) \rangle_{q(\lambda,\rho)} \tag{4.5.22}$$

$$= \mu\, a_\lambda b_\lambda \left(-\frac{1}{2} m\mu + \sum_{j=1}^{m} \mu_{\rho_j} \right) + \mu\, \eta_0 \left(\mu_0 - \frac{1}{2}\mu \right). \tag{4.5.23}$$

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

$$\frac{\mathrm{d}I(\mu)}{\mathrm{d}\mu} = -\mu(m\, a_\lambda b_\lambda + \eta_0) + \mu_0\eta_0 + a_\lambda b_\lambda \sum_{j=1}^{m} \mu_{\rho_j} = 0 \tag{4.5.24}$$

$$\Longrightarrow \mu^* = \frac{\mu_0\eta_0 + a_\lambda b_\lambda \sum_{j=1}^{m} \mu_{\rho_j}}{\eta_0 + m\, a_\lambda b_\lambda} \tag{4.5.25}$$

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

$$\mu_\mu = \mu^* \quad \text{and} \tag{4.5.26}$$

$$\eta_\mu = -\left.\frac{\mathrm{d}I^2(\mu)}{\mathrm{d}\mu^2}\right|_{\mu=\mu^*} = m\, a_\lambda b_\lambda + \eta_0 \tag{4.5.27}$$

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

As can be seen above, the approach we adopt here does not optimize all sufficient statistics of the approximate posterior. Instead, we only optimize the mean, while enforcing the variance to equate the inverse curvature at the mean. This procedure is a Laplace 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 approximation rather than a global optimization.

As long as the distributional family of the approximate density is similar to the true posterior, the Laplace approximation confers two advantages: (i) it is computationally efficient (see discussion in Section 4.8); and (ii) it typically gives rise to interpretable update equations. In the case of the first variational energy, for example, one 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.

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:

$$\operatorname{logit}\mathcal{N}\left(\xi \mid \mu_\mu, \eta_\mu\right) \tag{4.5.28}$$

$$= \frac{1}{\xi(1-\xi)}\sqrt{\frac{\eta_\mu}{2\pi}}\exp\left(-\frac{\eta_\mu}{2}\left(\sigma^{-1}(\xi) - \mu_\mu\right)^2\right) \tag{4.5.29}$$

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

$$I_2(\lambda) = \langle \ln p(k, \mu, \lambda, \rho) \rangle_{q(\mu,\rho)} \tag{4.5.30}$$

$$= \frac{m}{2} \ln \lambda - \frac{\lambda}{2} \sum_{j=1}^{m} \left(\left(\mu_{\rho_j} - \mu_\mu \right)^2 + \eta_{\rho_j}^{-1} + \eta_\mu^{-1} \right) + (a_0 - 1) \ln \lambda - \frac{\lambda}{b_0} + \text{const.} \tag{4.5.31}$$

The above expression already has the form of a log-Gamma distribution with parameters

$$a_\lambda = \frac{1}{2} m + a_0 \quad \text{and} \tag{4.5.32}$$

$$b_\lambda = \left(\frac{1}{2} \sum_{j=1}^{m} \left(\left(\mu_{\rho_j} - \mu_\mu \right)^2 + \eta_{\rho_j}^{-1} + \eta_\mu^{-1} \right) + \frac{1}{b_0} \right)^{-1}. \tag{4.5.33}$$

From this we can see that the posterior shape a_λ is a weighted sum between prior shape a_0 and data m. When viewing the second parameter as a 'rate' coefficient b_λ^{-1} (rather than as a shape coefficient b_λ), we can furthermore see 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}).

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

$$I_3(\rho) = \langle \ln p(k, \mu, \lambda, \rho) \rangle_{q(\mu,\lambda)} \tag{4.5.34}$$

$$= \sum_{j=1}^{m} k_j \ln \sigma(\rho_j) + (n_j - k_j) \ln \left(1 - \sigma(\rho_j)\right) - \frac{1}{2} a_\lambda b_\lambda \left(\rho_j - \mu_\mu \right)^2 + \text{const.} \tag{4.5.35}$$

Since an analytical expression for the maximum of this energy does not exist, we resort to an iterative Gauss-Newton (GN) scheme. For this, we

begin by considering the Jacobian

$$\left(\frac{\mathrm{d}I(\rho)}{\mathrm{d}\rho}\right)_j = \frac{\partial \mathrm{I}(\rho)}{\partial \rho_j} \tag{4.5.36}$$

$$= k_j - n_j\sigma(\rho_j) + a_\lambda b_\lambda(\mu_\mu - \rho) = 0 \tag{4.5.37}$$

and the Hessian

$$\left(\frac{\mathrm{d}^2I(\rho)}{\mathrm{d}\rho^2}\right)_{jk} = \frac{\partial^2 I(\rho)}{\partial\rho_j\partial\rho_k} \tag{4.5.38}$$

$$= -\delta_{jk}\left[n_j\sigma\left(\rho_j\right)\left(1-\sigma\left(\rho_j\right)\right) + a_\lambda b_\lambda\right], \tag{4.5.39}$$

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 (4.5.20). Each GN iteration performs the update

$$\rho \leftarrow \rho^* - \left[\left.\frac{\mathrm{d}^2I(\rho)}{\mathrm{d}\rho^2}\right|_{\rho=\rho^*}\right]^{-1} \times \left.\frac{\mathrm{d}I(\rho)}{\mathrm{d}\rho}\right|_{\rho=\rho^*} \tag{4.5.40}$$

until the vector ρ^* converges. Using this maximum, we can use a second-order Taylor expansion (i.e., the Laplace approximation) to set the moments of the approximate posterior:

$$\mu_\rho = \rho^* \quad \text{and} \tag{4.5.41}$$

$$\eta_\rho = -\left.\frac{\mathrm{d}^2I(\rho)}{\mathrm{d}\rho^2}\right|_{\rho=\rho^*} \tag{4.5.42}$$

VB algorithm

The expressions for the three variational energies depend on one another. This circularity can be resolved by looping over each expression in turn and updating the moments of the current approximate marginal given the current moments of the other marginals. This procedure maximizes the negative free-energy and leads to approximate marginals that are maximally similar to the exact marginals (within the boundaries of their parametric form).

It is worth noting that the algorithm does not strictly increase a lower bound to the negative free-energy on each iteration. This is due to the parametric assumptions, in particular the Laplace approximation which only retains terms up to the second order. Thus, the algorithm evolves an *approximation* to a lower bound to the log model evidence.

The algorithm terminates when the moments of all approximate posteriors have converged.

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 MCMC approach that is computationally much more expensive than variational Bayes but exact in the limit of infinite runtime.

The Gibbs sampler used here is described in detail in Appendix C.1. It is similar in structure to the algorithms in Appendices A.1 and B.1, but based on the new distributional assumptions specific to the univariate normal-binomial model. The algorithm 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. Importantly, unlike VB, which was based on a mean-field assumption, the posterior obtained through MCMC retains any conditional dependencies among the model parameters.

4.6 Variational Bayesian inference on the balanced accuracy

The twofold normal-binomial model. The 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 (Sections 4.4.1 and 4.4.3. Here, we adopt the approach of inferring 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 (Figure 4.15).

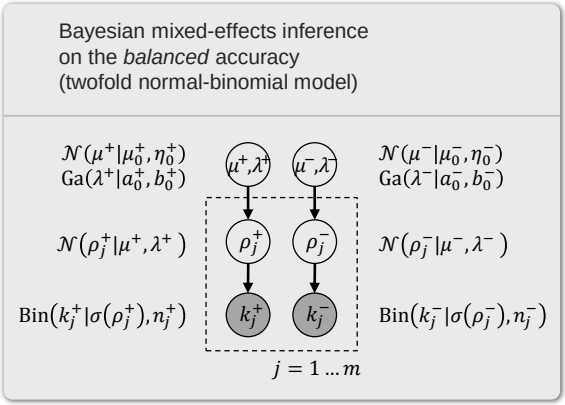

Figure 4.15: Inference on balanced accuracies. The univariate normal-binomial model (Figure 4.13b) can be easily extended to enable inference on the balanced accuracy. Specifically, the model is inverted separately for classification outcomes obtained on positive and negative trials. The resulting posteriors are then recombined (see main text).

Inference. 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 of accuracies on positive and negative trials, respectively.

Inverting the model proceeds by inverting its two separate parts independently. But in contrast to our previous treatment, 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,

$$p\left(\phi \mid k^+, k^-\right) = p\left(\frac{1}{2}\left(\sigma\left(\mu^+\right) + \sigma\left(\mu^-\right)\right) \,\middle|\, k^+, k^-\right). \tag{4.6.1}$$

Unlike the population mean accuracy (4.5.26), 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.

For this, we use a convolution analogous to Eqn. (3.2.6) on p. 53. Thus, we obtain the posterior distribution of the balanced accuracy as

$$p\left(\phi \mid k^+, k^-\right) = \int_0^{2\phi} p_{\sigma(\mu^+)}\left(2\phi - z \mid k^+\right)\, p_{\sigma(\mu^-)}\left(z \mid k^-\right) \mathrm{d}z, \qquad (4.6.2)$$

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

4.7 Applications

This section illustrates the sort of inferences that can be made using VB in a classification study. 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.

4.7.1 Application to simulated 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 (Figures 4.16 and 4.17).

The first synthetic dataset 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 $\mu = 1.1$ and a relatively high logit population precision of $\lambda = 4$ (Figure 4.16a). The corresponding population mean accuracy was 71%.

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 (Figure 4.16b). In practice, a good way of summarizing the posterior is to report a central 95% posterior probability interval, also sometimes referred to as a Bayesian credible interval. Although this

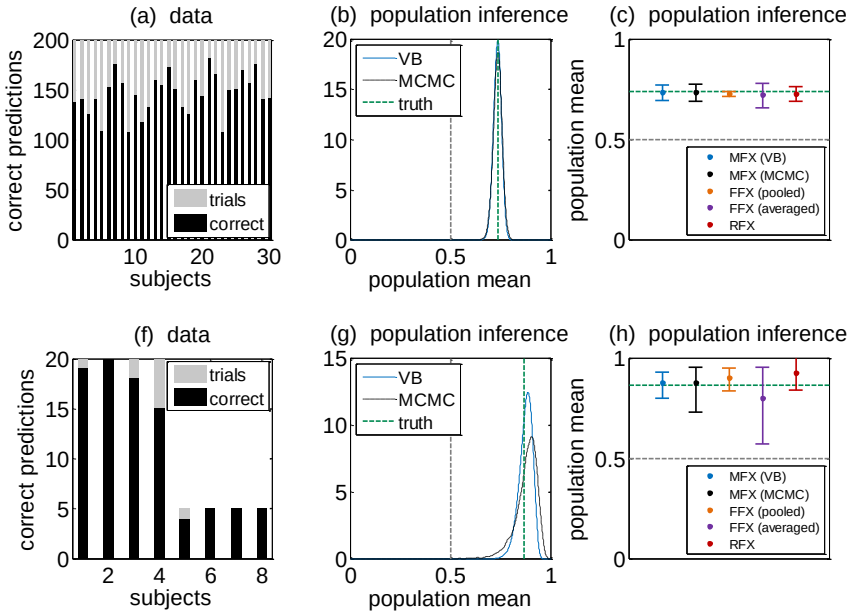

Figure 4.16: Application to simulated data I. Two simple synthetic datasets illustrative the sort of inferences that can be made using a mixed-effects model. (a) Simulated data, showing the number of trials in each subject (grey) and the number of correct predictions (black). (b) Resulting posterior 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. (f–h) 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. *Continued in Figure 4.17.*

interval is conceptually different from a classical (frequentist) 95% confidence interval, in this particular case the two intervals agreed very closely (Figure 4.16c), which is typical in the context of a diffuse prior and 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.

Another informative way of summarizing the posterior population mean is to report the posterior probability mass that is below chance (e.g., 0.5 for

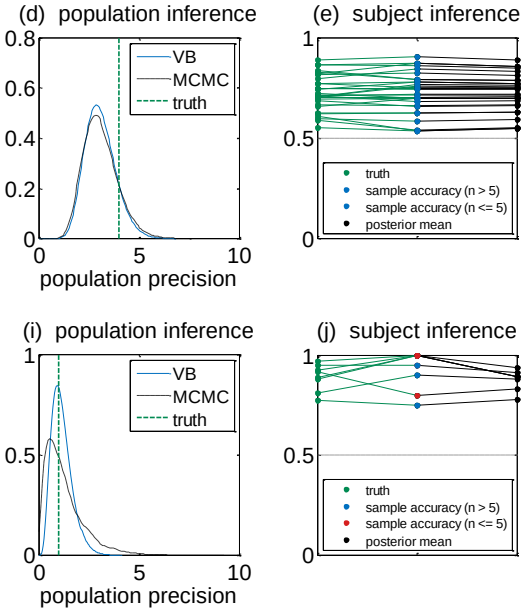

Figure 4.17: Application to simulated data II. *Continued from Figure 4.16.* (d) Posterior densities of the population precision (inverse variance). (e) The benefits of a mixed-effects approach in subject-specific inference can be visualized by contrasting the increase in dispersion (as we move from ground truth to sample accuracies) to 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 accounts for better estimates of ground truth (cf. Figure 4.20e,j). Shrinking may change the order of subjects since its extent depends on the subject-specific (first-level) posterior uncertainty. (i,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.

binary classification) which we refer to as the (posterior) *infraliminal probability* p of the classifier (cf. Section 4.3.2). Compared to a classical p-value, it has a deceptively similar but arguably more natural interpretation. Rather than representing the relative 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 (Figure 4.17e). 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 shrinkage effect is obtained naturally in a hierarchical model and, as we will see below, leads to systematically more accurate posterior inferences at the 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.[12] 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% (Figure 4.16f).

Comparing the resulting posteriors (Figures 4.16g,h and 4.17i,j) to those obtained on the first dataset, several differences are worth noting. Concerning the population parameters (Figures 4.16g and 4.17i), 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 (Figures 4.16b and 4.17d). This can be seen best from the credible intervals (Figure 4.16h, black). By comparison, a striking example of an unreasonable 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% (Figure 4.16h, red).

Another typical consequence of a small dataset with variable trial numbers can be seen in the shrinkage of subject-specific inferences (Figure 4.17j). 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. This attempt to restore the correct order of subjects

[12] Note that the heteroscedasticity in this dataset results both from the fact that subjects have different numbers of trials and from the (simpler) fact that they have different sample accuracies.

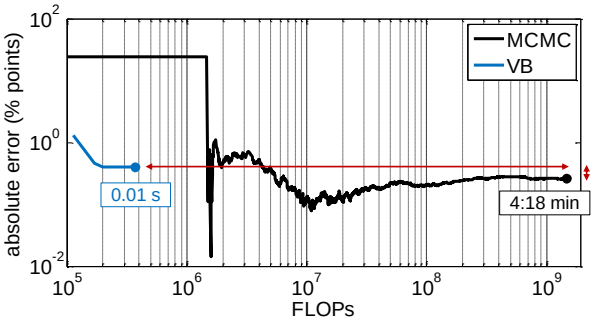

Figure 4.18: Estimation error and computational complexity. VB and MCMC differ in the way estimation error and computational complexity are balanced. The plot shows estimation error in terms of the absolute difference of the posterior mean of the population mean accuracy in percentage points (y-axis). Computational complexity is shown in terms of the number of floating point operations (FLOPs) consumed. VB converged after 370 000 FLOPs (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 (by 0.13 percentage points), VB was computationally more efficient by 4 orders of magnitude. It should be noted that the plot uses log-log axes which provide a conservative view on the differences between the two algorithms; they would be visually even more striking on a linear scale.

can become important, for example, when one wishes to relate subject-specific accuracies to other subject-specific measures, such as behavioural, 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 shown in Figure 4.16a. 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; Figure 4.18). We found that MCMC used 4 000 times more arithmetic operations to achieve an estimate that was better than VB by no more than 0.13 percentage points.

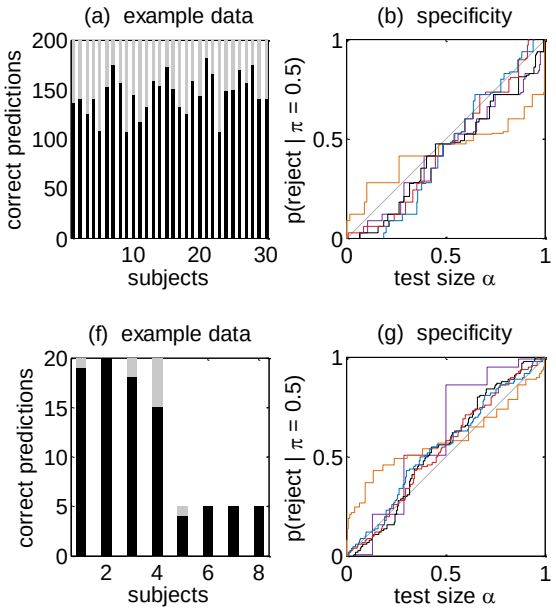

Figure 4.19: Application to a larger number of simulations I. (a) One example of 1 000 simulations of synthetic classification outcomes (for an individual analysis of this example, see Figures 4.16 and 4.16, top row). (b) Specificity of competing methods for testing whether the population mean accuracy is greater than chance, given a true population mean of 0.5. (f,g) Same analyses as above, but based on smaller experiments (cf. Figures 4.16 and 4.17, bottom row). For details, see main text. *Continued in Figure 4.20.*

4.7.2 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 1 000 simulations. This allowed us to examine the properties of our approach from a frequentist perspective (Figures 4.19 and 4.20).

In the first setting (Figures 4.19 and 4.20, top row), each simulation was based on synthetic classification outcomes from 30 subjects with 200 trials each, as described in the previous section. One of these simulations is shown as an example (Figure 4.19a; the same one as in Figure 4.16a), whereas all subsequent plots are based on 1 000 independent datasets generated in the

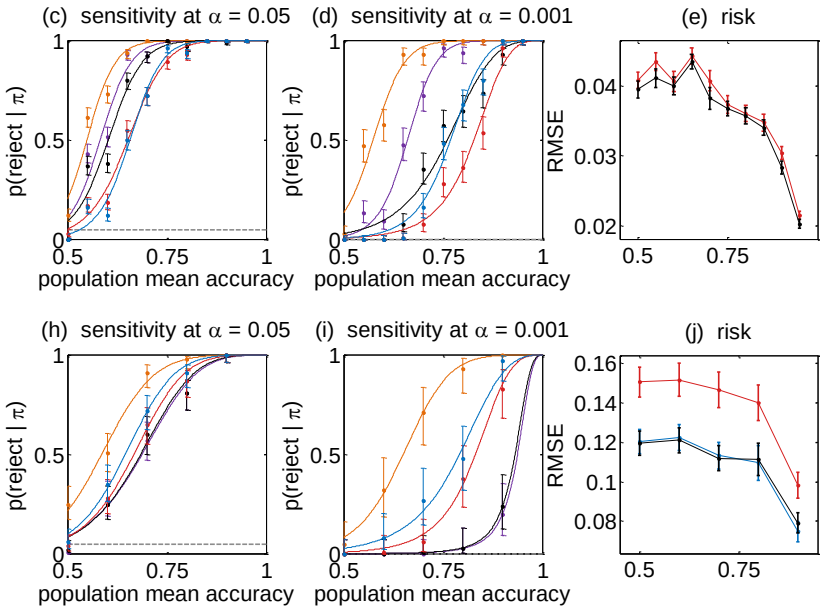

Figure 4.20: Application to a larger number of simulations II. *Continued from Figure 4.19.* (c,d) Power curves with different test sizes, testing whether the population mean accuracy is greater than chance, given different true population mean accuracies. (e) Comparison of maximum-likelihood estimator (blue), James-Stein estimator (red), and Bayes estimator (black) in terms of the mean squared difference between the estimate and ground truth. The three graphs show mean and standard errors across 1 000 simulations. (h–j) Same analyses as above, but based on smaller experiments (cf. Figures 4.16 and 4.17, bottom row). For details, see main text.

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 one of 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 MCMC (dotted black); (v) random-effects inference using a t-test on subject-specific sample accuracies (red).

The principal property of inferential conclusions (whether frequentist or Bayesian) 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 (Figure 4.19b). 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.

As can be seen from the plot, 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 is an invalid procedure for inference on the population mean accuracy.

A second property of inference schemes is *sensitivity* or statistical *power* (Figure 4.20c,d). 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. 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. We carried out 1 000 simulations for each level of true population mean accuracy $(0.5, 0.6, \ldots, 0.9)$ and plotted empirical rejection rates for two common test sizes: $\alpha = 0.05$ (Figure 4.20c); and $\alpha = 0.001$ (Figure 4.20d). The smaller the test size, the more striking the differences between different methods. This is because a small test size implies that the critical values are located in the tails of the null distribution, which is particularly poorly approximated by Student's t-distribution.[13]

Finally, we examined VB when estimating subject-specific accuracies. We compared three estimators: (i) posterior means of $\sigma(\rho_j)$ using VB; (ii) posterior means π_j using MCMC; and (iii) sample accuracies. The plot 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 (Figures 4.19 and 4.20, bottom row), we carried out the same analyses as above, but based on small datasets of just 8 subjects

[13]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.

with different numbers of trials (Figure 4.19f). Regarding test specificity, as before, we found fixed-effects inference to yield severely over-optimistic inferences at low test sizes (Figure 4.19g).

The same picture emerged when looking at sensitivities (Figure 4.20h,i). Again, fixed-effects inference on the pooled sample accuracy yielded over-confident results; it systematically rejected the null hypothesis too easily. By contrast, fixed-effects inference on averaged data led to overly pessimistic inferences, rejecting not frequently enough (violet). A conventional t-test is 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, subject-specific inferences benefitted substantially from a mixed-effects model when the data were limited in size (Figure 4.20j). This effect 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.

4.7.3 Accuracies versus balanced accuracies

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, that enables inference on balanced accuracies. To illustrate the differences between the two quantities, we replicated an analysis from a previous study in which we generated a typically imbalanced synthetic dataset and used a linear support vector machine (SVM) for classification (Figure 4.21; for details, see Section 4.4.6).

We observed that, as expected, the class imbalance caused the classifier to acquire a bias in favour of the majority class. This can be seen from the raw classification outcomes in which many more positive trials (green) than negative trials (red) were classified correctly, relative to their respective prevalence in the data (Figure 4.21a). 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 varies hugely (Figure 4.21b). In this setting, we found that the twofold normal-binomial model provided an excellent estimate of the true balanced accuracy (Figure 4.21c). In stark contrast, using the single

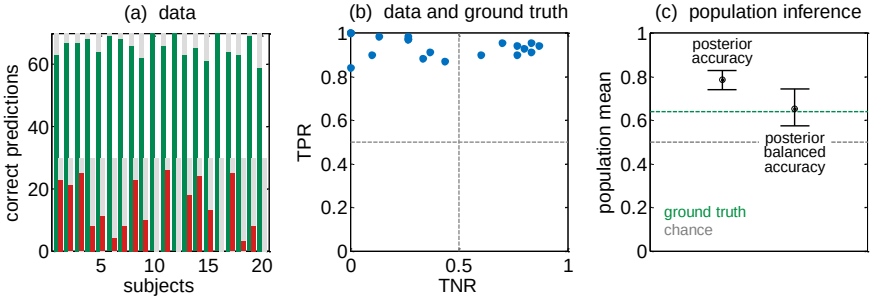

Figure 4.21: Imbalanced data and the balanced accuracy. (a) Classification outcomes obtained by applying a linear support vector machine (SVM) to synthetic data, using leave-one-out cross-validation. The plot shows, for each subject, the number of correctly classified positive (green) and negative (red) trials, as well as the respective total number of trials (grey). (b) Sample accuracies on positive (TPR) and negative classes (TNR). The underlying true population distribution is represented by a bivariate Gaussian kernel density estimate (contour lines). The plot shows that the population accuracy is high on positive trials and low on negative trials; the imbalance in the data has led the SVM to acquire a bias in favour of the majority class. (c) Central 95% posterior probability intervals of the population mean accuracy and the balanced accuracy. Inference on the accuracy is misleading, while the balanced accuracy interval provides a sharply peaked estimate of the true balanced accuracy.

normal-binomial model to infer on the population accuracy (as opposed to balanced accuracy) resulted in estimates that were severely optimistic and therefore misleading.

4.7.4 Application to fMRI data

To demonstrate the practical applicability of VB, we analysed data from an fMRI experiment involving 16 volunteers who participated in a simple decision-making task (Figure 4.22). During the experiment, subjects had to choose, on each trial, between two options that were presented on the screen. Decisions were indicated by button press (left/right index finger). Details on 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 tied to different buttons, we expected highly discriminative activity in the primary motor cortex.

Separately for each subject, a general linear model (Friston *et al.*, 1995) was used to create a set of parameter images representing trial-specific es-

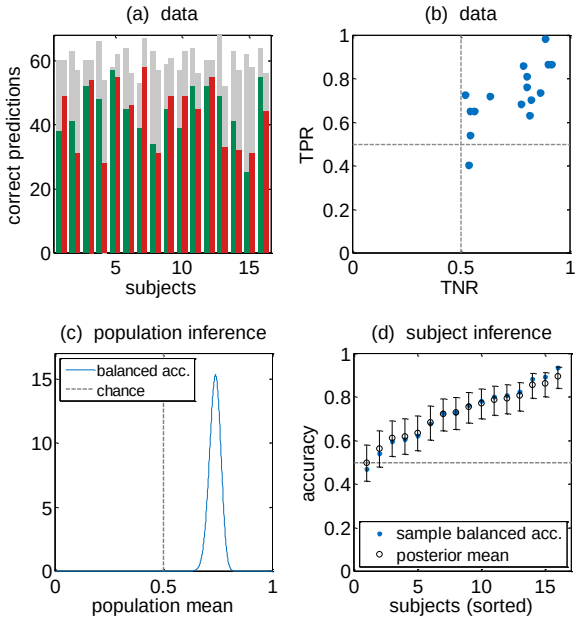

Figure 4.22: Application to empirical fMRI data I: overall classification performance. (a) Classification outcomes obtained by applying a linear SVM to trial-wise fMRI data from a decision-making task. (b) Posterior population mean accuracy, inferred on using variational Bayes. (c) Posterior population precision. (d) Subject-specific posterior inferences. The plot contrasts sample accuracies with central 95% posterior probability intervals, which avoid overfitting by shrinking to the population mean.

timates of evoked brain activity in each volume element. These images entered a linear support vector machine (SVM) that was trained and tested using leave-one-out cross-validation. Comparing predicted to actual choices resulted in 120 classification outcomes for each of the 16 subjects (Figure 4.22a).

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 (Figure 4.22b). The variational posterior population mean accuracy (posterior mean 73.7%; Figure 4.22c) agreed closely with an MCMC-based posterior (73.5%; not shown). Inference on subject-specific accuracies yielded fairly precise posterior intervals with a noticeable shrinking effect (Figure 4.22d).

The overall computation time for the above VB inferences was approximately 7 ms. This enormous 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 (Nandy and Cordes, 2003; Kriegeskorte *et al.*, 2006). In this analysis, we passed a sphere (radius 6 mm) across the brain and, at each location, trained and tested a linear SVM using leave-10-out cross-validation. We associated the voxel at the centre of the current sphere with the number of correct predictions in each subject (i.e., $k_{1:16} \in \mathbb{N}^{16}$). We then used our VB algorithm to compute a whole-brain *posterior accuracy map* (PAM; Figure 4.23). Comprising 220 000 voxels, the map needed no more than 7 min 18 s until completion (while an MCMC-based solution would have taken 31 days). The map shows the posterior population mean accuracy in those voxels with an infraliminal probability of less than 0.1%. Thus, the map highlights regions with a posterior probability of the classifier operating above chance at the group level that is at least 99.9%.

4.8 Discussion

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

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

The mixed-effects models considered in this chapter ensure that known

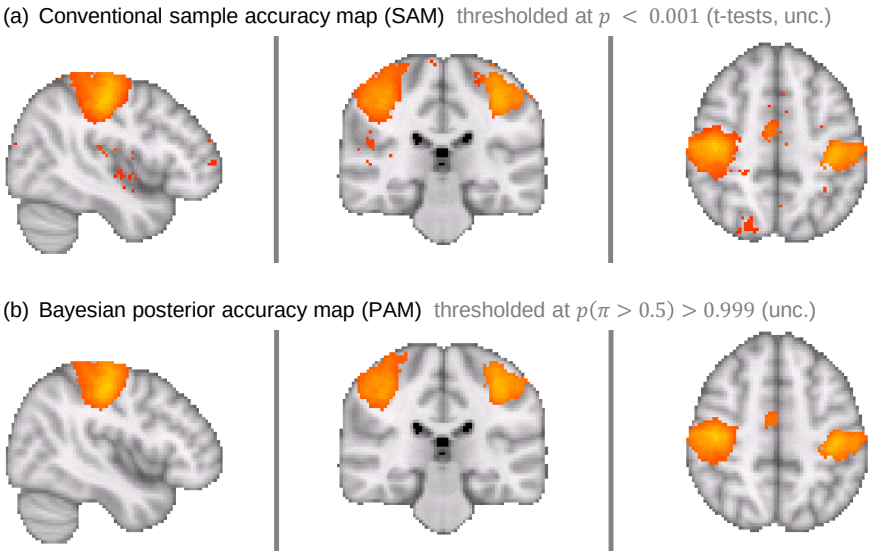

Figure 4.23: Application to empirical fMRI data II: posterior accuracy map. (a) A conventional sample accuracy map (SAM) highlights regions in which a one-tailed t-test on subject-specific sample accuracies yielded $p < 0.001$ (uncorrected). (b) Using the VB algorithm presented in this chapter, we can 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%.

dependencies between inferences on subject-specific accuracies are coherently accommodated within an internally consistent representation of the data. Specifically, the posterior distribution of the accuracy of one subject is partially influenced by the data from all other subjects, correctly weighted by their respective posterior precisions (see Section 4.3.3). Thus, the available group data are exploited to constrain individual inference appropriately. Non-hierarchical models, by contrast, risk being under-parameterized (i.e., their degrees of freedom are insufficient to fit data) or over-parameterized (i.e., they are prone to overfitting the data and generalize poorly). Hierarchical models overcome this problem in a natural way: they regularize the inversion problem by replicating the structural dependencies that govern the observed data.

The hierarchical models presented in this chapter are motivated by two-level designs that distinguish between inference at the subject level and

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

Finally, it is worth noting that mixed-effects models are not only useful when *evaluating* a classifier but also when *designing* it. For instance, Schelldorfer *et al.* (2010) proposed a linear mixed-effects model for classification that accounts for different sources of variation in the data. The model has been shown to improve classification performance in the domain of brain-computing interfaces (Fazli *et al.*, 2011).

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

An advantage of Bayesian inference is that a hierarchical data structure can be particularly easily translated into a corresponding hierarchical model, accomplishing mixed-effects inference naturally (see above).

An additional advantage of Bayesian approaches is that one can evaluate different models (e.g., alternative distributional assumptions) by comparing their respective model evidences, even when the models are non-nested. For example, in Section 4.4.5 we showed how alternative *a priori* assumptions about the population covariance of class-specific accuracies can be evaluated using Bayesian model selection.

In practice, our approach may also help avoid erroneous interpretations of inferential results. For example, the posterior infraliminal probability introduced in Section 4.3.2 has an arguably more natural and less error-prone interpretation than a classical p-value. Instead of denoting the probability of observing the data (or more extreme data) under the null hypothesis of a chance classifier (classical p-value), the infraliminal probability represents what we are ultimately interested in: the (posterior) probability that the classifier operates below (or above) chance given the data.

