Optimizing parameters and algorithms of multivariate pattern classification for hypothesis testing in high-density EEG

Dissertation

zur Erlangung des Grades eines
Doktors der Naturwissenschaften

der Mathematisch-Naturwissenschaftlichen Fakultät
und
der Medizinischen Fakultät
der Eberhard-Karls-Universität Tübingen

vorgelegt
von

Hamidreza Jamalabadi

aus Teheran, Iran

June – 2017

Tag der mündlichen Prüfung:	September 29, 2017
Dekan der Math.-Nat. Fakultät:	Prof. Dr. W. Rosenstiel
Dekan der Medizinischen Fakultät:	Prof. Dr. I. B. Autenrieth
1. Berichterstatter:	Prof. Dr. Steffen Gais
2. Berichterstatter:	Prof. Dr.-Ing. Moritz Grosse-Wentrup
Prüfungskommission:	Prof. Dr. Steffen Gais
	Prof. Dr.-Ing. Moritz Grosse-Wentrup
	Prof. Dr. Jan Born
	Prof. Dr. med. Martin Walter

Erklärung / Declaration:

Ich erkläre, dass ich die zur Promotion eingereichte Arbeit mit dem Titel:
„**Optimizing parameters and algorithms of multivariate pattern classification for hypothesis testing in high-density EEG**“

selbständig verfasst, nur die angegebenen Quellen und Hilfsmittel benutzt und wörtlich oder inhaltlich übernommene Stellen als solche gekennzeichnet habe. Ich versichere an Eides statt, dass diese Angaben wahr sind und dass ich nichts verschwiegen habe. Mir ist bekannt, dass die falsche Abgabe einer Versicherung an Eides statt mit Freiheitsstrafe bis zu drei Jahren oder mit Geldstrafe bestraft wird.

I hereby declare that I have produced the work entitled “**Optimizing parameters and algorithms of multivariate pattern classification for hypothesis testing in high-density EEG**”, submitted for the award of a doctorate, on my own (without external help), have used only the sources and aids indicated and have marked passages included from other works, whether verbatim or in content, as such. I swear upon oath that these statements are true and that I have not concealed anything. I am aware that making a false declaration under oath is punishable by a term of imprisonment of up to three years or by a fine.

Tübingen, den ....

Datum / Date Unterschrift /Signature

Abstract

Multivariate pattern analysis (MVPA) has come into widespread use for analysis of neuroimaging data in recent years and is gaining further momentum. Given the task of detecting a generalizable pattern in neural activity, MVPA allows to detect fine multidimensional spatiotemporal contrasts between two or more conditions and is thus able to take the full advantage of multivariate information encoded in the data. In particular, MVPA based approaches lend themselves very well to the analysis of electroencephalogram (EEG) data because, unlike the widely-used averaging methods, they consider the signal in its entirety and are thus less susceptible to the confounding effects of single points with abnormal amplitudes.

However, using MVPA for hypothesis testing purposes in high-density EEG data has remained a challenging issue. Although MVPA is getting more and more mainstream to detect information in neural activity, its behavior is not well understood, yet. EEG data are high dimensional, yet sample size is usually low in comparison. Moreover, due to the low signal-to-noise ratio, the effect size is small and differences between classes are hard to detect. In such cases, MVPA behaves unexpectedly which makes the overall accuracy of the classifier difficult to interpret. In addition, because MVPA is sensitive to any kind of structure in the data, confounding factors or additional variance within data can bias accuracy. Such complexities warrant extra caution when interpreting classification results, thereby requiring further investigation and guidelines. On the other hand, MVPA literature is mainly dominated by methods suited for fMRI data and most of the dedicated EEG methodology is developed for brain computer interfaces (BCI) or single trial analysis of event-related potentials. Specifically, decoding continuous EEG increasingly suffers from the curse of dimensionality because of the lack of clear prior knowledge on which frequency bands and time points carry relevant information, or an onset where the effect of stimulation can be expected.

In this thesis, we addressed the aforementioned challenges involved in using MVPA for decoding EEG data. Chapter 2 describes the statistical properties of MVPA in realistic neuroimaging data and provides important guidelines to interpret classification results. We show that the probability distribution of classification accuracies does not follow any known parametric distribution and can be strongly biased and skewed. We describe unexpected properties of the distribution of classification rates which forbid their use as estimates of the size of experimental effects. Importantly, we scrutinize the finding of below chance level classification rates, which often occur in low sample size, low effect size data and their implications on the shape of classification rates distribution.

Next, in chapter 3, we investigate neuroimaging data that, next to a main effect of class, additionally contains a nested subclass structure. We show that in these data sets, correct classification ratios are systematically biased from chance even in absence of class effect. We propose a nonparametric permutation algorithm which can detect the subclass bias and account for its effect by adjusting permutation tests to consider the subclass structure of the data, using subclass-level randomization.

Finally, in chapter 4, we used MVPA to decode continuous high-density EEG across subjects. We developed a classification framework along with a specific preprocessing procedure that is optimized for three purposes: 1) to increase signal-to-noise ratio, 2) to reduce the dimensionality of the data, and 3) to adapt the signal better to between-subject classification. Our algorithm uses a two-step classification procedure based on ensemble of linear support vector machines (SVM) which learns the spatial and temporal components of neural activity separately and then aggregates the two components of information to build a classification hyperplane using another linear SVM. We then use this method to see whether human sleep EEG contains any information about what has been learned before sleep.

TABLE OF CONTENTS

Chapter 1:

Synopsis

Hamidreza Jamalabadi

Introduction

Multivariate pattern analysis for hypothesis testing in neuroimaging

Multivariate pattern analysis (MVPA) has come into widespread use in recent years for classification, decoding and hypothesis testing (Haxby, et al., 2014; Haynes and Rees, 2006; Kamitani and Tong, 2005; Norman, et al., 2006; Tong and Pratte, 2012). Given the task of detecting a generalizable pattern in the neural activity, MVPA is often used as a replacement of parametric statistics and has been successfully applied in various neuroimaging studies in the field of cognitive neuroscience (Cox and Savoy, 2003; Horikawa, et al., 2013; Kay, et al., 2008; Mitchell, et al., 2008; Rissman, et al., 2010; Schaefer, et al., 2011; Schwarzlose, et al., 2008). These methods allow analyzing multivariate data in a way that takes into account the statistical power residing in the broad patterns of the data, instead of only searching for one or more features that individually allow to significantly distinguish between conditions. This enhances the sensitivity of the test in two important ways (Haxby, et al., 2014; Haynes, 2015; Norman, et al., 2006). First, variables with weak but reliable information which might not survive significance test will become discriminative due to the accumulation of information across all features. Second, variables which do not carry information might contribute to discrimination when they are jointly analyzed with another subset of features. In addition, because MVPA provides the possibility of analysis of data on a single trial basis, they often reduce the sample size required (Norman, et al., 2006). Taken together, pattern-based classification techniques provide a competent alternative with increased sensitivity compared to classical mass univariate techniques.

Importantly, MVPA represents an information-based instead of an activation-based approach (Kriegeskorte, et al., 2006). Accordingly, when employed on recorded data, results do not show which brain areas are most active during performance of a certain task, but rather point to the brain areas or patterns of

activity that contain the highest information about processing of the task. This can often answer the actual research question more precisely, because the interest mainly lies in the question whether a specific brain region participates in a cognitive process, rather than in knowing the actual strength of the activity within a certain region which is measured by the classical multiple comparison tests.

When using multivariate pattern analysis (see Figure 1), brain activity is analyzed at the level of patterns which are evoked by different stimuli or experimental conditions. Each pattern of activity is associated with a mental state (e.g. those elicited by viewing different images in Figure 1a) and can be expressed as a pattern vector which is represented by a group of variables in multidimensional feature space (e.g. voxels in fMRI, or channel-time points in EEG/MEG). The task of the classifier is to find a generalizable decision rule which can distinguish patterns of activity belonging to each experimental condition. For that, the classifier is trained on a subset of data and subsequently tested on an independent test set to predict the condition labels for new unseen data (see Figure 1e). If the classifier can successfully decode the stimuli (or experimental conditions) solely based on the patterns of brain activity, it can be concluded that some information relevant to the experimental manipulation exists in the data.

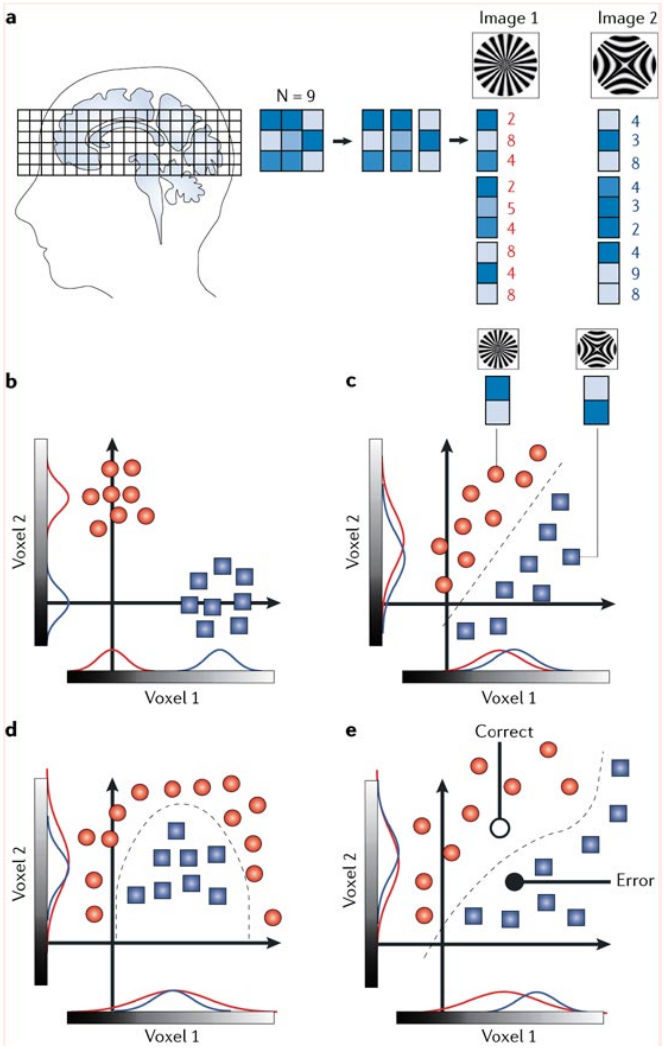


Figure 1: Patterns of brain activity can be analyzed using multivariate pattern analysis approach. (a) Brain responses to different stimuli (e.g. Image 1 and Image 2) are measured in different brain areas across time. Patterns of activations are represented in multiple dimensions (channels in different time instances for EEG/MEG, or voxels in fMRI) and can be expressed as pattern vectors. In b-e, only two dimensions are illustrated for simplicity (red and blue indicate the two dimensions). (b-c) Pattern vectors are distinguishable in single voxels if the marginal distributions are not overlapping (b). However, in many cases, the discrimination based on single voxels is impossible (due to largely overlapping distributions) while considering them together provides a perfect classification (c). (d) An example of a case where a nonlinear decision boundary (corresponding to a nonlinear classifier) is preferred to a linear one. (e) The performance of the classifier is evaluated by testing its predictions on an independent test set (not used in the training). Figure adapted and reprinted with permission from (Haynes & Rees 2006).

Multivariate pattern analysis as a supervised classification problem

MVPA involves finding a generalizable pattern in neural activity which allows discrimination of data across experimental conditions while at the same time avoiding overfitting (Duda, et al., 2000). The general idea is to find a 'rule' which correctly distinguishes the neural activity associated to different cognitive tasks based on the labels that indicate the corresponding experimental condition. Since the class structure is provided by the experimenter in this case, this problem is therefore a straightforward application of supervised pattern classification (Bishop, 1995; Duda, et al., 2000; Lemm, et al., 2011). In this terminology, a pattern classifier is a function that gets the values of different features (variables or predictors) in different samples (independent values for variables) as an input, and predicts the class label that a new data sample belongs to as an output. There are four essential steps in employing a pattern classifier for neuroimaging data, starting with feature extraction and preparing sets of data samples, proceeding through choice of classifier and accuracy estimation and ending up in evaluating and interpretation of the results. Although it is the interaction of these components which determine the overall performance of MVPA, it is practically more convenient to focus on each step separately.

Feature extraction and creating data samples

The first step concerns the choice of features which are used in the classification. Features are any set of variables or attributes that quantify the neural activity. These features define the representational space in which the classifier search for generalizable spatiotemporal patterns of brain activity (Haxby, et al., 2014). For example, in fMRI, single voxel activity or the average activity of several voxels in one or multiple region of interests can be considered as a feature (Mitchell, et al., 2004). In a hypothetical EEG/MEG experiment, features could be the amplitude of electrical brain activity recorded by one or a cluster of electrodes at different time points in a certain time interval. Other common

examples are EEG/MEG power spectral density as frequency domain features (Fuentemilla, et al., 2010; Newman and Norman, 2010), time-frequency representation (LaRocque, et al., 2013; Schulz, et al., 2012), and variance of EEG signal extracted from common spatial filters (Noh, et al., 2014). Most often, the features are selected based on the domain expert knowledge about which attributes of the signal might contain information or the a priori hypothesis about the experiment. For instance, it is common to use time domain information for classification of event-related potentials (ERPs) whereas the frequency domain measures might be more beneficial for analyzing continuous EEG/MEG data where there is no stimulus onset. In addition, it is principally possible to improve the performance of MVPA by applying a feature transformation (e.g. Principal Component Analysis or Independent Component Analysis) or feature selection (e.g. forward selection, backward elimination, etc.) technique. However, to ensure circular analysis is avoided, one should be careful not to use any algorithm that uses the label information before separating the training and test samples (Kriegeskorte, et al., 2009).

Classification framework

The second step is the choice of classifier and the proper training algorithm. The sensitivity and type of information which can be detected largely depend on the particular classifier used. Concerning the choice of classifier type, research in the machine learning field has put forward an enormous range of classification algorithms that can potentially be used in MVPA (Duda, et al., 2000). Classifiers that learn a mapping function are divided into two major categories based on the shape of their decision boundary; linear classifiers that decide class membership based on a hyperplane which is a linear combination of features, and nonlinear classifiers which classify data based on more complex nonplanar boundaries (see Figure 2). Some examples of linear classifiers are correlation-based classifiers (Haxby, et al., 2001), Linear Discriminant Analysis (LDA; Fisher, 1936) and linear Support Vector Machines (SVMs; for details see Duda, et al., 2000; Pereira, et al., 2009). All linear classifiers determine the class membership

by comparing a linear weighted sum of features to a threshold. LDA identifies these projection weights by maximizing the between-class to within-class variance while linear SVM identifies weights based on the maximum margin hyperplane (see Duda, et al., 2000 for details). If linear decision boundaries cannot partition the data sufficiently well, nonlinear classifiers like k-nearest neighbor (KNN), Gaussian Naïve Base (GNB) or SVMs with nonlinear kernels can be used. For comparing the relative performance of different classifiers on fMRI data, see (Misaki, et al., 2010).

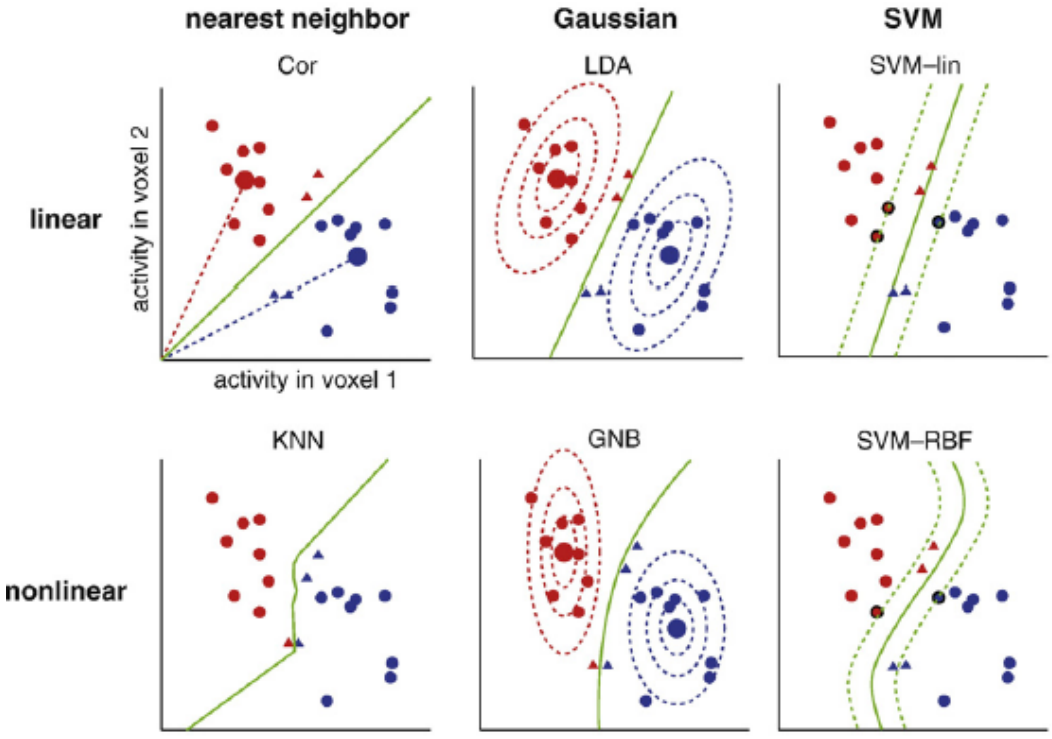


Figure 2: The main difference between linear and nonlinear classifiers is their shape of decision boundary. All linear classifiers compute a weighted sum of features which geometrically represents a hyperplane and separates trials belonging to two experimental conditions. Nonlinear classifiers define a nonplanar boundary which is more flexible and more complex than a hyperplane. However, because of more number of parameters that should be learned, they can be simply overfitted when sample size is low, a condition which is common in neuroimaging. Figure adapted and reprinted with permission from (Misaki, et al., 2010).

One key difference between nonlinear and linear classifiers is that nonlinear classifiers can, in principle, respond to high-level feature combinations in a way that differs from their response to individual features. In addition, they allow interactions between features to drive prediction. Therefore, the interpretation of the relation between features and the prediction outcome can become complicated when nonlinear classifiers are used (Pereira, et al., 2009). Moreover, nonlinear classifiers provide a more flexible decision boundary compared to linear ones. However, because of the low number of samples compared to the number of features available, which is common in neuroimaging, the boundary can become adapted to the noise which can obscure the classifier performance in generalizing to new data sets. An example of overfitting to training data with a higher-order polynomial classifier is shown in Figure 3. Although the tenth-order polynomial has a higher classification accuracy on the training set, the second-order polynomial generalizes better to the test set. This is because the tenth-order polynomial is more flexible than the second-order one due to the higher number of parameters and, therefore, can better adjust to the noise in the training data.

In practice, so far, linear classifiers have been the most successful. Particularly, linear SVMs which are developed based on the structural risk minimization principle have been used successfully in many neuroimaging studies and have often outperformed other classifiers (Haynes, 2015; Jamalabadi, et al., 2016; Misaki, et al., 2010; Mur, et al., 2009), including nonlinear SVMs (Cox and Savoy, 2003; LaConte, et al., 2005). Specifically, in two-category classification problems, the regularization technique embedded in SVM training holds down the effect of noisy and correlated features when sample size is low relative to the dimensionality of the feature vector, a condition which is common in neuroimaging data. Taken together, in low sample size data, linear SVM shows the strongest generalizability across linear and nonlinear classifiers (Attoor and Dougherty, 2004; Pereira, et al., 2009).

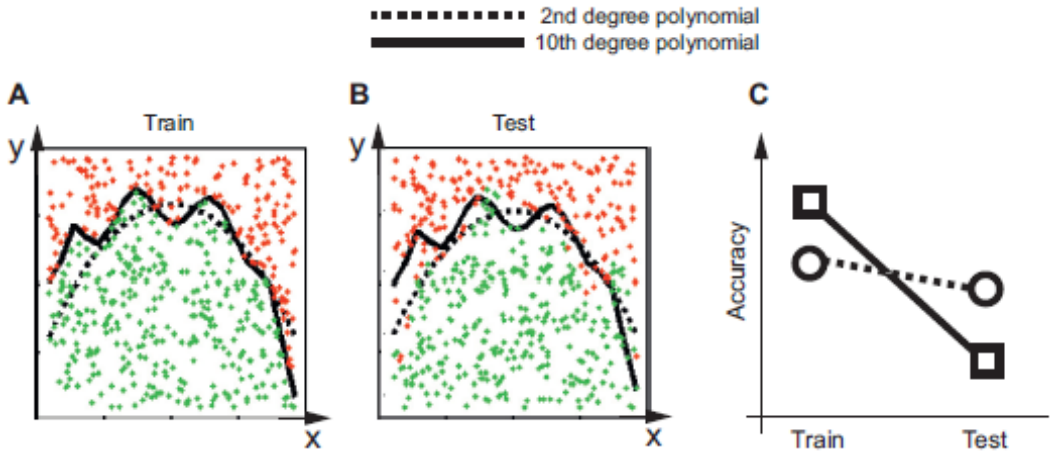


Figure 3: An example of overfitting to training data. (A, B) Nonlinear classification boundaries from two nonlinear classifiers, one second-order (dashed line) and one tenth-order (solid line) polynomial, are applied to test data. (C) The tenth-order polynomial performs better on the training set, but worse on the test set. The difference between classification accuracy of training versus test set is a measure of how well the neural pattern found by MVPA is generalizable to new unseen samples. Ideally, a classifier should be as good in classifying new samples as in classifying the samples which are used to train it. However, because of random variances in data, classifier may indeed overfit to the noise inherited in the training data. Figure adapted and reprinted with permission from (Haynes, 2015).

Accuracy estimation and cross-validation

The third step is to test the generalizability of the classification. The accuracy of classifier is the probability that a random instance is classified correctly. That is, given a set of training samples and a classification training algorithm, how well the classification results can be extended to similar but new samples. This is often done by estimating correct classification rates (CCRs), which are defined as the percentage of correctly classified new items (trials, subjects, etc.) (Jamalabadi, et al., 2016). For a two-class classification problem with balanced number of items, the theoretical chance level is 50% (100/n for n classes) and any classification equal or below chance level shows lack of evidence to support existing information about the variables of interests either because classes do

not differ or the measure used to train the classifier was insensitive to a potential difference (Jamalabadi, et al., 2016).

In order to estimate the true accuracy of a classifier, training and test data sets should be ideally as large as possible (which assures minimal variance of estimation) and be mutually exclusive (which guarantees zero bias) (Braga-Neto and Dougherty, 2004; Kohavi, 1995; Rodriguez, et al., 2010). However, because of the low number of sample size in neuroimaging data sets, this often cannot be the case. Today, the most widely used technique for CCR estimation, which provides a tradeoff between bias and variance of error estimation (compared to e.g. hold-out and bootstrap) is *K*-fold cross-validation (Hastie, 2001; Kohavi, 1995). In this method, the data is randomly divided into *K* disjoint subsets of approximately equal size. The classifier is then trained on the union of *K-1* partitions and tested on the remaining left-out subset. The *K-fold* CCR is then the average of *K* classification accuracies, where each of them is obtained by testing the corresponding classifier on the left-out test set. This procedure is usually repeated several times to reduce the variance of the estimation. Notably, *K-fold* cross-validation is simple to implement and has been used to estimate classification performance in many applications (Braga-Neto and Dougherty, 2004; Lemm, et al., 2011; Mur, et al., 2009; Noirhomme, et al., 2014; Pereira, et al., 2009).

Evaluating significance of the results

CCRs - like any other statistic - are subject to random estimation error and with low number of samples higher variations is expected. Therefore, values above chance level should be compared to the distribution expected from chance, often referred to as null distribution, to test whether the classifier has indeed extracted generalizable information from the data (Haynes, 2015). Only if CCR lies significantly above the level expected by chance, the classifier was able to detect generalizable class-related information in the data. There are two common ways to determine the significance threshold in the literature: parametric tests and permutation tests. In the framework of parametric tests,

classification accuracies are assumed to follow a known distribution (e.g. binomial or t-student distribution) while permutation tests are nonparametric tests that estimate the distribution of null hypothesis directly from the data (Nichols and Holmes, 2002). The null distribution is computed by resampling the data a large number of times (several hundreds or thousands of times) with randomly assigned group labels. As the new labels are permuted randomly, the class-related information is cancelled out. Therefore, the resulting classification accuracies from data with shuffled labels are expected to constitute the null distribution. The *p*-value in this case is the fraction of samples which is greater or equal to the accuracy actually observed when using the correct labels.

Importantly, permutation tests are computationally more engaging to implement but provide more accurate estimation of the null distribution compared to binomial tests or t-tests (Nichols and Holmes, 2002; Pereira and Botvinick, 2011; Stelzer, et al., 2013). It is shown recently that parametric tests can result in strongly biased p-values and should be avoided (Jamalabadi, et al., 2016; Noirhomme, et al., 2014).

Challenges and potentials of using MVPA for hypothesis testing in EEG

EEG is the oldest brain imaging technique and at the same time one of the most methodologically expanding tools in cognitive neuroscience. EEG has some evident advantages over other neuroimaging tools which make it a method of choice to study cognitive brain functions: it has a high temporal resolution, is rather simple to use and has relatively modest price. Besides, EEG provides an excellent window to the brain because it measures electrical brain activity which is the main representative of brain dynamics. In fact, EEG can provide valuable insights into dynamic processes underlying a cognitive brain function because it can capture the rhythmic property of neuronal activity which reflects the actual mechanisms of the brain information processing (Lopes da Silva,

2013). It is shown that oscillatory fluctuations of electrical brain activity contain information about different cognitive functions, including perception (Doesburg, et al., 2009; Rodriguez, et al., 1999), memory consolidation (Duzel, et al., 2010; Rasch and Born, 2013), memory maintenance of events (Jafarpour, et al., 2013; Newman and Norman, 2010), and neuronal plasticity (Takeuchi, et al., 2014; Walker and Stickgold, 2006). These neuronal fluctuations can be well studied using EEG because it offers excellent temporal resolution (in millisecond) with high practical flexibility (e.g. long time uninterrupted recording during sleep or wakefulness), which provides the possibility to investigate human brain activity while performing complex cognitive tasks. Therefore, the extent of questions which can be addressed using EEG spans almost all aspects of brain information processing.

However, the amount of information which can be extracted from EEG is often limited by several methodological challenges. One of the main challenges imposed by EEG signal is its low signal-to-noise ratio. The analysis of EEG suffers from the abundance of irrelevant brain activity as well as multiple sources of noise and distortions which make generalization of signals over subjects a difficult task. In the context of cognitive neuroscience, EEG analysis often heavily depends on the widely-used averaging of event-related potentials. In fact, because of the low signal-to-noise ratio in each single trial, the small signal is often not enough for trial-based analysis, which forces the experimenter to further average over trials. However, these spatiotemporal averaging methods can potentially cancel the useful information encoded in the distributed patterns of activity. Here especially, information based approaches lend themselves very well to the analysis of EEG data. MVPA methods consider the signal in its entirety and discover those parts that are relevant for a specific cognitive activity. Whereas averaging methods can be easily influenced by single data points with abnormal amplitudes, which often occur in EEG signals, machine learning algorithms, which were designed to detect consistent differences between classes of stimuli, allow the detection of fine multidimensional spatiotemporal

contrasts between two or more number of conditions and are thus able to take the full advantage of multivariate information encoded in EEG. MVPA methods consider only informative signals which can be transferred to new data, and are thus less vulnerable to outliers. This allows investigating EEG on a single trials-basis. Furthermore, due to the large amount of variability across subjects, it is rather challenging to directly compare EEG from different subjects. The electrical brain activities recorded from different subjects often show varying signal amplitudes and frequency spectrum structure (e.g. shifted or suppressed alpha peak (Haegens, et al., 2014)). Therefore, averaging-based methods can be strongly influenced and skewed by those cases having higher amplitudes. Here, between-subject classification, where data from different subjects are cross validated to estimate the performance, is a powerful method to detect the underlying general principles and reassures that the findings can be generalized to new subjects.

Importantly however, most of the dedicated MVPA based EEG methodology is developed for motor imagery, brain computer interfaces (BCI), and single trial analysis of EEG event related potential (ERPs). Because hypothesis testing in continuous EEG data has different requirements than individual item identification in BCI, methods optimized for the latter purpose are not necessarily the best for the former. Unlike ERP analysis, for many research questions, the lack of precisely defined time onsets or the length of the planned EEG recording make the usage of advanced MVPA methods mandatory. In high-density EEG, signals are recorded from 128 electrodes with a high temporal resolution (500Hz or more). In lack of a clear prior knowledge on which frequency bands and time points carry relevant information, all the above-mentioned challenges become even trickier to address. More specifically, extracting information from continuous EEG across multiple subjects and sessions is more difficult when data does not show a clear event related potential (ERP) or an onset where the effect of stimulation can be expected. Therefore, in studies with continues EEG data, the application of MVPA which

can analyze the data in its entirety is particularly invaluable to further our understanding of the underlying neural mechanism of brain responses.

Aside from the methodological difficulties involved in decoding spontaneous brain activity across subjects, the interpretation of classification accuracies, when MVPA is used for hypothesis testing, poses another challenge for effective integration of MVPA methodology into EEG analysis. Typical EEG data sets, which are recorded with the purpose of hypothesis testing, are high dimensional, yet sample size is usually low in comparison. This often leads into incomplete training of the classifiers and also a large variance of their classification accuracies. Moreover, due to the low signal-to-noise ratio, the effect size is small and differences between classes are hard to detect. In such cases, where the data is low sample size, low effect size and high dimensional, cross-validated classification behaves unexpectedly, which makes the overall accuracy of the classifier difficult to interpret. Furthermore, because MVPA is sensitive to any kind of structure in the data, any confounding factor or variance within data can affect classification accuracy. Such variations are common in EEG data sets and are mainly introduced if stimuli or types of stimuli are presented repeatedly, if multiple subjects or experimental sessions are included into one analysis, or if some secondary attributes (e.g. physical properties, familiarity, task difficulty etc.) are shared among subclasses of trials. Importantly, these confounding factors can bias classification accuracies but are not related to the effect under investigation. Thus, their effect should be discarded from the analysis. Such complexities concerning interpretation of classification accuracies and their relation to the size of effect under study require further investigation and guidelines.

Aims of this thesis

Following the aforementioned challenges of using MVPA for decoding EEG data, the main aim of this thesis is twofold. Firstly, to further investigate the behavior

and statistical properties of cross-validated MVPA in realistic life-science data and to develop methods and provide important guidelines that can be used to interpret classification results when MVPA is employed for the purpose of hypothesis testing in neuroimaging data. Secondly, to develop an effective classification framework to decode continuous high-density EEG data across subjects. To address these issues, the current thesis is divided into three chapters:

Chapter 2 deals with the overlooked statistical properties of MVPA when it is used for detection of information and generally, for hypothesis testing purposes in neuroimaging data. When MVPA is used as a replacement of parametric statistics, it is used to decode stimuli or experimental conditions to test if the neural activity contains information about them. However, despite the widespread use of MVPA, its behavior is still not fully understood. Often, higher classification accuracies are necessarily interpreted as larger effects, without taking the properties of data (e.g. sample size, number of features), cross-validation or the classifier type into account. This is particularly important because, with respect to these parameters, life-science data sets usually stand in one specific corner of the parameter space. The number of independent observations are in orders of ten (constrained by the number of subjects and trials) and in many cases less than the number of features under study (e.g. number of channels times number of frequency bins or time points in EEG recordings) (Button, et al., 2013; Jamalabadi, et al., 2016). Also, in many data sets that are recorded with the purpose of hypothesis testing, the effect size is low which naturally leads to low classification accuracy. In such low sample size/low effect size settings, understanding the behavior of MVPA by investigating the cumulative effects of cross-validation properties, classifier type, and data specifications (e.g. sample size, dimensionality) on the classification results could provide more information about reliability of the MVPA results. Here, we provide a set of guidelines that should be observed when MVPA is used for

hypothesis testing. We used a combination of simulations with synthetic data, mathematical modeling, and classification of EEG data in this chapter.

In chapter 3, we explore the consequences of high sensitivity of MVPA for differences found between subgroups of trials in data with nested subclasses. MVPA methods are more susceptible to confounding factors due to their increased sensitivity compared to conventional mass univariate methods. In particular, MVPA algorithms can use any variations within data including subject level differences or random variabilities in an effect (Haynes, 2015; Todd, et al., 2013). Here, we investigate the behavior of MVPA for neuroimaging data which, next to a main effect of class, additionally contain a nested subclass structure. In such setting, we show that classification accuracies are systematically biased and parametric testing fails critically to determine significance of classification outcomes, and that trial-wise permutation gives too liberal estimates. In order to control for the confounding contribution of subclass variance, we propose a nonparametric permutation algorithm which can account for the subclass bias by adjusting permutation tests to consider the subclass structure of the data, using subclass-level randomization. We give practical EEG examples of how to modify permutation testing for a range of common experimental designs. We further use simulations with synthetic data to study MVPA behavior and provide analytical description of bias and its relation to variances in the data.

Finally, in chapter 4, we develop a classification framework which can exploit the multidimensional pattern of brain activity in continuous EEG to find generalizable information across multiple subjects. There are mainly two major difficulties in designing a classifier for continuous EEG. First, in lack of a clear prior knowledge on which frequency bands and time points carry relevant information, or an onset where the effect of stimulation can be expected, applying MVPA becomes increasingly tricky because of the curse of dimensionality. This leads into incomplete training of the classifiers and inaccurate testing of their accuracies. This might leave important but small to

medium effect sizes undiscovered. Second, electrical brain activity recorded by EEG from different subjects show marked differences. Such variabilities are often larger than the size of effects which are under investigation. Therefore, generalizing the neural pattern using a cross-validation across subjects often becomes impractical. Here, we directly address these problems associated with analysis of high density continuous EEG when used to classify data across multiple subjects. We then use this method to see whether human sleep EEG contains any information about what has been learned before sleep.

