1 Classifying Event-Related Desynchronization in EEG, ECoG and MEG signals.

N. Jeremy Hill,[1] Thomas Navin Lal,[1] Michael Schröder,[2] Thilo Hinterberger,[3] Guido Widman,[4] Christian E. Elger,[4] Bernhard Schölkopf,[1] and Niels Birbaumer[3,5]

[1] Max Planck Institute for Biological Cybernetics, Tübingen, Germany
{jez,navin,bs}@tuebingen.mpg.de
[2] Eberhard Karls University, Dept. of Medical Psychology and Behavioral Neurobiology, Tübingen, Germany
{thilo.hinterberger,niels.birbaumer}@uni-tuebingen.de
[3] University of Bonn, Department of Epileptology, Bonn, Germany
{guido.widman,christian.elger}@ukb.uni-bonn.de
[4] Eberhard Karls University, Dept. of Computer Engineering, Tübingen, Germany schroedm@informatik.uni-tuebingen.de
[5] NIH: NINDS, Human Cortical Physiology, Bethesda, USA

Abstract

We present the results from three motor-imagery-based Brain-Computer Interface experiments. Brain signals were recorded from 8 untrained subjects using EEG, 4 using ECoG and 10 using MEG. In all cases, we aim to develop a system that could be used for fast, reliable preliminary screening in the clinical application of a BCI, so we aim to obtain the best possible classification performance in a short time. Accordingly, the burden of adaptation is on the side of the computer rather than the user, so we must adopt a machine-learning approach to the analysis. We introduce the required machine-learning vocabulary and concepts, and then present quantitative results that focus on two main issues. The first is the effect of the number of trials—how long does the recording session need to be? We find that good performance could be achieved, on average, after the first 200 trials in EEG, 75–100 trials in MEG, or 25–50 trials in ECoG. The second issue is the effect of spatial filtering—we compare the performance of the original sensor signals with that of the outputs of Independent Component Analysis and the Common Spatial Pattern algorithm, in each of the three sensor types. We find that spatial filtering does not help in MEG, helps a little in ECoG, and improves performance a great deal in EEG. The unsupervised ICA algorithm performed at least as well as the supervised CSP algorithm in all cases—the latter suffered from poor generalization performance due to overfitting in ECoG and MEG, although this could be alleviated by reducing the number of sensors used as input to the algorithm.

1.1 Introduction

Many different recording technologies exist today for measuring brain activity. In addition to electroencephalography (EEG) and invasive microelectrode recording techniques which have been known for some time, research institutes and clinics now have access to electrocorticography (ECoG), magnetoencephalography (MEG), near-infrared spectrophotometry (NIRS), positron emission tomography (PET) and functional magnetic resonance imaging (fMRI), any of which might be potentially useful in the design and implementation of Brain-Computer Interface systems. Each technology has its own particular set of advantages and limitations as regards spatial and temporal resolution, as well as cost, portability and risk to the user. Comparative studies are required in order to guide development, to explore the trade-offs between these factors.

Bulky, expensive systems (PET, fMRI, MEG) cannot be deployed as day-to-day BCI systems in users' homes, but may offer advantages in the early stages of BCI use. For example, they may be valuable for conducting *screening* procedures, in which a potential user is scanned for one or two sessions in order to ascertain what patterns of brain activity can be most clearly measured and most easily modulated by voluntary intention. An ideal screening would give clinicians the best possible basis on which to decide which task/stimulus setting the user should invest time training with, and (if invasive methods are being considered) where electrodes should be implanted. Regular periodic visits to the scanner might also be a valuable part of early BCI training. However, to justify the cost of screening and training in this way, we would need to know whether the technology yields advantages, for example in terms of signal quality, efficiency, or precision of source localization, that could not otherwise be obtained with cheaper methods.

Here we present a comparative study of motor-imagery BCI experiments based on EEG, ECoG and MEG. In all three, our goal is to develop techniques of analysis that could be used for efficient screening using a simple binary synchronous (trial-based) paradigm, to determine whether subsequent lengthy training in motor imagery might be worthwhile. This requires that we obtain good classification performance as quickly as possible, ideally within the duration of a single recording session. In longer-term user training regimes it might be desirable to fix the mapping between brain activity and output a priori, with users learning to adjust their brain activity such that the mapped recordings meet the desired output. However, in this shorter-term setting, users arrive untrained and so do not necessarily know how to control their brain activity in the optimal manner, which will often have a very user-specific (and usually not well describable) subjective character. Users have relatively little time to adjust and optimize their performance, yet we must still achieve the best results we can. Therefore for current purposes the burden of adaptation in brain-computer communication lies on the side of the computer—we follow the same principle of "letting the machines learn" that guides the Berlin Brain-Computer Interface project (Krauledat *et al.*, 2004). We envisage screening as consisting of

multiple discrete *trials* in which the user is repeatedly asked to produce brain-states of different classes. The mapping from brain states to the desired output is not known and has to be inferred from this limited set of example mappings—a problem of *empirical inference* for which a *machine learning* approach is well suited.

After briefly describing the neurological basis of our studies, the recording technologies and experimental setup, we will introduce some of the machine-learning concepts, terms and tools we will need. We then describe our analysis procedure, present results and conclude. In particular we will be interested in the question of how many trials are necessary to yield good classification performance—in other words, how soon could we have broken off the testing session, and still have obtained comparable results?

1.2 Neurological Phenomena of Imagined Movement

When a person is neither moving nor about to move, the electrical activity of the motor cortex is dominated by frequencies in the 8-12Hz (α-band) and 18–22Hz (β-band) ranges. These signal components are often referred to as μ-rhythms, or more generally as sensory-motor rhythms (SMR).

At the beginning of the planning phase about 1–1.5 seconds before a movement is executed, the SMR gradually diminishes, an effect known as event-related desynchronization (ERD). Slower shifts and deflections in electrical signal, known as movement-related potentials (MRP), can also be observed at roughly the same time. Both neurological phenomena can be recorded best over the motor cortex contralateral to the movement.

It is known that ERD is also present when movements are only imagined (e.g. Pfurtscheller. *et al.*, 1998) or attempted (Kauhanen *et al.*, 2004). Unfortunately, not all users show ERD in motor imagery, although it is possible to train healthy subjects (Guger *et al.*, 2003) as well as patients with ALS (Kübler *et al.*, 2005) to control their SMR such that the recorded activity becomes more classifiable. When present, ERD can be detected relatively easily and is therefore used in the majority of BCI studies.

Using both aspects—MRP and ERD—of the recorded signal leads to improved classification performance (Dornhege *et al.*, 2003a), a result supported by the work of Babiloni *et al.* (1999) who argue that MRP and ERD represent different aspects of cortical processing. In the current study, however, only a very small minority of our subjects showed useable MRPs in our imagined movement task—for simplicity, we therefore focus our attention on ERD.

1.3 Recording Technology

Since our focus is on ERD, we can only consider recording methods that have sufficient temporal resolution to capture changes in the α and β bands. This

rules out technologies such as PET, fMRI and NIRS that rely on the detection of regional changes in cerebral blood oxygenation levels. We briefly introduce the three recording systems we have used: EEG, ECoG and MEG.

1.3.1 EEG

Extracranial electroencephalography is a well-studied recording technique for cerebral activity that has been practised since its invention by Hans Berger in 1929. It measures electrical activity, mainly from the cortex, non-invasively: electrical signals of the order of 10^{-4} Volts are measured by passive electrodes (anything from a single electrode to about 300) placed on the subject's head, contact being made between the skin and the electrode by a conducting gel. EEG shows a very high temporal resolution of tens of milliseconds but is limited in its spatial resolution, the signals being spatially blurred due to volume conduction in the intervening tissue.

EEG experiments account for the large majority of BCI studies due to the hardware's low cost, low risk and portability. For a selection of EEG motor imagery studies, see Wolpaw *et al.* (1997); Birch *et al.* (2003); McFarland *et al.* (1997); Guger *et al.* (1999); Dornhege *et al.* (2004a); Lal *et al.* (2004).

1.3.2 ECoG

Electrocorticography or intracranial EEG is an invasive recording technique in which an array of electrodes, for example an 8-by-8 grid, is placed surgically beneath the skull, either outside or underneath the dura. Strips containing smaller numbers of electrodes may also be inserted into deeper regions of the brain. Unlike invasive microelectrode recording techniques, ECoG measures activity generated by large cell populations—ECoG measurements are thus more comparable to extracranial EEG, but the electrode's close proximity to the cortex and the lack of intervening tissue allows for a higher signal-to-noise ratio, better response at higher frequencies, and a drastic reduction in spatial blurring between neighbouring electrode signals and contamination by artefacts.

Naturally intracranial surgery is performed at some risk to the patient. Today, ECoG implantation is not widespread, but is mostly carried out as a short-term procedure for the localization of epileptic foci, prior to neurosurgical treatment of severe epilepsy. Patients typically have electrodes implanted for one or two weeks for this purpose, a window of opportunity that is being exploited to perform a variety of brain research, including motor-imagery BCI (Graimann *et al.*, 2004; Leuthardt *et al.*, 2004; Lal *et al.*, 2005c).

1.3.3 MEG

Magnetoencephalography is a non-invasive recording technique for measuring the tiny magnetic field fluctuations, of the order of 10^{-14} Tesla, induced by the electrical

activity of populations of cerebral neurons—mainly those in the cortex, although it has been reported that it is also possible to measure activity from deeper subcortical structures (Llinas *et al.*, 1999; Tesche and Karhu, 1997; Baillet *et al.*, 2001). Relative to fMRI, the spatial resolution of MEG is rather low due to the smaller number of sensors (100-300), but it has a high temporal resolution comparable to that of EEG, in the tens of milliseconds.

Due to the extremely low amplitude of the magnetic signals of interest, MEG scanners must be installed in a magnetically shielded room to avoid the signals being swamped by the earth's magnetic field, and the sensors must be cooled, usually by a large liquid helium cooling unit. MEG scanners are consequently rather expensive and non-portable.

Kauhanen *et al.* (2004) presented an MEG study of sensory-motor rhythms during attempted finger movements by tetraplegic patients. Very recently we introduced an online motor-imagery-based BCI using MEG signals (Lal *et al.*, 2005a).