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

An advantage of the above frequentist scheme would be that it no longer requires an assumption common to all other approaches considered in this chapter: the assumption that trial-wise classification outcomes y_i are conditionally independent and identically distributed (i.i.d.) given a subject-specific accuracy π. This is typically justified by assuming that, in a classification analysis, test observations are i.i.d. themselves, conditional on the parameters of the latent process that generated the data. The situation is less clear in a cross-validation setting, where, strictly speaking, classification outcomes are no longer independent of one another (Gustafsson *et al.*, 2010; Kohavi, 1995; Wickenberg-Bolin *et al.*, 2006). However, because violations of i.i.d. assumptions lead to conservative inference when controlling false positive rates, the i.i.d. assumption has generally not been a major concern in the literature. If trial-by-trial dependence is an issue, then one possibility is to resort to a single-split scheme, by training on one half of the data, and testing on the other.

Replacing the accuracy by the balanced accuracy. The third feature of our approach is its flexibility with regard to performance measures.

While it is common to compare algorithms with regard to their accuracy, the limitations of this metric are well-known. For example, when a classifier is tested on an imbalanced dataset, the accuracy may be inflated and lead to false conclusions about the classifier's performance. There are different potential solutions to this problem (Akbani *et al.*, 2004; Chawla *et al.*, 2002; Japkowicz and Stephen, 2002). One can, for example, restore balance by undersampling the large class or by oversampling the small class, or modify the costs of misclassification (Zhang and Lee, 2008). A more generic safeguard is to replace the accuracy with the balanced accuracy, defined as the arithmetic mean of the class-specific accuracies. Unlike the measure described by Velez *et al.* (2007), the balanced accuracy is symmetric with respect to the type of class. If desired, this symmetry assumption can of course be dropped by introducing class-specific misclassification costs.

Fundamentally, accuracies and balanced accuracies differ because they address different scientific questions. Inference on the accuracy is asking: what is the probability of making a correct prediction on a trial randomly drawn from a distribution with the same imbalance as that present in the current training set? Inference on the balanced accuracy, by contrast, is asking: what is the probability of a correct prediction on a trial that is equally likely, *a priori*, to belong to either class? This is what we are typically interested in when assessing the presence of a statistical link between data features and labels: the expected accuracy under a flat prior over classes.

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

Using the example of the balanced accuracy, we have described how hierarchical models enable Bayesian inference on performance measures other than the accuracy. We also demonstrated that there may be multiple plausible models *a priori*. In this case, Bayesian model selection can be used to decide between competing models. Alternatively, Bayesian model averaging produces predictions which account for posterior model uncertainty. This approach can be adopted with any other performance measure of interest.[14]

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

The choice of a versatile yet convenient parameterization of the distributions for class-specific accuracies π^+ and π^- has been a recurring theme in the literature. Whereas early treatments adopted an empirical Bayes approach (e.g., Albert, 1984; Good, 1956; Griffin and Krutchkoff, 1971), the more recent literature has discussed various fully hierarchical approaches (see Agresti and Hitchcock, 2005, for an overview). For instance, Leonard (1972) proposed to replace independent Beta priors on each element of π such as those in (4.4.1) by independent normal priors on each element of $\text{logit}(\pi)$. While this is analytically convenient, it requires independence assumptions in relation to the elements of π. This limitation was addressed by Berry and Christensen (1979), who placed a Dirichlet process prior on the elements of π. A related approach was proposed by Albert and Gupta (1983), who used Beta priors on the components of π such that their degree of correlation could be controlled by a common hyperparameter. As mentioned above, a principled way of evaluating such different propositions rests upon Bayesian model comparison (MacKay, 1992; Madigan and York, 1997; Penny *et al.*, 2004), which we illustrated by deciding between alternative parameterizations for inference on the balanced accuracy.

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

Replacing stochastic by deterministic inference. Continual advances in computing power might suggest that the importance of computational ef-

ficiency should become less critical; but the converse is true. New analysis ideas keep increasing the importance of fast algorithms. One example is provided by large-scale analyses such as searchlight approaches (Nandy and Cordes, 2003; Kriegeskorte *et al.*, 2006), in which we must potentially evaluate as many classification outcomes as there are voxels in a whole-brain scan.

Using variational Bayes, as we did in this chapter, makes it possible to create whole-brain maps of posterior mean accuracies, thresholded by infraliminal probabilities (Figure 4.23). 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 minutes when using variational Bayes, as we did in Figure 4.23.

Sampling approaches to Bayesian inference come with a range of other practical challenges, such as: how to select the number of required samples; how to check for convergence; how long to run the burn-in period for; 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 much fewer practical engineering considerations. Rather, they are based on a set of distributional assumptions that can be comprehensively captured in a simple graphical model (cf. Figures 4.13 and 4.15) and compared to competing assumptions by means of Bayesian model comparison.

Thus, the second half of this chapter 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 can be safely accepted in return for huge increases in computational efficiency.

Mixed-effects inference in other analysis domains. 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 analysis (Figure 4.24). In particular, as touched upon in Section 4.1 on p. 61, mass-univariate fMRI analyses based on the general linear model, too, were initially evaluated using fixed-effects models before these were replaced by random-effects and full mixed-effects alternatives (Holmes and Friston, 1998; Friston *et al.*, 1999; Beckmann *et al.*, 2003; Woolrich *et al.*, 2004; Friston *et al.*, 2005; Mumford

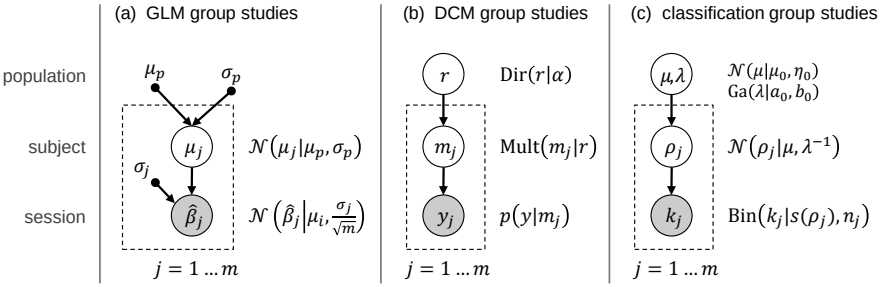

Figure 4.24: Analogies between mixed-effects models in neuroimaging. (a) The first broadly adopted models for mixed-effects inference and Bayesian estimation in neuroimaging were developed for mass-univariate fMRI analyses based on the general linear model. The figure shows a graphical representation of the summary-statistics approximation to mixed-effects inference. (b) Mixed-effects inference have subsequently also been developed for group studies based on dynamic causal modelling. (c) The present chapter addresses very similar issues, but in a different context, that is, in group analyses based on trial-by-trial classification. For details on variables, see references in the main text.

and Nichols, 2009). 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, 2002).

Another example are group analyses on the basis of dynamic causal modelling (Friston *et al.*, 2003), where fixed-effects inference has been supplemented by random-effects inference that is more appropriate when different models are optimal in characterizing different subjects in a group (Stephan *et al.*, 2009a). The present chapter addresses very similar issues, but in a different context, that is, in group analyses based on trial-by-trial classification. 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 that disregards variability at the subject or group level.

Summary of present results and conclusions. We hope that the models for Bayesian mixed-effects analyses introduced in this chapter will find widespread use. The VB approach in particular, proposed in the second half of this chapter, is as easy to use as a t-test, but conveys multiple advantages over contemporary 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' 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).

To facilitate the use of our approach, an open-source implementation of all models discussed in this chapter, including a step-by-step documentation, is available for download.[15] With this toolbox we hope to assist in improving the statistical sensitivity and correct interpretation of results in future classification group studies.

[15]For an implementation in MATLAB, see http://mloss.org/software/view/407/. An R package is currently in preparation.

Chapter 5

Model-based classification

In order to establish its potential utility for dissecting spectrum disorders, we must critically demonstrate that generative embedding may indeed be used to accurately relate measures of neural activity to an external label. This is the ambition of model-based classification. In this chapter, we propose and put to test a concrete implementation using a combination of dynamic causal models (DCM) and support vector machine (SVM) classifiers. For corresponding publications, see Brodersen *et al.* (2011a) and Brodersen *et al.* (2011b).

Model-based classification (or decoding) based on generative embedding comprises six conceptual steps (Figure 5.1). As described in Chapter 1, the first three steps are: the extraction of time series; inversion of a generative model; and embedding in a generative score space (see p. 21). The remaining three steps are analysis-specific, as described below.

4. A classification algorithm is trained and tested on a group of trials or subjects. Crucially, the only features submitted to the algorithm are parameter estimates provided by model inversion, e.g., posterior means. One could extend this, for example, by considering the full set of sufficient statistics of the conditional densities, e.g., by including the covariance matrix of a multivariate Gaussian density.

5. The classification accuracy of the approach is evaluated and compared to alternative algorithms. The accuracy can be viewed as the degree to which the biologically informed model has captured differences between classes.

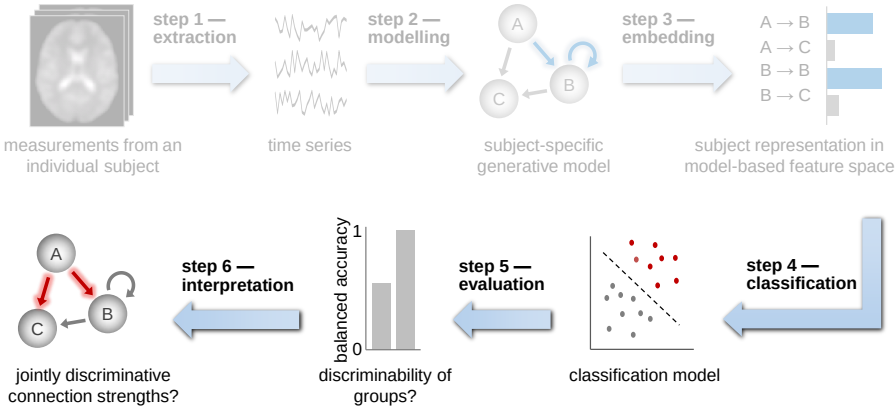

Figure 5.1: Model-based classification. The concrete implementations proposed in this chapter will use DCM as a generative model and a linear SVM as a discriminative classifier. While DCM is a natural (and presently the only) candidate for obtaining model-based estimates of synaptic plasticity (cf. Stephan *et al.*, 2008; den Ouden *et al.*, 2010), the most widely used approach to classification relies on discriminative methods, such as support vector machines (SVM; Müller *et al.*, 2001; Schölkopf and Smola, 2002). Together, DCM and SVM methods thus represent natural building blocks for classification of disease states.

6. The weights which the classifier has assigned to individual features are reconstructed and interpreted as the degree to which individual biophysical model parameters have proven informative (in the context of all features considered) in distinguishing between classes.

Advantages over conventional classification schemes. Generative embedding for model-based classification may offer three substantial advantages over conventional classification methods. First, because the approach aims to fuse the strengths of generative models with those of discriminative methods, it may outperform conventional voxel-based schemes, especially in those cases where crucial discriminative information is encoded in 'hidden' quantities such as directed (synaptic) connection strengths.

Second, the construction of the feature space is governed and constrained by a biologically motivated systems model. As a result, feature weights can be interpreted mechanistically in the context of this model. Incidentally, the curse of dimensionality faced by many conventional feature-extraction

methods may turn into a blessing when using generative embedding: the higher the temporal and spatial resolution of the data, the more precise the estimation of the parameters of the generative model, leading to better discriminability.

Third, our approach can be used to compare alternative generative model architectures in situations where evidence-based approaches, such as Bayesian model selection, are not applicable. We will deal with these three points in more detail in Chapter 7.

Overview. This chapter is organized as follows. We begin by detailing two methodological aspects of model-based classification: the incorporation of a generative kernel into a discriminative classifier (Section 5.1), and the reconstruction of feature weights (Section 5.2).

As an initial proof of concept, we illustrate the utility of model-based classification in the context of two independent electrophysiological datasets obtained in rats. The first dataset is based on a simple whisker stimulation experiment (Section 5.3); the second dataset is an auditory mismatch-negativity (MMN) paradigm (Section 5.4). In both cases, the aim is to predict, based on single-trial neural activity, which type of stimulus was administered on each trial.

We then turn to a clinical example based on an fMRI dataset acquired from moderately aphasic patients and healthy controls. We illustrate that our approach, now applied to a subject-by-subject classification setting, enables more accurate classification and deeper mechanistic insights about disease processes than conventional classification methods (Section 5.5). Finally, we discuss the key features of the proposed methods and outline future directions (Section 5.6).

5.1 Classification using a generative kernel

While a kernel describes how two subjects can be compared using a generative model of their fMRI data (cf. Section 2.5), it does not specify how such a comparison could be used for making predictions. This gap is filled by discriminative classification methods. A natural choice is the ℓ_2-norm soft-margin support vector machine (SVM), which currently represents the most widely used kernel method for classification (Boser *et al.*, 1992).

Many classification methods attempt to find a function that separates examples as accurately as possible in a space of features (e.g., voxel-wise

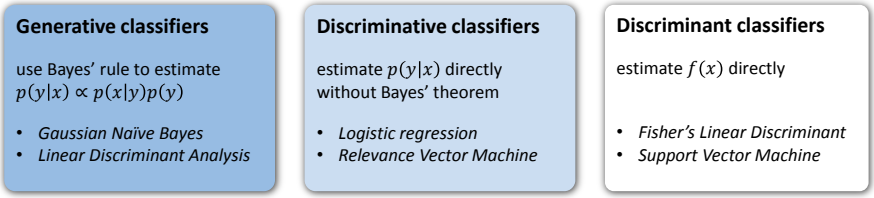

Figure 5.2: Generative, discriminative, and discriminant classifiers. Generative classifiers model the full joint probability $p(x, y)$ of data x and labels y and then derive $p(y \mid x)$ using Bayes' theorem. Discriminative classifiers can be conceptually split up into those relying on a discriminative model $p(y \mid x)$ which is estimated directly, without requiring Bayes' theorem; and those which find a discriminant function for mapping an example x onto a class label y directly, without invoking probability theory altogether. In this chapter, we will mostly use discriminant classifiers such as the support vector machine.

measurements). Such *discriminative* classification methods (sometimes referred to as *discriminant models*) differ from *generative* methods in two ways (Figure 5.2). First, rather than trying to estimate the joint density of observations and class labels (which is not needed for classification) or trying to estimate class-conditional probability densities (which can be difficult) discriminative classifiers directly model the class an example belongs to.

In binary classification, for instance, we are given a training set of n examples $x_i \in \mathbb{R}^d$ along with their corresponding labels $y_i \in \{-1, +1\}$. A learning algorithm might attempt to find a discriminant function $f \in \mathcal{F}$ from some hypothesis space $\mathcal{F}$ such that the classifier

$$h(x) := \operatorname{sgn}(f(x)) \tag{5.1.1}$$

minimizes the overall loss $\sum_{i=1}^{n} \ell(y_i, f(x_i))$. The loss function $\ell(y, f(x))$ is usually designed to approximate the unknown expected loss (or risk)

$$R[f] = \mathbb{E}_{X,Y}[\ell(Y, f(X)] \tag{5.1.2}$$

where X and Y denote the random variables of which the given examples (x_i, y_i) are realizations.

A frequent characteristic of discriminative methods is that many of them are designed not to operate on examples themselves. Rather, they are based on the similarity between any two examples, expressed as the inner product between their feature vectors. This 'kernel trick' provides an elegant way of transforming a linear classifier into a more powerful nonlinear one.

The most popular classification algorithm of the above kind is the ℓ_2-norm soft-margin SVM (Müller *et al.*, 2001; Schölkopf and Smola, 2002; Boser *et al.*, 1992; Ben-Hur *et al.*, 2008). This algorithm learns to discriminate between two groups of subjects by estimating a separating hyperplane in their feature space. The only way in which examples $x_i \in \mathbb{R}^d$ enter an SVM is in terms of an inner product $x_i^{\mathrm{T}} x_j$. This product can be replaced by the evaluation $k(x_i, x_j)$ of a kernel function

$$k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}, \tag{5.1.3}$$

which implicitly computes the inner product between the examples in a new feature space, $\phi(x_i)^{\mathrm{T}} \phi(x_j)$.

5.2 Reconstruction of feature weights

Most classification algorithms can not only be used to obtain predictions and an estimate of the generalization error that may be expected on new data. Once trained, most algorithms also indicate in one way or another which features contributed most to the overall performance attained. In cognitive neuroscience, such feature weights can be of much greater interest than the classification accuracy itself.

In contemporary decoding approaches applied to fMRI, for example, features usually represent individual voxels. Consequently, a map of feature weights projected back onto the brain (or, in the case of searchlight procedures, accuracies obtained from local neighbourhoods) may, in principle, reveal which voxels in the brain the classifier found informative (cf. Kriegeskorte *et al.*, 2006). However, this approach is often limited to the degree to which one can overcome the two challenges outlined in Chapter 1: the problem of feature selection and the problem of meaningful interpretation. Not only is it very difficult to design a classifier that manages to learn the feature weights of a whole-brain feature space with a dimensionality of 100 000 voxels; it is also not always clear how the frequently occurring salt-and-pepper information maps should be interpreted.

In contrast, using a feature space of biophysically motivated parameters provides a new perspective on feature weights. Since each parameter is associated with a specific biological role, their weights can be naturally interpreted in the context of the underlying model.

In the case of a soft-margin SVM (Figure 5.3), reconstruction of the feature weights is straightforward, especially when features are non-overlapping.

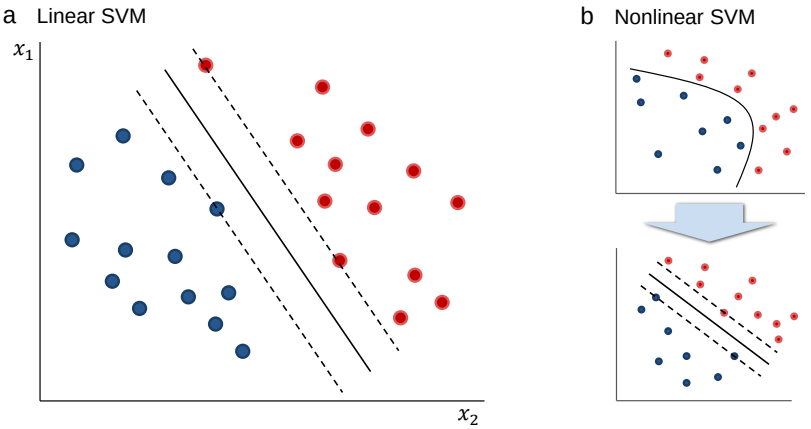

Figure 5.3: Linear and nonlinear support vector machine. (a) A linear SVM finds a maximum-margin hyperplane to separate two groups of data points. (b) A nonlinear SVM solves the exact same problem but operates in a transformed feature space that is implicitly induced by a kernel function. The advantage of a linear SVM is that there is a simple one-to-one relationship between features and feature weights, which facilitates interpretability.

Here, we briefly summarize the main principles to highlight issues that are important for generative embedding (for further points, see Ben-Hur *et al.*, 2008).

We begin by recalling the optimization problem that the algorithm solves during training:

$$\min_{w,b} w^{\mathrm{T}} w + C \sum_{i=1}^{n} \xi_i \tag{5.2.1}$$

$$\text{s.t. } \xi_i \geq 1 - y_i(w^{\mathrm{T}} x_i + b) \quad \forall i = 1, \ldots, n \tag{5.2.2}$$

$$\xi_i \geq 0, \tag{5.2.3}$$

where w and b specify the separating hyperplane, ξ_i are the slack variables that relax the inequality constraints to tolerate misclassified examples, and C is the misclassification penalty. The soft-margin minimization problem

can be solved by maximizing the corresponding Lagrangian,

$$\max_{w,b,\lambda,\alpha} \mathcal{L}(w,b,\lambda,\alpha) = \frac{1}{2} w^{\mathrm{T}} w + C \sum_{i=1}^{n} \xi_i + \sum_{i=1}^{n} \alpha_i \left(1 - y_i \left(w^{\mathrm{T}} x_i + b\right) - \xi_i\right) + \lambda^{\mathrm{T}}(-\xi). \quad (5.2.4)$$

In order to solve the Lagrangian for stationary points, we require its partial derivatives to vanish:

$$\frac{\partial \mathcal{L}}{\partial w} = w - \sum_{i=1}^{n} y_i \alpha_i x_i = 0 \quad (5.2.5)$$

$$\frac{\partial \mathcal{L}}{\partial b} = -\sum_{i=1}^{n} y_i \alpha_i = 0 \quad (5.2.6)$$

Rearranging the condition in (5.2.5) shows that the vector of feature weights w can be obtained by summing the products $y_i \alpha_i x_i$,

$$w = \sum_{i=1}^{n} y_i \alpha_i x_i, \quad (5.2.7)$$

where $x_i \in \mathbb{R}^d$ is the i^{th} example of the training set, $y_i \in \{-1, +1\}$ is its true class label, and $\alpha_i \in \mathbb{R}$ is its support-vector coefficient. More generally, when using a kernel $k(x, y) = \phi(x)^{\mathrm{T}} \phi(y)$ with an explicit feature map $\phi(x)$ that translates the original feature space into a new space, the feature weights are given by the d-dimensional vector

$$w = \sum_{i=1}^{n} y_i \alpha_i \phi(x_i). \quad (5.2.8)$$

For example, in the case of a polynomial kernel of degree p, the kernel function

$$k(x, y) = (a x^{\mathrm{T}} y + b)^p \quad (5.2.9)$$

with real coefficients a and b transforms a d-dimensional variable space into a feature space with

$$d' = \binom{d+p}{p} - 1 \quad (5.2.10)$$

dimensions that are not a constant (cf. Shawe-Taylor and Cristianini, 2004). In the case of two-dimensional examples $x = (x_1, x_2)^{\mathrm{T}}$ and a polynomial kernel of degree $p = 2$, for instance, the resulting explicit feature map is given by

$$\phi_2(x) = \left(a, \sqrt{2ab}x_i,\ \sqrt{2ab}x_2,\ bx_1^2,\ \sqrt{2}bx_1x_2,\ bx_2^2\right)^{\mathrm{T}}. \tag{5.2.11}$$

Features constructed in this way do not always provide an intuitive understanding. Even harder to interpret are features resulting from kernels such as radial basis functions (RBF). With these kernels, the transformation from a coordinate-like representation into a similarity relation presents a particular obstacle for assessing the relative contributions of the original features to the classification (cf. Schölkopf and Smola, 2002). We will therefore employ learning machines with linear kernels only (i.e., $p = 1$). This allows us to report the relative importance of a hyperplane component w_q in terms of, for instance, its normalized value

$$f_q := \frac{w_q}{\sum_{j=1}^{d'} |w_j|} \ \in [-1, 1],\ q = 1 \ldots d', \tag{5.2.12}$$

such that larger magnitudes correspond to higher discriminative power, and all magnitudes sum to unity.

Alternatives to the above procedure include the use of a sparse classifier (see Section 5.5.3, p. 158) or the use of a permutation test (see Section 5.6, p. 171).

5.3 Application to somatosensory LFPs

The most commonly investigated question in multivariate decoding is to predict from neuronal activity what type of sensory stimulus was administered on a given experimental trial. In order to investigate the applicability of generative embedding to this class of experiments, we analysed local field potentials (LFP) acquired from rats in the context of a simple sensory stimulation paradigm.

The electrophysiological recordings under consideration are highly resolved in time (here: 1 kHz). This property makes it possible to fit a neurobiologically inspired network model to individual experimental trials and hence construct a model-based feature space for classification.

The dataset considered in this section is based on a somatosensory stimulation paradigm. Using a single-shank electrode with 16 recording sites, we acquired LFPs from barrel cortex in anaesthetized rats while on each trial one of two whiskers was stimulated by means of a brief deflection. The goal was to decode from neuronal activity which particular whisker had been stimulated on each trial.

5.3.1 Experimental paradigm and data acquisition

Two adjacent whiskers were chosen for stimulation that produced reliable responses at the site of recording (dataset A1: whiskers E1 and D3; dataset A2: whiskers C1 and C3; datasets A3 and A4: whiskers D3 and β). On each trial, one of these whiskers was stimulated by a brief deflection of a piezo actuator. The experiment comprised 600 trials (Figure 5.4).

Data were acquired from 3 adult male rats. In one of these, an additional experimental session was carried out after the standard experiment described above. In this additional session, the actuator was very close to the whiskers but did not touch it, serving as a control condition to preclude experimental artifacts from driving decoding performance. After the induction of anaesthesia and surgical preparation, animals were fixated in a stereotactic frame. A multi-electrode silicon probe with 16 channels was introduced into the barrel cortex. On each trial, voltage traces were recorded from all 16 sites, approximately spanning all cortical layers (sweep duration 2 s). Local field potentials were extracted by band-pass filtering the data ($1 - 200$ Hz). All experimental procedures were approved by the local veterinary authorities.

5.3.2 Conventional decoding

Before constructing a model-based feature space for classification, we carried out two conventional decoding analyses. The purpose of the first analysis was to characterize the temporal specificity with which information could be extracted from raw recordings, whereas the second served as a baseline for subsequent model-based decoding.

We characterized the temporal evolution of information in the signal by training and testing a conventional decoding algorithm on individual time bins. Specifically, we used a nonlinear ℓ_2-norm soft-margin support vector machine (SVM) with a radial basis kernel to obtain a cross-validated estimate of generalization performance at each peristimulus time point (Chang

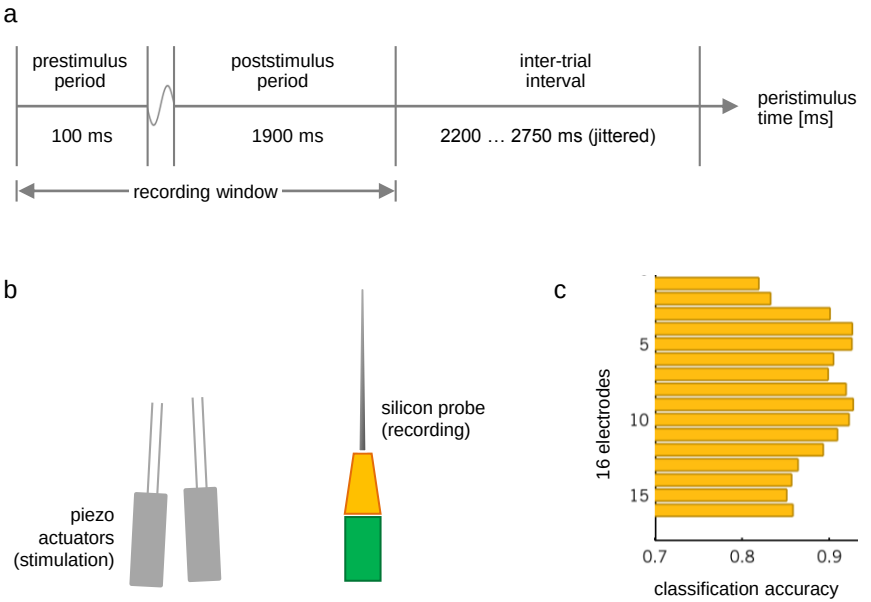

Figure 5.4: Experimental design (LFP dataset 1). The first experiment was based on a simple whisker-stimulation paradigm. (a) On each trial, after a short prestimulus period, a brief cosine-wave tactile stimulus was administered to one of two whiskers both of which had been confirmed to produce reliable responses at the site of recording. Each trial lasted for 2 s, followed by a jittered inter-trial interval. (b) Stimuli were administered using piezo actuators. Local field potentials were recorded from barrel cortex using a 16-channel silicon probe. (c) A conventional decoding analysis, applied to signals from each channel in turn, revealed a smooth profile of discriminative information across the cortical sheet. For each electrode, the diagram shows the prediction accuracy obtained when using a pattern-recognition algorithm to decode the type of whisker that was stimulated on a given trial.

and Lin, 2011). Since it is multivariate, the algorithm can pool information across all 16 channels and may therefore yield above-chance performance even at time points when no channel shows a significant difference between signal and baseline. This phenomenon was found in two out of three datasets (see arrows in Figure 5.5). Particularly strong decoding performance was obtained in dataset A2, in which, at the end of the recording window, 800 ms after the application of the stimulus, the trial type could still be deciphered from individual time bins with an accuracy of approx. 70%.

In order to obtain a baseline level for overall classification accuracies,

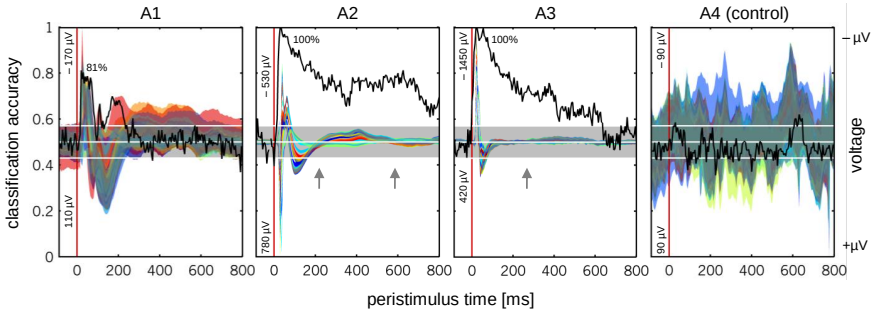

Figure 5.5: Temporal evolution of discriminative information (LFP dataset 1). The evolution of discriminative information over time can be visualized by training and testing a conventional decoding algorithm separately on the data within each peristimulus time bin. Here, time bins were formed by sampling the data at 200 Hz, and all 16 channels were included in the feature space. The black curve represents the balanced accuracy obtained within each time bin (left y-axis). Inset percentages (e.g., 81% in A1) indicate peak accuracies. Chance levels along with an uncorrected 95% significance margin are shown as white horizontal lines. Raw recordings have been added as a coloured overlay (right y-axis). Each curve represents, for one particular channel, the difference between the averaged signals from all trials of one class versus the other. The width of a curve indicates the range of 2 standard errors around the mean difference, in μV. Separately for each dataset, raw recordings were rescaled to match the range of classification accuracies, and were plotted on an inverse y-scale, i.e., points above the midline imply a higher voltage under stimulus A than under stimulus B. Minimum and maximum voltage differences are given as inset numbers on the left. As expected, since the significance margins around the chance bar are not corrected for multiple comparisons, even the control dataset occasionally achieves above-chance accuracies (as well as below-chance accuracies). The diagram shows that the classifier systematically performs well whenever there is a sufficient signal-to-noise ratio. Notably, high accuracies can be achieved even when no individual channel mean on its own shows a particularly notable difference from its baseline (arrows).

we examined how accurately a conventional decoding approach could tell apart the two trial types (see Figure 5.6). The algorithm was based on the same linear SVM that we would subsequently train and test on model-based features. Furthermore, both conventional and model-based classification were supplied with the same single-channel time series (channel 3), sampled at 1000 Hz over a $[-10, 290]$ ms peristimulus time interval. Thus using 300 data features, we found a highly significant above-chance posterior mean accuracy of 95.4% (infraliminal probability $p < 0.001$) at the group level of experimental data (A1-A3), while no significance was attained in the case of the control (A4).

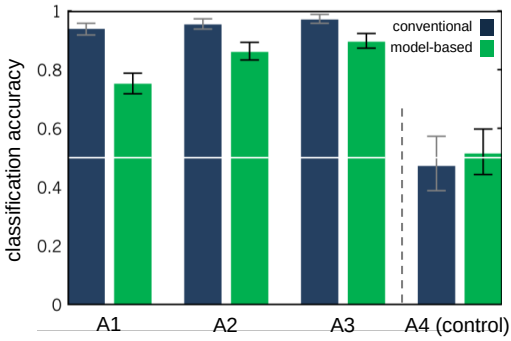

Figure 5.6: Conventional vs. model-based decoding performance (LFP dataset 1). The diagram shows overall classification accuracies obtained on each dataset, contrasting conventional decoding (blue) with model-based decoding (green). Bars represent balanced accuracies along with 95% credible intervals. Consistent in both conventional (posterior mean at the group level: 95.4%) and model-based decoding (83.6%), all accuracies are significantly above chance (infraliminal probabilities $p < 0.001$) on the experimental datasets (A1–A3). By contrast, neither method attains significance at the 0.05 level on the control dataset in which no physical stimuli were administered (A4). Despite a massively reduced feature space, model-based decoding does not perform much worse than the conventional approach and retains highly significant predictive power in all cases.