Conclusions and general discussion

Classification accuracy estimated by cross-validation is a random variable and is affected by the data properties (e.g. sample size and number of features), classifier properties (e.g. linear or nonlinear), and the number of folds. Although errors estimated by cross-validation are unbiased, they demonstrate large variability when sample size is low (Braga-Neto and Dougherty, 2004; Isaksson, et al., 2008). More importantly, the distribution of classification accuracies estimated by cross-validation do not follow well known parametric distributions which makes interpretation of cross-validation by means symmetric confidence intervals obsolete (Jamalabadi, et al., 2016). Here, we investigated the properties of accuracy distributions which are expected from cross-validation in common experimental setting, i.e. low sample size (LSS) – Low effect size (LES) data sets. We report a number of intriguing observations regarding the safety of employing cross-validation for data sets in neuroimaging. We propose a few guidelines which simplify use of MVPA and provide lower false positive rate and higher statistical power.

Below chance classification accuracy

When using k-fold cross-validation to estimate CCR, data are divided into training and test data sets *k* times. Therefore, training and test data sets used to

estimate accuracy are not exclusive. The dependency between training and test data induces a number of counterintuitive properties, especially when the effect size and sample size is low. Particularly, in chapter 2, we reported a prominent anomaly of cross-validated classification. We showed that in any linear classification problem using cross-validation, even in one-dimensional data sets, classification results below random guessing levels can occur and that expected values of the classification rates for LSS-LES data are below chance levels for all values of k. We analytically proved and found evidence in simulations that systematic below chance predictions occur when the effect size is low. More specifically, our analytical results state that the probability of correct classification for data sets with no effect, will always be below the chance level regardless of the data distribution. This can be explained by anticorrelation of means of training and test sets in cross-validation when the effect size is low, which makes any linear classifier to decide wrong.

Moreover, we showed that in addition to CCR, area under the curve (AUC) is similarly affected by the dependence of the sub-sample means. Using classification thresholds as it is done in AUC does not prohibit negative correlations between test and training means. Since every point on the receiver operating characteristics (ROC) curve corresponds to one CCR, below chance CCRs represent ROC curves mostly below the chance level division of true positive rate (TPR) = false positive rate (FPR). Therefore, the corresponding AUC will be below chance level.

The frequency and depth of below chance classification rates changes as a function of number of folds (k) in cross-validation. Generally, increasing k results in fewer, but more extreme below chance classifications. In particular, the below chance classification accuracies are less likely to happen when using LOO, however, if they do, they are often much lower than for 2-fold procedure.

Skewed accuracy distributions

Our results in chapter 2 indicate that when classifiers are used for decoding LSS-LES data sets, the classification outcome does not follow a binomial or symmetric distribution, but rather has a skewed probability distribution whose properties depend mainly on the sample size, the number of folds in cross-validation, and the choice of classifier. The variance of the distribution increases with decreasing sample size and effect size, and thus distorting the CCR distribution even further. The size of skewness depends on the choice of classifier. It is more prominent for linear classifiers compared to nonlinear ones. The sign of skewness changes disproportionally from positive with low number of folds (e.g. 2-fold) to negative with increasing number of folds (e.g. LOO).

Because of the skewness of the CCR distribution, the peak and the mean of the distribution are different. In other words, the mode of the distribution is shifted to above or below its expected value. For example, the mode of the CCR distribution for LOO which shows a strong skewness to the left is above 50%, resulting in spuriously high CCRs even if there is no effect in the data. As a result of the skewness, the majority of classification accuracies from data sets with the same true effect size will systematically either overestimate or underestimate the expected CCR. The deviation from the expected value depend on many properties and cannot be directly estimated from parametric tests. In fact, any significance test that does not take the skewed distribution of CCRs into account has a high risk to result in a false positive finding. Therefore, it is crucial not to interpret CCRs in absolute terms, but to compare them to a suitable null distribution which is based on nonparametric resampling approaches (Jamalabadi, et al., 2016).

Accuracy and sensitivity depend on the number of folds

Number of folds in cross-validation affects classification accuracy, most strongly in LES-LSS data. For small to medium effect sizes, classification accuracies based on cross-validation with lower number of folds (e.g. 2-fold) are on average lower

compared to when higher number of folds (e.g. LOO) is used. Although this is desirable in a context of single item classification, when the presence of a class difference is known, when MVPA is used for the purpose of hypothesis testing and therefore, the presence of a class difference is yet to be tested, higher accuracies do not necessarily support more robust or more statistically powerful finding. In fact, because CCRs have lower variance in 2-fold cross-validation, especially in null-distribution, they are often more sensitive and reach significance threshold with smaller effect sizes compared to when LOO is used. In addition, although CCRs from higher number of folds are higher on average when they are above chance, they can be much lower than 2-fold if they are below chance which strongly biases group averaging over a set of subjects, session, or data sets. Taken together, cross-validated classification with lower number of folds shows higher sensitivity in detecting an effect and is therefore, preferable for hypothesis testing purposes (Jamalabadi, et al., 2016).

Biased accuracies in data with nested subclasses

In the context of decoding neural activity, when data are recorded from two or more experimental conditions, the main interest is to detect class-related effect which generalizes well over trials, sessions, and subjects. In other words, only the contribution of those class-related information which are not specific to a subclass of data should be considered to reject the null hypothesis. However, when data are analyzed using MVPA, the classification algorithm leverages all the information contained in the data to maximize CCR. Any systematic differences between subclasses of trials (e.g. shared physical properties among a subgroup of stimuli, shared experimental settings, trials specific to the same session or subject, etc.) form distinct subclasses within each class, particularly in high dimensional feature space. MVPA algorithms which are sensitive to any kind of structure in the data use such groupings to increase classification accuracies even if there is no difference between the classes on a group level (Jamalabadi, et al., in prep; Alizadeh, et al., 2017). This can be specifically problematic for studies where MVPA is used for hypothesis testing. Importantly,

because the increased classification accuracy in such data sets is not based on class differences, the null distribution should be adjusted to account for accuracy biases. Notably, the biased accuracy is specific to nested subclasses of trials, i.e., when subclasses in two experimental conditions are class specific.

The CCR bias in data sets with nested subclasses is directly related to the number of subclasses and the intraclass correlation (ICC) which is defined as the ratio of subclass-to-trial variance. According to our simulations and the analytical solution, CCRs are most biased when the number of subclasses in each class is low and ICC is high. This bias is more prominent when the effect size is low. We proposed a method that can account for subclass bias by adjusting permutation tests to consider the subclass structure of the data, using subclass-level randomization. Our proposed nonparametric resampling algorithm provides exact p-values when the number of subclasses are more than 5. We noticed that for data sets with 5 or less number of subclasses, an exact p-value can only be estimated on a group level. All in all, our results suggest that subclass effects should be a general concern for all neuroimaging studies. Even in cases where the nuisance effect is balanced across conditions, they can drive the classification accuracy to significantly higher than chance. Therefore, we propose that studies which use data with subclasses and employ MVPA to decode brain activity, should be evaluated with an adjusted null distribution which addresses concerns about specificity and invariance of the findings.

Hypothesis testing based on classification accuracy

Often, correct classification rates (CCRs) are interpreted as a measure of how well the classifier performs or of how strong the effect under investigation is. However, the question of what it means if an experimental condition can be successfully decoded from the control should be approached with care (Haynes, 2015; Jamalabadi, et al., 2016; Jamalabadi, et al., in prep). Here, we propose a few important guidelines that should be observed when MVPA is used for hypothesis testing. Most importantly, we propose that the existence of an effect

should not be determined by the classification rate, but rather by statistical significance (quantified by p-value), and significance should not be based on parametric tests, but on Monte Carlo methods which are fully customized to accommodate not only the structure of the data (Jamalabadi, et al., 2016), but also the experimental design (Jamalabadi, et al., in prep). We find out that CCRs systematically deviate from their expected value, making the interpretation of CCRs in terms of absolute values impossible. In fact, when classification is used for hypothesis testing, the absolute height of CCRs can be misleading about the existence of class-related effect in two ways. First, null-distributions from linear classification which are combined with cross-validation are skewed and have their peak accuracy (mode) different from their mean. The skewness depends on the number of folds and also on the covariance structure of the data which often cannot be correctly estimated in the context of low sample size data set (Jamalabadi, et al., 2016). Second, CCRs in nested experimental designs, which are common in neuroscience, largely overestimate the true size of effect (Jamalabadi, et al., in prep). Subclasses of any kind (e.g. classes of stimuli, trials belonging to different subjects or sessions, etc) inject systematic dependencies in the data structure that leads to spuriously high CCRs. It can happen that a lower CCR represents a more robust result, showing a higher significance level when estimated from the unbiased null-distribution (Jamalabadi, et al., in prep).

In principle, we agree that under exact identical conditions, a higher CCR on average represents a larger difference between classes. However, because conditions (number of features, number of trials, sample size, covariance matrices, type of cross-validation, etc.) are in practice rarely identical, there are very few cases where interpretation or comparison of study results should be based on CCR. Therefore, comparing classifier performance between two experiments based on CCR will most likely lead to incorrect conclusions. Moreover, because of the skewness of the distribution, intuitive interpretations of CCRs will often be misleading, even if identical experimental designs and analyses are compared.

In summary, our analyses warrant the conclusion that CCRs do not reflect the size of the effect under investigation nor the classifier's sensitivity. We conclude that for the hypothesis testing purposes, the height of CCR is irrelevant and existence of an effect should only be determined by the statistical significance based on permutation test.

Decoding continuous sleep EEG across subjects

In chapter 4, we employed MVPA to investigate the neural signatures of material-specific memory reprocessing in human sleep EEG data (Schonauer, et al., 2017). Here, the aim of the study was to test whether EEG activity during sleep contains information about the kind of previously learned visual stimuli. We hypothesized that if the type of visual stimuli can be decoded only based on sleep EEG data, then EEG contains information about the materials learned prior to sleep.

One of the challenges of sleep data analysis is the high dimensionality of the data and the difficulty to record large sample sizes. As a result, while the data was extremely high dimensional (128 channels, recorded with 1 KHz sampling rate during 8 hours of sleep), the sample size was confined to 32 subjects. Another challenge was to generalize information across subjects. Because EEG activity differs greatly between different sleep stages and even more so between two nights of one subject, activity cannot be compared directly between these states. We therefore used between subject analyses to compare recordings from the same sleep stage by applying cross-validation on the subject level. This type of analysis ensures us that the pattern detected by MVPA, is only related to the conditions under study and is not driven by the confounding factors. However, since variabilities between subjects are much larger than the variabilities across conditions, finding a generalizable pattern of activity across subjects becomes increasingly difficult using the conventional MVPA methods. To address these problems, we developed a classification framework along with a specific preprocessing procedure that is optimized for three purposes: 1) to increase

signal-to-noise ratio, 2) to reduce the dimensionality of the data, and 3) to adapt the signal better to between-subject classification. More specifically, we proposed a two-step classification algorithm based on an ensemble of linear support vector machines (SVM) classifiers which learns the spatial and temporal components of neural activity separately and then aggregates the two components of information to build a classification hyperplane using a linear SVM. In fact, instead of training one linear SVM on the high dimensional spatiotemporal feature space which lead to overfitting, we used the spatial and temporal features in two successive stages that could serve as a feature reduction method and at the same time increases the signal-to-noise ratio. In addition, we devised a preprocessing technique which reduces the between-subject variabilities and therefore allows for a better comparison of EEG across subjects. The pipeline of data preprocessing involved down sampling EEG channels from 128 to 32, averaging over EEG trials from each subject/condition, and removing amplitude differences between channels and subjects. The latter was done by applying a spectral sharpening filter to remove the baseline spectrum and emphasize differences between neighboring frequencies. Importantly, both the preprocessing and two-step classification are vital for optimal performance and removing any of these steps results in a deteriorated classification accuracy.

We used this method to see whether human sleep EEG contains any information about what has been learned before sleep. We find significant and generalizable learning-related processing in the EEG in all sleep stages, which occurs during specific time windows (2 and 5 hours after sleep onset) and which also correlates with later recall performance. We track the reprocessing in both rapid eye movement (REM) and non-REM (NREM) sleep but its spatial distribution over the scalp and its frequency composition differ between NREM and REM sleep. Interestingly, reprocessing in both sleep stages is cyclic in nature, and may be timed to windows of maximal synaptic efficacy. We showed that it is possible to classify long continuous EEG data recorded from sleep. In addition, we have

further used our method in a number of data sets dealing with short terms memory and our results show applying this method significantly improves classification accuracy (Schonauer, et al., 2017; Schonauer, et al., in prep).

Limitations and outlook

In this thesis, we studied the use of MVPA in high-density EEG data for the purpose of hypothesis testing. We put forward a new algorithm for decoding low sample size continuous EEG data and proposed a few guidelines to better interpret classification results. However, there are two related questions whose answers go beyond the scope of this thesis and can considered as interesting follow ups for the present research.

Interpretation of accuracy maps

MVPA is repeatedly criticized for being a poor tool to localize information (Anderson and Oates, 2010). Although we have a convincing answer to the question of "is there information in the neural activity about the experimental conditions", but we lack a proper procedure to answer questions like "where is the main components of the information encoded". That is, in case of successful decoding (significant above chance CCR) on the whole set of available features, it remains an open question to pinpoint which subset of features are most strongly indicative of the quality and quantity of decoded information (Haynes, 2015). At the moment, there are three main procedures in MVPA literature for this purpose: First is to use a searchlight on the spatiotemporal feature space (Kriegeskorte, et al., 2006), second is to use classification weights (Haufe, et al., 2014; Lee, et al., 2010), and the third is to employ permutation statistics on a subset of features (Ojala and Garriga, 2010). However, the results of these tools depend on the choice of classifier or the size of feature input and often their outputs show marked differences.

I think a convincing approach should explicitly address two main issues: First, given that the rationale behind MVPA is that information is necessarily encoded in a multivariate pattern, to what extend does it make sense to distinctively localize information? Second, any proper method for information localization should provide a procedure to properly discriminate noise from information. That is, assuming that adding a certain feature improves overall classification accuracy, how to determine if this feature is informative per se or it contributes to the classification by reducing the noise?

Effects of correlation

EEG time series and features are often correlated because of the intrinsic correlation in brain activity and because of correlated noise (Averbeck, et al., 2006). Due to this correlation, the covariance structure is of the data would not be diagonal and the features should not be analyzed independently. However, when decoding experimental conditions, correlations between features and trials are often neglected. In this thesis (chapters 2-3), we showed that these correlations affect the distribution of classification accuracies and therefore should be accounted accordingly when the results are interpreted (Jamalabadi, et al., 2016; Jamalabadi, et al., in prep.).

In principle, these correlations are part of the information encoded in the brain activity and therefore carry independent information (Averbeck, et al., 2006) which can be exploited by the classifier. Although the performance of classification algorithms in data with uncorrelated features asymptotically approaches chance level when the number of features increase (Bickel and Levina, 2004; Clarke, et al., 2008; Fan and Fan, 2008; Hall, et al., 2005), adding more correlated features may actually improve decoding performance (Averbeck, et al., 2006). However, the exact effects of correlation on the amount of information is case specific and depends on the structure of noise and class related effect (see Figure 4). Therefore, aside from correlation effects on the statistical interpretation of MVPA results, it is a crucial step to investigate the

boundary conditions of when and how the correlations should be used for decoding.

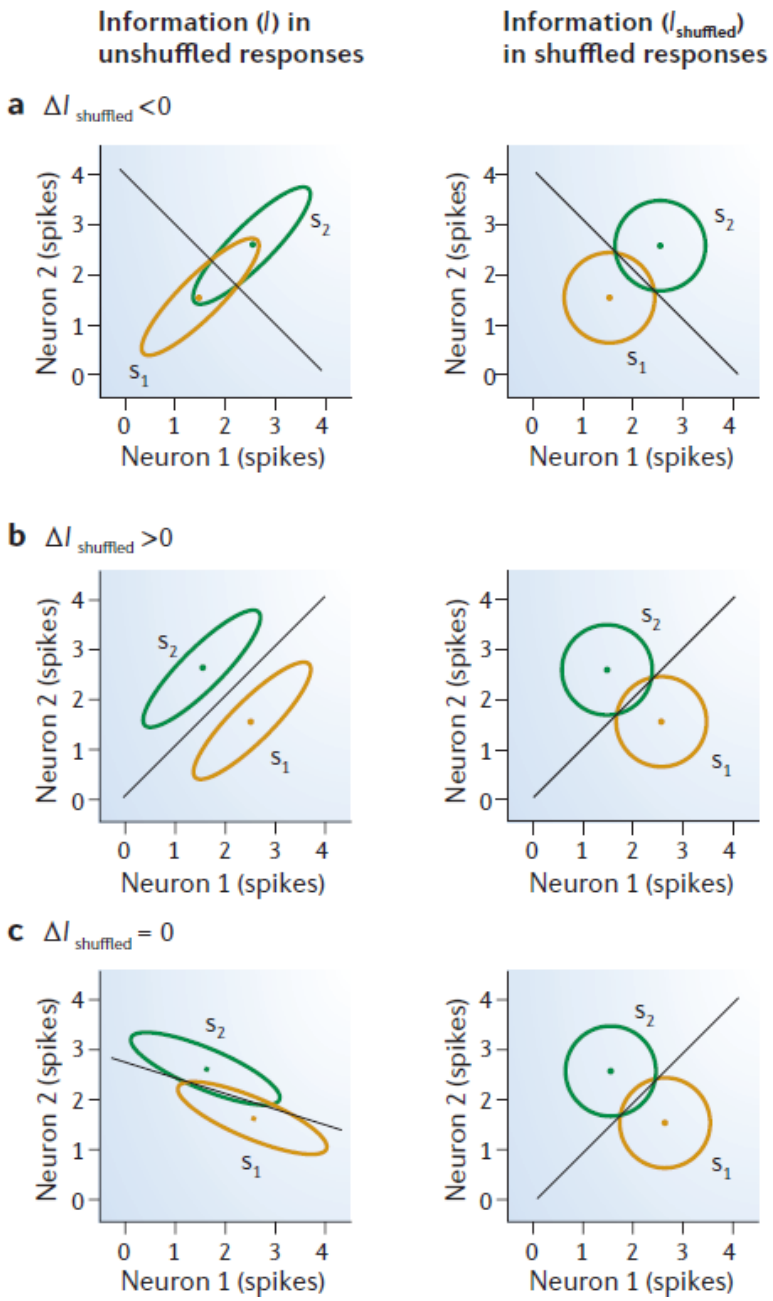


Figure 4: Correlation affects the amount of information which is encoded in the data. Denoting the information in correlated data by I, and the information in uncorrelated data by $I_{shuffled}$, the difference ($\Delta I_{shuffled} = I - I_{shuffled}$) demonstrates how correlation changes the amount of information. Importantly, $\Delta I_{shuffled}$ can take any values (positive, zero, negative) depending on the structure of noise. Therefore, ignoring correlation might result in suboptimal decoding of the experimental conditions. Figure reprinted with permission from (Averbeck, et al., 2006).

References

Alizadeh, S., Jamalabadi, H., Schonauer, M., Leibold, C., Gais, S. (2017) Decoding cognitive concepts from neuroimaging data using multivariate pattern analysis. NeuroImage, 159:449-458.

Anderson, M.L., Oates, T. (A critique of multi-voxel pattern analysis). In; 2010. p 1511-16.

Attoor, S.N., Dougherty, E.R. (2004) Classifier performance as a function of distributional complexity. Pattern Recogn, 37:1641-1651.

Averbeck, B.B., Latham, P.E., Pouget, A. (2006) Neural correlations, population coding and computation. Nat. Rev. Neurosci., 7:358-366.

Bickel, P.J., Levina, E. (2004) Some theory for Fisher's linear discriminant function, 'naive Bayes', and some alternatives when there are many more variables than observations. Bernoulli, 10:989-1010.

Bishop, C.M. (1995) Neural Networks for Pattern Recognition. Oxford University Press, Inc. 482 p.

Braga-Neto, U.M., Dougherty, E.R. (2004) Is cross-validation valid for small-sample microarray classification? Bioinformatics, 20:374-80.

Button, K.S., Ioannidis, J.P., Mokrysz, C., Nosek, B.A., Flint, J., Robinson, E.S., Munafo, M.R. (2013) Power failure: why small sample size undermines the reliability of neuroscience. Nat. Rev. Neurosci., 14:365-76.

Clarke, R., Ressom, H.W., Wang, A., Xuan, J., Liu, M.C., Gehan, E.A., Wang, Y. (2008) The properties of high-dimensional data spaces: implications for exploring gene and protein expression data. Nat. Rev. Cancer, 8:37-49.

Cox, D.D., Savoy, R.L. (2003) Functional magnetic resonance imaging (fMRI) "brain reading": detecting and classifying distributed patterns of fMRI activity in human visual cortex. Neuroimage, 19:261-70.

Doesburg, S.M., Green, J.J., McDonald, J.J., Ward, L.M. (2009) Rhythms of consciousness: binocular rivalry reveals large-scale oscillatory network dynamics mediating visual perception. PLoS One, 4:e6142.

Duda, R.O., Hart, P.E., Stork, D.G. (2000) Pattern Classification (2nd Edition). Wiley-Interscience.

Duzel, E., Penny, W.D., Burgess, N. (2010) Brain oscillations and memory. Curr Opin Neurobiol, 20:143-9.

Fan, J.Q., Fan, Y.Y. (2008) High Dimensional Classification Using Features Annealed Independence Rules. Ann. Stat., 36:2605-2637.

Fisher, R.A. (1936) The use of multiple measurements in taxonomic problems. Ann. Eugen., 7:179-188.

Fuentemilla, L., Penny, W.D., Cashdollar, N., Bunzeck, N., Duzel, E. (2010) Theta-coupled periodic replay in working memory. Curr. Biol., 20:606-12.

Haegens, S., Cousijn, H., Wallis, G., Harrison, P.J., Nobre, A.C. (2014) Inter- and intra-individual variability in alpha peak frequency. NeuroImage, 92:46-55.

Hall, P., Marron, J.S., Neeman, A. (2005) Geometric representation of high dimension, low sample size data. J. R. Stat. Soc. Ser. B-Stat. Methodol., 67:427-444.

Hastie, T.T., R.; Friedman, J. (2001) The Elements of Statistical Learning. New York. Springer.

Haufe, S., Meinecke, F., Gorgen, K., Dahne, S., Haynes, J.D., Blankertz, B., Biessgmann, F. (2014) On the interpretation of weight vectors of linear models in multivariate neuroimaging. NeuroImage, 87:96-110.

Haxby, J.V., Connolly, A.C., Guntupalli, J.S. (2014) Decoding Neural Representational Spaces Using Multivariate Pattern Analysis. Annu Rev Neurosci, 37:435-456.

Haxby, J.V., Gobbini, M.I., Furey, M.L., Ishai, A., Schouten, J.L., Pietrini, P. (2001) Distributed and overlapping representations of faces and objects in ventral temporal cortex. Science, 293:2425-30.

Haynes, J.D. (2015) A Primer on Pattern-Based Approaches to fMRI: Principles, Pitfalls, and Perspectives. Neuron, 87:257-270.

Haynes, J.D., Rees, G. (2006) Decoding mental states from brain activity in humans. Nat. Rev. Neurosci., 7:523-34.

Horikawa, T., Tamaki, M., Miyawaki, Y., Kamitani, Y. (2013) Neural Decoding of Visual Imagery During Sleep. Science, 340:639-642.

Isaksson, A., Wallman, M., Goransson, H., Gustafsson, M.G. (2008) Cross-validation and bootstrapping are unreliable in small sample classification. Pattern Recogn Lett, 29:1960-1965.

Jafarpour, A., Horner, A.J., Fuentemilla, L., Penny, W.D., Duzel, E. (2013) Decoding oscillatory representations and mechanisms in memory. Neuropsychologia, 51:772-80.

Jamalabadi, H., Alizadeh, S., Schonauer, M., Leibold, C., Gais, S. (2016) Classification based hypothesis testing in neuroscience: Below-chance level classification rates and overlooked statistical properties of linear parametric classifiers. Hum. Brain Mapp., 37:1842-55.

Jamalabadi, H., Alizadeh, S., Schonauer, M., Leibold, C., Gais, S. (In prep.) Adjusting permutation tests in multivariate anaylsis of neuroimaging data with subclasses: introduction of biased null distributions.

Kamitani, Y., Tong, F. (2005) Decoding the visual and subjective contents of the human brain. Nat. Neurosci., 8:679-85.

Kay, K.N., Naselaris, T., Prenger, R.J., Gallant, J.L. (2008) Identifying natural images from human brain activity. Nature, 452:352-5.

Kohavi, R. (1995) A study of cross-validation and bootstrap for accuracy estimation and model selection. Proceedings of the 14th international joint conference on Artificial intelligence - Volume 2. Montreal, Quebec, Canada: Morgan Kaufmann Publishers Inc. p 1137-1143.

Kriegeskorte, N., Goebel, R., Bandettini, P. (2006) Information-based functional brain mapping. Proc. Natl. Acad. Sci. U. S. A., 103:3863-8.

Kriegeskorte, N., Simmons, W.K., Bellgowan, P.S., Baker, C.I. (2009) Circular analysis in systems neuroscience: the dangers of double dipping. Nature neuroscience, 12:535-40.

LaConte, S., Strother, S., Cherkassky, V., Anderson, J., Hu, X. (2005) Support vector machines for temporal classification of block design fMRI data. Neuroimage, 26:317-29.

LaRocque, J.J., Lewis-Peacock, J.A., Drysdale, A.T., Oberauer, K., Postle, B.R. (2013) Decoding attended information in short-term memory: an EEG study. J Cogn Neurosci, 25:127-42.

Lee, S., Halder, S., Kubler, A., Birbaumer, N., Sitaram, R. (2010) Effective functional mapping of fMRI data with support-vector machines. Human brain mapping, 31:1502-11.

Lemm, S., Blankertz, B., Dickhaus, T., Muller, K.R. (2011) Introduction to machine learning for brain imaging. NeuroImage, 56:387-99.

Lopes da Silva, F. (2013) EEG and MEG: relevance to neuroscience. Neuron, 80:1112-28.

Misaki, M., Kim, Y., Bandettini, P.A., Kriegeskorte, N. (2010) Comparison of multivariate classifiers and response normalizations for pattern-information fMRI. NeuroImage, 53:103-18.

Mitchell, T.M., Hutchinson, R., Niculescu, R.S., Pereira, F., Wang, X.R., Just, M., Newman, S. (2004) Learning to decode cognitive states from brain images. Machine Learning, 57:145-175.

Mitchell, T.M., Shinkareva, S.V., Carlson, A., Chang, K.M., Malave, V.L., Mason, R.A., Just, M.A. (2008) Predicting human brain activity associated with the meanings of nouns. Science, 320:1191-1195.

Mur, M., Bandettini, P.A., Kriegeskorte, N. (2009) Revealing representational content with pattern-information fMRI--an introductory guide. Social cognitive and affective neuroscience, 4:101-9.

Newman, E.L., Norman, K.A. (2010) Moderate excitation leads to weakening of perceptual representations. Cereb. Cortex, 20:2760-70.

Nichols, T.E., Holmes, A.P. (2002) Nonparametric permutation tests for functional neuroimaging: a primer with examples. Hum Brain Mapp, 15:1-25.

Noh, E., Herzmann, G., Curran, T., de Sa, V.R. (2014) Using single-trial EEG to predict and analyze subsequent memory. NeuroImage, 84:712-23.

Noirhomme, Q., Lesenfants, D., Gomez, F., Soddu, A., Schrouff, J., Garraux, G., Luxen, A., Phillips, C., Laureys, S. (2014) Biased binomial assessment of cross-validated estimation of classification accuracies illustrated in diagnosis predictions. NeuroImage. Clinical, 4:687-94.

Norman, K.A., Polyn, S.M., Detre, G.J., Haxby, J.V. (2006) Beyond mind-reading: multi-voxel pattern analysis of fMRI data. Trends Cogn. Sci., 10:424-30.

Ojala, M., Garriga, G.C. (2010) Permutation Tests for Studying Classifier Performance. J. Mach. Learn. Res., 11:1833-1863.

Pereira, F., Botvinick, M. (2011) Information mapping with pattern classifiers: a comparative study. Neuroimage, 56:476-96.

Pereira, F., Mitchell, T., Botvinick, M. (2009) Machine learning classifiers and fMRI: a tutorial overview. NeuroImage, 45:S199-209.

Rasch, B., Born, J. (2013) About sleep's role in memory. Physiol. Rev., 93:681-766.

Rissman, J., Greely, H.T., Wagner, A.D. (2010) Detecting individual memories through the neural decoding of memory states and past experience. Proc. Natl. Acad. Sci. U. S. A., 107:9849-54.

Rodriguez, E., George, N., Lachaux, J.P., Martinerie, J., Renault, B., Varela, F.J. (1999) Perception's shadow: long-distance synchronization of human brain activity. Nature, 397:430-3.

Rodriguez, J.D., Perez, A., Lozano, J.A. (2010) Sensitivity analysis of k-fold cross validation in prediction error estimation. IEEE transactions on pattern analysis and machine intelligence, 32:569-75.

Schaefer, R.S., Farquhar, J., Blokland, Y., Sadakata, M., Desain, P. (2011) Name that tune: decoding music from the listening brain. Neuroimage, 56:843-9.

Schonauer, M., Alizadeh, S., Jamalabadi, H., Abraham, A., Pawlizki, A., Gais, S. (2017) Decoding material-specific memory reprocessing during sleep in humnas. Nat. Commun., 8:15404.

Schonauer, M., Alizadeh, S., Jamalabadi, H., Emmersberger, M., Gais, S. (In prep) Decoding retrieval success and memory content during short-term memory maintenance.

Schulz, E., Zherdin, A., Tiemann, L., Plant, C., Ploner, M. (2012) Decoding an individual's sensitivity to pain from the multivariate analysis of EEG data. Cerebral cortex, 22:1118-23.

Schwarzlose, R.F., Swisher, J.D., Dang, S., Kanwisher, N. (2008) The distribution of category and location information across object-selective regions in human visual cortex. Proc. Natl. Acad. Sci. U. S. A., 105:4447-52.

Stelzer, J., Chen, Y., Turner, R. (2013) Statistical inference and multiple testing correction in classification-based multi-voxel pattern analysis (MVPA): random permutations and cluster size control. Neuroimage, 65:69-82.

Takeuchi, T., Duszkiewicz, A.J., Morris, R.G. (2014) The synaptic plasticity and memory hypothesis: encoding, storage and persistence. Philos Trans R Soc Lond B Biol Sci, 369:20130288.

Todd, M.T., Nystrom, L.E., Cohen, J.D. (2013) Confounds in multivariate pattern analysis: Theory and rule representation case study. Neuroimage, 77:157-165.

Tong, F., Pratte, M.S. (2012) Decoding Patterns of Human Brain Activity. Annu. Rev. Psychol., 63:483-509.

Walker, M.P., Stickgold, R. (2006) Sleep, memory, and plasticity. Annu Rev Psychol, 57:139-66.

List of publications in this thesis

H. Jamalabadi, S. Alizadeh, M. Schonauer, C. Leibold, S. Gais (2016), "Classification based hypothesis testing in neuroscience: below-chance level classification rates and overlooked statistical properties of linear parametric classifiers", *Human Brain Mapping*, 37:1842-55.

H. Jamalabadi, S. Alizadeh, M. Schonauer, C. Leibold, S. Gais, "Adjusting permutation test for multivariate analysis of neuroimaging data with subclasses: introduction of biased null distributions", in submission process.

M. Schonauer*, S. Alizadeh*, **H. Jamalabadi**, A. Abraham, A. Pawlizki, S. Gais (2017), "Decoding material-specific memory reprocessing during sleep in humans". *Nature Communications,* 8:15404 (*equal contribution)

Statement of contributions

H. Jamalabadi, S. Alizadeh, M. Schonauer, C. Leibold, S. Gais (2016), "Classification based hypothesis testing in neuroscience: below-chance level classification rates and overlooked statistical properties of linear parametric classifiers", *Human Brain Mapping*, 37:1842-55.

HJ implemented the simulations and analyzed the synthetic data sets. MS and SG collected the EEG data. HJ and SA analyzed the data from EEG experiments. HJ and CL with the help of SA, MS, and SG developed the mathematical framework. CL provided the proofs in appendix A and B. HJ provided the proof in appendix C. HJ wrote the first draft of the manuscript. SA, MS, CL, and SG revised the manuscript.

H. Jamalabadi, S. Alizadeh, M. Schonauer, C. Leibold, S. Gais, "Adjusting permutation test for multivariate analysis of neuroimaging data with subclasses: introduction of biased null distributions", in submission process.

HJ, SA, and SG recognized biased classification accuracies in data with subclasses and developed a method to account for this bias by adjusting the respective null distribution. HJ modeled the subclass effect using a data structure with three-levels variance. HJ implemented the simulations and analyzed the synthetic data sets. MS and SG collected the EEG data. HJ and SA analyzed the data from EEG experiments. HJ and CL with the help of SA, SG, and MS defined the problem in a mathematical framework. CL provided the proof in Appendix A. HJ provided the proof in Appendix B. HJ wrote the first draft of the manuscript. SA, MS, CL, and SG revised the manuscript.

M. Schonauer*, S. Alizadeh*, **H. Jamalabadi**, A. Abraham, A. Pawlizki, S. Gais (2017), "Decoding material-specific memory reprocessing during sleep in humans". *Nature Communications,* 8:15404 (*equal contribution)

MS, AP, and SG planned and designed the experiments. MS, AA, and AP collected the data. HJ and SA developed and implemented the two-step classification framework for the between-subject classification. SA with the help of HJ developed and implemented the companion preprocessing steps dedicated for between-subject classification. SA and HJ classified the data and performed permutation test. MS and SG analyzed the behavioral data. SA and MS provided the analysis to relate behavioral data and machine learning output. HJ with the help of SA implemented the time domain analysis. SA wrote the first draft of the method part of the paper. MS and SG wrote the first draft of the content part of the manuscript. Revisions were done by HJ, SA, MS, and SG.