1.4 Experimental Setup

Three experiments form the basis for this chapter: one using EEG (described in more detail by Lal *et al.*, 2004), one using ECoG (Lal *et al.*, 2005c), and one based on MEG recordings (Lal *et al.*, 2005a).

There were 8 healthy subjects in the EEG experiment, seated in an armchair in front of a computer monitor. 10 healthy subjects participated in the MEG experiment, seated in the MEG scanner in front of a projector screen. In the ECoG experiment, 4 patients with epilepsy took part, seated in their hospital bed facing a monitor.

Table 1.1 contains an overview over the three experimental setups. Depending on the setup, subjects performed up to 400 trials. Each trial began with a small fixation cross displayed at the centre of the screen, indicating that the subject should not move, and blink as little as possible. One second later the randomly chosen task cue was displayed for 500 msec, instructing the subject to imagine performing one of two different movements: these were left hand and right hand movement[1] for the EEG study, and movement of either the left little finger or the tongue[2] for the MEG and the ECoG studies (ECoG grids were implanted on the right cerebral hemisphere). The imagined movement phase lasted at least 3 seconds, then the fixation point was extinguished, marking the end of the trial. Between trials was a short relaxation phase of randomized length between 2 and 4 seconds.

1. Visual cues: a small left- or right-pointing arrow, near the centre of the screen.
2. Visual cues: small pictures of either a hand with little finger extended, or of Einstein sticking his tongue out.

Table 1.1: Overview of the three experiments.

	EEG	ECoG	MEG
SUBJECTS	8	4	10
TRIALS PER SUBJECT	400	100–200	200
SENSORS	39	64–84	150
SAMPLING RATE (HZ)	256	1000	625

1.5 Machine-Learning Concepts and Tools

The problem is one of binary classification, a very familiar setting in machine learning. Here we introduce some of the vocabulary of machine learning, in the context of BCI, in order to explain the tools we use. For a more thorough introduction to machine learning in BCI, see Müller *et al.* (2004).

For each subject, we have a number of *data points*, each associated with one of two *target labels*—this is just an abstract way of stating that we have a number of distinct trials, each of which is an attempt by the subject to communicate one of two internal brain states. Each data point is a numerical description of a trial, and its target label denotes whether the subject performed, for example, imagined finger movement or imagined tongue movement on that trial. Classification is the attempt to extract the relevant information from one subset of the data points (the training subset, for which labels are given), to be able to predict as accurately as possible the labels of another subset (the test subset, for which label information is withheld until the time comes to evaluate final classification accuracy). Extraction of the relevant information for prediction on unseen data is termed *generalization* to the new data.

Each data point can be described by a large number of *features*, each feature being (for the current purposes) a real number. The features are the dimensions of the space in which the data points lie. We can choose the feature representation by selecting our preprocessing: a single trial, measured and digitized as t time samples from each of s sensors, may for example be fully described by the s times t discrete sample values, and this feature representation may or may not be useful for classification. An alternative feature representation might be the values that make up the amplitude spectra of the s sensor readings—the same data points have now been mapped into a different *feature space*, which may or may not entail an improvement in the ease of classification.

Note that both of these feature representations specify the positions of data points in very high-dimensional spaces. Successful generalization using a small number of data points in a relatively high-dimensional space is a considerable challenge (Friedman, 1988).

1.5.1 Support Vector Machines

For classification, we choose a Support Vector Machine (SVM) which has proven its worth in a very diverse range of classification problems from medical applications (Lee *et al.*, 2000) and image classification (Chapelle *et al.*, 1999) to text categorization (Joachims, 1998), and bioinformatics (Zien *et al.*, 2000; Sonnenburg *et al.*, 2005). Its approach is to choose a *decision boundary* between classes such that the *margin*, i.e. the distance in feature space between the boundary and the nearest data point, is maximized—intuitively, one can see that this might result in a minimized probability that a point, its position perturbed by random noise, might stray over onto the wrong side of the boundary. Figure 1.1 shows a two-dimensional (i.e. two-feature) example.

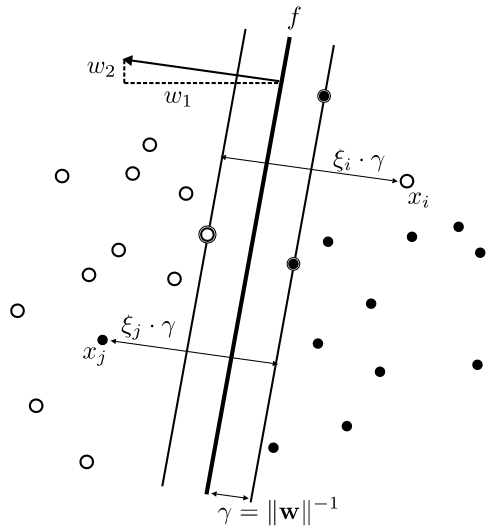

Figure 1.1: Linear SVM. The data are separated by a hyperplane with the largest possible margin γ. For the separable case (ignoring misclassified points x_i and x_j) the three ringed points lying exactly on the margin would be the support vectors (SVs). For non-separable data sets, slack variables ξ_k are introduced—depending on the scaling of these, more points will become SVs.

When one has more features to work on than data points, it is often all too easy to find a decision boundary that separates the training data perfectly into two classes, but which *overfits*. This means that the *capacity* of the classifier (loosely, its allowable complexity—see Vapnik (1998) for the theoretical background) is too large for the data, with the result that the classifier then models too precisely the specific training data points it has seen, and does not generalize well to new test data. Rather than attempting to separate all the data points perfectly, we may obtain better generalization performance if we allow for the possibility that some of the data points, due to noise in the measurement or other random factors, are

simply on the wrong side of the decision boundary. For the SVM, this leads to the *soft-margin* formulation:

$$f : \mathbb{R}^d \to \{-1, 1\}, \qquad \mathbf{x} \mapsto sign(\mathbf{w}^* \cdot \mathbf{x} + b^*)$$

$$(\mathbf{w}^*, b^*) = \underset{\mathbf{w}\in\mathbb{R}^d,\, b\in\mathbb{R}}{\operatorname{argmin}} \|\mathbf{w}\|_2^2 + C\sum_{k=1}^{n} \xi_k^2 \text{ subject to } y_k(\mathbf{w}\cdot\mathbf{x}_k + b) \geq 1 - \xi_k, \quad (k = 1, ..., n)$$

where $\mathbf{x}_k$ is the d-dimensional vector of features describing the k^{th} data point and y_k is the corresponding label, either -1 or $+1$. f is the classifying function, whose parameters $\mathbf{w}$ (the normal vector to the separating hyperplane) and b (a scalar bias term) must be optimized. The constraint $y_k(\mathbf{w} \cdot \mathbf{x}_k + b) \geq 1$ would result in a hard-margin SVM—the closest point would then have distance $\|\mathbf{w}\|^{-1}$ to the hyperplane, so minimizing $\|\mathbf{w}\|$ under this constraint maximizes the margin. The solution for the hyperplane can be written in terms of the *support vectors* which, in the hard-margin case, are the points lying exactly on the margin (highlighted points in figure 1.1). A soft margin is implemented by incorporating a penalty term ξ_k for each data point that lies on the wrong side of the margin, and a *regularization* parameter C which specifies the scaling of these penalty terms relative to the original criterion of margin maximization. Depending on C, the optimal margin will widen and more points will become support vectors.

For a given C there is a unique SVM solution, but a suitable value for C must somehow be chosen. This a question of *model selection* which is often addressed by *cross-validation*: the available training data points are divided randomly into (say) ten non-overlapping subsets of equal size. For each of these ten subsets (or *test folds*, the model is trained on the other 90% (the *training fold*) and tested on the test fold. The average proportion of mistakes made across the ten test folds is taken as the *cross-validation error*, and the model (in this case, the choice of C) with the smallest cross-validation error wins.

One of the SVM's noteworthy features is that it is a *kernel algorithm* (see Schölkopf *et al.*, 1998; Schölkopf and Smola, 2002), i.e. one which does not require an explicit representation of the features, but which can work instead using only a *kernel matrix*, a symmetric square matrix K with each element K_{ij} equal to some suitable measure of similarity between data point i and data point j. This has two advantages. The first is that the time and memory requirements for computation depend more on the number of data points than on the number of features—a desirable property in a trial-based BCI setting since recording a few hundred trials is relatively time-consuming, whereas each trial may be described by a relatively large number of features. The second advantage is that one may use non-linear similarity measures to construct K, which is equivalent to performing linear classification on data points that have been mapped into a higher-dimensional feature space and which can consequently yield a more powerful classifier, *without* the requirement that the feature-space mapping be known explicitly (the so-called *kernel trick*). However, it has generally been observed in BCI classification

applications (for example, see Müller *et al.*, 2003) that, given a well-chosen sequence of preprocessing steps (an explicit feature mapping), a further implicit mapping is usually unnecessary: thus a linear classifier, in which K_{ij} is equal to the dot-product between the feature representations of data points i and j, performs about as well as any non-linear classifier one might attempt. This is often the case in situations in which the number of data points is low, and indeed we find it to be the case in the current application.

Thus we use a linear SVM for the current study, and this has the advantage of interpretability: the decision boundary is a hyperplane, so its orientation may be described by its normal vector $\mathbf{w}$, which is directly interpretable in the explicitly chosen feature space (for example, in the space of multi-channel amplitude spectra). This vector gives us a measure of the relative importance of our features[3] and as such is useful in *feature selection*. In figure 1.1, where we have just two features, the horizontal component of the hyperplane normal vector $\mathbf{w}$ is larger than the vertical, which tells us what we can already see from the layout of the points, namely that horizontal position (feature 1) is more important than vertical position (feature 2) in separating the two classes. Some features may be entirely irrelevant to classification (so the corresponding element of $\mathbf{w}$ should be close to 0). Although the SVM can be formulated as a kernel algorithm and thus does not require explicit feature representation, the number of relevant features relative to the number of irrelevant features is still critical: we would prefer each dot product K_{ij} to be dominated by the sum of the products of relevant features, rather than this information being swamped by the products of irrelevant (noise) features. When one has a large number of features, good feature selection can make a large difference to classification performance.

See Burges (1998), Müller *et al.* (2001) or Schölkopf and Smola (2002) for a more comprehensive introduction to SVMs.