5.3.3 Generative embedding

We designed a simple DCM and used its parameter space to train and test a support vector machine (for the full model specification, see Brodersen *et al.*, 2011b, Supplement S3). Since the data were recorded from a single cortical region, the model comprised just one region. For trial-by-trial model inversion we used the recorded signal from electrode channel 3, representing activity in the supragranular layer.

Using the trial-by-trial estimates of the posterior means of the neuronal model parameters, we generated a 7-dimensional feature space (for a visualization, see Figure 5.7). We then trained and tested a linear SVM to predict, based on this model-based feature space, the type of stimulus for each trial (Figure 5.6). We found high cross-validated accuracies in all three experimental datasets (posterior mean accuracy at the group level: 83.6%, $p < 0.001$), whereas prediction performance on the control dataset was not significantly different from chance.

These results show that although the feature space was reduced by two

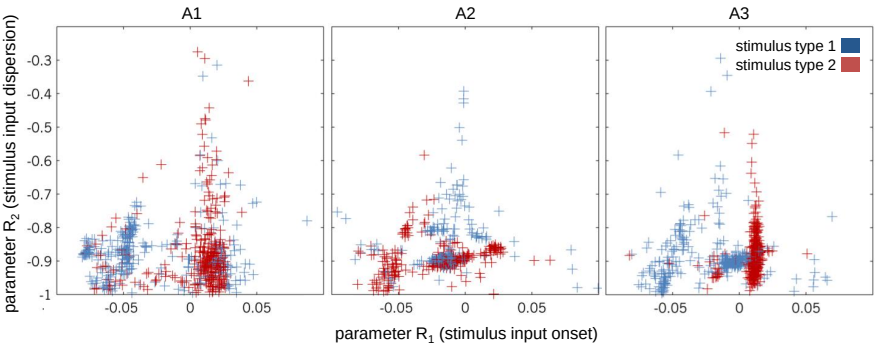

Figure 5.7: Generative score space (LFP dataset 1). The three panels show two-dimensional projections of the (7-dimensional) generative score space in the three experimental animals A1, A2, and A3. As may be intuited from the scatter plots, classification performance was strongest in A3 (see Figure 5.6).

orders of magnitude (from 300 to 7 features), model-based decoding still achieved convincing classification accuracies, all of which were significantly above chance. We next tested whether the model-based approach would yield feature weights that were neurobiologically interpretable and plausible.

5.3.4 Reconstruction and interpretation of discriminative parameters

In order to obtain discriminative feature weights, we trained our linear SVM on the entire dataset and used Eqn. (5.2.8) to reconstruct the resulting hyperplane. Thus, we obtained an estimate of the relative importance of each DCM parameter in distinguishing the two trial types. These estimates revealed a similar pattern across all three experiments (Figure 5.8). Specifically, the parameter encoding the onset of sensory inputs to the cortical population recorded from (R_1) was attributed the strongest discriminative power in all datasets.

Feature weights revealed a strikingly similar pattern across all three experiments. In particular, as described above, the model parameter representing the onset of sensory inputs to the cortical population recorded from (R_1) made the strongest contribution to the classifier's discriminative power in all datasets. This finding makes sense because in our experiment stimulation of the two whiskers induced differential stimulus input to the single electrode used. For whisker stimulation directly exciting the barrel

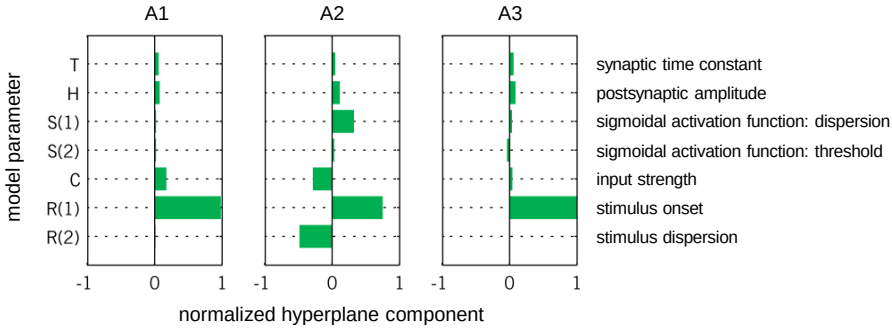

Figure 5.8: Reconstructed feature weights (LFP dataset 1). In order to make predictions, a discriminative classifier finds a hyperplane that separates examples from the two types of trial. The components of this hyperplane indicate the joint relative importance of individual features in the algorithm's success. The diagram shows the normalized value of the hyperplane component (x-axis) for the posterior expectation of each model parameter (y-axis). Feature-weight magnitudes sum to unity, and larger values indicate higher discriminative power (see main text). Consistent across all three experiments, the parameter encoding the stimulus onset (R_1) was attributed the strongest discriminative power.

recorded from, a shorter latency can be expected between sensory stimulus and neuronal response as input is directly received from thalamus. In contrast, for stimulation of the other whisker, afferent activity is expected to be relayed via cortico-cortical connections.

Similarly, a stimulus directly exciting the barrel recorded from, should be stronger and less dispersed in time than a stimulus coming from a neighbouring whisker. This is reflected by the finding that the parameters representing stimulus strength (C) and stimulus dispersion (R_2), respectively, were also assigned noticeable classification weights, although not for all three datasets. The pattern of informative features was confirmed by the 2D scatter plot (Figure 5.7), in which R_1 and R_2 play key roles in delineating the two stimulus classes.

5.4 Application to auditory LFPs

In order to explore the utility of model-based decoding in a second domain, we made an attempt to decode auditory stimuli from neuronal activity in behaving animals, using an oddball protocol that underlies a phenomenon

known as auditory mismatch negativity.

In this paradigm, two tones with different frequencies were repeatedly played to an awake, behaving rat: a frequent *standard* tone; and an occasional *deviant* tone. The goal was to decode from neuronal activity obtained from two locations in auditory cortex whether a standard tone or a deviant had been presented on a given trial (Section 5.4).[1]

5.4.1 Experimental design

The presented sequence consisted of frequent standard tones and occasional deviant tones of a different frequency (Figure 5.9a). Tone frequencies and deviant probabilities were varied across experiments. A tone was produced by bandpass-filtered noise of carrier frequencies between 5 and 18 kHz and a length of 50 ms (Figure 5.9b). Standard and deviant stimuli were presented pseudo-randomly with deviant probabilities of 0.1 (datasets B1 and B3) and 0.2 (dataset B2). The three datasets comprised 900, 500, and 900 trials, respectively.

For the present analyses we used data that was acquired from 3 animals in a sound-attenuated chamber (cf. Jung *et al.*, 2009). In order to record event-related responses in the awake, unrestrained animal, a telemetric recording system was set up using chronically implanted epidural silverball electrodes above the left auditory cortex. The electrodes were connected to an EEG telemetry transmitter that allowed for wireless data transfer. During the period of data acquisition, rats were awake and placed in a cage that ensured a reasonably constrained variance in the distance between the animal and the speakers (see Figure 5.9c). All experimental procedures were approved by the local governmental and veterinary authorities.

A robust finding in analyses of event-related potentials during the auditory oddball paradigm in humans is that deviant tones, compared to standard ones, lead to a significantly more negative peak between 150–200 ms post-stimulus, the so-called *mismatch negativity* (MMN; Näätänen *et al.*, 2001; Garrido *et al.*, 2009). Although the MMN-literature in rodents is almost exclusively concerned with animals under anaesthesia, the observed

[1] It should be noted that the second dataset was acquired using epidural silverball electrodes whose recording characteristics differ from those of the intracortical probes used in the first dataset. For the sake of simplicity, we will refer to both types of data as local field potentials (LFPs) and model both datasets using the forward model described in the Methods section.

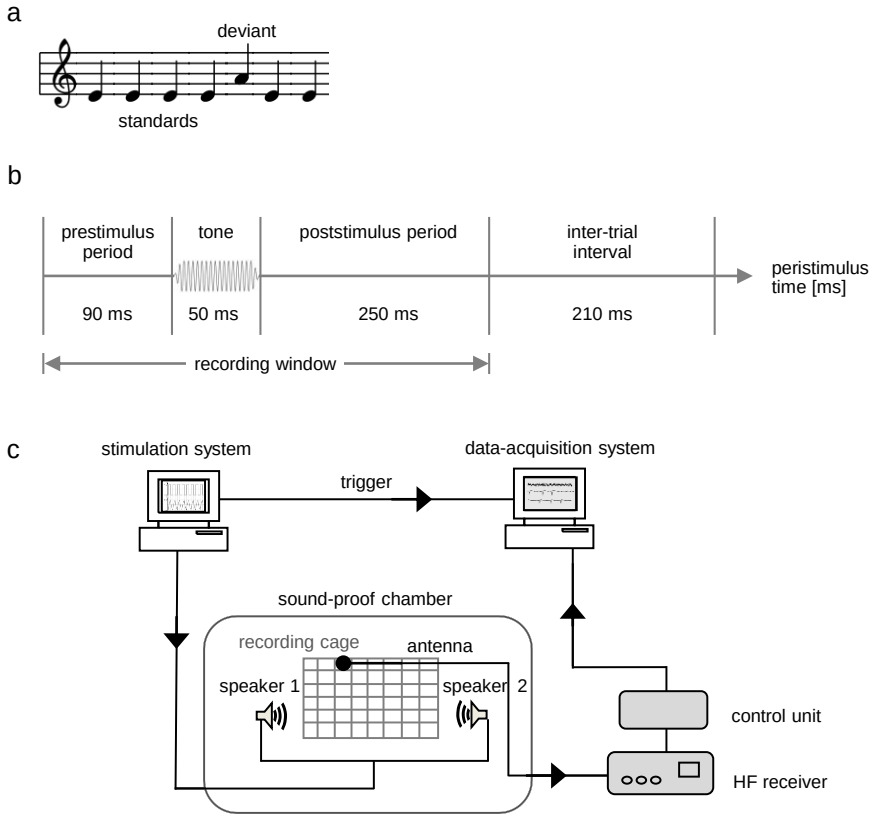

Figure 5.9: Experimental design (LFP dataset 2). (a) On each trial, the animal was presented either with a standard tone or, less frequently, with a deviant of a different frequency. (b) Each trial lasted for 600 ms, with a stimulus onset 90 ms after the beginning of a sweep. Recordings comprised 390 ms in total and were followed by an inter-trial interval of 210 ms. (c) Data were acquired in the awake, behaving animal using a wireless high-frequency receiver.

difference signals in our data are highly consistent with similar studies in rats (e.g., von der Behrens *et al.*, 2009), showing a negative deflection at approximately 30 ms and a later positive deflection at 100 ms (shaded overlay in Figure 5.10).

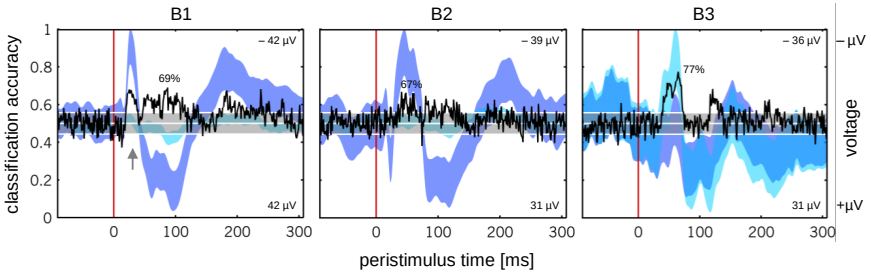

Figure 5.10: Temporal evolution of discriminative information (LFP dataset 2). By analogy with Figure 5.5, the diagram shows the temporal evolution of discriminative information in dataset 2. Time bins were formed by sampling the data from both channels at 1000 Hz. The black curve represents the balanced accuracy obtained within each time bin. The coloured overlay shows, separately for both channels, the mean signal from all deviant trials minus the mean signal from all standard trials. The diagram shows that the most typical situation in which the trial type can be decoded with above-chance accuracy is when at least one channel significantly deviates from its baseline (e.g., grey arrow in B1), though such deviations alone are not always sufficient to explain multivariate classification accuracies.

5.4.2 Conventional decoding

By analogy with Section 5.3.2, we first ran two conventional decoding analyses. For temporal classification, we used a nonlinear support vector machine with a radial basis function kernel (Chang and Lin, 2011) and characterized the temporal evolution of information in the signal by training and testing the same algorithm on individual time bins. In this initial temporal analysis, above-chance classification reliably coincided with the average difference between signal and baseline (see Figure 5.10).

In order to obtain baseline performance levels for subsequent model-based decoding, we ran a conventional trial-wise classification analysis based on a powerful polynomial kernel over all time points (see Figure 5.11). In order to ensure a fair comparison, we supplied the algorithm with precisely the same data as used in the subsequent analysis based on a model-induced feature space (see below). Specifically, each trial was represented by the time series of auditory evoked potentials from both electrodes, sampled at 1000 Hz, over a $[-10, 310]$ ms peristimulus time interval (resulting in 320 features). Across the three datasets we obtained an above-chance posterior mean prediction accuracy of 81.2% at the group level ($p < 0.001$).

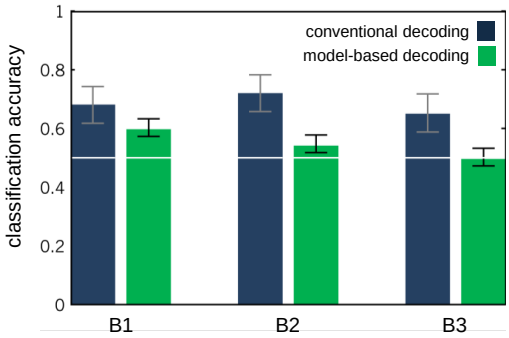

Figure 5.11: Conventional vs. model-based decoding performance (LFP dataset 2). The diagram contrasts conventional decoding (blue) with model-based decoding (green) in terms of overall classification accuracies obtained on each auditory mismatch dataset. Model-based accuracies tend to be lower than conventional accuracies, but they remain significantly above chance in 2 out of 3 cases (59.7% and 54.1%, $p < 0.05$ each). All results are given in terms of balanced accuracies (see Section 3.2) along with 95% credible intervals of the generalization performance.

5.4.3 Generative embedding

In this experiment, data from two electrodes and regions were available, enabling the construction of a two-region DCM. As the exact locations of the electrodes in auditory cortex were not known, we initially evaluated three alternative connectivity layouts between the two regions: (i) a model with forward connections from region 1 to region 2, backward connections from region 2 to region 1, and stimulus input arriving in region 1; (ii) a model with forward connections from region 2 to region 1, backward connections from region 1 to region 2, and stimulus input arriving in region 2; (iii) a model with lateral connections between the two regions and stimulus input arriving in both regions.

For each model, we created a 13-dimensional feature space based on the posterior expectations of all neuronal and connectivity parameters. We dealt with the problem of testing multiple hypotheses by splitting the data from all animals into two halves, using the first half of trials for model selection and the second half for reporting decoding results.[2]

Based on the first half of the data within each animal, we found that

[2]Cross-validation across animals, as opposed to within animals, would not provide a sensible alternative here since variability in the location of the electrodes precludes the assumption that all data stem from the same distribution.

the best discriminability was afforded by the model that assumes forward connections from region 2 to region 1 and backward connections from region 1 to 2. We then applied this model to the second half of the data, in which the auditory stimulus administered on each trial could be decoded with moderate but highly significant accuracies ($p < 0.001$) in 2 out of 3 datasets (B1 and B2; see Figure 5.11).

5.4.4 Reconstruction and interpretation of discriminative parameters

Feature weights are only meaningful to compute when the classifier performs above chance. Thus, separately for datasets B1 and B2, we trained the same SVM as before on the entire dataset and reconstructed the resulting hyperplane; see Eqn. (5.2.8). A similar pattern of weights was again found across the datasets (see Figure 5.12). In particular, the two model parameters with the highest joint discriminative power for both datasets were the parameters representing the strength of the forward and backward connections, respectively (A_F and A_B). Noticeable weights were also assigned to the extrinsic propagation delay ($D_{1,2}$) and to the dispersion of the sigmoidal activation function (S_1) (see Figure 5.12).

Very much like in the first dataset (Section 5.3), a similar pattern of feature weights was again found across the two datasets in which significant classification results had been obtained (Figure 5.11). This is not a trivial prediction, given that all results are based on entirely independent experiments with inevitable deviations in electrode positions. Nevertheless, several model parameters were found with consistent, non-negligible discriminatory power.

These consistently discriminative parameters included the strength of the forward and backward connections between the two areas (A_F and A_B) and the dispersion of the sigmoidal activation function (S_1). Other noticeable parameters included the synaptic time constants (T_1 and T_2) and the extrinsic propagation delays (D).

These findings are in good agreement with previous studies on the mechanisms of the MMN (Baldeweg, 2006; Garrido *et al.*, 2008; Kiebel *et al.*, 2007). In brief, these earlier studies imply that two separate mechanisms, i.e., predictive coding and adaptation, are likely to contribute to the generation of the MMN. While the latter mechanism relies on changes in post-synaptic responsiveness (which can be modelled through changes in the sigmoidal activation function and/or synaptic time constants), the former

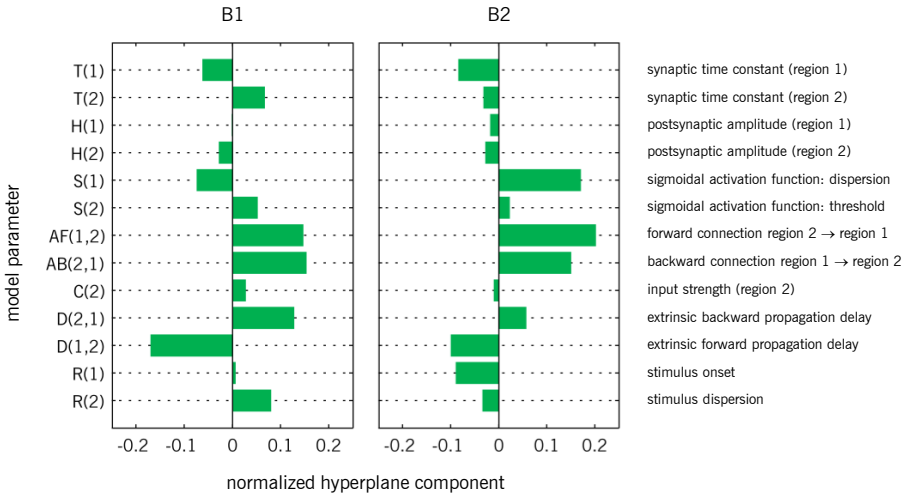

Figure 5.12: Reconstructed feature weights (LFP dataset 2). By analogy with Figure 5.8, the diagram shows the normalized hyperplane component magnitudes (x-axis) for all model parameters (y-axis). Larger values indicate higher discriminative power when considering the corresponding feature as part of an ensemble of features. One experiment (B3) was excluded from this analysis since its classification accuracy was not significantly above chance (see Figure 5.11). The sum of the feature weights of the two parameters coding for the strength of forward and backward connections (parameters A_F and A_B) was highest in both remaining datasets (B1 and B2).

highlights the importance of inter-regional connections for conveying information about prediction errors. The results of our model-based classification are consistent with this dual-mechanism view of the MMN.

5.4.5 Sensitivity analysis

Comparison between generative embedding and conventional analyses based on DCM. Model-based decoding may serve as an alternative to established procedures such as Bayesian model selection (BMS) in situations where log-evidence-based approaches are not applicable (see Section 5.6). However, it might also be worth investigating whether model-based classification offers higher or lower sensitivity than log-evidence-based approaches in situations where both could be used. Specifically, one could compare infraliminal p-values obtained from model-based classification to (equivalents of) p-values derived from Bayes factors in the context of con-

ventional DCM and BMS. In the DCM analysis, one would model the differences in class means in terms of changes in specified parameters, and then compare this model to a null model in which no changes in parameters (and thus no differences between class means) are allowed. Here, the equivalent of a p-value can be derived from the posterior model probabilities (i.e., 1 − the conditional probability that the alternate model was better than the null model).

Such a comparison is feasible but must be qualified carefully since the two approaches differ in several aspects. BMS-based p-values are the result of a fitting procedure that uses all available data, while classification operates on a strongly reduced feature space. Thus, one might generally expect model-based classification to be less sensitive than evidence-based model comparison. On the other hand, in the case of current DCM implementations for evoked responses, only a few parameters are allowed to change for explaining differences in observed responses (i.e., extrinsic connections strengths and the amplitude of excitatory postsynaptic potentials), whereas classification in a model-based feature space may utilize all parameters for identifying differences between trial types. In addition, a nonlinear classifier may allow for trial-type separation when no significant difference is revealed by class means alone. These considerations imply that the relative sensitivity of DCM/BMS vs. model-based classification may vary depending on the particular data set and model in question.

Indeed, when carrying out the comparison on our two datasets, as described below, we obtained mixed results (see Table 5.1). For the first (somatosensory) dataset, we found decoding-based p-values to be smaller than the p-values derived from the log Bayes factor in the conventional DCM analysis in two out of three cases, and both values were indistinguishable from zero in one case. In contrast, for the second (mismatch negativity) dataset, we found that in all three animals DCM-based p-values were smaller than the p-values provided by our model-based approach.

In summary, the relative sensitivity of DCM/BMS and model-based classification for establishing differences between trial types (or subject classes) is difficult to determine in full generality; rather, it likely depends on the data observed and the particular model used.

Comparison between generative embedding and Hotelling's T^2-test. Since generative embedding strongly reduces the dimensionality of the feature space, one may ask whether two trial types can be discriminated without invoking a cross-validation scheme and using a conventional

Animal	Bayesian model comparison (BMS)		Model-based classification	Result
A1	0.9445	$>$	0	decoding more sensitive
A2	0.5002	$>$	0	decoding more sensitive
A3	0	$\approx$	0	indistinguishable
A4*	0.2193	$<$	0.589	decoding more specific
B1	0	$<$	0.0113	BMS more sensitive
B2	0	$<$	0.0023	BMS more sensitive
B3	0.5046	$<$	0.9585	BMS more sensitive

Table 5.1: Comparison of p-values. The table compares infraliminal p-values obtained from model-based classification to (equivalents of) p-values derived from Bayes factors in the context of conventional DCM and BMS. The asterisk (*) denotes a control animal in which no discriminability is expected.

Animal	Hotelling's T^2-test		Model-based classification	Result
A1	0	$\approx$	0	indistinguishable
A2	0	$\approx$	0	indistinguishable
A3	0	$\approx$	0	indistinguishable
A4*	0.17	$<$	0.31	decoding more specific
B1	6.8×10^{-6}	$\approx$	3.1×10^{-6}	indistinguishable
B2	4.5×10^{-4}	$\approx$	1.2×10^{-4}	indistinguishable
B3	0.001	$<$	0.18	Hotelling's more sensitive

Table 5.2: Comparison of p-values. The table compares the significance of above-chance decoding accuracies to the outcome of Hotelling's T^2-test, the multivariate generalization of Student's t-test. The asterisk (*) denotes a control animal in which no discriminability is expected.

encoding model instead. Specifically, we compared the significance of above-chance decoding accuracies to the outcome of Hotelling's T^2-test, the multivariate generalization of Student's t-test. In our context, the null hypothesis states the absence of any difference between class-conditional means of model parameter estimates.

In the case of decoding, we computed p-values as the probability of obtaining the observed balanced accuracy under the null hypothesis that the classifier operates at chance. In the case of Hotelling's T^2-test, we computed p-values as the probability of the T^2-statistic being equal or greater than the observed value under the null hypothesis of the between-condition Mahalanobis distance being zero (Table 5.2).

Given that our data represent averages and should conform to parametric assumptions by the central limit theorem, the Neyman-Pearson lemma states that Hotelling's T^2-test should provide the most powerful test. How-

ever, it can only be applied when there are fewer features than examples, which means that the decoding scheme described in the main text has a greater domain of application.

For the first dataset, p-values were numerically indistinguishable from zero in all experimental cases (A1–A3); in the control case where no stimuli were applied (A4) and where no significant p-value is expected, neither method yielded a false positive result. For the second dataset, there was no meaningful difference between decoding-based p-values and Hotelling's p-values in two out of three cases, while only Hotelling's p-value was significant for the third animal. These anecdotal results are consistent with the notion that Hotelling's T^2-test provides the most powerful test when applicable.

5.4.6 Interim conclusions

In the preceding sections, we set out to demonstrate the utility of generative embedding for local field potentials (LFP). We analysed two independent datasets: one based on multichannel-electrode recordings from rat barrel cortex during whisker stimulation under anaesthesia (Section 5.3); and one based on two-electrode recordings from two locations in auditory cortex of awake, behaving rats during an auditory oddball paradigm (Section 5.4).

In both datasets, we used a state-of-the-art SVM algorithm in a conventional manner (applying it to approx. 300 'raw' data features, i.e., measured time points) and compared it to a model-based alternative (which reduced the feature space by up to two orders of magnitude). Specifically, we designed a model-based feature space using trial-by-trial DCMs; of course, other modelling approaches could be employed instead. Although generative embedding did not quite achieve the same accuracy as conventional methods, the results were significant in all but one instance. Importantly, it became possible to interpret the resulting feature weights from a neurobiological perspective.

Thus, we have provided a proof-of-concept demonstration for the practical applicability of model-based feature construction. The application domain we have chosen here is the trial-by-trial decoding of distinct sensory stimuli, using evoked potentials recorded from rat cortex. This method may be useful for guiding the formulation of mechanistic hypotheses that can be tested by neurophysiological experiments. For example, if a particular combination of parameters is found to be particularly important for distinguishing between two cognitive or perceptual states, then future experiments could test the prediction that selective impairment of the asso-

ciated mechanisms should maximally impact on the behavioural expression of those cognitive or perceptual states.

A more important step, from our perspective, however, will be to employ the same approach to subject-by-subject classification on the basis of human fMRI data. This particular domain may hold great potential for clinical applications, as will be examined in the next section.

5.5 Application to fMRI

It has been argued that the construction of biologically plausible and mechanistically interpretable models are critical for establishing diagnostic classification schemes that distinguish between pathophysiologically distinct subtypes of spectrum diseases, such as schizophrenia (e.g., Stephan *et al.*, 2009b). The model-based classification approach presented in this thesis could be an important component of this endeavour, particularly in cases where conventional BMS cannot be applied for discrimination of clinical (sub)groups.

5.5.1 Strategies for unbiased model specification and inversion

For conventional fMRI classification procedures, good-practice guidelines have been suggested for avoiding an optimistic bias in assessing classification performance (O'Toole *et al.*, 2007; Pereira *et al.*, 2009). Generally, to obtain an unbiased estimate of generalization accuracy, a classifier must be applied to test data that have not been used during training. In generative embedding, this principle implies that the specification of the generative model cannot be treated in isolation from its use for classification. In this section, we structure different strategies in terms of a decision tree and evaluate the degree of bias they invoke (see Figure 5.13).

The first distinction is based on whether the regions of interest (ROIs) underlying the DCM are defined *anatomically* or *functionally*.

When ROIs are defined exclusively on the basis of anatomical masks (Figure 5.13a), the selection of voxels is independent of the functional data. Using time series from these regions, the model is inverted separately for each subject. Thus, given n subjects, a single initial model-specification step is followed by n subject-wise model inversions. The resulting parameter

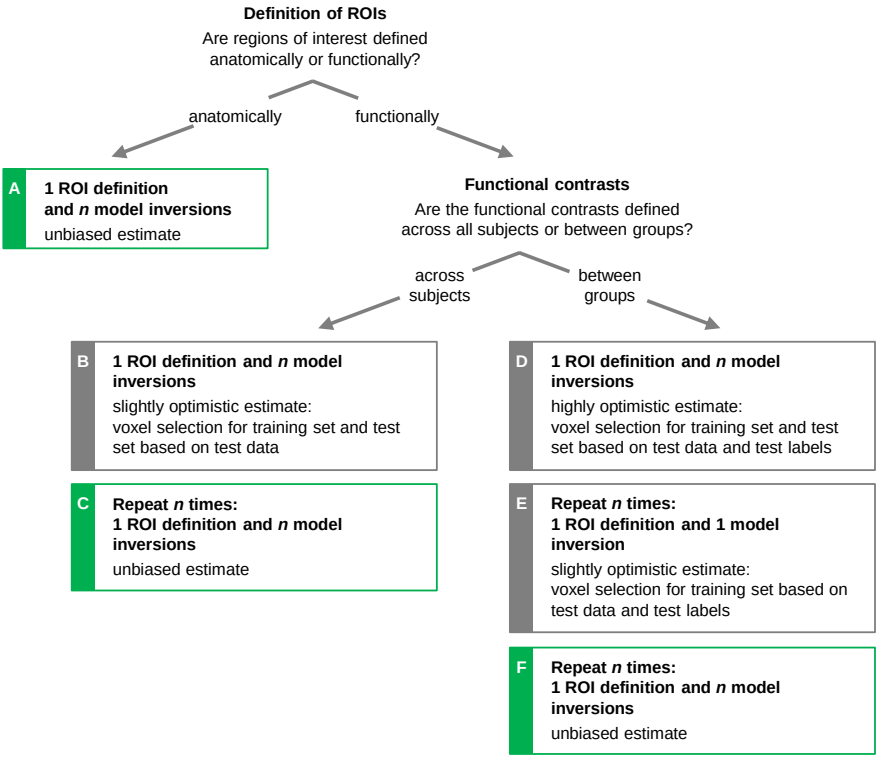

Figure 5.13: Strategies for unbiased DCM-based generative embedding. This figure illustrates how generative embedding can be implemented using dynamic causal modelling. Depending on whether regions of interest are defined anatomically, based on across-subjects functional contrasts, or based on between-group contrasts, there are several possible practical procedures. Some of these procedures may lead to biased estimates of classification accuracy (grey boxes). Procedures a, c, and f avoid this bias, and are therefore recommended (green boxes). The analysis of the illustrative dataset described in this chapter follows procedure c.

estimates can be safely submitted to a cross-validation procedure to obtain an unbiased estimate of classification performance.

Whenever *functional* contrasts have played a role in defining ROIs, subsequent classification may no longer be unbiased, since a functional contrast introduces statistics of the data into voxel selection. In this case, we ask whether contrasts are defined in an *across-subjects* or a *between-groups* fash-

ion.

In the case of an across-subjects contrast (which does not take into account group membership), one might be tempted to follow the same logic as in the case of anatomical ROI definitions: a single across-subjects contrast, computed for all subjects, guides the selection of voxels, and the resulting DCM is inverted separately for each subject (Figure 5.13b). Unfortunately, this procedure is problematic. When using the resulting parameter estimates in a leave-one-out cross-validation scheme, in every repetition the features would be based on a model with regions determined by a group contrast that was based on the data from all subjects, including the left-out test subject. This means that training the classifier would no longer be independent of the test data, which violates the independence assumption underlying cross-validation, a situation referred to as *peeking* (Pereira *et al.*, 2009). In consequence, the resulting generalization estimate may exhibit an optimistic bias.

To avoid this bias, model specification must be integrated into cross-validation (Figure 5.13c). Specifically, in each fold, we leave out one subject as a test subject and compute an across-subjects group contrast from the remaining $n - 1$ subjects. The resulting choice of voxels is then used for specifying time series in each subject and the resulting model is inverted separately for each subject, including the left-out test subject. This procedure is repeated n times, each time leaving out a different subject. In total, the model will be inverted n^2 times. In this way, within each cross-validation fold, the selection of voxels is exclusively based on the training data, and no peeking is involved. This strategy is adopted for the dataset analysed in this section, as detailed in Section 5.5.3.

When functional contrasts are not defined across all subjects but between groups, the effect of peeking may become particularly severe. Using a between-groups contrast to define regions of interest on the basis of all available data, and using these regions to invert the model for each subject (Figure 5.13d) would introduce information about group membership into the process of voxel selection. Thus, feature selection for both training and test data would be influenced by both the data and the label of the left-out test subject.