Chapter 2:

Classification based hypothesis testing in neuroscience: Below-chance level classification rates and overlooked statistical properties of linear parametric classifiers

Hamidreza Jamalabadi, Sarah Alizadeh, Monika Schönauer, Christian Leibold, and Steffen Gais

Published in Human Brain Mapping (March 2016)

H. Jamalabadi, S. Alizadeh, M. Schonauer, C. Leibold, S. Gais (2016), "Classification based hypothesis testing in neuroscience: below-chance level classification rates and overlooked statistical properties of linear parametric classifiers", *Human Brain Mapping*, 37:1842-55.

Abstract

Multivariate pattern analysis (MVPA) has recently become a popular tool for data analysis. Often, classification accuracy as quantified by correct classification rate (CCR) is used to illustrate the size of the effect under investigation. However, we show that in low sample size (LSS), low effect size (LES) data, which is typical in neuroscience, the distribution of CCRs from cross-validation of linear MVPA is asymmetric and can show classification rates considerably below what would be expected from chance classification. Conversely, the mode of the distribution in these cases is above expected chance levels, leading to a spuriously high number of above chance CCRs. This unexpected distribution has strong implications when using MVPA for hypothesis testing. Our analyses warrant the conclusion that CCRs do not well reflect the size of the effect under investigation. Moreover, the skewness of the null-distribution precludes the use of many standard parametric tests to assess significance of CCRs. We propose that MVPA results should be reported in terms of p-values, which are estimated using randomization tests. Also, our results show that cross-validation procedures using a low number of folds, e.g. 2-fold, are generally more sensitive, even though the average CCRs are often considerably lower than those obtained using a higher number of folds.

Introduction

Multivariate pattern analysis (MVPA) is becoming more and more mainstream for classification, decoding and hypothesis testing in neuroscientific data analysis (Damarla and Just, 2013; Deuker, et al., 2013; Duarte, et al., 2014; Haynes and Rees, 2006; Kamitani and Tong, 2005; Norman, et al., 2006; Staresina, et al., 2013), with linear classifiers being the most successful ones (Clarke, et al., 2008; Lemm, et al., 2011). Whereas classical statistical approaches search for one or more features in a data set that independently allow to distinguish between experimental conditions or groups, multivariate classification algorithms analyze data sets in a way that takes into account all the available information contained therein. Therefore, they provide increased sensitivity compared to classical methods, which are based on multiple univariate comparisons. Because a classifier is usually trained on one portion of the data and then validated on another, it provides an estimation of the generalizability of the learned classification rule and can thus be used for data-driven exploratory analyses as well as for hypothesis-driven testing. These properties make multivariate pattern classification an attractive analysis tool for the neurosciences, where experiments often generate large amounts of multivariate data (Kriegeskorte, et al., 2009; Norman, et al., 2006).

Performance of MVPA algorithms is frequently quantified in terms of correct classification rates (CCRs), which is defined as the percentage of correctly classified items. The most widely used algorithm for CCR estimation is cross-validation (Hastie, 2001). It has a low bias, is simple to implement (Kohavi, 1995) and has been used to estimate classification performance in many applications (Braga-Neto and Dougherty, 2004; Lemm, et al., 2011; Noirhomme, et al., 2014). Cross-validation makes efficient use of all the available data by repeatedly partitioning data into complementary training and test subsets. Thus, this method is especially suitable when there are only few available

samples. Because it strictly separates classifier training from error estimation, it avoids overfitting the classifier to the data set. It also precludes confirming in the same data the hypotheses generated by the classifier during training and thus prevents circularity (Kriegeskorte, et al., 2009). Although these characteristics in principle should guarantee the safe employment of classification with cross-validation for hypothesis testing purposes, it produces only a point estimate of classification accuracy, and its relation to the size of the effect in question is unclear.

Beyond the CCR, we need the confidence interval or better still the distribution of this estimation. Variance of cross-validated CCRs depends on classifier type, on experimental design parameters like sample size and data dimensionality, and on signal-to-noise ratio inherent in the data (Azuaje, 2003; Clarke, et al., 2008; Dougherty, 2001; Raudys and Jain, 1991; Rodriguez, et al., 2010). Neuroscience data usually has a number of properties that cause classification accuracies to have particularly large variances. First, they frequently contain hundreds or thousands of features (e.g. number of voxels in fMRI studies, number of channels times number of frequency bins or time points in EEG recordings, number of genes in cDNA microarray studies, etc.), which leads to a phenomenon known as noise accumulation. Noise accumulation is the increasing difficulty to determine the centroid of data in a space with increasing dimensionality. Second, features often contain only a small amount of information, i.e. the signal-to-noise ratio or effect size is small and differences between classes are hard to detect. Third, although data sets often consist of large numbers of features, the number of samples is often on the order of tens, because it is practically limited by the number of subjects that can participate in a study and by the number of trials each subject can complete within a reasonable time (see Button, et al., 2013 for a related discussion). This again makes accurate estimation of class centroids problematic.

Although MVPA is getting more and more attention in neuroimaging, its intricacies are not yet completely understood. Often, higher CCRs are essentially

interpreted as representing larger effects, generally without further taking the properties of the data (e.g. sample size, number of features) or the classifier into account. To investigate the reliability of MVPA results in hypothesis testing applications, we explore the properties of cross-validated classification accuracies in typical neuroscientific data, which we model as low sample size/low effect size data (LSS-LES). We use simulations and an analytical approach to describe the distribution of classification results with systematically varying sample size, effect size, and number of cross-validation folds. From our findings we draw conclusions regarding the use of cross-validation in MVPA with linear classifiers for hypothesis testing. Our findings show that cross-validation in LSS-LES data possesses some counterintuitive properties that can critically bias interpretation of findings. In particular, we show that CCRs should not be used to display or compare class differences, because a higher CCR does not necessarily represent a larger difference between classes if not all the properties of the data sets are comparable (sample size, number of features, number of trials, type of cross-validation, covariance matrices, etc.)

Method and Results

Case study: Classification of EEG data

The main premise of classification is that expected CCR of a given classifier has a monotonic, albeit nonlinear relation with the amount of signal in the data, i.e. classification of a data set with more information results in higher CCR (Raudys and Pikelis, 1980). In particular, if the collected data from two experimental conditions represent identical distributions, i.e., the null hypothesis is true and thus the effect size is zero, one would expect CCRs near 50% (chance level). However, although this is true provided an asymptotically infinite number of samples, this does not seem to hold for low sample size data, which is usually available for hypothesis testing purposes. We noticed in the literature

and in our own preliminary experiments that classification rates are often much lower than what could be expected even from chance performance, sometimes even approaching 0% (Deuker, et al., 2013; Etzel, et al., 2013; Fuentemilla, et al., 2010; Gisselbrecht, et al., 2013; Noirhomme, et al., 2014). These extreme below chance level CCRs are not generally matched by similarly high CCRs when an analysis is repeated multiple times within a series of analyses, which already hints at an asymmetry in the distribution of CCRs.

First, we consider the analysis of a study that aimed at investigating event related potentials (ERP) elicited by two kinds of visual stimuli in a visual learning task (see Figure 1). In this experiment, EEG was recorded from 20 healthy subjects while two types of stimuli, which were photographs belonging to different semantic categories, were presented. Presentation time was 100 ms, the EEG was recorded from 100 ms before to 900 ms after onset of stimulus presentation. EEG was recorded using an active 128 channel Ag/AgCl-electrode system (ActiCap, Brain products, Gilching, Germany) with 1 kHz sampling frequency and a high-pass filter of 0.1 Hz. Electrodes were placed according to the extended international 10–20 electrode system. Artefacts were rejected in a semiautomatic process using custom MATLAB scripts that made sure that only a minimal number of artefacts remained in the data. Epochs with artefacts were removed from the dataset, channels that contained too many epochs with artefacts were removed and then interpolated using routines provided by EEGLAB (Delorme and Makeig, 2004). We used 30 artefact-free trials of each stimulus category per subject for our within-subject classification procedure. The goal of this study was to investigate whether and which aspects of the ERP in terms of spatial locations and time windows have discriminating power between the two types of stimuli. To answer this question, we used a so-called searchlight approach (Kriegeskorte, et al., 2006), sweeping the spatiotemporal feature space with a window size of 3 cm on-scalp radius for spatial and 20 ms for temporal features, respectively, resulting in 1600 searchlights with approximately 100 features on average. We classified the data in each

searchlight with diagonal Linear Discriminant Analysis (LDA) using Leave-One-Out (LOO) cross-validation. Figure 1b shows the histogram of CCRs for different searchlights for an exemplar subject. It becomes immediately obvious that this distribution has a heavy tail to the left, which is expected for neither zero nor positive effect sizes. To describe the distribution of CCRs quantitatively, we use three different measures: Below Chance Percentage (BCP) denotes the percentage of results that have less than the chance level of 50% correct classifications. If the null hypothesis is true and CCRs are symmetrically distributed, 50% of below chance findings can be expected. A deviation from 50% indicates either a non-zero effect (BCP>50%) or a skewed distribution of CCRs (BCP<>50%). $CCR_{0.05}$ and $CCR_{0.95}$ represent the classification rates of the 5^{th} and 95^{th} percentile of the distribution. Although the average CCR is 52.7%, only about a quarter of CCRs fall below the expected chance level (BCP = 27.5%). On the other hand, the distribution shows a heavier tail on its left than its right side, i.e. below chance values are more extreme than above chance values. Together with the peak of the distribution (mode), which lies to the right of the mean at 55%, these data clearly demonstrate the asymmetry of distributions of classification outcomes: a few extremely low values are weighed against a higher number of above average values.

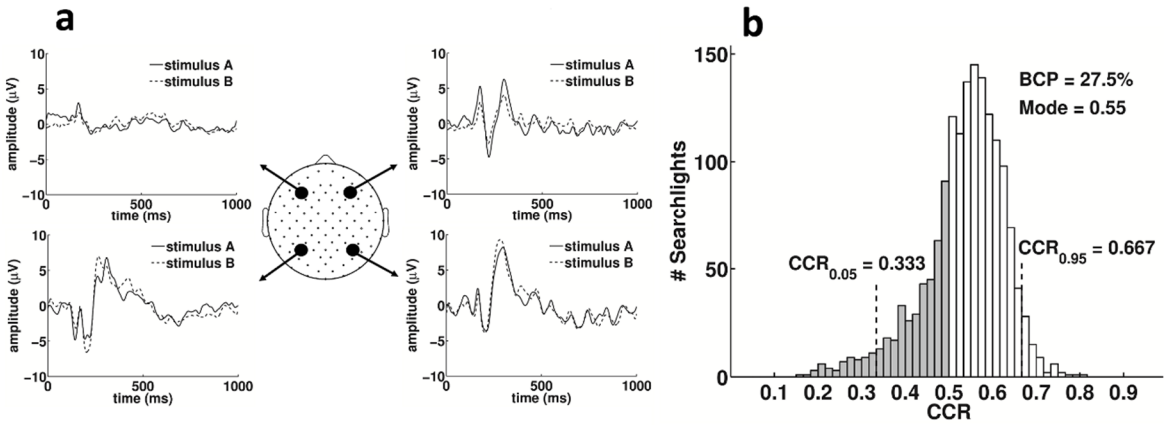


Figure 1: Classification results in an EEG visual learning task. (A) A topo-plot showing the standard electrode locations in high-density EEG recording with four examples of average event-related potentials from 2 frontal and 2 occipital channels. (B) Histogram of correct classification rates for 1600 classifications using a searchlight procedure on spatiotemporal features of the EEG. The distribution shows a strong skewness with a heavy tail on the left.

Classification rates below the level expected for chance

Using synthetically generated data, we have the opportunity to manipulate various parameters, we studied the actual shape of classification rate distributions in 21 series of 10,000 synthetically generated, one dimensional, two-class experiments. Each simulated experiment consisted of $N = 15$ observations per condition which were randomly sampled from two normally distributed populations. Populations had identical variances and a priori determined true effect sizes. The effect size is a measure of the amount of signal, here given as the mean difference between two classes in standard deviation units $\left(d = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{({s_1}^2 + {s_2}^2)/2}}\right)$. For the behavioral sciences, Cohen (1988) defined a value of $d = 0.2$ as a small effect. We classified the data from each experiment using cross-validation and LDA classification. For each particular true effect size, we repeated the whole sampling procedure 10,000 times.

We then pursued two complementary approaches to estimate the expected value of classification accuracies. In the first approach, we averaged over all the CCRs of those experiments that are generated from equal a priori, true effect size (Fig. 2a). In the second one, we averaged over CCRs of experiments with equal a posteriori, estimated effect sizes (Fig. 2b). Thus, we differentiate between the population-based true effect size (d) and the sample-based estimated effect size ($\hat{d}$), which is observed in a specific finite sample. This is particularly relevant for any real data set, where we do not have access to the true effect size. Whereas intuitively, one may assume that expected CCRs are identical for true and estimated effect sizes, this turns out to be true only for large effect sizes. Our simulations on low sample size data exhibit large differences when effect sizes are small. We see in Fig. 2a that expected values of CCRs start from chance level (here 50%) for a true effect size of $d = 0$ and increase nonlinearly with increasing effect size. However, Figure 2b, which shows expected values of CCRs as a function of the observed, sample-based effect sizes, demonstrates that average CCRs in low sample size data can drop far below 50% when the estimated effect size is low, which could explain the unexpectedly high number of very low CCRs around 20% in the empirical data in Figure 1b. In other words, classification of data sets with small estimated effect sizes results on average in below chance level CCRs. For the experiments simulated in Figure 2b, an estimated effect size of $\hat{d} = 0$ results in CCRs between 0% and 65%, and CCRs below 50% occur up to an effect size of $\hat{d} = 1$. This effect size represents a mean difference between conditions of one standard deviation and is already interpreted as a large effect in some fields (Cohen, 1988). Only for even larger effects, expected CCRs for true and estimated effect sizes converge. We further investigated the distribution of estimated effect sizes and plot CCRs for all the experiments with an a priori true effect size of $d = 0$. As expected, estimated effect sizes $\hat{d}$ are symmetrically distributed around zero (Fig. 2c). However, CCRs follow another distribution which is neither normal nor binomial and not even symmetric (Fig. 2d). This observation shows that there is no simple

relation between estimated effect size and CCR and that CCR therefore does not well reflect the size of the effect under study.

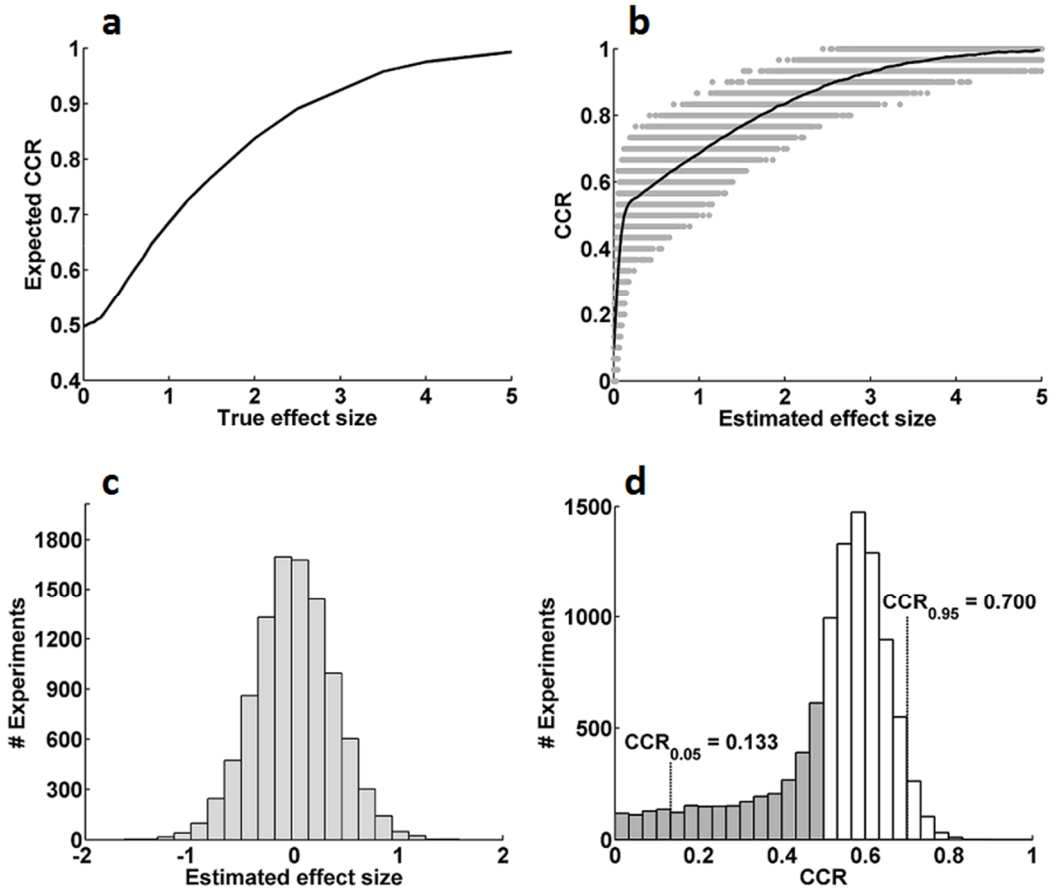


Figure 2: Below chance classification rates in LSS-LES data when classified using LDA with LOO cross-validation. (A) Expected CCRs as a function of true effect size $d \in \{0,0.04,0.08,0.12,0.16,0.20,0.24,0.32,0.44,0.50,0.6,0.7,0.8,1,1.2,1.5,2,2.5,3.5,4,5\}$ provided infinite sample size. (B) The solid line shows the expected CCR as a function of the estimated effect size. Grey dots represent individual experiments. CCRs can only obtain discrete values. (C) Distribution of estimated effect sizes of experiments with a true effect size of zero. (D) Distribution of CCRs that result from experiments with a true effect size of zero. The grey area shows the experiments with classification rates below 50%.

The simulations in Figure 2 show that CCRs reach far below chance levels when sample size and estimated effect size is low. To further investigate the underlying causes of this observation, we develop an analytical description of cross-validated classification rates in a one-dimensional linear classification

paradigm. We assume a data set that consists of N observations for each of two classes A and B with empirical means $\vec{m} = [m_A, m_B]^T$. During k-fold cross-validation the data set is divided into a training and a test set, with means $\vec{\mu} = [\mu_A, \mu_B]$ for the training set and $\vec{v} = [v_A, v_B]$ for the test set. Below chance classification rates can be understood from the dependence of the subsample means. For given sample mean $\vec{m} = (m_A, m_B)^T$, subsample means are mutually dependent. $m_{A,B} = \frac{(k-1)\mu_{A,B}+v_{A,B}}{k}$ yields a negative correlation between test and training means. Thus, if the test mean is a little above the sample mean, the training mean must be a little below and vice versa. If the means of both classes are very similar, the difference of the training means must necessarily have a different sign than the difference of the test means. This effect does not average out across folds, because accuracy in every fold is below 50% irrespective of the direction of shift in group mean. Figure 3 illustrates this effect in a simple one-dimensional, two-class problem. (See Theorem 1 in Appendix A and B for further details and examples.) From this directly follows Corollary 1 (see Appendix C), which states that the probability of correct classification for data sets with no effect ($m_A = m_B$) will always be below the level expected for chance classification.

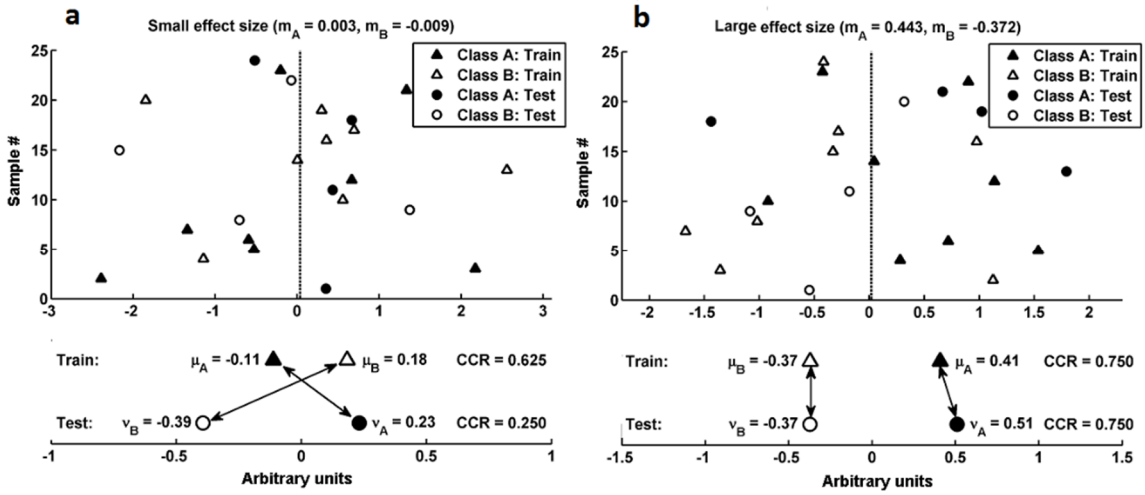


Figure 3: An illustrative example showing how dependency between training and test sets in cross-validation in LSS-LES data results in below chance classification rates. Two experiments with small (A) and large (B) effect size. Each consists of 12 observations per

condition. They were generated by random sampling from two univariate normal distributions. In each experiment, two-thirds of the data is used for training the classifier. Below-chance classification rates are obvious for $m_A \approx m_B$. This can be understood because during cross-validation, the means $\mu_{A,B}$ and $v_{A,B}$ of training and test set are anticorrelated for fixed $m_{A,B}$. If total sample means $m_A \approx m_B$ and training means $\mu_A \gg \mu_B$, then test means must obey $v_A < v_B$. Thus, any classifier relying on linear averages would predict wrong.

Spuriously high classification rates as side effect of below chance CCRs

Theorem 1 implies that CCRs resulting from cross-validation on a given data set are a function of mean difference between classes, number of observations per condition (sample size), and number of folds in cross-validation. Dimensionality of data and classifier type are two additional factors influencing CCRs beyond our analytical derivation. In a series of Monte Carlo experiments with various sample sizes $N \in \{10,15,30,60\}$ and 21 different effect sizes $0 \leq d \leq 5$ (see above), we classified the data using cross-validation with five frequently used classifiers: LDA, linear SVM, Classification And Regression Tree (CART), and 1-Nearest Neighbor classification (1NN). For each particular set of effect size, sample size and classifier type, we repeated the classification procedure 10,000 times once using leave-one-out (LOO) and once using 2-fold cross-validation, which represent both extremes of k-fold cross-validation. Figure 4 summarizes results.

Figures 4a and 4b demonstrate the relationship of BCP, $CCR_{0.05}$ and $CCR_{0.95}$ with true effect size and sample size for LOO and 2-fold cross-validation of LDA classification, respectively. In both cases, as either true effect size or sample size decreases, the probability of below chance classifications increases. In small samples, this value is already high for medium effect sizes (i.e. a mean difference of 0.7 standard deviation units following the conventions of Cohen, 1988). Moreover, the "depth" of below chance classification rates as denoted by $CCR_{0.05}$ and the range of CCRs increases with decreasing true effect size or sample size. Thus, skewness and width of the distribution changes as a function of true effect

size and sample size. Because of the asymmetry of the distribution (see Fig. 2d), some very low CCRs necessarily mean that there must be a larger number of moderately above chance results, even in the no effect case of $d = 0$. Because the expected average CCR over all experiments with $d = 0$ is 50%, each experiment with a CCR of 0% must be counterbalanced by 10 experiments with a CCR of 55%. This high number of positive results is obviously spurious and can be misleading if one is unaware of the skewness of the CCR distribution.

Comparison of different classifiers

Not all classifiers are susceptible to below chance classification rates (Fig. 4c). Only parametric linear classifiers (i.e. linear SVM and LDA) show expected values of their CCRs below the chance level when estimated effect sizes are low. The other two classifiers (Nearest Neighbor and CART), which do not depend on a linear metric, do not show average CCRs below chance level. The disadvantage of these nonlinear classifiers compared to the linear ones becomes obvious at larger effect sizes (here for $0.5 < \hat{d} < 3$), where they have, on average, distinctly lower classification rates.

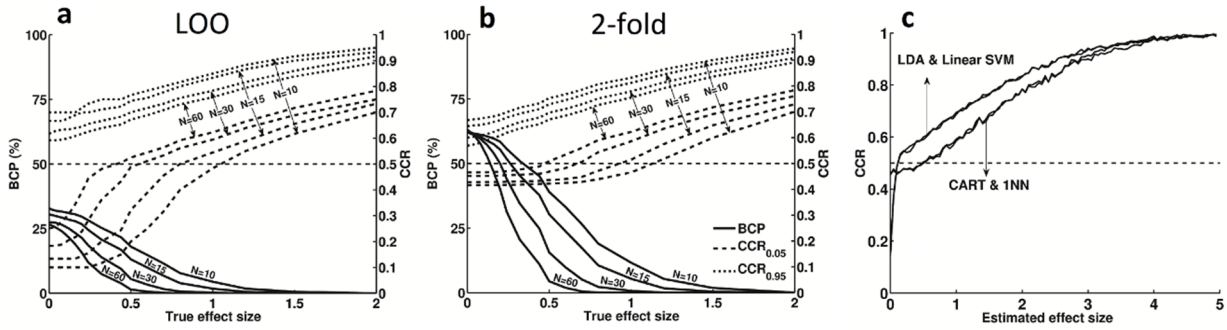


Figure 4: The effect of sample size, effect size, and classifier type on cross-validated classification rates. (A) BCP, $CCR_{0.05}$, and $CCR_{0.95}$ as a function of true effect size and sample size for LDA with LOO cross-validation. BCP is indicated on the left vertical axis, $CCR_{0.05}$ and $CCR_{0.95}$ on the right. (B) BCP, $CCR_{0.05}$, and $CCR_{0.95}$ as a function of true effect size and sample size for LDA with 2-fold cross-validation. (C) Classification results for 3 parametric linear and 2 nonlinear classifiers using LOO cross-validation. Linear classifiers show classification rates considerably below the chance level for small effect sizes, whereas the other two classifiers do not exhibit this phenomenon.

Classification rate versus statistical significance

Our results show that for 2-fold cross-validation and small to medium effect sizes, more than 50% of experiments result in below chance level classification (Figs. 4a and b). On the other hand, CCRs for LOO can be much lower than for 2-fold cross-validation. In the next set of experiments we look at whether a higher or a lower number of cross-validation folds k should be recommended for the use in hypothesis testing, in particular in LSS-LES data. We studied the effect of k on classification performance (CCR) and on its statistical significance. For this purpose, we generated synthetic data sets with the same procedure as before with $N = 15$ observations per condition, which were drawn from two normally distributed populations. Population effect size varied from 0 to 1, estimated effect size was determined from the samples. The whole process was repeated 10,000 times for each set of parameters. Statistical significance was determined comparing the estimated CCR to a null distribution obtained from the 10,000 repetitions where population effect size was 0.

Figures 5a-d show the null distributions for 2-fold, 5-fold, 10-fold, and LOO cross-validation, respectively. All cross-validations were stratified except for LOO where only one trial was removed. Generally, increasing k results in fewer, but more extreme below chance classifications. (LOO results in only 30.5% below chance classifications, 2-fold results in 62% BCP). In contrast, CCRs for LOO can reach 0%, whereas they remain above 40% for 2-fold. Because the expected values of these distributions, i.e. the mean CCRs over a large number of experiments, are fixed at 50%, most experiments using LOO cross-validation must result in CCRs above 50%. From this consideration, it becomes clear that for $d = 0$ LOO will result in more spuriously high CCRs than 2-fold cross-validation.

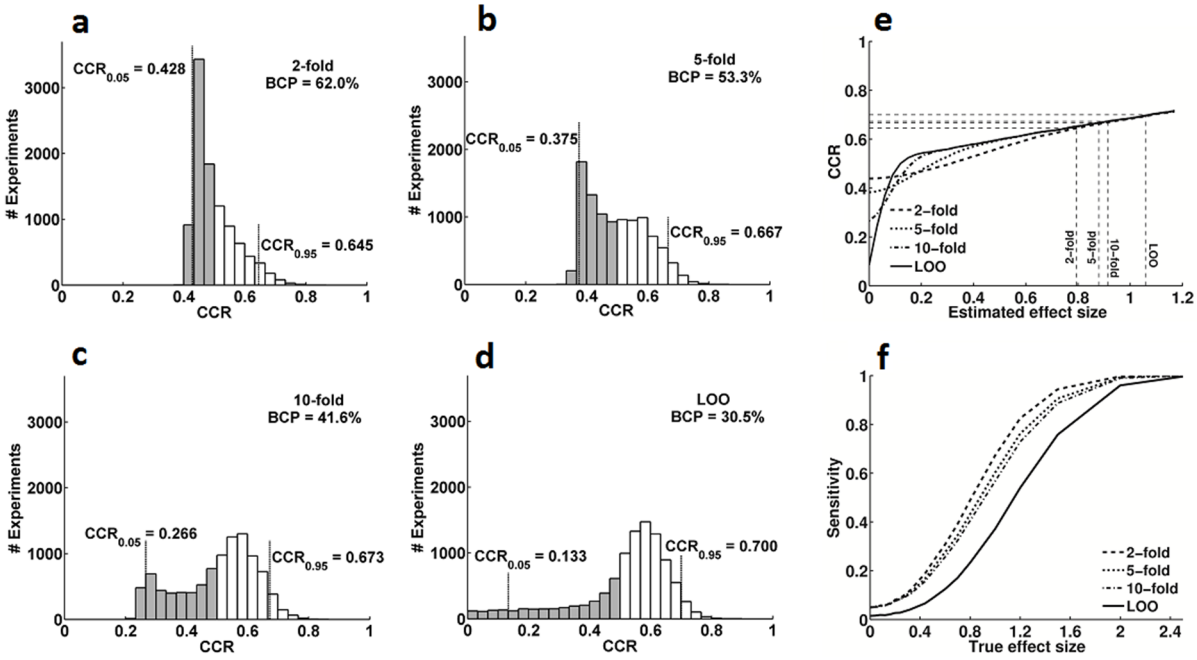


Figure 5: Effect of k on CCRs of LDA with cross-validation. (A-D) Null distributions for different values of k. Data come from 10,000 simulated experiments with a priori true effect size of $d = 0$. They are classified with LDA using 2-fold, 5-fold, 10-fold, and LOO cross-validation, respectively. The grey areas show the experiments with classification rates below 50%. Because these distributions represent true effect sizes of $d = 0$, mean CCRs are exactly 50%. (E) CCRs as a function of estimated effect size for various values of k. Estimated effect sizes and CCRs where significance is reached are indicated by dashed lines. (F) Sensitivity shows the ability of different cross-validation procedures to correctly detect effects when true effect size is nonzero. For a given effect size d, the significance threshold of $p \leq 0.05$ will be reached for smaller estimated effect sizes for 2-fold than for LOO cross-validation and therefore, 2-fold shows higher sensitivity. For LOO, sensitivity can fall below 0.05 because its null-distribution can only assume discrete values.

But, are the higher CCRs in LOO than in 2-fold cross-validation a sign of a more sensitive test procedure in cases of $d > 0$? Classification algorithms are typically optimized to reach high CCRs. This makes sense if classes are known to be distinct ($d \gg 0$), and the focus is on correctly identifying novel items. For hypothesis testing purposes, however, the existence and size of a possible effect are unknown. Algorithms should therefore be optimized for highest sensitivity regarding the distinction between $d = 0$ and $d > 0$. Let us therefore assert that

a reliable distinction is possible if a given CCR is higher than the $CCR_{0.95}$ of its null distribution. Figure 5e shows CCRs as a function of the estimated effect size $\hat{d}$ for different values of k (see also Appendix D for comparison in terms of AUC). For very large estimated effect sizes, the influence of k can be neglected. For $0.1 < \hat{d} < 1$, in this example, LOO shows higher CCRs than 2-fold cross-validation. However, when testing for significance against the null distribution (Figs. 5a-d), the 2-fold procedure, although it needs on average larger effect sizes to attain the same CCR, reaches the threshold for significance ($CCR_{0.95}$) earlier than LOO (2-fold: $\hat{d}_{thres} = 0.8$, LOO: $\hat{d}_{thres} = 1.1$). Together, LOO not only produces more spuriously high CCRs in the no effect case, it is also less sensitive than 2-fold cross-validation when $\hat{d} > 0$. We can conclude that 2-fold cross-validation should be preferred over LOO for hypothesis testing. CCRs should only be interpreted in relation to a suitable null distribution.

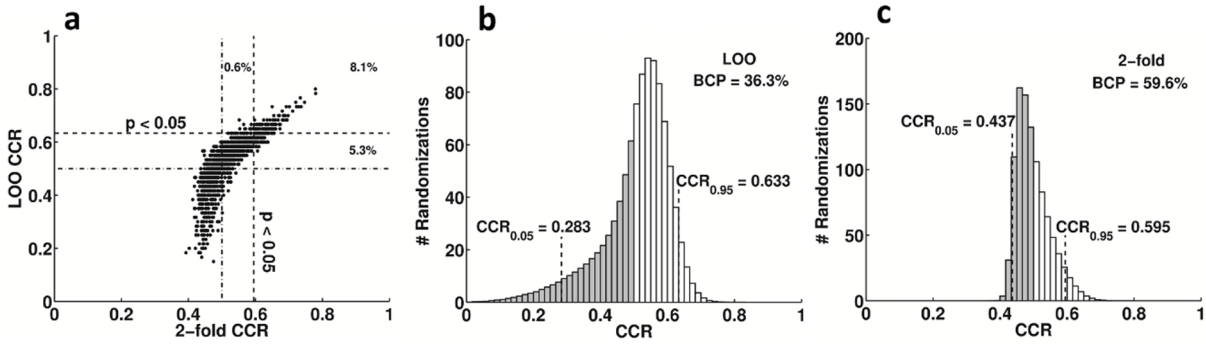


Figure 6: Searchlight classification of EEG data. (A) CCRs from LOO versus CCRs from a 2-fold procedure. Each dot represents the classification result of a single searchlight. Dashed lines indicate the average 5% significance threshold based on randomization tests. LOO results generally in higher CCRs, but 2-fold reaches significance for lower CCR values thus having a higher sensitivity. Percentages represent the proportion of tests that are below p<0.05 for 2-fold and LOO, respectively. (B, C) Average null distribution of EEG searchlight classification obtained by random shuffling of class labels when classified using LOO and 2-fold. LOO results in more spuriously high CCRs, higher significance thresholds, and below chance classification rates close to 0%. 2-fold produces more CCRs slightly below 50% under the null distribution and has a lower significance threshold and thus higher sensitivity.