1.5.2 Receiver Operating Characteristic curves and the AUC measure

A Receiver Operating Characteristic (ROC) curve is a plot of a one-dimensional classifier's "hit" rate (for example, probability of the correct identification of a finger-movement trial) against its "false alarm" rate (for example, probability of misidentification of a tongue trial as a finger trial). As one varies the threshold of the classifier, one moves along a curve in this two-dimensional space (a lower threshold for classifying trials as finger trials results in more "hits", but also more "false alarms"). The area under the curve (AUC) is a very informative statistic for the evaluation of performance of classification and ranking algorithms, as well as for the analysis of the usefulness of features. For example, we might order all our

3. Many authors use Linear Discriminant Analysis for this purpose—we choose to use weight vector from the SVM itself, appropriately regularized, since in theory the SVM's good performance relies less on parametric assumptions about the distribution of data points, and in practice this results in a better track record as a classifier.

data points according to their value on a particular single feature axis (say, the amount of band power in a band centred on 10 Hz, measured by a particular sensor at a particular time after the start of the trial) and compute the AUC score of this ordering. An AUC of 1 indicates perfect separability: all the finger trials lie above the highest of the tongue trials on this axis. An AUC of 0 also indicates perfect separability: all the finger trials lie below the lowest of the tongue trials. Thus a value close to 0 or 1 is desirable,[4] whereas a value of 0.5 would indicate that the chosen feature axis is entirely uninformative for the purposes of separating the two classes.

ROC analysis gives rise to many attractive statistical results (see Flach, 2004, for details and references). One attractive property of the AUC score as a measure of feature usefulness is that it is a bounded scale, on which the three values 0, 0.5 and 1 have very clear intuitive interpretations. Another is that it is entirely insensitive to monotonic transformations of the feature axis, relying only on the ordering of the points, and is thus free of any parametric assumptions about the shapes of the class distributions.

Note, however, that we use AUC scores to evaluate features in isolation from each other, which may not give the full picture: it is easy to construct situations in which two highly correlated features each have AUC scores close to 0.5, but in which the sum of the two features separates classes perfectly. Therefore, analysis of individual feature scores should go hand-in-hand with the examination of optimal directions of separation in feature space, by examining the weight vector of a suitably trained classifier. For the current data sets, we find that the two views are very similar, so we plot only the AUC picture.

1.6 Preprocessing and Classification

Starting 500 msec after offset of the visual task cue, we extract a window of length 2 seconds. For each trial and each sensor, the resulting time-series is low-pass-filtered by a zero-phase-distortion method with a smooth falloff between 45 and 50 Hz, downsampled at 100 Hz, and then linearly detrended.

Due to the downsampling, signal components at frequencies higher than 50 Hz are no longer represented in the data. This is no great loss in EEG, since EEG cannot in general be expected to yield much useful information at frequencies higher than this, but it might have been possible to obtain good higher-frequency information in ECoG and MEG. However, based on an examination of the AUC scores of individual

4. In most classical formulations, AUC scores are rectified about 0.5, there being no sense in reporting that a classifier performs "worse than chance" with a score lower than 0.5. However, since here it is entirely arbitrary to designate a "hit" as the correct detection of a finger trial rather than a tongue trial, a value of 0 can be considered just as good as a value of 1, and retaining the unrectified score in the range $[0, 1]$ aids us in interpreting the role of a given feature.

frequency features in each subject's data set, and also of the weight vector of a linear classifier trained on the data, we did not find any indication that this information helped in separating classes in the current task. Figure 1.2 shows typical patterns of AUC scores *before* filtering and downsampling (one representative subject for each of the three sensor types). For all frequencies, there is some "noise" in the AUC values—depending on the number of trials available, values between about 0.4 and 0.6 will be obtained by chance. For all three sensor types, it is only below about 40–50 Hz that we see meaningful patterns in which AUC scores differ significantly from 0.5. While the AUC representation only considers each feature in isolation, an almost identical pattern was observed (for all subjects) in the weights of the linear classifier, which takes linear combinations of features into account.

Therefore, our analysis is restricted to a comparison of the extent to which class-relevant information in the 0–50 Hz range can be recovered using the different recording techniques. It would certainly be interesting to examine the potential use of higher-frequency information—perhaps class-relevant *non-linear* combinations of high-frequency features might be discovered using non-linear classification techniques, or perhaps the higher frequencies might be useful for classification when represented in different ways, other than as amplitude spectra. However, such a study is likely to require considerably larger data sets for individual subjects than those we have currently available, and is beyond the scope of this chapter.

For each number of trials n from 25, in steps of 25, up to the maximum available, we attempt to classify the first n trials performed by the subject. Classification performance is assessed using 10-fold cross-validation, conducted twice with different random seeds. On each of these 20 folds, only the training fold (roughly 90% of the n trials) is used for training, and for feature and model selection—the label information from the remaining $n/10$ trials is used only to compute a final test accuracy estimate for the fold. Where necessary, model and feature selection was performed by a second level of 10-fold cross-validation, *within* the training fold of the outer cross-validation, as described by Müller *et al.* (2004) and Lal *et al.* (2005c,a). Final performance is estimated by averaging the proportion of correctly classified test trials across the 20 outer folds.

Before classification, a spatial filter is computed (see below), and applied to both the training and test trials. Then, amplitude spectra are computed by the short-time Fourier transform (STFT) method of Welch (Welch, 1967): a time series is split into 5 segments each overlapping the next by 50%, a temporal Hanning window is applied to each, and the absolute values of the discrete Fourier transforms of the 5 windowed segments are averaged. For each trial, this gives us a vector of 65 values per sensor (or rather, per spatially filtered linear combination of sensors, which we will call a "channel") as inputs to the classifier.

We use a linear support vector machine as the classifier. First, the regularization parameter is optimized using 10-fold cross validation within the training trial subset. Then we employ the technique of recursive channel elimination (RCE) first described by Lal *et al.* (2004). This is a variant of recursive feature elimination (RFE), an embedded feature-selection method proposed by Guyon *et al.* (2002) in

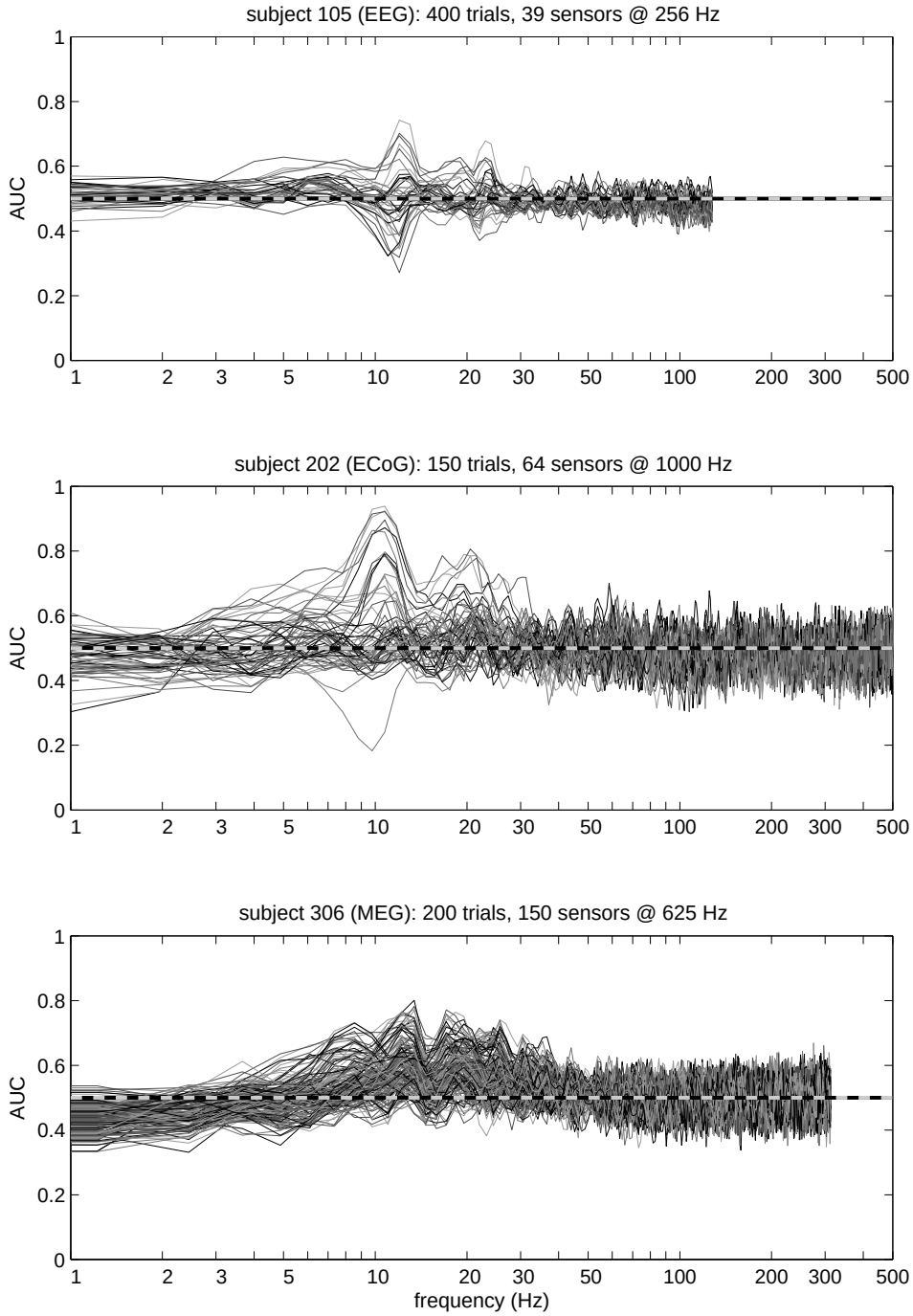

Figure 1.2: AUC scores for multichannel amplitude spectra: representative examples from EEG, ECoG and MEG. Each curve shows the AUC scores corresponding to the frequency-domain features from one of the available sensor outputs.

which an SVM is trained, its resulting weight vector is examined, and the subset of features with the lowest sum squared weight is eliminated (the features being grouped, in our case, into subsets corresponding to channels). Then the procedure is repeated, re-training and re-eliminating for ever-decreasing numbers of features. We run RCE once on the complete training subset, the reverse order of elimination giving us a rank order of channel importance. Then we perform 10-fold cross-validated RCE within the training subset, testing every trained SVM on the inner test fold in order to obtain an estimate of performance as a function of the number of features. Based on the rank order and the error estimates, we reduce the number of channels: we choose the minimum number of channels for which the estimated error is within 2 standard errors of the minimum (across all numbers of features). This procedure is described in more detail in Lal *et al.* (2005a), and embedded feature selection methods are treated in more depth by Lal *et al.* (2005b). Finally, the regularization parameter is re-optimized on the data set after channel rejection and the classifier is ready to be trained on the training subset of the outer fold, in order to make predictions on the test subset.