One way of decreasing the resulting bias is to integrate model specification into cross-validation (Figure 5.13e). In this procedure, the between-groups contrast is computed separately for each training set (i.e., based on $n - 1$ subjects), and the resulting regions are used to invert the model for the test subject. Consequently, the class label of the test subject is no

longer involved in selecting features for the test subject. However, the test label continues to influence the features of the training set, since these are based on contrasts defined for a group that included the test subject.This bias can only be removed by adopting the same laborious procedure as with across-subjects contrasts: by using a between-groups contrast involving $n-1$ subjects, inverting the resulting model separately for each subject, and repeating this procedure n times (Figure 5.13f). This procedure guarantees that neither the training procedure nor the features selected for the test subject were influenced by the data or the label of the test subject.

In summary, the above analysis shows that there are three practical strategies for the implementation of generative embedding that yield an unbiased cross-validated accuracy estimate. If regions are defined anatomically, the model is inverted separately for each subject, and the resulting parameter estimates can be safely used in cross-validation (Figure 5.13a). Otherwise, if regions are defined by a functional contrast, both the definition of ROIs and model inversion for all subjects need to be carried out separately for each cross-validation fold (Figure 5.13c,f).

5.5.2 Experimental design, data acquisition, and preprocessing

In order to illustrate the utility of generative embedding for fMRI, we used data from two groups of participants (patients with moderate aphasia vs. healthy controls) engaged in a simple speech-processing task. The conventional SPM and DCM analyses of these data are published elsewhere; we refer to Leff *et al.* (2008) and Schofield *et al.* (2012) for detailed descriptions of all experimental procedures.

The two groups of subjects consisted of 26 right-handed healthy participants with normal hearing, English as their first language, and no history of neurological disease (12 female; mean age 54.1 years; range 26–72 years); and 11 patients diagnosed with moderate aphasia due to stroke (1 female; mean age 66.1; range 45–90 years). The patients' aphasia profile was characterized using the Comprehensive Aphasia Test (Swinburn *et al.*, 2004). As a group, they had scores in the aphasic range for: spoken and written word comprehension (single word and sentence level); single word repetition; and object naming. It is important to emphasize that the lesions did not affect any of the temporal regions which we included in our model described below (see Schofield *et al.*, 2012, for detailed information on lesion localization).

Subjects were presented with two types of auditory stimulus: (i) normal

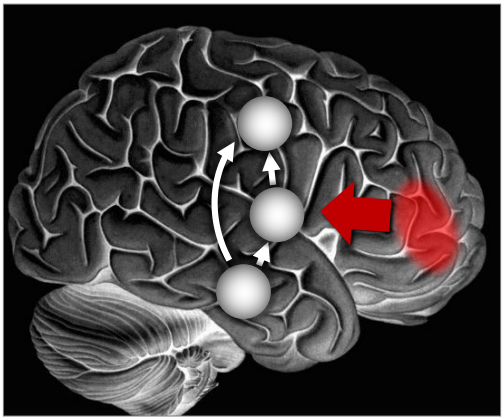

Figure 5.14: Detecting a remote lesion. To illustrate generative embedding for fMRI, we aimed to distinguish between stroke patients and healthy controls, based on non-lesioned regions involved in speech processing. In other words, we asked whether the downstream consequences of a remote lesion on a network of healthy brain regions could be picked up by a model-based classifier.

speech; and (ii) time-reversed speech, which is unintelligible but retains both speaker identity and the spectral complexity of normal speech. Subjects were given an incidental task, to make a gender judgment on each auditory stimulus, which they indicated with a button press.

Functional T2*-weighted echo-planar images (EPI) with BOLD contrast were acquired using a Siemens Sonata 1.5 T scanner (in-plane resolution 3 mm $\times$ 3 mm; slice thickness 2 mm; inter-slice gap 1 mm; TR 3.15 s). In total, 122 volumes were recorded in each of 4 consecutive sessions. In addition, a T1-weighted anatomical image was acquired. Following realignment and unwarping of the functional images, the mean functional image of each subject was coregistered to its high-resolution structural image. This image was spatially normalized to standard Montreal Neurological Institute (MNI152) space, and the resulting deformation field was applied to the functional data. These data were then spatially smoothed using an isotropic Gaussian kernel (FWHM 8 mm). In previous work, these data have been analysed using a conventional general linear model (GLM; Friston *et al.*, 1995) and (DCM; Friston *et al.*, 2003); the results are described in Schofield *et al.* (2012). Here, we re-examined the dataset using the procedure shown in Figure 5.13c, as described in detail in the next subsection.

5.5.3 Implementation of generative embedding

First-level analysis. The first level of our statistical analysis employed a mass-univariate analysis in each subject. Each auditory stimulus was modelled as a separate delta function, and the resulting trains of auditory events were convolved with a canonical haemodynamic response function. The first regressor in the design matrix contained all auditory events (i.e., normal and time-reversed speech stimuli); the second regressor modelled intelligibility (normal vs. time-reversed speech) as a parametric modulation. Beta coefficients were estimated for all voxels using a GLM. To identify regions responding to auditory stimulation *per se*, we used an 'all auditory events' contrast based on the first regressor (i.e., a contrast between auditory stimuli and background scanner noise), designed to find early auditory regions required for the perception of any broad-band stimulus, whether it is speech or speech-like.

Second-level (group) analysis. The second-level analysis served to select regions whose voxels entered the subject-specific DCMs (in terms of the first eigenvariate of their time series). In the previous study of these data (Schofield *et al.*, 2012), a set of 512 alternative DCMs had been compared that embodied competing hypotheses about the architecture of the thalamo-temporal network processing speech-like stimuli *per se*. Here, we focused on the model which was found to have the highest evidence in this previous study, i.e., the model providing the best trade-off between accuracy and complexity in explaining the data (Raftery, 1995; Stephan *et al.*, 2007a, 2009a). Note that this selection procedure is ignorant of subject labels, which prevents test labels from influencing the training procedure.[3] In addition, the selection of time series remains independent of the test data.

The DCM we used contains 6 regions (medial geniculate body, MGB; Heschl's gyrus, HG; planum temporale, PT), three in each hemisphere, and 14 interregional connections (see Figure 5.16). Note that this model concerned processing of acoustic stimuli with speech-like spectral properties *per se*, not differentiating between normal and time-reversed speech; therefore, it did not contain modulatory inputs (corresponding to all-zero $B^{(j)}$ matrices in Eqn. (2.4.2) on p. 40). Critically, instead of identifying regions functionally by a group contrast, we pre-defined large anatomical masks

[3] An alternative, computationally more expensive approach would be to select the model that affords the best classification accuracy, and integrate this selection step into an overall cross-validation scheme, as we did in Sections 5.3 and 5.4.

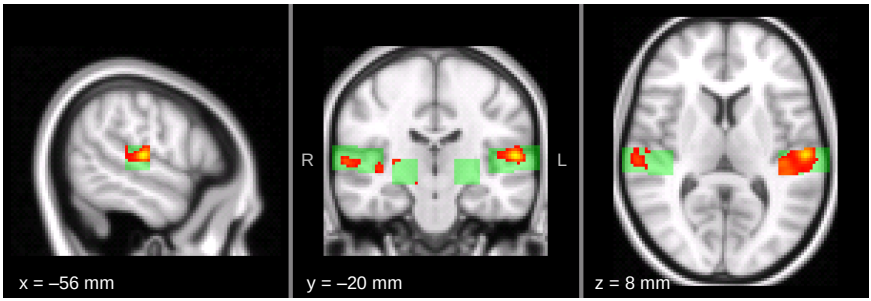

Figure 5.15: Regions of interest and searchlight classification result. (1) In order to illustrate generative embedding for fMRI, a DCM was constructed on the basis of 6 anatomical regions of interest. As described in the main text, the exact location of these regions was determined on the basis of an $n-1$ group contrast and hence was allowed to vary between cross-validation folds. Regions were defined by 16 mm × 16 mm × 16 mm cubes centred on the group maxima (see Table 5.3). The figure shows the location and extent of the anatomical masks (green) that were used to define fold-specific DCM regions. (2) A conventional searchlight analysis (Kriegeskorte *et al.*, 2006) was carried out to illustrate the degree to which a given voxel and its local spherical environment (radius 4 mm) allowed for a separation between aphasic patients and healthy controls. The map is thresholded at $p = 0.05$, uncorrected, and illustrates which regions were most informative.

L.MGB	left medial geniculate body	−23 mm, −23 mm, −1 mm
L.HG	left Heschl's gyrus (A1)	−47 mm, −26 mm, 7 mm
L.PT	left planum temporale	−64 mm, −23 mm, 8 mm
R.MGB	right medial geniculate body	22 mm, −21 mm, −1 mm
R.HG	right Heschl's gyrus (A1)	48 mm, −24 mm, 6 mm
R.PT	right planum temporale	65 mm, −22 mm, 3 mm

Table 5.3: Regions of interest. Speech processing was modelled using a DCM with 6 regions. The table lists these regions in terms of MNI152 coordinates defining the centre of the rough anatomical masks (16 mm × 16 mm × 16 mm) that guided the specification of the exact location and extent of the regions of interest underlying model inversion. For an illustration of these masks, see Figure 5.15.

(16 mm × 16 mm × 16 mm) that specified only the rough location of the 6 regions of interest (see Table 5.3). These masks served to guide the selection of time series, using a leave-one-out approach to feature selection as described below.

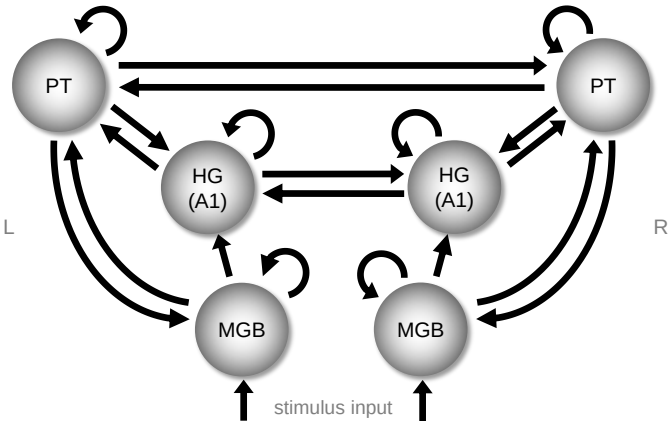

Figure 5.16: Dynamic causal model of speech processing. The diagram illustrates the specific dynamic causal model (DCM) that was used for the illustrative application of generative embedding in this study. It consists of 6 regions (circles), 15 interregional connections (arrows between regions), 6 self-connections (circular arrows), and 2 stimulus inputs (straight arrows at the bottom). The specific set of connections shown here is the result of Bayesian model selection that was carried out on the basis of a large set of competing connectivity layouts (for details, see Schofield *et al.*, 2012).

Model specification. To specify the exact location and extent of our 6 regions of interest, and thus the exact time series that would be modelled by the DCM, we used a leave-one-out approach to feature selection. For this purpose, we carried out n separate second-level analyses, each time leaving out one subject, and then used a conventional summary-statistics approach (Friston *et al.*, 2005) across the remaining $n-1$ subjects to find voxels that survived a one-sample 'all auditory events' t-test with a statistical threshold of $p = 0.001$, uncorrected, across all subjects, within the anatomical masks described above. Note that this contrast is agnostic about diagnostic status (corresponding to Figure 5.13c).[4] Within each leave-one-out repetition, our procedure yielded 6 voxel sets, one for each region of interest. We used the first eigenvariate over voxels as a representative time series for each region in DCM.

[4]With the cross-validation scheme used here, a between-group contrast could have been used as well without risking bias; see Section 5.5.1. This case would correspond to Figure 5.13f.

Model inversion. Inversion of the DCM was carried out independently for each subject, and separately for each cross-validation fold (i.e., each group contrast). With regions (and thus modelled time series) differing each time depending on the current set of $n - 1$ subjects, this procedure resulted in a total of $n^2 = 1\,369$ fitted DCMs.

Characterization of the feature space. The low dimensionality of the model-based feature space makes it possible to visualize subjects in a radial coordinate system, where each axis corresponds to a particular model parameter (see Figure 5.17). When using parameters that represent directed connection strengths, this form of visualization is reminiscent of the notion of 'connectional fingerprints' for characterizing individual cortical regions (Passingham *et al.*, 2002). In our case, there is no immediately obvious visual difference in fingerprints between aphasic patients and healthy controls. On the contrary, the plot gives an impression of the large variability across subjects and suggests that differences might be subtle and possibly jointly encoded in multiple parameters.

One way of characterizing the discriminative information encoded in individual model parameters more directly is to estimate class-conditional univariate feature densities (see Figure 5.18). Here, densities were estimated in a nonparametric way using a Gaussian kernel with an automatically selected bandwidth, making no assumptions about the distributions other than smoothness (Scott, 1992). While most densities are heavily overlapping, a two-sample t-test revealed significant group differences in four model parameters (denoted by stars in Figure 5.18): the self-connection of L.HG (parameter 4); the influence that L.HG exerts over L.PT (parameter 5); the influence R.MGB on R.PT (parameter 13); and the influence of R.HG on L.HG (parameter 14). All of these were significant at the 0.001 level while no other parameter survived $p = 0.05$.

Kernel construction. A generative score space was constructed on the basis of the posterior mean estimates of the neuronal model parameters (θ_n in (2.4.2) on p. 40). The resulting space contained 22 features: 20 interregional connection strengths (A matrix), no modulatory parameters (as the B matrix was empty in the DCM we used), and 2 input parameters (C vector). All feature vectors were normalized to unit length. To minimize the risk of overfitting and enable a clear interpretation of feature weights, we used a linear kernel. Consequently, the similarity between two subjects

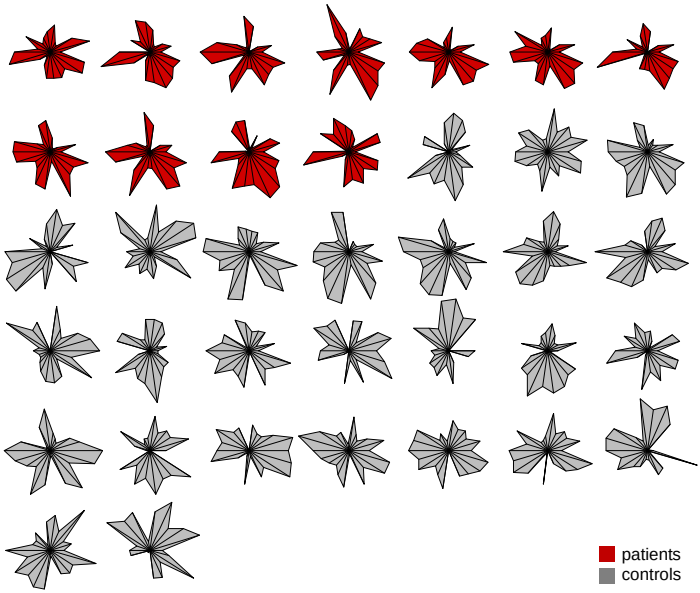

Figure 5.17: Connectional fingerprints. Given the low dimensionality of the model-induced feature space, subjects can be visualized in terms of 'connectional fingerprints' (Passingham *et al.*, 2002) that are based on a simple radial coordinate system in which each axis corresponds to the posterior mean estimate of a particular model parameter. The plot shows that the difference between aphasic patients (red) and healthy controls (grey) is not immediately obvious, suggesting that it might be subtle and potentially of a distributed nature.

was defined as the inner product between the normalized vectors of the posterior means of their model parameters.

Classification. An ℓ_2-norm soft-margin linear support vector machine (SVM) was trained and tested using leave-one-out cross-validation. Specifically, in each fold j, the classifier was trained on all subjects except j, on the basis of the DCM parameter estimates obtained from fitting the voxel time series selected by the group analysis based on all subjects except j. The classifier was then tested by applying it to DCM parameter estimates for the time series from subject j (using the same voxels as the rest of the group). Crucially, in this way, test data and test labels were neither used for model specification nor for classifier training, preventing optimistic estimates of

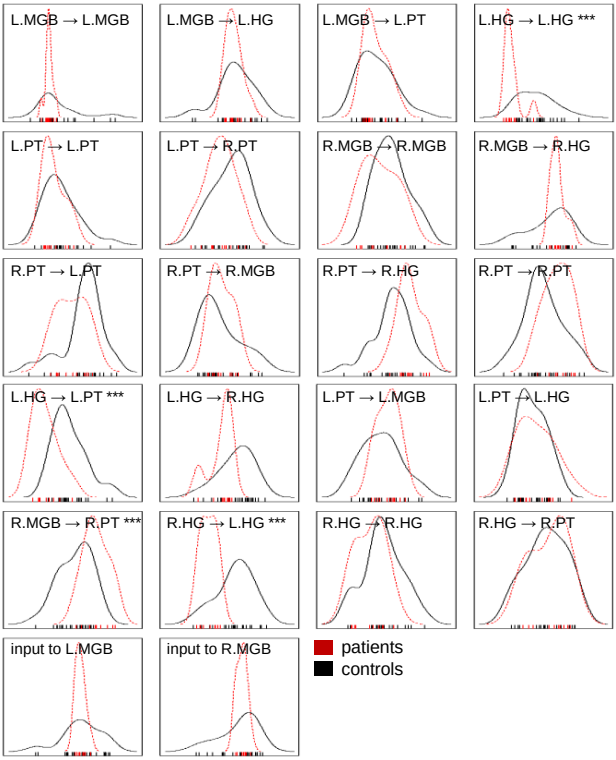

Figure 5.18: Univariate feature densities. Separately for patients and healthy controls, the figure shows nonparametric estimates of the class-conditional densities of the posterior mean estimates of model parameters. The estimates themselves are shown as a rug along the x-axis. The results of individual liberal two-sample t-tests, thresholded at $p = 0.05$, uncorrected for multiple testing, are indicated in the title of each panel. Three stars (***) correspond to $p < 0.001$, indicating that the associated model parameter assumes very different values for patients and controls.

classification performance (Figure 5.19).

Jointly discriminative features. The ℓ_2-norm SVM is a natural choice when the goal is maximal prediction accuracy. However, it usually leads to a dense solution (as opposed to a sparse solution) in which almost all features are used for classification. This dense estimation result is suboptimal when one wishes to understand which model parameters contribute most

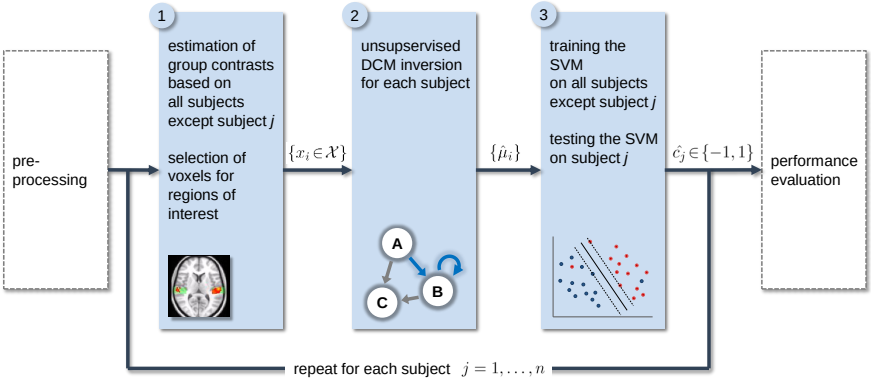

Figure 5.19: Practical implementation of generative embedding for fMRI. This figure summarizes the three steps involved in model-based classification for fMRI, designed to integrate the inversion of a generative model into cross-validation. In step 1, within a given repetition $j = 1, \ldots, n$, the model is specified using all subjects except j. This yields a set of time series $\{x_i \in \mathcal{X}\}$ for each subject $i = 1, \ldots, n$. In step 2, the model is inverted independently for each subject, giving rise to a set of subject-specific posterior parameter means $\{\hat{\mu}_i\}$. In step 3, these parameter estimates are used to train a classifier on all subjects except j and test it on subject j, which yields a prediction about the class label of subject j. After having repeated these three steps for all $j = 1, \ldots, n$, the set of predicted labels can be compared with the true labels, which allows us to estimate the algorithm's generalization performance (Chapters 3 and 4). In addition, parameters that proved jointly discriminative can be interpreted in the context of the underlying generative model. The sequence of steps shown here corresponds to the procedure shown in Figures 5.13c and 5.13f.

to distinguishing groups, which will be the focus in Section 5.5.6 where we will interpret jointly discriminative features in the context of the underlying model. In this case, an SVM that enforces feature sparsity may be more useful. One simple way of inducing sparsity is to penalize the number of non-zero coefficients by using an ℓ_0-regularizer. Unlike other regularizers, the ℓ_0-norm (also known as the counting norm) reduces the feature-selection bias inherent in unbounded regularizers such as the ℓ_1- or ℓ_2-norm. The computational cost of optimizing an ℓ_0-SVM objective function is prohibitive, because the number of subsets of d items which are of size k is exponential in k. We therefore replace the ℓ_0-norm by a capped ℓ_1-regularizer which has very similar properties (Zhang, 2009). One way of solving the resulting optimization problem is to use a bilinear programming approach (Peleg and Meir, 2008). Here, we use a more efficient difference-

of-convex-functions algorithm (Ong and An, 2012).

In summary, we will use two types of SVM. For the purpose of classification, we aim to maximize the potential for highly accurate predictions by using an ℓ_2-norm SVM. For the purpose of feature selection and interpretation, we will focus on feature sparsity by using an approximation to an ℓ_0-norm SVM, which will highlight those DCM parameters jointly deemed most informative in distinguishing between groups.

5.5.4 Comparative analyses

We compared the performance of generative embedding to a range of alternative approaches. To begin with, we examined several conventional activation-based classification schemes. The first method was based on a feature space composed of all voxels within the predefined anatomical masks used for guiding the specification of the DCMs. As above, we used a linear SVM, and all training sets were balanced by oversampling. We will refer to this approach as *anatomical feature selection*.

The second method, in contrast to the first one, was not only based on the same classifier as in generative embedding but also used exactly the same voxels. Specifically, voxels were selected on the basis of the same 'all auditory events' contrast as above, which is a common approach to defining a voxel-based feature space in subject-by-subject classification (Ford *et al.*, 2003; Fan *et al.*, 2007; Pereira *et al.*, 2009). In every cross-validation fold, only those voxels entered the classifier that survived a t-test ($p = 0.001$, uncorrected) in the current set of $n-1$ subjects. Training sets were balanced by oversampling. We will refer to this method as *contrast feature selection*.

The third activation-based method used a locally multivariate 'searchlight' strategy for feature selection. Specifically, in each cross-validation fold, a sphere (radius 4 mm) was passed across all voxels contained in the anatomical masks described above (Kriegeskorte *et al.*, 2006). Using the training set only, a nested leave-one-out cross-validation scheme was used to estimate the generalization performance of each sphere using a linear SVM with a fixed regularization hyperparameter ($C = 1$). Next, all spheres with an accuracy greater than 75% were used to form the feature space for the current outer cross-validation fold, which corresponds to selecting all voxels whose local neighbourhoods allowed for a significant discrimination between patients and healthy controls at $p = 0.01$. Both outer and inner training sets were balanced by oversampling. We will refer to this method as *searchlight feature selection*. To illustrate the location of the most infor-

mative voxels, we carried out an additional searchlight analysis, based on the entire dataset as opposed to a subset of size $n-1$, and used the results to generate a discriminative map (Figure 5.15).

The fourth conventional method was based on a principal component analysis (PCA) to reduce the dimensionality of the feature space constructed from all voxels in the anatomical masks described above. Unlike generative embedding, PCA-based dimensionality reduction finds a linear manifold in the data without a mechanistic view of how those data might have been generated. We sorted all principal components in decreasing order of explained variance. By retaining the 22 top components, the resulting dimensionality matched the dimensionality of the feature space used in generative embedding.

In addition to the above activation-based methods, we compared generative embedding to several approaches based on undirected regional correlations. We began by averaging the activity within each region of interest to obtain region-specific representative time series. We then computed pairwise correlation coefficients to obtain a 15-dimensional feature space of functional connectivity. Next, instead of computing spatial averages, we summarized the activity within each region in terms of the first eigenvariate. Thus, in this approach, the exact same data was used to estimate functional connectivity as was used by DCM to infer effective connectivity. Finally, as suggested in (Craddock *et al.*, 2009), we created yet another feature space by transforming the correlation coefficients on eigenvariates into z-scores using the Fisher transformation (Fisher, 1915).

In addition to conventional activation- and correlation-based approaches, we also investigated the dependence of generative embedding on the structure of the underlying model. Specifically, we repeated our original analysis on the basis of three alternative models. For the first model, we constructed a feedforward system by depriving the original model of all feedback and interhemispheric connections (Figure 5.20a); while this model could still, in principle, explain neuronal dynamics throughout the system of interest, it was neurobiologically less plausible. For the second and third model, we kept all connections from the original model but modelled either only the left hemisphere (Figure 5.20b) or only the right hemisphere (Figure 5.20c).

In summary, we compared the primary approach proposed in this section to 4 conventional activation-based methods, 3 conventional correlation-based methods, and 3 generative-embedding analyses using models which, in comparison to the original models, were reduced and biologically less plausible.

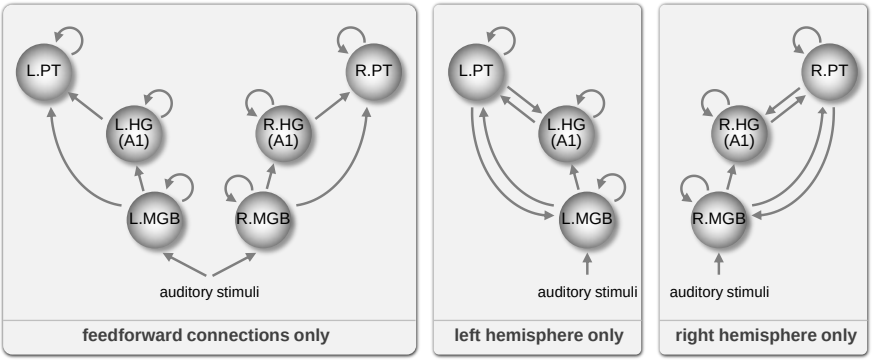

Figure 5.20: Biologically less plausible models. To illustrate the specificity of generative embedding, the analysis described in the main text was repeated on the basis of three biologically less plausible models. In contrast to the full model shown in Figure 5.16, these alternative models either (a) contained no feedback or interhemispheric connections, (b) accounted for activity in the left hemisphere only, or (c) focussed exclusively on the right hemisphere. For results, see Figures 5.21 and 5.22.

5.5.5 Classification performance

The classification performance of generative embedding was evaluated using the procedure described in Figures 5.13c and 5.19. This procedure was compared to several conventional activation-based and correlation-based approaches. As an additional control, generative embedding was carried out on the basis of three biologically ill-informed models. In all cases, a leave-one-subject-out cross-validation scheme was used to obtain the posterior distribution of the balanced accuracy as well as smooth estimates of the underlying receiver-operating characteristic (ROC) and precision-recall (PR) curves (Brodersen *et al.*, 2010b). Results are presented in Figures 5.21 and 5.22.

The strongest classification performance was obtained when using generative embedding with the full model shown in Figure 5.16. The approach correctly associated 36 out of 37 subjects with their true disease state, corresponding to a balanced accuracy of 98%.

Regarding conventional activation-based methods, classification based on anatomical feature selection did not perform significantly above chance (balanced accuracy 62%, $p \approx 0.089$). Contrast feature selection (75%, $p \approx 0.003$), searchlight feature selection (73%, $p \approx 0.006$), and PCA-based dimensionality reduction (80%, $p < 0.001$) did perform significantly above

chance; however, all methods were outperformed significantly by generative embedding ($p \approx 0.003$, $p \approx 0.001$, and $p \approx 0.045$, paired-sample Wald test). Regarding conventional correlation-based methods, all three approaches performed above chance, whether based on correlations amongst the means (70%, $p \approx 0.011$), correlations amongst eigenvariates (83%, $p < 0.001$), or z-transformed correlations amongst eigenvariates (74%, $p \approx 0.002$). Critically, however, all were significantly outperformed by generative embedding ($p < 0.001$, $p \approx 0.045$, $p \approx 0.006$).

Regarding generative embedding itself, when replacing the original model shown in Figure 5.16 by a biologically less plausible feedforward model (Figure 5.20a) or by a model that captured the left hemisphere only (Figure 5.20b), we observed a significant decrease in performance, from 98% down to 77% ($p \approx 0.002$) and 81% ($p \approx 0.008$), respectively, although both accuracies remained significantly above chance ($p \approx 0.001$ and $p < 0.001$). By contrast, when modelling the right hemisphere only (Figure 5.20c), performance dropped to a level indistinguishable from chance (59.3%, $p \approx 0.134$).

5.5.6 Reconstruction and interpretation of discriminative parameters

In order to provide a better intuition as to how the generative model shown in Figure 5.16 created a score space in which examples were much better separated than in the original voxel-based feature space, we produced two scatter plots of the data (Figure 5.23). The first plot is based on the peak voxels of the three most discriminative clusters among all regions of interest, evaluated by a searchlight classification analysis. The second plot, by analogy, is based on the three most discriminative model parameters, as measured by two-sample t-tests in the (normalized) generative score space. This illustration shows how the voxel-based projection (left) leads to classes that still overlap considerably, whereas the model-based projection (right) provides an almost perfectly linear separation of patients and controls.

To understand which DCM parameters jointly enabled the distinction between patients and controls, we examined the frequency with which features were selected in leave-one-out cross-validation when using an SVM with a sparsity-inducing regularizer (Peleg and Meir, 2008; Zhang, 2009).

We found that the classifier favoured a highly consistent and sparse set of 9 (out of 22) model parameters (see Figure 5.25); the corresponding synaptic connections are highlighted in red in Figure 5.26. Notably, this

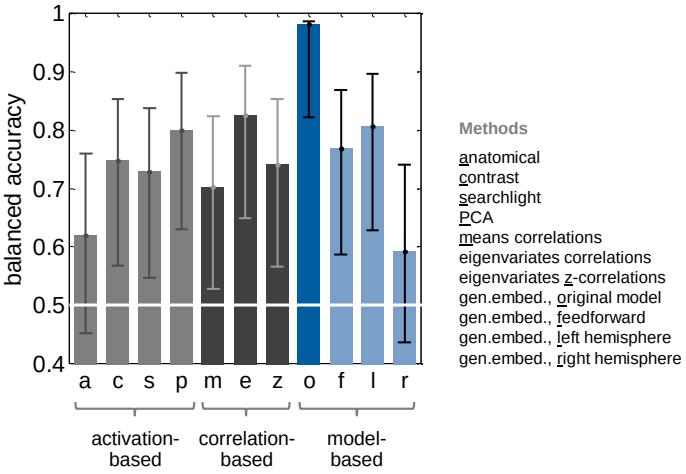

Figure 5.21: Classification performance I. Classification based on generative embedding using the model shown in Figure 5.16 was compared to ten alternative methods: anatomical feature selection, contrast feature selection, searchlight feature selection, PCA-based dimensionality reduction, regional correlations based on region means, regional correlations based on eigenvariates, regional z-transformed correlations based on eigenvariates, as well as generative embedding using three biologically unlikely alternative models. The balanced accuracy and its central 95% posterior probability interval show that all methods performed significantly better than chance (50%) with the exception of classification with anatomical feature selection and generative embedding using a nonsensical model. Differences between activation-based methods (light grey) and correlation-based methods (dark grey) were largely statistically indistinguishable. By contrast, using the full model shown in Figure 5.16, generative embedding (blue) significantly outperformed all other methods, except when used with biologically less plausible models (Figure 5.20).

9-dimensional feature space, when used with the original ℓ_2-norm SVM, yielded the same balanced classification accuracy (98%) as the full 22-dimensional feature space, despite discarding more than two thirds of its dimensions.

The above representation disclosed interesting potential mechanisms. For example, discriminative parameters were restricted to cortico-cortical and thalamo-cortical connection strengths, whereas parameters representing auditory inputs to thalamic nuclei did not contribute to the distinction between patients and healthy controls.