Generalization to high-dimensional data

Finally, we studied how dimensionality of data affects classification performance. For this purpose, we analyzed the real EEG data set described above and compared it with synthetically generated data. We designed two series of 100 dimensional Monte Carlo experiments containing 30 observations per condition, which is comparable to the EEG data set. Simulations had a true effect size of zero to simulate the null distribution. In the first series, we assumed independent features; in the second series, features were correlated, as it is typically the case in neuroscience, e.g. in electrophysiological and imaging data. To obtain multivariate normal, correlated data, we generated 10,000 normally distributed, uncorrelated data sets, which were then multiplied by the Cholesky decomposition of a randomly generated covariance matrix. To derive the null distribution of our real EEG data, we used randomly shuffled class labels (Nichols and Holmes, 2002). The three data sets were then classified with 3 popular linear classifiers, once using LOO and once using 2-fold cross-validation.

Figure 6a shows the CCR distribution of real EEG data classified using LOO versus 2-fold (data for diagonal LDA, results for other classifiers are summarized in Table 1). CCRs for LOO are generally higher than for 2-fold, but in the below chance range LOO is also more extreme (CCR_{min} = 15%) than 2-fold (CCR_{min} = 39%). In spite of the lower CCRs, 2-fold cross-validation reaches significance sooner (59% in 2-fold vs. 63% in LOO) and shows higher sensitivity (13% of classifications significant with 2-fold, 9% of classifications significant with LOO). Looking at the null distribution obtained by shuffling EEG data labels shows that LOO and 2-fold procedures result in different distributions. LOO shows a strong skewness to the left giving CCRs as low as 0%, and a large number of spuriously high CCRs, with only 36.3% of classification rates below chance. 2-fold on the other hand shows a strong skewness to the right with a BCP of 59.6%. Together, these findings emphasize that 2-fold cross-validation results less often in CCRs above 50%, but at the same time shows higher sensitivity than LOO cross-validation (Figs. 6b-c).

Comparing different linear classifiers (nearest centroid classifier [NCC], diagonal LDA, and linear SVM) shows that all of the linear classifiers tested here show asymmetric distributions for synthetic correlated data and real EEG data, i.e., higher than expected BCP for 2-fold, lower than expected BCP for LOO with asymmetric range (see Table 1). Both confirm the findings described in the previous sections.

Table 1: BCP and range of null distribution for different classifiers based on real and synthetic EEG data when classified using LOO and 2-fold.

	Real EEG data				Correlated data (synthetic)				Uncorrelated data (synthetic)			
	2-fold		LOO		2-fold		LOO		2-fold		LOO	
	BCP	Range	BCP	Range	BCP	Range	BCP	Range	BCP	Range	BCP	Range
NCC	59.1	43.4–59.6	36.4	26.7–63.3	62.0	44.7–60.2	28.5	16.7–65.0	50.4	41.3–58.9	49.5	33.3–65.0
DLDA	59.6	43.4–59.6	36.3	26.7–63.3	62.0	44.7–60.2	28.6	16.7–65.0	50.4	41.5–58.9	49.8	33.3–65.0
SVM	61.5	41.9–57.2	48.6	30.0–66.7	56.0	43.8–58.5	42.8	33.3–63.3	51.0	42.3–58.3	51.7	35.0–65.0

NCC: nearest centroid classifier, DLDA: diagonal LDA, SVM: support vector machine, BCP: below-chance percentage, Range: [$CCR_{0.05}$ – $CCR_{0.95}$].

Unlike correlated data, the null distributions of uncorrelated data show an almost symmetric structure, i.e. BCP of approximately 50% with nearly symmetric range. The effect of correlation can be understood by looking at multivariate effect sizes. In multidimensional data, the Mahalanobis distance $D = \sqrt{d^T R^{-1} d}$ between two classes can be taken as the multivariate effect size. d represents the vector of univariate effect sizes in each dimension, and R is the correlation matrix of the data. When features are independent ($R = I$), the Mahalanobis distance reduces to $\widehat{D} = \sqrt{\hat{d}_1^2 + \hat{d}_2^2 + \cdots + \hat{d}_p^2}$, with the empirical effects $\hat{d}$ consisting of any true effect d plus the measurement error e. Thus, the

distance $\widehat{D}$ increases with increasing dimensionality even when no true effect is present at all. Therefore, inversions between training and test classification like we described in Figure 3a become unlikely and less asymmetry in below chance classification rates occurs. A larger distance $\widehat{D}$ between conditions also allows the classifier to separate classes more easily, but, because no actual effect exists between conditions, this leads to overfitting the training data and no generalization to test data. Thus, CCRs for uncorrelated high-dimensional data follow a symmetric distribution with BCP of 50%.

Discussion

When we draw random samples from two normal distributions that do not differ in their means, the estimated differences of these means will be normally distributed around zero. When using MVPA on these data, one could assume that the distribution of classification rates is also distributed symmetrically around the chance level of 50%, following an approximately normal distribution. However, we show that our intuitive evaluation of results, based on assumptions of symmetry and normality, is misleading for classification outcomes. For classification with cross-validation in typical life-science data, i.e., small sample size data holding small effects, the distribution of classification rates is neither normal nor binomial (Noirhomme, et al., 2014). It can be strongly asymmetric and in some cases even bimodal. Its mode can lie either below or above the level expected by chance. Particularly for small effect sizes, we notice that the distribution of results can be strongly skewed and CCRs can sometimes fall far below chance level. Moreover, variance of CCR changes with effect size, disproportionally increasing for decreasing effect sizes and thus distorting CCR distribution further. This irregular distribution of CCRs implies that interpretation of CCRs as an indicator of presence or size of an effect is not advisable.

The aim of using MVPA for hypothesis testing is to detect the presence of differences between two or more conditions. CCR itself cannot be used for this purpose without a suitable critical value to compare it to. The level of this critical value can differ greatly depending on classifier, cross validation, number of features, sample size, and effect size in the data. Therefore, CCRs cannot be compared between different analyses in most cases. Instead, p-values must be provided, which determine whether a certain CCR indicates a significant difference between classes. Significance testing always relies on an accurate estimation of the distribution of data under the null hypothesis. Importantly, the irregular shape of the CCR distribution forbids the use of parametric tests, which rely on normal or binomial null distributions. Instead, the null distribution should be established from surrogate data using Monte Carlo or randomization methods (Manly, 2007). Furthermore, an maximization of CCR, which is perfectly reasonable when (infinite) novel stimuli must be identified, should never be done when cross-validation is used in MVPA for hypothesis testing. This can be easily understood when looking at the comparison of 2-fold and LOO cross validation. Although the LOO procedure results in more above chance classifications, it is less sensitive, i.e., it requires a higher effect size to reach significance.

CCRs are unbiased, i.e. on average they are a good measure of classifiability. However, a problem occurs because of the skewness of the distribution, which shifts the mode of the distribution above or below its expected value. One particular danger when interpreting CCRs is that the presence of a few very strong below chance level CCRs must necessarily lead to a larger number of moderately above chance CCRs, even in data that do not contain any effect. This leads to several complications. First, although two methods (e.g. 2-fold and LOO) may have the same expected value for a certain true effect size, one will result in most cases in a higher CCR than the other (see Fig. 5). Any interpretation of this finding based on a small number of experiments will be misleading. Even if a large number of independent experiments is performed, the number of

positive results will be similarly misleading. Only averaging CCRs over this larger number of experiments would generate a result close to the correct expected value. What is even worse, if those rare experiments or subjects resulting in very low CCRs are discarded as outliers or experimental failures and findings are not published, the average CCR of published studies will be clearly above 50%. Very few unpublished studies can then disproportionally distort conclusions of meta analyses. Similarly, if a number of identical analyses are carried out on a number of voxels, electrodes, genes etc., the number of above 50% findings can be much higher than the number of below 50% findings although the null hypothesis is valid. If results (class differences, classifiability) are presented in terms of CCR, this can lead to the erroneous assumption that most dimensions (voxels, electrodes or genes) carry class information. Any significance test that does not take the skewed distribution of CCRs into account has a high risk to result in a false positive finding.

Instead of using CCR to display classification results, we propose the use of p-values for this purpose. While CCR behaves in a rather unintuitive fashion because of the skewness of its distribution when effect sizes are small, the distribution of p-values is simple and most readers are familiar with their interpretation. This proposal follows the same logic that is used with fMRI analyses, which also usually present statistical maps rather than actual measures of hemodynamic responses. P-values are a combined measure of central tendency, variability and sample size, and represent the strength of evidence against the null hypothesis. Given identical sample size, they provide a standardized way to compare results of different experiments, conditions and analyses. In addition, when using first and second level models, statistical values can be used to aggregate data over groups of subjects or compare data coming from different experimental conditions. If a suitable measure of multivariate effect size is available (see e.g. Allefeld and Haynes, 2014), then this should be presented and significances indicated. When no suitable measure of multivariate effect size is available, p-values seem clearly preferable over CCRs to represent classification results.

The occurrence of classification rates far below the chance level in classification with cross-validation has been observed in a number of studies before. It has been reported that LOO cross-validation will in specific cases (e.g. linear SVM with a dimensionality approaching infinity) result in zero percent classification accuracy in finite data sets (Hall, et al., 2005; Verleysen, 2003). The same observation was reported for majority inducers irrespective of dimensionality (Kohavi, 1995). We present analysis and simulations that help understand the causes underlying below chance classification. Occurrence of below chance classification is a direct result of the dependence of test and training means and thus in essence denotes that an effect is too small to be detected by the classifier. This strength of the dependence between training and test means depends on sample size and is also governed by *k* in k-fold cross validation.

The number of folds *k* in cross-validation affects the variance of classification accuracy and its significance, in particular, when sample size and estimated effect size are low. CCRs obtained by LOO are less likely to be below the chance level, but if they are, they are usually lower than for a 2-fold procedure. For medium estimated effect sizes LOO gives higher CCRs than 2-fold cross-validation on average. This is desirable in a context of single item classification, when the presence of a class difference is known and the focus is on accurately classifying individual items. Previous studies therefore concluded that LOO should be preferred over 2-fold cross-validation because of the latter's conservative bias (Kohavi, 1995; Rodriguez, et al., 2010). However, in a hypothesis testing context, when the presence of an effect is uncertain, 2-fold cross-validation has a smaller variance, especially in the null-distribution, and is therefore more reliable and reaches significance already with much smaller effect sizes than LOO. Using a 2-fold procedure is therefore preferable for hypothesis testing, because of its higher sensitivity, especially in LSS-LES data.

Dimensionality of data is another important factor which affects the behavior of classification algorithms. Increasing the number of features impairs performance of classification algorithms, a fact also known as the "curse of

dimensionality" (Bickel and Levina, 2004; Clarke, et al., 2008; Fan and Fan, 2008). Hall et al. showed that when the size of the feature vector increases with a fixed number of samples, linear SVM will asymptotically approach chance performance (Hall, et al., 2005). Jin et al. demonstrated that in data sets with few and weak relevant features classification can be impossible if feature selection is not done prior to classification (Donoho and Jin, 2008; Jin, 2009). We showed in simulations and real EEG data that asymmetric distributions of CCRs with strong below chance classification rates and many spuriously high classification rates occur in high-dimensional data as well as in low-dimensional data, especially if features are correlated.

In this paper, we have investigated the behavior of cross-validation and MVPA in realistic life-science data. This kind of data is often characterized by small effect sizes, small sample sizes, but a large number of features. We show that there are a few important guidelines that should be observed. Most importantly, the existence of an effect should not be determined by the classification rate, but rather by statistical significance, and significance should not be based on parametric tests, but on Monte Carlo methods. Furthermore, because hypothesis testing has different requirements than individual item identification, methods optimized for the latter purpose are not necessarily the best for the former. Therefore, although LOO cross-validation results in higher classification rates, 2-fold cross-validation is more suitable for hypothesis testing because its smaller variance makes it more sensitive. If these guidelines are observed, we believe that MVPA is an excellent method that allows dealing with the problems of multivariate data in the life-sciences.

Acknowledgements

This research was supported by DFG grant GA730/3-1 and BMBF Bernstein Center grant 01 GQ 1004A - B4.

Appendix A: Theorem 1

We assume a data set that consists of N observations for each of two classes A and B with empirical means $\vec{m} = [m_A, m_B]^T$. The means $m_{A,B}$ themselves are determined by the probability distributions $p_A(x_A)$ and $p_B(x_B)$ for the individual observations $x_{A,B}$. We fix m_A and m_B, thereby assuming the data set to be one specific realization of the random processes governed by p_A and p_B. During k-fold cross-validation the data set is divided into a training and a test set, with means $\vec{\mu} = [\mu_A, \mu_B]$ for the training set and $\vec{\nu} = [\nu_A, \nu_B]$ for the test set.

Under the definitions and assumptions above, the probability $\pi(\vec{m})$ of a correct classification, conditional to the sample means $\vec{m} = [m_A, m_B]^T$, is given by

$$\pi(\vec{m}) = \frac{1}{2} - \frac{1}{2}\left(\prod_{i=A,B} \int d\mu_i \Pr(\mu_i|m_i)\right)(\phi_A - \phi_B) sign(\mu_A - \mu_B)$$

in which ϕ_A is defined as $\phi_A =: \Pr\left(x_A \leq \frac{\mu_A + \mu_B}{2} \middle| \mu_A < \mu_B, \vec{\mu}, \vec{m}\right)$ and ϕ_B is defined analogously.

Proof

For given sample means $m_{A,B}$, subsample means $\mu_{A,B}$ and $\nu_{A,B}$ are statistically dependent stochastic variables, since

$$m_{A,B} = \frac{\mu_{A,B}(k-1) + \nu_{A,B}}{k} \tag{1}$$

LDA in 1-d is expressed by the two threshold conditions for correct classification of test observations x_A and x_B from the two classes

$$(\mu_A - \mu_B)(x_A - t) \geq 0, \ (\mu_A - \mu_B)(x_B - t) < 0 \tag{2}$$

Here, the discrimination threshold is the training mean:

$$t = \frac{\mu_A + \mu_B}{2} \tag{3}$$

For given sample means $\vec{m} = (m_A, m_B)^T$ the probability π_A of a correct classification of a test observation from class A (x_A) is thus obtained as

$$\pi_A(\vec{m}) = \int d\vec{\mu}\,[\Pr(x_A \geq t, \mu_A \geq \mu_B, \vec{\mu}, \vec{m}) + \Pr(x_A \leq t, \mu_A \leq \mu_B, \vec{\mu}, \vec{m})] \quad (4)$$

which can be further transformed as follows:

$$\pi_A(\vec{m}) = \int d\vec{\mu}[\Pr(x_A \geq t|\mu_A \geq \mu_B, \vec{\mu}, \vec{m})\Pr(\mu_A \geq \mu_B, \vec{\mu}|\vec{m}) + \Pr(x_A \leq t|\mu_A < \mu_B, \vec{\mu}, \vec{m})\Pr(\mu_A < \mu_B, \vec{\mu}|\vec{m})] \quad (5)$$

The two factors implementing the threshold connection are probabilities of complementary events and thus we can write:

$$\Pr(x_A \geq t|\mu_A \geq \mu_B, \vec{\mu}, \vec{m}) =: 1 - \phi_A \quad (6)$$

$$\Pr(x_A \leq t|\mu_A < \mu_B, \vec{\mu}, \vec{m}) =: \phi_A \quad (7)$$

Using equations (1) and (2), the probability ϕ_A can be expressed as the cumulative distribution:

$$\phi_A = \int_{-\infty}^{\frac{\mu_A+\mu_B}{2}} dx_A \Pr(x_A|\upsilon_A) = \int_{-\infty}^{\frac{\mu_A+\mu_B}{2}} dx_A \frac{\Pr(\upsilon_A|x_A)p_A(x_A)}{\Pr(\upsilon_A)} = \int_{-\infty}^{\frac{\mu_A+\mu_B}{2}} dx_A \frac{\Pr(\sum_{\tau=1}^{T-1} x_A^\tau = T\upsilon_A - x_A)p_A(x_A)}{\Pr(\upsilon_A)} \quad (8)$$

in which $T = N/k$ is the number of test observations.

The two remaining factors in equation (5) can again be evaluated using conditional probabilities:

$$\Pr(\mu_A \geq \mu_B, \vec{\mu}|\vec{m}) = \Pr(\mu_A \geq \mu_B|\vec{\mu}, \vec{m})\ \Pr(\vec{\mu}|\vec{m}) = \Theta(\mu_A \geq \mu_B)\Pr(\mu_A|m_A)\Pr(\mu_B|m_B) \quad (9)$$

Here Θ denotes the Heavyside step function (indicator), and factorization of $\Pr(\vec{\mu}|\vec{m})$ accounts for the independence of the classes.

In analogy, one obtains

$$\Pr(\mu_A < \mu_B, \vec{\mu}|\vec{m}) = \Theta(\mu_A < \mu_B)\Pr(\mu_A|m_A)\Pr(\mu_B|m_B) \quad (10)$$

The probability π_A of correct classification of a class A observation equals

$$\pi_A(\vec{m}) = \left(\prod_{i=A,B} \int d\mu_i \Pr(\mu_i|m_i)\right)[(1-\phi_A)\Theta(\mu_A \geq \mu_B) + \phi_A\Theta(\mu_A < \mu_B)] \quad (11)$$

Correspondingly, for class B:

$$\pi_B(\vec{m}) = \left(\prod_{i=A,B} \int d\mu_i \Pr(\mu_i|m_i)\right)[(1-\phi_B)\Theta(\mu_A < \mu_B) + \phi_B\Theta(\mu_A \geq \mu_B)] \quad (12)$$

Thus the probability of correct classification is

$$\pi(\vec{m}) = \frac{\pi_A(\vec{m})+\pi_B(\vec{m})}{2} = \frac{1}{2} - \frac{1}{2}\left(\prod_{i=A,B} \int d\mu_i \Pr(\mu_i|m_i)\right)(\phi_A - \phi_B)sign(\mu_A - \mu_B) \quad (13)$$

Which completes the proof.

Appendix B: Simplification for normal distributions

Example for the case with no effect

As an extreme example, we use Gaussian distributed data with no signal, i.e.

$$p_A(x) = p_B(x) = \mathcal{N}(x) = \frac{1}{\sqrt{2\pi}} exp\left(-\frac{x^2}{2}\right) \tag{14}$$

Applying equation (8) for the Gaussian distributions $p_{A,B}$, the cumulative distribution is that of a Gaussian with variance $(T-1)/T$ and mean

$$v_A = km_A - (k-1)\mu_A \tag{15}$$

This yields

$$\phi_A - \phi_B = \Phi\left(k(\mu_A - m_A) + \frac{\mu_B - \mu_A}{2}\right) - \Phi\left(k(\mu_B - m_B) + \frac{\mu_A - \mu_B}{2}\right) \tag{16}$$

with $\Phi(z) = \frac{1}{2}[1 + \mathrm{erf}(z/\sqrt{2(T-1)/T})]$.

The conditional probabilities $\Pr(\mu_i|m_i)$ are inferred from Bayes' law as

$$\Pr(\mu_i|m_i) \propto \Pr(m_i|\mu_i)\Pr(\mu_i) = \Pr[v_i = km_i - (k-1)\mu_i]\Pr(\mu_i) \tag{17}$$

The means being sums of Gaussian variables, their distributions $\Pr(\mu_i)$ and $\Pr(v_i)$ are also Gaussians with variance $k/[(k-1)N]$ andk/N, respectively, and thus

$$\Pr(\mu_i|m_i) = \sqrt{\frac{N(k-1)}{2\pi}} \exp\left[\frac{N(k-1)}{2}(\mu_i - m_i)^2\right] \tag{18}$$

Finally, we calculated the distributions of π for $\vec{m}$ sampled from a Gaussian distribution

$$\Pr(m_i) = \sqrt{\frac{N}{2\pi}} \exp[-m_i^2(N/2)] \tag{19}$$

Example for the case with an effect of $d > 0$

The previous Gaussian example can be easily generalized to finite signal strength d by letting

$$p_A(x) = \mathcal{N}\left(x + \frac{d}{2}\right),\ p_B(x) = \mathcal{N}\left(x - \frac{d}{2}\right) \tag{20}$$

In this case both the difference $\phi_A - \phi_B$ from equation (16) and $\Pr(\mu_i|m_i)$ from equation (18) are unchanged, since they only depend on differences of means. Thus equation (13) for $\pi(\vec{m})$ still holds in this case. The only difference between the case with and without effect therefore derives from the sampling of $\vec{m}$. As a result of $d \neq 0$ the part of the function $\pi(\vec{m})$ that is sampled is further away from the valley of below-chance classification rates and therefore the distribution of π is less skewed than in the no effect case.

To validate our analytical results, we generated a series of simulations with different effect sizes and calculated CCRs once using k-fold cross-validation and once with equation (13). As Figure B1 shows, the results are fairly similar.

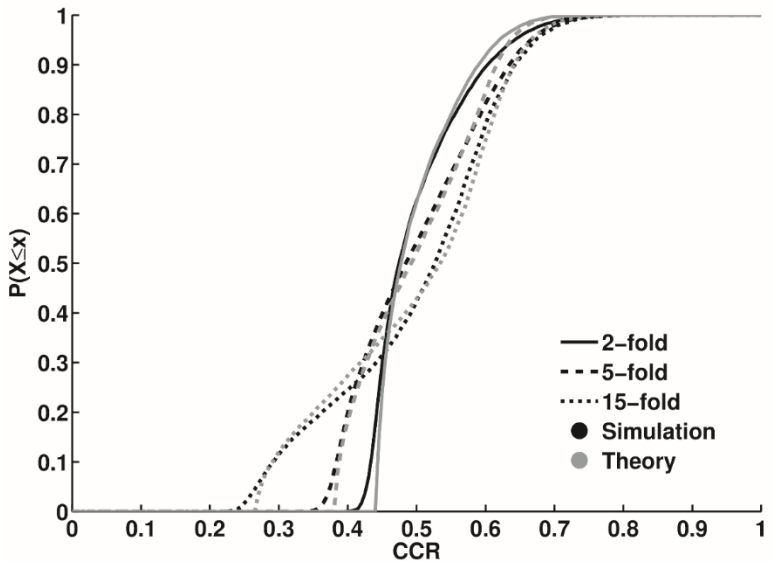


Figure B1: Cumulative distribution function of CCRs for LDA ($N = 15$) using different values of k calculated from equation 13 (grey lines) and determined using simulation results (black lines). The results show that both equation 13 and simulations produce very similar results.

Appendix C: Corollary 1

Corollary 1: The probability of correct classification in LDA with cross-validation for no effect data sets ($m_A = m_B$) must always be below the level expected for chance classification. This result is independent of the data distribution and the number of cross-validation folds k.

Proof

From Theorem 1 we know:

$$\pi(\vec{m}) = \frac{\pi_A(\vec{m})+\pi_B(\vec{m})}{2} = \frac{1}{2} - \frac{1}{2}\left(\prod_{i=A,B} \int d\mu_i \Pr(\mu_i|m_i)\right)(\phi_A - \phi_B)sign(\mu_A - \mu_B) \quad (21)$$

The integral on the right side represents the deviation from chance level, i.e., 0.5. Here we show that for data with an estimated effect size of zero, the integrand will be always positive thus leading to below chance values of $\pi(\vec{m})$. To prove this it suffices to show that $(\phi_A - \phi_B)sign(\mu_A - \mu_B) > 0$ because $\Pr(\mu_i|m_i) > 0$ by definition.

According to equation (8), for $m = m_A = m_B$ we can write

$\phi_A - \phi_B = \int dx\,[Pr(x|km - (k-1)\mu_A) - Pr(x|km - (k-1)\mu_B)]$.

If $\mu_A > \mu_B$the left summand under the integral is a left-shifted version of the right summand. Since being probabilities, both summands are positive and normalized, the difference $\phi_A - \phi_B$ is positive. Conversely, for $\mu_A < \mu_B$ the difference $\phi_A - \phi_B$ is negative. Thus, in both cases $(\phi_A - \phi_B)sign(\mu_A - \mu_B) > 0$ and the corollary is proved.

Appendix D: Area Under the Curve (AUC)

AUC is similarly affected by the dependence of the sub-sample means. Using classification thresholds as it is done in AUC does not prohibit negative correlations between test and training means. Appendix Figure 2 replicates Figure 5e for AUC.

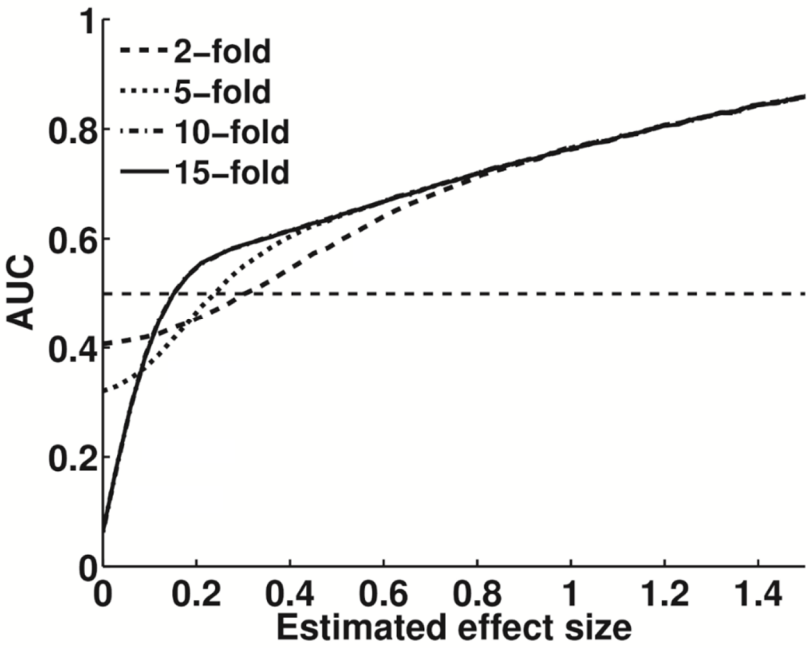


Figure D1: The dependence of the sub-sample means affects performance of an LDA classifier largely independently of performance measure. Using a signal detection approach and replacing CCR with the area under the curve (AUC) from receiver operating characteristics (ROC) curves results in AUCs below 0.5.

References

Allefeld, C., Haynes, J.D. (2014) Searchlight-based multi-voxel pattern analysis of fMRI by cross-validated MANOVA. NeuroImage, 89:345-357.

Azuaje, F. (2003) Genomic data sampling and its effect on classification performance assessment. BMC bioinformatics, 4:5.

Bickel, P.J., Levina, E. (2004) Some theory for Fisher's linear discriminant function, 'naive Bayes', and some alternatives when there are many more variables than observations. Bernoulli, 10:989-1010.

Braga-Neto, U.M., Dougherty, E.R. (2004) Is cross-validation valid for small-sample microarray classification? Bioinformatics (Oxford, England), 20:374-80.

Button, K.S., Ioannidis, J.P., Mokrysz, C., Nosek, B.A., Flint, J., Robinson, E.S., Munafo, M.R. (2013) Power failure: why small sample size undermines the reliability of neuroscience. Nature reviews. Neuroscience, 14:365-76.

Clarke, R., Ressom, H.W., Wang, A., Xuan, J., Liu, M.C., Gehan, E.A., Wang, Y. (2008) The properties of high-dimensional data spaces: implications for exploring gene and protein expression data. Nature reviews. Cancer, 8:37-49.

Cohen, J. (1988) Statistical Power Analysis for the Behavioral-Sciences. New Jersey. Lawrence Erlbaum Associates.

Damarla, S.R., Just, M.A. (2013) Decoding the representation of numerical values from brain activation patterns. Human brain mapping, 34:2624-34.

Delorme, A., Makeig, S. (2004) EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. Journal of neuroscience methods, 134:9-21.

Deuker, L., Olligs, J., Fell, J., Kranz, T.A., Mormann, F., Montag, C., Reuter, M., Elger, C.E., Axmacher, N. (2013) Memory consolidation by replay of stimulus-specific neural activity. The Journal of neuroscience : the official journal of the Society for Neuroscience, 33:19373-83.

Donoho, D., Jin, J.S. (2008) Higher criticism thresholding: Optimal feature selection when useful features are rare and weak. Proceedings of the National Academy of Sciences of the United States of America, 105:14790-14795.

Dougherty, E.R. (2001) Small sample issues for microarray-based classification. Comparative and functional genomics, 2:28-34.

Duarte, J.V., Ribeiro, M.J., Violante, I.R., Cunha, G., Silva, E., Castelo-Branco, M. (2014) Multivariate pattern analysis reveals subtle brain anomalies relevant to the cognitive phenotype in neurofibromatosis type 1. Human brain mapping, 35:89-106.

Etzel, J.A., Zacks, J.M., Braver, T.S. (2013) Searchlight analysis: promise, pitfalls, and potential. NeuroImage, 78:261-9.

Fan, J.Q., Fan, Y.Y. (2008) High Dimensional Classification Using Features Annealed Independence Rules. Ann. Stat., 36:2605-2637.

Fuentemilla, L., Penny, W.D., Cashdollar, N., Bunzeck, N., Duzel, E. (2010) Theta-coupled periodic replay in working memory. Current biology : CB, 20:606-12.

Gisselbrecht, S.S., Barrera, L.A., Porsch, M., Aboukhalil, A., Estep, P.W., 3rd, Vedenko, A., Palagi, A., Kim, Y., Zhu, X., Busser, B.W., Gamble, C.E., Iagovitina, A., Singhania, A., Michelson, A.M., Bulyk, M.L. (2013) Highly parallel assays of tissue-specific enhancers in whole Drosophila embryos. Nature methods, 10:774-80.

Hall, P., Marron, J.S., Neeman, A. (2005) Geometric representation of high dimension, low sample size data. J. R. Stat. Soc. Ser. B-Stat. Methodol., 67:427-444.

Hastie, T.T., R.; Friedman, J. (2001) The Elements of Statistical Learning. New York. Springer.

Haynes, J.D., Rees, G. (2006) Decoding mental states from brain activity in humans. Nature reviews. Neuroscience, 7:523-34.

Jin, J.S. (2009) Impossibility of successful classification when useful features are rare and weak. Proc. Natl. Acad. Sci. U. S. A., 106:8859-8864.

Kamitani, Y., Tong, F. (2005) Decoding the visual and subjective contents of the human brain. Nature neuroscience, 8:679-85.

Kohavi, R. (1995) A study of cross-validation and bootstrap for accuracy estimation and model selection. Proceedings of the 14th international joint conference on Artificial intelligence - Volume 2. Montreal, Quebec, Canada: Morgan Kaufmann Publishers Inc. p 1137-1143.

Kriegeskorte, N., Goebel, R., Bandettini, P. (2006) Information-based functional brain mapping. Proceedings of the National Academy of Sciences of the United States of America, 103:3863-8.

Kriegeskorte, N., Simmons, W.K., Bellgowan, P.S., Baker, C.I. (2009) Circular analysis in systems neuroscience: the dangers of double dipping. Nature neuroscience, 12:535-40.

Lemm, S., Blankertz, B., Dickhaus, T., Muller, K.R. (2011) Introduction to machine learning for brain imaging. NeuroImage, 56:387-99.

Manly, B.F.J. (2007) Randomization, bootstrap, and Monte Carlo methods in biology. Boca Raton, FL. Chapman & Hall/ CRC.

Nichols, T.E., Holmes, A.P. (2002) Nonparametric permutation tests for functional neuroimaging: a primer with examples. Human brain mapping, 15:1-25.

Noirhomme, Q., Lesenfants, D., Gomez, F., Soddu, A., Schrouff, J., Garraux, G., Luxen, A., Phillips, C., Laureys, S. (2014) Biased binomial assessment of cross-validated estimation of classification accuracies illustrated in diagnosis predictions. NeuroImage. Clinical, 4:687-94.

Norman, K.A., Polyn, S.M., Detre, G.J., Haxby, J.V. (2006) Beyond mind-reading: multi-voxel pattern analysis of fMRI data. Trends in cognitive sciences, 10:424-30.

Raudys, S., Pikelis, V. (1980) On dimensionality, sample size, classification error, and complexity of classification algorithm in pattern recognition. IEEE transactions on pattern analysis and machine intelligence, 2:242-52.

Raudys, S.J., Jain, A.K. (1991) Small sample size effects in statistical pattern recognition: recommendations for practitioners. IEEE transactions on pattern analysis and machine intelligence, 13:252-264.

Rodriguez, J.D., Perez, A., Lozano, J.A. (2010) Sensitivity analysis of k-fold cross validation in prediction error estimation. IEEE transactions on pattern analysis and machine intelligence, 32:569-75.

Staresina, B.P., Alink, A., Kriegeskorte, N., Henson, R.N. (2013) Awake reactivation predicts memory in humans. Proceedings of the National Academy of Sciences of the United States of America, 110:21159-64.

Verleysen, M. (2003) Learning high-dimensional data. Limitations and Future Trends in Neural Computation 22.