We summarize the procedure in algorithm 1.

Algorithm 1 Summary of error estimation procedure using nested cross-validation

Require: preprocessed data of one subject
for ($n = 25$ to maximum available in steps of 25) **do**
  take first n trials performed by the subject
  for (outer fold = 1 to 20) **do**
    split data: 90% training set, 10% test set
    with training set do:
      compute spatial filter W
      10-fold inner CV: train SVMs to find regularization parameter C
      10-fold inner CV: RCE to estimate error as a function of number of channels
      RCE on whole training set to obtain channel rank order
        reduce number of channels
        10-fold inner CV: train SVMs to find regularization parameter C
        train SVM S using best C
      with test set do:
        apply spatial filter W
        reject unwanted channels
        test S on test set
        save error
  end for
end for
Output: estimated generalization error (mean and standard error across outer folds)

1.6.1 Spatial Filtering

A spatial filter is a vector of weights specifying a linear combination of sensor outputs. We can represent our signals as an s-by-t matrix X, consisting of s

time series, each of length t, recorded from s different sensors. Spatial filtering amounts to a premultiplication $X' = WX$, where W is an r-by-s matrix consisting of r different spatial filters. If an appropriate spatial filter is applied before any non-linear processing occurs (such as the non-linear step of taking the absolute values of a Fourier transform to obtain an amplitude spectrum), then classification performance on the resulting features will often improve. This is illustrated in figure 1.3, where the AUC scores of the amplitude spectra from one subject in the EEG experiment are considerably better on both training and test folds if the correct spatial filters have been applied. We compare three spatial filtering conditions: no spatial filtering (where W is effectively the identity matrix, so we operate on the amplitude spectra of the raw sensor outputs), Independent Components Analysis (described in section 1.6.1.1) and Common Spatial Pattern filtering (described in section 1.6.1.2).

.

1.6.1.1 Independent Component Analysis (ICA)

Concatenating the n available trials to form s long time series, we then compute a (usually square) separating matrix W that maximizes the independence of the r outputs. This technique is popular in the analysis of EEG signals because it is an effective means of linear blind source separation, in which differently weighted linear mixtures of the signals of interest ("sources") are measured, and must be "de-mixed" to estimate the sources themselves: since EEG electrodes measure the activity of cortical sources through several layers of bone and tissue, the signals are spatially quite "blurred" and the electrodes measure highly correlated (roughly linear) mixtures of the signals of interest. To find a suitable W we use an ICA algorithm based on the Infomax criterion (as implemented in EEGLAB—see Delorme and Makeig (2004)) which we find to be comparable to most other available first-order ICA algorithms in terms of resulting classification performance, while at the same time having the advantage of supplying more consistent spatial filters than many others. Note that, due to the large amount of computation required in the current study, we compute W based on all n trials rather than performing a separate ICA for each outer training/test fold. Target label information is not used by ICA, so there is no overfitting as such, but it could potentially be argued that the setting has become unrealistically "semi-supervised" since the (computationally expensive) algorithm training cannot start until the novel input to be classified has been measured. However, by performing a smaller set of pilot experiments (two values of n for each subject, and only 10 outer folds instead of 20) in which ICA *was* recomputed on each outer fold, we were able to verify that this did not lead to any appreciable difference in performance, either for individual subjects or on average.

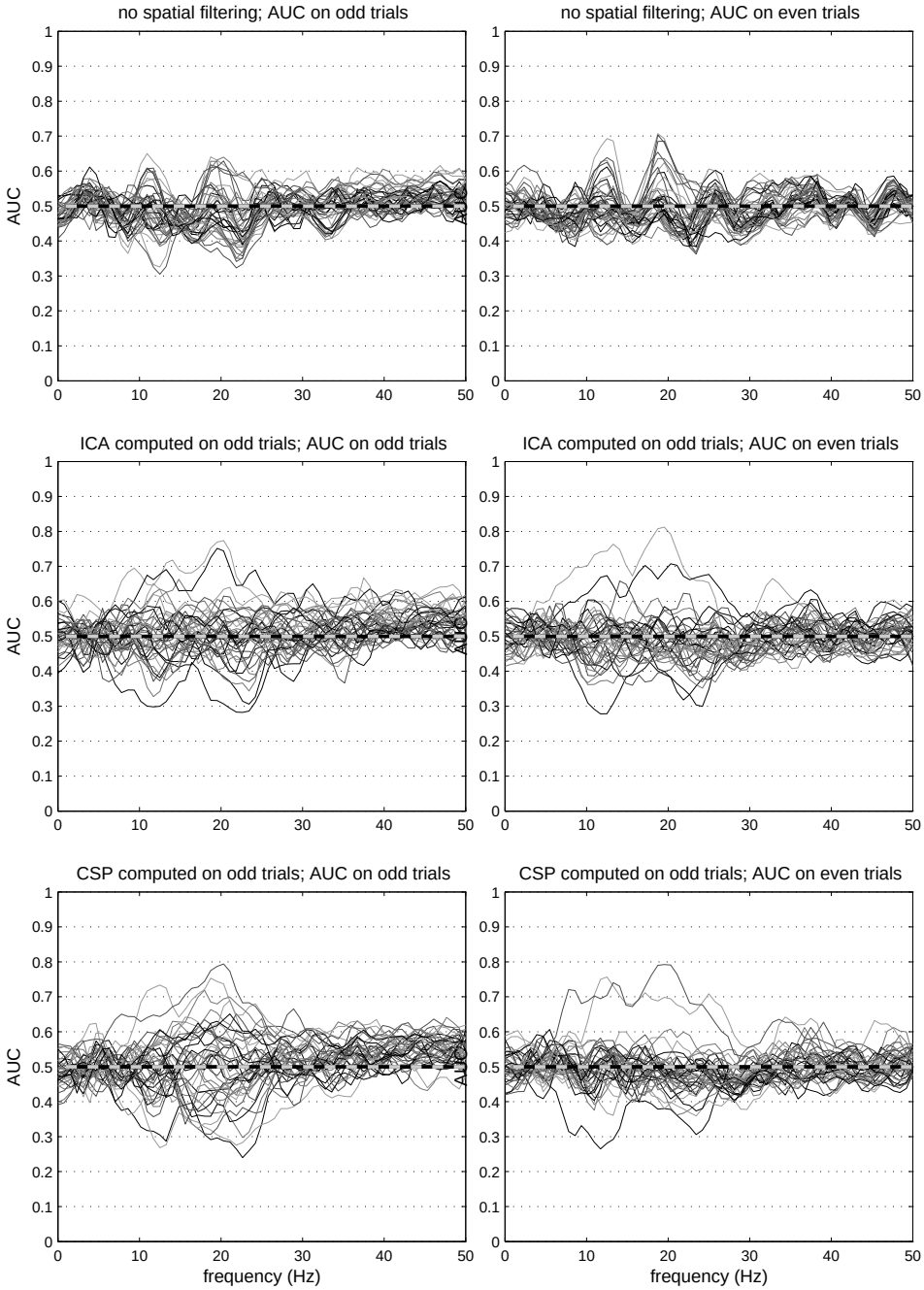

Figure 1.3: Effects of spatial filtering on subject 104 in the EEG experiment. In the left-hand column, we see AUC scores for the amplitude spectra of the odd-numbered trials (a total of 200), and on the right we see AUCs on the even-numbered trials (also 200). In the top row there is no spatial filtering, in the middle there we have applied a square filter matrix W obtained by ICA (section 1.6.1.1) on the odd-numbered trials, and in the bottom row we have applied a square W obtained by CSP (section 1.6.1.2) on the odd-numbered trials

1.6.1.2 Common Spatial Pattern (CSP) analysis

This technique (due to Koles *et al.*, 1990) and related algorithms (Wang *et al.*, 1999; Dornhege *et al.*, 2003b; Lemm *et al.*, 2004; Dornhege *et al.*, 2004b) are supervised methods for computing spatial filters whose outputs have maximal between-class differences in variance. For this to be useful, the input to the algorithm must be represented in such a way that class-dependent changes in the signal are reflected in a change in signal variance: for event-related desynchronization in motor imagery, this can be achieved by applying a zero-phase-distortion band-pass filter which captures the part of the spectrum in which sensorimotor rhythms are expressed: the variance of the filtered signal, which has zero mean, is a measure of amplitude in the chosen band. Here we use a bandpass filter between 7 and 30 Hz (we generally found that this broad band performed approximately as well as any specifically chosen narrow band). Often, the variances of the spatially filtered channels themselves (forming a feature vector $\mathbf{v} = [v_1 \dots v_r]$) are used as features for classification. This makes sense given that the algorithm aims specifically to maximize class differences in this statistic, and it is a convenient way of reducing the dimensionality of the classification problem. In section 1.7.3 we will adopt this approach, discarding the subsequent channel selection stage to save processing time. However, we were able to obtain slightly better performance on the EEG data sets by applying the spatial filters obtained by CSP to the whole (unfiltered) signal, then computing Welch spectra and classifying as described above. Therefore we report the latter results in section 1.7.1.

Since CSP uses label information, it *must* be performed once for each outer training/test fold, using the training subset only. The drawback to CSP is its tendency to overfit, as illustrated in figure 1.4 where we have taken 200 trials from one subject in the 39-channel EEG experiment (upper panel), and 200 trials from the same subject in the 150-channel MEG experiment (lower panel). In each case we have trained the CSP algorithm on half of the available data, and applied the resulting spatial filters W to the other half. We retain the maximum number of spatial patterns, $r = s$, and plot the AUC scores of the features $v_1 \dots v_r$, lighter bars denoting separation of the training trials and darker bars denoting separation of the test trials. In the lower panel we see that, when the algorithm is given a larger number of channels to work with, it finds many linear combinations of channels whose amplitude in the 7–30 Hz band separates the classes nearly perfectly (AUC scores close to 0 or 1). However, the large majority of these patterns tell us nothing about the test trials—only the last two spatial patterns separate the test trials well. In the EEG context, we see that overfitting occurs, but to a lesser extent.[5]

5. The overfitting effect in EEG can also be seen by comparing the left and right panels in the bottom row of figure 1.3, paying particular attention to the centre of the desired 7–30 Hz band.