This finding implies that, as one would expect, low-level processing of

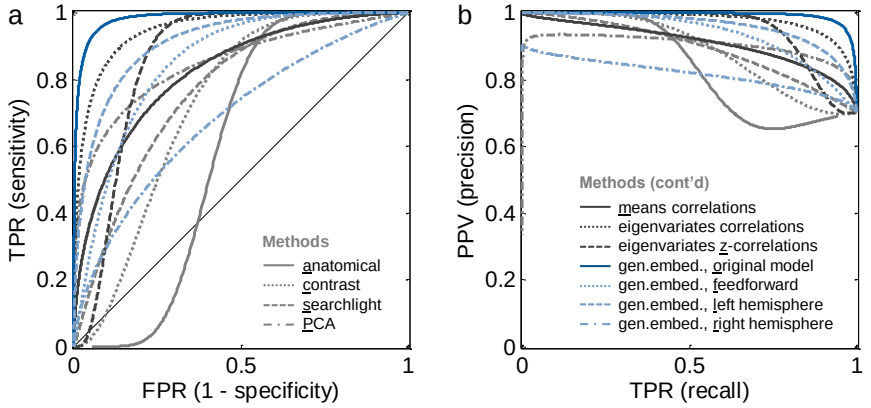

Figure 5.22: Classification performance II. (a) Receiver-operating characteristic (ROC) curves of the eleven methods illustrate the trade-off between true positive rate (sensitivity) and false positive rate (1 − specificity) across the entire range of detection thresholds. (b) Precision-recall (PR) curves illustrate the trade-off between positive prediction value (precision) and true positive rate (recall). As with ROC curves, a larger area under the curve is better. Smooth ROC and PR curves were obtained using a binormal assumption on the underlying decision values (see Brodersen *et al.*, 2010b, for details).

auditory stimuli, from brain stem to thalamus, is unimpaired in aphasia and that processing deficiencies are restricted to thalamo-cortical and cortico-cortical networks. Discriminative connections included, in particular, the top-down connections from planum temporale to Heschl's gyrus bilaterally; the importance of these connections had also been highlighted by the previous univariate analyses of group-wise DCM parameters in the study by Schofield *et al.* (2012).

Furthermore, all of the connections from the right to the left hemisphere were informative for group membership, but none of the connections in the reverse direction. This pattern is interesting given the known specialization of the left hemisphere in language and speech processing and previous findings that language-relevant information is transferred from the right hemisphere to the left, but not vice versa (Stephan *et al.*, 2007c). It implies that aphasia leads to specific changes in connectivity, even in non-lesioned parts of the language network, with a particular effect on inter-hemispheric transfer of speech information.

This specificity is seen even more clearly when considering only those

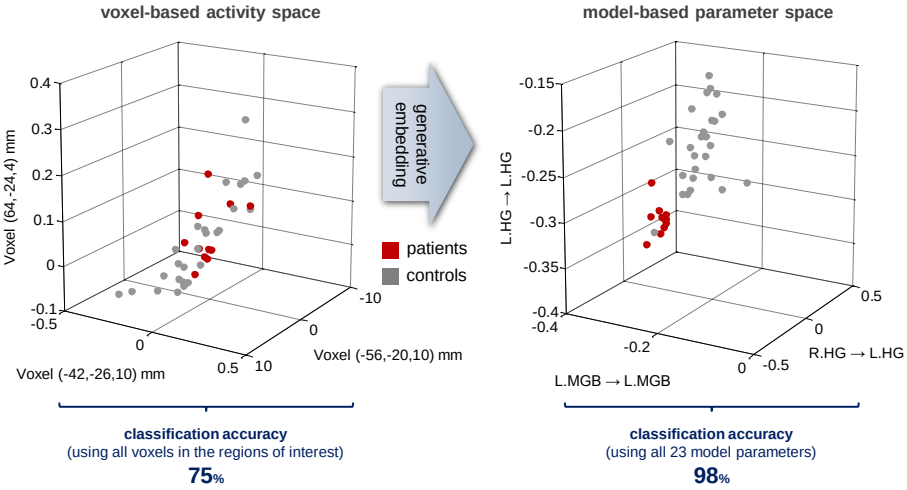

Figure 5.23: Induction of a generative score space. This figure provides an intuition of how a generative model transforms the data from a voxel-based feature space into a generative score space (or model-based feature space), in which classes become more separable. The left plot shows how aphasic patients (red) and healthy controls (grey) are represented in voxel space, based on t-scores from a simple 'all auditory events' contrast (see main text). The three axes represent the peaks of those three clusters that showed the strongest discriminability between patients and controls, based on a locally multivariate searchlight classification analysis. They are located in L.PT, L.HG, and R.PT, respectively (cf. Table 5.3). The right plot shows the three individually most discriminative parameters (two-sample t-test) in the (normalized) generative score space induced by a dynamic causal model of speech processing. The plot illustrates how aphasic patients and healthy controls become almost perfectly linearly separable in the new space. Note that this figure is based on normalized examples (as used by the classifier), which means the marginal densities are not the same as those shown in Figure 5.18 but are exactly those seen by the classifier.

three parameters which were selected 100% of the time (i.e., in all cross-validation folds) and are thus particularly meaningful for classification (bold red arrows in Figure 5.25). The associated connections mediate information transfer from the right to the left hemisphere and converge on the left planum temporale which is a critical structure for processing of language and speech (Price, 2010; Dehaene *et al.*, 2010).

In summary, all selected features represented connectivity parameters (as opposed to stimulus input); their selection was both sparse and highly consistent across resampling repetitions; and their combination was suffi-

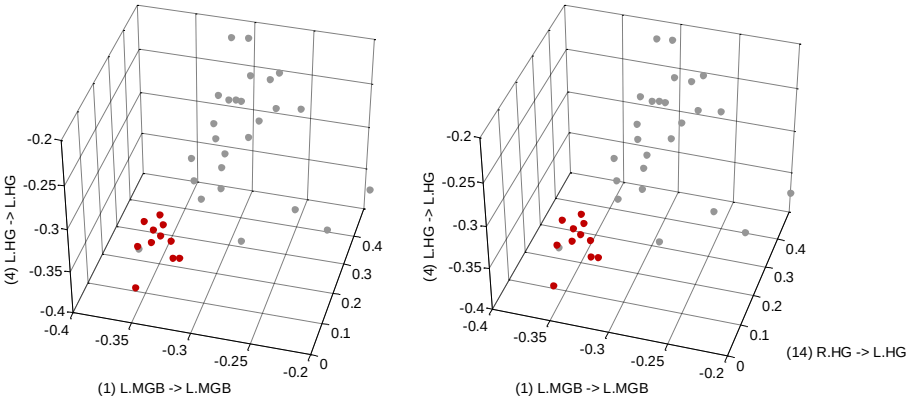

Figure 5.24: Stereogram of the generative score space. Based on the generative score space illustrated in Figure 5.23b, we here show the same plot from two slightly different angles. Readers are invited to try and focus an imaginary point behind the two plots, or use a stereoscope, to recover a fully three-dimensional impression.

cient to afford the same classification accuracy as the full feature set.

5.6 Discussion

Recent years have seen a substantial increase in research that investigates the neurophysiological encoding problem from an inverse perspective, asking how well we can decode a cognitive or clinical state from neuronal activity. Here, we have proposed a new classification approach based on generative embedding. This approach involves (i) trial-wise or subject-wise inversion of a biophysically interpretable model of neural responses, (ii) classification in parameter space, and (iii) interpretation of the ensuing feature weights.

Summary of findings. While our results on electrophysiological recordings (Sections 5.3 and 5.4) provide an initial proof of concept, the primary focus of this chapter is on our analysis of fMRI data, which provided two novel results. First, we found strong evidence in favour of the hypothesis that aphasic patients and healthy controls may be distinguished on the basis of differences in the parameters of a generative model alone. Generative embedding did not only yield a near-perfect balanced classification accuracy (98%). It also significantly outperformed conventional activation-

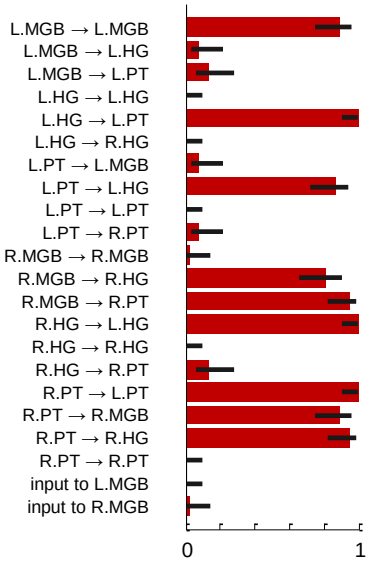

Figure 5.25: Discriminative features. A support vector machine with a sparsity-inducing regularizer (capped ℓ_1-regularizer) was trained and tested in a leave-one-out cross-validation scheme, resulting in n subsets of selected features. The figure summarizes these subsets by visualizing how often each feature (printed along the y-axis) was selected across the n repetitions (given as a fraction on the x-axis). Error bars represent central 95% posterior probability intervals of a Beta distribution with a flat prior over the interval $[0, 1]$ (cf. Section 3.2). A group of 9 features was consistently found jointly informative for discriminating between aphasic patients and healthy controls. Crucially, since each feature corresponds to a model parameter that describes one particular interregional connection strength, the group of informative features can be directly related back to the underlying dynamic causal model (see highlighted connections in Figure 5.16).

based methods, whether they were based on anatomical (62%), contrast (75%), searchlight feature selection (73%), or on a PCA-based dimensionality reduction (80%).

Similarly, our approach outperformed methods that used correlations as features, whether they were based on regional means (70%) or regional eigenvariates (74%–83%). Furthermore, it is interesting to observe that group separability was reduced considerably when using a less plausible feedforward model (77%). Finally, performance decreased significantly when modelling only the left hemisphere (81%), and it dropped to chance when considering the right hemisphere by itself (60%), which is precisely what

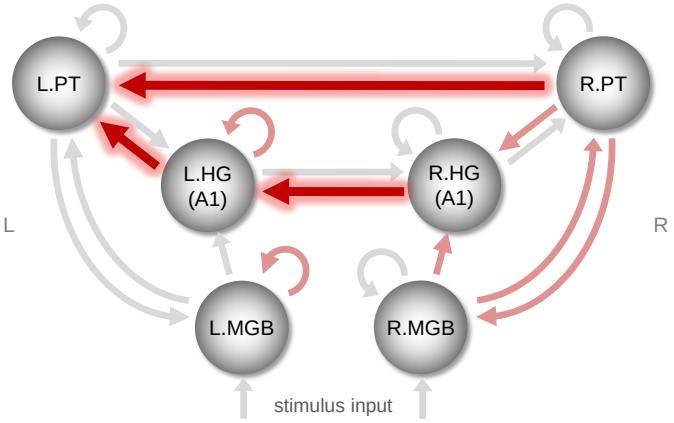

Figure 5.26: Interpretation of jointly discriminative features. A sparse set of 9 out of 23 connectivity and input parameters (see Figure 5.25) was found to be sufficiently informative to distinguish between aphasic patients and healthy controls with near-perfect accuracy (see Figure 5.21). The connections corresponding to these 9 parameters are highlighted in red. Only three parameters were selected in all cross-validation folds and are thus particularly meaningful for classification (bold red arrows); these refer to connections mediating information transfer from the right to the left hemisphere, converging on left PT, which is a key structure in speech processing.

one would expect under the view that the left hemisphere is predominantly, but not exclusively, implicated in language processing.

Taken together, our findings provide strong support for the central idea of this thesis—that critical differences between groups of subjects may be expressed in a highly nonlinear manifold which remains inaccessible by methods relying on activations or undirected correlations, but which can be unlocked by the nonlinear transformation embodied by an appropriate generative model.

The second result that we obtained from our analysis concerned interpretability. Since features correspond to model parameters, our approach allowed us to characterize a subset of features (Figure 5.25) that can be interpreted in the context of the underlying model (Figure 5.16). This subset showed four remarkable properties.

- Discriminative parameters were restricted to cortico-cortical and thalamo-cortical connection strengths. On the contrary, parameters representing auditory inputs to thalamic nuclei did not contribute to the

distinction between patients and healthy controls.

- We observed a high degree of stability across resampling folds. That is, the same 9 (out of 22) features were selected on almost every repetition.

- The set of discriminative parameters was found to be sparse, not just within repetitions (which is enforced by the underlying regularizer) but also across repetitions (which is not enforced by the regularizer). At the same time, the set was considerably larger than what would be expected from univariate feature-wise t-tests (Figure 5.18).

- The sparse set of discriminative parameters proved sufficient to yield the same balanced classification accuracy (98%) as the full set.

These results are consistent with the notion that a distinct mechanism, and thus few parameters, are sufficient to explain differences in processing of speech and speech-like sounds between aphasic patients and healthy controls. In particular, all of the connections from the right to the left hemisphere were informative with regard to group membership, but none of the connections in the reverse direction.

This asymmetry resonates with previous findings that language-relevant information is transferred from the right hemisphere to the left, but not vice versa (Stephan *et al.*, 2007c), and suggests that in aphasia connectivity changes in non-lesioned parts of the language network have particularly pronounced effects on inter-hemispheric transfer of speech information from the (non-dominant) right hemisphere to the (dominant) left hemisphere.

It is worthwhile briefly commenting on how the present findings relate to those of the original DCM study by Schofield *et al.* (2012). Two crucial differences are that the previous study (i) applied Bayesian model averaging to a set of 512 models and (ii) statistically examined each of the resulting average connection strengths in a univariate fashion. They found group differences for most connections, highlighting in particular the top-down connections from planum temporale to primary auditory cortex bilaterally.

In our multivariate analysis, these two connections were also amongst the most informative ones for distinguishing patients from controls (Figure 5.16). Schofield *et al.* (2012) also found group differences for inter-hemispheric connection strengths between left and right Heschl's gyrus, but their univariate approach did not demonstrate any asymmetries. In contrast, our multivariate approach yielded a sparser set of discriminative

connections, highlighting the asymmetries of interhemispheric connections described above (Figure 5.16).

Inference on discriminative models parameters. The above interpretation is based on SVM feature weights; these can be dependent on task-unrelated sources of variance in the data and may not always be interpretable as such. One way of addressing this issue is to relate weights to their empirical *null* distributions, i.e., those distributions that one would obtain if no statistical relationship between model parameters and diagnostic category existed. These distributions can be obtained by randomly permuting subject-specific labels and re-estimating the model N times based on the new labels, where N is a large number. A p-value for each model parameter can then be obtained as: the rank of the original feature weight within the distribution of feature weights based on permuted labels, divided by the number of permutations. Thus, to obtain a result at a given significance level α, we must repeat model estimation at least $N = 1/\alpha$ times. If this is is computationally infeasible, it is possible to run fewer permutations and summarize the null distribution in terms of the mean and variance of a Gaussian. We can then compute, for each model parameter θ_i, a score

$$t_i = \frac{w_i - \hat{\mu}_i}{\hat{\sigma}_i} \sim t_{N-1}, \tag{5.6.1}$$

where $\hat{\mu}_i$ and $\hat{\sigma}_i$ denote the sample mean and standard deviation of weights for parameter θ_i across all random permutations, and t_{N-1} is Student's t-distribution on $N - 1$ degrees of freedom. We will explore this approach in more detail in future studies; for an initial application, see Brodersen, Wiech, Lomakina *et al.* (*under review*).

Optimal decoding. The model-based classification approach described in this chapter employs a biophysically and neurobiologically meaningful model of neuronal interactions to enable a mechanistic interpretation of classification results. This approach departs fundamentally from more generic decoding algorithms that operate on raw data, which may be considered one end of a spectrum of approaches. At the other end lies what is often referred to as *optimal decoding*.

In optimal decoding, given an encoding model that describes how a cognitive state of interest is represented by a particular neuronal state, the cognitive state can be reconstructed from measured activity by inverting

the model. Alternatively, if the correct model is unknown, decoding can be used to compare the validity of different encoding models. Recent examples of this sort include the work by Naselaris *et al.* (2009) and Miyawaki *et al.* (2008), who demonstrated the reconstruction of a visual image from brain activity in visual cortex. Other examples include Paninski *et al.* (2007) and Pillow *et al.* (2008), who inverted a generalized linear model for spike trains. The power of this approach derives from the fact that it is model-based: if the presumed encoding model is correct, the approach is optimal (cf. Paninski *et al.*, 2007; Pillow *et al.*, 2008; Naselaris *et al.*, 2009; Miyawaki *et al.*, 2008). However, there are two reasons why it does not provide a feasible option in most practical questions of interest.

The first obstacle in optimal decoding is that it requires an encoding model to begin with. In other words, an optimal encoding model requires one to specify exactly and *a priori* how different cognitive states translate into differential neuronal activity. Putting down such a specification may be conceivable in simple sensory discrimination tasks; but it is not at all clear how one would achieve this in a principled way in the context of more complex paradigms. In contrast, a modelling approach such as DCM for LFPs is agnostic about a pre-specified mapping between cognitive states and neuronal states. Instead, it allows one to construct competing models of neuronal responses to external perturbations (e.g., sensory stimuli, or task demands), compare these different hypotheses, select the one with the highest evidence, and use it for the construction of a feature space.

The second problem in optimal decoding is that even when the encoding model is known, its inversion may be computationally intractable. This limitation may sometimes be overcome by restricting the approach to models such as generalized linear models, which have been proposed for spike trains (e.g. Paninski *et al.*, 2007; Pillow *et al.*, 2008); however, such restrictions will only be possible in special cases. It is in these situations where model-based classification using generative embedding could provide a useful alternative.

Choice of classifier. Generative embedding is compatible with any type of classifier, as long as its design makes it possible to reconstruct feature weights, that is, to estimate the contribution of individual features to the classifier's success. For example, an SVM with a linear or a polynomial kernel function is compatible with this approach, whereas in other cases (e.g., when using a radial basis function kernel), one might have to resort to computationally more expensive alternatives (such as a leave-one-feature-out comparison of overall accuracies).

It should also be noted that feature weights are not independent of the algorithm that was used to learn them. In this chapter, for example, we illustrated model-based classification using an SVM. Other classifiers (e.g., a linear discriminant analysis) might differ in determining the separating hyperplane and could thus yield different feature weights. Also, when the analysis goal is not prediction but inference on underlying mechanisms, alternative methods could replace the use of a classifier (e.g., feature-wise statistical testing).

Dimensionality of the feature space. Since it is model based, our approach involves a substantial reduction of the dimensionality of the original feature space. Ironically, depending on the specific scientific question, this reduction may render decoding and cross-validation redundant, since reducing the feature space to a smaller dimensionality may result in having fewer features than observations. In this situation, if one is interested in demonstrating a statistical relationship between the pattern of parameter estimates and class labels, one could use conventional encoding models and eschew the assumptions implicit in cross-validation schemes.

In the case of the first LFP dataset, for example (Section 5.3), having summarized the trial-specific responses in terms of seven parameter estimates, we could perform multiple linear regression or an ANCOVA using the parameter estimates as explanatory variables and the class label as a response variable. In this instance, the ANCOVA parameter estimates reflect the contribution of each model parameter to the discrimination and play the same role as the weights in a classification scheme. In the same vein, we could replace the p-value obtained from a cross-validated accuracy estimate by a p-value based on Hotelling's T^2-test, the multivariate generalization of Student's t-test.

In principle, according to the Neyman-Pearson lemma, this approach should be more sensitive than the cross-validation approach whenever there is a linear relationship between features and class labels. However, in addition to assuming linearity, it depends upon parametric assumptions and a sufficient dimensionality reduction of feature space, which implies that the classification approach has a greater domain of application.

An open question is how well our approach scales with an increasing number of model parameters. For example, meaningful interpretation of feature weights might benefit from using a classifier with sparseness properties: while the ℓ_2-norm support vector machine used here, by design, typically leads to many features with small feature weights, other approaches such

as sparse nonparametric regression (Caron and Doucet, 2008), sparse linear discriminant analysis (Grosenick *et al.*, 2009), group-wise regularization (van Gerven *et al.*, 2009), or sparse logistic regression (Ryali *et al.*, 2010) might yield results that enable even better interpretation. One could also attempt to directly estimate the mutual information between the joint distribution of combinations of model parameters and the variable of interest. These questions will be addressed in future studies.

Dynamic and structural model selection. An important aspect in generative embedding is the choice of a model. For the second LFP dataset described in this chapter, for example, there was a natural choice between three different connectivity layouts. The better the model of the neuronal dynamics, the more meaningful the interpretation of the ensuing feature weights should be. But what constitutes a 'better model?'

Competing models can be evaluated by Bayesian model selection (BMS; Friston *et al.*, 2007; Penny *et al.*, 2004; Stephan *et al.*, 2009a). In this framework, the best model is the one with the highest (log) model evidence, that is, the highest probability of the data given the model (MacKay, 1992). BMS has been very successful in model-based analyses of neuroimaging and electrophysiological data. It also represents a generic and powerful approach to model-based classification whenever the trial- or subject-specific class labels can be represented by differences in model structure (Figure 5.27, panel 1). However, there are two scenarios in which BMS is problematic and where the approach suggested by this chapter may represent a useful alternative.

The first problem is that BMS requires the explananda (i.e., the data features to be explained) to be identical for all competing models. This requirement is fulfilled, for example, for DCMs of EEG or MEG data, where the distribution of potentials or fields at the scalp level does not change with model structure. In this case, BMS enables both dynamic model selection (i.e., concerning the parameterization and mathematical form of the model equations) and structural model selection (i.e., concerning which regions or nodes should be included in the model). However, when dealing with fMRI or invasive recordings, BMS can only be applied if the competing models refer to the same sets of brain regions or neuronal populations; this restriction arises since changing the regions changes the data (Friston, 2009). At present, BMS thus supports dynamic, but not structural, model selection for DCMs of fMRI and invasive recordings. This restriction, however, would disappear once future variants of DCM also optimize spatial parameters of

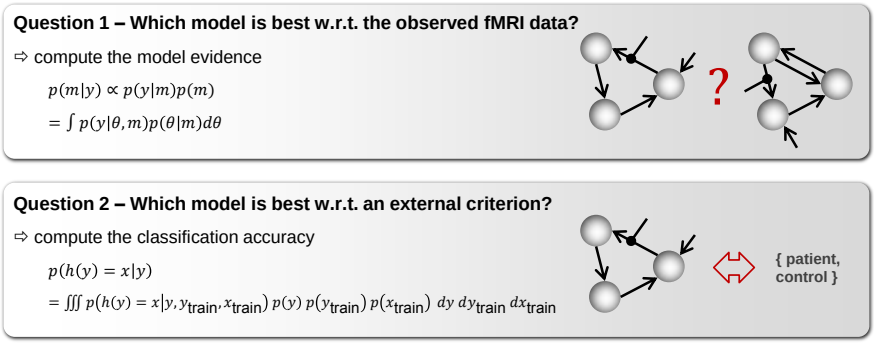

Figure 5.27: Different perspectives on model selection. (1) Bayesian model selection can be used to decide which model provides the best explanation of the data. Competing models typically differ in their structure, while the data must be the same. (2) A complementary perspective is provided by generative embedding, by asking which model is best with respect to an external criterion. Incidentally, this enables model selection even in those cases where the data (e.g., regions of interest) differ between models.

brain activity.

Secondly, with regard to model-based classification, BMS is limited when the class labels to be discriminated cannot be represented by models of different structure, for example when the differences in neuronal mechanisms operate at a finer conceptual scale than can be represented within the chosen modelling framework. In this case, discriminability of trials (or subjects, respectively) is not afforded by differences in model structure, but may be provided by different patterns of parameter estimates under the same model structure (an empirical example of this case was described recently by Allen *et al.*, 2010). In other words, differences between trials (or subjects, respectively) can be disclosed by using the parameter estimates of a biologically informed model as summary statistics.

In both above scenarios, the approach proposed in this chapter enables model comparison (Figure 5.27, panel 2), since model-based feature construction can be viewed as a method for biologically informed dimensionality reduction, and the performance of the classifier is related to how much class information was preserved by the estimates of the model parameters. In other words, training and testing a classifier in a model-induced feature space means that classification accuracies can now be interpreted as the degree to which the underlying model has preserved discriminative information

about the features of interest. This view enables a classification-based form of model comparison even when the underlying data (e.g., the chosen regional fMRI time series or electrophysiological recordings) are different, or when the difference between two models lies exclusively in the pattern of parameter estimates.

If discriminability can be afforded by patterns of parameter estimates under the same model structure, one might ask why not simply compare models in which the parameters are allowed to show trial-specific (or subject-specific) differences using conventional model comparison? One can certainly do this, however the nature of the inference is different in a subtle but important way: the differences in evidence between trials (or subjects) afforded by BMS are not the same as the evidence for differences between trials (or subjects). In other words, a *difference in evidence* is not the same as *evidence of difference*. This follows from the fact that the evidence is a nonlinear function of the data. This fundamental distinction means that it may be possible to establish significant differences in parameter estimates between trials (or subjects) in the absence of evidence for a model of differences at the within-trial (or within-subject) level. This distinction is related intimately to the difference between random- and fixed-effects analyses. Under this view, the approach proposed in this chapter treats model parameters as random effects that are allowed to vary across trials (or subjects); it can thus be regarded as a simple random-effects approach to inference on dynamic causal models.

In summary, our approach is not meant to replace or outperform BMS in situations when it can be applied. In fact, given that BMS rests on computing marginal-likelihood ratios and thus accords with the Neyman-Pearson lemma, one may predict that BMS should be optimally sensitive in situations where it is applicable (for an anecdotal comparison of BMS and model-based classification, see Section 5.4.5). Instead, the purpose of the present chapter is to introduce an alternative solution for model comparison in those situations where BMS is not applicable, by invoking a different criterion of comparison: in model-based classification, the optimal model is the one that generalizes best (in a cross-validation sense) with regard to discriminating trial- or subject-related class labels of interest.

Chapter 6

Model-based clustering

There is increasing pressure in the field of neuroimaging to translate basic research into clinical applications, as reflected, for instance, by the growing number of classification approaches that aim to provide automated diagnostic tools. In particular, an increasing number of studies have begun to describe neurobiological markers for psychiatric disorders (Davatzikos *et al.*, 2005, 2008a,b; Fu *et al.*, 2008; Misra *et al.*, 2009; Nenadic *et al.*, 2010; Klöppel *et al.*, 2008, 2009, 2012).

These studies have received much attention; but they also suffer from two limitations. Firstly, conventional classification approaches are 'blind' to domain knowledge and typically do not convey mechanistic insights. Secondly, their utility is limited in the sense that they are often replicating diagnostic categories which are flawed themselves. In the domain of spectrum disorders, for example, diagnostic labels are based on purely symptomatic descriptions rather than underlying pathophysiological differences. As a result, these labels do not convey mechanistic insights, and they are poor predictors of drug response or treatment outcome.

The examples described in the previous chapter have shown that generative embedding can be used to infer a diagnostic state using measures of brain activity and, at the same time, convey mechanistic insights. However, we have not yet addressed the issue of meaningful labels. In our fMRI application, for example, the diagnostic status of each subject was known without doubt and the networks involved in speech processing are well characterized. This circumstance allowed us to provide an initial proof of principle for the utility of generative embedding when labels are known.

Above and beyond, however, we hope to use generative embedding for addressing clinical problems of high practical relevance. In particular, we hope that generative embedding may prove useful for dissecting psychiatric spectrum disorders, such as schizophrenia, into physiologically defined subgroups (Stephan *et al.*, 2009b). We therefore turn to cases where the mechanisms underlying a given collection of symptoms are poorly understood, and we will complement classification analyses by an additional, an *exploratory* or *unsupervised* perspective.

This chapter proposes a model-based clustering approach that might help discover mechanistically defined subgroups that are not known *a priori*. We suggest a concrete implementation for fMRI data and apply our method to a large group of patients diagnosed with schizophrenia and healthy controls ($n = 83$). Full details will be provided in Brodersen *et al.* (*in preparation*).

Model-based clustering by generative embedding comprises six conceptual steps: (i) extraction of time series, (ii) modelling and model inversion, (iii) embedding in a generative score space, (iv) clustering, (v) validation with respect to clinical facts, and (vi) interpretation of the identified subgroups. The following sections briefly describe these steps (Figure 6.1).

6.1 Clustering and model selection

The critical difference between the approach proposed here and the procedures presented in Chapter 5 is that a supervised classification algorithm is replaced by an unsupervised clustering algorithm. This gives rise to a new class of applications that turn from hypothesis testing towards exploratory data analysis.

6.1.1 Extraction of time series

The first step in a model-based clustering analysis concerns the extraction of data features that will be subject to modelling. Here, we use the same approach to time-series extraction as described in Section 2.4 and used in Section 5.5. Specifically, we begin by specifying a set of regions of interest (ROIs) defined anatomically and by means of a functional contrast. We then compute, for each region, the first eigenvariate based on all voxels contained in that region, which yields a region-specific representative time course of BOLD activity.

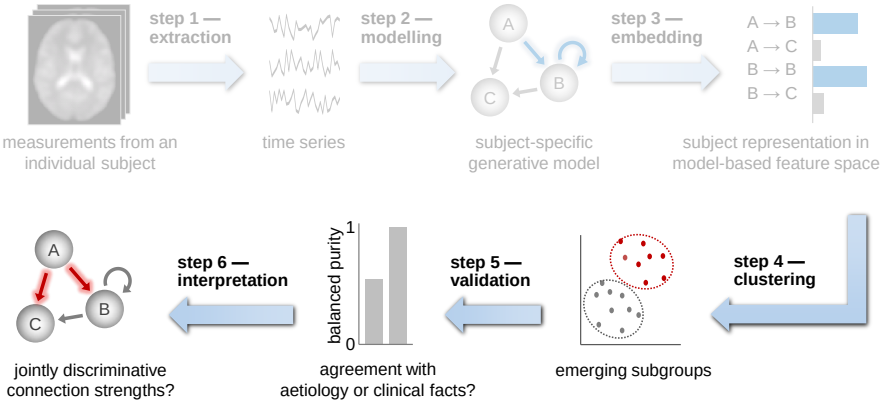

Figure 6.1: Model-based clustering. This schematic illustrates how generative embedding enables model-based clustering of fMRI data. First, BOLD time series are extracted from a number of regions of interest, separately for each subject. Second, subject-specific time series are used to estimate the parameters of a generative model. Third, subjects are embedded in a generative score space, in which each dimension represents a specific model parameter. This space implies a similarity metric under which any two subjects can be compared. Fourth, a clustering algorithm is used to identify salient substructures in the data. Fifth, the resulting clusters are validated against known external (clinical) variables. In addition, a clustering solution can, sixth, be interpreted mechanistically in the context of the underlying generative model.

In order to enable an unbiased clustering analysis, it is important that the selection of time series is not based on group information. This means in particular that between-group contrasts should not be used for the definition of regions of interest (cf. Figure 5.13).

6.1.2 Modelling and model inversion

Generative embedding rests on the specification and inversion of a generative model of the data. As detailed in Chapters 2 and 5, we use a dynamic causal model (DCM; Friston *et al.*, 2003). DCM provides a mechanistic model for explaining measured time series of brain activity as the outcome of hidden dynamics in an interconnected network of neuronal populations and its experimentally induced perturbations. Inverting such a model means to infer the posterior distribution of the parameters of both a neuronal and a forward model layer from observed responses within a given subject.

6.1.3 Embedding in a generative score space

To obtain a clustering of subjects, one might represent each subject as a vector of voxel-wise activity measurements over time. Such a feature space would retain all information we have measured within each region of interest. However, it would disregard the spatio-temporal structure of the data as well as the process that generated them, which has motivated the search for more natural ways of representing functional datasets. One such way is to embed the data in a feature space that is constructed using a generative model. This generative score space embodies a model-guided dimensionality reduction of the observed data.

As pointed out previously, a straightforward way of creating a generative score space, using DCM, is to consider the posterior expectations of model parameters of interest (e.g., parameters encoding synaptic connection strengths). More formally, we can define a mapping $\mathcal{M}_\Theta \rightarrow \mathbb{R}^d$ that extracts a subset of point estimates $\hat{\mu} := \langle \theta | x, m \rangle$ from the posterior distribution $p(\theta|x, m)$. This simple d-dimensional vector space represents a summary of the information encoded in the connection strengths *between* regions, as opposed to activity levels *within* these regions. Alternatively, one could also incorporate elements of the posterior covariance matrix into the vector space.