Chapter 3:

Adjusting permutation tests in multivariate analysis of neuroimaging data with subclasses: introduction of biased null distributions

Hamidreza Jamalabadi, Sarah Alizadeh, Monika Schönauer, Christian Leibold, and Steffen Gais

Manuscript in submission process

Abstract

Multivariate pattern analysis (MVPA) is considered a powerful method for detecting systematic effects in large datasets. When used for hypothesis testing, a classifier is trained on one part of the data and performance is tested on separate data. Significance of such cross-validated classification is determined using permutation testing, estimating the distribution of classification results when systematic information about class assignment is removed by randomly reshuffling class labels. Traditionally, this relabeling occurs on a trial-by-trial basis. We show here that in data which, next to a main effect of class (e.g. visual presentation of letter vs. number), additionally contains a nested subclass structure (individual digits and letters), trial-level-randomization gives too liberal estimates of significance because subclasses introduce systematic information that generally improves separability of the classification problem. We analytically prove that this subclass bias systematically affects correct classification rates (CCRs), even in the absence of a main effect. In simulations, we demonstrate that subclass bias is highest for low between-class effect size and high subclass variance, but can be reduced by increasing the total number of subclasses. Moreover, we can account for the subclass bias by adjusting permutation tests to consider the subclass structure of the data, using subclass-level randomization. In several experiments recording human brain electrical activity, we demonstrate that parametric testing fails critically to determine significance of classification outcomes, and that trial-wise permutation gives too liberal estimates. To avoid false positive results due to subclass biases, we give practical examples of how to modify permutation testing for a range of common experimental designs with subclasses.

Introduction

Multivariate pattern analysis (MVPA) combined with cross-validation and permutation testing allows the use of machine learning algorithms to detect differences between classes of data for statistical hypothesis testing (Haxby, et al., 2014; Jamalabadi, et al., 2016; Stelzer, et al., 2013). Whereas classical statistical approaches search for individual features in a data set that allow to distinguish two experimental conditions, MVPA analyzes data sets as a whole, searching for distinguishing multi-dimensional patterns. Therefore, it can provide increased sensitivity compared to classical multiple-univariate testing methods in high-dimensional data sets (Haynes, 2015; Norman, et al., 2006; Woolgar, et al., 2014; Alizadeh, et al., 2017).

When MVPA is used for hypothesis testing an algorithm (a classifier, e.g. a support vector machine [SVM]) is trained on a portion of a data set to separate data belonging to different classes (e.g. different experimental conditions, different groups of patients, etc.). Then, the ability to classify new data is tested on the remaining part of the data. This results in a percentage of accurate classifications (correct classification rate [CCR]). To improve estimation accuracy of this percentage, CCRs are usually determined using a cross-validation procedure, which assures that all parts of the data are used for training as well as testing on repeated iterations of the analysis (Efron, 1993). If accuracy of classification lies significantly above the level expected by chance (e.g. 50% for a two-class problem), it can be concluded that there is class-related information in the data, and classes can be concluded to differ significantly. To determine the significance threshold non-parametric permutation statistics should be used, because CCRs often do not follow any known distribution (Jamalabadi, et al., 2016). These tests determine the null distribution by resampling the data a large number of times with randomly assigned group labels (Nichols and Holmes, 2002).

Here, we will present evidence that there are even cases where parametric tests fail critically, and trial-based randomization tests are systematically biased towards false positive results. The problem arises because classification accuracy is sensitive to any kind of structure in the data. In particular, when the data contain distinct subclasses, the obtained classification accuracies can be systematically higher than the expected chance level, even when data from both conditions are sampled from the same distribution, i.e. the null hypothesis is true. A simplified example is illustrated in Figure 1. Here, classes A and B are each comprised of four distinguishable subclasses. The average CCR will be above 50% (here 70.9%, Fig. 1a) even if no systematic differences between classes A and B exist, e.g. centroids of class A and B are identical, and all differences pertain to random differences between subclasses. In our example separability of eight subclasses along two feature dimensions leads to an average CCR of 71.1% over all possible random attributions of subclasses to classes A and B (Fig. 1b). As we will work out below, this behavior can be observed to a varying degree in every data set in which classes consist of distinct subclasses (e.g. types of stimuli, groups of subjects, multiple recording sessions, blocks of fMRI recording, etc.).

There are three sources of variance that are of interest in the present considerations (Galbraith, et al., 2010) as can be described in the following model: $y_{ijk} = C_i + S_{ij} + \epsilon_{ijk}$. y_{ijk} are individual measurements, e.g. physiological brain responses to certain stimuli. $C_{i\in[1,2]}$ represents the class centroids, e.g. the influence of an experimental manipulation, the difference between patient and control group, or the responses to different conditions. S_{ij} are the centroids of the j^{th} subclass within class i. The variance σ_S^2 of the set of centroids reflect differences that are unrelated to the studied categories and in univariate analysis represent classical confounds. ϵ_{ijk} reflects the deviation from the subclass mean arising from the variance σ_W^2 of the data in each subclass. It represents measurement noise. The ratio of subclass-to-trial-variance, defined as the intraclass correlation ($ICC = \frac{\sigma_S^2}{\sigma_S^2 + \sigma_W^2}$), determines the extent to

which subclass variation affects statistical conclusions (Aarts, et al., 2014). Because the influence of subclasses often cannot be distinguished from the main effects (see for example (Malone, et al., 2016; Todd, et al., 2013), experiments and statistical analyses must be designed to avoid these confounds.

In classical statistics, it has been demonstrated that failing to accommodate for the effect of non-zero subclass variance can produce large false positive rates (Aarts, et al., 2014). In multivariate analyses, however, error variances do not average out but accumulate over features (Fan and Fan, 2008). Because multivariate linear classifiers take the differences over all features into account, subclasses will critically affect the result of an MVPA. In special cases, it is possible that both classes have identical means on all feature dimensions independently, but still CCRs systematically diverge from theoretical chance level because of random subclass differences (Fig. 1a). It is the common practice to ignore subclasses (Aarts, et al., 2014; Galbraith, et al., 2010; Lazic, 2010). We will show using real EEG data as well as synthetic data sets that subclass variance spuriously increases classification accuracy. In the present paper, we will investigate the boundary conditions and consequences of this phenomenon and describe a method to circumvent false positive results.

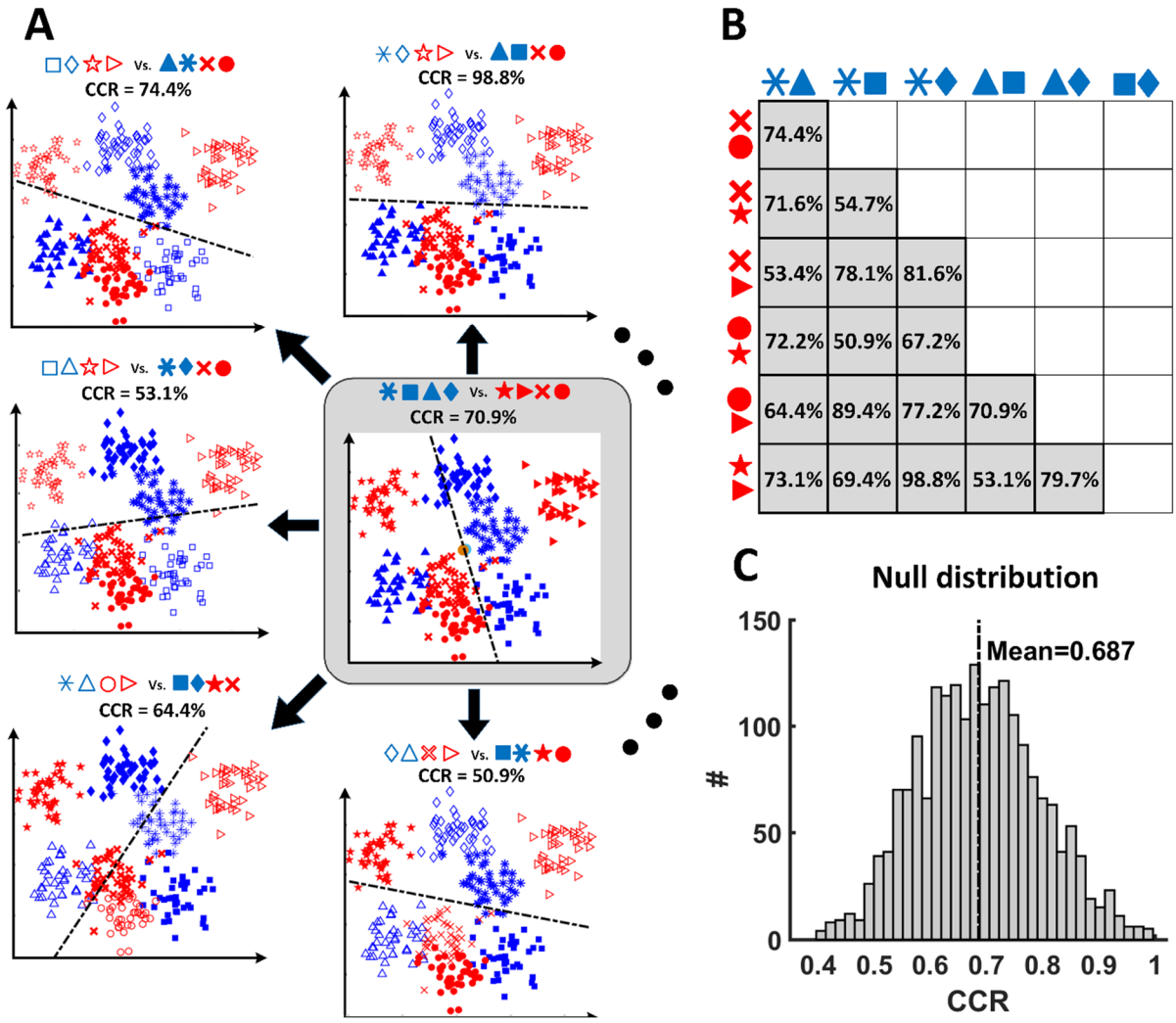


Figure 1: Classification accuracy in data with subclasses can exceed chance level even if data are randomly attributed to two conditions. (a) Center: An exemplary data set with 4 subclasses per class (blue and red). Although classes have almost identical means in both dimensions (represented by two filled circles in the center), LDA with 2-fold cross validation leads to 70.9% classification accuracy. Surrounding plots represent other random subclass-class relations (open vs. closed symbols). Note that only few of these random assignments show close to chance level CCRs. (b) The 8 subclasses can be randomly divided into two classes of four subclasses in 18 ways. The table here shows all 18 possible configurations with their respective CCRs. The average CCR is 71.1%. (c) Simulating 1000 data sets with the same structure as in A (8 subclasses randomly assigned to 2 classes, no overall difference between classes), results in a null distribution with a mean CCR of 68.7%.

Experimental and theoretical work

Practical example 1: Biased classification results in two-way designs with nested subclasses

The following example illustrates how a nested factor can influence the expected chance level in an experiment that investigates the EEG responses to the presentation of digits and letters (Alizadeh, et al., 2017). 20 stimuli (10 digits [class A], 10 letters [class B]) were presented in a working memory task to 19 subjects for 100 ms with an interstimulus interval of 900 ms. 128-channel EEG was recorded during the task using an active 128-channel Ag/AgCl-electrode system (ActiCap, Brain products, Gilching, Germany) with 1 kHz sampling frequency and a high-pass filter of 0.1 Hz. Electrodes were placed according to the extended international 10-20 electrode system. Continuously recorded EEG data was low-pass filtered offline at 40 Hz and divided into epochs of one second, starting 50 ms before stimulus onset. Artefact rejection was done in a semiautomatic process using custom MATLAB scripts. Epochs containing artefacts were removed from the data set, channels that contained too many epochs with artefacts were removed and interpolated using routines provided by EEGLAB (Delorme and Makeig, 2004). Single trial classification of digits and letters resulted in a significant mean classification accuracy of 54.1% over all 19 subjects in two sessions (trial-wise permutation test per subject, Fisher's method for aggregation over subjects: $p < 8 \times 10^{-10}$ and $p < 10^{-10}$ for two sessions respectively). However, as suggested in Fig. 1, because the experimental stimuli split the data set into 2×10 subclasses (digits and letters) and the existence of subclasses can induce detectable differences between conditions, it cannot be assumed that the chance level CCR is 50%.

To adjust for this possible bias, the null distribution for data with intact subclass structure, but without information about the actual classes must be determined. To do so, data can be permuted in a way that keeps subclasses together but still assigns random class labels, thus effectively removing any class-related

information. This procedure uses randomization at the subclass level instead of the usual trial-level randomization. Trial-level randomization treats every trial as an independent observation and will remove any systematic information from the data. This will result in the null distribution of the data assuming that no systematic relation between trials exists. If the data contains subclasses, trials within these subclasses are, by definition, systematically related. Because here we want to control for the influence of this structure at the subclass level, the dependencies at the subclass level must be kept intact while removing class-level information. To achieve this, we assign 5 digits and 5 letters to one class and the other 5 digits and 5 letters to the other class. We can draw randomly from $\frac{1}{2}\binom{10}{5}^2 = 31752$ possible permutations of these random assignments if labels of subclasses are permuted between classes in a balanced fashion. If classification results are on average above 50%, this must be related to the subclass structure of the data. Here, such classification over 1000 random permutations results in an average CCR of 50.9% (95% CI: [50.2%, 51.6%]. Note that the shift from 50% is statistically significant. To obtain an adjusted significance of the CCR with classes intact, we compared it with the distribution of results with the subclass structure intact, by randomly replacing class labels for half the digit and letter stimuli such that all instances of an exemplary digit A_i were assigned letter class labels (subclass-level randomization, Fig. 2A). The adjusted probability of the actual CCR occurring by chance in the presence of subclasses turns out to be $p = p < 5 \times 10^{-6}$ and $p < 10^{-6}$(subclass-wise permutation test per subject, Fisher's method for aggregation over subjects) compared with unadjusted $p < 8 \times 10^{-10}$ and $p < 10^{-10}$ (see above).

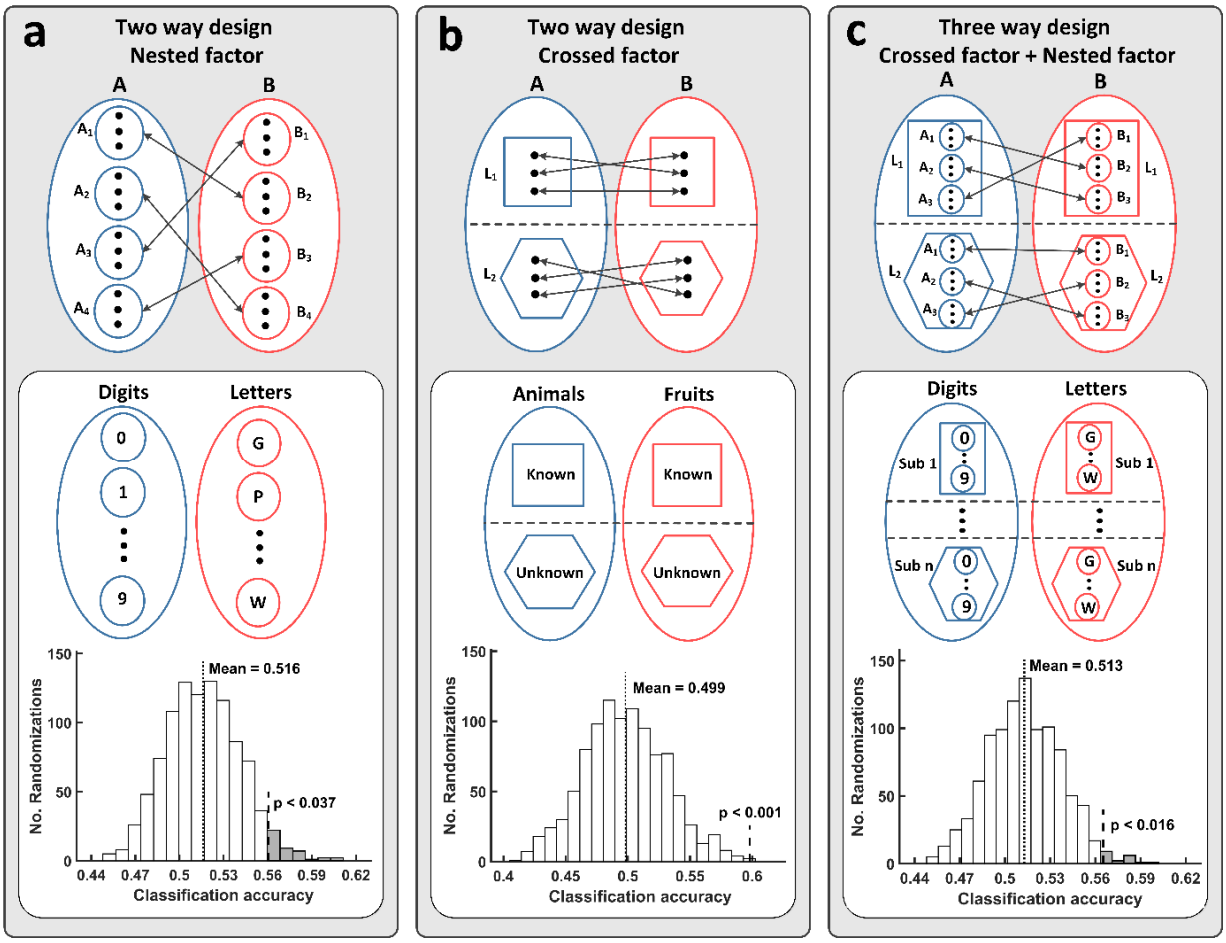


Figure 2: Subclasses can be either crossed or nested. The key difference between these cases in permutation tests is the way in which labels should be shuffled. (a) If subclasses are nested, blocked relabeling on the subclass level should be used. The histogram shows the null distribution and p-value for a subject of Experiment 1. (b) For crossed subclasses, labels should be shuffled within each level of the crossed factor. The histogram shows the null distribution and p-value for a subject of Experiment 3A. (c) Crossed and nested factors can be both present simultaneously. In such cases, the permutation principle remains the same as before. That is, the trials within nested subclasses must be kept together during permutation and shuffling should occur on the levels of the crossed factor. The histogram shows the null distribution and p-value for data of Experiment 4.

Simulating biased classification results in data with nested subclasses

To investigate the effect of nested subclasses on classification accuracies systematically, we used synthetically generated data with varying relations between subclass and class variance. Nested subclasses are subclasses that do not overlap between the two classes (e.g. 10 letters and 10 digits as in Experiment 1 above). As we will show, in this kind of experimental design, subclasses biases classification results most strongly. We studied the distribution of CCRs in two series of 100-dimensional, two-class experiments where each class contained either 2 or 10 subclasses per class. Each data set consisted of 120 observations per class in a nested two-way design. Data was sampled from normally distributed populations with identical trial variance ($\sigma_W^2 = I$) and varying subclass variance ($\sigma_S^2 = a \times I, a \in [0, 1]$). In addition, we varied the size of the main effect. We classified data from each simulated experiment with linear SVM (with cost parameter $C = 1$) using 2-fold cross-validation. For each set of parameters, we repeated the whole sampling and classification procedure 5000 times to achieve a stable estimate of CCRs (Fig. 3).

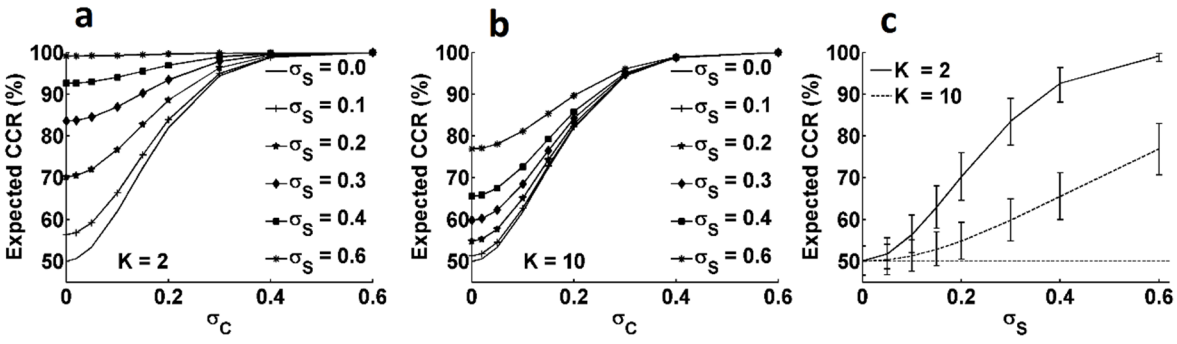


Figure 3: Expected CCRs when data contains subclasses (a, b) Expected CCRs for data sets with a constant trial variance ($\sigma_W = I$) and varying class and subclass variance (σ_C and σ_S) with $K = 2$ and $K = 10$ subclasses per class. (c) Expected CCRs and standard deviations for data sets with zero effect size ($\sigma_C = 0$).

If the classes are indistinguishable (i.e. main effect is zero, class centroids $C_1 = C_2$), subclass-effects contribute most strongly to classification accuracy. With increasing between-class variance, the relative influence of subclass variance

diminishes (Fig. 3a). A higher number of subclasses also mitigates the influence of the subclass effect (Fig. 3b). From these graphs, it is obvious that a high CCR per se does not indicate the presence of class-related information in the data, and that a significance test cannot assume a chance level of 50% when nested subclasses are present in the data. To test for significance, the actual CCR obtained from the data must be compared to a null distribution that removes the class-level information, but retains subclass-level structure. For our simulations, Fig. 3c depicts the mean of these null distributions for varying levels of subclass variance and different numbers of subclasses. It becomes apparent that subclasses bias the expected CCR, and that the bias is higher for data with fewer subclasses per class.

The simulations in Fig. 3 show that CCRs depend on subclass variance and on the number of subclasses. To investigate the implications of this observation further, we develop an analytical description of classification rates when data with subclasses are analyzed using linear classifications. We assume our data set consists of two sets of $N \times K$ independent random vectors $\vec{x}_n^{(k)}, \vec{y}_n^{(k)}$ where $k, k' \in \{1, \ldots, K\}$ labels the subclasses and $n, n' \in \{1, \ldots, N\}$ identifies the sample index in each of the subclasses. The task of the linear classifier is to separate x and y into two categories. As a model for the linear classifier, we use Linear Discriminant Analysis (LDA). Data distribution within the subclasses is assumed to be Gaussian with variance σ_W^2, the distribution of the subclass means is also assumed to be Gaussian with variance σ_S^2 and expected values $\hat{\mu}, \hat{\nu}$. Under these conditions, we determine the expected CCR to be as described in Theorem 1 (Appendix A). From this theorem directly follows corollary 1 (Appendix B), which states that the estimated CCR for data sets with no effect ($\hat{\mu} = \hat{\nu}$) is a decreasing function of the number of subclasses (K) and an increasing function of subclass variance (σ_S). It also shows that for zero effect size, CCR only depends on the intraclass correlation ICC and the number of subclasses K. In particular, CCR is 50% only when the intraclass correlation $ICC = 0$ and is a monotonically increasing function of ICC (and of σ_S^2 if σ_W^2 remains fixed). Note that in the case

when $ICC = 0$, by definition no subclasses exist in the data, an assumption that is mostly not tested in but typically could be present in experiments which a) compare different kinds of stimuli, b) test more than one subject, or c) collect data in multiple sessions. This in turn means that each subclass introduced to the experimental design will have a measurable impact on classification results.

Adjusting permutation tests to correct for subclass bias

Since subclass differences inflate classification accuracy, significance tests must take this subclass-related bias into account. Here, we propose a permutation strategy, which addresses this problem by adjusting the null distribution. The general idea of permutation tests is to shuffle data labels in such a way that all information related to the classes under investigation is removed. Typically, this is achieved by shuffling labels on the level of individual trials. However, this method also removes any subclass-related structure from the data, i.e. subclasses no longer have distinct centroids. Thus, the null-distribution obtained in this way has no bias, contrary to the actual data. We therefore have to remove the information pertaining to the classes while preserving subclass-related information. To achieve this, we permute class association on the subclass-level instead of the trial level (see Fig. 2 for illustrated examples of such permutation procedures). In the case of our practical example above (Experiment 1: two-way design with nested subclasses) we would randomly replace class labels for digit and letter stimuli in such a way that all members of an individual subclass are consistently relabeled (see Fig. 2a). This method assumes that trials of a subclass are not independent and subclass centers are determined by random or systematic, class-unrelated influences. It retains the dependence between trials of a subclass while it removes the relation between subclass and class centroids. If classification accuracy for the real data is higher than that for permuted data, it can be concluded that there is a systematic difference between the classes.

To remove all class related information, we randomly switch class labels on the subclass level, with consistent relabeling of all items belonging to a subclass. The distribution of CCRs for these relabeled data represents the null distribution with which the actual CCR must be compared. The total number of possible permutations is $\frac{1}{2}\binom{K}{K/2}^2$, where K represents the number of subclasses per class. Importantly, the number of possible resampling is only determined by the number of subclasses and is independent of the total number of trials. When K is large, the null distribution of CCRs can be sampled sufficiently well. However, if K is small, this is not possible. For instance, for $K = 4$, the maximum number of permutations is 18. In particular, the number of possible permutations for $K < 6$ does not allow to reach significances with $\alpha < 0.05$, because too few points of the random distribution can be estimated. In these cases, when the same statistic is available for multiple subjects or sessions, we propose to use the group null distribution which can be obtained by the non-parametric method described in (Stelzer, et al., 2013). In this method, the mean CCR from real data over all the subjects is tested against a null distribution, which is obtained by repeatedly averaging randomly sampled CCRs from the subclass-level permutations from each subject.

Quantification of significance bias in data with nested subclasses

To systematically study how subclasses variance affects significance tests, we used the simulated data sets described above and produced their respective null distributions once using adjusted subclass-wise and once using unadjusted trial-wise permutation tests. We calculated the expected p-values for varying sizes of class and subclass effects (Figs. 4a-b). Whereas the false positive rate is defined as the percentage of falsely significant results when no class-related effect is present ($\sigma_C = 0$), Figure 4b illustrates that there are also more significant results than expected when there is an extant but small class effect. Here, for any given class- and subclass-variance, we define a measure of significance bias (SB) as the normalized difference between the number of data sets with significant

p-values and the number of data sets with significant p-values with the same amount of class-variance but subclass variance is not present. In contrast to false positive rates, this value can also be calculated when there is an actual effect. The intuition behind this measure is to quantify the contribution of nuisance effects (e.g. the effect of subclasses) on the performance of a significance test that should only be sensitive to the primary effect (e.g. the class effect). Figures 4c-d show significance biases for adjusted subclass-wise and unadjusted trial-wise permutation tests, respectively. It becomes obvious that trial-wise permutation does not correct for the confounding effects of subclass variance and results in liberally biased p-values when a subclass effect exists and the class-related effect is small or not present. On the other hand, the adjusted test successfully limits the false positive rate and is only slightly conservative (negative SB) when subclass variance is high compared to class variance.

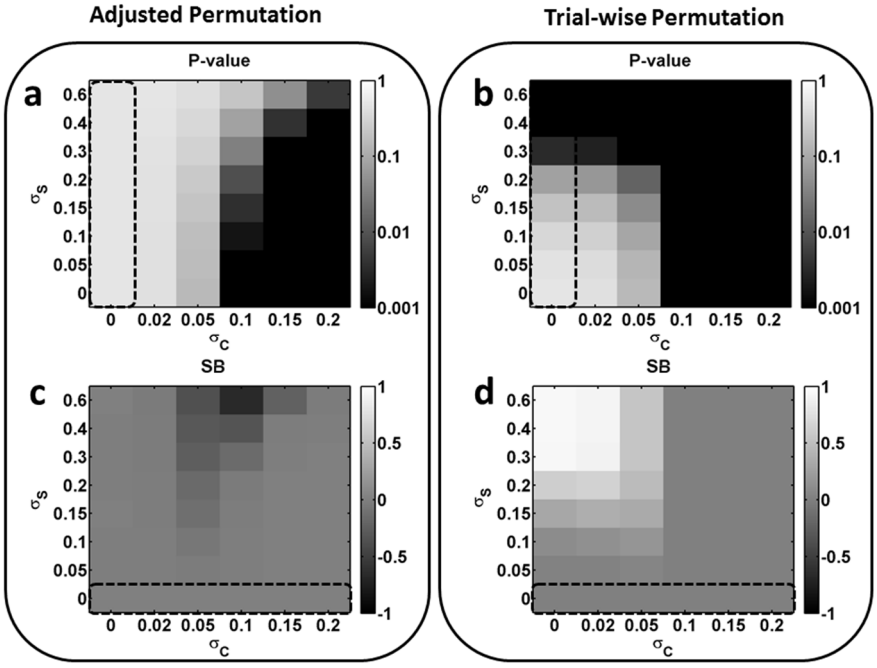


Figure 4: Randomization results for data with subclasses ($K = 10$). (a, b) The area delimited by the dashed rectangle shows the expected p-values for a main effect of zero. Importantly, even when CCRs are strongly biased because of nonzero subclass variance, p-values remain constant for the adjusted subclass-wise randomization test. The trial-wise permutation test fails to remove the bias introduced by subclass variance. Note that even small subclass effects result in falsely positive significance tests (dark grey and

black squares for $\sigma_C = 0$). In the ideal case, the presence of subclass variance should not affect the rate of significant findings, i.e. p-values should be identical within each column over σ_C. Obviously, for increasing values of σ_S, trial-wise randomization leads to too many significant results for $\sigma_C = 0$ whereas subclass-wise permutation leads to a reduced number of significant findings for small values of σ_C. (c, d) To illustrate the difference between the number of significant results when subclass variance is not present (dashed rectangle) and when it is present, we calculate normalized difference between the number of data sets with significant p-values and the number of data sets with significant p-values with the same amount of class-variance but subclass variance is not present (significance bias, SB). This value shows that adjusted permutation is unbiased for small and sufficiently large class effects and shows only a small conservative bias (negative SB) when subclass variance is substantially larger than class variance. Testing with the unadjusted, trial-wise permutation, on the other hand, is too liberal when the class effect is small or null and even a subclass effect is present. This test is therefore leading to larger number of false positive results.

Considerations for designs with crossed factors

Factors in experimental designs can be nested or crossed. For nested factors the levels of the subordinate factor differ for each level of the superordinate factor (e.g. different subjects are tested for two experimental condition, different sets of stimuli are used in two experimental conditions, …). Crossed factors are formed when the second factor coexists in both classes, e.g. the same subjects are tested under two types of stimulation, the same set of stimuli is used in two experimental conditions, …). Unlike in nested designs, in data sets with crossed factors, the subclasses structure is the same within all classes (centroids of subclasses show the same distances and relations). Therefore, classification accuracy should remain unbiased as long as there is no interaction between the factors class × subclass, which can often be assumed. In the example described below, we would expect that the subclass manipulation well-known vs. unknown member exerts the same effect on brain activity for both members of the classes fruits and animals, or in other words: we would expect a main effect on brain activity of both the subclass factor well-known vs. unknown, as well as the class factor fruit vs. animal (see Fig. 2b). However, in crossed designs,

subclass centroids within each class can still distribute in a way that e.g. masks differences between the classes of interest (Hohne, et al., 2016). Unwanted effects of subclass variance can be avoided if permutation is done on a trial-by-trial basis within each subclass (see also (Anderson and Ter Braak, 2003; Gonzalez and Manly, 1998; Manly, 2006). Figure 2b shows the appropriate randomization strategies for crossed subclass factors.

Practical example 2: Two-way designs with crossed factors

As an example, we analyzed data from a two-factorial experiment that recorded ERP responses to the presentation of 60 pictures of fruits and 60 pictures of animals in Experiment 2. Half of the animals and fruits were common, well-known objects, the other half were rare and unknown. 128-channel EEG was recorded during stimulus presentation from 19 healthy subjects as described in Experiment 1. Every picture was presented for 300 ms, followed by a black screen of 1.5 s. The subjects were then asked to decide if the presented picture was familiar or unfamiliar. Familiarity was defined as knowing the animal or fruit by name. Between trials, a fixation cross appeared for 400 ms. ERPs were calculated for epochs of 1 s starting at stimulus onset. Data was prepared as in Experiment 1. We classified data within each subject with linear SVM using 2-fold cross-validation. The average CCR to classify animal and fruit trials over all subjects was 58.1%. To test for significance, we randomly relabeled trials for each subject within the 'known' and 'unknown' subclasses (Fig. 2b). This procedure results in null distributions that had a mean of 50.08%. Comparing the CCRs obtained from the actual data with these randomized distributions shows that classification is significant in 12 subjects. To test for population significance, we aggregated individual significances using Fisher's method, resulting in a population $p < 6 \times 10^{-12}$. Note that if permutation is not restricted within subclasses, subclass distribution between classes becomes uneven, and the null distribution becomes biased (here, this would lead to an average null distribution mean of 51.0%).

Effects of subject and session variability on classification accuracy

Brain responses recorded from different subjects and different sessions show marked differences. Here, we show how these variabilities can be addressed within the framework of subclasses. If not treated properly, session and subject variability can result in biased CCRs. It is possible to use all data from multiple subjects or sessions simultaneously to train a classifier if the null distribution is adjusted properly for significance testing. If all classes are repeated in each subject/session, the subject/session variability is crossed over the main effect and the permutation can be performed as in Experiment 2, i.e. randomization must be done within each subject/session. It must be noted, however, that this method can be very insensitive if there are large subject/session differences. In this case, it might be more useful to analyze each session separately {paper from Berlin on crossed factors}. On the other hand, if each subject/session consists of data from only one class (nested design), then each subject/session has to be treated as an independent subclass analogous to Experiment 1, i.e. randomization must be done on the level of subclasses (subjects/sessions) consistently switching all labels of data belonging to one subclass (see Fig. 2a, A_i in this case referring to different subjects or sessions). In this way, variability between subjects/sessions will be properly considered as a source of unwanted positive classification bias.

For the following Experiment 3, which explores the effect of subject and session variability on CCRs, we recorded 128-channel ERP responses during the presentation of pictures of faces and houses. EEG was recorded from 20 healthy subjects in two sessions with the same parameters as described in Experiment 1. Presentation time was 100 ms, the ERP was calculated from 100 ms before to 900 ms after onset of stimulus presentation. We classified the data with linear SVM using 2-fold cross-validation.

Practical example 3A: Dealing with session variability as a crossed factor

To assess effects of session variability, we used 30 trials per class from each of the two sessions in Experiment 3A. We classified data within each subject, collapsing trials from both sessions. Average classification accuracy over all subjects was 64.6% (see Figure 5A). Since both classes are present in each session, we permuted data within each session (see Fig. 2b, L_1 and L_2 denoting the different sessions). The mean of null distributions averaged over all 20 subjects is 50.0%. To calculate the p-value, we use a group significance test as described above (Stelzer, et al., 2013), which indicates the classification accuracy is significantly above chance ($p < 10^{-5}$).