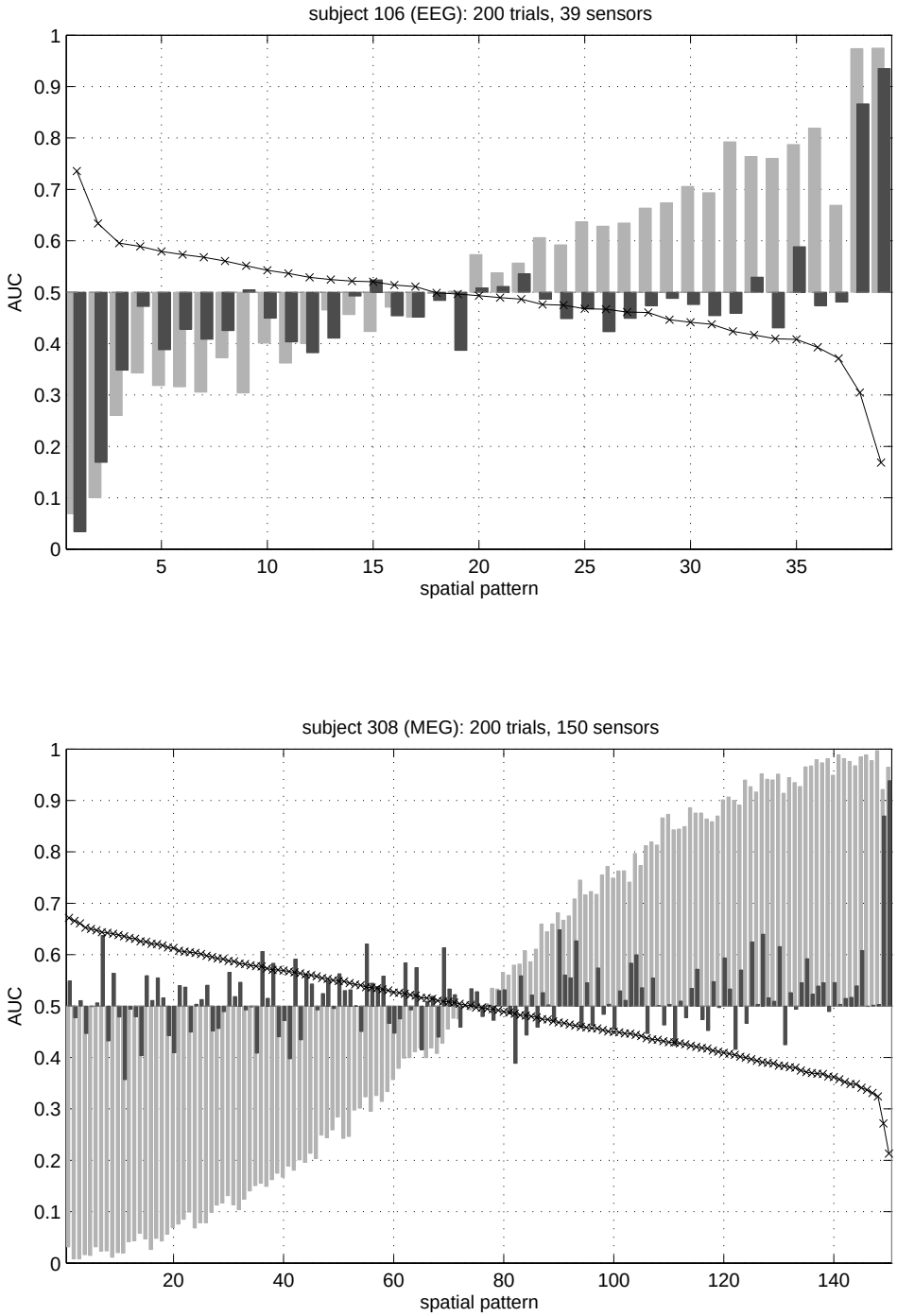

Figure 1.4: Large differences in performance of the CSP algorithm on training data (grey bars) and test data (black bars), as indicated by an AUC measure of class separability computed separately for the projected variance on each spatial pattern. This overfitting effect is extremely pronounced in 150-channel MEG (lower panel), and less so in 39-channel EEG (upper panel). Crosses show the eigenvalues corresponding to each spatial pattern in the CSP decomposition.

The lines of crosses indicate the eigenvalues returned by the CSP algorithm's diagonalization of the whitened class covariance matrix (see Müller-Gerking *et al.*, 1999; Lemm *et al.*, 2004, for an accessible account of the algorithm details). These are in the range $[0, 1]$ and are an indicator of the amount of between-class difference in variance that each spatial pattern is able to account for. Values close to 0.5 indicate the smallest differences and values close to 0 or 1 denote the largest differences and therefore potentially the most useful spatial patterns. The eigenvalues tell us something related, but not identical, to the AUC scores. In the end we are interested in the classifiability of single trials in as-yet unseen data. The eigenvalues are only an indirect measure of single-trial classifiability, since they tell us about the fractions of variance accounted for across many trials. Variance is not a robust measure, so a large eigenvalue *could* arise from very high SMR-modulation in just a small minority of the trials, with the majority of the trials being effectively inseparable according to the spatial pattern in question. AUC scores, on the other hand, are a direct measure of trial separability according to individual features. Hence, AUC scores on the data that the CSP algorithm has not seen (black bars in figure 1.4) are our standard for evaluating the generalization performance of each spatial pattern. By contrast, AUC scores computed from the training trials alone (grey bars) show a grossly inflated estimate of performance, which illustrates the overfitting effect. The eigenvalues show a somewhat intermediate picture. On the one hand, they are computed only on the training trials, and accordingly their magnitude is clearly better predicted by looking at the AUC on training trials than at the AUC on test trials. On the other hand, they also contain information that, in the current examples, allows us to identify which components are really useful according to our standard (tallest black bars). First, by sorting the spatial patterns by eigenvalue, we have correctly sorted the useful components to the extreme ends of the plot. Second, the useful patterns are identifiable by an acceleration in the eigenvalue spectrum towards the ends (c.f. Wang *et al.*, 1999).

In practice, the eigenvalues are a fairly good and often-used predictor of the generalization performance of each spatial pattern. Some such predictor is necessary, since CSP's overfitting will often lead to poor generalization performance. Standard remedies for this employ a feature selection stage after CSP, with the aim of retaining only those spatial patterns that are likely to be useful. Selection strategies may vary: one common approach is to take only the first k in patterns, in the order of preference indicated by the eigenvalues, number k being either fixed, or determined by cross-validation of the CSP algorithm within the training set. The results reported in section 1.7.1 employ this strategy with k fixed at 5, which we found to produce results roughly as good as a cross-validation strategy.[6]

6. One may also attempt to perform channel selection after CSP, *without* using the eigenvalues or cross-validating the CSP algorithm itself, but this is hampered by the fact that the training data have been transformed by an algorithm that overfits on them: cross-validation error rates in subsequent model and feature selection tend to be uninformatively close to 0, and classifiers end up under-regularized. To investigate a possible workaround

In section 1.7.3 we employ an additional tactic: since the degree of overfitting is determined largely by the number of free parameters in the optimization, and the algorithm finds one scaling parameter per sensor in each spatial pattern, it makes sense to attempt to reduce the number of sensors used as input to the CSP algorithm. We do this using a preliminary step in which Welch spectra of the raw sensor outputs are computed, an SVM is trained (cross-validating to find the best regularization parameter) and the weight vector is used to provide a measure of relative channel importance, as in RCE. Going back to the time-domain representation, the top 10, 25, 39 and (in ECoG and MEG) 55 sensors found by this method were then passed into CSP. Spatial patterns were then chosen by a cross-validation method: CSP was run on each of 10 inner training folds and variances $v_1 \dots v_r$ were computed on the corresponding test fold and saved. At the end of cross validation, each trial then had a new representation $\mathbf{v}$, and AUC scores corresponding to each of these features could be computed on the whole outer training fold, and these are useful for selection since they generally correlate well with the AUC scores on unseen data. The significance of the AUC values was expressed in terms of the standard deviation expected from random orderings of a data set of the same size. Eigenvalue positions with AUC scores more than 2 standard deviations away from 0.5 were retained in the outer CSP.

1.7 Results

1.7.1 Performance of spatial filters using all available sensors

In figure 1.5, classification accuracy is plotted as a function of n for each subject, along with average performance in each of the three experiments (EEG, ECoG and MEG). We plot the time-course of the *overall* effectiveness of the experimental setup, subject and classifier taken all together: our curves are obtained by computing performance on the *first* 25 trials performed by the subject, then recomputing based on the first 50 trials, and so on (instead of on a random 25 trials, then a random 50 trials). As a result the observed changes in performance with increasing n reflect not only effect of the amount of input on classifier performance, but also

for this, we tried splitting each training set into two partitions: one to be used as input to CSP to obtain spatial filters W, and the other to be transformed by W and then used in channel selection and classifier training as described above. We experimented with a 25:75 percent partition, as well as 50:50 and 75:25, of which 50:50 was found to be the best for nearly all values of n. However, the resulting performance was worse than in the simpler strategy of performing CSP on the whole training set and taking the best 5 eigenvalues—the reduction in the number of trials available for CSP exacerbates the overfitting problem to an extent that is not balanced out by the improvement in feature and model selection. The results of the partition experiments are not shown.

changes in the subjects' performance, whether due to practice, fatigue or transient random influences.

Note that, for 2 out of 8 subjects in the EEG condition (subjects 101 and 102), and 1 out of 10 in MEG (subject 303), we were never able to classify at significantly better than chance level. These subjects were omitted from the averaging process and from the further analysis of section 1.7.3. The strength of sensorimotor rhythms, and the degree to which their modulation with imagined movement is measurable, varies from person to person. One must expect that some subjects will be unable to use a motor-imagery-based BCI at all, and that performance of the remainder will vary between individuals. Given the necessarily small size of our three subject groups, we are unable to draw strong conclusions as to the effect of recording technology on absolute performance level, to say for example whether MEG is a significantly better option than EEG. Another effect of between-subject variation is that, though we find certain individual subjects in all three groups who are able to attain high performance levels (say, $> 90\%$), average performance is poor. However, it should be borne in mind that, with one exception,[7] the subjects had never taken part in a motor-imagery BCI experiment before, and that performance is therefore based on a maximum of three hours' experience with the paradigm, and without feedback.