It is worth emphasizing that, because a generative-embedding approach rests upon a mechanistically motivated dynamic systems model, a model-based feature space of the sort described above is implicitly based on a highly nonlinear mapping: from individual measurement time points to posterior parameter expectations of an underlying dynamical system. In addition, the generative score space is not just driven by the data themselves; because the generative model is inverted in a fully Bayesian fashion, the resulting space incorporates domain knowledge and information from previous experiments that drove the specification of the prior. These aspects may be critical when aiming for an interpretable clustering solution, as described next.

6.1.4 Clustering

By model-based clustering we refer to the notion of using a clustering algorithm in a generative score space, i.e., in a space in which each dimension represents a model parameter. Here, we use a Gaussian mixture model (GMM) for clustering, which we invert using a variational Bayes approach (see Penny *et al., in preparation*, for a detailed description). This approach

has two strengths. Firstly, it provides an approximation to the model evidence which can be used for model-order selection, i.e., for deciding on the number of clusters (conditional on their Gaussianity). Secondly, it produces probabilistic output; this is in contrast to the support vector classifiers used in Chapter 5 which only provided point estimates of class membership. One could go even one step further and make the model order part of the model itself. One common class of clustering models that support this are rooted in nonparametric Bayesian inference (see Rasmussen, 2000; Dubey *et al.*, 2004; Iwata *et al.*, 2012, for the underlying theory, an application, and a more recent development).

In brief, a Gaussian mixture model defines a likelihood of the data $x_j \in \mathbb{R}^d$ of a given subject j as

$$p(x_j \mid \mu, \Sigma, \pi) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x_j \mid \mu_k, \Sigma_k). \tag{6.1.1}$$

The form of this likelihood is based on a model with K clusters. Each cluster is defined in terms of a mean μ_k and a covariance matrix Σ_k, and so the model as a whole is defined in terms of cluster means $\mu = (\mu_1, \ldots, \mu_K)$ and covariance matrices $\Sigma = (\Sigma_1, \ldots, \Sigma_k)$. The data are modelled as belonging to cluster k with probability π_k; cluster membership k itself is defined as an indicator variable, i.e., $\pi_k = p(k_j = k \mid \pi)$. Thus, the log likelihood of a full dataset of i.i.d. subjects $j = 1 \ldots m$ is given by

$$\ln p(x \mid \mu, \Sigma, \pi) = \sum_{j=1}^{m} \ln \sum_{k=1}^{K} \pi_k \mathcal{N}(x_j \mid \mu_k, \Sigma_k). \tag{6.1.2}$$

The model can be inverted using the EM algorithm to find maximum-likelihood estimates of cluster assignments as well as the cluster means and covariances themselves. This approach is simple and efficient. However, it can be prone to singularities when a Gaussian component collapses on a single data point, causing the log likelihood to diverge to infinity. Moreover, a maximum-likelihood formulation of GMM assumes that the optimal number of the Gaussian components be known *a priori*.

These limitations can be overcome using a variational Bayes approach (Penny *et al., in preparation*). This approach entails the same advantages we built on in Chapter 4 (Sections 4.5 and 4.6). It eschews the problem of singularities by introducing regularizing priors over all parameters. In addition, VB enables us to determine the optimal number of clusters by

means of Bayesian model selection. Specifically, we can compute a free-energy bound to the log model evidence

$$\ln p(x) = \ln \iiint p\left(x \mid \mu, \Sigma, \pi\right)\ p(\mu, \Sigma, \pi)\, \mathrm{d}\mu\, \mathrm{d}\Sigma\, \mathrm{d}\pi, \tag{6.1.3}$$

which enables model comparison. The model is re-estimated several times, each time using a different number of clusters. The (approximate) log model evidence is then used to decide on the optimal number of clusters, from which we obtain the final clustering solution.

6.2 Validation

The clustering solution with the highest model evidence yields the most likely substructure given the data and the GMM assumptions. However, any clustering solution remains an untested hypothesis unless we explicitly validate it against known structure that is external to the clustering model itself. We therefore explicitly assess whether a given clustering solution matches the structure implied by external variables.

External variables are often categorical. For instance, each subject might be associated with a symptom-based diagnostic category, such as schizophrenia. In this case, we wish to assess how well the clustering solution matches diagnostic categories. This goal can be achieved by computing the *purity* of the solution.

Informally, purity $\in [0, 1]$ measures how homogeneous the obtained clusters are. A perfectly homogeneous cluster contains only subjects from the same class; whereas a heterogeneous cluster contains a mixture of data points from different classes. Homogeneous clusters indicate that the clustering solution has picked up the implicit grouping structure defined by an external variable which, critically, was unavailable at the time of clustering. Another way of interpreting purity is by asking: what would the classification accuracy of an algorithm be that assigned each example to the majority class within its cluster? This accuracy is the purity of the solution. Under this view, purities can be meaningfully compared with accuracies.

To compute the purity of a solution, all data points are conceptually assigned to the class label that occurs most frequently in the associated

cluster. Purity is then calculated as

$$\text{purity}(\Omega, \mathbb{C}) := \frac{1}{n} \sum_{k=1}^{K} \max_{j} |\omega_k \cap c_j| \,, \tag{6.2.1}$$

where n is the number of subjects, $\Omega = (\omega_1, \omega_2, \ldots, \omega_k)$ is the set of cluster assignments, and $\mathbb{C} = (c_1, c_2, \ldots, c_j)$ is the set of true classes. The term $|\omega_k \cap c_j|$ represents the number of subjects in cluster k with external label j. Thus, purity is a number between 0 and 1 and indicates the degree to which the obtained clustering solution agrees with grouping structure implied by an external categorical variable.

One limitation of the purity in (6.2.1) is its misleading nature when applied to imbalanced datasets. The underlying issue is exactly the same as with classification accuracy, which is a misleading measure of classification performance when the data are not perfectly balanced. In these cases, the balanced accuracy is a more useful performance measure as it removes the bias that may arise when applying a classification algorithm to an imbalanced dataset. Here, we introduce the same idea to provide bias correction for the purity of a clustering solution. Specifically, we define the *balanced purity* as

$$\text{bp}(\Omega, \mathbb{C}) := \left(1 - \frac{1}{n}\right) \left(\frac{\text{purity}(\Omega, \mathbb{C}) - \xi}{1 - \xi}\right) + \frac{1}{n}. \tag{6.2.2}$$

In the above expression, ξ is the degree of imbalance in the data, defined as the fraction of subjects associated with the largest class. When cluster assignments perfectly agree with the external variable, the balanced purity is 1. By contrast, when cluster assignments are random, the quantity drops to $1/K$. In this way, the balanced purity can be interpreted in the same way as the (balanced) accuracy of a classification algorithm. It indicates the probability with which a new subject with label $\tilde{y}$ would be assigned to a cluster in which the majority of subjects have the same label $\tilde{y}$.

External variables may be continuous rather than categorical. For example, we might want to assess to what extent an obtained clustering solution is related to a (continuous) measure of symptoms or clinical outcome. In this case, the concept of purity no longer applies. Instead, we could validate a solution, for instance, by testing the (null) hypothesis, using a one-way ANOVA, that the distribution of the external variable has the same mean in all clusters. Examples of both categorical and continuous external variables for validation will be considered in the next section.

6.3 Application to synthetic fMRI data

The application of model-based clustering using generative embedding will be described in the context of two separate datasets. First, we generate synthetic fMRI data to illustrate the individual analysis steps and to clarify the conceptual difference between model selection and model validation. Second, we apply our approach to an fMRI dataset acquired in schizophrenia patients and healthy controls ($n = 83$) and demonstrate the nature of insights that can be gained using a generative-embedding approach.

Data generation

To illustrate the key features of our analysis approach, we generated four synthetic fMRI datasets and applied a model-based clustering analysis to each of them (Figure 6.2).

To begin with, we specified ground-truth connection parameters for $K = 2$ groups with 40 subjects each, totalling $n = 80$ subjects (Figure 6.2, top left plot within each panel). The two groups differed only in the strength of the modulatory input on regions 1 and 2. In particular, the influence of region 1 on region 2 was strongly influenced by external modulatory input in group 1 but not group 2 (encoded by the model parameter B_{21}). In group 2, conversely, this influence affected the effective connectivity from region 2 to region 1 (model parameter B_{12}). To induce population variability, connections were sampled from group-specific population distributions:

$$p\left(\begin{pmatrix}B_{21}\\B_{12}\end{pmatrix}\,\middle|\,k_1\right) = \mathcal{N}\left(\begin{pmatrix}B_{21}\\B_{12}\end{pmatrix}\,\middle|\,\begin{pmatrix}-1\\1\end{pmatrix},\begin{pmatrix}\psi & 0\\0 & \psi\end{pmatrix}\right) \quad \text{for group 1;} \tag{6.3.1}$$

$$p\left(\begin{pmatrix}B_{21}\\B_{12}\end{pmatrix}\,\middle|\,k_1\right) = \mathcal{N}\left(\begin{pmatrix}B_{21}\\B_{12}\end{pmatrix}\,\middle|\,\begin{pmatrix}1\\-1\end{pmatrix},\begin{pmatrix}\psi & 0\\0 & \psi\end{pmatrix}\right) \quad \text{for group 2.} \tag{6.3.2}$$

We induced different degrees of group separability by varying the population variance ψ between 1/2 (Figure 6.42, left column) and 1/20 (right column).

To obtain a synthetic BOLD signal, we generated a neuronal trajectory for each subject and added Gaussian observation noise. We induced different signal-to-noise ratios (SNR) by varying the noise variance between 1 and 10. We then used the conventional full Bayesian approach implemented in SPM8/DCM10 to (re)estimate the underlying parameters from the synthetic BOLD time series. Model inversion was carried out independently for each subject and, critically, was uninformed by group membership.

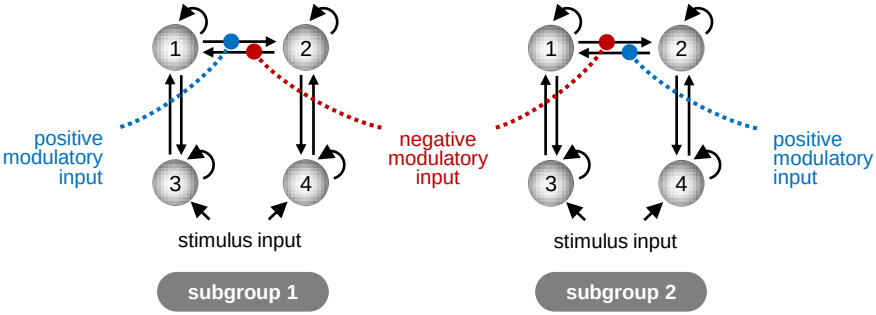

Figure 6.2: Dynamic causal model underlying synthetic fMRI data. Data were generated for two subgroups using two simple four-region models. The only difference between the groups concerned the modulatory connections (B_{21} and B_{12}).

Differences in population variability and SNR are reflected by the accuracy of the parameter estimates (Figure 6.3, bottom left plot within each panel). Posterior means agreed nicely with the true parameters when the SNR was high or groups were clearly separated. By contrast, group boundaries became less well-defined when low group separability was combined with a low SNR ratio. In all simulations, parameter estimates display a clearly noticeable shrinkage effect towards zero, as induced by the conservative shrinkage prior on modulatory connections.

Model-based clustering

Separately for each synthetic dataset, we applied model-based clustering for $K = 1 \ldots 5$ clusters by fitting a Gaussian mixture model to DCM parameter estimates. In the case of low group separability and low SNR (Figure 6.3c), the highest model evidence was obtained by a clustering solution with 3 distinct clusters, that is, inference was biased towards too large a number of groups (Figure 6.3c). In all other scenarios, the clustering algorithm correctly inferred the true number of clusters ($K = 2$; Figure 6.3a,b,d). To minimize the effects of algorithm initialization, each analysis was repeated 10 times with randomly sampled initial cluster locations.

To assess how well the obtained clustering solutions agreed with the true group structure, we computed the balanced purity[1] for each clustering

[1] In this particular case, since the simulated data were perfectly balanced across groups, the balanced purity reduced to the conventional *purity* (see Section 6.2 on p. 182).

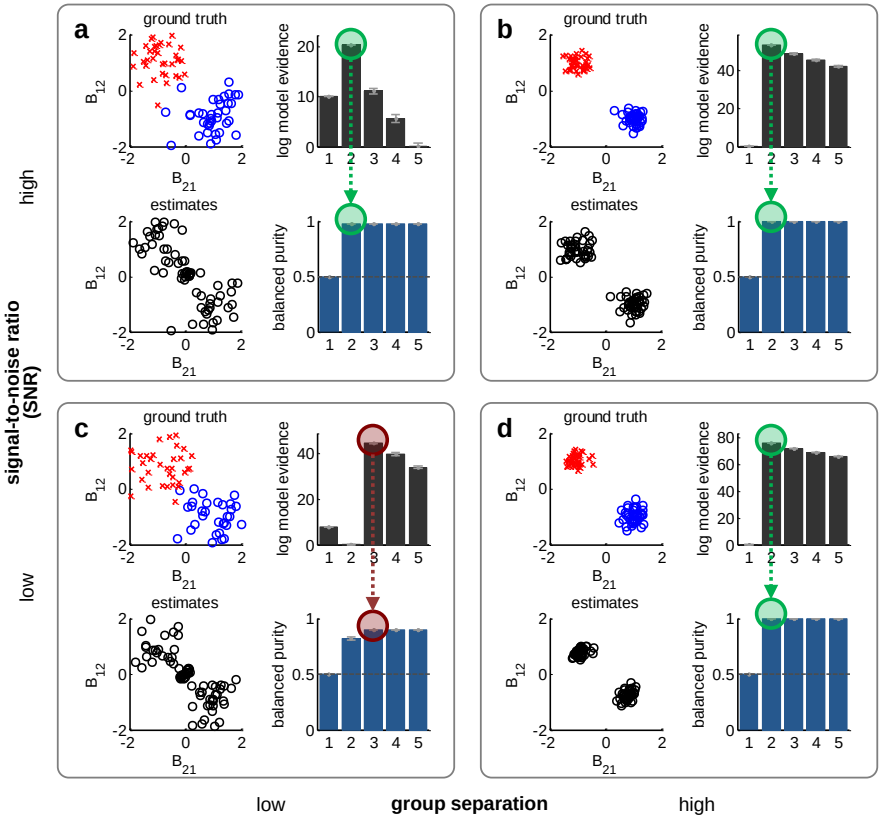

Figure 6.3: Model-based clustering results on synthetic fMRI data. Panels a–d represent simulations based on different signal-to-noise ratios and degrees of group separation. The four plots within each panel show: the true parameters (top left); their posterior mean estimates based on generated fMRI data (bottom left); the log model evidence of clustering solutions with different numbers of clusters (i.e., model selection; top right); and the balanced purity of these clustering solutions with respect to true group membership (i.e., model validation; bottom right).

solution. We observed that in those cases where the correct number of clusters had been inferred, the corresponding clustering solution showed a perfect purity of 100%.

The above analyses do not guarantee the feasibility of model-based clustering in its full generality; but they do illustrate how, in principle, salient

structure can be successfully discovered by generative embedding in situations of sufficient group separation and signal-to-noise ratio. We will next apply the same approach to empirical fMRI data.

6.4 Application to schizophrenia

To demonstrate the utility of model-based clustering in a clinical setting, we analysed a large fMRI dataset ($n = 83$) based on (i) a group of 41 patients diagnosed with schizophrenia (10 female; mean age 34.1 years; SD 10.4); and (ii) a group of 42 healthy controls (19 female; mean age 35.4; SD 12.2). In this study, subjects were engaged in a simple working memory task while undergoing fMRI. A brief summary of the task, data acquisition, and preprocessing is provided below; we refer to Deserno *et al.* (2012) for details.

In brief, functional imaging data were acquired on a 1.5 T MRI scanner (Siemens Magnetom Vision) using whole-brain gradient-echo echo-planar imaging (TR 2600 ms; TE 40 ms; flip angle 90°; matrix 64×64; voxel size $4 \times 4 \times 5.5\text{mm}^3$). Volumes were realigned to a mean image of both functional time series to correct for between-scan movements. The mean image was normalized to the MNI standard EPI template, and the parameters obtained in the normalization matrix were applied to the realigned images, which were resliced with a voxel size of $4 \times 4 \times 4\text{mm}^3$. The resulting images were then spatially smoothed using an isotropic Gaussian kernel (FWHM 8 mm).

In previous work, these data have been analysed using a conventional general linear model (GLM) and DCM; the results are described in Deserno *et al.* (2012). Here, we re-examined the dataset using the procedure shown in Figure 6.1 and the DCM shown in Figure 6.4.

Model inversion was carried out, separately for each subject, in an unsupervised fashion, i.e., without reference to the subjects' diagnostic status. We constructed a generative score space on the basis of the posterior means of all neuronal model parameters. The resulting space contained 12 features: 6 interregional connections as well as 3 self-connections (A matrix); 2 modulatory parameters (B matrix); and 1 input parameter (C matrix). Using this feature space, we asked whether the difference between patients and healthy controls would emerge as the most salient structure in the generative score space.

Without further processing of this space, it is possible that analysis results are driven by salient features in the data other than clinical variables such as diagnostic status. We therefore regressed out sex, handedness, and

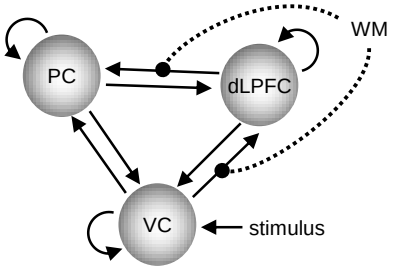

Figure 6.4: Dynamic causal model of working-memory activity. The model consists of three fully-connected nodes representing activity in the visual cortex (VC), dorsolateral prefrontal cortex (dlPFC), and parietal cortex (PC). Stimulus input enters the system in the visual cortex, whereas the working-memory condition in the experiment may modulate bottom-up effective connectivity from VC to dlPFC and PC.

age, using a separate multiple linear regression model for each model parameter. Thus, model-based classification was carried out on the residuals of parameter estimates after removing demographic confounds.

Model-based classification

Before turning to model-based clustering, we adopted a *supervised* approach and used model-based classification (Chapter 5) to distinguish between patients and healthy controls. Specifically, we trained and tested a linear support vector machine (Chang and Lin, 2011) on subject-specific connectivity patterns, i.e., using the posterior expectations of DCM parameters (Figure 6.5a).

This model-based classification algorithm was able to separate patients and controls with a balanced accuracy of 78% (infraliminal probability $p < 0.01$). We compared this result to an alternative approach in which the classifier operated on estimates of (undirected) functional connectivity rather than posterior means of effective connectivity. Following the same procedure as in Chapter 5, functional connectivity was computed in terms of Pearson correlation coefficients among eigenvariates of BOLD time series (cf. Figure 5.21e). This approach yielded a classification accuracy of 61%. While this was still significantly above chance ($p < 0.05$), it was significantly outperformed by generative embedding ($p < 0.01$). Thus, the overall accuracy of 78% provided a baseline measure of how well patients and healthy controls can be separated when the identity of each subject in the training

data is known *a priori*.

Importantly, however, it is not classification analyses of this sort where generative embedding may prove maximally useful. Rather, we want to ask whether substructures would emerge even in the absence of *a priori* knowledge about their existence.

Model-based clustering of all subjects

Using the variational clustering algorithm described above, we performed a model-based clustering analysis of all subjects, based on their posterior parameter estimates. The purpose of this first clustering analysis was to assess whether our model would be sufficiently sensitive to detect differences between patients and healthy controls in the absence of any *a priori* knowledge about the very existence of these two groups, let alone their clinical validity.

We found that the highest model evidence was obtained by a Gaussian mixture model with two clusters (Figure 6.5b). This model outperformed the next-best model by a log Bayes factor (BF) of 66.0, providing very strong evidence that a model with two clusters provided the best explanation of the data within our hypothesis class of Gaussian mixture models.

Asking to what degree the best clustering solution matched known structure in the data, we obtained a balanced purity of 71% with regard to the known diagnostic distinction between schizophrenia and healthy controls. In other words, without any *a priori* knowledge about the existence of schizophrenia among the group of participants, the above clustering analysis concluded that there are two groups in the data, and, notably, these groups largely matched the difference between healthy participants and those diagnosed with the disease.

This result is reassuring; however, it ultimately only provides the confirmation of a diagnostic category that is already known and that can be obtained much easier by means of a conventional clinical questionnaire. This motivated the final analysis, described next, in which we focused exclusively on the group of patients, leaving aside healthy controls. In this analysis, we asked whether model-based clustering would yield substructures which are not yet captured by current diagnostic schemes and which might have been masked by the more salient difference between patients and healthy controls in the clustering analysis above.

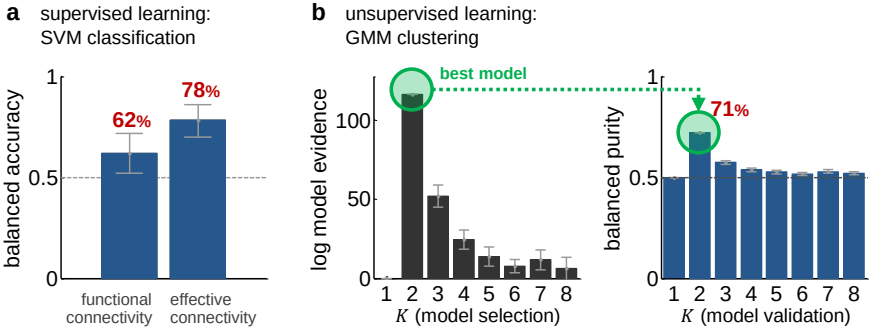

Figure 6.5: Model-based classification and clustering results. (a) In an initial supervised approach, we compared a model-based classification approach (based on parameter estimates of a DCM) with a more conventional approach (based on undirected estimates of functional connectivity). With an accuracy of 78%, generative embedding outperformed this alternative approach and enabled a clear separation between participants diagnosed with schizophrenia and healthy controls. (b) In a subsequent unsupervised approach, a Gaussian mixture model (GMM) was used to evaluate what structure emerged in the absence of any *a priori* knowledge about diagnostic status. The highest evidence was obtained for a model with two clusters. Notably, the clustering implied by this model matched known diagnostic categories with a purity of 71%, which is a rather competitive degree of discrimination even in relation to supervised approaches.

Model-based clustering of patients

Using model-based clustering on the group of participants diagnosed with schizophrenia, we obtained the highest model evidence for a clustering solution with three clusters (Figure 6.6a). There was very strong evidence that this solution was better than the next best model which contained only two clusters (log BF = 29.1). This result indicates that the absence of healthy controls may indeed have unmasked a more subtle distinction among patients, disclosing a group structure with three distinct clusters.

Since the clusters were identified in a DCM-based generative score space, differences between clusters can be examined in terms of their implied dynamic systems models. Here, we investigated the structure of the DCM corresponding to each posterior cluster mean (Figure 6.6b).

Before returning to the interpretation of these clusters, we examined their potential clinical validity. We carried out a one-way ANOVA separately for two clinical variables of interest. The first one encoded chlorpromazine equivalents (CPZ) which represent a drug-independent measure of

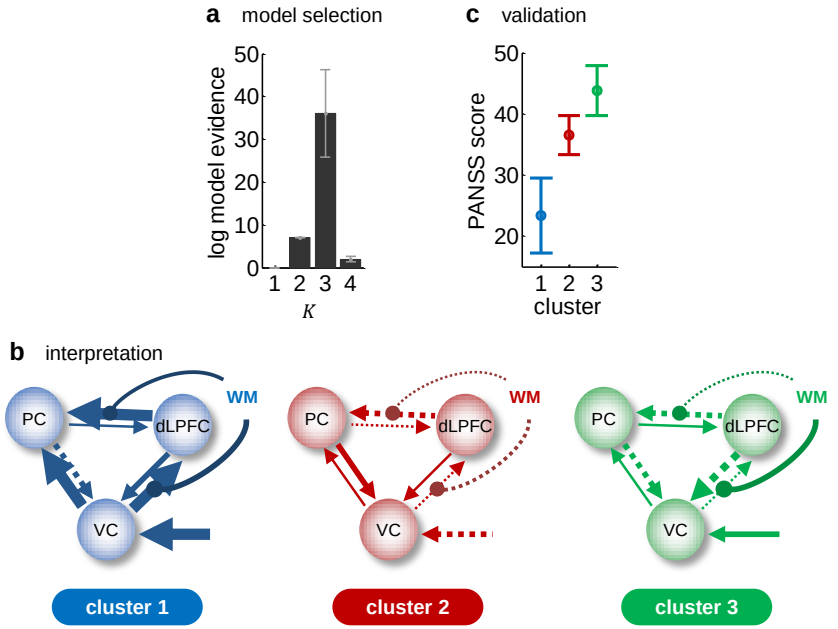

Figure 6.6: Model-based clustering results on patients. (a) When focusing on the group of patients diagnosed with schizophrenia, the highest model evidence is obtained for a model with three clusters. (b) The centroid of each cluster can be interpreted in terms of the underlying system. (c) Notably, the three clusters identified here differ significantly in terms of the positive and negative symptom score (PANSS) of schizophrenia (one-way ANOVA, $p < 0.05$).

medication. The second variable encoded scores on the traditional 'positive and negative symptom scale' (PANSS; Kay *et al.*, 1987) which represents a measure of symptom severity in patients with schizophrenia.

No significant differences in CPZ equivalents were found between clusters. By contrast, and this result was striking, PANSS scores differed significantly between clusters ($p < 0.05$). Very simply speaking, cluster 1 comprised patients with symptom scores between 20 and 30; cluster 2 contained those with scores between 30 and 40; and cluster 3 those between 40 and 50 (Figure 6.6c). One might have decided to split patients across these boundaries based on intuition or experience; but here both the number of subgroups and their exact boundaries emerged automatically, based on salient mechanistic differences in effective connectivity.

These differences between clusters disclosed potentially interesting mechanisms. Patients in the first cluster were characterized by strong effective connectivity throughout, with medium modulation by working-memory task demands. Patients in the second cluster, in contrast, displayed much lower connectivity on average and only a small amount of modulatory influences. In the third cluster, finally, connectivity was fairly low throughout except for the influence that activity in the visual cortex exerted onto the dorsolateral prefrontal cortex which was strongly modulated by working memory (cf. Figure 6.6b).

In summary, there was no obvious linear relationship between model parameters and external clinical scores. Rather, the clustering algorithm had picked up patterns of model parameters which seemed to exhibit a nonlinear relationship with clinical symptom scores and which delineated these into three groups. To avoid repetition, we will evaluate the implications of this initial proof of concept in conjunction with an overall discussion of the material presented in this thesis in the final chapter.

Chapter 7

Conclusions

Generative embedding provides several characteristics that set it apart from conventional analyses of high-dimensional time series, as discussed below.

Model-based classification

Generative embedding can be used as a basis for model-based classification algorithms. When used in this way, the approach combines the explanatory strengths of generative models with the classification power of discriminative classifiers. Thus, in contrast to purely discriminative or purely generative methods, generative embedding is a hybrid approach. It fuses a feature space capturing both the data and the underlying generative process with a classifier that finds, for instance, the maximum-margin boundary of class separation.

It is possible to define a problem-specific kernel and combine it with a general-purpose algorithm for discrimination. This makes our approach modular and widely applicable, for instance to different acquisition modalities.

Intuitively, our results exploit the idea that differences in the generative process between two examples (observations) provide very rich discriminative information for accurate predictions. In the case of DCM for fMRI, this rationale should pay off whenever the directed connection strengths between brain regions contain more information about a disease state than regional activations or undirected correlations.

This is indeed what we found in our initial analysis of stroke patients

(Figure 5.21). It is also what we found in our subsequent analysis of patients diagnosed with schizophrenia (Figure 6.5a). It has long been suspected that a DCM-informed data representation might prove particularly relevant in psychiatric disorders, such as schizophrenia or depression, where aberrant effective connectivity and synaptic plasticity are central to the disease process (Castren, 2005; Stephan *et al.*, 2009b). Our results provide support for this hypothesis.

Interpretability

Generative embedding enables a mechanistic and intuitive interpretation of features and their weights, an important property not afforded by most conventional classification methods (Lao *et al.*, 2004; Thomaz *et al.*, 2004). By using parameter estimates from a mechanistically interpretable model for constructing a feature space, the subsequent classification no longer yields 'black box' results but allows one to assess the relative importance of different mechanisms for distinguishing groups (e.g., whether or not synaptic plasticity alters the strengths of certain connections in a particular context).

Put differently, generative embedding embodies a shift in perspective: rather than representing sequential data in terms of high-dimensional and potentially extremely noisy trajectories, we are viewing the data in terms of the coefficients of a much more well-behaved model of system dynamics.

In this sense, models like DCM, when used in the context of generative embedding, turn the curse of dimensionality faced by conventional classification methods into a blessing: the higher the spatial and temporal resolution of the underlying fMRI data, the more precise the resulting DCM parameter estimates, leading to more accurate predictions.

It is also worth pointing out that generative embedding does not need to rest on a generative model of *brain activity*. Other types of data could be represented using the exact same techniques. In particular, given a generative model of *behaviour*, one could create a generative score space in which each feature corresponds to a model parameter that represents a particular latent feature of reasoning, learning, or decision making. We previously suggested an example for a generative model of this sort in the domain of decision making (Brodersen *et al.*, 2008). This model combined computational aspects of a Bayesian learner with the potential physiological mechanisms supporting these computations. It thus represented an example of what one might view as a *neurocomputational* model. The model was sufficiently descriptive to provide a unique characterization of all three

pilot subjects participating in the study. The more recent literature in this domain includes the models by Daunizeau *et al.* (2010) and Mathys *et al.* (2011). Using such models, generative embedding might provide a strategy for coercing subject-specific model parameters into neurocomputational fingerprints which can then be submitted to group analyses.

Model-based clustering

It is envisaged that an increasingly relevant facet of generative embedding will be its utility for model-based clustering. Specifically, when a reliable and meaningful grouping structure has not yet been established, model-based clustering approaches, such as the one presented in Chapter 6, make it possible to generate an initial hypothesis about subgroups. Any such hypothesis can subsequently be validated against an external variable representing, for instance, known clinical facts. Because the approach is based on a biophysical model of the data, it is conceivable that this approach has a higher likelihood of leading to clinically valid clusters.

Model comparison

Finally, generative embedding is tightly related to questions of model comparison. For any given dataset, there is an infinite number of possible dynamic causal models, differing in the number and location of nodes, in connectivity structure, and in their parameterization (e.g., priors). Competing models can be compared using Bayesian model selection (BMS; Penny *et al.*, 2004; Stephan *et al.*, 2007a; Friston *et al.*, 2007; Stephan *et al.*, 2009a), where the best model is the one with the highest model evidence, that is, the highest probability of the data given the model (MacKay, 1992). However, there are two scenarios in which BMS is problematic and where classification based on generative embedding may represent a useful alternative.

First, BMS requires the data to be identical for all competing models. Thus, in the case of current implementations of DCM for fMRI, BMS enables dynamic model selection (concerning the parameterization and mathematical form of the model equations) but not structural model selection (concerning which regions or nodes should be included in the model).

Second, BMS is limited when different groups cannot be mapped onto different model structures, for example when the differences in neuronal mechanisms operate at a finer conceptual scale than can be represented within the chosen modelling framework. In this case, discriminability of

subjects may be afforded by differences in (combinations of) parameter estimates under the same model structure (see Allen *et al.*, 2010, for a recent example).

In both these scenarios, the approach proposed in this thesis may provide a solution, in that the unsupervised creation of a generative score space can be viewed as a method for biologically informed feature extraction, and classification accuracy or clustering purity can be viewed as measures of how much class information is encoded in the model parameters. This view enables a form of model comparison in which the best model is the one that enables the highest discriminability (cf. Figure 5.27).