Practical example 3B: Dealing with subject variability as a nested factor

In another analysis of the same data (Experiment 3B), we used 30 'face' stimuli from 10 subjects and 30 'house' stimuli from another set of 10 subjects, resulting in 300 trials per class. We classified the data with trials collapsed across subjects, resulting in a classification accuracy of 82.4% (see Figure 5B). To test for significance, we treated data from different subjects as subclasses and randomized the class assignment of the subjects, keeping all trials of a subject together. This results in a null distribution with a mean of 78.5% and a 95% confidence interval of [74.6 – 82.2%] and $p < 0.045$. Note that although classification accuracy is much higher than in the within-subject analysis of Experiment 3A, the p-value is actually worse. This indicates that there are large differences between subjects, which strongly biases classification accuracy. In contrast, trail-wise randomization would underestimate the p-value by several orders of magnitude (95% confidence interval: [47.5 52.5].

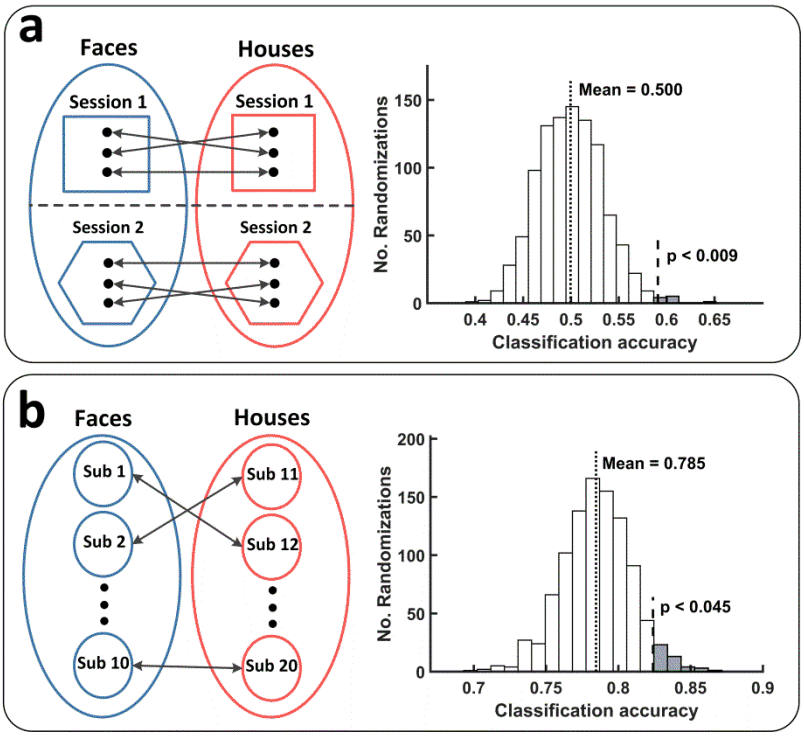


Figure 5: Permutation strategies for data sets that contain recordings from multiple subjects or sessions. (a) When both classes are present in each session (i.e. crossed design), CCR remains unbiased and the null distribution is unchanged. Permutation is done within each session separately. The histogram shows the null distribution and p-value for one of the 20 subjects in Experiment 3A. (b) Data from different subjects form subclasses with non-zero subclass variance. Although the actual CCR is quite high (82.4%), the p-value is only barely significant (p = 0.045). Note that if trial-wise permutation is applied, the mean of the randomization distribution will be 50% and the p-value becomes misleadingly small. The histogram shows the adjusted null distribution and p-value for data in Experiment 3B.

Adjusting permutations test in three-way designs with both nested and crossed factors

Adjusted permutation testing can also be used in more complex experimental designs, e.g. when a data set has both nested and crossed subclasses (see Figure 2c). In the following practical example (Experiment 4), we illustrate how to adjust permutation tests in an experiment that compared the effect of digit and

letter presentation in a group of subjects. We took data from three subjects of Experiment 1, selecting trials during which four digits ('2', '3', '8', and '0') and four letters ('B', 'D', 'W', and 'Y') were presented. This data set has three factors: the primary factor is the two classes 'letters' and 'numbers', the second factor is the subjects, which is crossed over conditions, and the third factor is the individual digits and letters, which is nested within the second factor (see Figure 2c). We classified data with linear SVM using 2-fold cross-validation, which resulted in a classification accuracy of 57.6%. To test for significance, we permuted the labels of the data in a way that the individual numbers and letters remained together as subclasses and data from different subjects were permuted separately. We shuffled the data 5000 times resulting in a null distribution with an average CCR of 51.3% and $p < 0.019$. Importantly, simple permutation on the trial level would largely overestimate the significance level ($p < 0.003$)

Discussion

Many neuroscience data sets comprise subclasses, either because of requirements of the experimental design (multiple subjects, sessions, recording sites, laboratories etc.) or because of the nature of experimental stimuli (Aarts, et al., 2014). We show that these subclasses can systematically bias classification accuracy and induce classifiability even when no actual class-related effect exists. This happens because subclasses have distinct centroids, particularly in high-dimensional space (Fan and Fan, 2008; Jamalabadi, et al., 2016). These differences can be detected by MVPA, even when they cancel out in the classes as a whole (see Figure 1). Therefore, when testing for significance, it is necessary to account for biased chance levels. We propose a ways of permutation testing that can provide the correct null distribution for classification rates.

Subclasses are formed by groups of trials that have a common covariance that is not the class-related covariance under investigation (Lazic, 2010; Todd, et al.,

2013). This occurs when data is gathered in distinct blocks or when there are clusters of trials with similar physical (e.g. color) or cognitive (e.g. concepts, emotions) properties. In univariate analyses, these subclasses represent classical confounds. Their effects usually cancel out and become irrelevant when subclasses are randomly distributed around the class means. In a higher dimensional space, noise introduced by subclasses accumulates over dimensions and contributes to the distinctness of the subclasses. This effect increases sharply with an increasing number of dimensions (due to the so-called "curse of dimensionality"). It is therefore vital to adjust for the bias in classification rates introduced by subclasses in multivariate analyses.

The relation between trial-variance and subclass-variance determines the extent to which the within-subclass covariance affects conclusions from the data. This ratio of trial-to-subclass-variance can be quantified in terms of the intraclass correlation ICC. It can be shown that even a small ICC of 0.01 can increase the false positive rate to more than 20% when it is expected to be at $\alpha = 0.05$ (Aarts, et al., 2014). When MVPA is employed, the deviation of CCRs from expected chance levels depends only on ICC (see Appendix B), and an ICC as small as 0.1 can spuriously increase CCR by 10% (see Figure 3c). When average ICCs in neuroscience have been found to average 0.19 with a range between 0.00 and 0.74 (Aarts, et al., 2014), it can be assumed that reported CCRs in data with subclasses can be biased by up to 50%. Because parametric tests cannot account for this bias, statistical significance has to be determined using a permutation procedure that does not eliminate the bias.

Making subclasses less prominent reduces the bias. According to our simulations and the analytical solution in Appendix B, this can be achieved in two ways: either by decreasing the subclass variance or by increasing the number of subclasses. While decreasing the subclass variance might be difficult in real-world experiments, increasing the number of subclasses is often a possibility. Although a higher number of subclasses is preferable to a lower number (Aarts, et al., 2014; Anderson and Ter Braak, 2003; Lazic, 2010), 6

subclasses already allow reasonable permutation testing. However, for a number of subclasses of $K < 6$, p-values better than p = 0.05 cannot be reached. Therefore, a proper significance test for such data sets cannot be conducted on a single subject level. In these cases, we propose to estimate the group level p-value, which can be estimated by aggregating permutations over subjects (Stelzer, et al., 2013).

The absolute height of CCRs is not informative about classification success when cross-validated classification is used for hypothesis testing. Next to the fact that null-distributions can be skewed (Jamalabadi, et al., 2016), the present simulations and experiments show that in nested experimental designs, which are common in the life sciences, subclasses inject systematic dependencies in the data structure that can lead to spuriously high CCRs. When permutation testing is adjusted to respect these dependencies, the null-distribution may no longer be centered around 50%, but can even deviate significantly from the estimated chance level if the same dependencies are ignored. In the practical example of Experiment 3B, we show that between-subject variance, when subjects are a nested factor, can dramatically increase CCR from 64.6% (within-subject crossed factors design) to 82.4% (between-subject nested factors design), when the number of trials is kept constant. Still, the lower CCR represents a more robust result, showing a higher significance level when estimated from the unbiased null-distribution obtained by subclass-level randomization.

In particular cases, including subclasses in the classification procedure can actually be used to improve classification accuracy (Hastie and Tibshirani, 1996; Hohne, et al., 2016; Zhu and Martinez, 2006). Thus, issues of subclass variance can be avoided by performing classification within subclasses. This method, however, can only be employed in data sets with crossed subclasses, i.e. when every subclass is repeated in all classes. In nested designs, which were the main focus of the present paper, randomization must always keep the subclass structure intact and classification must thus include all subclasses.

In this article, we explored the use of MVPA for data sets with subclasses. We show classification accuracies can be strongly biased even with small amounts of subclass-related variance. We therefore suggest that statistical significance should be tested with nonparametric permutation tests that accommodate for the bias in CCR induced by these subclasses. A more diverse range of stimuli can also be used to mitigate the bias and result in more reliable classification results.

Appendix A: Theorem 1

We assume our data set consists of two sets of $N \times K$ independent random vectors $\vec{x}_n^{(k)}, \vec{y}_n^{(k)}$where $k, k' \in \{1, \ldots, K\}$ labels the subclasses and $n, n' \in \{1, \ldots, N\}$ identifies the sample index in each of the subclasses. The task of the linear classifier is to separate x and y into two categories. As a model for the linear classifier, we use LDA. We therefore can map the d-dimensional vectors$\vec{x}_n^{(k)}, \vec{y}_{n'}^{(k')}$ onto the coordinates $\xi_n^{(k)}$and $\eta_{n'}^{(k')}$ w.r.t to the axis defined by the difference of the mean values of the two classes. Furthermore, we label the empirical means of the classes as:

$$\mu^{(k)} = N^{-1} \sum_n \xi_n^{(k)}, \nu^{(k)} = N^{-1} \sum_n \eta_{n'}^{(k')}$$

The distributions of within the subclasses is assumed to be Gaussian with variance σ_W^2 , the distribution of the subclass means is also assumed to be Gaussian with variance σ_S^2 and generally different expected values $\hat{\mu}, \hat{\nu}$. In the case of the two categories are undistinguishable the two are identical (no signal), $\hat{\mu} = \hat{\nu}$,. For every realization the means of $\mu^{(k)}, \nu^{(k')}$ will be different from $\hat{\mu}, \hat{\nu}$, and thus we also introduce the empirical means $\bar{\mu}, \bar{\nu}$, which underlie the estimated signal $\delta = \bar{\mu} - \bar{\nu}$.

Besides, we can compute the whole variance of the data set as:

$$\sigma^2 = \frac{1}{2}[var(\xi) + var(\eta)] = var(\xi) = < [\xi - \hat{\mu}]^2 >$$

$$= < \left[\left(\xi - \mu^{(k)} \right) + (\mu^{(k)} - \hat{\mu}) \right]^2 > = \sigma_w^2 + \sigma_s^2$$

Under these conditions, assuming that the whole data variance $\sigma^2$is constant, the expected CCR for such data can be estimated as:

$$CCR = \int_{-\infty}^{+\infty} dq \frac{sign(\left(1+\frac{\rho}{K}\right)q+\tilde{\delta})}{2} N(q)[erf\left(\frac{q\frac{\rho}{K}+\frac{\tilde{\delta}}{2}}{\sqrt{2}\lambda}\right) + erf(\frac{q+\frac{\tilde{\delta}}{2}}{\sqrt{2}\lambda})]$$

With $N(q)$ denoting the normal distribution, $\rho = \frac{\sigma_s^2}{\sigma^2}$, $\tilde{\delta} = \frac{\delta}{\sigma}\sqrt{1+\frac{\rho}{K}}$, and $\lambda^2 = \frac{\rho}{2K}(1-\frac{\rho}{K})$.

Proof.

For LDA, the Correct Classification Rate (CCR) can be computed as the probability that during testing, a data point of class x is on the same side of the classification threshold $\theta = \frac{\bar{\mu}+\bar{\nu}}{2}$ as the empirical mean $\hat{\mu}$, and a data point of class y is on the opposite side:

$$CCR = [p(\xi > \theta, \bar{\mu} > \bar{\nu}) + p(\xi < \theta, \bar{\mu} < \bar{\nu})]p\xi + [p(\eta > \theta, \bar{\mu} > \bar{\nu}) + p(\eta < \theta, \bar{\mu} < \bar{\nu})]p\eta \ (1)$$

Under the assumption of symmetry between class labels, i.e., $p\xi = p\eta = \frac{1}{2}$, equally distributed subclass means, and equally within-class distributions equation (1) must be symmetrical with respect to exchanging x and y and thus we can obtain the CCR from

$$CR = p(\xi > \theta, \bar{\mu} > \bar{\nu}) + p(\xi < \theta, \bar{\mu} < \bar{\nu}) (2)$$

Denoting $\vec{\mu} = \left(\mu^{(1)}, \dots, \mu^{(K)}\right)^T$ and $\vec{\nu} = \left(\nu^{(1)}, \dots, \nu^{(K)}\right)^T$, we thus can write

$$CCR = \int d\vec{\mu} d\vec{\nu} p\left(\xi > \theta, \bar{\mu} > \bar{\nu} | \vec{\mu}, \vec{\nu}\right) p(\vec{\mu}, \vec{\nu}) + p\left(\xi < \theta, \bar{\mu} < \bar{\nu} | \vec{\mu}, \vec{\nu}\right) p(\vec{\mu}, \vec{\nu})$$

$$= \int d\vec{\mu} d\vec{\nu} p(\vec{\mu}, \vec{\nu}) [\int_{\theta}^{\infty} d\xi p(\xi | \vec{\mu}, \vec{\nu}) H(\bar{\mu} - \bar{\nu}) + \int_{-\infty}^{\theta} d\xi p(\xi | \vec{\mu}, \vec{\nu}) H(\bar{\nu} - \bar{\mu}) \ ] (3)$$

with H denoting the Heaviside step function. Substituing $\xi = \theta + t$, the integrals of ξ can be transformed to

$$CCR = \int_0^\infty dt \int d\vec{\mu} d\vec{\nu} p(\vec{\mu}, \vec{\nu}) \times [p(t+\theta|\vec{\mu}, \vec{\nu}) H(\bar{\mu} - \bar{\nu}) + p(-t+\theta|\vec{\mu}, \vec{\nu}) H(\bar{\nu} - \bar{\mu})] \quad (4)$$

The subsample means are independent $p(\vec{\mu}, \vec{\nu}) = \prod_k p(\mu^{(k)}) \prod_{k'} p(\nu^{(k')})$. Moreover, $\vec{\nu}$ only affects the integral with its mean $\bar{\mu}$, and thus $d\vec{\nu} p(\vec{\nu}) = d\bar{\nu} p(\bar{\nu})$.

All distributions are assumed to be Gaussians. We therefore can express all probabilities by Gaussian distribution G. In particular

$$p(\xi|\vec{\mu}, \vec{\nu}) = \frac{1}{K} \sum_k G\left(\xi - \mu^{(k)}, \sigma_w\right)$$

$$p(\bar{\nu}) = G(\bar{\nu} - \hat{\nu}, \sigma_s/\sqrt{K})$$

$$p\left(\mu^{(k)}\right) = G(\mu^{(k)} - \hat{\mu}, \sigma_s) \quad (5)$$

Inserting eqs. (5) into eq. (4), we obtain:

$$CCR = \frac{1}{K} \sum_k \int_0^\infty dt \left(\prod_k d\mu^{(k)} G(\mu^{(k)} - \hat{\mu}, \sigma_s)\right)$$

$$\times \left[\int_{-\infty}^{\bar{\mu}} d\,\bar{\nu} G(\bar{\nu} - \hat{\nu}, \sigma_s/\sqrt{K}) G(t + \theta - \mu^{(k)}, \sigma_w)\right.$$

$$\left. + \int_{\bar{\mu}}^\infty d\,\bar{\nu} G(\bar{\nu} - \hat{\nu}, \sigma_s/\sqrt{K}) G(-t + \theta - \mu^{(k)}, \sigma_w)\right] \quad (6)$$

Substituting $u = \bar{\nu} - \bar{\mu}$ the integrals over $\bar{\nu}$ can be combined such that

$$CCR = \frac{1}{K} \sum_k \int_0^\infty dt \int_0^\infty du \left(\prod_k d\mu^{(k)} G(\mu^{(k)} - \hat{\mu}, \sigma_s)\right)$$

$$\times [G(\bar{\mu} - \hat{\nu} - u, \sigma_s/\sqrt{K}) G(\bar{\mu} - \mu^{(k)} + t - u/2, \sigma_w)$$

$$+ G(\bar{\mu} - \hat{\nu} - u, \sigma_s/\sqrt{K}) G(\bar{\mu} - \mu^{(k)} - t + u/2, \sigma_w)] \quad (7)$$

Completing squares in the last two Gaussian distributions and integrating over all $\mu^{(k)}$ with $\kappa \neq k$, yields

$$CCR = \frac{1}{K}\sum_{k}\int_0^\infty dt \int_0^\infty du \int d\mu^{(k)} G(\mu^{(k)} - \hat{\mu}, \sigma_s)$$

$$\times K[G(\mu^{(k)}\left(1 - K\left(\frac{\sigma_s}{\sigma^*}\right)^2\right) + (K-1)\hat{\mu} + \widetilde{C_+}(u,t), \Sigma^*) \times G(\mu^{(k)}$$

$$- B_+(u,t), \sigma^*/\sqrt{K})$$

$$+K[G(\mu^{(k)}\left(1 - K\left(\frac{\sigma_s}{\sigma^*}\right)^2\right) + (K-1)\hat{\mu} + \widetilde{C_-}(u,t), \Sigma^*) \times G(\mu^{(k)} - B_-(u,t), \sigma^*/$$

$\sqrt{K})]$ (8)

With $(\sigma^*)^2 = K\sigma_w^2 + \sigma_s^2$, $(\Sigma^*)^2 = (K-1)\sigma_s^2 + (K\sigma_w\sigma_s/\sigma^*)^2$, $B_\pm(u,t) = \hat{v} \pm (t + u/2)$,

And

$$\widetilde{C_\pm}(u,t) = (\pm K\left(t - \frac{u}{2}\right)\sigma_s^2 - K^2(\hat{v} \pm u)\sigma_w^2)/(\sigma^*)^2$$

Again, completing squares of the first and third Gaussian distribution results in

$$G\left(\mu^{(k)} - \hat{\mu}, \sigma_s\right)G\left(\mu^{(k)} - B_\pm(u,t), \frac{\sigma^*}{\sqrt{K}}\right)$$

$$= G(\hat{\mu} - B_+(u,t), \sigma^{**}/\sqrt{K}) \times G(\mu^{(k)} - D_\pm(u,t), \frac{\sigma^*\sigma_s}{\sigma^{**}})$$

With $(\sigma^{**})^2 = K\sigma_s^2 + (\sigma^*)^2$ and

$$D_\pm(u,t) = \hat{\mu}\left(\frac{\sigma^*}{\sigma^{**}}\right)^2 + KB_\pm(u,t)\left(\frac{\sigma_s}{\sigma^{**}}\right)^2$$

Solving the integral over $\mu^{(k)}$ in eq. (8) as a convolution of two Gaussians, we end up at

$$CCR = \int_0^\infty dt \int_0^\infty du\, [G(\hat{\mu} - \hat{v} - (t + u/2), \sigma^{**}/\sqrt{K})$$

$$\times G(\alpha(\hat{\mu}-\hat{v})-\beta u+\gamma t,\Sigma^{**}/\sqrt{K})$$

$$+G(\hat{\mu}-\hat{v}+(t+u/2),\sigma^{**}/\sqrt{K})\times G(\alpha(\hat{\mu}-\hat{v})+\beta u+\gamma t,\Sigma^{**}/\sqrt{K})] \quad (9)$$

With $\alpha=1-2\left(\frac{\sigma_S}{\sigma^{**}}\right)^2$, $\beta=1-\left(\frac{\sigma_S}{\sigma^{**}}\right)^2$, $\gamma=2\left(\frac{\sigma_S}{\sigma^{**}}\right)^2$, and

$$(\Sigma^{**})^2=(\Sigma^{*})^2+\left((1-K\left(\frac{\sigma_S}{\sigma^{*}}\right)^2)\frac{\sigma^{*}\sigma_S}{\sigma^{**}}\right)^2$$

Introducing the signal $\delta=\hat{\mu}-\hat{v}$ and substituting $v=t+u/2\mp\delta$ yields

$$CCR=2\int_{-\infty}^{\infty}dt\int_{0}^{\infty}dv\,G(v,\sigma^{**}/\sqrt{K})$$

$$\times[H(v-t+\delta)G(t-\beta v-\delta/2,\Sigma^{**}/(2K))$$

$$+H(v-t-\delta)G(t-\beta v+\delta/2,\Sigma^{**}/(2K))]\frac{1}{2}$$

$$=\int_{-\infty}^{+\infty}dv\frac{sign(v+\delta)}{2}G(v,\sigma^{**}/\sqrt{K})[erf\left(\frac{\gamma v+\delta}{\sqrt{2\Sigma^{**}/K}}\right)+erf(\frac{(2-\gamma)v+\delta}{\sqrt{2\Sigma^{**}/K}})]$$

$$=\int_{-\infty}^{+\infty}dq\frac{sign(\left(1+\frac{\rho}{K}\right)q+\tilde{\delta})}{2}N(q)[erf\left(\frac{q\frac{\rho}{K}+\frac{\tilde{\delta}}{2}}{\sqrt{2}\lambda}\right)+erf\left(\frac{q+\frac{\tilde{\delta}}{2}}{\sqrt{2}\lambda}\right)] \quad (10)$$

With $N(q)$ denoting the normal distribution, $\rho=\frac{\sigma_S^2}{\sigma^2}$, $\tilde{\delta}=\frac{\delta}{\sigma}\sqrt{1+\frac{\rho}{K}}$, and $\lambda^2=\frac{\rho}{2K}(1-\frac{\rho}{K})$.

Appendix B: Corollary 1

The expected correct classification using LDA for data sets with no effect ($\hat{\mu} = \hat{v}$) is an increasing function of subclass variance σ_S, a decreasing function of number of subclasses K, and is 50% only when $ICC = 0$.

Proof. Substituting for equation 10 with $\delta = 0$ we have

$$CCR = \int_{-\infty}^{+\infty} dq \frac{sign\left(\left(1+\frac{\rho}{K}\right)q\right)}{2} N(q)[erf\left(\frac{q\frac{\rho}{K}}{\sqrt{2}\lambda}\right) + erf\left(\frac{q}{\sqrt{2}\lambda}\right)$$

$$= \int_{-\infty}^{+\infty} dq \frac{sign(q)}{2} N(q)\left[erf\left(\frac{q\frac{\rho}{K}}{\sqrt{2}\lambda}\right) + erf\left(\frac{q}{\sqrt{2}\lambda}\right)\right]$$

Noting that $sign(q), erf(q)$ are odd functions and $N(q)$ is and even function of q, we have

$$CCR = \frac{1}{2}\int_{0}^{+\infty} dq\, N(q)\left[erf\left(\frac{q\frac{\rho}{K}}{\sqrt{2}\lambda}\right) + erf\left(\frac{q}{\sqrt{2}\lambda}\right)\right]$$

$$= \frac{1}{2}\int_{0}^{+\infty} dq\, N(q) erf\left(\frac{q\frac{\rho}{K}}{\sqrt{2}\lambda}\right) + \frac{1}{2}\int_{0}^{+\infty} dq\, N(q) erf\left(\frac{q}{\sqrt{2}\lambda}\right)$$

$$= \frac{1}{2} - \frac{1}{\pi} arctan(\lambda) + \frac{1}{2} - \frac{1}{\pi} arctan\left(\frac{\lambda K}{\rho}\right)$$

$$= 1 - \frac{1}{\pi}(arctan(\lambda) + arctan\left(\frac{\lambda K}{\rho}\right)) = 1 - \frac{1}{\pi} arctan\left(\sqrt{2(\frac{K}{\rho} - 1)}\right) (11)$$

Note that the closed form of CCR in data sets with no effect size depend only on $ICC = \rho$ and K.

To complete the proof, since arctan is a monotonic increasing function for positive elements, we have to show that $\sqrt{2(\frac{K}{\rho}-1)}$ is increasing function of K and decreasing function of ρ which is already evident.

To validate our analytical results, we generated a series of simulations with totally 4, 8, or 16 subclasses and varied ICC. We classified these data sets with LDA and compared the CCRs with results of Equation 11. Figure Appendix B1 shows that results of the analytical solution and simulations are almost identical.

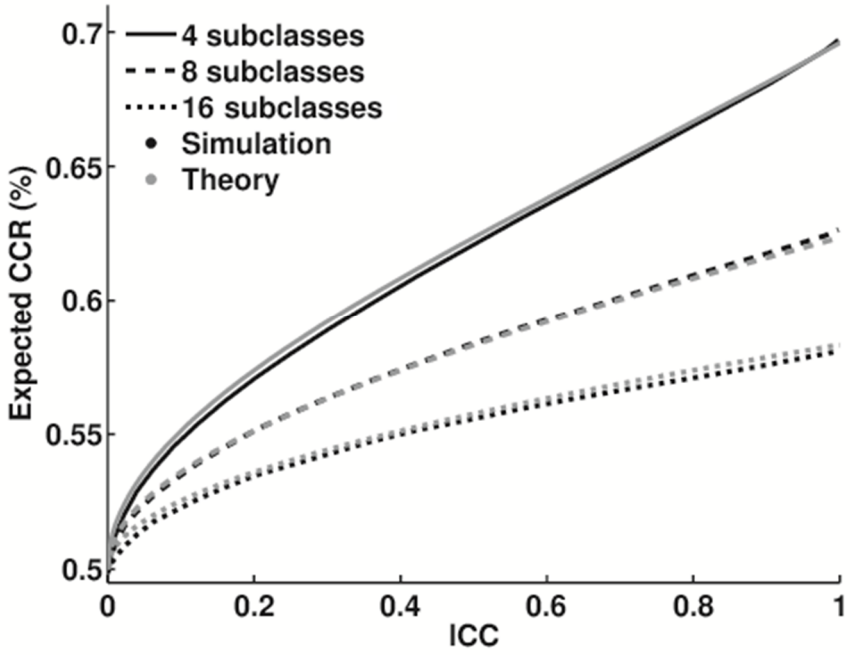


Figure appendix B1: Expected CCRs for data sets with nested subclasses when the size of main effect is zero using equation 11 (gray lines) and simulated (black lines). The figure confirms that the analytical solution and simulations produce very similar results.

References

Aarts, E., Verhage, M., Veenvliet, J.V., Dolan, C.V., van der Sluis, S. (2014) A solution to dependency: using multilevel analysis to accommodate nested data. Nature neuroscience, 17:491-6.

Alizadeh, S., Jamalabadi, H., Schonauer, M., Leibold, C., Gais, S. (2017) Decoding cognitive concepts from neuroimaging data using multivariate pattern analysis. NeuroImage, 159:449-458.

Anderson, M.J., Ter Braak, C.J.F. (2003) Permutation tests for multi-factorial analysis of variance. J Stat Comput Sim, 73:85-113.

Delorme, A., Makeig, S. (2004) EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. Journal of neuroscience methods, 134:9-21.

Efron, B.T., R. . (1993) An introduction to bootstrap. Chapman and Hall.

Fan, J.Q., Fan, Y.Y. (2008) High Dimensional Classification Using Features Annealed Independence Rules. Ann. Stat., 36:2605-2637.

Galbraith, S., Daniel, J.A., Vissel, B. (2010) A study of clustered data and approaches to its analysis. The Journal of neuroscience : the official journal of the Society for Neuroscience, 30:10601-8.

Gonzalez, L., Manly, B.F.J. (1998) Analysis of variance by randomization with small data sets. Environmetrics, 9:53-65.

Hastie, T., Tibshirani, R. (1996) Discriminant analysis by Gaussian mixtures. J Roy Stat Soc B Met, 58:155-176.

Haxby, J.V., Connolly, A.C., Guntupalli, J.S. (2014) Decoding Neural Representational Spaces Using Multivariate Pattern Analysis. Annu Rev Neurosci, 37:435-456.

Haynes, J.D. (2015) A Primer on Pattern-Based Approaches to fMRI: Principles, Pitfalls, and Perspectives. Neuron, 87:257-270.

Hohne, J., Bartz, D., Hebart, M.N., Muller, K.R., Blankertz, B. (2016) Analyzing neuroimaging data with subclasses: A shrinkage approach. NeuroImage, 124:740-751.

Jamalabadi, H., Alizadeh, S., Schonauer, M., Leibold, C., Gais, S. (2016) Classification based hypothesis testing in neuroscience: Below-chance level classification rates and overlooked statistical properties of linear parametric classifiers. Human brain mapping, 37:1842-55.

Lazic, S.E. (2010) The problem of pseudoreplication in neuroscientific studies: is it affecting your analysis? BMC neuroscience, 11:5.

Malone, P.S., Glezer, L.S., Kim, J., Jiang, X., Riesenhuber, M. (2016) Multivariate Pattern Analysis Reveals Category-Related Organization of Semantic Representations in Anterior Temporal Cortex. The Journal of neuroscience : the official journal of the Society for Neuroscience, 36:10089-96.

Manly, B.F. (2006) Randomization, bootstrap and Monte Carlo methods in biology. CRC Press.

Nichols, T.E., Holmes, A.P. (2002) Nonparametric permutation tests for functional neuroimaging: a primer with examples. Human brain mapping, 15:1-25.

Norman, K.A., Polyn, S.M., Detre, G.J., Haxby, J.V. (2006) Beyond mind-reading: multi-voxel pattern analysis of fMRI data. Trends in cognitive sciences, 10:424-30.

Stelzer, J., Chen, Y., Turner, R. (2013) Statistical inference and multiple testing correction in classification-based multi-voxel pattern analysis (MVPA): random permutations and cluster size control. NeuroImage, 65:69-82.

Todd, M.T., Nystrom, L.E., Cohen, J.D. (2013) Confounds in multivariate pattern analysis: Theory and rule representation case study. NeuroImage, 77:157-65.

Woolgar, A., Golland, P., Bode, S. (2014) Coping with confounds in multivoxel pattern analysis: what should we do about reaction time differences? A comment on Todd, Nystrom & Cohen 2013. NeuroImage, 98:506-12.

Zhu, M.L., Martinez, A.M. (2006) Subclass discriminant analysis. Ieee T Pattern Anal, 28:1274-1286.

Chapter 4:

Decoding material-specific memory reprocessing during sleep in humans

Monika Schönauer*, Sarah Alizadeh*, Hamidreza Jamalabadi, Annette Abraham, Annedore Pawlizki, & Steffen Gais

* These authors contributed equally to this work

Published in Nature Communications (May 2017)

M. Schonauer*, S. Alizadeh*, H. Jamalabadi, A. Abraham, A. Pawlizki, S. Gais (2017), "Decoding material-specific memory reprocessing during sleep in humans". *Nature Communications,* 8:15404. (*equal contribution)

Abstract

Neuronal learning activity is reactivated during sleep but the dynamics of this reactivation in humans are still poorly understood. Here we use multivariate pattern classification to decode electrical brain activity during sleep, and determine what type of images participants had viewed in a preceding learning session. We find significant patterns of learning-related processing during rapid eye movement (REM) and non-REM (NREM) sleep, which are generalizable across subjects. This processing occurs in a cyclic fashion during time windows congruous to critical periods of synaptic plasticity. Its spatial distribution over the scalp and relevant frequencies differ between NREM and REM sleep. Moreover, only the strength of reprocessing in slow-wave sleep influenced later memory performance, speaking for at least two distinct underlying mechanisms between these states. We thus show that memory reprocessing occurs in both NREM and REM sleep in humans, and that it pertains to different aspects of the consolidation process.

Introduction

Sleep helps us retain new memories [1,2]. A reactivation of newly encoded memory traces in the sleeping brain is thought to underlie this effect. Replay of learning-related neuronal firing patterns has been observed in single cell recordings of the hippocampus and neocortex in animals [3-6]. Importantly, this sleep-dependent activation of neurons has recently been shown to promote synaptic plasticity [7]. Reactivation of neuronal ensembles involved in motor learning is associated with changes in the task-related spiking behavior of these neurons in the rodent brain [8]. Furthermore, oscillation related to memory replay during sleep have been linked to greater memory strength and precision in rats [9]. The dynamics of this memory trace reactivation in humans, however, are still poorly understood. When memory content was associated with auditory or olfactory cues during learning, a re-exposure to these cues during sleep can improve later recall performance [10,11]. Moreover, activity on the level of brain areas suggests reactivation during sleep [12,13]. It is unclear whether this re-expression of learning related activity reflects the specific content of a previous learning task. Recent advances in multivariate pattern classification (MVPC) methods have made it possible to investigate covert cognitive processes in continuous brain activity. Using such methods on brain activity measured with fMRI, Horikawa et al. [14] have recently shown that it is possible to decode the content of visual imagery occurring at sleep onset. In the present study, we used MVPC to test whether the human sleep electroencephalogram (EEG) contains information about what has previously been learned, and thus indicates reprocessing of memory content.

In our experiment, participants learned pictures of either faces or houses before sleeping in the laboratory for a whole night. During this time, brain activity was recorded using high-density EEG. We then employed MVPC methods to detect information about the previously learned material in electrical brain activity

during sleep (Fig. 1, also see Materials and Methods). We investigated continuous sleep EEG instead of evoked activity, because we were specifically interested in spontaneous information processing in sleep. Cued reactivation, which has already been demonstrated in humans with functional MRI, shows that stimulus processing in sleep can lead to memory improvement. Previous studies, however, have not shown that such activity actually occurs spontaneously in humans. After demonstrating the existence of such an activity, we were also interested in the time course of memory reprocessing across the night and in sleep-stage specific activity. It has been discussed previously whether such reactivation occurs during NREM or REM sleep, and both have been implicated in memory reactivation and consolidation [12,13,15,16]. Furthermore, activity that is present only at specific times during the night indicates that the underlying process is related to discrete periods of reprocessing rather than prolonged ongoing activity.