In the EEG experiment, both ICA (grey asterisks) and CSP (open diamonds) allow very large improvements in performance relative to the condition in which no spatial filtering was used (filled circles). This effect is clear in the averaged data as well as in the individual subject plots. In ECoG, the difference between ICA and no spatial filtering is slight, although ICA is at least as good as no spatial filtering for all four subjects; CSP is consistently a little worse than either. In MEG, there is no consistent benefit or disadvantage to ICA over the raw sensor outputs, and again CSP is worse, this time by a larger margin.

The failure of CSP in ECoG and MEG is likely to be related to the overfitting effect discussed above. This is clearest for subject 310 when 200 trials are used: although spatial filters exist (and have been found by ICA) which can improve classification performance, CSP fails to find any patterns which help to classify the data, because useless (overfitted) spatial patterns dominate the decomposition of the class covariance matrices.

Overall, maximum performance can be achieved using about 200 trials in EEG and 75–100 trials in MEG. For ECoG, though it is harder to draw strong conclusions due to the smaller number of subjects and trials, it generally appears that the curves are even flatter: the best results can already be obtained with only 25–50 trials.

One curious feature of the results is the strikingly good performance without spatial filtering for some subjects (103, 104, 107, 108, 302, and 308) when only the first 25 trials are tested, quickly dropping to much poorer performance when more

7. The exception is subject 308 in the MEG condition, who had previously taken part in the EEG condition—106 and 308 are the same person.

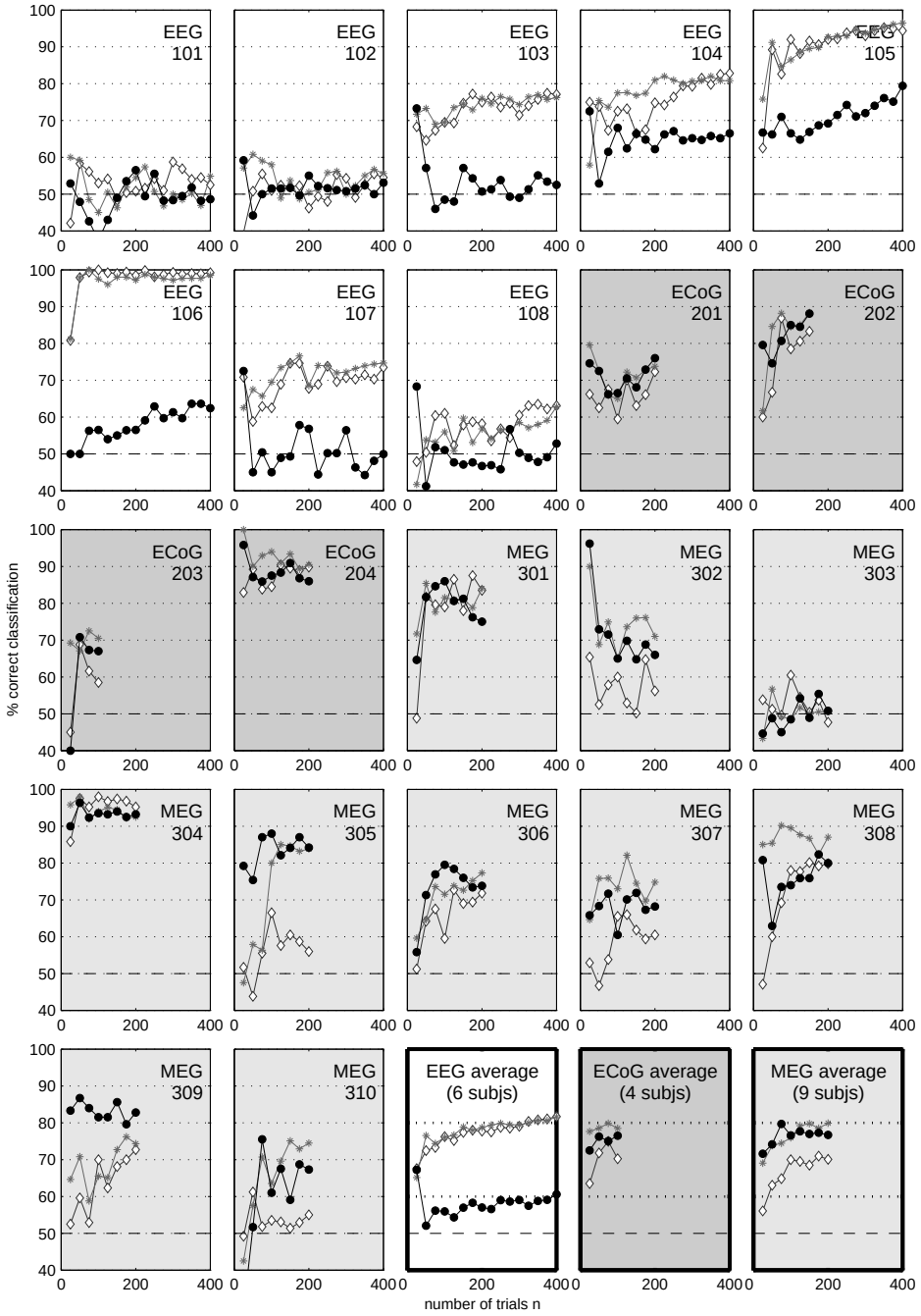

Figure 1.5: For each subject, classification accuracy is plotted as a function of the number of trials performed and the spatial filtering method employed: filled circles denote no spatial filtering, asterisks denote ICA, and open diamonds denote CSP. The last three plots show averages, for the EEG, ECoG and MEG experiments respectively, across all subjects for whom classification had been possible at all.

trials are taken. A possible explanation for this lies in the fact that the test trials on each outer fold were drawn uniformly and randomly from the first n trials—when n is very small, this means that the test trials were performed, on average, closer in time to the training trials than when n is larger. If the subjects' signals exhibit properties that are non-stationary over time, this may lead to an advantage when the training and test trials are closer together. Such effects merit a more in-depth analysis which is beyond the scope of this report.

1.7.2 Topographic interpretation of results

Figure 1.6 shows topographic maps of the features selected by our analysis, for seven of our subjects. Sensor ranking scores were obtained by Recursive Channel Elimination on data that had not been spatially filtered: each of the 20 outer training/test folds of the analysis returned a channel ranking, and these ranks were averaged across folds and then divided by their standard deviation across folds. The result indicates which channels were ranked highly most consistently (darker colours indicating channels ranked as more influential). We also plot spatially interpolated projected amplitudes[8] for the top two independent components (selected by Recursive Channel Elimination in the first outer training/test fold) and the first two spatial patterns (indicated by the best two eigenvalues in the first outer fold).

In general, we see that ICA and CSP recover very similar patterns of activation which are consistent with the modulation of activity in motor and pre-motor cortical areas. In EEG, both algorithms recover patterns centred on C4/CP4 in the right hemisphere (where we would expect modulation associated with imagined left hand movement) and C3/CP3 in the left (imagined right hand movement). In MEG, we see patterns consistent with parietal-central and central-frontal dipoles in the right hemisphere where we would expect to see modulation associated with imagined left hand movement. Subject 308 appears to use sources in both hemispheres—possibly a bilateral representation of tongue movement. In the ECoG, the top two independent components and the top spatial pattern are all highly localized, activation in each case being focused on just three or fewer electrodes located above the motor cortex.

For subjects 202, 304 and 306, the second spatial pattern shows a more complicated topography. Given that CSP generally performs less well than ICA for these subjects, we may suspect that this is a reflection of overfitting. Presented with a large number of sensors, the algorithm can account for class differences in signal variance by combining sensors in spatial configurations that are more complicated than necessary, which in turn results in poorer generalization performance.

---

8. Each map is spline-interpolated from a single column of the mixing matrix W^{-1}, the inverse of the spatial filter matrix. The column corresponding to a given source tells us the measured amplitude of that source as a function of sensor location.