Notably, this procedure can be applied even when (i) the underlying data (e.g., the chosen regional fMRI time series) are different, or when (ii) the difference between two models lies exclusively in the pattern of parameter estimates. In this thesis, we have illustrated both ideas: structural model selection to decide between a full model and two reduced models that disregard one hemisphere; and dynamic model selection to distinguish between different groups of subjects under the same model structure.

In summary, BMS evaluates the goodness of a model with regard to its generalizability for explaining the data, whereas generative embedding evaluates a model in relation to an external criterion, i.e., how well it allows for inference on group membership of any given subject. This difference is important as it highlights that the concept of a 'good' model can be based on fundamentally different aspects, and one could imagine scenarios where BMS and generative embedding arrive at opposing results. If, for example, discriminability of groups relies on a small subspace of the data, then one model (which provides a good accuracy-complexity trade-off for most of the data except that subspace) may have higher evidence, but another model that describes this subspace particularly well but is generally worse for the rest of the data may result in better classification performance or clustering purity. We will examine the relation and complementary nature of BMS and generative-embedding approaches in future work.

Inference on mechanisms for clinical applications

We hope that the approach presented in this thesis will be useful for addressing clinical problems of high practical relevance, for instance for dissecting psychiatric spectrum disorders, such as schizophrenia, into physiologically defined subgroups (Stephan *et al.*, 2009b), or for predicting the response of individual patients to specific drugs. While an increasing number of stud-

ies have tried to describe neurobiological markers for psychiatric disorders (Davatzikos *et al.*, 2005, 2008a,b; Fu *et al.*, 2008; Misra *et al.*, 2009; Nenadic *et al.*, 2010; Klöppel *et al.*, 2008, 2009, 2012), we argue that these studies should be complemented by model-based approaches for inferring biologically plausible mechanisms.

These could become useful in three ways: (i) to generate clinical hypotheses by dissecting a group of patients with similar symptoms into mechanistically defined subgroups; (ii) to harvest the potentially rich discriminative information encoded in aspects of synaptic plasticity or neuromodulation to build classifiers which, perhaps in conjunction with behavioural learning paradigms and pharmacological challenges, distinguish between different subtypes of a psychiatric disorder on a physiological basis; and (iii) to decide between competing hypotheses about neural mechanisms, based on the model evidence, on classification accuracy, or on clustering purity.

In the case of the two fMRI datasets analysed in this thesis, generative embedding yielded stronger classification performance than conventional methods, whether they were based on activations or regional correlations. One might at first imagine that this superior ability to accurately group individual subjects determines the clinical value of the approach. Instead, we argue that its clinical value will ultimately depend on whether patients that share the same symptoms can be differentially treated according to the underlying pathophysiology of the disorder. Generative embedding, using biologically plausible and mechanistically interpretable models, may prove critical in establishing diagnostic classification schemes that distinguish between pathophysiologically distinct subtypes of spectrum diseases and allow for predicting individualized behavioural and pharmacological therapy.

Statistical paradigms

Methodologically, the research presented in this thesis began with elements of classical parametric statistics (e.g., the general linear model and t-tests for univariate feature selection), nonparametric statistics (e.g., permutation tests for feature interpretation), and statistical learning theory (e.g., kernels and support vector machines).

The more recent building blocks, in contrast, were based on Bayesian inference with its elements of probabilistic graphical models (e.g., mixed-effects inference), stochastic inference methods (e.g., MCMC) and variational approximations (e.g., VB). These methods have proven extremely powerful, flexible, and efficient. Our future research will continue to take

full advantage of the Bayesian paradigm and increasingly move from algorithmic pipelines of independent building blocks towards integrated models and their efficient inversion.

References

Agresti, A. and Hitchcock, D. B. (2005). Bayesian inference for categorical data analysis. *Statistical Methods and Applications*, **14**(3), 297–330.

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

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

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

Allen, P., Stephan, K. E., Mechelli, A., Day, F., Ward, N., Dalton, J., Williams, S. C., and McGuire, P. (2010). Cingulate activity and fronto-temporal connectivity in people with prodromal signs of psychosis. *NeuroImage*, **49**(1), 947–955.

Baldeweg, T. (2006). Repetition effects to sounds: evidence for predictive coding in the auditory system. *Trends in Cognitive Sciences*, **10**(3), 93–94. PMID: 16460994.

Bayes, T. and Price, R. (1763). An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society*, **53**, 370–418.

Beal, M. (2003). *Variational algorithms for approximate Bayesian inference*. Ph.D. thesis, University College London, United Kingdom.

Beckmann, C. F., Jenkinson, M., and Smith, S. M. (2003). General multilevel linear modeling for group analysis in fMRI. *NeuroImage*, **20**(2), 1052–1063.

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

Ben-Hur, A., Ong, C. S., Sonnenburg, S., Schölkopf, B., and Rätsch, G. (2008). Support vector machines and kernels for computational biology. *PLoS Comput Biol*, **4**(10), e1000173.

Berry, D. A. and Christensen, R. (1979). Empirical Bayes estimation of a binomial parameter via mixtures of Dirichlet processes. *The Annals of Statistics*, **7**(3), 558–568.

Bicego, M., Murino, V., and Figueiredo, M. A. (2004). Similarity-based classification of sequences using hidden Markov models. *Pattern Recognition*, **37**(12), 2281–2291.

Bicego, M., Cristani, M., Murino, V., Pekalska, E., and Duin, R. (2009a). Clustering-based construction of hidden Markov models for generative kernels. In *Proceedings of the 7th International Conference on Energy Minimization Methods in Computer Vision and Pattern Recognition*.

Bicego, M., Pekalska, E., Tax, D. M., and Duin, R. P. (2009b). Component-based discriminative classification for hidden Markov models. *Pattern Recognition*, **42**(11), 2637–2648.

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

Blankertz, B., Lemm, S., Treder, M., Haufe, S., and Müller, K.-R. (2011). Single-trial analysis and classification of ERP components - a tutorial. *NeuroImage*, **15**(56), 814–25.

Bles, M. and Haynes, J.-D. (2008). Detecting concealed information using brain-imaging technology. *Neurocase*, **14**(1), 82–92. PMID: 18569734.

Bode, S. and Haynes, J.-D. (2009). Decoding sequential stages of task preparation in the human brain. *NeuroImage*, **45**(2), 606–613.

Bosch, A., Zisserman, A., and Munoz, X. (2006). Scene classification via pLSA. In *ECCV*, pages 517–530.

Bosch, A., Zisserman, A., Pujol, M., *et al.* (2008). Scene classification using a hybrid generative/discriminative approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 2008, vol. 30, p. 712-727*.

Boser, B. E., Guyon, I. M., and Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. In *Proceedings of the 5th Annual Workshop on Computational Learning Theory*, pages 144–152.

Brodersen, K. H., Penny, W. D., Harrison, L. M., Daunizeau, J., Ruff, C. C., Duzel, E., Friston, K. J., and Stephan, K. E. (2008). Integrated Bayesian models of learning and decision making for saccadic eye movements. *Neural Networks*, **21**(9), 1247–1260.

Brodersen, K. H., Ong, C. S., Stephan, K. E., and Buhmann, J. M. (2010a). The balanced accuracy and its posterior distribution. In *Proceedings of the 20th International Conference on Pattern Recognition*, pages 3121–3124. IEEE Computer Society.

Brodersen, K. H., Ong, C. S., Stephan, K. E., and Buhmann, J. M. (2010b). The binormal assumption on precision-recall curves. In *Proceedings of the 20th International Conference on Pattern Recognition*, pages 4263–4266. IEEE Computer Society.

Brodersen, K. H., Schofield, T. M., Leff, A. P., Ong, C. S., Lomakina, E. I., Buhmann, J. M., and Stephan, K. E. (2011a). Generative embedding for model-based classification of fMRI data. *PLoS Comput Biol*, **7**(6), e1002079.

Brodersen, K. H., Haiss, F., Ong, C. S., Jung, F., Tittgemeyer, M., Buhmann, J. M., Weber, B., and Stephan, K. E. (2011b). Model-based feature construction for multivariate decoding. *NeuroImage*, **56**(2), 601–615.

Brodersen, K. H., Wiech, K., Lomakina, E. I., Lin, C.-S., Buhmann, J. M., Bingel, U., Ploner, M., Stephan, K. E., and Tracey, I. (2012). Decoding the perception of pain from fMRI using multivariate pattern analysis. *NeuroImage*, **63**, 1162–1170.

Brodersen, K. H., Lin, Z., Gupta, A., Deserno, L., Penny, W. D., Schlagenhauf, F., Buhmann, J. M., and Stephan, K. E. (*in preparation*). Model-based clustering.

Brodersen, K. H., Mathys, C., Chumbley, J. R., Daunizeau, J., Ong, C., Buhmann, J. M., and Stephan, K. E. (*in press*). Mixed-effects inference on classification performance in hierarchical datasets. *Journal of Machine Learning Research*.

Brodersen, K. H., Daunizeau, J., Mathys, C., Chumbley, J. R., Buhmann, J. M., and Stephan, K. E. (*under review*). Variational Bayesian mixed-effects inference for classification studies.

Caron, F. and Doucet, A. (2008). Sparse Bayesian nonparametric regression. In *Proceedings of the 25th international conference on Machine learning*, pages 88–95, Helsinki, Finland. ACM.

Castren, E. (2005). Is mood chemistry? *Nature Reviews Neuroscience*, **6**(3), 241–246.

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

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

Chen, C. C., Kiebel, S. J., and Friston, K. J. (2008). Dynamic causal modelling of induced responses. *NeuroImage*, **41**(4), 1293–1312. PMID: 18485744.

Chu, C., Ni, Y., Tan, G., Saunders, C. J., and Ashburner, J. (2010). Kernel regression for fMRI pattern prediction. *NeuroImage*. PMID: 20348000.

Cooper, G. F. and Herskovits, E. (1992). A Bayesian method for the induction of probabilistic networks from data. *Machine Learning*, **9**(4), 309–347.

Cormack, G. V. (2008). Email spam filtering: a systematic review. *Foundations and Trends in Information Retrieval*, **1**(4), 335–455.

Cox, D. D. and 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**(2), 261–270.

Craddock, R. C., III, P. E. H., Hu, X. P., and Mayberg, H. S. (2009). Disease state prediction from resting state functional connectivity. *Magnetic Resonance in Medicine*, **62**(6), 1619–1628.

Cuturi, M., Fukumizu, K., and Vert, J. (2006). Semigroup kernels on measures. *Journal of Machine Learning Research*, **6**(2), 1169.

Daunizeau, J., Kiebel, S. J., and Friston, K. J. (2009). Dynamic causal modelling of distributed electromagnetic responses. *NeuroImage*, **47**(2), 590–601. PMID: 19398015.

Daunizeau, J., den Ouden, H. E. M., Pessiglione, M., Kiebel, S. J., Stephan, K. E., and Friston, K. J. (2010). Observing the observer (I): meta-bayesian models of learning and decision-making. *PLoS ONE*, **5**(12), e15554.

Davatzikos, C., Ruparel, K., Fan, Y., Shen, D., Acharyya, M., Loughead, J., Gur, R., and Langleben, D. (2005). Classifying spatial patterns of brain activity with machine learning methods: Application to lie detection. *NeuroImage*, **28**(3), 663–668.

Davatzikos, C., Fan, Y., Wu, X., Shen, D., and Resnick, S. M. (2008a). Detection of prodromal Alzheimer's disease via pattern classification of magnetic resonance imaging. *Neurobiology of Aging*, **29**(4), 514–523.

Davatzikos, C., Resnick, S., Wu, X., Parmpi, P., and Clark, C. (2008b). Individual patient diagnosis of AD and FTD via high-dimensional pattern classification of MRI. *NeuroImage*, **41**(4), 1220–1227.

David, O. and Friston, K. J. (2003). A neural mass model for MEG/EEG: coupling and neuronal dynamics. *NeuroImage*, **20**(3), 1743–1755.

David, O., Kiebel, S. J., Harrison, L. M., Mattout, J., Kilner, J. M., and Friston, K. J. (2006). Dynamic causal modeling of evoked responses in EEG and MEG. *NeuroImage*, **30**(4), 1255–1272.

Deely, J. J. and Lindley, D. V. (1981). Bayes empirical Bayes. *Journal of the American Statistical Association*, **76**(376), 833–841.

Dehaene, S., Pegado, F., Braga, L. W., Ventura, P., Nunes Filho, G., Jobert, A., Dehaene-Lambertz, G., Kolinsky, R., Morais, J., and Cohen, L. (2010). How learning to read changes the cortical networks for vision

and language. *Science (New York, N.Y.)*, **330**(6009), 1359–1364. PMID: 21071632.

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

den Ouden, H. E. M., Daunizeau, J., Roiser, J., Friston, K. J., and Stephan, K. E. (2010). Striatal prediction error modulates cortical coupling. *Journal of Neuroscience*, **30**(9), 3210–3219. PMID: 20203180.

Deserno, L., Sterzer, P., Wüstenberg, T., Heinz, A., and Schlagenhauf, F. (2012). Reduced prefrontal-parietal effective connectivity and working memory deficits in schizophrenia. *The Journal of Neuroscience*, **32**(1), 12–20.

Dixon, P. (2008). Models of accuracy in repeated-measures designs. *Journal of Memory and Language*, **59**(4), 447–456.

Dubey, A., Hwang, S., Rangel, C., Rasmussen, C. E., Ghahramani, Z., and Wild, D. L. (2004). Clustering protein sequence and structure space with infinite gaussian mixture models. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, pages 399–410. PMID: 14992520.

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

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

Fan, Y., Rao, H., Hurt, H., Giannetta, J., Korczykowski, M., Shera, D., Avants, B. B., Gee, J. C., Wang, J., and Shen, D. (2007). Multivariate examination of brain abnormality using both structural and functional MRI. *NeuroImage*, **36**(4), 1189–1199.

Fan, Y., Resnick, S. M., Wu, X., and Davatzikos, C. (2008a). Structural and functional biomarkers of prodromal Alzheimer's disease: A high-dimensional pattern classification study. *NeuroImage*, **41**(2), 277–285.

Fan, Y., Gur, R. E., Gur, R. C., Wu, X., Shen, D., Calkins, M. E., and Davatzikos, C. (2008b). Unaffected family members and schizophrenia patients share brain structure patterns: A high-dimensional pattern classification study. *Biological Psychiatry*, **63**(1), 118–124.

Fazli, S., Danoczy, M., Schelldorfer, J., and Müller, K.-R. (2011). L1-penalized linear mixed-effects models for high dimensional data with application to BCI. *NeuroImage*, **56**(4), 2100–2108.

Felleman, D. J. and van Essen, D. C. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex (New York, N.Y.: 1991)*, **1**(1), 1–47. PMID: 1822724.

Fisher, R. A. (1915). Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika*, **10**(4), 507–521.

Ford, J., Farid, H., Makedon, F., Flashman, L. A., McAllister, T. W., Megalooikonomou, V., and Saykin, A. J. (2003). Patient classification of fMRI activation maps. In *Medical Image Computing and Computer-Assisted Intervention*, pages 58–65. MICCAI.

Formisano, E., De Martino, F., Bonte, M., and Goebel, R. (2008). Who is saying what? Brain-based decoding of human voice and speech. *Science*, **322**(5903), 970–973.

Friston, K. (2009). Dynamic causal modeling and Granger causality. comments on: The identification of interacting networks in the brain using fMRI: model selection, causality and deconvolution. *NeuroImage*, **58**(2), 303–305.

Friston, K., Mattout, J., Trujillo-Barreto, N., Ashburner, J., and Penny, W. (2007). Variational free energy and the Laplace approximation. *NeuroImage*, **34**(1), 220–234. PMID: 17055746.

Friston, K., Chu, C., Mourao-Miranda, J., Hulme, O., Rees, G., Penny, W., and Ashburner, J. (2008). Bayesian decoding of brain images. *NeuroImage*, **39**(1), 181–205.

Friston, K. J. (2002). Bayesian estimation of dynamical systems: An application to fMRI. *NeuroImage*, **16**(2), 513–530. PMID: 12030834.

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

Friston, K. J., Holmes, A. P., and Worsley, K. J. (1999). How many subjects constitute a study? *NeuroImage*, **10**, 1–5.

Friston, K. J., Harrison, L., and Penny, W. (2003). Dynamic causal modelling. *NeuroImage*, **19**(4), 1273–1302.

Friston, K. J., Stephan, K. E., Lund, T. E., Morcom, A., and Kiebel, S. (2005). Mixed-effects and fMRI studies. *NeuroImage*, **24**(1), 244–252.

Fu, C. H., Mourao-Miranda, J., Costafreda, S. G., Khanna, A., Marquand, A. F., Williams, S. C., and Brammer, M. J. (2008). Pattern classification of sad facial processing: Toward the development of neurobiological markers in depression. *Biological Psychiatry*, **63**(7), 656–662.

Garrido, M. I., Friston, K. J., Kiebel, S. J., Stephan, K. E., Baldeweg, T., and Kilner, J. M. (2008). The functional anatomy of the MMN: a DCM study of the roving paradigm. *NeuroImage*, **42**(2), 936–944. PMID: 18602841.

Garrido, M. I., Kilner, J. M., Stephan, K. E., and Friston, K. J. (2009). The mismatch negativity: a review of underlying mechanisms. *Clinical Neurophysiology*, **120**(3), 453–463. PMID: 19181570.

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

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

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

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

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

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

Grosenick, L., Greer, S., and Knutson, B. (2008). Interpretable classifiers for fMRI improve prediction of purchases. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, **16**(6), 539–548.

Grosenick, L., Klingenberg, B., Greer, S., Taylor, J., and Knutson, B. (2009). Whole-brain sparse penalized discriminant analysis for predicting choice. *NeuroImage*, **47**(Supplement 1), S58.

Gustafsson, M. G., Wallman, M., Wickenberg Bolin, U., Göransson, H., Fryknäs, M., Andersson, C. R., and Isaksson, A. (2010). Improving Bayesian credibility intervals for classifier error rates using maximum entropy empirical priors. *Artificial Intelligence in Medicine*, **49**(2), 93–104.

Hampton, A. N. and O'Doherty, J. P. (2007). Decoding the neural substrates of reward-related decision making with functional MRI. *Proceedings of the National Academy of Sciences*, **104**(4), 1377–1382.

Harrison, S. A. and Tong, F. (2009). Decoding reveals the contents of visual working memory in early visual areas. *Nature*, **458**(7238), 632–635.

Hassabis, D., Chu, C., Rees, G., Weiskopf, N., Molyneux, P. D., and Maguire, E. A. (2009). Decoding neuronal ensembles in the human hippocampus. *Current Biology*, **19**(7), 546–554. PMID: 19285400.

Haussler, D. (1999). Convolution kernels on discrete structures. Technical report.

Haynes, J.-D. and Rees, G. (2005). Predicting the orientation of invisible stimuli from activity in human primary visual cortex. *Nature Neuroscience*, **8**(5), 686–691.

Haynes, J.-D. and Rees, G. (2006). Decoding mental states from brain activity in humans. *Nature Reviews Neuroscience*, **7**(7), 523–534.

Haynes, J.-D., Sakai, K., Rees, G., Gilbert, S., Frith, C., and Passingham, R. E. (2007). Reading hidden intentions in the human brain. *Current Biology*, **17**(4), 323–328.

Hein, M. and Bousquet, O. (2005). Hilbertian metrics and positive definite kernels on probability measures. In *Proceedings of AISTATS*, volume 2005.

Hofmann, T. (2000). Learning the similarity of documents: An information-geometric approach to document retrieval and categorization. In *Advances in Neural Information Processing Systems*, volume 12, pages 914–920.

Holmes, A. P. and Friston, K. J. (1998). Generalisability, random effects and population inference. *Fourth Int. Conf. on Functional Mapping of the Human Brain, NeuroImage*, **7**, S754.

Holub, A. D., Welling, M., and Perona, P. (2005). Combining generative models and fisher kernels for object recognition. In *IEEE International Conference on Computer Vision*, volume 1, pages 136–143, Los Alamitos, CA, USA. IEEE Computer Society.

Howard, J. D., Plailly, J., Grueschow, M., Haynes, J.-D., and Gottfried, J. A. (2009). Odor quality coding and categorization in human posterior piriform cortex. *Nature Neuroscience*, **12**(7), 932–938.

Iwata, T., Duvenaud, D., and Ghahramani, Z. (2012). Nonparametric bayesian clustering via infinite warped mixture models. Technical Report arXiv:1206.1846 [stat.ML], University of Cambridge, Cambridge, UK.

Jaakkola, T., Diekhans, M., and Haussler, D. (1999). Using the Fisher kernel method to detect remote protein homologies. In *Proceedings of the 7th International Conference on Intelligent Systems for Molecular Biology*, pages 149–158.

Jaakkola, T. S. and Haussler, D. (1999). Exploiting generative models in discriminative classifiers. In *Advances in Neural Information Processing Systems*, pages 487–493.

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

Jansen, B. H. and Rit, V. G. (1995). Electroencephalogram and visual evoked potential generation in a mathematical model of coupled cortical columns. *Biological Cybernetics*, **73**(4), 357–366. PMID: 7578475.

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

Jebara, T., Kondor, R., and Howard, A. (2004). Probability product kernels. *Journal of Machine Learning Research*, **5**, 844.

Jensen, F. V. and Nielsen, T. D. (2007). *Bayesian networks and decision graphs*. Springer.

Jung, F., Tittgemeyer, M., Kumagai, T., Moran, R., Stephan, K. E., Endepols, H., and Graf, R. (2009). Detection of auditory evoked potentials and mismatch negativity-like responses in the awake and unrestrained rat.

Kamitani, Y. and Tong, F. (2005). Decoding the visual and subjective contents of the human brain. *Nature Neuroscience*, **8**(5), 679–685.

Kamitani, Y. and Tong, F. (2006). Decoding seen and attended motion directions from activity in the human visual cortex. *Current Biology*, **16**(11), 1096–1102.

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

Kay, K. N., Naselaris, T., Prenger, R. J., and Gallant, J. L. (2008). Identifying natural images from human brain activity. *Nature*, **452**(7185), 352–355.

Kay, S. R., Fiszbein, A., and Opler, L. A. (1987). The positive and negative syndrome scale (PANSS) for schizophrenia. *Schizophrenia Bulletin*, **13**(2), 261–276.

Kiebel, S. J., Garrido, M. I., and Friston, K. J. (2007). Dynamic causal modelling of evoked responses: The role of intrinsic connections. *NeuroImage*, **36**(2), 332–345.

Kiebel, S. J., Garrido, M. I., Moran, R., Chen, C.-C., and Friston, K. J. (2009). Dynamic causal modeling for EEG and MEG. *Human Brain Mapping*, **30**(6), 1866–1876. PMID: 19360734.

Klöppel, S., Stonnington, C. M., Chu, C., Draganski, B., Scahill, R. I., Rohrer, J. D., Fox, N. C., Jack, C. R., Ashburner, J., and Frackowiak, R. S. J. (2008). Automatic classification of MR scans in Alzheimer's disease. *Brain*, **131**(Pt 3), 681–689. PMID: 18202106 PMCID: 2579744.

Klöppel, S., Chu, C., Tan, G. C., Draganski, B., Johnson, H., Paulsen, J. S., Kienzle, W., Tabrizi, S. J., Ashburner, J., and Frackowiak, R. S. J. (2009). Automatic detection of preclinical neurodegeneration: Presymptomatic Huntington disease. *Neurology*, **72**(5), 426–431. PMID: 19188573.

Klöppel, S., Abdulkadir, A., Jack Jr., C. R., Koutsouleris, N., Mourão-Miranda, J., and Vemuri, P. (2012). Diagnostic neuroimaging across diseases. *NeuroImage*, **61**(2), 457–463.

Knops, A., Thirion, B., Hubbard, E. M., Michel, V., and Dehaene, S. (2009). Recruitment of an area involved in eye movements during mental arithmetic. *Science (New York, N.Y.)*, **324**(5934), 1583–1585.

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

Koutsouleris, N., Meisenzahl, E. M., Davatzikos, C., Bottlender, R., Frodl, T., Scheuerecker, J., Schmitt, G., Zetzsche, T., Decker, P., Reiser, M., Möller, H.-J., and Gaser, C. (2009). Use of neuroanatomical pattern classification to identify subjects in at-risk mental states of psychosis and predict disease transition. *Archives of General Psychiatry*, **66**(7), 700–712. PMID: 19581561.

Kozel, F. A., Johnson, K. A., Mu, Q., Grenesko, E. L., Laken, S. J., and George, M. S. (2005). Detecting deception using functional magnetic resonance imaging. *Biological Psychiatry*, **58**(8), 605–613.

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

Kriegeskorte, N., Goebel, R., and Bandettini, P. (2006). Information-based functional brain mapping. *Proceedings of the National Academy of Sciences*, **103**(10), 3863–3868.

Kriegeskorte, N., Formisano, E., Sorger, B., and Goebel, R. (2007). Individual faces elicit distinct response patterns in human anterior temporal cortex. *Proceedings of the National Academy of Sciences*, **104**(51), 20600–20605.

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

Lao, Z., Shen, D., Xue, Z., Karacali, B., Resnick, S. M., and Davatzikos, C. (2004). Morphological classification of brains via high-dimensional shape transformations and machine learning methods. *NeuroImage*, **21**(1), 46–57.

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

Lasserre, J. A., Bishop, C. M., and Minka, T. P. (2006). Principled hybrids of generative and discriminative models. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, pages 87–94, Los Alamitos, CA, USA. IEEE Computer Society.

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

Leff, A. P., Schofield, T. M., Stephan, K. E., Crinion, J. T., Friston, K. J., and Price, C. J. (2008). The cortical dynamics of intelligible speech. *Journal of Neuroscience*, **28**(49), 13209–13215. PMID: 19052212.

Leonard, T. (1972). Bayesian methods for binomial data. *Biometrika*, **59**(3), 581 –589.

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

Madigan, D. and Raftery, A. E. (1994). Model selection and accounting for model uncertainty in graphical models using Occam's window. *Journal of the American Statistical Association*, **89**(428), 1535–1546.

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

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

Martins, A., Bicego, M., Murino, V., Aguiar, P., and Figueiredo, M. (2010). Information theoretical kernels for generative embeddings based on hidden markov models. *Structural, Syntactic, and Statistical Pattern Recognition*, pages 463–472.

Mathys, C., Daunizeau, J., Friston, K. J., and Stephan, K. E. (2011). A Bayesian foundation for individual learning under uncertainty. *Frontiers in Human Neuroscience*, **5**. PMID: 21629826 PMCID: PMC3096853.

Mercer, J. (1909). Functions of positive and negative type, and their connection with the theory of integral equations. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, **209**, 415–446.

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

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

Minka, T. (2005). Discriminative models, not discriminative training. Technical Report MSR-TR-2005-144, Microsoft Research.

Misra, C., Fan, Y., and Davatzikos, C. (2009). Baseline and longitudinal patterns of brain atrophy in MCI patients, and their use in prediction of short-term conversion to AD: results from ADNI. *NeuroImage*, **44**(4), 1415–1422.

Mitchell, T. M., Hutchinson, R., Just, M. A., Niculescu, R. S., Pereira, F., and Wang, X. (2003). Classifying instantaneous cognitive states from fMRI data. *Annual Symposium Proceedings*, **2003**, 465–469. PMC1479944.

Mitchell, T. M., Shinkareva, S. V., Carlson, A., Chang, K.-M., Malave, V. L., Mason, R. A., and Just, M. A. (2008). Predicting human brain activity associated with the meanings of nouns. *Science*, **320**(5880), 1191–1195.

Miyawaki, Y., Uchida, H., Yamashita, O., Sato, M.-a., Morito, Y., Tanabe, H. C., Sadato, N., and Kamitani, Y. (2008). Visual image reconstruction from human brain activity using a combination of multiscale local image decoders. *Neuron*, **60**(5), 915–929.

Moran, R., Stephan, K., Kiebel, S., Rombach, N., O'Connor, W., Murphy, K., Reilly, R., and Friston, K. (2008). Bayesian estimation of synaptic physiology from the spectral responses of neural masses. *NeuroImage*, **42**(1), 272–284.

Moran, R. J., Stephan, K. E., Seidenbecher, T., Pape, H.-C., Dolan, R. J., and Friston, K. J. (2009). Dynamic causal models of steady-state responses. *NeuroImage*, **44**(3), 796–811. PMID: 19000769.

Mourao-Miranda, J., Bokde, A., Born, C., Hampel, H., and Stetter, M. (2005). Classifying brain states and determining the discriminating activation patterns: Support vector machine on functional MRI data. *NeuroImage*, **28**(4), 980–995.

Mumford, J. A. and Nichols, T. (2009). Simple group fMRI modeling and inference. *NeuroImage*, **47**(4), 1469–1475. PMID: 19463958.

Müller, K. R., Mika, S., Ratsch, G., Tsuda, K., and Schölkopf, B. (2001). An introduction to kernel-based learning algorithms. *IEEE Transactions on Neural Networks*, **12**(2), 181–201.

Nandy, R. R. and Cordes, D. (2003). Novel nonparametric approach to canonical correlation analysis with applications to low CNR functional MRI data. *Magnetic Resonance in Medicine*, **50**(2), 354–365. PMID: 12876712.

Naselaris, T., Prenger, R. J., Kay, K. N., Oliver, M., and Gallant, J. L. (2009). Bayesian reconstruction of natural images from human brain activity. *Neuron*, **63**(6), 902–915.

Nenadic, I., Sauer, H., and Gaser, C. (2010). Distinct pattern of brain structural deficits in subsyndromes of schizophrenia delineated by psychopathology. *NeuroImage*, **49**(2), 1153–1160.

Norman, K. A., Polyn, S. M., Detre, G. J., and Haxby, J. V. (2006). Beyond mind-reading: Multi-voxel pattern analysis of fMRI data. *Trends in Cognitive Sciences*, **10**(9), 424–30. PMID: 16899397.

Näätänen, R., Tervaniemi, M., Sussman, E., Paavilainen, P., and Winkler, I. (2001). Primitive intelligence in the auditory cortex. *Trends in Neurosciences*, **24**(5), 283–288. PMID: 11311381.

Olivetti, E., Veeramachaneni, S., and Nowakowska, E. (2012). Bayesian hypothesis testing for pattern discrimination in brain decoding. *Pattern Recognition*, **45**(6), 2075–2084.

Ong, C. S. and An, L. T. H. (2012). Learning sparse classifiers with difference of convex functions algorithms. *Optimization Methods and Software*, pages 1–25.

O'Toole, A. J., Jiang, F., Abdi, H., Penard, N., Dunlop, J. P., and Parent, M. A. (2007). Theoretical, statistical, and practical perspectives on pattern-based classification approaches to the analysis of functional neuroimaging data. *Journal of Cognitive Neuroscience*, **19**(11), 1735–1752.

Paninski, L., Pillow, J., and Lewi, J. (2007). Statistical models for neural encoding, decoding, and optimal stimulus design. *Progress in Brain Research*, **165**, 493–507. PMID: 17925266.

Passingham, R. E., Stephan, K. E., and Kotter, R. (2002). The anatomical basis of functional localization in the cortex. *Nature Reviews Neuroscience*, **3**(8), 606–616.

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

Peleg, D. and Meir, R. (2008). A bilinear formulation for vector sparsity optimization. *Signal Processing*, **88**, 375–389.

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

Pereira, F., Mitchell, T., and Botvinick, M. (2009). Machine learning classifiers and fMRI: a tutorial overview. *NeuroImage*, **45**(1, Supplement 1), S199–S209.

Perina, A., Cristani, M., Castellani, U., Murino, V., and Jojic, N. (2010). A hybrid generative/discriminative classification framework based on free-energy terms. In *IEEE 12th International Conference on Computer Vision*, pages 2058–2065.

Pillow, J. W., Shlens, J., Paninski, L., Sher, A., Litke, A. M., Chichilnisky, E. J., and Simoncelli, E. P. (2008). Spatio-temporal correlations and visual signalling in a complete neuronal population. *Nature*, **454**(7207), 995–999. PMID: 18650810.