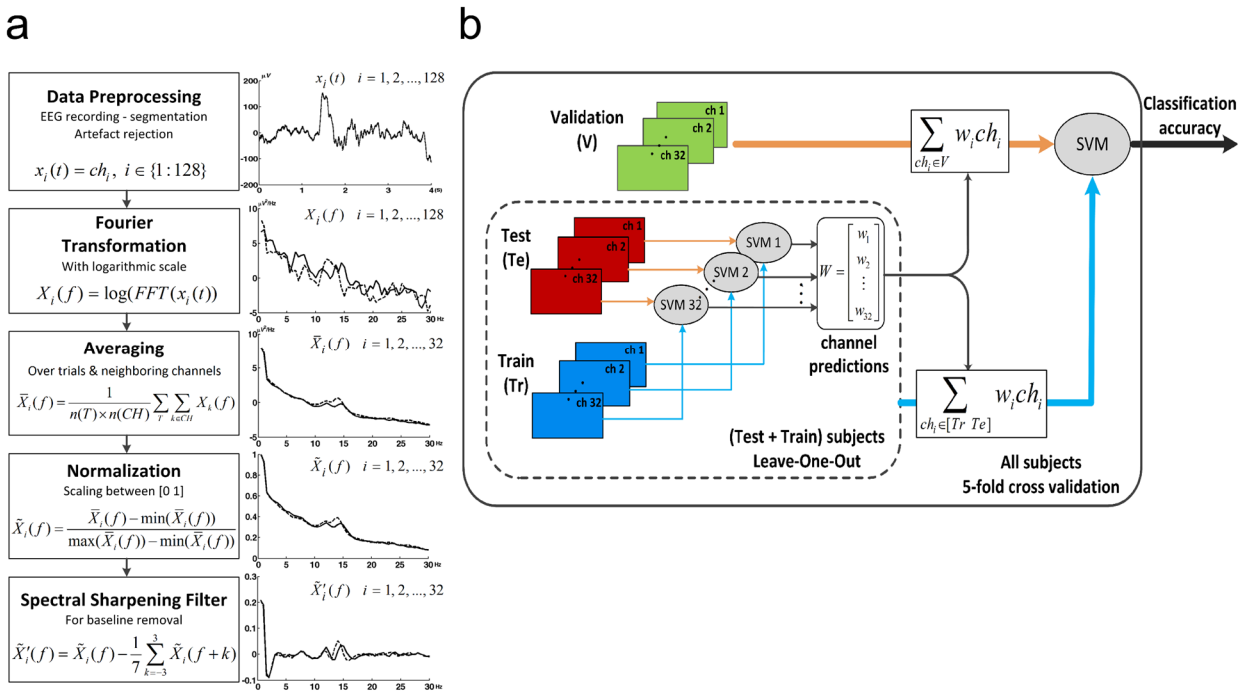


Figure 1: Data preprocessing and MVPC analysis. (a) After artefact rejection, data from the remaining 4-s trials of 128-channel sleep EEG data was frequency transformed. To reduce the dimensionality of the data and to increase the signal-to-noise ratio, spectra were averaged over trials and neighboring channels. Next, spectra of all channels were

normalized separately to make them comparable, and a spectral sharpening filter was applied to remove the baseline spectrum and enhance differences between neighboring frequency bins. (b) Training data was strictly separated from validation data in all MVPC analyses. Dimensionality of the data was further reduced in a two-step training procedure. Individual channel performance was determined using separate single-channel classifiers. An average of data from all channels weighted by their standalone performance was then used to train a classifier to distinguish between face and house stimulus conditions. Finally, classification was tested on independent validation data.

Results

Detecting memory reprocessing using MVPC

We tested whether MVPC can decode from the sleeping brain's activity what has been learned beforehand. Instead of looking for a single feature that can distinguish between conditions, MVPC methods take into account and compare the whole temporospatial pattern of activity. Given their multivariate nature, they are more suitable to deal with this kind of high-dimensional problem than is classical statistics, which usually relies on multiple univariate testing. Because EEG activity differs greatly between sleep stages and even more so between sleep and wakefulness, activity cannot be compared directly between these states. We therefore used between subject analyses to compare recordings from the same sleep state, i.e. the classifier was trained and tested on sleep data. If MVPC can determine from the sleep recording which type of visual stimulus a subject has learned before sleep, this implies that stimulus-specific reprocessing of the learned material occurs during sleep.

Our results show that human sleep EEG contains information about which kind of visual stimuli was learned before sleep (Fig. 2a). Classification accuracies for this distinction exceed classification rates expected from chance guessing of the classifier, as determined by randomization statistics, during two of the four 90-min segments (Fig. 2b). Thus, the sleep EEG reflects previous learning during

these intervals. Moreover, both NREM and REM sleep contain relevant information (Fig. 2a, b and c).

We used two different approaches to ensure that findings are significant and generalizable. First, we generated randomly labeled data, which, per se, cannot contain any information, and compared the performance of the classifier on these random data with its performance on the original observed data (see Supplementary Fig. 1). This test allows to determine the probability of an outcome by chance given that the data contain no actual information and thus provides exact significance values. Because this process, which repeats the whole analysis for each random iteration, is computationally intensive, we could complete only 1001 repetitions, which allows significance testing with a lower limit of precision of p=0.001. In the case of REM sleep of the 2nd 90-min sleep segment, none of these 1001 random iterations produced higher classification rates than the real data, thus allowing the conclusion of p<0.001.

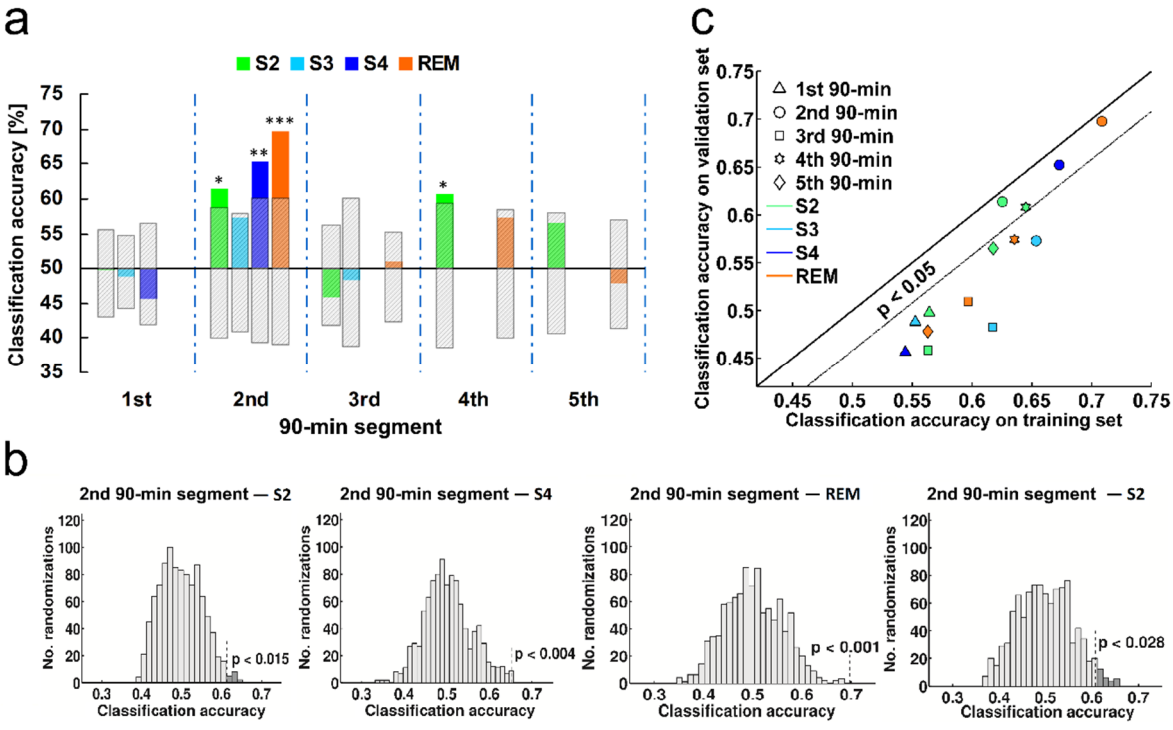


Figure 2: Classification results. (a) The content of a previous learning experience can be determined from sleep EEG during the second and fourth 90-min segment of the night. At these times, classification accuracy for all sleep stages is significant or approaches

significance. The hatched area shows the 95% confidence interval. Classification accuracies for S4 sleep as well as REM sleep in the second sleep segment remain significant after Bonferroni-Holm correction considering all tests (S4: p = 0.048, REM: p = 0.014). (b) Significance was assessed using permutation tests to ensure that classification rates are higher than can be expected from data sets with random labeling of the data, i.e. not containing any information. To estimate the displayed null-distribution from which exact significance levels of classification results can be determined, the MVPC analysis was repeated 1001 times on the actual data with randomly shuffled condition labels. Dark grey areas show those randomizations during which classification accuracy on randomly labeled data exceeded accuracy obtained on correctly labeled data. (c) If classification accuracies are similar between the training and validation sets, generalizable information could be extracted and the classifier was not overfitted on the training data set. This was the case for all analyses that were significant, i.e. for data from the second (circles) and fourth (stars) 90-min segments of the night. Here, patterns detected in one set of subjects during classifier training can be generalized to data from a new set of subjects. Data from the first (triangles) and third (squares) 90-min segments show low training accuracy low accuracy on validation data, indicating that the classifier could not extract information about previous learning content from these periods of the night.

The second approach to ensure generalizability was to compare classification accuracies of training and validation sets. If accuracy is higher during training than during validation testing, the classifier was overfitted to the training data set and uses random feature characteristics that allow separating classes only in the training data , which are not predictive for new data, and thus cannot be generalized. Ideally, classification rates for the validation data should resemble those for the training data. This shows that the classifier can extract meaningful information from the training set, and that the learned pattern can be generalized to new data. It can be seen in Fig. 2b that for data from the 1st (triangles) and 3rd (squares) 90-min sleep segment training accuracy was low (<0.625), but classification accuracy for the validation set was still worse. Thus, EEG from these periods does not seem to contain information pertaining to previous learning experience. On the other hand, EEG from the second (circles) and fourth (stars) 90-min sleep segment consistently shows higher training and

validation accuracies, and in some cases shows nearly perfect generalization between training and validation.

Relating reprocessing to behavioral memory performance

Participants showed good recognition performance in both the face and house learning conditions (see Supplementary Table 1). We did not observe forgetting across the night. This result is in line with other studies on declarative memory consolidation that have shown stable maintenance of memory performance over sleep but significant decline of memory performance after sleep-deprivation or daytime wakefulness [17,18]. Memory consolidation, i.e. the overnight change in performance, was positively correlated with time spent in sleep stage S4 (r_{64} = 0.254, p = 0.043; Supplementary Table 2), confirming that sleep was related to the consolidation of this task. We also tested the relation of memory consolidation with the strength of memory reprocessing, which was inferred from the classification probability estimates provided by the classifier. We find that memory reprocessing during SWS shows a positive relation with memory consolidation (r_{64} = 0.329, p = 0.008; Supplementary Table 3 and Fig. 3). This correlation remained significant after removing the three most influential values determined by leverage statistics (r_{61} = 0.28, p = 0.030). Memory reprocessing during sleep stage S2 and REM sleep were not related to memory performance (S2: r_{64} = 0.099, p = 0.436; REM: r_{56} = -0.199, p = 0.142). A regression model including strength of reprocessing in S2, SWS and REM sleep as predictors for memory consolidation found that only reprocessing during SWS correlated significantly with memory consolidation (β = 0.339, p = 0.020, explaining 9.7% of the variance), reprocessing in S2 and REM sleep was no significant predictor (S2: β = -0.064, p = 0.656, explaining 0.3% of the variance; REM: β = -0.112, p = 0.436, explaining 1% of the variance). Slopes differed significantly between SWS and REM sleep (strength of reprocessing × sleep stage interaction: p = 0.008), indicating that memory reprocessing in these sleep stages is differentially related to memory consolidation and could thus have different functions.

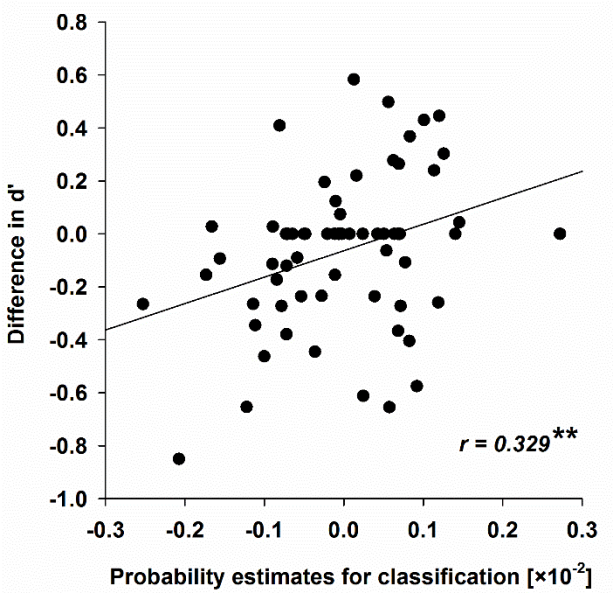


Figure 3: Correlation between classification probability estimates and overnight memory consolidation during SWS sleep. The more confident the classifier was in placing each subject in the correct condition, the more positive the change in memory performance during later recall. Spearman's rho is reported.

We then controlled whether general sleep features such as time spent in deep sleep could possibly account for an increase in both behavioral performance as well as classifiability of the data. Entering strength of reprocessing in SWS and time spent in this sleep stage in a regression model, we found that only strength of reprocessing in SWS was a significant predictor of memory consolidation and explained a larger part of the variance ($\beta = 0.335$, $p = 0.006$, explaining 11.2% of the variance), whereas duration of SWS was only marginally significant ($\beta = 0.214$, $p = 0.074$, explaining 5.2% of the variance). Strength of reprocessing in SWS was independent of time spent in that sleep stage ($r_{64} = -0.025$, $p = 0.423$) and the partial correlations support the view that strength of reprocessing in SWS and duration of SWS are independent predictors of overnight memory consolidation (partial correlation with strength of reprocessing during SWS controlling for the duration: $r_{64} = 0.342$, $p = 0.006$; partial correlation with duration of SWS controlling for strength of

reprocessing: $r_{64} = 0.226$, $p = 0.074$). Analogous regression analyses for strength of reprocessing and time spent in S2 and REM sleep yielded no significant results, as could be expected from the general lack of association with overnight memory consolidation (all $p > 0.143$).

While the proportion of variance in overnight memory consolidation that is explained by memory reprocessing during SWS is low in absolute terms, it should be noted that factors such as alertness or individual differences can introduce considerable variance in memory performance. Classifier performance similarly provides a measure of reprocessing strength that is affected by many sources of between-subject variance as it is estimated based on other participants' sleep EEG characteristics. Despite these difficulties, we demonstrate that memory reprocessing during SWS is significantly related to overnight memory retention, suggesting a robust underlying effect.

Temporal dynamics of reprocessing

We detected processing of learning material during sleep in the second and fourth 90-min segment of the night (Fig. 2). To investigate this pattern on a more fine-grained scale, we split the night into smaller intervals and analyzed the time course of classification accuracy across the night with a resolution of 4.5 min, using the same procedure as above. Again, we find two periods of the night during which brain processing seems to be more strongly related to previous learning, congruent with the time windows reported above. During other periods, no learning-related information was detected (Fig. 4).

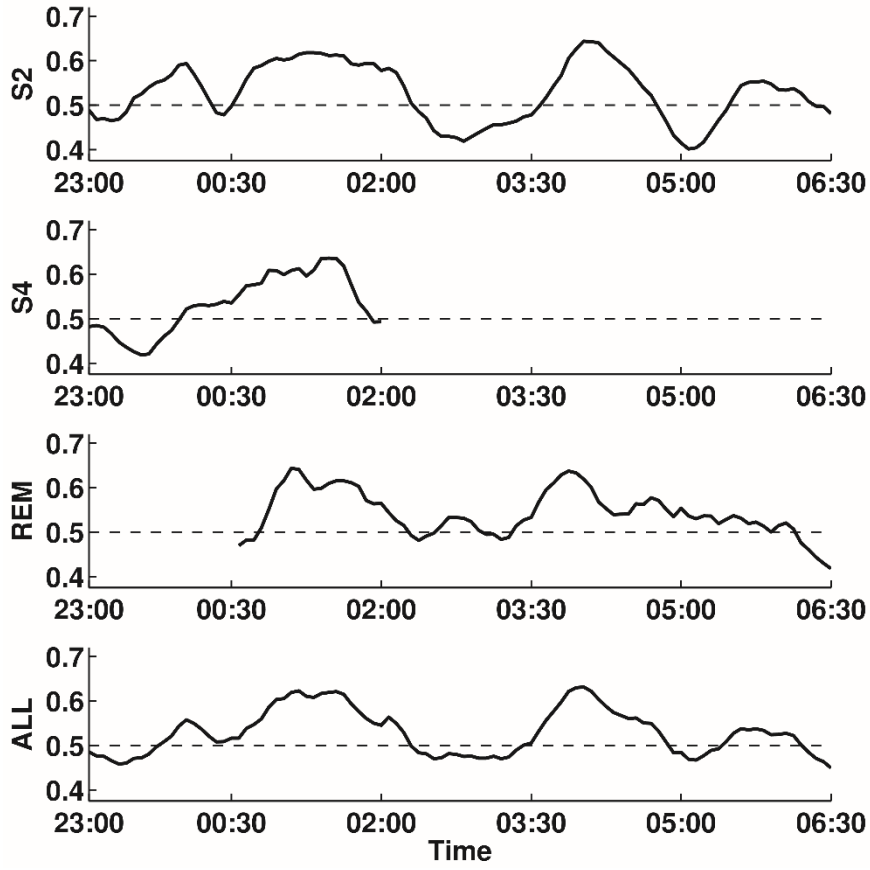


Figure 4: Time course of classification accuracy across the night. Separate analyses were performed for sleep stages S2, S4, and REM sleep. Classification performance follows an oscillatory pattern and peaks around three and six hours after learning in all stages. Timing therefore is more relevant to when memory reprocessing occurs than sleep stage

Spatial characteristics of reprocessing and frequency contributions

Brain activity in REM and NREM sleep is not alike. It is thus reasonable to assume that also information processing in these states will take different forms. To investigate this, the relative contribution of each frequency band to classification can be assessed in terms of classification weights and compared between sleep stages (Fig. 5). Our results show that the frequencies that are important for identifying previous learning content differ between sleep stages. Activity in the range of sleep spindles (11-16 Hz) can distinguish previous

learning conditions only in NREM sleep (Fig. 5a). Theta-band activity (4-8 Hz), on the other hand, has higher discriminative power in REM sleep. Slow frequencies below 4 Hz were informative in both NREM and REM sleep, but their topographies differ (Fig. 5b). Although there is some resemblance between the feature weight plots and power spectra of sleep, it has to be noted that the feature weights do not follow the typical 1/f logarithmic decrease of EEG power spectra, but remain essentially constant after a linear decrease in delta frequencies. Moreover, actual classifier input was not the power spectra but the preprocessed data seen in the lower panel of Fig. 1a.

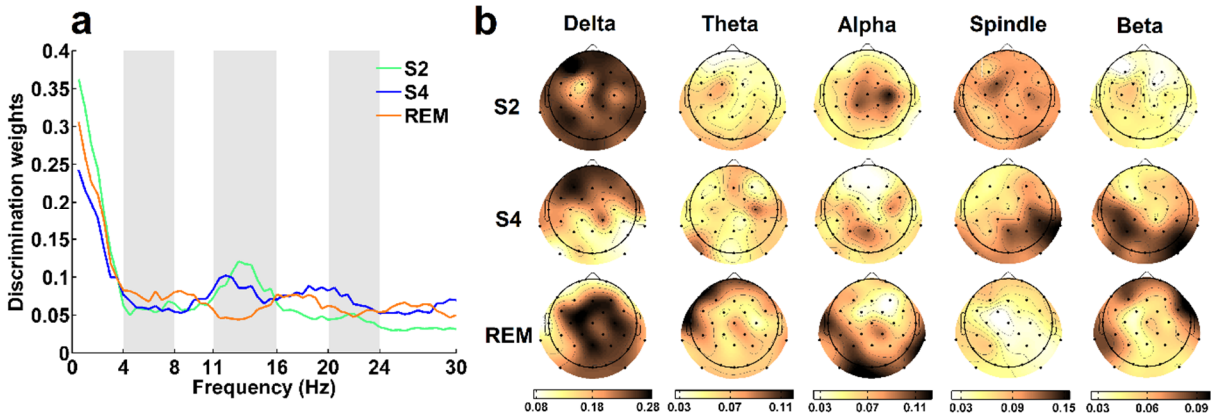


Figure 5: Frequency contributions to memory reprocessing in NREM and REM sleep. (a) Discrimination weights show that in NREM sleep stages S2 and S4 spindle activity in the frequency range between 11 and 16 Hz is predictive for learning content. In REM sleep, theta, alpha, and higher beta frequencies contributed more to correct classification. Slow frequencies below 4 Hz were informative in all sleep stages. (b) The topography of predictive channels clearly differs between NREM and REM sleep. In NREM sleep stage S2, mainly delta and spindle frequencies contributed to correct classification. Similarly, frontal delta power and right parieto-temporal spindle activity were most informative for classification during NREM sleep stage S4, together with posterior higher frequency activity. REM sleep shows a more complex pattern. Here, slow oscillations of central electrodes, frontal and temporal theta as well as occipital alpha contributed most to discrimination between learning conditions.

Discussion

We show that memory processing of a single memory task occurs during all stages of sleep. Reprocessing in REM and NREM sleep, however, has different effects on later memory performance. Although a large number of studies in rodents have observed the occurrence of spontaneous memory reactivation during NREM sleep [4-6,19,20], linking this reactivation with improvements in behavioral performance has remained a challenge. Contrary to rodents, task difficulty and training time can be easily adjusted in studies on humans, giving greater power to analyses on behavioral effects. It has early been suggested that memory reactivation during sleep has functional significance for strengthening new memories [21]. Indirect evidence for this assumption has accumulated over the last years [10,11,22-24]. A recent study in rats found that sleep-dependent reactivation of neurons involved in a simple motor learning task is associated with changes in the task-related spiking behavior of the same neurons [8]. In this way, reactivation may be related to later improvements in performance. We now show that content-related reprocessing of declarative learning material during NREM sleep influences later memory strength in humans. Conversely, memory reprocessing during REM sleep does not show this graded relation with overnight memory retention.

A number of animal studies detected reactivation of learning activity also in REM sleep [25,26], yet empirical evidence for this has remained ambiguous. We find that memory content is reprocessed during both NREM and REM sleep. The differential significance of memory reprocessing for behavioral performance between these states points towards at least two different mechanisms underlying memory reprocessing during sleep.

Already early on, it has been suggested that memory is formed in a two-stage process. Labile memory traces are formed during exploratory behavior, when theta power is high. Later, during rest or sleep, long-lasting traces are formed [9,21]. Similarly, it has been proposed that during sleep, slow-wave-related NREM

activity and theta-related REM activity have complementary, mutually dependent functions [27]. We find that reprocessing occurs in both NREM and REM sleep. Interestingly, we can demonstrate a correlation between reprocessing and later memory performance only for NREM sleep. This supports the view that reprocessing during REM sleep and NREM sleep serves distinct functions. Our finding is in line with previous studies, which show no behavioral benefit of reactivating memories by cueing during REM sleep [10]. Interestingly, memory replay observed during REM sleep has also been shown to have different characteristics than that in NREM sleep, including a smaller time-compression factor, which is less suited for the induction of long-term potentiation [20,25].

A number of recent studies stress the importance of either light NREM sleep, SWS or REM sleep for memory consolidation, respectively [2,27,28]. Based on these findings, theoretical accounts have suggested that NREM and REM sleep may interact during memory consolidation, emphasizing different aspects of this process. The sequential hypothesis of sleep stresses that different sleep stages have to occur in succession to effectively influence memory function. It assigns specific and substantially different, but interdependent roles to NREM and REM sleep regarding the processing of memories [29]. Other accounts suggest the different processes contributing to memory processing during NREM and REM sleep are separate and independent. Thus, the function of NREM and REM sleep in consolidation is assumed to pertain to different aspects or forms of memory [30]. We find that relevant activity occurs in close temporal proximity over different stages, and that a single memory task triggers learning-related activity in both NREM and REM sleep EEG. It therefore seems possible that both sleep stages cooperate in the processing of memories. The differential function of NREM and REM sleep stages is still controversial [7,16,31]. One recent hypothesis is that cortical activity and long-range connectivity differs between sleep stages, allowing local memory reactivation and potentiation in SWS, and network-wide information integration in REM sleep [32,33]. This view fits with our findings.

Our data indicate that memory processing in sleep is cyclic in nature and its occurrence might depend more strongly on timing than on the stage of sleep. Instead of occurring in SWS throughout the whole night, reprocessing was detected in S2, S4 as well as REM sleep in the 2nd 90-min period, but not in the 1st or 3rd. Whether this consolidation window depends on time after learning, time after sleep onset, or circadian rhythm cannot be determined in the present study, because these were not varied independently.

Because reprocessing peaks during distinct times of the night, it is unlikely that the detected activity simply reflects ongoing reverberation of learning-related activity or selective fatigue in the involved brain areas. Instead, it must reveal a process that is selectively initiated at specific points during sleep. The finding that reprocessing is strongest around three and around six hours after learning fits well with experiments that found critical periods during memory consolidation, during which memory is particularly sensitive to disruption [34]. Thus, inhibiting protein synthesis 15 min and 3 h after learning, but not 1 h after learning impairs hippocampal one-trial avoidance learning [35]. Similarly, in drosophila, different behavioral memories and corresponding neuronal traces develop during different time windows over several hours after conditioning [36], a process that has been linked to systems memory consolidation in humans [37].

Moreover, our finding of discrete periods for memory reprocessing is reminiscent of previously reported 'sleep windows', i.e. times during which sleep has to occur after learning to strengthen memory [38,39]. Along the same lines, Stickgold et al. have found that, for consolidation of a visual discrimination task, mainly duration of SWS and REM sleep in the first and the last quartile of the night, respectively, are most critical parts of the night [40]. Although that task presumably does not rely on hippocampal memory reactivation and might therefore follow a different temporal trajectory, the similarities suggest the possibility of a common mechanism. Further behavioral, electrophysiological and molecular investigations are required to elucidate this underlying mechanism. Moreover, it has still to be ascertained whether the other periods of

the night have memory-related functions that cannot be detected by our method.

Because the amount of signal related to memory reprocessing across the whole night is very small compared to the unrelated noise, we used MVPA, which is a very sensitive method to detect systematic differences between large sets of data. However, multivariate approaches are not better suited to supply information about univariate hypotheses than classical tests. Using feature weights and individual channel accuracies (Fig. 5) can to some extent illustrate the features that are carrying relevant information. However, these features must be seen within the entire pattern. The following discussion of individual physiologic features should therefore be seen as a starting point for studies focusing on a smaller feature search space.

When looking at the frequencies contributing to correct classification, we find that spindle activity during NREM sleep contributes to the distinction of previous learning conditions. This is consistent with the fact that sleep spindles increases after learning [41] and correlate with performance [42]. Parietal sleep spindles accompany task specific reactivation seen in fMRI [43]. Moreover, frontal slow-waves, as they appear in our analysis for NREM sleep, have previously been shown to correlate with performance gains observed after memory reactivation induced by cueing during sleep [44].

Apart from confirming that learning-related information resides in frequency bands that have previously been implicated in memory consolidation, such as NREM spindles and slow oscillations, our results hint at promising objects for future study. We suggest that particular attention should be given to the function of REM sleep theta. Frontal theta power increases during successful memory encoding and retrieval, and theta is also involved in memory processing during wakefulness, such as in controlling, maintaining and storing memory content [45]. Theta has been linked to memory and sleep for a long time, but has only recently received renewed attention [16,46]. For instance, theta band activity during sleep has been shown to support formation of imprinting memory in chicks [47]. In

humans, another recent study found increased frontal theta power after presentation of cues related to a verbal learning task during sleep [44,48]. Moreover, frontal theta in REM sleep is predictive of successful dream recall [49]. These findings stress the active role of theta activity in memory reprocessing during sleep.

It is difficult to demonstrate reactivation directly in humans. Electroencephalographic activity during sleep differs greatly from that during wakefulness in both the time domain and the frequency domain. Thus, amplitude changes over time, as well as power spectral density cannot be compared between these states. This is owing to different modes of generation and transmission of electrical activity during sleep [50,51]. Previous data have shown that reactivation can be both time-compressed as well as changing in location (e.g. neocortical replay following hippocampal activity) [19,52]. Markers reflecting reactivation of neuronal firing patterns observed during learning can thus be altered by a large number of operations, which renders the search space virtually infinite. Because this makes wake-to-sleep classification problematic, and a within-subject design would have to rely on between-session classification that is confounded by various session differences (e.g. recording artefacts), we instead opted for a between-subject classification approach. This allowed us to detect information pertaining to a previous learning experience in data recorded in the same state of consciousness. Previous attempts to observe memory reactivation during off-line periods succeeded in showing memory reprocessing during wakefulness, but not during sleep [53-55]. Using an approach that trains and tests the classifier in the same state of consciousness made it possible for us to observe material-specific memory reprocessing during sleep and study its dynamics and relation to later behavioral performance.

We used multivariate pattern classification to decode the content of a previous learning experience from electrical brain activity during sleep. By linking brain activity during sleep with the content of previous learning, our findings bridge studies from multicell recordings in animals, which show learning-related

reactivation, to human imaging studies, which show reactivation of brain regions during sleep. Pattern classification methods are powerful tools for investigating the covert mechanisms that link electrical brain activity and behavior, and can thus contribute to our understanding of these complexities.

Materials and Methods

Subjects. In this study, we recorded EEG data from 32 healthy subjects with no history of neurological or psychiatric disorders. All participants were students, between 18 and 30 years old, native German speakers and non-smokers. They were right handed as measured by Edinburgh Handedness Inventory-test [56]. Chronotype was assessed via the Munich Chronotype Questionnaire [57] and experimental timing was adjusted to participants' usual sleep times (sleep midpoint 03:56h ± 01:33h [mean ± SD]). Subjects were regular sleepers with a habitual sleep duration of 6-9 h. They did not report any chronic or acute sleep-related problems in an initial interview. Moreover, they did no shift work and did not change time zones in the six weeks leading up to the experiment. Participants were told to refrain from drinking alcohol, coffee and tea on the days of the experiment and did not take any drugs that affect the central nervous system. All experimental procedures were approved by the local ethics committee (Department of Psychology, Ludwig-Maximilians-Universität München). Informed consent was obtained from all subjects.

Experimental Design. Participants slept in our laboratory on three different nights. The first of these served as an adaption night, to accustom subjects to the environment and to sleeping under the experimental conditions (e.g. wearing an EEG cap). In the subsequent two experimental nights, subjects completed an intensive image learning task, during which they studied pictures of either faces or houses. For an exemplary subject, learning took place from 8:30 p.m. to 10 p.m. after the EEG electrodes had been attached, and memory was tested immediately afterwards. The subject then went to bed at 11 p.m. for an 8-h sleep

period. Memory was tested once more in the morning. The times of the experiment were advanced or delayed such that time to bed corresponded to the individual habitual bedtime of the participants. All subjects participated in two experimental nights, each time learning only one type of images, in a counterbalanced fashion. The two nights were spaced at least 5 days apart. Sleepiness was tested with a visual analog scale in the evening and after sleep in the morning (Supplementary Table 4).

Learning Task. Subjects studied a set of 100 images of faces or houses in 30 repetitions. Pictures were shown in random order and individual images were always presented in one of the four quadrants of the screen. Participants had to remember the individual pictures and learn to associate the images with the quadrant in which it was presented. Participants were tested once immediately after learning and again in the next morning after a full night of sleep. During both immediate and delayed testing, 100 learned images were presented together with a set of 50 new images in random order. Participants first had to indicate via keypress whether they had seen the image before (with left hand on main keyboard: 1-sure, 2-probably, 3-probably not, 4-surely not. Responses 1 and 2 were counted as a "yes" response, responses 3 and 4 were counted as a "no" response). For "yes"-responses, also the quadrant in which the image had been presented was probed (with right hand on numerical pad: 1-lower left, 3-lower right, 7-upper left, 9-upper right). Image material was derived from two different sources: 300 pictures of houses were taken from German online real estate sites, 300 pictures of neutral faces were taken from Minear & Park [58].

This task was chosen because it is a declarative task that is supposed to involve the hippocampus, and sleep-related reactivation has mainly been shown in the hippocampus [10,19]. Face and house processing are clearly different in event-related EEG potentials and fMRI [59]. Face processing activates the mid-fusiform gyrus (fusiform face area) and the occipital face area in the occipito-temporal cortex as well as other temporal areas [60], whereas processing of houses activates the parahippocampal place area and the lateral occipital gyrus [61,62].

EEG Recording. Sleep EEG was recorded using an active 128 channel Ag/AgCl-electrode system (ActiCap, Brain products, Gilching, Germany) with 1 kHz sampling frequency and a high-pass filter of 0.1 Hz. Electrodes were positioned according to the extended international 10–20 electrode system. For sleep scoring, recordings were split into 30-s epochs and sleep stages were determined on electrodes C3/C4 according to standard rules by two independent raters [63]. Average sleep durations are reported in Supplementary Table 5.