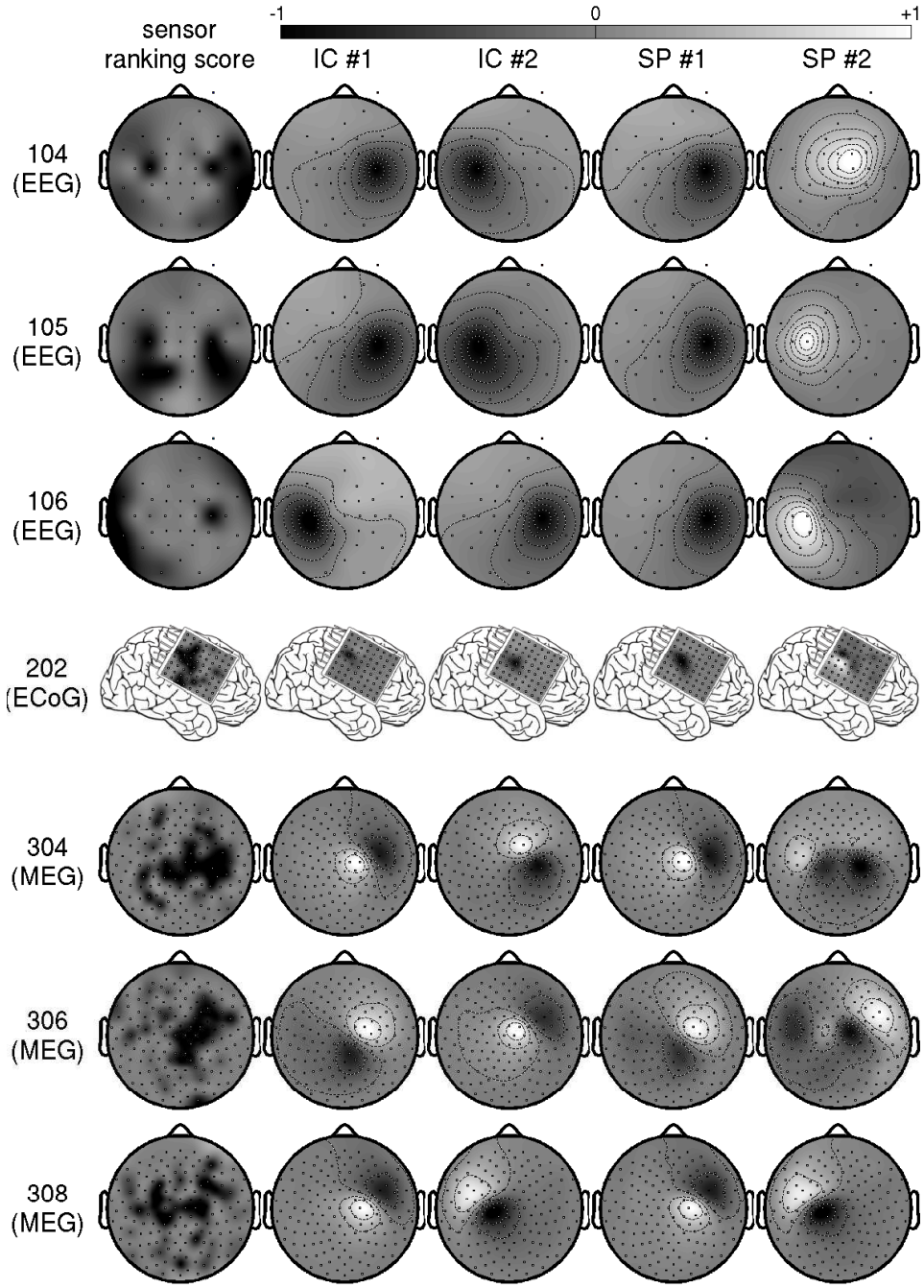

Figure 1.6: Topographic maps showing the ranking or weighting of sensors at different spatial locations, for three EEG subjects, one ECoG subject, and three MEG subjects. Sensor ranking scores (first column) are obtained by Recursive Channel Elimination on the data when no spatial filtering is used. The top two independent components (columns 2–3) are selected by Recursive Channel Elimination after Independent Component Analysis. The top two spatial patterns (columns 4–5) are selected using the eigenvalues returned by the CSP algorithm. Topographic maps are scaled from -1 (black) through 0 (grey) to 1 (white) according to the maximum absolute value in each map.

Finally, we note that the ranking scores of the raw sensors, while presenting a somewhat less tidy picture, generally show a similar pattern of sensor importance to that indicated by the ICA and CSP maps (note that the ranking score patterns may reflect information from influential sources beyond just the first two components that we have shown). The sensors ranked most consistently highly are to be found in lateralized central and pre-central regions, bilaterally for the EEG experiment and for subject 308, and with a right-hemisphere bias for the others. For further examination of the performance of Recursive Channel Elimination in the identification of relevant source locations, see Lal *et al.* (2004), Lal *et al.* (2005c) and Lal *et al.* (2005a).

1.7.3 Effect of sensor subsetting

In figure 1.7 we show average classification accuracy at $n = 25$, 50, 100 and 200 (respectively in the four rows from top to bottom) for EEG, ECoG and MEG (left to right). Classification performance of CSP is shown as a function of the number of sensors the algorithm is permitted to work with ("more" denoting the maximum available: 64, 74 or 84 in ECoG, and 150 in MEG).

First we note that, in our EEG data, performance is better the more sensors are used, up to the maximum of 39 available in the current study. For ECoG and MEG, this trend is reversed when the number of available trials is small. This is in line with our intuition about overfitting: we suffer when attempting to recombine too many channels based on a small number of data points. For $n = 25$, $s = 10$ is the best number of sensors to choose, and CSP performance may then equal (and even exceed, although the difference is not significant) the best classification previously possible with ICA (in ECoG) or with the raw sensors outputs (in MEG). As the number of trials n increases to 50 and beyond, the peak shifts to the left (it is useful to have more sensors available as the number of trials increases) and the slope becomes shallower as the difference between CSP and the raw sensors diminishes (overfitting becomes less of an issue).

1.8 Summary

We have compared the classifiability of signals obtained by EEG, ECoG and MEG in a binary, synchronous motor-imagery-based Brain-Computer Interface. We held the time interval, and (after failing to find any information useful for the classification in frequencies above 50 Hz) also the sampling frequency, constant across sensor types, and classified event-related desynchronization effects in the signals' amplitude spectra using regularized Support Vector Machines and automatic feature selection.

We varied the number of trials used, in order to see how quickly we might reach maximum classification performance with our unpractised subjects. Maximum performance, averaged across subjects, was roughly equal across sensor types at around 80%, although subject groups were small and between-subject variation was

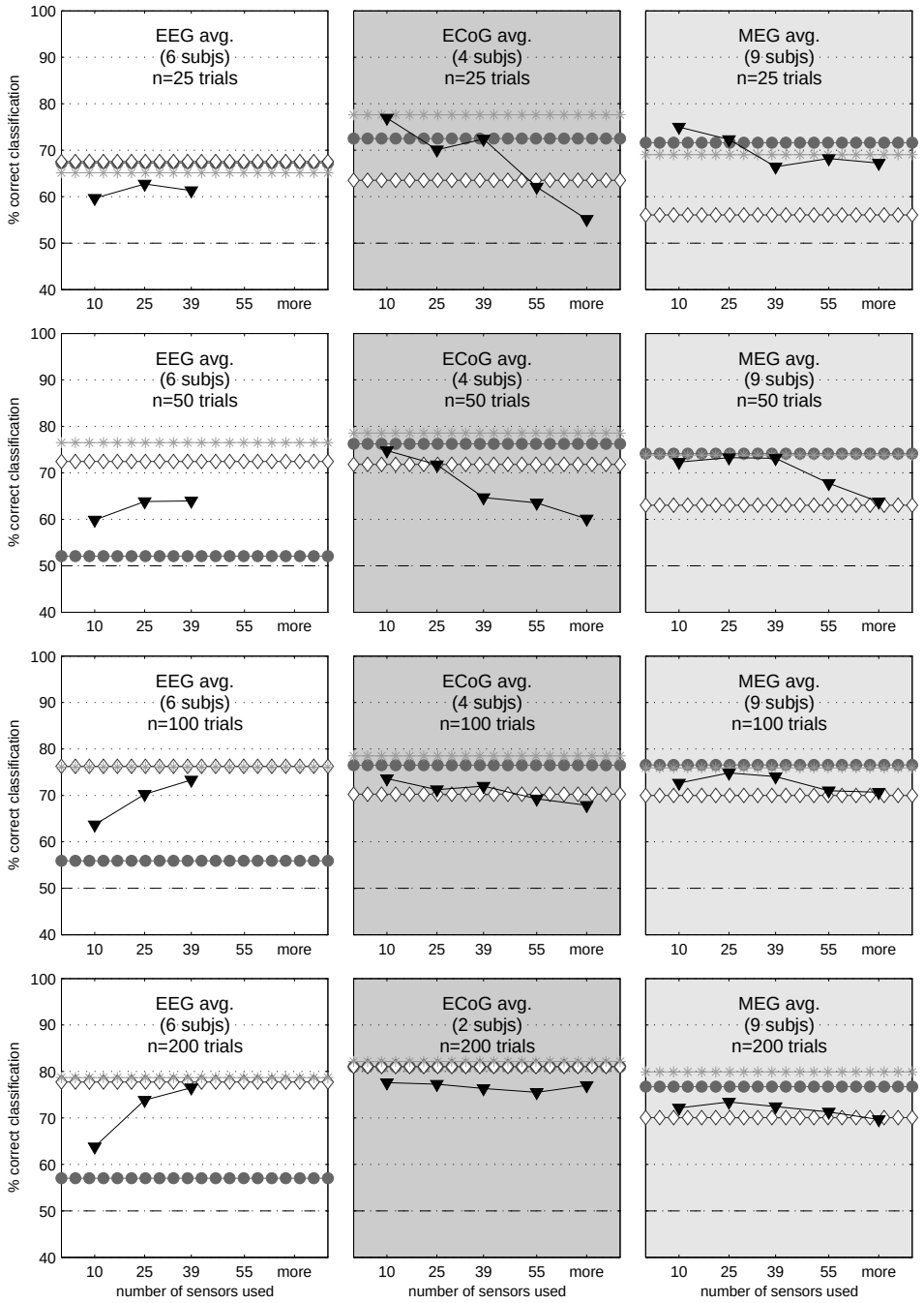

Figure 1.7: Filled triangles indicate average classification accuracy of the CSP algorithm for each sensor type (EEG, ECoG and MEG respectively in the three columns from left to right), as a function of the number of sensors used as input to the algorithm, and the number of trials (25, 50, 100 and 100 respectively in the four rows from top to bottom). For comparison, horizontal chains of symbols denote the average performance levels reported in the previous analysis of figure 1.5: filled circles for no spatial filtering, asterisks for ICA, open diamonds for CSP with a fixed number of patterns $k = 5$.

large, so we attach no particular weight to this observation. Maximum performance was attained after about 200 trials in EEG, 75–100 trials in MEG, and 25–50 trials in ECoG.

Performance was affected by spatial filtering strategy in a way that depended on the recording hardware. For EEG, where signals are highly spatially blurred, spatial filtering is crucial: large gains in classification accuracy were possible using either first-order Independent Component Analysis or the Common Spatial Pattern algorithm, the performance of these two approaches being roughly equal. For ECoG and MEG, as one might expect from systems that experience less cross-talk between channels, spatial filtering was less critical: the MEG signals were the "cleanest" in this regard, in that there was no appreciable difference in performance between classification of the raw sensor outputs and classification of any of the linear combinations of sensors we attempted. First-order spatial filtering would appear to be become largely redundant for the detection of event-related desynchronization as the volume conduction problem diminishes (down to the level at which it is still present in ECoG and MEG).

Across all three conditions, ICA was the best (or roughly equal-best) spatial filtering strategy. CSP suffered badly from overfitting in the ECoG and MEG conditions when large numbers of sensors (> 40) were used, resulting in poor generalization performance. This could be remedied by sparsification of the spatial filters, where a subset of the sensors was selected and the rest discarded—a strategy that was particularly effective when the number of trials was very small, but which never resulted in a significant overall win for optimized spatial filters as against raw sensor outputs. We did not find a convincing advantage, with any of the three sensor types, of supervised optimization of the spatial filters over blind source separation.