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

Polyn, S. M., Natu, V. S., Cohen, J. D., and Norman, K. A. (2005). Category-specific cortical activity precedes retrieval during memory search. *Science*, **310**(5756), 1963–1966.

Price, C. J. (2010). The anatomy of language: a review of 100 fMRI studies published in 2009. *Annals of the New York Academy of Sciences*, **1191**, 62–88. PMID: 20392276.

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology*, **25**, 111–163.

Rasmussen, C. E. (2000). The infinite Gaussian mixture model. *Advances in neural information processing systems*, **12**(5.2), 2.

Rasmussen, C. E. and Ghahramani, Z. (2003). Bayesian Monte Carlo. *Advances in neural information processing systems*, **15**, 489–496.

Robert, C. P. (2007). *The Bayesian choice: from decision-theoretic foundations to computational implementation*. Springer.

Ryali, S., Supekar, K., Abrams, D. A., and Menon, V. (2010). Sparse logistic regression for whole-brain classification of fMRI data. *NeuroImage*, **51**(2), 752–764.

Schelldorfer, J., Bühlmann, P., and van de Geer, S. (2010). Estimation for high-dimensional linear mixed-effects models using L1-penalization. *arXiv.org preprint 1002.3784*.

Schofield, T. M., Penny, W. D., Stephan, K. E., Crinion, J. T., Thompson, A. J., Price, C. J., and Leff, A. P. (2012). Changes in auditory feedback connections determine the severity of speech processing deficits after stroke. *Journal of Neuroscience*, **32**(12), 4260–4270. PMID: 22442088.

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

Schölkopf, B. and Smola, A. J. (2002). *Learning with kernels: Support vector machines, regularization, optimization, and beyond*. MIT Press.

Scott, D. W. (1992). Kernel density estimators. In *Multivariate Density Estimation: Theory, Practice, and Visualization*. Wiley.

Serences, J. T. and Boynton, G. M. (2007). The representation of behavioral choice for motion in human visual cortex. *Journal of Neuroscience*, **27**(47), 12893–12899.

Shawe-Taylor, J. and Cristianini, N. (2004). *Kernel methods for pattern analysis*. Cambridge University Press.

Shen, H., Wang, L., Liu, Y., and Hu, D. (2010). Discriminative analysis of resting-state functional connectivity patterns of schizophrenia using low dimensional embedding of fMRI. *NeuroImage*, **49**(4), 3110–3121.

Sitaram, R., Caria, A., Veit, R., Gaber, T., Rota, G., Kuebler, A., and Birbaumer, N. (2007). fMRI brain-computer interface: A tool for neuroscientific research and treatment. *Computational Intelligence and Neuroscience*, **2007**. Article ID 25487.

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

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

Smith, N. and Niranjan, M. (2000). Data-dependent kernels in SVM classification of speech patterns. In *Proceedings of the 6th International Conference on Spoken Language Processing*.

Soon, C., Namburi, P., Goh, C., Chee, M., and Haynes, J.-D. (2009). Surface-based information detection from cortical activity. *NeuroImage*, **47**(Supplement 1), S79.

Stephan, K., Penny, W., Moran, R., den Ouden, H., Daunizeau, J., and Friston, K. (2010). Ten simple rules for dynamic causal modeling. *NeuroImage*, **49**(4), 3099–3109.

Stephan, K. E., Weiskopf, N., Drysdale, P. M., Robinson, P. A., and Friston, K. J. (2007a). Comparing hemodynamic models with DCM. *NeuroImage*, **38**(3), 387–401. PMID: 17884583.

Stephan, K. E., Harrison, L. M., Kiebel, S. J., David, O., Penny, W. D., and Friston, K. J. (2007b). Dynamic causal models of neural system dynamics: Current state and future extensions. *Journal of Biosciences*, **32**(1), 129–44. PMID: 17426386.

Stephan, K. E., Marshall, J. C., Penny, W. D., Friston, K. J., and Fink, G. R. (2007c). Interhemispheric integration of visual processing during task-driven lateralization. *Journal of Neuroscience*, **27**(13), 3512–3522. PMID: 17392467.

Stephan, K. E., Kasper, L., Harrison, L. M., Daunizeau, J., den Ouden, H. E., Breakspear, M., and Friston, K. J. (2008). Nonlinear dynamic causal models for fMRI. *NeuroImage*, **42**(2), 649–662.

Stephan, K. E., Penny, W. D., Daunizeau, J., Moran, R. J., and Friston, K. J. (2009a). Bayesian model selection for group studies. *NeuroImage*, **46**(4), 1004–1017.

Stephan, K. E., Friston, K. J., and Frith, C. D. (2009b). Dysconnection in schizophrenia: From abnormal synaptic plasticity to failures of self-monitoring. *Schizophrenia Bulletin*, **35**(3), 509–527. PMC2669579.

Swinburn, K., Porter, G., and Howard, D. (2004). *Comprehensive Aphasia Test*. Psychology Press, New York.

Thomaz, C. E., Boardman, J. P., Hill, D. L., Hajnal, J. V., Edwards, D. D., Rutherford, M. A., Gillies, D. F., and Rueckert, D. (2004). Using a maximum uncertainty LDA-based approach to classify and analyse MR brain images. In *Medical Image Computing and Computer-Assisted Intervention*, pages 291–300.

Tong, F. and Pratte, M. S. (2012). Decoding patterns of human brain activity. *Annual review of psychology*, **63**, 483–509. PMID: 21943172.

Valente, G., De Martino, F., Esposito, F., Goebel, R., and Formisano, E. (2010). Predicting subject-driven actions and sensory experience in a virtual world with relevance vector machine regression of fMRI data. *NeuroImage*. PMID: 20888922.

van Gerven, M., Hesse, C., Jensen, O., and Heskes, T. (2009). Interpreting single trial data using groupwise regularisation. *NeuroImage*, **46**(3), 665–676.

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

von der Behrens, W., Bäuerle, P., Kössl, M., and Gaese, B. H. (2009). Correlating stimulus-specific adaptation of cortical neurons and local field potentials in the awake rat. *Journal of Neuroscience*, **29**(44), 13837–13849. PMID: 19889995.

Wang, Y., Fan, Y., Bhatt, P., and Davatzikos, C. (2010). High-dimensional pattern regression using machine learning: From medical images to continuous clinical variables. *NeuroImage*, **50**(4), 1519–1535.

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

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

Woolrich, M. W., Behrens, T. E. J., Beckmann, C. F., Jenkinson, M., and Smith, S. M. (2004). Multilevel linear modelling for FMRI group analysis using bayesian inference. *NeuroImage*, **21**(4), 1732–1747. PMID: 15050594.

Yamashita, O., Sato, M.-a., Yoshioka, T., Tong, F., and Kamitani, Y. (2008). Sparse estimation automatically selects voxels relevant for the decoding of fMRI activity patterns. *NeuroImage*, **42**(4), 1414–1429.

Zhang, D. and Lee, W. S. (2008). Learning classifiers without negative examples: A reduction approach.

Zhang, T. (2009). Some sharp performance bounds for least squares regression with L1 regularization. *Annals of Statistics*, **37**(5A), 2109–2144.

Appendix A

Inversion of the beta-binomial model

A.1 Algorithm for stochastic approximate inference

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

$$\omega^{(0)} = \left(\ln\left(\frac{\alpha^{(0)}}{\beta^{(0)}}\right), \ln\left(\alpha^{(0)} + \beta^{(0)}\right)\right)^{\mathrm{T}}. \tag{A.1.1}$$

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

$$q_\tau\left(\omega^* \,\middle|\, \omega^{(\tau-1)}\right) = \mathcal{N}_2\left(\omega^* \,\middle|\, \omega^{(\tau-1)}, \begin{pmatrix} 1/8 & 1 \\ 1 & 1/8 \end{pmatrix}\right). \tag{A.1.2}$$

This candidate sample ω^* is accepted with probability

$$\min\left\{1,\ \frac{p(k_{1:m} \mid \alpha^*, \beta^*)\, p(\alpha^*, \beta^*)}{p(k_{1:m} \mid \alpha^{(\tau-1)}, \beta^{(\tau-1)})\, p(\alpha^{(\tau-1)}, \beta^{(\tau-1)})}\right\} \tag{A.1.3}$$

$$= \min\left\{1,\ \exp\left(\sum_{j=1}^{m} f(\alpha^*, \beta^*, k_j) - f(\alpha^{(\tau-1)}, \beta^{(\tau-1)}, k_j)\right)\right\} \tag{A.1.4}$$

where (4.3.6) and (4.3.9) (main text) were used in defining

$$f(\alpha, \beta, k) := \ln \mathrm{Bb}(k \mid \alpha, \beta) + \ln p(\alpha, \beta). \tag{A.1.5}$$

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

$$\mathbb{E}\left[\frac{\alpha}{\alpha+\beta} \;\middle|\; k_{1:m}\right] \approx \frac{1}{c}\sum_{\tau=1}^{c} \frac{\alpha^{(\tau)}}{\alpha^{(\tau)}+\beta^{(\tau)}}. \tag{A.1.6}$$

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

$$p = P\left(\frac{\alpha}{\alpha+\beta} \leq 0.5 \;\middle|\; k_{1:m}\right) \approx \frac{1}{c}\,\#\left\{\frac{\alpha^{(\tau)}}{(\alpha^{(\tau)}+\beta^{(\tau)}} \leq 0.5\right\}, \tag{A.1.7}$$

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

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

$$\mathrm{Beta}\left(\pi_j^{(\tau)} \;\middle|\; \alpha^{(\tau)} + k_j,\; \beta^{(\tau)} + n_j - k_j\right). \tag{A.1.8}$$

This can be derived by relating the full conditional

$$p(\pi_j \mid \alpha, \beta, \pi_{1:j-1, j+1:m}, k_{1:m}) \tag{A.1.9}$$

to the closed-form posterior in (4.3.8) (see main text; cf. Gelman *et al.*, 2003).

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

$$p(\tilde{\pi} \mid k_{1:m}), \tag{A.1.10}$$

in which $\tilde{\pi}$ denotes the classification accuracy in a new subject drawn from the same population as the existing group of subjects with latent accuracies $\pi_1, \ldots, \pi_m$.[1] Unlike the posterior on $\alpha/(\alpha + \beta)$, the posterior predictive density on $\tilde{\pi}$ reflects both the mean and the variance of the performance achieved in the population.[2]

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

$$p(\tilde{\pi} \mid k_{1:m}) = \iint p(\tilde{\pi} \mid \alpha, \beta)\, p(\alpha, \beta \mid k_{1:m})\, \mathrm{d}\alpha\, \mathrm{d}\beta, \tag{A.1.11}$$

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

$$\mathrm{Beta}\left(\tilde{\pi}^{(\tau)} \;\middle|\; \alpha^{(\tau)}, \beta^{(\tau)}\right). \tag{A.1.12}$$

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

$$p(\tilde{\pi} \leq 0.5) \approx \frac{1}{c}\, \#\{\pi^{(\tau)} \leq 0.5\}, \tag{A.1.13}$$

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

A.2 Classical shrinkage using the James-Stein estimator

When inferring on subject-specific accuracies π_j, the beta-binomial model uses data from the entire group to inform inferences in individual subjects.

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

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

Effectively, subject-specific posteriors are 'shrunk' to the population mean. This is in contrast to using sample accuracies $\hat{\pi} = k_j/n_j$ as individual estimates. In classical inference, a similar shrinkage effect can be achieved using the positive-part James-Stein estimator (James and Stein, 1961). It is given by

$$\hat{\pi}^{\mathrm{JS}}_{1:m} = (1 - \xi)\bar{\hat{\pi}}_{1:m} + \xi\hat{\pi}_{1:m} \tag{A.2.1}$$

$$\xi = \left(1 - \frac{(m-2)\,\hat{\sigma}^2_m(\hat{\pi}_{1:m})}{\|\hat{\pi}_{1:m} - \bar{\hat{\pi}}_{1:m}\|^2_2}\right)^+ \tag{A.2.2}$$

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

Appendix B

Inversion of the bivariate normal-binomial model

B.1 Algorithm for stochastic approximate inference

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

$$\kappa_m = \kappa_0 + m \tag{B.1.1}$$

$$\nu_m = \nu_0 + m \tag{B.1.2}$$

$$\mu_m = \frac{\kappa_0}{\kappa_m}\mu_0 + \frac{m}{\kappa_m}\bar{\rho}^{(\tau-1)} \tag{B.1.3}$$

$$S = \Sigma_{j=1}^{m}\left(\rho_j^{(\tau-1)} - \bar{\rho}^{(\tau-1)}\right)\left(\rho_j^{(\tau-1)} - \rho^{(\tau-1)}\right)^{\mathrm{T}} \tag{B.1.4}$$

$$\Lambda_m = \Lambda_0 + S + \frac{\kappa_0\, m}{\kappa_m}\left(\bar{\rho}^{(\tau-1)} - \mu_0\right)\left(\bar{\rho}^{(\tau-1)} - \mu_0\right)^{\mathrm{T}}, \tag{B.1.5}$$

[1] Here, we define one iteration as an update of all latent variables. Alternatively, one could update just one variable (or a subset of variables) per iteration, chosen randomly or systematically, as long as each variable is updated periodically.

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

$$\Sigma^{(\tau)} \sim \text{Inv-Wishart}_{\nu_m}\left(\Sigma^{(\tau)} \;\middle|\; \Lambda_m^{-1}\right). \tag{B.1.6}$$

We then complete the first step by drawing

$$\mu^{(\tau)} \sim \mathcal{N}_2\left(\mu^{(\tau)} \;\middle|\; \mu_m,\, \Sigma^{(\tau)}/\kappa_m\right), \tag{B.1.7}$$

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

$$\phi^{(\tau)} := \frac{1}{2}\left(\mu_1^{(\tau)} + \mu_2^{(\tau)}\right). \tag{B.1.8}$$

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

$$p\left(\rho_j^{(\tau)} \;\middle|\; k_{1:m}^{+}, k_{1:m}^{-}, \rho_{1:j-1}^{(\tau)}, \rho_{j+1:m}^{(\tau-1)}, \mu^{(\tau)}, \Sigma^{(\tau)}\right) \tag{B.1.9}$$

$$= p\left(\rho_j^{(\tau)} \;\middle|\; k_j^{+}, k_j^{-}, \mu^{(\tau)}, \Sigma^{(\tau)}\right), \tag{B.1.10}$$

which we have simplified by omitting all variables that are not part of the Markov blanket of ρ_j (cf. Figure 4.8b). Because we cannot sample from this distribution directly, we generate a candidate from a symmetric proxy distribution

$$q(\rho_j^*) = \mathcal{N}_2\left(\rho_j^* \;\middle|\; \rho_j^{(\tau-1)}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}^{\mathrm{T}}\right), \tag{B.1.11}$$

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

$$p\left(\rho_j^* \;\middle|\; k_j^{+}, k_j^{-}, \mu^{(\tau)}, \Sigma^{(\tau)}\right) \tag{B.1.12}$$

$$\propto \tilde{p}\left(\rho_j^* \;\middle|\; k_j^{+}, k_j^{-}, \mu^{(\tau)}, \Sigma^{(\tau)}\right) \tag{B.1.13}$$

$$= p\left(k_j^{+}, k_j^{-} \;\middle|\; \rho_j^*, \mu^{(\tau)}, \Sigma^{(\tau)}\right)\; p\left(\rho_j^* \;\middle|\; \mu^{(\tau)}, \Sigma^{(\tau)}\right) \tag{B.1.14}$$

$$= p\left(k_j^{+}, k_j^{-} \mid \rho_j^*\right)\; p\left(\rho_j^* \;\middle|\; \mu^{(\tau)}, \Sigma^{(\tau)}\right) \tag{B.1.15}$$

$$= p\left(k_j^{+} \mid \rho_{j,1}^*\right)\; p\left(k_j^{-} \mid \rho_{j,2}^*\right)\; p\left(\rho_j^* \;\middle|\; \mu^{(\tau)}, \Sigma^{(\tau)}\right) \tag{B.1.16}$$

$$= \text{Bin}\left(k_j^{+} \mid \sigma(\rho_{j,1}^*)\right)\; \text{Bin}\left(k_j^{-} \mid \sigma(\rho_{j,2}^*)\right)\; \mathcal{N}_2\left(\rho_j^* \;\middle|\; \mu^{(\tau)}, \Sigma^{(\tau)}\right), \tag{B.1.17}$$

where (B.1.13) places our focus on the unnormalized density, (B.1.14) uses Bayes' theorem, (B.1.15) is based on the Markov blanket, (B.1.16) exploits the conditional independence of class-specific outcomes k_j^+ and k_j^-, and (B.1.17) relies on the model assumptions introduced in (4.3.2) and (4.4.9) (main text). We can use this result to accept the candidate sample ρ_j^* with probability

$$\min\{1, \exp(r)\}, \tag{B.1.18}$$

where

$$\begin{aligned} r &= \ln \frac{\tilde{p}\left(\rho_j^* \mid k_j^+, k_j^-, \mu^{(\tau)}, \Sigma^{(\tau)}\right)}{\tilde{p}\left(\rho_j^{(\tau-1)} \,\middle|\, k_j^+, k_j^-, \mu^{(\tau)}, \Sigma^{(\tau)}\right)} \qquad \text{(B.1.19)} \\ &= \ln \mathrm{Bin}\left(k_j^+ \mid \sigma\left(\rho_{j,1}^*\right)\right) + \ln \mathrm{Bin}\left(k_j^- \mid \sigma\left(\rho_{j,2}^*\right)\right) \\ &\quad + \ln \mathcal{N}_2\left(\rho_j^* \,\middle|\, \mu^{(\tau)}, \Sigma^{(\tau)}\right) \\ &\quad - \ln \mathrm{Bin}\left(k_j^+ \,\middle|\, \sigma\left(\rho_{j,1}^{(\tau-1)}\right)\right) - \ln \mathrm{Bin}\left(k_j^- \,\middle|\, \sigma\left(\rho_{j,2}^{(\tau-1)}\right)\right) \\ &\quad - \ln \mathcal{N}_2\left(\rho_j^{(\tau-1)} \,\middle|\, \mu^{(\tau)}, \Sigma^{(\tau)}\right). \qquad \text{(B.1.20)} \end{aligned}$$

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

$$\pi_j^{(\tau)} = \sigma\left(\rho_j^{(\tau)}\right). \tag{B.1.21}$$

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

$$\phi_j^{(\tau)} := \frac{1}{2}\left(\pi_{j,1}^{(\tau)} + \pi_{j,2}^{(\tau)}\right). \tag{B.1.22}$$

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

$$\tilde{\rho}^{(\tau)} \sim \mathcal{N}_2\left(\tilde{\rho}^{(\tau)} \,\middle|\, \mu^{(\tau)}, \Sigma^{(\tau)}\right), \tag{B.1.23}$$

and then obtain the desired sample using

$$\tilde{\pi}^{(\tau)} = \sigma\left(\tilde{\rho}^{(\tau)}\right), \tag{B.1.24}$$

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

$$\tilde{\phi}^{(\tau)} := \frac{1}{2}\left(\tilde{\pi}_1^{(\tau)} + \tilde{\pi}_2^{(\tau)}\right). \tag{B.1.25}$$

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

The approximate expression for the model evidence in (4.4.22) can be obtained as follows:

$$\begin{aligned}
&\ln p(k^+_{1:m}, k^-_{1:m} \mid M_{nb}) && \text{(B.1.26)}\\
&= \ln \int p(k^+_{1:m}, k^-_{1:m} \mid \rho_{1:m})\; p(\rho_{1:m} \mid M_{nb})\; \mathrm{d}\rho_{1:m} && \text{(B.1.27)}\\
&= \ln \left\langle p(k^+_{1:m}, k^-_{1:m} \mid \rho_{1:m}) \right\rangle_{\rho_{1:m}} && \text{(B.1.28)}\\
&= \ln \left\langle \prod_j^m p(k^+_j, k^-_j \mid \rho_j) \right\rangle_{\rho_{1:m}} && \text{(B.1.29)}\\
&= \ln \left\langle \prod_j^m p\left(k^+_j \mid \rho_{j,1}\right)\; p\left(k^-_j \mid \rho_{j,2}\right) \right\rangle_{\rho_{1:m}} && \text{(B.1.30)}\\
&\approx \ln \frac{1}{c} \sum_{\tau=1}^{c} \prod_j^m p\left(k^+_j \mid \rho^{(\tau)}_{j,1}\right)\; p\left(k^-_j \mid \rho^{(\tau)}_{j,2}\right) && \text{(B.1.31)}\\
&= \ln \frac{1}{c} \sum_{\tau=1}^{c} \prod_j^m \mathrm{Bin}\left(k^+_j \;\middle|\; \sigma\left(\rho^{(\tau)}_{j,1}\right)\right)\; \mathrm{Bin}\left(k^-_j \;\middle|\; \sigma\left(\rho^{(\tau)}_{j,2}\right)\right) && \text{(B.1.32)}
\end{aligned}$$

B.2 Bivariate normal prior

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

$$p_\pi(\pi \mid \mu, \Sigma) = p_\rho(\sigma^{-1}(\pi) \mid \mu, \Sigma) \left|\frac{\mathrm{d}\sigma}{\mathrm{d}\rho}\right|^{-1}, \tag{B.2.1}$$

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

$$\frac{\mathrm{d}\sigma}{\mathrm{d}\rho} = \begin{pmatrix} \sigma'(\rho_1) & 0 \\ 0 & \sigma'(\rho_2) \end{pmatrix}, \tag{B.2.2}$$

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

$$\left|\frac{\mathrm{d}\sigma}{\mathrm{d}\rho}\right|^{-1} = \frac{1}{\sigma'(\rho_1)\,\sigma'(\rho_2)}. \tag{B.2.3}$$

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

$$p_\pi(\pi \mid \mu, \Sigma) \tag{B.2.4}$$

$$= \mathcal{N}_2\left(\sigma^{-1}(\pi) \;\middle|\; \mu, \Sigma\right) \frac{1}{\sigma'(\sigma^{-1}(\pi_1))\,\sigma'(\sigma^{-1}(\pi_2))} \tag{B.2.5}$$

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

$$p_\pi(\pi \mid \mu, \Sigma) \tag{B.2.6}$$

$$\propto \frac{1}{\pi_1\,\pi_2\,(1-\pi_1)(1-\pi_2)} \times \exp\left\{-\frac{1}{2}\left(\sigma^{-1}(\pi) - \mu\right)^{\mathrm{T}}\,\Sigma^{-1}\,\left(\sigma^{-1}(\pi) - \mu\right)\right\}, \tag{B.2.7}$$

where we have used the fact that $\sigma'(x) = \sigma(x)(1 - \sigma(x))$. This derivation allows us to illustrate the degrees of freedom of our family of prior distributions over μ and Σ (see Figure 4.9).

Appendix C

Inversion of the univariate normal-binomial model

C.1 Algorithm for stochastic approximate inference

Given a set of classification outcomes $k \equiv k_{1:m} \equiv (k_1, \ldots, k_m)$, to obtain a sample from the first posterior of interest, $p(\mu \mid k)$, we draw from the full-conditional distribution $p\left(\mu \mid \lambda^{(\tau-1)}, \rho^{(\tau-1)}\right)$. Since the Gaussian prior $p(\mu)$ is conjugate with respect to the likelihood $p(\rho_j \mid \mu, \lambda)$, the full-conditional posterior (i.e., the distribution from which $\mu^{(\tau)}$ is sampled) is available in closed form,

$$\mu^{(\tau)} \leftarrow \mathcal{N}\left(\mu^{(\tau)} \;\middle|\; \frac{\eta_0}{\eta_0 + m\lambda^{(\tau-1)}}\mu_0 + \frac{m\lambda^{(\tau-1)}}{\eta_0 + m\lambda^{(\tau-1)}}\bar{\rho}^{(\tau-1)},\, \eta_0 + m\lambda^{(\tau-1)}\right). \tag{C.1.1}$$

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 $\bar{\rho}^{(\tau-1)}$ 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

$p\left(\lambda \mid \mu^{(\tau)}, \rho^{(\tau-1)}\right)$. As above, the choice of a conjugate prior yields a closed-form posterior,

$$\lambda^{(\tau)} \leftarrow \mathrm{Ga}\left(\lambda^{(\tau)} \;\middle|\; a_0 + \frac{m}{2}, b_0 + \frac{1}{2}\sum_{j=1}^{m}\left(\rho_j^{(\tau-1)} - \mu^{(\tau)}\right)^2\right), \tag{C.1.2}$$

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 (C.1.1).

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

$$\rho_j^* \leftarrow \mathcal{N}\left(\rho_j^* \;\middle|\; \rho_j^{(\tau-1)}, 2^2\right), \tag{C.1.3}$$

where the choice of variance of the proxy density was guided empirically to balance exploration and exploitation of the resulting Markov chain (cf. p. 72 in Section 4.3.3). The sample drawn in (C.1.3) is accepted as the next $\rho_j^{(\tau)}$ with probability

$$\min\left\{1, \frac{\mathrm{Bin}\left(k_j \mid \sigma(\rho_j^*), n_j\right)\mathcal{N}\left(\rho_j^* \mid \mu^{(\tau)}, \lambda^{(\tau)}\right)}{\mathrm{Bin}\left(k_j \;\middle|\; \sigma\left(\rho_j^{(\tau-1)}\right), n_j\right)\mathcal{N}\left(\rho_j^{(\tau-1)} \;\middle|\; \mu^{(\tau)}, \lambda^{(\tau)}\right)}\right\}. \tag{C.1.4}$$

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. 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 in Section 4.7).

Acknowledgments

I want to thank my supervisors Klaas Enno Stephan and Joachim Buhmann for their superb guidance and truly dedicated support. Being part of their research programmes has been an exceptional pleasure and honour.

It has been an incredibly enjoyable and enlightening experience to work with my fellow doctoral students Alberto-Giovanni Busetto, Morteza Chehreghani, Sandra Iglesias, Lars Kasper, Kate Lomakina, Christoph Mathys, and the postdoctoral researchers Andreea Diaconescu, Jakob Heinzle, and Gabor Stefanics.

I want to thank Justin Chumbley, Jean Daunizeau, Cheng Soon Ong and, once again, Christoph Mathys, all of whom have been outstandingly knowledgeable, generous, and truly delightful colleagues to work with.

Several of the results presented in this thesis directly gave rise to the projects of Ajita Gupta and Zhihao Lin who I had the great pleasure of supervising during their Master's studies. Some results are tightly interlocked with the projects of Falk Lieder, Saee Paliwal, and Vera Schäfer; I truly enjoyed gaining an insight into their research.

I am indebted to Eva Gschwend, Rita Klute, Ursi Meier, Denise Spicher, Suzanne Wilde, and Karin Zeder for their administrative support.

My profound thanks go to the many collaborators who were involved in the studies presented in this thesis: Lorenz Deserno and Florian Schlagenhauf (Berlin); James Rowe and Laura Hughes (Cambridge); Fabienne Jung, Ulrich Pfeiffer, Leo Schillbach, Marc Tittgemeyer, Kai Vogeley (Cologne); Paul Allen, Alex Leff, Philip McGuire, Rosalyn Moran, Tom Schofield, and Will Penny (London); Chia-shu Lin, Irene Tracey, and Katja Wiech (Oxford); as well as Florent Haiss and Buno Weber (Zurich).

I am greatly and variously indebted to the many colleagues and former supervisors who have contributed to this research through discussions and correspondences: Carlton Chu, Karl Friston, Janaina Mourao-Miranda, and Ged Ridgway (London); Tim Behrens, Adrian Groves, Laurence Hunt, Saad Jbabdi, Matthew Rushworth, and Mark Woolrich (Oxford); Tom Nichols (Warwick); as well as Anne Broger, Ernst Fehr, and Christian Ruff (Zurich).

I want to thank my parents for their unwavering support, and Mia for her warmth, kindness, and grace.

This thesis was supported by the University Research Priority Program 'Foundations of Human Social Behaviour' at the University of Zurich, by the SystemsX.ch project 'Neurochoice', and by the NCCR 'Neural Plasticity.'

Kurzfassung

Multivariate Zeitreihen lassen sich mit Differenzialgleichungen modellieren, die beschreiben, wie die Bestandteile eines unterliegenden dynamischen Systems zeitlich interagieren. Ein besonders herausforderndes Anwendungsfeld ist die Neurowissenschaft, in der zunehmend *dynamic causal models* verwendet werden, um die Mechanismen hinter multivariaten Zeitreihen von Hirnaktivität im gesunden und erkrankten menschlichen Gehirn zu beleuchten. Die vorliegende Dissertation stellt einen Ansatz vor, solche Modelle in klinische Anwendungen zu übertragen, den wir als *generative embedding* bezeichnen. Unser Ansatz basiert auf der Idee, dass eine mechanistisch interpretierbare Beschreibung eines Systems wesentlich bessere Einsichten ermöglicht als die beobachteten Zeitreihen selbst. Konzeptionell beginnen wir mit der Entwicklung eines Verfahrens zur *modellbasierten Klassifikation*; es beruht auf der Kombination eines generativen Modells mit einem diskriminativen Klassifikator. Wir zeigen, dass dieser Ansatz signifikant genauere diagnostische Klassifikationen und tiefere mechanistische Einsichten ermöglicht als bisherige Methoden. Die Verwendung eines Klassifikationsalgorithmus auf hierarchichen Daten erfordert neue Antworten auf die Frage nach dessen statistischer Evaluation. Wir führen Bayesianische Modelle ein, die die verschiedenen Quellen von Variabilität richtig zueinander in Bezug setzen, um optimale statistische Inferenz zu ermöglichen. Wir schlagen vor, die konventionelle Klassifikationsgenauigkeit durch die sog. balancierte Genauigkeit zu ersetzen, wenn die Daten nicht selbst balanciert sind. Wir veranschaulichen die Eigenschaften unserer Modelle anhand von stochastischer approximativer Inferenz auf der Basis von Markov chain Monte Carlo. Wir leiten anschließend mittels Bayes'scher Variationsrechnung eine hocheffiziente deterministische Näherung her. Komplementär zur Anwendung in der Klassifikation ermöglicht *generative embedding* die Entdeckung mechanistisch interpretierbarer, a priori unbekannter Untergruppen. Wir entwickeln ein Verfahren zum *modellbasierten Clustering*, mit dem wir eine Gruppe von Schizophrenie-Patienten in Untergruppen mit klinischer Validität zerlegen. Zusammengefasst erkundet die vorliegende Dissertation mit *generative embedding* und Bayes'scher Inferenz die konzeptionellen, statistischen, und rechnerischen Grundlagen für den Einsatz von modellbasierten Klassifikations- und Clustering-Verfahren im klinischen Kontext. Wir erwarten, dass zukünftige Anwendungen unserer Methodik es ermöglichen werden, Gruppen von Patienten mit ähnlichen Symptomen in sich pathophysiologisch unterscheidende Untergruppen aufzuspalten.

Curriculum vitae

Kay Henning Brodersen
born 9 December 1982, Flensburg, Germany

Education

04/2009 – 10/2012	**ETH Zurich**, Switzerland PhD in Computer Science
09/2007 – 04/2009	**University of Oxford**, UK Master of Science (MSc) in Neuroscience
09/2005 – 07/2006	**University of Cambridge**, UK Visiting undergraduate in Computer Science
09/2003 – 09/2007	**University of Muenster**, Germany Master of Science (MSc) in Information Systems Bachelor of Science (BSc) in Information Systems
08/1993 – 04/2002	**Foerde-Gymnasium Flensburg**, Germany Abitur

Research experience

06/2012 – 12/2012	**Google**, Paris, France / Mountain View, USA
04/2008 – 04/2009	**Oxford Centre for fMRI of the Brain (FMRIB)** University of Oxford, UK
07/2006 – 10/2006	**Wellcome Trust Centre for Neuroimaging (FIL)** UCL, London, UK
03/2005 – 04/2005	**Juelich Supercomputing Centre**, Germany
04/2003 – 07/2003	**IBM**, Duesseldorf, Germany