Methodological Considerations. One of the challenges in sleep research is the difficulty of recording large sample sizes and the large amount of data that is recorded. The goal of classical analyses, which use multiple univariate comparisons (e.g. classical fMRI analysis), is to find single features that are strong enough independently to distinguish between conditions. Such features are unlikely to exist in high-density all-night EEG recordings, which thus present a problem better addressed by a multivariate approach. In multivariate analyses, it is of interest whether the overall pattern of data contains information that is relevant to distinguish conditions. A prominent method that can deal with large numbers of data dimensions is MVPC. However, high dimensional, low sample size data, like EEG recordings, pose specific problems for classical statistical testing as well as for MVPC [64,65]. For this kind of data, it is important to minimize the number of features. If the signal across features is highly correlated, as in EEG data, this can be achieved by averaging, which reduces dimensionality of the data and at the same time increases signal-to-noise ratio. We developed a two-step procedure that uses spatial averaging and a channel-based weighted average to improve classifiability of our data (Fig. 1). These steps are described in detail in the sections Data Preparation and Multivariate Pattern Classification (MVPC) below.

Data Preparation. For artefact rejection and further analysis, EEG data was split into 4-s trials. Artefact rejection was done in a semiautomatic process using custom MATLAB scripts. Based on the distributions of different parameters of

the raw data and power spectrum, rejection thresholds were chosen for each recording individually to make sure that only a minimal number of artefacts remained in the data. We tested for disconnected electrodes (outliers in overall spectral power), sudden jumps of the signal (outliers in amplitude changes) and muscle artefacts (outliers in spectral power between 110 and 140 Hz). Outlier thresholds were automatically suggested based on the variance of the data and manually confirmed upon visual inspection of parameter distributions and of the raw data. Trials containing artefacts were removed from the data set, channels that contained too many trials with artefacts were removed entirely and interpolated using routines provided by EEGLAB [66]. Whether individual epochs or channels were to be removed was determined automatically so that data loss was kept minimal. Artefact-free trials were then transformed into the frequency domain using Fourier transformation. To obtain smooth spectra, Welch's method was used for this, averaging over 10 Hamming windows of 2-s length with 95% overlap, resulting in a final data resolution of 0.5 Hz. Data was used up to a maximum frequency of 30 Hz.

The subsequent steps for data preparation were implemented to 1) increase signal-to-noise ratio, 2) reduce dimensionality of the data, and 3) adapt the signal for between-subject classification. First, we averaged power spectra across electrodes within a radius of approximately 3 cm around the 32 evenly spread locations of the extended 10-20-system to decrease the number of redundant features and increase signal-to-noise ratio as well as spatial similarity between subjects. We then separately averaged over all artefact-free trials available for each 90-min segment and sleep stage, to obtain a reliable estimate of spectral properties. This also ensures that an equal number of epochs per subject enters analysis, which is important for classification to remain unbiased. To remove amplitude differences between channels, which are caused by the distance of each channel to the reference electrode, spectra of all channels were separately normalized between zero and one. This also removed between-subject variability in general spectral power.

Because baseline EEG power spectra are highly similar and differences between conditions can be expected to be of smaller magnitude, these differences need to be enhanced within the spectra. We thus applied a spectral sharpening filter, which removes the baseline spectrum and emphasizes differences between neighboring frequencies in a final preparation step. To achieve this, we subtracted a moving average of six neighboring frequency bins (window size: 3 Hz) from the signal. This accentuates the smaller differences in power between frequencies within the spectrum. This is a valid procedure because neighboring data points in the power spectrum represent neighboring frequencies from the same signal and are therefore strongly correlated.

Subjects were only included in the analysis if they had at least 40 artefact-free trials within the respective sleep stage and segment (i.e. 160 s of data). Only segments and stages with at least 11 subjects were analyzed. The number of subjects and trials available for each 90-min segment and sleep stage can be found in Supplementary Table 6. As can be seen from that table, the amount of data available was unrelated to classifier performance.

Multivariate Pattern Classification (MVPC). In the present study, we tested whether electrical brain activity during sleep holds information about the content of previously learned visual stimuli. Instead of the typically used multiple univariate tests, we employed a multivariate classification approach, which can detect information contained in the overall pattern of brain activity, but is not distinguishable from single features.

Sleep EEG recordings from 64 nights (32 subjects, two conditions each) were analyzed using a classification algorithm developed on the basis of linear support vector machines (SVM). The aim was to detect material-specific information in the data. Please note that whereas the experiment followed a within-subject design, classification was done between subjects, with both nights of each participant (face and house conditions) assigned either to the training, test, or validation set. All analyses were done with the Matlab implementation of libsvm 3.1 (http://www.csie.ntu.edu.tw/~cjlin/libsvm). EEG

recordings pose problems typical of high dimensional, low sample size data (potential feature space of 128 channels times 60 frequency bins). We thus preprocessed the data to reduce the number of features and increase signal-to-noise ratio (see Fig. 1 and Data Preparation), averaging over neighboring channels to lower the number of channels to 32. To further enhance relevant features, we used a stepwise procedure for classification, which first regarded every channel as an independent classifier and then combined outcomes of this first step for the final analysis.

We split data into independent training and validation sets. In a first step, one linear SVM was trained for each of the 32 averaged EEG channels on all but one subject of the training set to see how much each channel contributes to distinguishing the content of learning conditions ('face' learning or 'house' learning). This channel-based classification was cross-validated in a leave-one-out procedure on each subject, and the obtained classification accuracies were averaged over all cross-validation runs. In the second step, this average classification accuracy from each channel was used as a weight to obtain a weighted average of the 32 channels. The main SVM was then trained on this weighted training set and classification accuracy tested on the independent validation set. The main reason for weighted averaging of channels was to reduce feature space dimensionality, because feature weights cannot be reliably determined if sample size is much smaller than the number of features [67]. Apart from this, weighted averaging can amplify relevant information in the data. This two-step classification process was cross-validated on independent data using 280 repetitions of a 5-fold procedure, which covers the whole data set with five independent validation sets.

We used permutation tests to assess significance. These tests sample the distribution of the null hypothesis by random shuffling of the original data, which is repeated a large number of times. To obtain the correct null-distribution for our data, we randomly shuffled condition labels, i.e. the two conditions of each subject were randomly labeled as 'face'/'house' or as

'house'/'face', effectively removing all relevant data pertaining to the effect of interest, while keeping other dependencies in the data constant. We then calculated classification accuracies for the randomly labeled data to estimate the random distribution. This was repeated 1001 times. Significance was calculated by determining the percentage of times that classification on randomly labeled data produced accuracies that were equal to or higher than the classification accuracy obtained from the actual data. If randomly labeled data did not result in a classification accuracy equal to or higher than the actual data, then the p value was determined by the number of random repetitions that were calculated (see Supplementary Fig. 1).

To assess whether reprocessing occurs uniformly across time, we split the night, starting from time to bed, into five 90-min segments, which are likely to include a whole sequence of sleep stages (S2, S3, S4, and REM sleep; see Supplementary Table 5 for details of sleep stage distribution). In this first analysis, we classified separately for all segments and sleep stages to assess the temporal dynamics of memory reprocessing. To determine a more fine-grained time course of classification accuracy, we moved a sliding window with a width of 22.5 min in steps of 4.5 min across the night. We then estimated classification accuracy within each window using the same two-step classification procedure as before. Analysis was done separately for each sleep stage and the same inclusion criteria were applied as in the main analysis.

To assess which features of the sleep EEG are particularly predictive, we analyzed classification weights. To assess which features of the sleep EEG are particularly predictive, we analyzed classification weights. The absolute value of the weights are informative about how much each frequency band and channel contributes to successful distinction. We averaged the classification weights over all repetitions of the training procedure, resulting in an averaged 32 (channels) × 60 (frequency bins) weight matrix. To examine frequency contributions to memory reprocessing, we further averaged the absolute values of these weights over all channels (see Fig. 5a). The topography of predictive

channels (see Fig. 5b) was obtained by averaging absolute values of classification weights for each channel over different frequency bands (delta: 0.5-3.5 Hz, theta: 4-7.5 Hz, alpha: 8-10.5 Hz, spindle: 11-15.5 Hz, beta: 16-30 Hz). We chose to analyze classification weights for frequencies obtained in the inner train-test loop (Fig. 1) because they can give additional information on the topography of predictive channels. These frequency weights are confirmed by weights from the outer validation loop (Fig. 1). Frequency contributions to classification assessed from both loops show the same pattern (see Supplementary Fig. 2).

Behavioral Performance. For assessment of memory performance, we calculated the memory sensitivity index d' as the difference of z-values between correctly recognized old items vs. falsely recognized new items (z[hits] – z[false alarms]). Performance pre and post sleep, as well as memory consolidation across the nights is reported in Supplementary Table 1. We correlated overnight memory consolidation with time spent in different sleep stages (see Supplementary Table 2). To examine whether memory reprocessing during sleep is associated with better memory performance, we correlated the probability estimates for classification given by the classifier with overnight memory consolidation measured as the difference between post sleep and pre sleep d' values. No such correlation was found for encoding or retrieval performance per se (see Supplementary Table 3). For each subject, results of all 280 repetitions of the 5-fold cross-validation procedure were averaged. We conducted this analysis separately for different sleep stages. All correlations report Spearman's rho.

Data availability

All data and codes are available from the corresponding authors upon request.

References

1 Walker, M. P. & Stickgold, R. Sleep, memory, and plasticity. *Annu. Rev. Psychol.* **57**, 139-166, doi:10.1146/annurev.psych.56.091103.070307 (2006).

2 Rasch, B. & Born, J. About sleep's role in memory. *Physiol. Rev.* **93**, 681-766, doi:10.1152/physrev.00032.2012 (2013).

3 Ribeiro, S. *et al.* Long-lasting novelty-induced neuronal reverberation during slow-save sleep in multiple forebrain areas. *PLoS Biol.* **2**, E24 (2004).

4 Lansink, C. S., Goltstein, P. M., Lankelma, J. V., McNaughton, B. L. & Pennartz, C. M. Hippocampus leads ventral striatum in replay of place-reward information. *PLoS Biol.* **7**, e1000173, doi:10.1371/journal.pbio.1000173 (2009).

5 Wilson, M. A. & McNaughton, B. L. Reactivation of hippocampal ensemble memories during sleep. *Science* **265**, 676-679 (1994).

6 Peyrache, A., Khamassi, M., Benchenane, K., Wiener, S. I. & Battaglia, F. P. Replay of rule-learning related neural patterns in the prefrontal cortex during sleep. *Nat. Neurosci.* **12**, 919-926, doi:10.1038/nn.2337 (2009).

7 Yang, G. *et al.* Sleep promotes branch-specific formation of dendritic spines after learning. *Science* **344**, 1173-1178, doi:10.1126/science.1249098 (2014).

8 Ramanathan, D. S., Gulati, T. & Ganguly, K. Sleep-Dependent Reactivation of Ensembles in Motor Cortex Promotes Skill Consolidation. *PLoS Biol.* **13**, e1002263, doi:10.1371/journal.pbio.1002263 (2015).

9 Girardeau, G., Benchenane, K., Wiener, S. I., Buzsaki, G. & Zugaro, M. B. Selective suppression of hippocampal ripples impairs spatial memory. *Nat. Neurosci.* **12**, 1222-1223, doi:10.1038/nn.2384 (2009).

10 Rasch, B., Büchel, C., Gais, S. & Born, J. Odor cues during slow-wave sleep prompt declarative memory consolidation. *Science* **315**, 1426-1429, doi:10.1126/science.1138581 (2007).

11 Rudoy, J. D., Voss, J. L., Westerberg, C. E. & Paller, K. A. Strengthening individual memories by reactivating them during sleep. *Science* **326**, 1079, doi:10.1126/science.1179013 (2009).

12 Peigneux, P. *et al.* Learned material content and acquisition level modulate cerebral reactivation during posttraining rapid-eye-movements sleep. *Neuroimage.* **20**, 125-134 (2003).

13 Maquet, P. *et al.* Experience-dependent changes in cerebral activation during human REM sleep. *Nat. Neurosci.* **3**, 831-836, doi:10.1038/77744 (2000).

14 Horikawa, T., Tamaki, M., Miyawaki, Y. & Kamitani, Y. Neural decoding of visual imagery during sleep. *Science* **340**, 639-642, doi:10.1126/science.1234330 (2013).

15 Marshall, L., Helgadottir, H., Molle, M. & Born, J. Boosting slow oscillations during sleep potentiates memory. *Nature* **444**, 610-613 (2006).

16 Grosmark, A. D., Mizuseki, K., Pastalkova, E., Diba, K. & Buzsaki, G. REM sleep reorganizes hippocampal excitability. *Neuron* **75**, 1001-1007, doi:10.1016/j.neuron.2012.08.015 (2012).

17 Himmer, L., Müller, E., Gais, S. & Schönauer, M. Sleep-mediated memory consolidation depends on the level of integration at encoding. *Neurobiol. Learn. Mem.* **137**, 101-106, doi:10.1016/j.nlm.2016.11.019 (2017).

18 Gais, S., Lucas, B. & Born, J. Sleep after learning aids memory recall. *Learn. Mem.* **13**, 259-262 (2006).

19 Ji, D. & Wilson, M. A. Coordinated memory replay in the visual cortex and hippocampus during sleep. *Nat. Neurosci.* **10**, 100-107, doi:10.1038/nn1825 (2007).

20 Euston, D. R., Tatsuno, M. & McNaughton, B. L. Fast-forward playback of recent memory sequences in prefrontal cortex during sleep. *Science* **318**, 1147-1150, doi:10.1126/science.1148979 (2007).

21 Buzsáki, G. Two-stage model of memory trace formation: a role for "noisy" brain states. *Neuroscience* **31**, 551-570 (1989).

22 Peigneux, P. *et al.* Are spatial memories strengthened in the human hippocampus during slow wave sleep? *Neuron* **44**, 535-545 (2004).

23 Dewald, J. F., Meijer, A. M., Oort, F. J., Kerkhof, G. A. & Bogels, S. M. The influence of sleep quality, sleep duration and sleepiness on school performance in children and adolescents: A meta-analytic review. *Sleep Med Rev* **14**, 179-189, doi:10.1016/j.smrv.2009.10.004 (2010).

24 Schönauer, M., Geisler, T. & Gais, S. Strengthening procedural memories by reactivation in sleep. *J. Cogn. Neurosci.* **26**, 143-153, doi:10.1162/jocn_a_00471 (2014).

25 Louie, K. & Wilson, M. A. Temporally structured replay of awake hippocampal ensemble activity during rapid eye movement sleep. *Neuron* **29**, 145-156 (2001).

26 Pavlides, C. & Winson, J. Influences of hippocampal place cell firing in the awake state on the activity of these cells during subsequent sleep episodes. *J. Neurosci.* **9**, 2907-2918 (1989).

27 Giuditta, A. *et al.* The sequential hypothesis of the function of sleep. *Behav. Brain Res.* **69**, 157-166 (1995).

28 Genzel, L., Kroes, M. C., Dresler, M. & Battaglia, F. P. Light sleep versus slow wave sleep in memory consolidation: a question of global versus local processes? *Trends Neurosci.* **37**, 10-19, doi:10.1016/j.tins.2013.10.002 (2014).

29 Ambrosini, M. V. & Giuditta, A. Learning and sleep: the sequential hypothesis. *Sleep Medicine Reviews* **5**, 477-490 (2001).

30 Ackermann, S. & Rasch, B. Differential effects of non-REM and REM sleep on memory consolidation? *Curr. Neurol. Neurosci. Rep.* **14**, 430, doi:10.1007/s11910-013-0430-8 (2014).

31 Abel, T., Havekes, R., Saletin, J. M. & Walker, M. P. Sleep, plasticity and memory from molecules to whole-brain networks. *Curr. Biol.* **23**, R774-788, doi:10.1016/j.cub.2013.07.025 (2013).

32 Boly, M. *et al.* Hierarchical clustering of brain activity during human nonrapid eye movement sleep. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 5856-5861, doi:10.1073/pnas.1111133109 (2012).

33 Sterpenich, V. *et al.* Memory reactivation during rapid eye movement sleep promotes its generalization and integration in cortical stores. *Sleep* **37**, 1061-1075, 1075A-1075B, doi:10.5665/sleep.3762 (2014).

34 Bourtchouladze, R. *et al.* Different training procedures recruit either one or two critical periods for contextual memory consolidation, each of which requires protein synthesis and PKA. *Learn. Mem.* **5**, 365-374 (1998).

35 Igaz, L. M., Vianna, M. R., Medina, J. H. & Izquierdo, I. Two time periods of hippocampal mRNA synthesis are required for memory consolidation of fear-motivated learning. *J. Neurosci.* **22**, 6781-6789, doi:20026642 (2002).

36 Davis, R. L. Traces of Drosophila memory. *Neuron* **70**, 8-19, doi:10.1016/j.neuron.2011.03.012 (2011).

37 Dubnau, J. & Chiang, A. S. Systems memory consolidation in Drosophila. *Curr. Opin. Neurobiol.* **23**, 84-91, doi:10.1016/j.conb.2012.09.006 (2013).

38 Smith, C. Sleep states and memory processes. *Behav. Brain Res.* **69**, 137-145 (1995).

39 Prince, T. M. *et al.* Sleep deprivation during a specific 3-hour time window post-training impairs hippocampal synaptic plasticity and memory. *Neurobiol. Learn. Mem.* **109**, 122-130, doi:10.1016/j.nlm.2013.11.021 (2014).

40 Stickgold, R., Whidbee, D., Schirmer, B., Patel, V. & Hobson, J. A. Visual discrimination task improvement: a multi-step process occurring during sleep. *J. Cogn. Neurosci.* **12**, 246-254 (2000).

41 Scholz, J., Klein, M. C., Behrens, T. E. & Johansen-Berg, H. Training induces changes in white-matter architecture. *Nat. Neurosci.* **12**, 1370-1371, doi:10.1038/nn.2412 (2009).

42 Schabus, M. *et al.* Sleep spindles and their significance for declarative memory consolidation. *Sleep* **27**, 1479-1485 (2004).

43 Bergmann, T. O., Molle, M., Diedrichs, J., Born, J. & Siebner, H. R. Sleep spindle-related reactivation of category-specific cortical regions after learning face-scene associations. *Neuroimage* **59**, 2733-2742, doi:10.1016/j.neuroimage.2011.10.036 (2012).

44 Schreiner, T. & Rasch, B. Boosting Vocabulary Learning by Verbal Cueing During Sleep. *Cereb. Cortex*, doi:10.1093/cercor/bhu139 (2014).

45 Lisman, J. E. & Jensen, O. The theta-gamma neural code. *Neuron* **77**, 1002-1016, doi:10.1016/j.neuron.2013.03.007 (2013).

46 Walker, M. P. & van der Helm, E. Overnight therapy? The role of sleep in emotional brain processing. *Psychol. Bull.* **135**, 731-748, doi:10.1037/a0016570 (2009).

47 Jackson, C. *et al.* Dynamics of a memory trace: effects of sleep on consolidation. *Curr. Biol.* **18**, 393-400, doi:10.1016/j.cub.2008.01.062 (2008).

48 Schreiner, T., Lehmann, M. & Rasch, B. Auditory feedback blocks memory benefits of cueing during sleep. *Nat. Commun.* **6**, 8729, doi:10.1038/ncomms9729 (2015).

49 Marzano, C. *et al.* Recalling and forgetting dreams: theta and alpha oscillations during sleep predict subsequent dream recall. *J. Neurosci.* **31**, 6674-6683, doi:10.1523/JNEUROSCI.0412-11.2011 (2011).

50 Steriade, M., McCormick, D. A. & Sejnowski, T. J. Thalamocortical oscillations in the sleeping and aroused brain. *Science* **262**, 679-685 (1993).

51 Massimini, M. *et al.* Breakdown of cortical effective connectivity during sleep. *Science* **309**, 2228-2232, doi:10.1126/science.1117256 (2005).

52 Nadasdy, Z., Hirase, H., Czurko, A., Csicsvari, J. & Buzsaki, G. Replay and time compression of recurring spike sequences in the hippocampus. *J. Neurosci.* **19**, 9497-9507 (1999).

53 Staresina, B. P., Alink, A., Kriegeskorte, N. & Henson, R. N. Awake reactivation predicts memory in humans. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 21159-21164, doi:10.1073/pnas.1311989110 (2013).

54 Deuker, L. *et al.* Memory consolidation by replay of stimulus-specific neural activity. *J. Neurosci.* **33**, 19373-19383, doi:10.1523/JNEUROSCI.0414-13.2013 (2013).

55 Tambini, A. & Davachi, L. Persistence of hippocampal multivoxel patterns into postencoding rest is related to memory. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 19591-19596, doi:10.1073/pnas.1308499110 (2013).

56 Oldfield, R. C. The assessment and analysis of handedness: the Edinburgh inventory. *Neuropsychologia* **9**, 97-113 (1971).

57 Roenneberg, T. *et al.* Epidemiology of the human circadian clock. *Sleep medicine reviews* **11**, 429-438, doi:10.1016/j.smrv.2007.07.005 (2007).

58 Minear, M. & Park, D. C. A lifespan database of adult facial stimuli. *Behav. Res. Methods* **36**, 630-633 (2004).

59 Iidaka, T., Matsumoto, A., Haneda, K., Okada, T. & Sadato, N. Hemodynamic and electrophysiological relationship involved in human face processing: evidence from a combined fMRI-ERP study. *Brain Cogn.* **60**, 176-186, doi:10.1016/j.bandc.2005.11.004 (2006).

60 Atkinson, A. P. & Adolphs, R. The neuropsychology of face perception: beyond simple dissociations and functional selectivity. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **366**, 1726-1738, doi:10.1098/rstb.2010.0349 (2011).

61 O'Craven, K. M., Downing, P. E. & Kanwisher, N. fMRI evidence for objects as the units of attentional selection. *Nature* **401**, 584-587, doi:10.1038/44134 (1999).

62 Pourtois, G., Schwartz, S., Spiridon, M., Martuzzi, R. & Vuilleumier, P. Object representations for multiple visual categories overlap in lateral occipital and medial fusiform cortex. *Cereb. Cortex* **19**, 1806-1819, doi:10.1093/cercor/bhn210 (2009).

63 Rechtschaffen, A. & Kales, A. *A manual of standardized terminology, techniques and scoring system for sleep stages of human subjects*. (Brain Information Service, University of California, 1968).

64 Fan, J. & Fan, Y. High dimensional classification using features annealed independence rules. *Ann. Stat.* **36**, 2605-2637 (2008).

65 Norman, K. A., Polyn, S. M., Detre, G. J. & Haxby, J. V. Beyond mind-reading: multi-voxel pattern analysis of fMRI data. *Trends Cog. Sci.* **10**, 424-430, doi:10.1016/j.tics.2006.07.005 (2006).

66 Delorme, A. & Makeig, S. EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *J. Neurosci. Methods* **134**, 9-21, doi:10.1016/j.jneumeth.2003.10.009 (2004).

67 Jamalabadi, H., Alizadeh, S., Schonauer, M., Leibold, C. & Gais, S. Classification based hypothesis testing in neuroscience: Below-chance level classification rates and overlooked statistical properties of linear parametric classifiers. *Hum. Brain Mapp.* **37**, 1842-1855, doi:10.1002/hbm.23140 (2016).

Acknowledgements

This work was supported by Deutsche Forschungsgemeinschaft grant GA730/3-1.

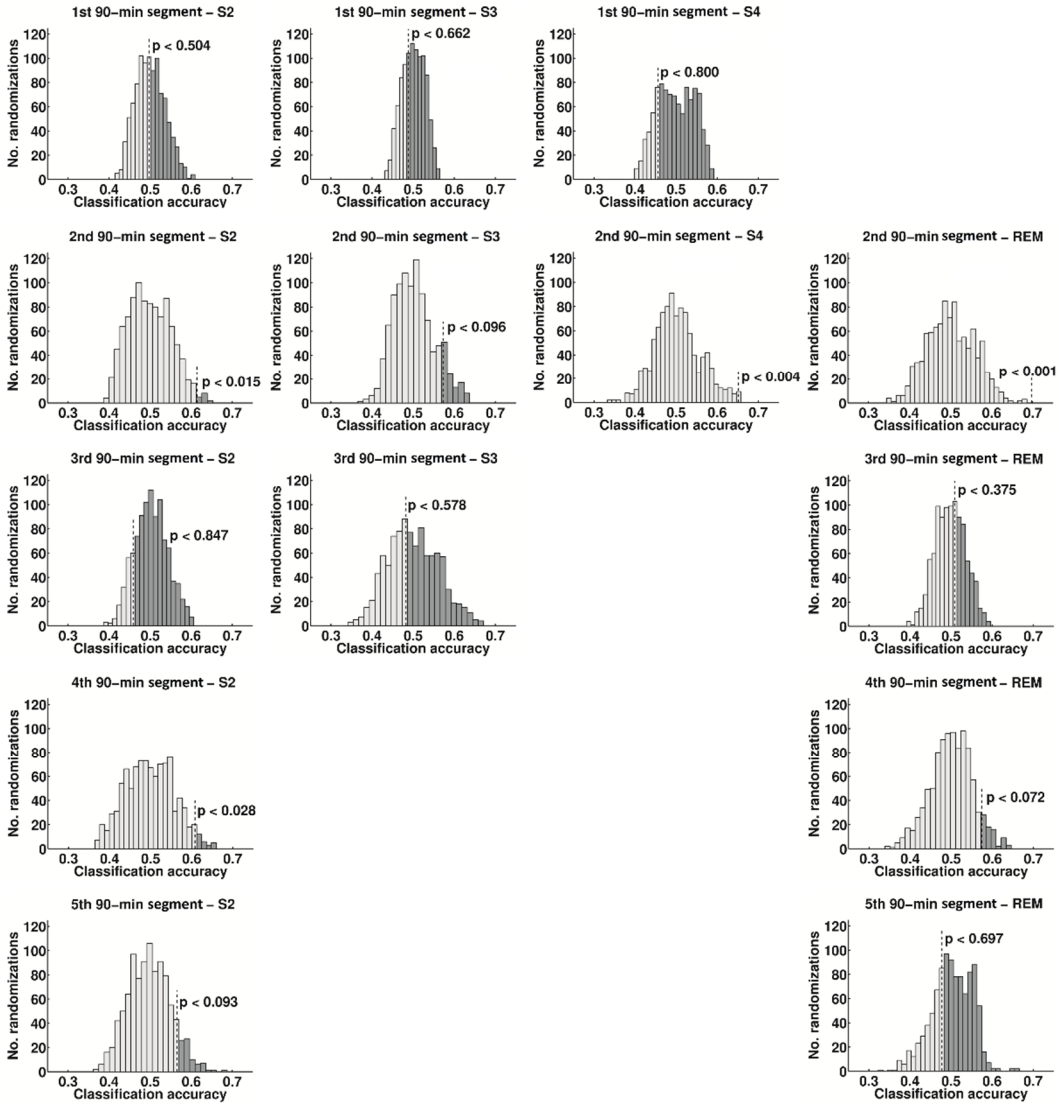
Author contributions

MS, AP, and SG designed the experiments. MS, AA, and AP collected the data. MS, SA, and HJ, analyzed the data. MS, SA, HJ, and SG wrote the manuscript.

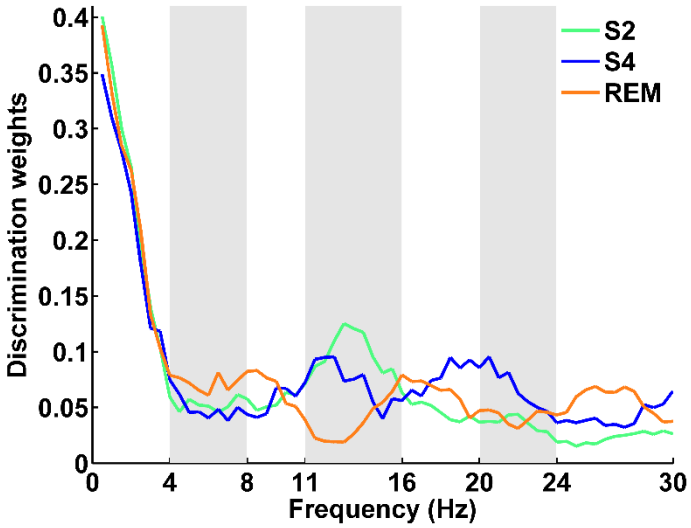
Competing financial interests

The authors declare no conflict of interest.

Supplementary Information



Supplementary Figure 1: Randomization statistics for classification in all segments (rows) and sleep stages (columns). Dark grey areas indicate those randomizations where classification accuracy for randomly labeled data exceeded the classification accuracy obtained with correctly labeled data.



Supplementary Figure 2: Absolute classification weights for the outer loop SVM. Note that weights estimated in the outer loop closely resemble those obtained in the inner loop of the two-step classification procedure (Figure 5).

Supplementary Table 1. Memory sensitivity d' in the face and house learning conditions

	pre	post	difference	p-value
Face pictures	3.72 ± 0.12	3.66 ± 0.12	-0.07 ± 0.04	0.116
House pictures	3.42 ± 0.13	3.34 ± 0.14	-0.08 ± 0.05	0.167

Values are given as mean ± SEM. Two sided t-test for dependent measures is reported. Note that no significant forgetting occurred across the night.

Supplementary Table 2. Correlations between total time spent in sleep stages and memory consolidation (difference in d' post-pre) over sleep for all available nights

	r	*p*	*n*
S2	-0.139	0.272	64
S3	0.106	0.405	64
S4	**0.254***	**0.043**	64
REM	-0.048	0.707	64

**Significant two-sided test at threshold of α < 0.05; Spearman's rho is reported.*

Supplementary Table 3. Correlations between classifier performance (probability estimates for classification) and memory consolidation (difference in d' post-pre) over sleep for all available nights

	difference		**pre**		**post**		
	r	*p*	*r*	*p*	*r*	*p*	*n*
S2 sleep	0.099	0.436	0.023	0.859	0.044	0.733	64
SWS sleep	**0.329****	**0.008**	-0.055	0.667	0.065	0.608	64
REM sleep	-0.199	0.142	0.069	0.611	-0.036	0.791	56

*** Significant two-sided test at threshold of α < 0.01; Spearman's rho is reported.*

Supplementary Table 4. Levels of fatigue in the face and house learning conditions

	Face night	House night	p-value
evening	5.3 ± 2.0	5.5 ± 1.8	0.772
morning	3.7 ± 1.9	3.6 ± 1.6	0.924

Values are given as mean ± SD. Participants were asked to rate their sleepiness on a visual analogue scale with the end points 0 (not tired at all) and 10 (very tired). Two sided t-test for dependent measures is reported.

Supplementary Table 5. Sleep data (mean ± SD)

	W	S1	S2	S3	S4	REM
1st 90-min segment	20.3 ± 11.8	4.8 ± 2.7	29.9 ± 11.8	14.2 ± 6.7	17.9 ± 13.8	2.4 ± 3.3
2nd 90-min segment	3.5 ± 7.8	2.1 ± 1.9	50.9 ± 12.8	11.1 ± 6.6	10.0 ± 8.7	11.0 ± 6.3
3rd 90-min segment	4.2 ± 10.9	2.2 ± 2.0	48.5 ± 10.9	8.0 ± 5.1	5.1 ± 5.7	20.3 ± 7.4
4th 90-min segment	6.9 ± 12.8	2.7 ± 2.2	49.0 ± 13.4	5.6 ± 5.1	1.8 ± 3.8	21.0 ± 8.2
5th 90-min segment	6.9 ± 11.4	4.9 ± 3.8	42.4 ± 12.0	3.3 ± 4.4	1.5 ± 4.1	26.4 ± 11.2
total	48.2 ± 41.5	18.7 ± 8.9	237.7 ± 40.4	42.5 ± 15.0	36.4 ± 23.7	96.0 ± 23.8

Average sleep latency was 20.1 ± 17.0 min (mean ± SD). Please note that total time does not correspond to the sum of 90-min segment values because participants slept slightly longer than five 90-min sleep segments.

Supplementary Table 6. Number of participants and trials that entered classification in different segments and sleep stages. Only data points with N ≥ 11 and number of trials ≥ 40 for both the face and house learning conditions were entered into analysis in each segment and stage.

	S2		S3		S4		REM	
	N	*trials*	*N*	*trials*	*N*	*trials*	*N*	*Trials*
1st 90-min segment	31	472 ± 47	30	355 ± 100	18	455 ± 84	3	279 ± 118
2nd 90-min segment	32	494 ± 33	20	321 ± 102	12	375 ± 93	18	360 ± 74
3rd 90-min segment	29	483 ± 46	16	300 ± 121	6	344 ± 111	24	417 ± 89
4th 90-min segment	24	478 ± 53	9	252 ± 110	2	257 ± 148	19	443 ± 59
5th 90-min segment	20	454 ± 94	0		0		18	415 ± 115

Values for total number of trials collapsed over the face and house conditions that entered classification, given as mean ± SD.

Acknowledgements

First and foremost, I would like to express my gratitude to my supervisor, Steffen Gais, for the many things I learned from him and his support of my work. We had many interesting and long-lasting discussions over a wide range of scientific and nonscientific topics which I enjoyed a lot.

Special thanks also go to Christian Leibold, for his collaboration in my project and for his inspirational mathematical modeling and insightful comments which significantly enhanced the quality of my work.

I am grateful to all my co-authors. Special thanks go to Monika Schönauer, for her continuous support and keen interest in machine learning which significantly improved the results of my work.

My sincere thank goes to my friends and fellow lab mates Andreas Ray, Farid Shiman, Ander Ramos, Jingyi Wang, Frederik Weber, Lea Himmer, Svenja Brodt, Monika Schönauer, Paulo Rogerio, and Thiago Figueiredo with whom I shared so many hours of frustrations, received refreshing and gentle support, and had a lot of fun during the last few years.

I am grateful to my mother who was always so interested in my work. She provided me through moral and emotional support which energized me along the way.

I also would like to thank BCCN and LMU Munich, UKT and GTC Tübingen for the funding and supports. Furthermore, I would like to thank Jan Born and Moritz Grosse-Wentrup who were part of my thesis advisory board committee.

Last but certainly most of all, I express my gratitude towards my wife and colleague Sarah Alizadeh. This thesis would not have been possible without her support which is beyond any words.