Acknowledgements

We would like to thank Hubert Preissl, Jürgen Mellinger, Martin Bogdan, Wolfgang Rosenstiel, and Jason Weston for their help with this work, as well as the two anonymous reviewers whose careful reading of the chapter and helpful comments enabled us to make significant improvements to the manuscript.

The authors gratefully acknowledge the financial support of the *Max-Planck-Gesellschaft*, the *Deutsche Forschungsgemeinschaft* (SFB550/B5 and RO1030/12), the European Community IST Programme (IST-2002-506778 under the PASCAL Network of Excellence), and the *Studienstiftung des deutschen Volkes* (grant awarded to T.N.L.).

References

Babiloni, C., Carducci, F., Cincotti, F., Rossini, P., Neuper, C., Pfurtscheller, G., and Babiloni, F. (1999). Human movement-related potentials vs desynchronization of EEG alpha rhythm: A high-resolution EEG study. *NeuroImage*, **10**, 658–665.

Baillet, S., Mosher, J., and Leahy, R. (2001). Electromagnetic brain mapping. *IEEE Singal Processing Magazine*, **18**(6), 14–30.

Birch, G., Mason, S., and Borisoff, J. (2003). Current trends in brain-computer interface research at the Neil Squire Foundation. *IEEE Transactions on Rehabilitation Engineering*, **11**(2), 123–126.

Burges, C. (1998). A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, **2**(2), 121–167.

Chapelle, O., Haffner, P., and Vapnik, V. (1999). SVMs for histogram based image classification. *IEEE Transactions on Neural Networks*, **9**.

Delorme, A. and Makeig, S. (2004). EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics. *Journal of Neuroscience Methods*, **134**, 9–21.

Dornhege, G., Blankertz, B., Curio, G., and Müller, K.-R. (2003a). Combining features for BCI. In S. T. S. Becker and K. Obermayer, editors, *Advances in Neural Information Processing Systems*, volume 15, Cambridge, MA, USA. MIT Press.

Dornhege, G., Blankertz, B., and Curio, G. (2003b). Speeding up classification of multi-channel brain-computer interfaces: Common spatial patterns for slow cortical potentials. *Proceedings of the 1st International IEEE EMBS Conference on Neural Engineering*, pages 591–594.

Dornhege, G., Blankertz, B., Curio, G., and Müller, K.-R. (2004a). Boosting bit rates in noninvasive eeg single-trial classifications by feature combination and multiclass paradigms. *IEEE Transactions on Biomedical Engineering*, **51**(6), 993–1002.

Dornhege, G., Blankertz, B., Curio, G., and Müller, K.-R. (2004b). Increase information transfer rates in bci by csp extension to multi-class. In S. Thrun, L. Saul, and B. Schölkopf, editors, *Advances in Neural Information Processing Systems*, volume 16. MIT Press, Cambridge, MA.

Flach, P. A. (2004). The many faces of ROC analysis in machine learning. Tutorial presented at the 21st International Conference on Machine Learning. Available from: `http://www.cs.bris.ac.uk/~flach/ICML04tutorial/`.

Friedman, J. H. (1988). Fitting functions to noisy data in high dimensions. In E. Wegman, D. Gantz, and J. Miller, editors, *Computing Science and Statistics: Proceedings of the 20th Symposium on the Interface*, pages 13–43, Alexandria, VA, USA. American Statistical Association.

Graimann, B., Huggins, J. E., Levine, S. P., and Pfurtscheller, G. (2004). Towards a direct brain interface based on human subdural recordings and wavelet packet analysis. *IEEE Transactions on Biomedical Engineering*, **51**(6), 954–962.

Guger, C., Harkam, W., Hertnaes, C., and Pfurtscheller, G. (1999). Prosthetic control by an EEG-based brain-computer interface (BCI). In *Proceedings if the 5th European Conference for the Advancement of Assistive Technology (AAATE)*, Düsseldorf, Germany.

Guger, C., Edlinger, G., W.Harkam, Niedermayer, I., and Pfurtscheller, G. (2003). How many people are able to operate an EEG-based brain-computer interface (BCI). *IEEE Transactions on Rehabilitation Engineering*, **11**(2), 145–147.

Guyon, I., Weston, J., Barnhill, S., and Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine Learning*, **46**, 389–422.

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. *European Conference on Machine Learning*.

Kauhanen, L., Rantanen, P., Lehtonen, J. A., Tarnanen, I., Alaranta, H., and Sams, M. (2004). Sensorimotor cortical activity of tetraplegics during attempted finger movements. *Biomedizinische Technik*, **49**(1), 59–60.

Koles, Z. J., Lazar, M. S., and Zhou, S. Z. (1990). Spatial patterns underlying population differences in the background EEG. *Brain Topography*, **2**(4), 275–284.

Krauledat, M., Dornhege, G., Blankertz, B., Curio, G., and Müller, K.-R. (2004). The berlin brain-computer interface for rapid response. *Biomedizinische Technik*, **49**(1), 61–62.

Kübler, A., Nijboer, F., Mellinger, J., Vaughan, T., Pawelzik, H., Schalk, G., McFarland, D., Birbaumer, N., and Wolpaw, J. (2005). Patients with ALS can use sensorimotor rhythms to operate a braincomputer interface. *Neurology*, **64**, 1775–1777.

Lal, T., Schröder, M., Hinterberger, T., Weston, J., Bogdan, M., Birbaumer, N., and Schölkopf, B. (2004). Support vector channel selection in BCI. *IEEE Transactions on Biomedical Engineering. Special Issue on Brain-Computer Interfaces*, **51**(6), 1003–1010.

Lal, T., Schröder, M., Hill, J., Hinterberger, T., Mellinger, J., Rosenstiel, W., Hofmann, T., Birbaumer, N., and Schölkopf, B. (2005a). A brain computer interface with online feedback based on magnetoencephalography. In *Proceedings*

of the 22nd International Conference on Machine Learning (ICML), pages 465 – 472.

Lal, T., Chapelle, O., Weston, J., and Elisseeff, A. (2005b). Embedded methods. In I. Guyon, S. Gunn, M. Nikravesh, and L. Zadeh, editors, *Feature extraction, foundations and applications*. Springer. In press.

Lal, T. N., Hinterberger, T., Widman, G., Schröder, M., Hill, N. J., Rosenstiel, W., Elger, C. E., Schölkopf, B., and Birbaumer, N. (2005c). Methods towards invasive human brain computer interfaces. In L. K. Saul, Y. Weiss, and L. Bottou, editors, *Advances in Neural Information Processing Systems 17*. MIT Press, Cambridge, MA.

Lee, Y.-J., Mangasarian, O. L., and Wolberg, W. H. (2000). Breast cancer survival and chemotherapy: A support vector machine analysis. *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, **55**, 1–10.

Lemm, S., Blankertz, B., Curio, G., and Müller, K.-R. (2004). Spatio-spectral filters for robust classification of single trial EEG. *IEEE Transactions on Biomedical Engineering*, **52**(9), 993 – 1002.

Leuthardt, E., Schalk, G., Wolpaw, J., Ojemann, J., and Moran, D. (2004). A brain computer interface using electrocorticographic signals in humans. *Journal of Neural Engineering*, **1**, 63–71.

Llinas, R., Ribary, U., Jeanmonod, D., Kronberg, E., and Mitra, P. (1999). Thalamocortical dysrhythmia: A neurological and neuropsychiatric syndrome characterized by magnetoencephalography. *Proceedings of the National Academy of Science*, **96**(26), 15222–15227.

McFarland, D., McCane, L., David, S., and Wolpaw, J. (1997). Spatial filter selection for EEG-based communication. *Electroencephalography and Clinical Neurophysiology*, **103**, 386–394.

Müller, K.-R., Mika, S., Rätsch, G., Tsuda, K., and Schölkopf, B. (2001). An introduction to kernel-based learning algorithms. *IEEE Transactions on Neural Networks*, **12**(2), 181–201.

Müller, K.-R., Anderson, C. W., and Birch, G. E. (2003). Linear and nonlinear methods for brain-computer interfaces. *IEEE Transactions on Rehabilitation Engineering*, **11**(2), 165–169.

Müller, K.-R., Krauledat, M., Dornhege, G., Curio, G., and Blankertz, B. (2004). Machine learning techniques for brain-computer interfaces. *Biomedical Engineering*, **49**(1), 11–22.

Müller-Gerking, J., Pfurtscheller, G., and Flyvbjerg, H. (1999). Designing optimal spatial filters for single-trial EEG classification in a movement task. *Clinical Neurophysiology*, **110**(5), 787–98.

Pfurtscheller., G., Neuper, C., Schlögl, A., and Lugger, K. (1998). Separability of EEG signals recorded during right and left motor imagery using adaptive autoregressive parameters. *IEEE Transactions on Rehabilitation Engineering*,

6(3), 316–325.

Schölkopf, B. and Smola, A. (2002). *Learning with Kernels*. MIT Press, Cambridge, USA.

Schölkopf, B., Smola, A., and Müller, K.-R. (1998). Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, **10**(5), 1299–1319.

Sonnenburg, S., Rätsch, G., and Schölkopf, B. (2005). Large scale genomic sequence SVM classifiers. *Proceedings of the International Conference on Machine Learning*.

Tesche, C. and Karhu, J. (1997). Somatosensory evoked magnetic fields arising from sources in the human cerebellum. *Brain Research*, **744**, 23–31.

Vapnik, V. N. (1998). *Statistical Learning Theory*. John Wiley and Sons, New York, USA.

Wang, Y., Berg, P., and Scherg, M. (1999). Common spatial subspace decomposition applied to analysis of brain responses under multiple task conditions: a simulation study. *Clinical Neurophysiology*, **110**(4), 604–614.

Welch, P. (1967). The use of fast fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms. In *IEEE Trans. Audio Electroacoustics*, volume AU-15, pages 70–73.

Wolpaw, J., Flotzinger, D., Pfurtscheller, G., and McFarland, D. (1997). Timing of EEG-based cursor control. *Journal of Clinical Neurophysiology*, **14**, 529–538.

Zien, A., Rätsch, G., Mika, S., Schölkopf, B., Lengauer, T., and Müller, K.-R. (2000). Engineering support vector machine kernels that recognize translation initiation sites. *Bioinformatics*, **16**(9), 799–807.